Adaptation of Deep Object Detectors for New Modalities

by

Heitor RAPELA MEDEIROS

MANUSCRIPT-BASED THESIS PRESENTED TO ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
IN PARTIAL FULFILLMENT FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
Ph.D.

MONTREAL, NOVEMBER 18, 2025

ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
UNIVERSITÉ DU QUÉBEC

BOARD OF EXAMINERS

THIS THESIS HAS BEEN EVALUATED

BY THE FOLLOWING BOARD OF EXAMINERS

Mr. Marco Pedersoli, Thesis supervisor
Department of Systems Engineering, École de technologie supérieure

Mr. Eric Granger, Thesis Co-Supervisor
Department of Systems Engineering, École de technologie supérieure

Mr. Jose Dolz, Chair, Board of Examiners
Department of Software Engineering and IT, École de technologie supérieure

Mr. Matthew Toews, Member of the Jury
Department of Systems Engineering, École de technologie supérieure

Mr. Angel D. Sappa, External Independent Examiner
Computer Vision Center, Barcelona, Spain
ESPOL Polytechnic University, Guayaquil, Ecuador

THIS THESIS WAS PRESENTED AND DEFENDED

IN THE PRESENCE OF A BOARD OF EXAMINERS AND THE PUBLIC

ON OCTOBER 16, 2025

AT ÉCOLE DE TECHNOLOGIE SUPÉRIEURE

ACKNOWLEDGEMENTS

I would like to thank my wife, Lorenna, for being with me on this journey. Thanks so much for making my path easier; without you, I would not have finished this Ph.D. She knows how hard I worked to finish this Ph.D. and how far we came together. I love you. Thanks for being part of my family and for all the support you have given me on this journey. Lorenna, your journey from a small city in Brazil to becoming a top medical doctor and pediatrician from the best places in our country, studying abroad, and working in hospitals in Canada, inspires me every day to work harder.

Furthermore, I would like to thank my mother, Angela Cristina, who was the first Ph.D. in our family, and as a professor, she inspired me a lot to try this journey. I would also like to thank her for all the investment in my education, and thanks for all the good moments shared with me, thanks for the best mother in the world. Thanks also to my father, Ivan Bezerra, who is the best father in the world, who always did what he could for me and my brothers to receive the best education, from getting us into school, to providing the best food, and to securing us in the bad times. I love my mother and father so much that everything that I did in my life was just to give back what they gave me, from education to all the effort to win. I hope one day I will be able to help them. I would also like to thank my sister Camila, who is the best sister that I could have, and also my brother Guilherme, who is also the best. I love them so much. Thanks also to Penha, my second big mother, and my grandparents (in memoriam), Andres, Angelina, and Severina. They made my childhood so much happier and special. I will always remember you for being such a role model for me.

I would also like to thank Tia Marcilene and Tio Antonio Carlos, mother and father of Lorenna and Lidia (her sister), for all the support and help over the years. Thanks for considering me as part of your family.

I would like to thank my advisors, Marco and Eric, for the experience and opportunity since I joined the team. I learned a lot about how to improve my research skills and engage in a nice

debate to improve and sharpen my ideas and projects. Thanks for all the effort in making this journey a success.

This Ph.D. was a strong collaboration between many people, which I would like to thanks for all the help and effort: First, to Paulo Freitas, which invited me to apply for ETS for the first time to do a Ph.D. in communication and then become a nice friend in Montreal, then to Fidel which invited me to apply for ETS Montreal for a Ph.D. in Computer Vision with Marco, them to Masih which was my closest collaborator and friend, which I complain everyday about how hard is the PhD life, sorry for that, and thanks for all the help... Then, I would like to thank Thomas Dubail, a good friend with whom I enjoyed discussing, and David Latortue for such an amazing collaboration in MiPa. Atif Belal, the master padawan, is a good friend and a nice collaborator. We had nice conversations. Also, thanks to Sri, who started slowly but is making good contributions to our team, and is a nice guy. Moreover, we had great moments together thanks to Simon, Clement, and Moetez, who were part of our team and good friends. Also, thanks to the new members, Kevin, Taha, and Adarley. I hope you guys continue our project with success. Therefore, I would like to thank Distech Controls members, in particular François Gervais, for the interesting meetings and discussions in their detection project that I was part of since the beginning, and for the financial support, as well as Mitacs, for the support in developing the research in this thesis.

I would like to thanks the friends were not involved directly on our project, but which helped me indirectly with fun, motivation, and good food (the order does not mean anything): Thanks to Eduardo, Thiago Chaves, Luiz Zanini, Paulo/Jessica, Bala, Gustavo, Julien, David Osowiechi, Yara, Saypra, Akhil, Marco Colussi, Sahar, Bingyuan, Fereshteh, Shivi, Boshra, Mohammad Hasan, Melike, Banafsheh, Masoumeh, Maxime Darrin, Maxime Zanella, Phillip, Mehrdad, Ali, Moslem, Nicholas, Shakeeb, Soufiane, Rishav, Djebril, Lorenzo, Leandro Ensina, Cristiano Garcia, Elisa, Juscimara, Kercia, Lucas, Mahdi, Farzad, Sina, Julio, Julia, Hugo, professor Rafael, Damien, Malik, Jerome.

Also, I would like to thank the RoboCIn group from Brazil, which is the robotic team that I co-founded. We had a lot of great experiences with my friends Lucas, Bebeto, Carlos, Cris, Gabriel, Renato, Marvson, Juliana, Walber, Gonça, Matheus, Douglas, Zé Junior, Zé Victor, João, Maggi, Felipe, Ceci, Sabino, Johny, Rapha, Pedrinho, Renie, Professor Hans, and Professor Edna, and all the new members who made a lot of impact. Thanks also to my friends from CETENE: Cambuim, Johny, Antonius, Diogo, Rafa Nunes, and Vanessa for all the great research experience and fun moments. From the Sica group, I would like to thank my friends Anderson Urbano, Diogo Rodrigues, Larissa Lages, Gustavo, and others.

Adaptation des détecteurs d'objets basés sur l'apprentissage profond à de nouvelles modalités

Heitor RAPELA MEDEIROS

RÉSUMÉ

Les performances des détecteurs d'objets profonds se détériorent considérablement lorsqu'ils sont déployés sur différentes modalités de capteurs, telles que le RGB, l'infrarouge et la profondeur. Cette dégradation provient du décalage entre les modalités, qui affecte drastiquement les performances des modèles. Les méthodes d'adaptation existantes, telles que la traduction au niveau des pixels ou l'alignement dans l'espace des caractéristiques, sont souvent limitées à des modalités spécifiques ou nécessitent un réentraînement important, entraînant un coût computationnel élevé et une perte potentielle des connaissances acquises auparavant. Contrairement à l'approche dominante consistant à adapter le modèle, nous explorons ici le potentiel d'adapter l'entrée, ou d'apporter des modifications mineures, afin de préserver autant que possible les connaissances préalables tout en incorporant celles propres à la nouvelle modalité. Dans ce contexte, cette thèse étudie des stratégies d'adaptation de modalités visant à combler l'écart entre les modalités tout en maintenant les performances de détection du modèle RGB préentraîné sur la source.

Dans cette thèse, nous présentons d'abord notre démarche et nos contributions. Ensuite, dans le premier chapitre, nous fournissons un cadre général permettant de comprendre les différentes stratégies abordées dans cette thèse, avec divers mécanismes utilisés pour adapter les détecteurs d'objets, allant du niveau d'entrée (modification de l'image) au niveau intermédiaire (mécanismes dans les épines dorsales) et au niveau de sortie (adaptation des boîtes ou modifications pseudo-niveau). Dans le deuxième chapitre, nous étudions comment incorporer la connaissance de deux modalités différentes dans un seul encodeur partagé et agnostique à la modalité pour les détecteurs, de manière efficace et performante. Ensuite, dans le troisième chapitre, nous explorons une adaptation progressive des modalités : d'abord, l'adaptation des connaissances du détecteur à partir des données RGB sources (par exemple, le jeu de données COCO) vers un jeu de données RGB cible (par exemple, LLVIP RGB), puis l'adaptation de l'entrée à partir des données infrarouges (par exemple, LLVIP IR) vers une représentation pseudo-RGB à l'aide du retour du détecteur. Dans le quatrième chapitre, nous nous concentrons sur l'adaptation de la modalité au niveau de l'entrée, en préservant les connaissances du modèle préentraîné sur la source (par exemple, COCO) et en l'adaptant directement au jeu de données infrarouges (par exemple, LLVIP IR), sans l'étape intermédiaire du chapitre précédent, tout en cherchant à maximiser la performance de détection et à conserver les capacités zéro-shot de la source. Dans le cinquième chapitre, nous explorons comment intégrer le langage dans l'adaptation de la modalité d'entrée pour les détecteurs d'objets visuel-langage ; notre objectif était donc de préserver la connaissance zéro-shot du détecteur tout en comprenant comment intégrer des techniques puissantes d'adaptation de modalité visuelle, combinées à l'adaptation de prompts.

Nos principales contributions incluent : pour le deuxième chapitre, nous introduisons MiPa, une stratégie d'entraînement par patchs mixtes pour les détecteurs d'objets basés sur des

transformeurs, permettant à un encodeur partagé d'être agnostique aux modalités RGB et infrarouge. MiPa échantillonne et combine de manière stochastique des patchs complémentaires RGB/IR pendant l'entraînement, capturant efficacement l'information intermodale sans nécessiter les deux modalités à l'inférence. Dans le quatrième chapitre, nous introduisons ModTr, un cadre de traduction de modalités pour adapter les détecteurs d'objets RGB préentraînés à de nouvelles modalités, telles que l'infrarouge (IR), sans modifier les paramètres du détecteur. ModTr préserve les connaissances originales du détecteur, permettant à un seul modèle de gérer plusieurs modalités via des traducteurs dédiés, réduisant ainsi les coûts de mémoire et de calcul. ModTr introduit des stratégies de fusion simples mais efficaces, telles que la fusion basée sur le produit d'Hadamard, pour combiner les entrées traduites et originales. Dans le cinquième chapitre, nous introduisons ModPrompt, un cadre basé sur des prompts visuels pour adapter les détecteurs d'objets à vocabulaire ouvert (OV-OD) à de nouvelles modalités visuelles, telles que l'infrarouge, la profondeur et le LiDAR, sans compromettre leurs capacités zéro-shot. Contrairement aux stratégies de prompts au niveau des pixels utilisées en classification, ModPrompt emploie un module de prompt visuel encodeur-décodeur qui génère des prompts spécifiques à chaque image. Il propose également les Modality Prompt Decoupled Residuals (MPDR), qui améliorent l'adaptation en introduisant des paramètres résiduels légers et compatibles avec l'inférence, permettant un alignement entre modalités sans perte des connaissances langagières préentraînées. Enfin, dans la dernière partie de cette thèse, nous présentons une conclusion générale sur la manière de tirer parti des connaissances RGB préentraînées des détecteurs tout en les adaptant à de nouvelles modalités, ainsi que des recommandations pour les travaux futurs.

Mots-clés: Détection d'objets, Adaptation de modalités, Décalage de domaine, Modèles vision–langage, Information privilégiée, Apprentissage inter-modalités.

Adaptation of Deep Object Detectors for New Modalities

Heitor RAPELA MEDEIROS

ABSTRACT

The performance of deep object detectors significantly deteriorates when deployed across different sensing modalities, such as RGB, infrared, and depth. This degradation arises from the shift between modalities, which drastically affects the performance of the models. Existing adaptation methods, such as pixel-level translation and feature-space alignment, are often limited to specific modalities or require extensive retraining, leading to increased computational cost and potential loss of previously acquired knowledge. In contrast to the dominant approach of adapting the model, here we investigated the potential of adapting the input, or making minor changes, preserving as much prior knowledge while incorporating new modality knowledge. In this context, this thesis studies modality adaptation strategies to bridge the gap between modalities while preserving detection performance on the source pre-trained RGB model.

In this thesis, we first introduce our search and contributions. Then, in the first chapter, we provided a general background to understand the current different strategies presented in this thesis with different mechanisms used to adapt object detectors, ranging from input-level (image modification) to middle-level (mechanisms in backbones) to output-level (adaptation of boxes or pseudo-level modifications). In the second chapter, we study how to incorporate knowledge of two different modalities in a single modality-agnostic shared encoder for detectors in an efficient and powerful way. Then, in the third chapter, we explore progressive modality adaptation, first adapting the detector knowledge from the RGB source data (e.g., COCO dataset) to the target RGB dataset (e.g., LLVIP RGB) and then adapting the input from IR data (e.g., LLVIP IR) to a pseudo-RGB representation with this detector feedback. In the fourth chapter, we focused on input modality adaptation, preserving the knowledge of the source pre-trained model (e.g, COCO dataset) and adapting directly to the IR dataset (e.g., LLVIP IR), without the intermediate step of the prior chapter, and focusing on maximizing the detection performance while keeping the source zero-shot knowledge. In the fifth chapter, we explored how to incorporate language in the input modality adaptation for visual-language object detectors; therefore, our goal was still to preserve zero-shot knowledge of the detector, but also to understand how to incorporate powerful visual modality adaptation techniques, along with prompt adaptation.

Our main contributions include: for the second chapter, we introduced MiPa, a mixed-patch training strategy for transformer-based object detectors that enables a single shared encoder to be modality-agnostic to RGB and infrared inputs. MiPa stochastically samples and combines complementary RGB/IR patches during training, effectively capturing cross-modal information without requiring both modalities at inference. In the fourth chapter, we introduced ModTr, a modality translation framework for adapting pre-trained RGB object detectors to new modalities, such as infrared (IR), without changing the detector's parameters. ModTr preserves the detector's original knowledge, enabling a single model to serve multiple modalities through dedicated translators, reducing memory and computation costs. ModTr introduces simple yet effective

fusion strategies, such as the Hadamard product–based gating, to blend the translated and original inputs. In the fifth chapter, we introduced ModPrompt, a visual prompt–based framework for adapting open-vocabulary object detectors (OV-ODs) to new visual modalities, such as infrared, depth, and LiDAR, without compromising their zero-shot capabilities. Unlike pixel-level prompt strategies used in classification, ModPrompt employs an encoder–decoder visual prompt module that generates modality-specific prompts tailored to each input image. It further proposes Modality Prompt Decoupled Residuals (MPDR), which enhance adaptation by introducing lightweight, inference-friendly residual parameters, enabling modality alignment without losing pre-trained language knowledge. Finally, in the last part of this thesis, we provided an overall conclusion of our thesis and how we can leverage pre-trained RGB knowledge of detectors while we adapt to new modalities and recommendations for future work.

Keywords: Object Detection, Multimodal Learning, Modality Adaptation, Vision-Language Models, Privileged Information, Cross-Modality Learning.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AP	Average Precision
BBOX	Bounding Box
CLIP	Contrastive Language–Image Pretraining
COCO	Common Objects in Context
CV	Computer Vision
DDAOD	Deep Domain Adaptive Object Detection
DL	Deep Learning
DPM	Deformable Part-based Model
FLIR	Forward Looking Infrared
FPN	Feature Pyramid Networks
FPS	Frames Per Second
GAN	Generative Adversarial Network
GPU	Graphical Processing Unit
GT	Ground Truth
HoG	Histogram of Gradients
IoU	Intersection over Union
LLVIP	Low-Light Visible-Infrared Paired Dataset
LoRA	Low-Rank Adaptation
LUPI	Learning Using Privileged Information

MAE	Masked Autoencoder
MiPa	Mixed Patches
ModPrompt	Modality Prompt
ModTR	Modality Translation
MPDR	Modality Prompt Decoupled Residual
NMS	Non-Maximum Suppression
NYUv2	New York University Depth Dataset V2
RCNN	Regions with CNN
RGB	Red, Green, Blue (Visible Spectrum Image)
SPPNet	Spatial Pyramid Pooling Net
SSD	Single Shot MultiBox Detector
VGG	Visual Geometry Group
VLM	Vision-Language Model
WST	Weak Self-Training
YOLO	You Only Look Once

INTRODUCTION

In recent years, deep learning models have significantly outperformed traditional methods across a wide range of domains, including computer vision (Krizhevsky, Sutskever & Hinton, 2012), natural language processing (Vaswani *et al.*, 2017), and others. In particular, visual object detection (OD) has seen rapid progress, fueled by advancements in convolutional neural networks (CNNs) and, more recently, transformer-based architectures. These developments have enabled accurate and real-time detection capabilities, driving adoption in applications such as autonomous driving (Huang & Chen, 2020), video surveillance (Zhang *et al.*, 2018a; Chen, Li, Deng, Li & Yu, 2019a), and smart building environments (Dubail *et al.*, 2022).

Despite this progress, these models are primarily trained on a specific type of visual input modality, RGB images, which are acquired from standard RGB sensors. Natural images are typically acquired using those RGB sensors, which have become the dominant modality in computer vision due to the availability of large-scale annotated datasets, such as ImageNet (Deng *et al.*, 2009) for image classification and COCO (Lin *et al.*, 2014) for object detection. These datasets have served as standard benchmarks and have played a central role in the development and evaluation of deep learning models. Notably, many models were conceived in the context of these challenges with RGB images, including AlexNet (Krizhevsky *et al.*, 2012) for the ILSVRC 2010, VGG (Simonyan & Zisserman, 2014) in the 2014 ImageNet Challenge, and ResNet (He, Zhang, Ren & Sun, 2016), which achieved top performance in both the ILSVRC 2015 and COCO 2015 competitions. Architectures such as VGG and ResNet continue to serve as important components in modern deep learning pipelines, often used as backbone feature extractors for a wide range of vision tasks.

Alongside CNNs, more recently, the backbones such as vision transformer (Dosovitskiy *et al.*, 2020) and Swin transformer (Liu *et al.*, 2021b) have opened the possibility to advance the performance of detectors (e.g., DETR (Carion *et al.*, 2020), DINO (Zhang *et al.*, 2023c)) with

the scale of data. Transformer-based detectors are known to be data-hungry compared with CNN-based detectors, and plenty of novel detectors are targeting to speed up such detectors to make them as fast as CNN-based (Zhao *et al.*, 2024b). Nonetheless, a common thread across both CNN- and Transformer-based models is their strong reliance on RGB data for both pre-training and evaluation, and in fact, the vast majority of these methods have been developed and evaluated primarily on RGB images, relying on large-scale RGB datasets for pre-training and assuming consistent access to the same modality during deployment.

Challenge of Modality Adaptation. Adapting object detectors to diverse modalities, such as IR, depth, LiDAR, remains a challenging task due to the large distribution shift, where the data distribution seen during training (typically RGB) differs substantially from that at test time (Medeiros *et al.*, 2024c,b; Medeiros, Belal, Muralidharan, Granger & Pedersoli, 2024a). These shifts are especially problematic when labeled data in the target modality is scarce, when compared with the amount of data for a common RGB pre-training. Moreover, traditional adaptation methods often can lead to overfitting on the target modality while suffering from catastrophic forgetting of the source knowledge. Additionally, naive cross-modal application of detectors results in degraded performance, as models trained on RGB fail to generalize to the significantly different feature statistics of non-RGB modalities (Park, Vien & Lee, 2023).

Several strategies have emerged to address these challenges. *Unsupervised domain adaptation (UDA)* techniques aim to transfer knowledge from labeled source data to unlabeled target domains through pseudo-labeling, adversarial training, or feature alignment (Ganin & Lempitsky, 2015). At the same time, *image translation methods* based on generative models like CycleGAN (Zhu, Park, Isola & Efros, 2017a) or Pix2Pix (Isola, Zhu, Zhou & Efros, 2017b) have been employed to convert RGB images into target-style modalities for training. However, these methods often prioritize visual realism over task-specific utility, failing to directly optimize for detection performance.

Emergence of Vision-Language Models (VLMs). Recently, vision-language models (VLMs) such as CLIP (Radford *et al.*, 2021), Grounding DINO (Liu *et al.*, 2024a), and YOLO-World (Cheng *et al.*, 2024) have shown remarkable zero-shot capabilities by leveraging joint image-text embeddings. These models allow open-vocabulary detection, which can extend the classes the detector was trained for, and enable the localization of novel object categories without explicit supervision (Liu *et al.*, 2024a). However, their performance degrades significantly when applied to unseen visual modalities (e.g., IR or depth) since they are predominantly trained on RGB datasets (Medeiros *et al.*, 2024a). Preserving the generalization power of VLMs under modality shift remains an open problem.

Despite the rapid progress of deep learning, most object detectors remain restricted to RGB images, which limits their applicability in real-world conditions where other sensing modalities—such as infrared, depth, or LiDAR—are essential. This constraint prevents robust perception in low-light, safety-critical, or privacy-sensitive scenarios, such as surveillance, autonomous driving, or smart building environments. The main motivation of this thesis is to bridge the gap between pre-trained RGB-based models and real-world multimodal perception systems. By developing efficient and generalizable adaptation strategies, we aim to enable pre-trained detectors to operate reliably across diverse modalities, ensuring scalability, robustness, and reduced data dependency for deployment in heterogeneous environments.

Motivation: Why modality adaptation matters in the real world. While RGB images dominate in academic datasets and benchmarks, many real-world applications operate in environments where RGB sensing is suboptimal or even infeasible. For instance, infrared (IR) cameras are widely deployed at night-time or in low-lighting conditions in surveillance (Luo, Remillard & Hoetzer, 2010), military (Fokkinga *et al.*, 2025) applications, and in privacy-sensitive environments (Dubail *et al.*, 2022) such as hospitals or homes. IR imagery provides complementary information to RGB (Wang *et al.*, 2022), particularly under poor illumination

or adverse weather. Similarly, depth sensors and thermal cameras are increasingly used in autonomous driving (Li & Ibanez-Guzman, 2020) and robotics (Yang *et al.*, 2022a), where robust perception under varying environmental conditions is essential. However, collecting large-scale annotated datasets in these alternative modalities is often expensive, time-consuming, and sometimes impractical due to hardware limitations or privacy concerns.

This reality creates a fundamental tension: most state-of-the-art ODs, such as YOLO-World and Grounding DINO, are trained on RGB datasets (e.g., COCO, ImageNet) and cannot generalize well to non-RGB modalities. Training modality-specific detectors from scratch is often not viable due to the lack of large pre-training data on different modalities, and fine-tuning existing RGB detectors on small target-domain datasets typically leads to overfitting and catastrophic forgetting. As a result, there is a critical need for methods that can *efficiently adapt an OD pre-trained on RGB images to new modalities* while preserving their learned knowledge and ensuring strong performance in data-scarce conditions. This thesis tackles this challenge by developing scalable, detection-driven adaptation strategies tailored for both traditional detectors and modern vision-language detectors across multiple sensing modalities.

0.1 Problem statement

Thesis Objective and Contributions. In this thesis, we address the general problem of *modality adaptation for deep OD*. The main objective is to develop cost-effective deep OD models for modality adaptation, ensuring robustness across sensing modalities while minimizing computational overhead, exploring previous knowledge, and preserving detection accuracy in real-world deployment scenarios. We propose a suite of methods to enhance the performance of OD pretrained on RGB images to adapt to new modalities.

1. Design adaptation strategies that transfer knowledge from RGB-trained detectors to new modalities such as infrared, depth, or LiDAR, using minimal additional data.

2. Preserve detection accuracy on the source modality while improving robustness and generalization in the target modality.
3. Explore complementary adaptation mechanisms—input-level, representation-level, and prompt-based conditioning—without retraining or altering the base detector.
4. Demonstrate the practicality and scalability of these methods across various architectures and real-world sensing environments.

0.2 Contributions

As outlined above, this thesis focuses on modality adaptation of deep OD. Before presenting the contributions, we briefly situate them with respect to prior work. In Chapter 1, we reviewed the main families of object detection methods, including one-stage, two-stage, and vision-language detectors (Section 1.2). We also described deep domain adaptive OD (Section 1.3), highlighting approaches such as discrepancy-, adversarial-, reconstruction-based, and hybrid methods. In addition, we discussed methods using privileged information (Section 1.4), the challenges of adapting RGB-trained detectors to infrared benchmarks (Section 1.5), and strategies in multimodal learning (Section 1.6). These works reveal key challenges such as the reliance on large-scale annotated data, the difficulty of handling missing modalities, and the limited generalization of RGB-centric models. Therefore, we introduced novel methods for adapting deep OD to different modalities, challenging the previous approaches. The main contributions of this thesis are divided into four chapters:

- **Chapter 2**: We propose MiPa (Mixed Patches), a novel modality-agnostic training strategy for RGB/IR object detection that enables a single transformer-based encoder to generalize across both modalities while requiring only one during inference. Instead of relying on dual-modality fusion or modality-specific branches, MiPa introduces a stochastic patch mixing mechanism that combines RGB and IR patches into a single image-level mosaic during training. This is coupled with a patch-wise modality-agnostic (MA) module that

suppresses modality imbalance by encouraging indistinguishable representations across modalities. Our method improves generalization, avoids inference overhead, and achieves state-of-the-art results on standard RGB/IR detection benchmarks.

Related publication:

- Mixed patch visible-infrared modality agnostic object detection. **Medeiros, H. R.**, Latortue, D., Granger, E., & Pedersoli, M. (2025, February). In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)* (pp. 9023-9032). IEEE.

Authorship: First author with equal contribution as David Latortue. **Roles:** Contributed to the method inception, implementation, and validation, as well as writing the paper.

- **Chapter 3**: We introduce HalluciDet, a novel modality adaptation framework that performs IR-to-RGB image translation guided by detection performance, not pixel loss. Unlike classical image translation methods that reconstruct RGB images from IR inputs using pixel-level losses, HalluciDet leverages an RGB detector and formulates a hallucination loss based solely on object detection objectives. This design allows the model to exploit the privileged information encoded in the RGB detector, generating a hallucinated representation that enhances detection of IR images without requiring RGB data at test time. HalluciDet outperforms both traditional fine-tuning and state-of-the-art image translation techniques.

  Related publication:

  - HalluciDet: hallucinating RGB modality for person detection through privileged information. **Medeiros, H. R.**, Pena, F. A. G., Aminbeidokhti, M., Dubail, T., Granger, E.,& Pedersoli, M. (2024). In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (pp. 1444-1453).

Authorship: First author. **Roles:** Contributed to the method inception, implementation, validation, and paper writing.

- **Chapter 4**: We propose ModTr, a modality translation framework for adapting pre-trained RGB object detectors to infrared (IR) inputs without modifying their parameters or accessing RGB data. Unlike traditional fine-tuning methods that often erase prior knowledge, ModTr preserves the original detector's capabilities by learning a lightweight image-to-image translation network optimized purely for detection performance. This network transforms IR images into RGB-like representations through detection-driven training, allowing the frozen detector to operate seamlessly across modalities. We introduce multiple fusion strategies—including Hadamard product-based gating—to enhance representation alignment and enable efficient server-side deployment. ModTr consistently outperforms conventional fine-tuning and translation approaches on standard benchmarks, demonstrating strong generalization and robustness across architectures and datasets

  Related publication:

  - Modality translation for object detection adaptation without forgetting prior knowledge. **Medeiros, H. R.**, Aminbeidokhti, M., Peña, F. A. G., Latortue, D., Granger, E., & Pedersoli, M. (2024, September). In *European Conference on Computer Vision* (pp. 51-68). Cham: Springer Nature Switzerland.

  Authorship: First author. **Roles:** Contributed to the method inception, implementation, validation, and paper writing.

- **Chapter 5**: We introduce ModPrompt, a novel prompt-based adaptation method for open-vocabulary object detectors like YOLO-World and Grounding DINO. Unlike existing approaches that retrain large vision-language models (VLMs) or traditional adapters,

ModPrompt enables efficient adaptation to new input modalities such as infrared (IR) and depth through visual conditioning and embedding adaptation. Specifically, we propose a prompting mechanism that includes visual prompts injected into the image and residual embeddings appended to the embedded object categories. These prompts are trained to align multi-modal features without altering the base detector, preserving zero-shot capabilities. Our method is parameter-efficient, requires no paired training data, and significantly improves detection performance under modality shifts while maintaining scalability and compatibility with existing VLM architectures.

Related publication: Visual Modality Prompt for Adapting Vision-Language Object Detectors. **Medeiros, H. R.**, Belal, A., Muralidharan, S., Granger, E., & Pedersoli, M. (2024). Published in *International Conference on Computer Vision (ICCV)*, 2025.

Authorship: First author. **Roles:** Contributed to the method inception, implementation, validation, and paper writing.

Summary. This thesis introduces a series of methods, each designed to address specific challenges in modality adaptation for object detection (OD). The first work, MiPa, introduces new modality knowledge, such as IR, to pre-trained RGB transformer-based detectors without introducing any new parameters during inference. Then, the other chapters (HalluciDet, ModTr, and ModPrompt) bridge image translation, detection-driven optimization, and prompt-based adaptation for vision-language models. Collectively, the contributions presented in this thesis demonstrate that robust cross-modality adaptation can be achieved without full retraining, access to large-scale source data, or architectural modification. By combining detection-driven optimization, modality-invariant representation learning, and parameter-efficient prompting, this thesis establishes a comprehensive analysis of how we can improve adaptation of modern detectors to heterogeneous sensing conditions—paving the way for practical and generalizable multimodal perception.

0.2.1 Additional contributions

In addition to the main contributions where I was the first author, four other contributions were made during this thesis:

- WiSE-OD: Benchmarking Robustness in Infrared Object Detection. **Medeiros, H. R.**, Belal, A., Aminbeidokhti, M., Granger, E., & Pedersoli, M. (2025). Submitted to *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2026.

  Authorship: First author. **Roles**: Contributed to the method inception, implementation, and validation, as well as writing the paper.

- Re-basin via implicit sinkhorn differentiation, Peña, F. A. G., **Medeiros, H. R.**, Dubail, T., Aminbeidokhti, M., Granger, E., & Pedersoli, M. (2023). Published in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 20237-20246).

  Authorship: Second author. **Roles**: Contributed to discussions, validation, as well as writing the paper.

- Domain generalization by rejecting extreme augmentations. Aminbeidokhti, M., Pena, F. A. G., **Medeiros, H. R.**, Dubail, T., Granger, E., & Pedersoli, M. (2024). Published in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (pp. 2215-2225).

  Authorship: Third author. **Roles**: Contributed to discussions, as well as writing the paper.

- Privacy-preserving person detection using low-resolution infrared cameras. Dubail, T., Guerrero Peña, F. A., **Medeiros, H. R.**, Aminbeidokhti, M., Granger, E., & Pedersoli, M. (2022, October). Published in Real-World Surveillance: Applications and Challenges Workshop at *European Conference on Computer Vision* (pp. 689-702). Cham: Springer Nature Switzerland.

  Authorship: Third author. **Roles**: Contributed to discussions and wrote the paper.

0.2.2 External collaboration

This doctoral project was accompanied by two different external collaborations:

Research chair ÉTS - Distech Controls (2021-2025): This PhD project was co-funded from beginning to end through a Mitacs-Alliance grant with the Distech Controls Industrial Research Chair at ÉTS Montreal. The Chair focuses on developing advanced embedded AI techniques that enable smarter, more efficient, and privacy-preserving building environments. My research contributed to this mission by exploring robust object detection and modality adaptation strategies applicable to detecting objects under different sensors for smart building applications. This partnership provided valuable insights into real-world deployment constraints, fostering research with tangible industrial impact.

Machine Learning Researcher Intern - RBC Borealis (Jan 2025 – Apr 2025): During this internship, we developed methods for Test-Time Adaptation (TTA) in Time-Series Forecasting. This experience provided valuable cross-domain perspectives and underscored the broader relevance and applicability of the research conducted in this thesis, particularly in demonstrating its potential impact across diverse industry sectors.

CHAPTER 1

BACKGROUND

In this chapter, we provide the necessary background to contextualize the contributions of this thesis and connect them with the challenges addressed in the following chapters. We begin with an overview of *Multimodal image translation* (Section 1.1), which introduces the problem of distribution shifts between training and deployment modalities. Although many approaches exist at the image level, they are often limited in their ability to generalize across modalities, motivating the need for more flexible adaptation strategies explored later in this thesis. We then describe the foundations of *Object detection* (Section 1.2) based on deep learning, including one-stage detectors, two-stage detectors, and vision-language object detectors. These models form the backbone of most detection frameworks, yet they are typically designed and benchmarked only on RGB data, leaving a gap when transferring to modalities such as infrared or depth. During this thesis, we focus on deep object detectors, which are based on deep learning features. Previous detector methods were based on conventional machine learning handcrafted features with pattern matching, but these methods, although important for the field, are not within the scope of this thesis.

Next, we focus on *Deep domain adaptive object detection* (Section 1.3), which discusses mechanisms to mitigate domain shifts through adversarial learning and multi-step adaptation. While effective, these methods often require extensive retraining or large-scale annotated data, highlighting the need for lightweight and knowledge-preserving strategies such as those proposed in this work. We also review *Object detection using infrared images* (Section 1.5), which emphasizes the practical relevance of IR for applications like surveillance and autonomous driving, while exposing the difficulty of adapting RGB-trained detectors to IR benchmarks such as LLVIP and FLIR. This motivates the development of methods like HalluciDet and MiPa that specifically target modality adaptation.

Building on this, we present concepts of *Multimodal learning* (Section 1.6), which address the integration of complementary modalities, handling missing modalities, and designing effective

fusion strategies. Despite progress in this area, most existing approaches assume the availability of all modalities at inference, a limitation that this thesis tackles with methods such as ModTr and ModPrompt. Finally, we introduce the *Evaluation methodologies* (Section 1.7), which not only serve as the basis for evaluating our proposed methods, but also reveal gaps in current benchmarks and metrics.

1.1 Multimodal image translation

The field of multimodal image translation has grown with the advances in techniques like autoencoders (AEs) (Hinton & Zemel, 1993) and generative adversarial networks (GANs) (Goodfellow *et al.*, 2014). The objective of multimodal image translation is to adapt the image from the source modality to be closer to the image on the target modality to be used on the model initially trained on the target modality or to train a model without directly using the target modality as input, but, e.g., as additional information on the learning process.

Some techniques use supervision from the target domain samples to adapt the task model, also known as finetuning on the target domain. Unsupervised methods do not directly use the target domain samples to do finetuning but produce additional information for the machine learning technique, e.g., pseudo-labels, to guide the task model adaptation. The vast majority of methods of image-level domain adaptation are unsupervised. We will describe some techniques on Section 1.3, which are considered hybrid techniques that adapt to the image level and also the model level. This section will focus on pure image-level techniques (AEs and GANs).

Autoencoders (AE). AEs are a kind of artificial neural network that is used for dimensionality reduction and feature learning. During training, the AE aims to minimize the reconstruction error, i.e., the error between its input and the respective output, to produce an output close to the input. The learning phase of the AE does not require labels, only the input itself, so it is considered an unsupervised learning method. The hidden layer in the middle is usually constrained to be a narrow bottleneck producing a latent space. The system can minimize

the reconstruction error by ensuring the hidden units capture the most relevant aspects of the data (Murphy, 2012).

The AE consists of two parts: an encoder function ($f(x)$), in which x is the input data, that maps the input to the latent space and a decoder ($g(x)$) that produces a reconstruction $g(f(x)) = \hat{x}$, in which $\hat{x}$ is the reconstructed input data. At the end of the learning process, the AE learns to approximate the input to itself ($g(f(x)) \approx x$). The AE loss function is described with the following Equation 1.1:

$$\mathcal{L}(x, g(f(x))) = \|x - g(f(x))\|_2^2, \tag{1.1}$$

in which x is the input, $f(x)$ is the encoder, $g(x)$ is the decoder. When the AE has a constraint in which the latent dimension is smaller than the input x dimension, it forces the AE to learn the more important and useful input features to encode it. A fully connected AE is illustrated in Figure 1.1.

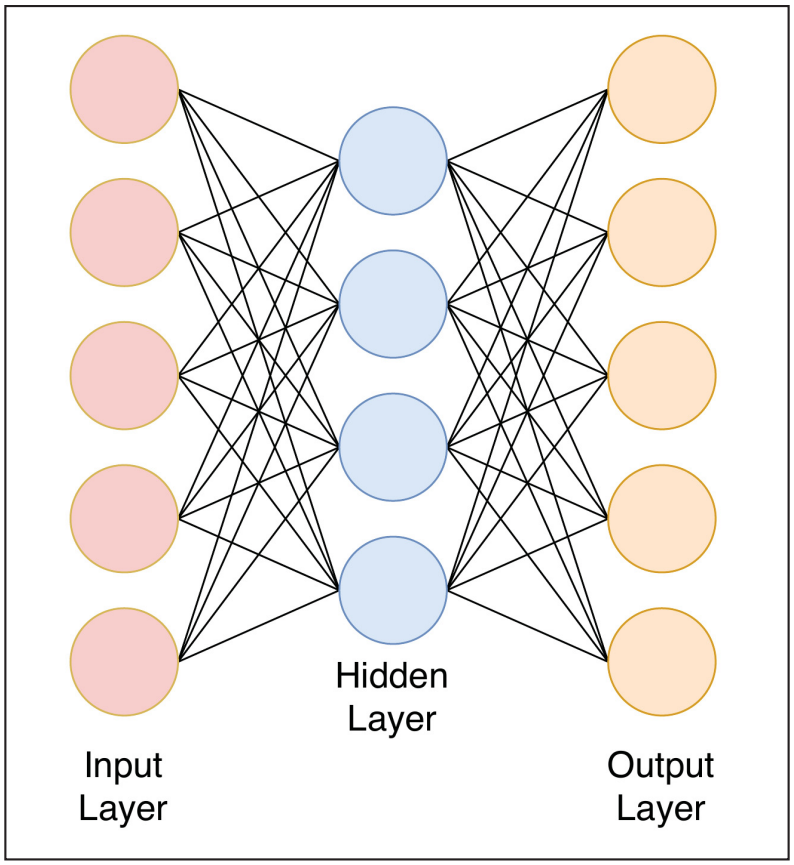


Figure 1.1 A simple example of AE. The neurons in red represent the input layer, the neurons in blue represent the hidden layer, and the neurons in orange represent the output layer

Generative Adversarial Networks (GANs). GANs are generative modeling approaches based on differentiable generator networks (Goodfellow *et al.*, 2014). The GAN is trained using the minimax game, in which there are two networks: the generator and the discriminator. The generator wants to produce fake samples similar to the input, while the discriminator wants to judge if the fake images are real or not. The minimax game for training a GAN is described by the following Equation 1.2:

$$\min_G \max_D V(D,G) = \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}[\log D(\boldsymbol{x})] + \mathbb{E}_{z\sim p_z(z)}[\log(1 - D(G(\boldsymbol{z})))], \tag{1.2}$$

in which D is the discriminator, G is the generator and $\mathbb{E}$ is the expected value, $\boldsymbol{x}$ is the real input set, and $\boldsymbol{z}$ is the generated input (fake input) set. The discriminator D is trained to maximize the probability of assigning the correct label for the samples from the real data x and the samples from G, while the generator G is trained to minimize it. Additionally, we can have other GANs such as BicycleGAN (Zhu *et al.*, 2017c). BicycleGAN is a bidirectional image-to-image translation framework combining a conditional VAE and a conditional latent regressor to generate diverse outputs while mitigating mode collapse. It employs a U-Net generator and PatchGAN discriminators trained with least-squares loss for stability. Extending this idea, CEGAN (Xiong, Wang & Gao, 2019) enforces tight consistency between latent and image spaces to model multiple modes and enhance visual realism.

Style Transfer. Style transfer aims to render an image in a particular artistic or domain-specific style while preserving its structural content. Early approaches relied on texture synthesis and non-photorealistic rendering (NPR) to manually reproduce stylistic effects such as brush strokes or shading. However, these methods were limited in flexibility and often constrained to a single predefined artistic domain (Jing *et al.*, 2019). The seminal work of Gatys, Ecker & Bethge (2016) revolutionized the field by introducing Neural Style Transfer (NST), which exploits convolutional neural networks (CNNs) to decouple and recombine content and style representations. Their Gram matrix–based formulation enabled transferring texture statistics from artworks to photographs, inspiring a vast range of follow-ups focused on efficiency, generalization, and perceptual realism. Image-optimization-based methods such as Gatys *et al.* (2016) produce high-quality stylizations but are computationally expensive. In contrast, model-optimization-based methods like Johnson, Alahi & Fei-Fei (2016) and Ulyanov, Vedaldi & Lempitsky (2016) train feed-forward networks for real-time stylization, while Adaptive Instance Normalization (AdaIN) by Huang & Belongie (2017) and subsequent variants allow flexible multi-style and arbitrary-style transfer (Cai, Ma, Wang & Li, 2023). Recent surveys highlight the field's rapid expansion beyond artistic stylization. According to Cai *et al.* (2023), neural style transfer has extended to industrial and scientific domains such as medicine, film, and architecture, driven by GAN- and Transformer-based frameworks that integrate attention mechanisms for improved local detail and global coherence.

Collectively, these advances have transformed style transfer into a general paradigm for cross-domain visual adaptation, bridging aesthetic, semantic, and modality gaps in both creative and technical applications.

In this section, we discussed AEs, GANs, and style transfer. Many variations of these methods can be used for image-to-image translation, as we mentioned previously, e.g., CycleGAN is a kind of GAN that can be used for this purpose. The methods that we will explore as part of the research direction will be discussed in the research direction section.

1.2 Object detection

In object detection (OD), we assume that there are often multiple objects in the image of interest. Different from simple classification tasks, in which we want only to recognize the object category, in object detection, we additionally want to know specific positions of the objects (Zhang *et al.*, 2023a). For the following definition, we consider only traditional visual RGB detectors, which are normally trained on large RGB datasets and do not include language mechanisms. The specific positions of each object are represented by a rectangle that contains the detected object. This rectangle receives the name of bounding box, which is determined by the x and y coordinates of the upper-left corner and lower-right corner of the rectangle (other representations like coordinates of the center and width and height of the box are also common). In machine learning, the training data contains the category of the object plus the bounding box positions, which we call ground-truth and are used to guide the training process. In the next part, we described in detail the problem definition.

Problem definition. The training set for OD is denoted as $\mathcal{D} = \{(x, Y)\}$, where $x \in \mathbb{R}^{W \times H \times C}$ represents an image in the dataset, with dimensions $W \times H$ and C channels. Subsequently, the OD model aims to identify N regions of interest within these images, denoted as $Y = \{(b_i, c_i)\}_{i=1}^{N}$. The top-left corner coordinates and the width and height of the object define each region of interest b_i. Additionally, a classification label c_i is assigned to each detected object, indicating its corresponding class within the dataset. In terms of optimization, the primary goal of the

detection task is to maximize detection accuracy, often measured using the average precision (AP) metric across all classes. An OD is formally represented as the mapping $f_\theta : \mathbb{R}^{W\times H\times C} \to \hat{Y}$, where θ denotes the parameter vector. To effectively train a detector, a differentiable surrogate for the AP metric, referred to as the detection cost function, $C_{det}(\theta)$, is employed. The typical structure of such a cost function involves computing the average detection loss over the dataset $\mathcal{D}$, denoted as $\mathcal{L}_{det}$, described as:

$$C_{det}(\theta) = \frac{1}{|\mathcal{D}|} \sum_{(x,Y)\in\mathcal{D}} \mathcal{L}_{det}(f_\theta(x), Y). \tag{1.3}$$

Categorization and review of detectors. Most of the early object detectors on computer vision were built based on handcrafted features (Zou, Chen, Shi, Guo & Ye, 2023). Viola-Jones Detector achieved real-time detection of human faces for the first time using the integral image with manually selected Haar filters and detection cascades (Dalal & Triggs, 2005). Additionally, we have the Histogram of Gradients (HoG) feature descriptor, which is considered an improvement of scale-invariant feature transform and shape contexts, and was primarily used for pedestrian detection. Furthermore, it is important to remember the Deformable Part-based Model (DPM) that was originally proposed by Felzenszwalb, McAllester & Ramanan (2008). The idea of DPM is that an object can be decomposed into its parts during the training, and the inference can consider the ensemble of detections on these different object parts. Some ideas from these works inspired many of the new deep learning methods. With the advances in deep learning, we can currently categorize the deep learning object detection methods into two main groups: Two-stage detectors and one-stage detectors. More recently, vision-language object detections were proposed, which we also described in more detail in the following sections.

1.2.1 One-stage detectors

The one-stage detectors focus on an end-to-end approach for object detection with deep learning. Here, these methods are visual object detectors that do not include a language mechanism.

In this scenario, the object detector has a single neural network to extract the features for the regression of the bounding box and give the class probabilities without an auxiliary network for the region proposals. The following object detectors described are the major one-stage deep learning object detectors.

You Only Look Once (YOLO). The YOLO proposed by Redmon *et al.* (2016) was the first deep learning one-stage object detector, see Figure 1.2.

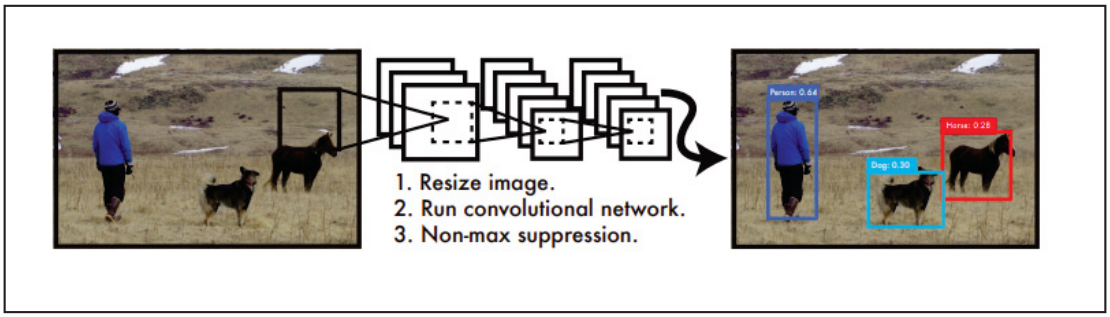


Figure 1.2 The YOLOv1 model (Redmon *et al.*, 2016). The YOLO model resizes the input image to 448×448, then runs a single convolutional network on the image, and thresholds the resulting detections by the model's confidence
Taken from Redmon *et al.* (2016)

The main idea of YOLO is to divide the image into regions that are given as input to a neural network to predict the probabilities and classes of the objects in a single forward pass, without the usage of region proposals, see Figure 1.3. The YOLO model is known for its detection speed. The subsequent versions of the model (Redmon & Farhadi, 2017, 2018) focused on improving the detection of a small object because the two-stage detectors were better in that scenario.

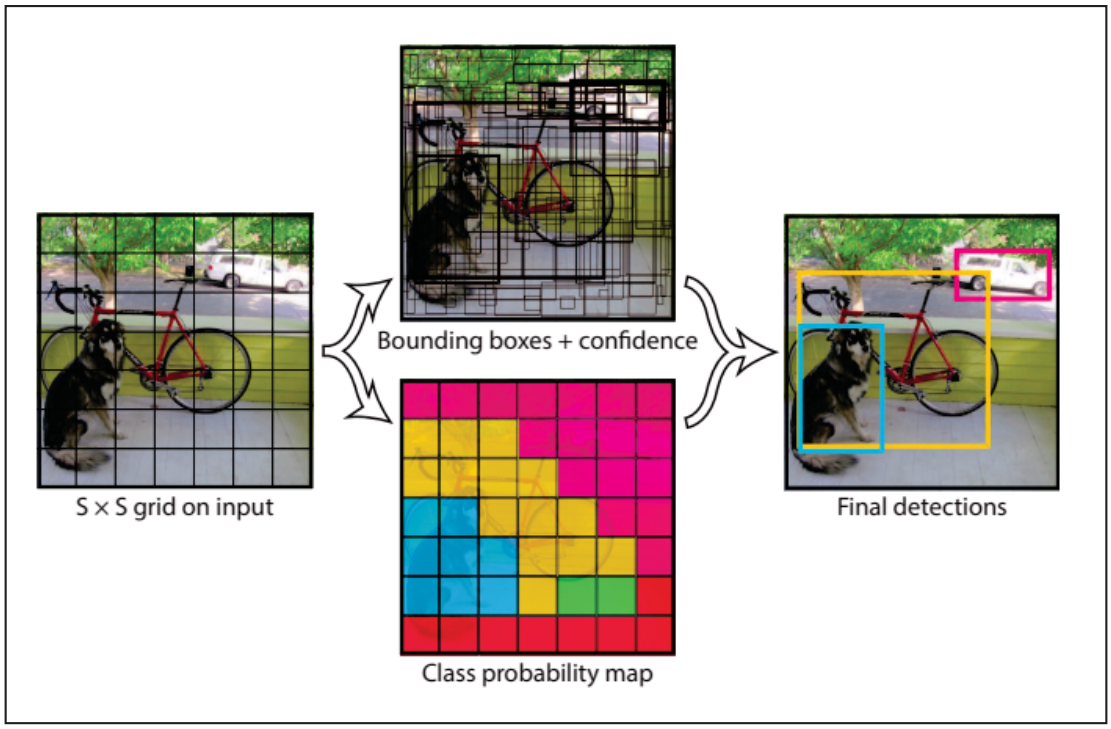


Figure 1.3 The YOLO deals with detection as a regression problem. It divides the image into an $S \times S$ grid and for each grid cell predicts B bounding boxes, confidence for those boxes, and C class probabilities. These predictions are encoded as an $S \times S \times (B * 5 + C)$ tensor Taken from Redmon *et al.* (2016)

RetinaNet. The work developed by Lin *et al.* (2017b) discovered that even one-stage detectors had high-speed inference and simplicity when compared with two-stage detectors, but they had a problem with extreme foreground-background class imbalance. The Focal Loss, see Figure 1.4, was proposed alongside the RetinaNet to address the imbalance problem. This loss is responsible for investing in misclassified examples during the training. This strategy leads the one-stage detectors to achieve results comparable to the two-stage detectors while keeping their detection speed.

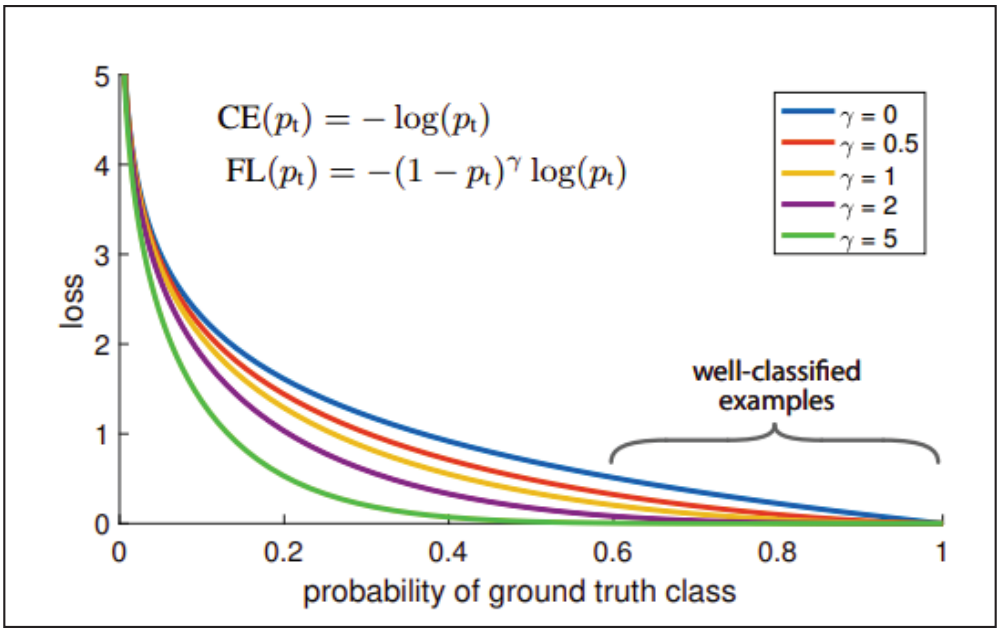


Figure 1.4 The Focal Loss adds a factor $(1 - pt)^{\gamma}$ to the standard cross entropy criterion. Setting $\gamma > 0$ reduces the relative loss for well-classified examples ($pt > .5$), putting more focus on hard, misclassified examples. The focal loss enables training highly accurate dense object detectors in the presence of vast numbers of background examples
Taken from Lin *et al.* (2017b)

The Retinanet also uses FPN combined with ResNet and classification and regression subnetworks, see Figure 1.5.

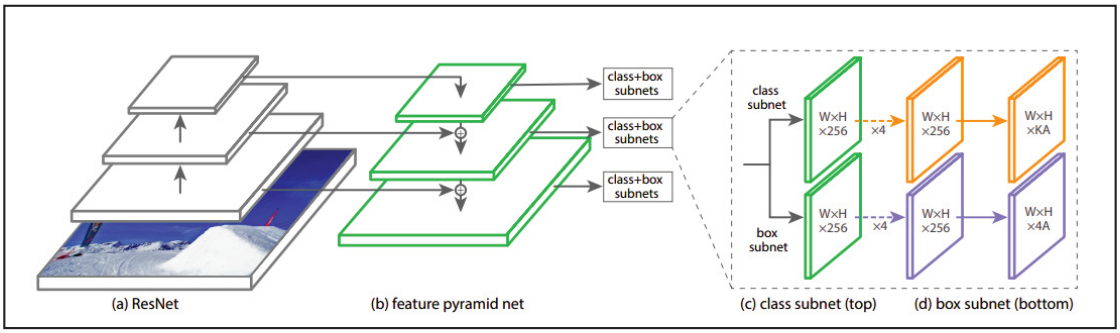


Figure 1.5 The one-stage RetinaNet network architecture uses a Feature Pyramid Network (FPN) backbone on top of a feedforward ResNet architecture with classification and regression subnets
Taken from Lin *et al.* (2017b)

1.2.2 Two-stage detectors

The Two-Stage Detectors' idea consists of firstly extracting regions of interest or proposals for a second-stage classifier. Then, the second stage is responsible for classifying whether there is an object in that region. The following object detectors are considered the principal two-stage deep learning object detectors.

Regions with CNN (R-CNN). The R-CNN by Girshick *et al.* (2015) extracts object proposals (object candidate boxes) using selective search (Uijlings, Van De Sande, Gevers & Smeulders, 2013), see Figure 1.6. Then, the proposals are rescaled and given to a CNN trained on ImageNet to extract features. Further, the features are given to a classifier to predict the presence or absence of the object.

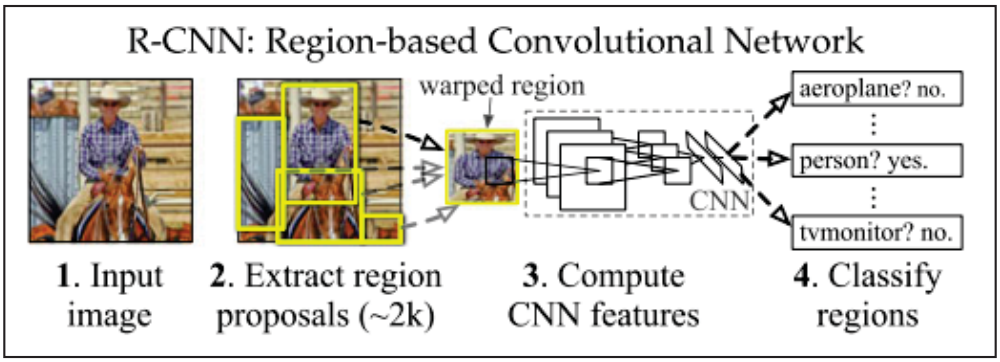


Figure 1.6 R-CNN takes an input image, then extracts around 2000 bottom-up region proposals, after it computes features for each proposal using a large CNN, and then classifies each region using class-specific linear SVMs
Taken from Girshick *et al.* (2015)

Spatial Pyramid Pooling Net (SPPNet). One drawback of the RCNN was that it required fixed-size input. The Spatial Pyramid Pooling (SPP) Net was proposed by He *et al.* (2015) to solve this problem. The SPP layer enables a generation of fixed-length representations without rescaling the input image or region of interest, see Figure 1.7.

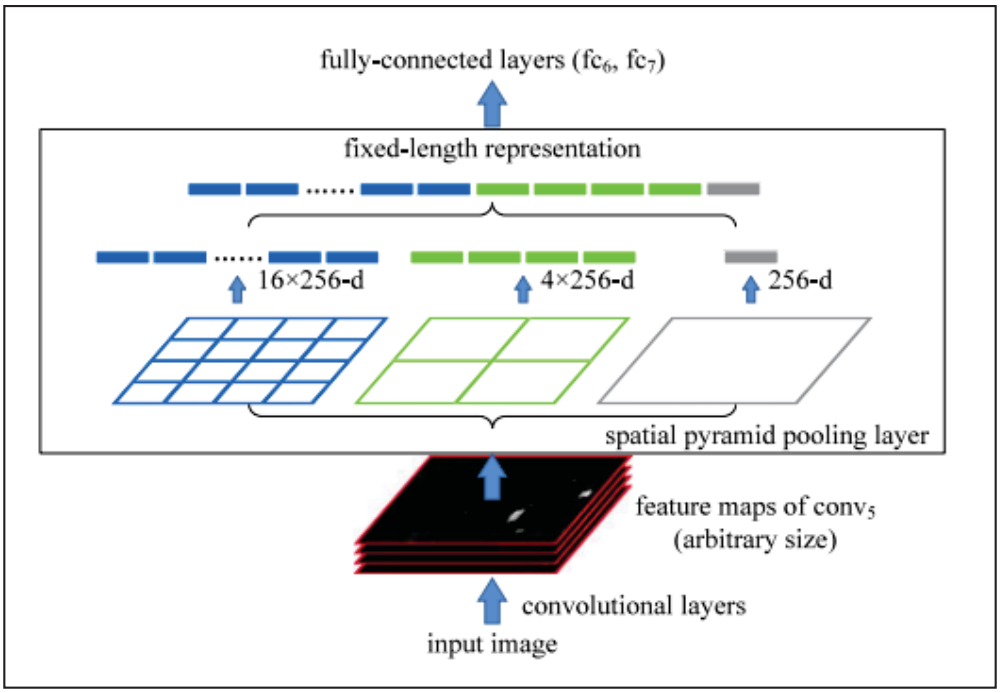


Figure 1.7 A network structure with a spatial pyramid pooling layer
Taken from He *et al.* (2015)

Fast-RCNN. Girshick (2015) proposed the Fast-RCNN method to speed up the RCNN. It improves the RCNN with SPPNet, which can train the detector and a bounding box regressor with the same network configurations, see the Figure 1.8. The proposals stage detections still limit the Fast-RCNN detection speed.

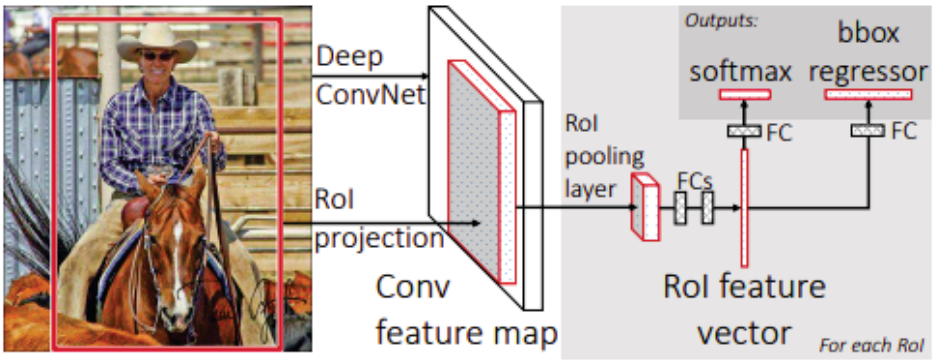


Figure 1.8 Fast R-CNN architecture. An input image and multiple regions of interest (RoIs) are input into a fully convolutional network. Each RoI is pooled into a fixed-size feature map and then mapped to a feature vector by fully connected layers (FCs) Taken from Girshick (2015)

Faster R-CNN. The Faster R-CNN proposed by Ren *et al.* (2015) is the first end-to-end deep learning object detector to reach real-time speed. The speedup was achieved by introducing the Region Proposal Network (RPN), a network responsible for the region proposals without impacting the computational performance compared with previous region proposals algorithms (Ren, He, Girshick & Sun, 2016). For detailed architecture of Faster R-CNN, see Figure 1.9.

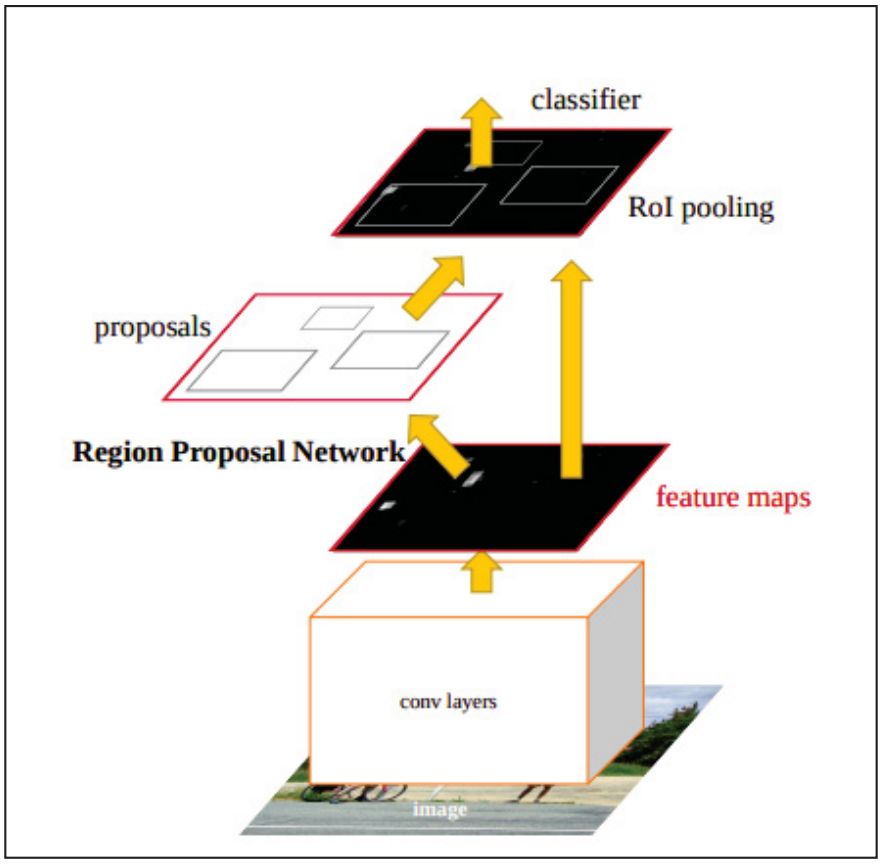


Figure 1.9 Faster R-CNN is a unified network for object detection. The RPN module serves as the attention of Faster R-CNN Taken from Ren *et al.* (2015)

Feature Pyramid Networks (FPN). The FPN method proposed by Lin *et al.* (2017a) provides a practical and fast solution for applying feature pyramids with deep learning object detectors. The proposed model was built using Faster R-CNN as a base detector. The FPN brings a top-down architecture with lateral building block connections responsible for building features for the objects at all scales.

1.2.3 Vision-language object detectors

Vision-Language Models (VLMs), including CLIP (Radford *et al.*, 2021), GLIP (Li *et al.*, 2022b), YOLO-World (Cheng *et al.*, 2024), and Grounding DINO (Liu *et al.*, 2024a), have

enabled open-vocabulary detection by aligning visual and textual embeddings. This capability allows the detection of novel classes without explicit training data.

CLIP. The CLIP (Radford *et al.*, 2021) architecture jointly trains an image encoder and a text encoder to predict the correct pairings of a batch of (image, text) training examples. At test time, the learned text encoder synthesizes a zero-shot linear classifier by embedding the names or descriptions of the target dataset's classes. Thus, the focus of CLIP is classification tasks. In Figure 1.10, we can see the CLIP architecture.

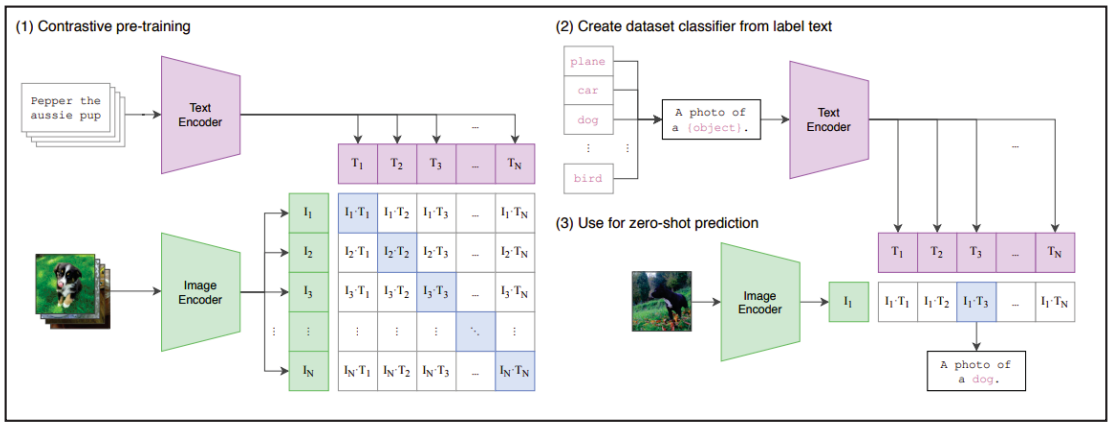


Figure 1.10 Overview of the CLIP architecture. The model is first trained with a contrastive learning objective that aligns image and text representations in a shared embedding space (1). In inference, dataset labels are converted into natural language prompts and encoded into the same space (2). Zero-shot classification is then performed by comparing the similarity between encoded images and encoded label prompts (3)
Taken from Radford *et al.* (2021)

GLIP. GLIP (Li *et al.*, 2022b) is a unified framework for detection and grounding, see Figure 1.11. In GLIP, the authors reformulate detection as a grounding task by aligning each region/box to phrases in a text prompt. GLIP jointly trains an image encoder and a language encoder to predict the correct pairings of regions and words. Then, they further add the cross-modality deep fusion to early fuse information from two modalities and to learn a language-aware visual representation.

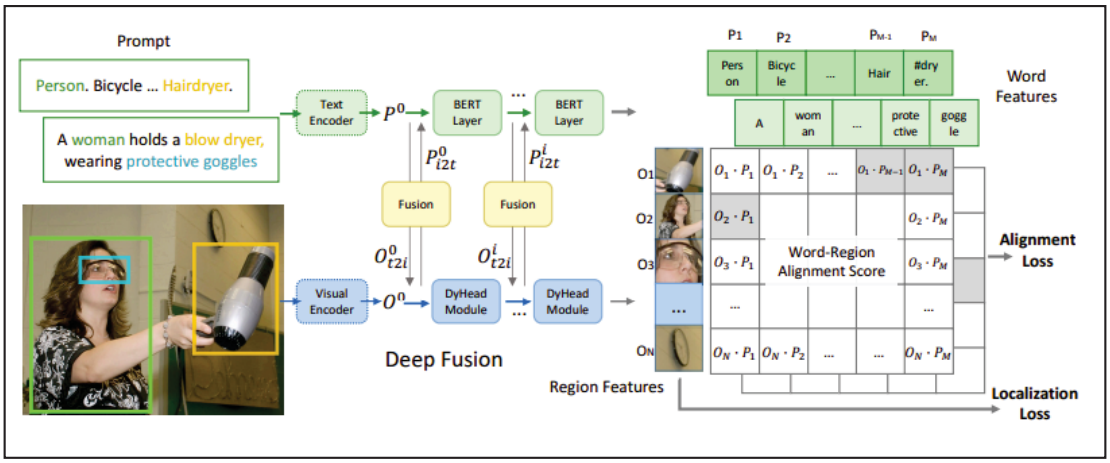


Figure 1.11 Overview of the GLIP architecture. GLIP unifies object detection and vision-language pretraining by encoding both region features and text prompts into a shared space. Text is processed with a BERT-based encoder, while image regions are extracted via a visual encoder and DyHead modules (Deep Fusion). The DyHead is a module responsible for combining different attention mechanisms such as scale-aware attention, spatial-aware attention, and task-aware attention. Word-region alignment is enforced through alignment loss, and detection quality is optimized with localization loss, enabling strong performance in both grounding and detection tasks
Taken from Li *et al.* (2022b)

Recent advancements in open-vocabulary object detection (OVD) leverage vision-language models (VLMs) with multimodal alignment at the region level to enable recognition of novel concepts. Early approaches, such as RegionCLIP (Zhong *et al.*, 2022), generate region-level pseudo-labels for contrastive pretraining, while Gu et al. (Gu, Lin, Kuo & Cui, 2022) cast OVD as a knowledge distillation problem, aligning detector embeddings with CLIP's region-text representations. Lin et al. (Lin *et al.*, 2023) employ bipartite matching for image-text alignment, and GLIP (Li *et al.*, 2022b) integrates deep multimodal fusion with self-generated box annotations, later extended by GLIPv2 (Zhang *et al.*, 2022) through region-grounded pretraining and contrastive learning.

More recent models, such as Grounding DINO (Liu *et al.*, 2024a), introduce a multiphase image-text fusion module that combines transformer-based detection with large-scale grounding

pretraining, excelling in both open-vocabulary and referring expression detection, see Figure 1.12.

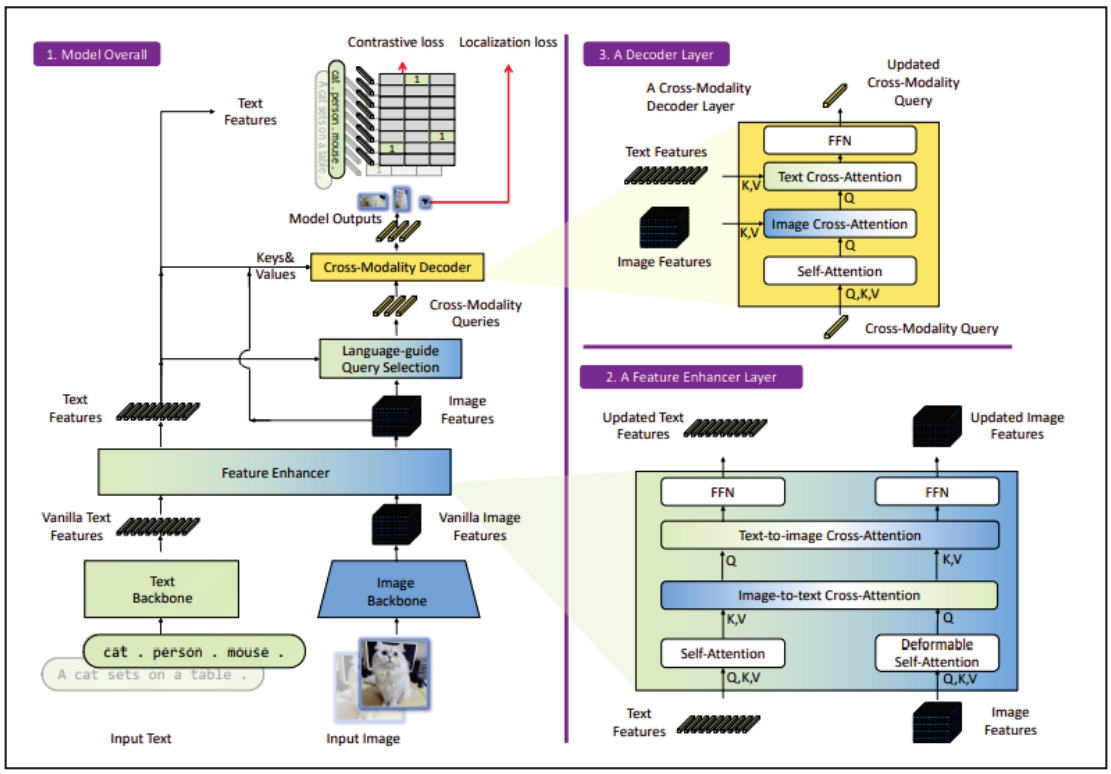


Figure 1.12 Overview of the Grounding DINO architecture. The model integrates text and image features through a feature enhancer and a cross-modality decoder. The feature enhancer refines vanilla text and image features using self-attention and cross-attention across modalities, while the decoder layer applies cross-modality queries to align text and image features. Training is supervised by both contrastive and localization losses, enabling unified grounding and detection with natural language
Taken from Liu *et al.* (2024a)

In contrast, YOLO-World (Cheng *et al.*, 2024) adapts the YOLOv8 architecture for real-time OVD via a re-parameterizable vision-language fusion network (RepVL-PAN) that efficiently merges CLIP-derived text embeddings with visual features, see Figure 1.13. Its "prompt-then-detect" paradigm allows offline text encoding for rapid deployment, achieving competitive zero-shot accuracy while maintaining low latency for edge applications.

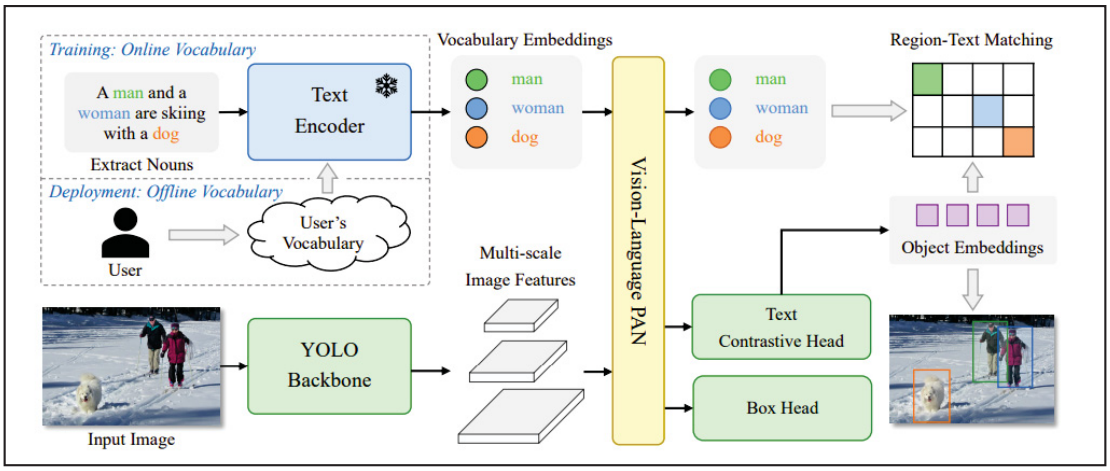


Figure 1.13 YOLO-World architecture. The text encoder first encodes the input text input text embeddings. Then the image encoder encodes the input image into multi-scale image features and the proposed RepVL-PAN exploits the multi-level cross-modality fusion for both image and text features. Finally, YOLO-World predicts the regressed bounding boxes and the object embeddings for matching the categories or nouns that appeared in the input text
Taken from Cheng *et al.* (2024)

1.3 Domain adaptation in object detection

In this work, we are going to use the definition by Li, Li, Luo, Wang *et al.* (2020b) for Deep Domain Adaptive Object Detection (DDAOD). DDAOD to train a robust OD using label-rich data of the source domain and label-agnostic or label-poor data of the target domain. In our scenario, domain adaptation is important to adapt the model to differences in camera setups and environmental changes. There are different categories that we can classify DDAOD works:

1. Mechanism to address domain shift;
2. One-step vs. multi-step adaptation methods;
3. Labeled data of the target domain;
4. Base object detector.

1.3.1 Mechanism to address domain shift

As described by Li *et al.* (2020b), we have methods that are discrepancy-based, adversarial-based, reconstruction-based, hybrid, and others. In this section, we detail each one.

Discrepancy-based. The idea of a discrepancy-based mechanism consists of decreasing the domain shift by fine-tuning the model with the usage of the target dataset with or without labels.

Adversarial-based. The main intention of the adversarial-based mechanism resides in using domain discriminators and adversarial training to encourage domain confusion between the source and target domains.

Reconstruction-based. The reconstruction-based mechanism presumes that reconstructing the source or target samples, creating a possible intermediate representation, helps the model adapt between different domains.

Hybrid methods. The hybrid mechanism is the most successful for obtaining better performance on DDAOD. Some works combine more than one idea above to build a complete solution for DDAOD tasks.

1.3.2 One-step vs. multi-step adaptation methods

When the distribution from source and target data is similar, we can leverage this by using a one-step adaptation by directly applying transfer learning techniques. While in multi-step adaptation methods, we can use several intermediate domain representations to help diminish the gap between the source and target domains. The multi-step domain adaptation methods can be diminished by applying prior knowledge that the intermediate domain will help the adaptation, or it can be learned automatically.

1.3.3 Labeled data of the target domain

This category considers the number of labels in the target domain. Generally, the vast majority of works in DDAOD are for unsupervised domain adaptation, but we can also have supervised, semi-supervised, weakly-supervised, and few-shot.

1.3.4 Base detectors

Generally, the works apply one type of detector, e.g., Faster RCNN or YOLO. The most common detector used for DDAOD is the Faster RCNN.

1.3.5 Review of methods

Many works on DDAOD deal with an unsupervised domain adaptation, in which the labels of the target domain are not provided. The work proposed by Saito *et al.* (2019) combines weak global alignment with strong local alignment (Strong-Weak domain Alignment Model) for unsupervised domain adaptation. The weak global alignment model uses adversarial alignment loss to give more importance to similar images at the global level and less to the different images. For the local alignment, a domain classifier is designed to spotlight the local features and align with the other domain. The context vector is extracted by the domain classifiers and is concatenated in the layer before the final fully-connected layer. The model is illustrated in Figure 1.14.

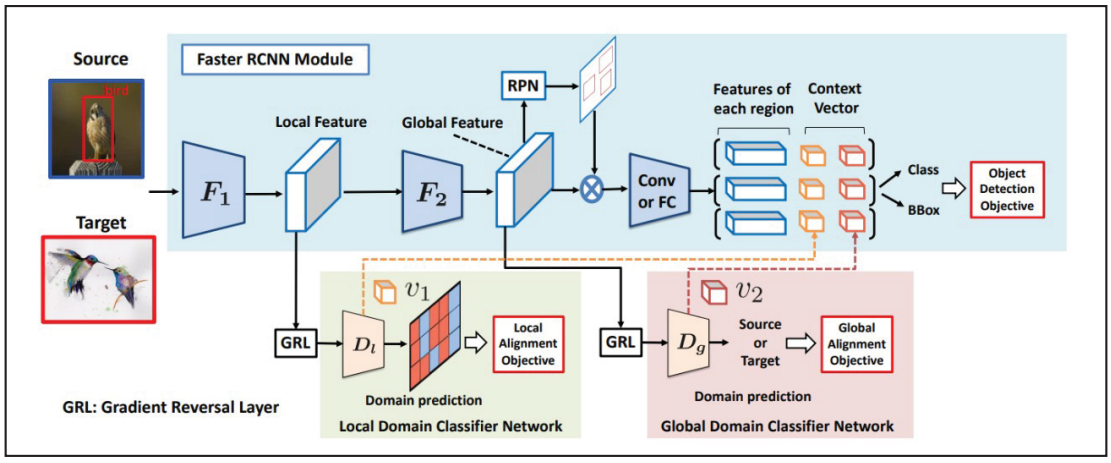


Figure 1.14 The model proposed by Saito *et al.* (2019) for unsupervised domain adaptive object detection. In blue is the detection module, in green is the local domain classifier, and in red is the global domain classifier
Taken from Saito *et al.* (2019)

Instead of working on alignment between global and local features, Khodabandeh *et al.* (2019) aims to tackle the problem of unsupervised DDAOD from the perspective of robust learning, in which the model is trained using noisy labels. To achieve this goal, a robust object detection framework (Figure 1.15) is developed, which adapts to the domain shift between source and target domains. In the first phase, an object detector is trained on source domain labeled data. Also, the detector is used for generating noisy labels on the target domain (pseudo-labeling). Then, in the second phase, the previous annotations are refined using a classification module. Finally, the detector is retrained using original labeled data and refined pseudo-labeling data in the third phase.

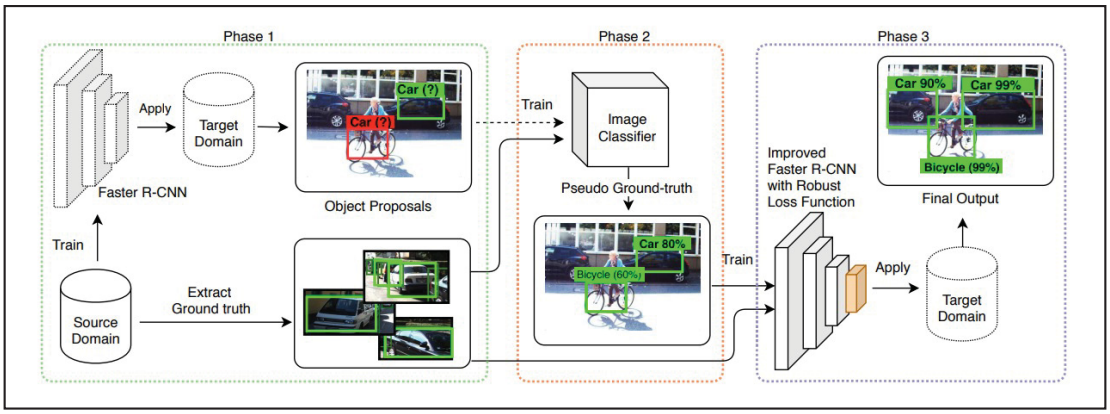


Figure 1.15 The robust learning object detection framework proposed by Khodabandeh *et al.* (2019) for unsupervised deep adaptive object detection. Phase 1 trains the object detector on the source domain and generates pseudo-labels. Phase 2 refines the labels with a classifier. Phase 3 retrains the detector with both source and refined target pseudo-labels
Taken from Khodabandeh *et al.* (2019)

Also, for unsupervised DDAOD, but with a different idea, Kim *et al.* (2019b) introduces a model to alleviate the imperfection translation problem of pixel-level adaptations and the source-biased discriminability problem of feature-level adaptations simultaneously. The proposed method is composed of two stages. One is domain diversification (DD) and multi-domain-invariant representation learning (MRL). The DD stage is responsible for diversity in the distribution of the label data, generating various distinctive shifted domains from the source domain using a CycleGAN with slight modifications to the loss. The MRL applies adversarial learning with a multi-domain discriminator to encourage features to be indistinguishable among the domains. In Figure 1.16 the proposed model is illustrated.

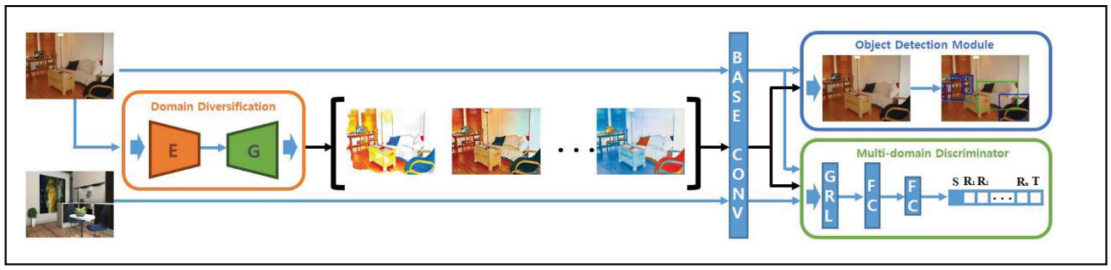


Figure 1.16 The architecture proposed by Kim *et al.* (2019b). In orange is the domain diversification (DD) module, in blue is the object detector module, and in green is the multi-domain invariant representation learning (MRL) module Taken from Kim *et al.* (2019b)

Furthermore, the idea of using CycleGAN was also investigated in other works. The work proposed by Hsu *et al.* (2020) consists of progressively adapting the model, creating an intermediate synthetic image between the source and the target domains using CycleGAN. Instead of learning to diminish the domain gap between the source and target domains, the model diminishes the gap between the synthetic representation and the target, as shown in Figure 1.17. The model has two stages. In the first stage, a CycleGAN is trained to adapt the model to a synthetic intermediate image representation. In the second stage, the discriminator of this CycleGAN is used to calculate a weight (W) that represents how close this intermediate image is to the target image. Then W is used in the adaptation network to adjust the detection loss.

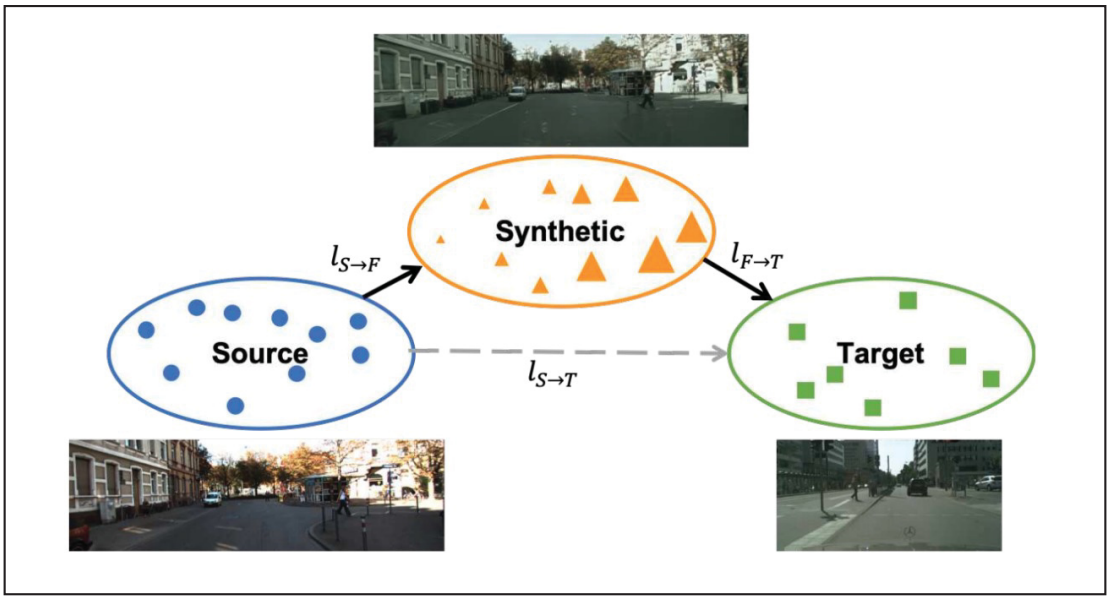


Figure 1.17 The architecture developed by Hsu *et al.* (2020) creates a synthetic representation to help with the unsupervised domain adaptation task through progressive adaptation
Taken from Hsu *et al.* (2020)

As well, Inoue *et al.* (2018) proposed a model using CycleGAN but for weakly-supervised object detection. In the paper, the task is defined as follows: A source domain (class labels and bounding box labels) and a target domain consisting only of instance-level annotations (class labels). The target domain classes are restricted to be a subset of the source domain classes. The final goal of the work is to detect common objects in target images without instance-level annotation on the target domain. The Figure 1.18 illustrates the problem.

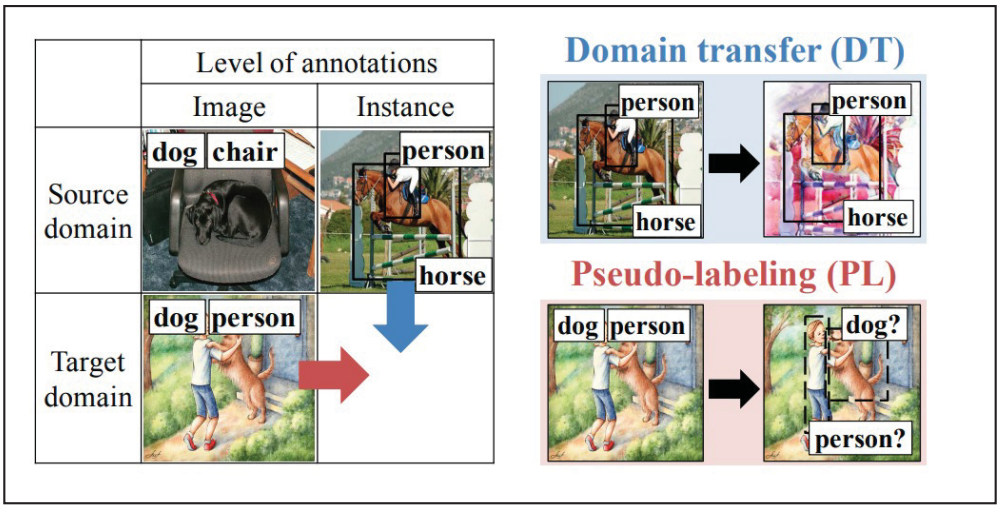


Figure 1.18 On the left is the Cross-Domain Weakly Supervised Object Detection Task. The proposed model with Domain Transfer (DT) and Pseudo-labeling (PL) is on the right
Taken from Inoue *et al.* (2018)

The workflow of the framework is described in Figure 1.19, in which in the first stage, the images in the source domain are used to train a detector, then in the second stage, the images are adapted using Domain Transfer (CycleGAN) to have a similar distribution of the target domain, and that images with bounding box annotations are used to fine-tune the detector. Moreover, this new detector is used to label the bounding box of the target images in the third stage, called pseudo-labeling (PL).

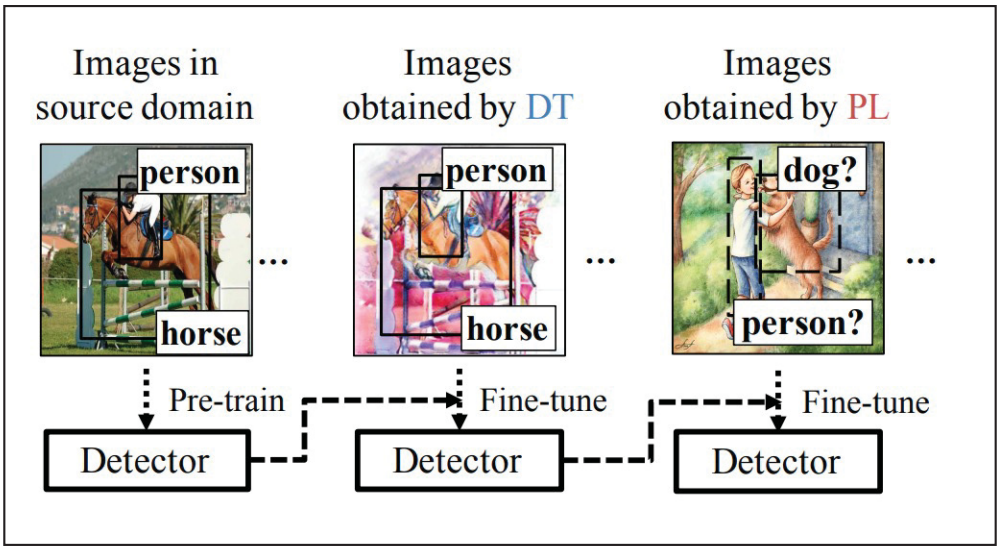


Figure 1.19 Cross-Domain Weakly Supervised Object Detection Framework
Taken from Inoue *et al.* (2018)

Likewise, using Pseudo-Labeling, but for unsupervised domain adaptive object detection, Kim *et al.* (2019a), proposed a weak self-training (WST) for a stable learning procedure and adversarial background score (BSR) regularization. The WST is responsible for minimizing the adverse effects of false positives and false negatives in pseudo-labels. The WST does this without a label on the target domain. The BST is used in the training phase to reduce the domain shift. In Figure 1.20, the training phase is in green with WST and BSR, and the test phase is in orange.

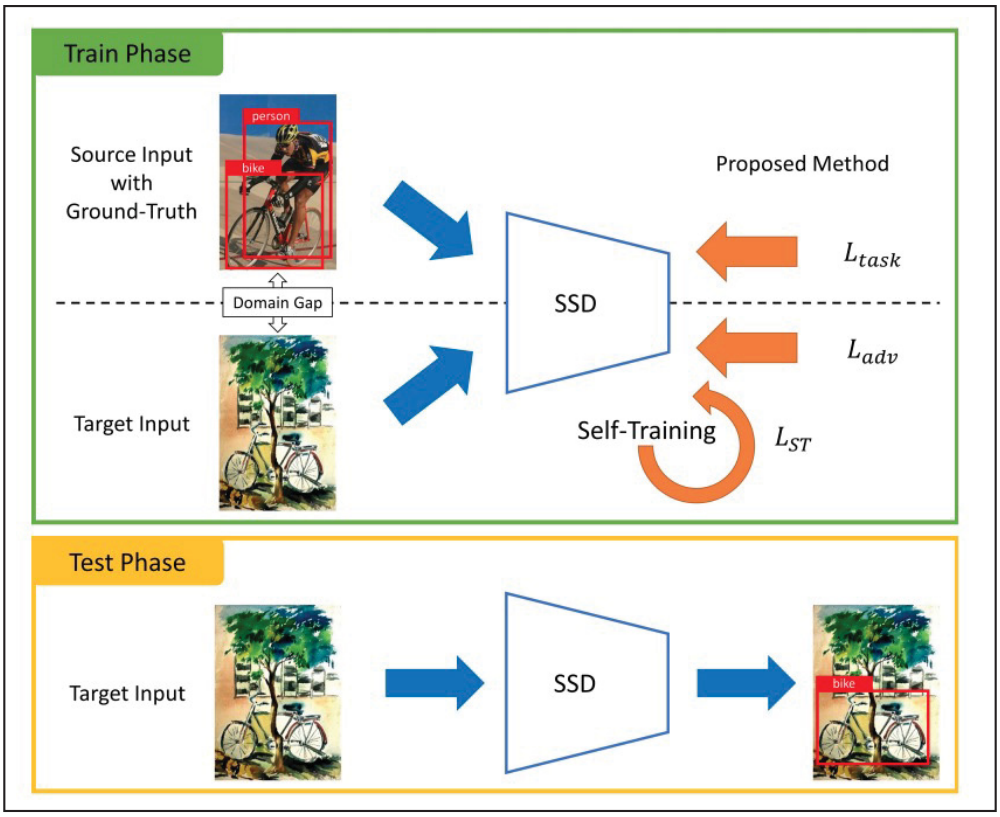


Figure 1.20 Illustration of unsupervised domain adaptive one-stage object detection. Training phase in green and testing phase in orange. The method improves the network's performance for target inputs Taken from Kim *et al.* (2019a)

These works are essential for adapting models for different conditions. Some of them compromise the alignment between feature levels, and some work on global or local alignment at the pixel level. In Figure 1.21, in which we can see that the detector gets lost when the target image domain gets noisy, like foggy. To deal with this problem, Xie *et al.* (2019) developed a multi-level domain adaptive learning framework.

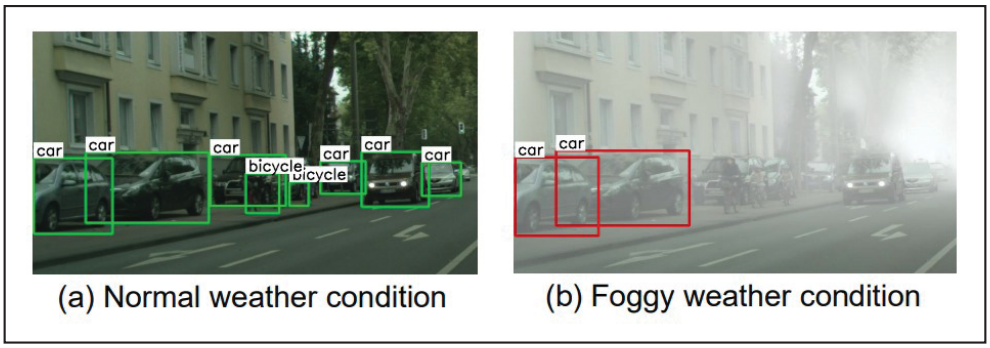


Figure 1.21 Different environmental conditions: a) normal weather condition, b) foggy weather condition
Taken from Xie *et al.* (2019)

The method proposed by Xie *et al.* (2019) consists of an end-to-end framework that learns how to adapt two different domains, making the local and global features of the source and target domains similar. They used local adaptations, with gradient reversal layers, after different levels of feature extractors and global adaptations near the classifier for this task. The Figure 1.22 describes the model.

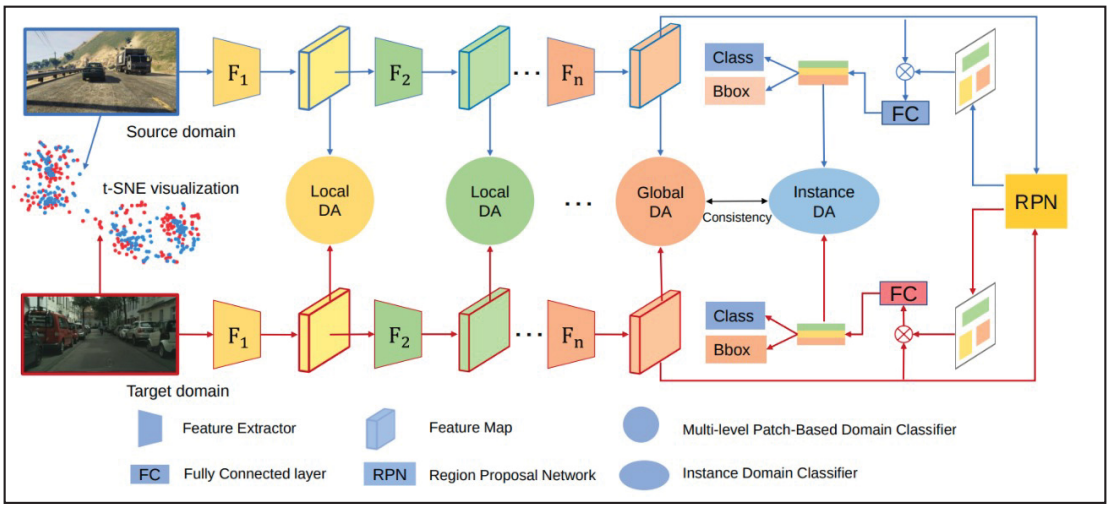


Figure 1.22 Multi-level Domain Adaptive learning for Cross-Domain Detection Framework
Taken from Xie *et al.* (2019)

Chen *et al.* (2020a) came up with a Hierarchical Transferability Calibration Network (HTCN) responsible for calibrating the feature representation of the model. The HTCN consists of three important components: (1) Importance Weighted Adversarial Training with input Interpolation (IWAT-I), (2) Context-aware Instance-Level Alignment (CILA), and (3) local feature masks. The (1) is responsible for strengthening the global discriminability by re-weighting the interpolated image-level features. The (2) enhances the local discriminability by capturing the underlying complementary effect between the instance-level feature and the global context information for the instance-level feature alignment. The (3) gives semantic guidance for the pattern alignment and can be seen as an attention-like module that captures the transferable regions. In Figure 1.23, the HTCN is illustrated.

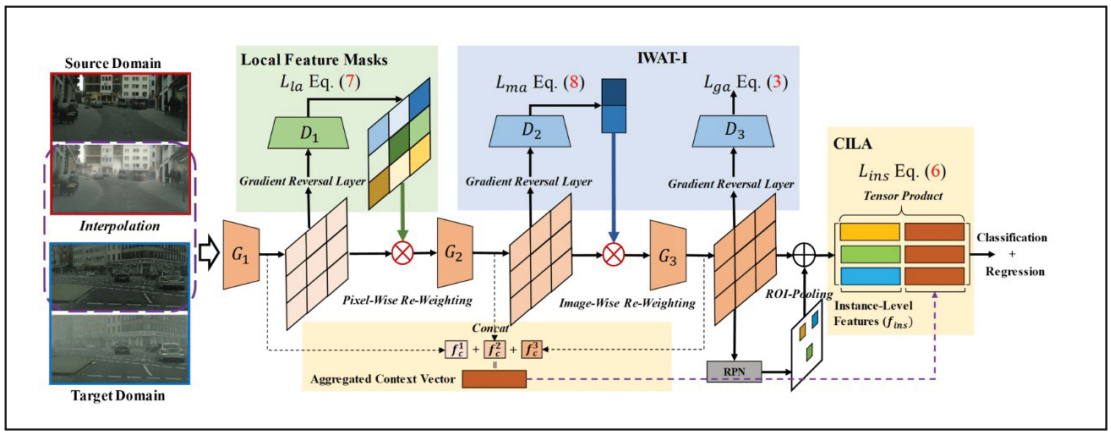


Figure 1.23 The overall structure of the proposed HTCN. D1 is a pixel-wise domain discriminator, while D2 and D3 are image-wise domain discriminators. G1, G2, and G3 denote the different level feature extractors Taken from Chen *et al.* (2020a)

Another interesting work that deals with object domain adaptation is the work proposed by RoyChowdhury *et al.* (2019), which adds temporal information. The work uses consecutive frames to help with object detection using tracking. The method also uses self-training with soft labels for the domain adaptation while using the cues from the tracker to help the model. This idea seems interesting for our problem because we could use temporal information to guide the model adaptation, improving its detections. A person who has been detected could keep being

detected if a tracker helped the model with some additional information, even if the environment changed.

This section gives good guidance on the techniques used in domain adaptation for object detection that our work could exploit. In our scenario, each room where we could install a smart device of our partner Distech Controls will have different setups, changes in environmental conditions, camera parameters, and points of view. The adaptation of one model trained in one setup could help our detector detect and track the person in the room. For example, CycleGAN approaches seem to be widely used to adapt between different domains and Pseudo-Labeling in unsupervised domain adaptation. Furthermore, the gradient reversal layer (GRL) for adversarial domain adaptation is a good strategy for generating domain features invariant that could boost performance on the target domain. Considering a multi-level approach, the GRL could also be exploited on different layers, managing to deal with more fine-grained representations (near classifier) or not, enriching the feature representation, aiming to diminish the gap between the domains.

1.4 Object detection using privileged information

In a scenario where we could have both RGB and IR labeled data available during training time, the assumption to work with Learning Using Privileged Information makes sense. In which one domain can be used as privileged information to improve the performance of a model trained on the other modality.

The classical machine learning paradigm considers the scheme to have a set of training examples used to find the best approximation function to an unknown decision rule, in which a teacher does not play an important role. However, in human learning, the role of a teacher is crucial, guiding the students with additional information, such as explanations, comparisons, and so on (Vapnik & Vashist, 2009). In the learning using privileged information (LUPI), during the training, we have additional information x^* to help the learning. The classical paradigm of

supervised machine learning for classification can be described as follows:

$$S = \{(x_1, y_1), ..., (x_l, y_l) \mid x_i \in X \wedge y_i \in Y\}, \tag{1.4}$$

in which S is the supervised data set composed of (x_i, y_i) pairs, X is the input set and Y is the set of labels. A machine learning model use the training set of pairs (x_i, y_i) to search a function in a set of functions $y = f(x, \theta)$, $\theta \in \Theta$, in which θ is the parameters of the model and Θ is the parameters space. The goal is to find a function $y = f(x, \theta)$ that minimizes the probability of incorrect classifications (the learnt model or function guarantees the smallest probability of incorrect classifications of y).

In the LUPI paradigm,

$$S^* = \{(x_1, x_1^*, y_1), ..., (x_l, x_l^*, y_l), \mid x_i \in X \wedge x_i^* \in X^* \wedge y_i \in Y\}, \tag{1.5}$$

in which S^* is the supervised data set composed of (x_i, x_i^*, y_i) triples, X is the input set from one domain, X^* is the input set from a different domain, and Y is the set of labels. The goal of the model is the same: to find the function $y = f(x, \theta)$, but now for each training set, we have an additional x^* that can be incorporated into the learning phase. The additional information, $x^* \in X^*$, belongs to the space X^*, which is different from the space of X. Since the additional information is available at the training stage but not during the test time, we call it privileged information (Vapnik & Vashist, 2009).

Lambert, Sener & Savarese (2018) propose the usage of privileged information to guide the variance of a Gaussian dropout. In a classification scenario, additional localization information of objects is used, and its results show that it improves the generalization, requiring fewer samples for the learning process (Lambert *et al.*, 2018). Motiian, Piccirilli, Adjeroh & Doretto (2016) designed a large-margin classifier using information bottleneck learning with privileged information for visual recognition tasks. These works demonstrated that privileged information could boost different machine learning tasks while reducing the number of training samples due to the additional information provided.

In the object detection problem, Hoffman *et al.* (2016) presents a modality hallucination framework, which incorporates the training RGB and Depth images, and during test time, RGB images are processed through the Multimodal framework to improve the performance of the detection. The modality hallucination network is responsible for mimicking depth mid-level features using RGB as input during the test phase. In Figure 1.24, the architecture of the modality hallucination network is illustrated. Also, Liu, Zhang & Zhang (2021a) used depth as privileged information for object detection with a Depth-Enhanced Deformable Convolution Network.

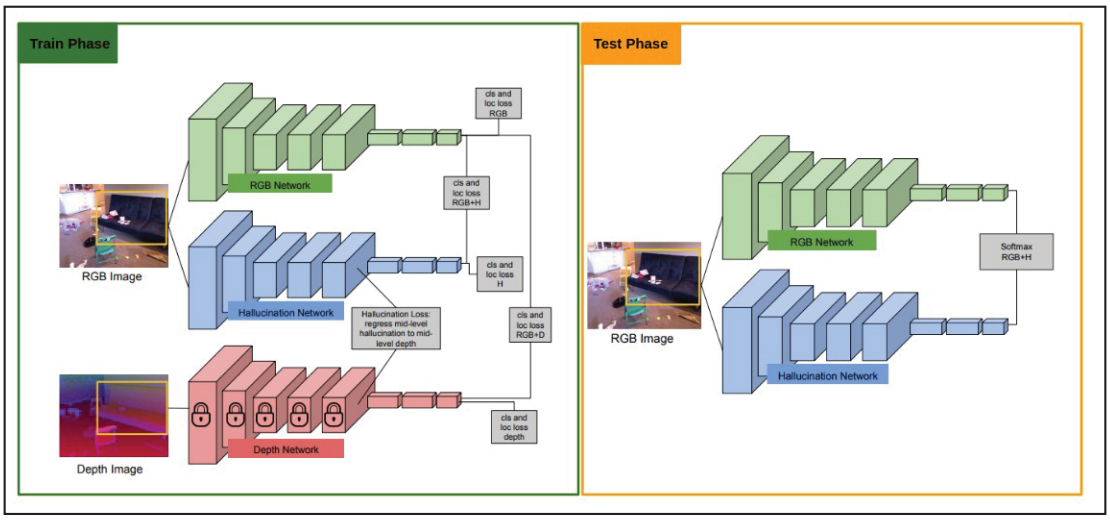


Figure 1.24 On green: Modality Hallucination Architecture during train phase.
On orange: Architecture during the test phase
Taken from Hoffman *et al.* (2016)

The LUPI setting was the core idea behind our work HalluciDet, which is going to be detailed in the next chapters and is one of our main contributions.

1.5 Object detection using infrared images

1.5.1 Infrared imaging

Infrared imaging (IR) is the process of capturing and representing images of objects based on their infrared radiation, which falls within the invisible portion of the electromagnetic

spectrum (Danaci & Akagunduz, 2022). In automotive applications, two types of infrared sensors are typically utilized: near-infrared (NIR) cameras, which capture radiation in the 0.7 to the 1.4-micrometer range, and long-wavelength infrared (LWIR) cameras, commonly referred to as thermal band cameras, which detect radiation in the 8 to the 15-micrometer range (Hwang *et al.*, 2015).

Table 1.1 The IR Spectrum
Taken from Danaci & Akagunduz (2022)

Wavelength	**Designation**
$10^{-6}\mu m$ to $10^{-2}\mu m$	x-rays
$10^{-2}\mu m$ to $0.4\mu m$	ultraviolet
$0.4\mu m$ to $0.7\mu m$	visible
0.7μm to 1.4μm	**NIR**
1.4μm to 3μm	**SWIR**
3μm to 8μm	**MWIR**
8μm to 15μm	**LWIR**
15μm to 1mm	**FIR**
1mm to 1m	microwaves
1m to 10km	radiowaves

The term FLIR (Forward Looking Infrared) is widely used in the literature. FLIR is a type of imaging technology that captures and represents the thermal (infrared) radiation emitted by objects in the environment. FLIR systems are utilized in several applications, such as military and security surveillance, building inspections, medical imaging, and automotive technology. These systems produce FLIR images, which display the relative temperature differences of objects and can be useful for detecting heat patterns and sources. The Kaist Multispectral dataset (Hwang *et al.*, 2015) illustrated in Figure 1.25 shows a FLIR image on the right part of the image.

Figure 1.25 Example of Forward Looking Infrared (FLIR) image from Kaist Multispectral Dataset. The FLIR operates on the long-wavelength infrared (LWIR) spectrum
Taken from Hwang *et al.* (2015)

The commonly available high-resolution datasets, infrared/visible, such as LLVIP, Kaist Multispectral dataset, and FLIR ADAS, are collected with FLIR technology. These datasets provide good-quality IR and visible images from people in urban scenarios. For video monitoring of human activity, the long-wavelength infrared (LWIR) spectral range (8-15 μm) is more suitable than the mid-wavelength infrared (MWIR) range (3-8 μm). This is because objects at ambient temperature (300K) emit radiation that peaks in the LWIR range, making it more effective for detecting human activity (St-Laurent, Maldague & Prévost, 2007). In particular, a human body at $37°C$ emits radiation with a dominant wavelength of 9.3μm, which falls within the LWIR range (St-Laurent *et al.*, 2007).

Recently, the cost of thermal sensors has decreased due to advances in new sensor technology. For example, a thermal array sensor developed with thermopile technology provides low-resolution

infrared sensing in a way capable of measuring humans' presence in indoor environments (Wang, Ali, Fan & Abuhmed, 2023a). A thermopile is an electronic device responsible for converting thermal energy into electrical energy (Wang & Ge, 2016). Thus, this device is used to measure indoor environments and capture thermal information that can be mapped to thermal images, as we can see on Figure 1.26.

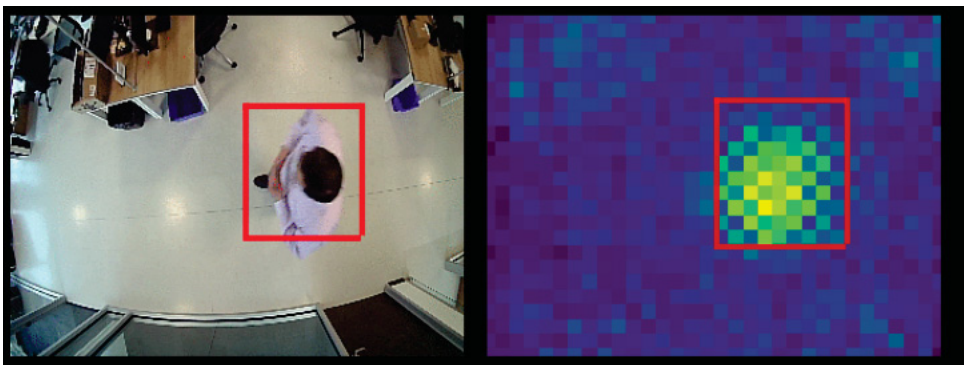


Figure 1.26 Visible (left) and infrared (right) images from the Distech system. The infrared image was acquired with a far-infrared thermal sensor array with a resolution of 32x24. The infrared image shown results from per-image normalization of the thermal sensor information, which introduces noise to the image and models trained on this image

The work of Wang *et al.* (2023a) is a recent work on real-time human detection that modified the YOLOv5 detection architecture for human detection on infrared images. Still, this work did not investigate the best way to adapt the thermal information for the YOLOv5 architecture. Since the code is not provided yet, we analyzed their provided dataset, which is the infrared colormap using per-image normalization. This colormap is not optimal for infrared, as we had already employed it in the Distech thermopile dataset, so the models trained on it suffer when the image is empty. In Dubail *et al.* (2022), the work proposes different uses of supervision for person localization. However, the work misses an important evaluation of the best way to deal with the infrared information provided by the thermal sensor and the best way to incorporate previous information from deep learning models that were developed for RGB data to work on this

infrared data. Another recent paper (Vandersteegen, Reusen, Beeck & Goedemé, 2023) proposes to use a modification of YOLOv2 with fewer parameters combined with background subtraction for person detection on low-cost devices. This work did not investigate how the backbone could directly benefit from the thermal information. Thus, the work applied a per-image normalization, as seen in their paper figure, and to deal with the noise of the per-image normalization, it incorporated background subtraction. Here, we see a clear gap in the literature about the real gains of pre-training foundation detector models and the best way to deal with the normalization of the infrared sensor information for these detectors' architectures. This section is important for understanding the details and physics definitions of infrared data. Additionally, the low resolution was really important for the initial stages of this thesis, as we developed models in collaboration with Distech Controls. Even though it is important to understand some details of this type of data, the following sections and the work on this thesis focus more on the deep learning aspect for high-resolution benchmarks composed of different modalities, where one modality is part of this benchmark.

1.5.2 Detection with infrared

In this section, we focus on object detection with infrared images, explaining some differences between visible images and infrared images while bringing a review of the state-of-the-art for object detection with deep learning. The following contextualization will help the development of our solution for infrared person detection, which comprises low-resolution infrared images.

Visible cameras may not capture valuable information under poor environmental conditions due to changes in illumination (Wang *et al.*, 2019). Thermal imaging, or infrared (IR) imaging, measures the heat of objects compared with other objects around them and translates those heat measurements into an image. The IR cameras are widely used for thermal imaging in video surveillance systems when the visible light is poor at night (Wu, Zheng, Yu, Gong & Lai, 2017). Even having some differences in the wavelength, the infrared range is more comprehensive than the thermal images. In this work, we will consider it the same, so that we will use both names (infrared and thermal) interchangeably.

Before the deep learning era, hand-crafted features like histogram oriented gradient (HoG) were used for pedestrian detection using infrared images (Suard, Rakotomamonjy, Bensrhair & Broggi, 2006; Miezianko & Pokrajac, 2008). Along with the advances of DL, we see works developing DL-based models ensuring improvements in different tasks using IR data. Bhattarai & Martinez-Ramon (2020) used thermal images with deep learning to help the detection of targets in real-time to improve firefighting. Meanwhile Janssens, Van de Walle, Loccufier & Van Hoecke (2017) used DL with an infrared thermal image for monitoring the condition of machines and their components. Moreover, Wang, Cai, Chen & Chen (2016) proposed a method for night-time vehicle detection in far infrared images using a local adaptive threshold and a deep belief network (DBN) based classifier. We see thermal images used in many different tasks, such as helping monitor different conditions to save lives.

Targeting object detection, Ghenescu *et al.* (2018); Brehar, Vancea, Marita, Vancea & Nedevschi (2019) used a YOLO-based approach for detection with high-resolution infrared images. Ryu & Kim (2018) designed a data-driven proposal and a convolutional neural network, which added a region proposal to a YOLO-based architecture for small infrared target detection. These works changed the adapted detector architecture for infrared detection but did not explore more details. Focusing on person detection, Chebrolu & Kumar (2019) developed a mechanism to perform pedestrian detection based on the light conditions. If the input consists of a nightly image, it selects a Faster R-CNN model trained on thermal images. Otherwise, it uses a Faster R-CNN trained on color images. Nataprawira, Gu, Goncharenko & Kamijo (2021) used multispectral images (paired RGB images and thermal concatenated) as input for pedestrian detection with a YOLO-based network and shows that the model performed better than using only thermal images.

1.6 Multimodal learning

1.6.1 Multimodal learning with images

In multimodal learning with images, we have network architectures that contain modality-specific layers and shared layers, which use images of different modalities (Dou *et al.*, 2020). Generally, multimodal approaches in the image domain commonly use architectures that are "Y"-shaped and "X"-shaped based on the U-Net (Ronneberger, Fischer & Brox, 2015) architecture, as illustrated in Figure 1.27.

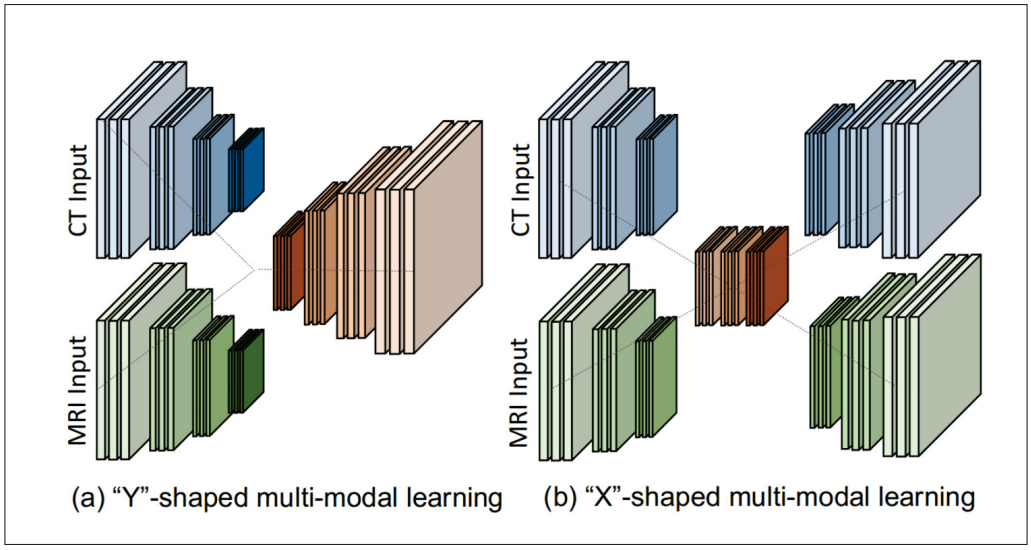


Figure 1.27 The image illustrates two different shapes used in Multimodal learning for the segmentation of brain imaging. In the image
a) “Y”-shaped that has a shared latent space and one decoder and
b) “X”-shaped that has a shared latent space with one decoder for each modality
Taken from Dou *et al.* (2020)

Multimodal learning can bring a robust shared latent space for the modalities. This space can synthesize the target domain during test time, which can bring additional information for object detection. Nonetheless, using this approach in object detection is still hard without paired images because the features have to be aligned in the latent space to bring useful information for

the detection part. Thus, if we want to use something similar in object detection, we should reduce the misalignment of the modalities to use with unpaired images. We were inspired by this to incorporate U-Net-based models in our subsequent works, thus adding new knowledge of modalities into the detector.

1.6.2 Learning with missing modalities

Learning with Missing Modalities is a common topic in segmentation tasks, especially for medical images (Azad, Khosravi, Dehghanmanshadi, Cohen-Adad & Merhof, 2022). Sometimes we have images from one modality (one sensor), and images from other sensors can be missing during the test. For example, to simulate the behavior of a missing modality, the work of Shen & Gao (2019) has an input batch composed of all modalities stacked, and they drop some modalities randomly at each step of the training. With a self-supervised training using similarity loss, they are capable of recovering the information on the missing channel during the test.

In the medical images field, we find surveys describing learning missing modalities for segmentation on Magnetic Resonance Imaging (MRI) (Azad *et al.*, 2022), in which they proposed a taxonomy of methods dealing with missing modality: a) Synthesis Models, b) Common Latent Space Models, c) Knowledge Distillation, d) Mutual Information Maximization and e) Generative Adversarial Networks.

a) Synthesis models. This class of models, as the name suggests, uses information on the missing modality to synthesize it during the test. Even though this class of models synthesizes the information, Azad *et al.* (2022) grouped here models that are more classical machine learning approaches, such as SVM-based, shallow Neural Networks (NNs) based models, and Restricted Boltzmann Machine (RBMs) based models.

b) Common latent space models. In this class of models, the modalities are mapped to a common latent space using deep learning. With this latent space, the modality that is missing is capable of being synthesized.

c) Knowledge distillation. The knowledge distillation models are models composed of teacher-student models, in which the teacher is trained in more modalities and distills the knowledge to the student. The student, during the test time, is capable of dealing with the missing information because they can learn the information provided by the soft labels of the teacher.

d) Mutual information maximization. The mutual information maximization class achieves minimal information loss in the missing modality situation by optimizing similarity metrics between available modalities during training.

e) Generative adversarial networks. In this class of models, generative adversarial networks (GANs) are used to do unsupervised translation. The idea behind using GANs is to generate the missing modalities with the generator.

We think that learning with missing modalities can help us understand how we can leverage the privileged information of the RGB during training, while we do not have access to it during the test. Still, we can synthesize (the missing modality) and use it in our model to increase the performance of the target modality. Furthermore, GAN-based models are strong baselines that we plan to add for comparison with our proposed model.

1.6.3 Multimodal fusion

Multimodal learning is associated with information provided by multiple sources (Ngiam *et al.*, 2011). The difference between multimodal and cross-modal relies on the presence of the modality during the learning phases. In multimodal learning, access to multiple sources of data is present during the training and also during the test. In the cross-modality setting, the multiple sources of data are presented only during the training phase. During the test phase, the model will be performed in one single modality (Ngiam *et al.*, 2011). In this work, we consider the different sources of data, e.g., infrared images and visible images, as different modalities that are going to be used in our models. There are different ways that can be used to fuse different modalities in a multimodal scenario. Following, we are going to define the different settings from the perspective of deep learning.

To fuse modalities on deep learning architectures for image classification, the work of D. Ramachandram and W. Graham (2017) Ramachandram & Taylor (2017) classify the approaches as Early fusion (input level or data-level fusion), late fusion (decision level), and middle fusion (feature level or intermediate fusion). This classification is illustrated in Figure 1.28.

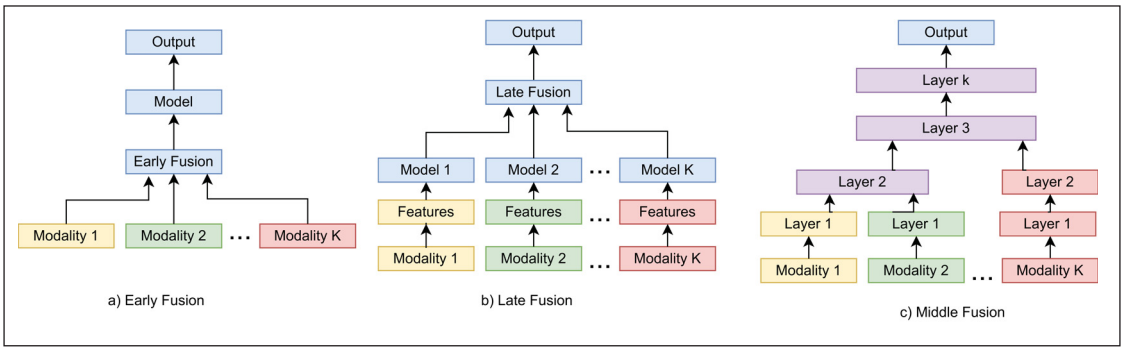


Figure 1.28 Different strategies to fuse multimodalities for learning models. a) early fusion, b) late fusion, c) middle fusion Taken from Ramachandram & Taylor (2017)

Early fusion. The focus on early fusion, as the name suggests, is to perform the fusion on the input level. In early fusion, the objective is to combine data from multiple sources, remove correlations between the modalities, and then fuse this data as input to the machine learning algorithm Ramachandram & Taylor (2017). In the object detection scene, we also see the early fusion approach with infrared and visible as multispectral, as the data is fused in the early phase of training. Wagner, Fischer, Herman, Behnke *et al.* (2016a) propose a study comparing early and late fusion in object detection for thermal and visible images. The work proposed by Knyaz, Vladimir (2019), for instance, performs multimodal data fusion for object recognition in images using infrared and visible images (Knyaz, 2019). In this work, multimodal recognition refers to using multimodal models to recognize (classify and localize) objects in images using deep learning models. In this work, Knyaz, Vladimir (2019) proposes an early fusion data approach using the stacking of infrared and visible data. This work also proposes a late fusion approach.

Middle fusion. The middle fusion, on the other hand, tries to embed the input in a feature space that combines the different modalities. Generally, this kind of fusion is performed by learning

algorithms in which the raw inputs are transformed into a higher-level representation by this algorithm, and this fused representation is used for the final task (Ramachandram & Taylor, 2017). Sharma et al. (2020) developed a YOLO-based object detection in multimodal remote sensing using middle-level fusion (Sharma *et al.*, 2020). The model has two different feature extractors that are fused in a middle layer of the proposed architecture. The work of Zhang *et al.* (2019b) proposed a weakly aligned cross-modal learning for multispectral (thermal/visible) pedestrian detection. Zhang *et al.* (2019b) focused on proposing feature fusion to provide a region feature alignment network using a two-stream framework, each one for one modality. In the work of Xu, Ouyang, Ricci, Wang & Sebe (2017a), they reconstruct the thermal data from the RGB during test time. Thus, in training, they trained a network to mimic some features of the IR, and during test time, they plugged another network together with this one to make the final inference. This work has two steps and two-stream networks. Thus, it requires some additional memory to mimic the missing modality. The goal of their work is to have the RGB data during test time, which is different from the actual trend of privacy-preserving detection, which tries to remove the RGB data during inference. We also find a similar trend of using two-stream feature fusion with other works like: Choi, Kim, Park & Sohn (2016), Li, Song, Tong & Tang (2018a) and Zhang *et al.* (2019a). Additionally, if we look at approaches that use generative models to create pseudo-RGB images, we see the work of Devaguptapu, Akolekar, M Sharma & N Balasubramanian (2019) that uses two-stream ResNets for fusing one branch provided by the thermal and other one provided by a cycle-gan image, the images created by the cycle-gan can be okay for the task of translation but does not have a guarantee that is the best for the subsequent detection task. Thus, our second research direction idea benefits from using a pseudo-RGB (hallucination modality) guided by the detection loss, not constraining the at al quality of the reconstruction but also giving some importance to what is better for the final detection task.

Late fusion. The late fusion is performed on the output space; for instance, if the problem is image classification, it is the fusion of the final probabilities of the models from each modality. The strategies used in late fusion deep object detection are similar to the ones used

in deep learning fusion due to the usage of backbone deep learning architectures to extract the features. The main difference is regarded as the way to perform the bounding-box fusion. D. Ramachandram and W. Graham (2017) describe late fusion as the aggregation of decisions from multiple classifiers, each trained on a separate modality. Xu et al. (2021) proposed an adaptive attention mechanism to perform 3d object detection. The input data of different modalities are fused based on the output of each specific modality network and the attention mechanism (Xu *et al.*, 2021). More recently, Chen *et al.* (2022) proposes a multimodal object detection using a probabilistic ensemble (ProbEn) approach that works on bounding-box fusion level using the Bayes rule. In Figure 1.29, the difference between previous works that fused on feature level and the work ProbEn that focused on very late fusion.

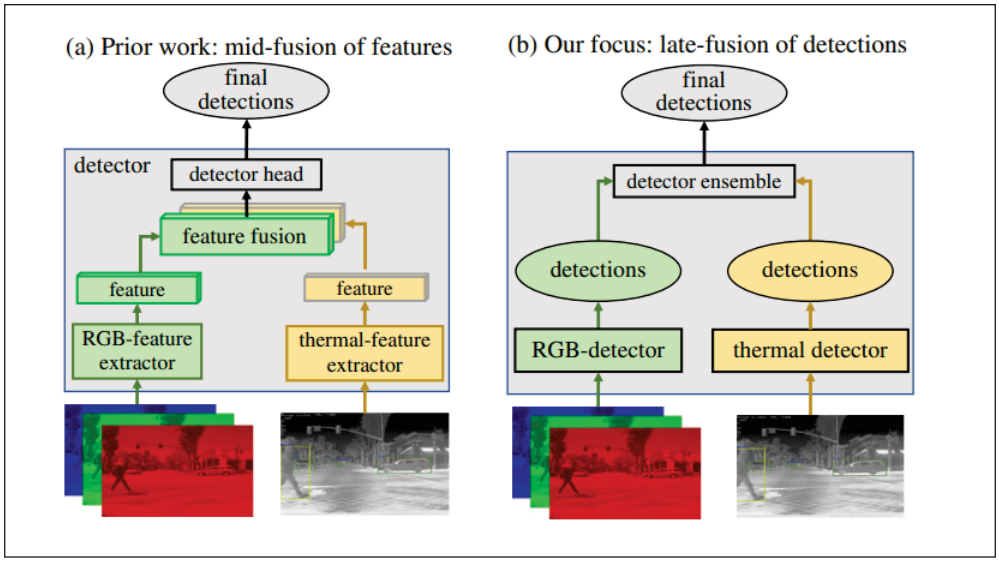


Figure 1.29 Comparison between different ways of doing fusion used on object detection. a) Previous works primarily focused on mid-fusion, concatenating features computed by single-modal feature extractors. b) The work proposed by Chen *et al.* (2022) focused on late fusion detector ensemble
Taken from Chen *et al.* (2022)

1.7 Evaluation methodologies

The choice of datasets and evaluation metrics plays a fundamental role in the development and assessment of object detection models. Robust evaluation requires both diverse and challenging datasets that reflect real-world conditions, and standardized metrics that allow fair comparisons across different methods. In this section, we detail the benchmark datasets used throughout this thesis, highlighting their main characteristics, modalities, and relevance to our experiments. These datasets cover a range of visual domains, from large-scale RGB benchmarks to infrared (IR), depth, event-based, and LiDAR modalities, enabling a comprehensive analysis of modality adaptation.

1.7.1 Benchmark datasets

COCO (Common Objects in Context). The COCO dataset (Lin *et al.*, 2014) is a large-scale benchmark widely used in computer vision, containing more than 330,000 images with over 1.5 million annotated object instances across 80 categories. It provides rich annotations, including bounding boxes, segmentation masks, and captions, and serves as one of the most comprehensive resources for training object detection models. In this work, COCO is employed as the main pre-training dataset for RGB detectors, ensuring a strong initialization before adapting to new modalities such as infrared (IR) and depth.

LLVIP (Low-Light Visible-Infrared Paired Dataset). The LLVIP dataset (Jia *et al.*, 2021) consists of $15,488$ paired RGB and infrared images ($12,025$ paired images in the training set and $3,463$ paired images in the test set) captured in low-light conditions, designed to facilitate pedestrian detection and related low-illumination vision tasks. The dataset is notable for its pixel-level alignment between RGB and IR images, enabling effective evaluation of modality adaptation methods. In this thesis, LLVIP is extensively used in the HalluciDet, ModTr, MiPa, and ModPrompt experiments as a core benchmark for RGB-to-IR adaptation.

FLIR Aligned. For the FLIR dataset, we adopt the sanitized and aligned paired version provided by Zhang, Fromont, Lefèvre & Avignon (2020b). This subset contains $4,129$ training IR

images and $1,013$ IR test images, each captured from the perspective of a front-facing camera mounted on a car, with a resolution of 640×512. The dataset includes annotations for four main categories: bicycles, cars, people, and dogs. Following the protocol of Cao, Bin, Hamari, Blasch & Liu (2023b), we exclude the "dog" class due to the limited number of annotations, which are insufficient for stable training. In this thesis, FLIR complements LLVIP as a primary IR benchmark.

NYU$_{v2}$ Depth. The NYU Depth V2 dataset (Silberman, Hoiem, Kohli & Fergus, 2012) is a large indoor RGB-D dataset captured with a Microsoft Kinect sensor. It comprises $1,449$ densely labeled images (795 for training and 654 for testing) at a resolution of 640×480, covering 19 semantic categories such as beds, chairs, and bookshelves. While the dataset contains both RGB and depth information, in this work, we focus on the depth modality for modality adaptation. NYUv2 is particularly employed in the ModPrompt experiments for vision-language object detectors, enabling the study of adaptation beyond visible and infrared modalities.

Additional modalities. To further test the generalization ability of the Modprompt, we also consider event-based and LiDAR modalities. The PEDRo dataset (Boretti *et al.*, 2023) provides event-camera data for pedestrian detection, capturing asynchronous brightness changes rather than conventional frames. In addition, the STCrowd dataset (Cong *et al.*, 2022) offers large-scale LiDAR point clouds annotated for pedestrians in crowded environments. Both datasets extend the evaluation beyond standard RGB, IR, and depth modalities, showcasing the adaptability of ModPrompt to a broader range of sensing technologies.

1.7.2 Evaluation metrics

In all experiments, we adopt the standard detection-based metrics established in the COCO benchmark (Lin *et al.*, 2014), which rely on the notion of Average Precision (AP) computed over predicted bounding boxes and ground-truth annotations. The AP metric summarizes the precision-recall curve, providing a single measure of detection quality that accounts for both false positives and false negatives. To compute AP, we first need to define the Jaccard

Index (Jaccard, 1901), also known as the Intersection over Union (IoU). The Jaccard Index measures the similarity between two sets, defined as:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}, \tag{1.6}$$

where A and B are sets, and $|*|$ is the cardinality. In object detection, A and B correspond to the areas of the predicted and ground-truth bounding boxes. Thus, IoU (illustrated in Figure 1.30) quantifies the overlap between bounding boxes: when two boxes are disjoint, IoU = 0, and when they perfectly match, IoU = 1.

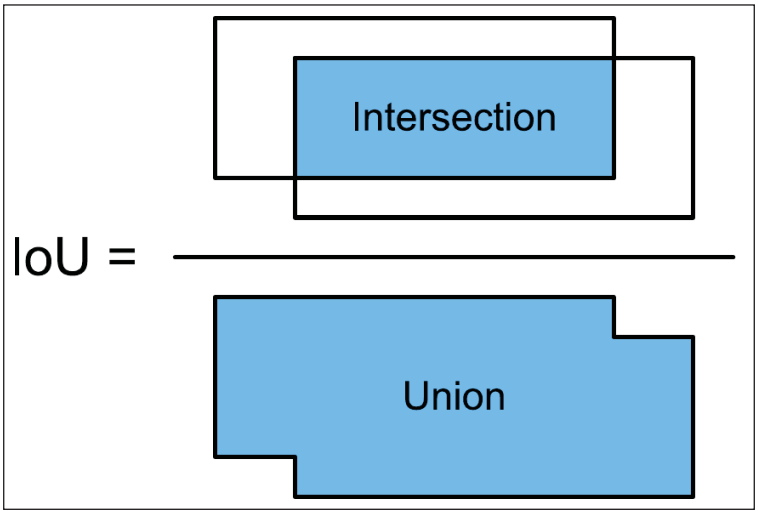


Figure 1.30 Intersection over Union (IoU): measures the overlap between predicted and ground-truth bounding boxes.
Taken from: Zhang *et al.* (2023a)

Using IoU with a threshold t, we can determine whether a detection is correct. A detection with $IoU \geq t$ is considered a true positive (TP), while one with $IoU < t$ is a false positive (FP). Ground-truth objects that remain undetected are counted as false negatives (FN). Based on these definitions, we can compute precision (P), recall (R), and the F_1-score as:

$$P = \frac{TP}{TP + FP} = \frac{TP}{all\ detections}, \tag{1.7}$$

$$R = \frac{TP}{TP + FN} = \frac{TP}{all\ ground\ truths}, \tag{1.8}$$

$$F_1 = \frac{TP}{TP + \frac{1}{2}(FP + FN)} = 2 * \frac{P * R}{P + R}. \tag{1.9}$$

Although F_1-score can be used, it is less common in object detection. Instead, the Average Precision (AP) is widely reported, defined as the area under the precision-recall curve. The Mean Average Precision (mAP) extends this definition by averaging the AP over all classes:

$$mAP = \frac{1}{N} \sum_{i=1}^{N} AP_i, \tag{1.10}$$

where N is the number of object categories, following the COCO convention (Lin *et al.*, 2014), AP is typically reported as $AP@[.5 : .95]$ (or just AP or mAP), which averages AP over ten IoU thresholds from 0.5 to 0.95 with increments of 0.05. This stricter evaluation provides a more robust measure of both classification accuracy and localization precision. In addition, it is common to report AP at fixed thresholds: $\mathbf{AP}_{50}$, corresponding to $IoU = 0.5$, measures coarse detection ability and emphasizes classification accuracy, while $\mathbf{AP}_{75}$, corresponding to $IoU = 0.75$, enforces stricter localization and highlights precise bounding box alignment. Together, $AP@[.5 : .95]$, AP_{50}, and AP_{75} provide complementary insights into model performance, balancing the evaluation of detection robustness and localization precision.

CHAPTER 2

MIXED PATCH VISIBLE-INFRARED MODALITY AGNOSTIC OBJECT DETECTION

Heitor R. Medeiros[a]*, David Latortue[a]*,
Eric Granger[a], Marco Pedersoli[a]

[a] Department of Systems Engineering, École de technologie supérieure,
1100 Notre-Dame West, Montreal, Quebec, Canada H3C 1K3

Paper published in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), February 2025*

Abstract

In real-world scenarios, using multiple modalities like visible (RGB) and infrared (IR) can greatly improve the performance of a predictive task such as object detection (OD). Multimodal learning is a common way to leverage these modalities, where multiple modality-specific encoders and a fusion module are used to improve performance. In this paper, we tackle a different way to employ RGB and IR modalities, where only one modality or the other is observed by a single shared vision encoder. This realistic setting requires a lower memory footprint and is more suitable for applications such as autonomous driving and surveillance, which commonly rely on RGB and IR data. However, when learning a single encoder on multiple modalities, one modality can dominate the other, producing uneven recognition results. This work investigates how to efficiently leverage RGB and IR modalities to train a common transformer-based OD vision encoder while countering the effects of modality imbalance. For this, we introduce a novel training technique to Mix Patches (MiPa) from the two modalities, in conjunction with a patch-wise modality agnostic module, for learning a common representation of both modalities. Our experiments show that MiPa can learn a representation to reach competitive results on traditional RGB/IR benchmarks while only requiring a single modality during inference. Our code is available at: https://github.com/heitorrapela/MiPa.

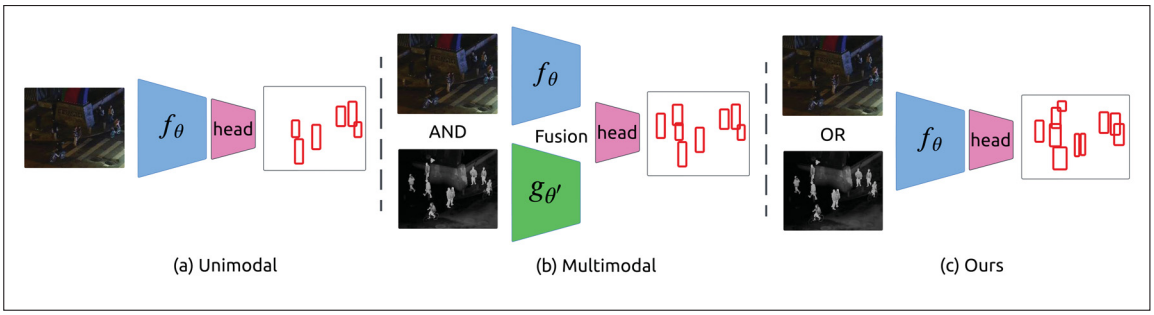


Figure 2.1 Differences in inputs for different modality learning. (a) *Unimodal* learning assumes that only one modality is used for both training and testing. (b) *Multimodal* learning requires multiple modalities and a special architecture to fuse them in order to improve performance. (c) *Ours* assumes that a model should be able to perform well for both modalities by using both for training but only one at a time for testing and with a shared vision encoder

2.1 Introduction

In recent years, the reducing costs in data acquisition and labeling have proportioned the advancements in multi-modality. Various fields are increasingly using this form of learning to enhance applications, such as surveillance (Chen *et al.*, 2019a; Alehdaghi, Josi, Cruz & Granger, 2022; Medeiros *et al.*, 2024c), industrial monitoring (Kong & Ge, 2021; Garillos-Manliguez & Chiang, 2021; Kini, Fleischer, Dave & Shah, 2023), smart buildings (Dubail *et al.*, 2022; Dayarathna *et al.*, 2023), self-driving cars (Stilgoe, 2018; Michaelis *et al.*, 2019; Medeiros *et al.*, 2024b), and robotics (Eitel, Springenberg, Spinello, Riedmiller & Burgard, 2015; Pierson & Gashler, 2017; Ivorra, Ortega, Alcañiz & Garcia-Aracil, 2018), due to their powerful ability to operate better in the presence of diverse environmental information (Tang, Liang & Zhu, 2023b). For instance, the combination of visible (RGB) and infrared (IR) has been showing promising results regarding such applications due to the difference in light spectrum sensing by different sensors, which provide not only additional but also complementary information (Wang *et al.*, 2022).

An unimodal learning (Figure 2.1a), utilizes data from a single modality, for instance, an object detector trained and used in production with RGB images. In multimodal learning (Figure 2.1b), the objective is to create a model able to incorporate information from multiple modalities,

such as RGB and IR, from different sensors and requires paired modalities for both training and inference. Although this multimodal learning covers a wide range of applications, as aforementioned, we have identified an underserved scenario where one might want an RGB/IR modality agnostic model that is trained on both modalities but is subjected to only either one or another during inference (Figure 2.1c). One example of that is a surveillance system where a server model is running all the time, and this model can provide detections for different RGB or IR sensors to address the need to make accurate detection in every lighting condition during different pre-defined conditions.

Despite the strong interest and business value in multimodal systems, most publicly available datasets and powerful pre-trained models are built around one modality: RGB. Furthermore, the lack of IR data gives additional motives to build a detector upon an already pre-trained unimodal RGB detector. However, the current methods proposed in research to incorporate dual-modality information into a model require dedicated components associated with each modality, making them incompatible with such RGB detectors. These methods are mainly based on fusion. For instance, these techniques adopt different modalities by distributing the RGB/IR across a four-channel input (three RGB followed by one for IR), in the case of early fusion (Wagner, Fischer, Herman & Behnke, 2016b)., or merging both modalities later in the model architectures (ZHANG, FROMONT, LEFEVRE & AVIGNON, 2021; Zhang, Fromont, Lefevre & Avignon, 2020c; Cao, Bin, Hamari, Blasch & Liu, 2023a) for mid-stage fusion or ensembling different unimodal modality detectors (Chen *et al.*, 2022) for late-stage fusion. This constrains the model to utilize both modalities during inference, which significantly increases inference speed compared to an unimodal architecture.

Typically, one probable phenomenon that can occur during multimodal training is modality imbalance. This happens when the strongest modality is leveraged more than the others, leading to better overall performance while discarding contributions from the others (Das *et al.*, 2023). In this work, we provide a way to train a single shared vision encoder to be agnostic to its input RGB/IR modality yet still extract its knowledge during training to attain results almost as good

on both modalities as if it was trained solely on each during testing/production. The naive solution for this type of task is to train a model with a dataset that blends both modalities.

Recently advances on patch-based transformers, such as ViT (Dosovitskiy *et al.*, 2020), and Multi-Modal Masked Autoencoders (Bachmann, Mizrahi, Atanov & Zamir, 2022) have steered us towards exploring patch-based architectures to build a powerful and yet simple training technique to create a RGB/IR modality agnostic vision encoder for object detection. Such approaches have been promising for multi-modal learning, which allows an efficient combination of different information (Geng *et al.*, 2022; Bachmann *et al.*, 2022). Our work investigates how to use RGB and IR modalities efficiently by using a patch-based transformer encoder. Thus, Mi(xed) Pa(tch) does not introduce any inference overhead during the testing phase while exploring an effective way to use the two modalities during the training. To accomplish such a task, we introduce a stochastic complementary patch mixing method, allowing the detector to explore each modality without having to rely on both of them simultaneously. This is possible by effectively sampling the optimal ratio of patches for each modality, which is then mixed using our technique. Subsequently, we enhance the training by suppressing the modality imbalances by proposing a modality-agnostic training technique, making the modalities indistinguishable from each other, a module inspired by Gradient Reversal Layer (GRL) (Ganin *et al.*, 2016) but with a novel design for patch based architectures. This approach is designed to allow low-cost inference in production while removing all requirements to know beforehand which modality the detector is going to be used with. Hence, in applications that run a detector all day, we can know beforehand that any of the modalities, RGB or IR, whenever they are being used, are going to perform optimally for the same shared vision encoder.

Our work provides empirical results alongside a theoretical explanation based on information theory describing the benefits of using MiPa with transformer-based backbones. Additionally, we study the ability of our MiPa to also be used as a regularization method for the more robust modality to boost the overall performance of the detector and we show that we can achieve competitive results on two traditional RGB/IR benchmarks: LLVIP and FLIR.

Our main contributions can be summarized as follows:

(1) We introduce MiPa, a novel mix patches RGB/IR modality agnostic training method for transformer-based object detectors, which learns how effectively sample the RGB and IR patches for best compressing the information of both modalities in a single encoder, without additional inference overhead.

(2) We propose a novel patch-wise modality agnostic module, which is inspired by the gradient reversal layer (GRL) for modality adaptation and is responsible for making the RGB/IR modalities invariant by the detector.

(3) We empirically demonstrate that the proposed method can also be used to improve the overall performance of detection when utilized as regularization for the strongest modality and achieve competitive results when compared with multimodal fusion methods, with less information during inference. Furthermore, MiPa can simply be applied to different transformer-based detectors, such as DINO (Zhang *et al.*, 2023b) and Deformable DETR (Zhu *et al.*, 2021).

2.2 Related Work

2.2.1 Patch-Based Vision Encoding

With the integration of Transformers in the vision field, researchers have started to deconstruct images into patches to allow the modeling of long-range relationships between patches (Dosovitskiy *et al.*, 2020). This powerful approach yielded great results and quickly became the norm amongst the top-performing models, ranking well on popular benchmarks such as ImageNet-1k (Russakovsky *et al.*, 2015). Multiple variants of the vision transformer have been proposed in recent years, for instance, ViT (Dosovitskiy *et al.*, 2020), DeiT (Touvron *et al.*, 2021), Swin (Liu *et al.*, 2021b), and VOLO (Yuan, Hou, Jiang, Feng & Yan, 2022). Alongside the new way of utilizing input images came a novel pretraining method for vision encoding: Masked Autoencoders (He *et al.*, 2022) (MAE). Indeed, this technique, which is simple to understand and easy to implement, consists of using a classifier as an encoder in an autoencoder architecture to generate images by only using a small fraction of the patches as input. This unsupervised method

has proven to be very useful in terms of improving results for downstream tasks. Furthermore, a similar idea has also been influential in the world of multi-modality models by building a multimodal MAE with one encoder and multiple decoders to reconstruct all the different modalities (Bachmann *et al.*, 2022). Recently, advances towards using Swin Transformer as a backbone of DINO (Zhang *et al.*, 2023b), an object detector descendant of the DETR (Carion *et al.*, 2020), were responsible for reaching competitive results in detection benchmarks, such as in COCO dataset (Lin *et al.*, 2015).

2.2.2 Multimodal Visible-Infrared Object Detectors

Regarding object detection, the primary methods of exploiting pairs of modalities, even when unaligned, are multimodal techniques; mainly fusion (Bayoudh, Knani, Hamdaoui & Mtibaa, 2021). Fusion is a technique where the advantage of multiple modalities is taken to better optimize one training objective by combining them to develop a multimodal representation (Pawłowski, Wróblewska & Sysko-Romańczuk, 2023). Fusion can be achieved at different stages, i.e., *early-stage fusion*, which concatenates the modalities across the channels, *mid-stage fusion*, where modalities are processed through dedicated decoders then merged e.g., Channel Switching and Spatial Attention (CSSA) (Cao *et al.*, 2023a), Halfway Fusion (Zhang *et al.*, 2020c), RSDet (Zhao, Yuan & Wei, 2024a), CrossFormer (Lee, Park & Park, 2024) or Guided Attentive Feature Fusion (GAFF) (ZHANG *et al.*, 2021), and finally *late-stage fusion*, where typically modalities are processed independently through different models and combined at the end using ensembling (Chen *et al.*, 2022), e.g. ProbEn (Chen *et al.*, 2022). The limitations of multimodal learning are that they require a custom architecture to handle each modality and are constrained to use both modalities during inference. A cross-modal with shared encoder vision models, however, are not affected by these limitations as the different modalities are only used during training and share the same encoder. This type of architecture unlocks the ability for detectors to have a higher degree of freedom for inference without compromising real-time applications.

2.2.3 Modality Imbalance

A potential obstacle to an RGB/IR modality-agnostic network is the phenomenon of modality imbalance. Given a dataset with multi-modal inputs, modality imbalance occurs when a model becomes more biased towards the contribution of one modality (Das *et al.*, 2023) than the others. To counter that, some methods have been proposed for classification, for instance, gradient modulation (Peng, Wei, Deng, Wang & Hu, 2022), Gradient-Blending (Wang, Tran & Feiszli, 2020), and Knowledge Distillation from the well-trained uni-modal model (Du *et al.*, 2021). In gradient modulation, Peng et al. proposed a mechanism to control the adaptive optimization of each modality by monitoring their contributions to the learning objective. In gradient blending, Wang et al. identified that multi-modal learning can overfit due to the increased capacity of the networks and proposed a mechanism to blend the gradients effectively (Wang *et al.*, 2020). Du et al. (Du *et al.*, 2021) show that training multi-modal models on joint training can suffer from learning inferior representations for each modality because of the imbalance of the modalities and the implicit bias of the common objectives in the fusion strategy. An effective approach to help on the modality imbalance in a shared encoder consists of using a Gradient Reversal Layer (GRL) (Ganin & Lempitsky, 2015), which was introduced for domain adaptation to reduce a network's reliance on a specific domain. GRL was exhaustively applied in object detection to create a shared domain; for instance, in the work of Chen et al. (Chen, Li, Sakaridis, Dai & Van Gool, 2018c), the GRL is used to adapt Faster R-CNN to distribution shifts in illumination or object appearance. The core idea of GRL involves training a classifier to identify the class of a data example during training. During backpropagation, the gradients are reversed to train the network to deceive the classifier.

In this work, we adapt this technique to address modality imbalance learning. Unlike typical cases where data belongs to a single domain/modality, a single training example of MiPa consists of a mosaic of the two modalities: RGB and IR. Therefore, our classifier is trained to predict a modality map instead. In our work, we tackle the imbalance with an adjustable balancing sampling, which learns how to effectively sample the RGB and IR patches during training, and a

patch-based GRL module responsible for encoding in the same vision encoder the information of both modalities while improving detection performance.

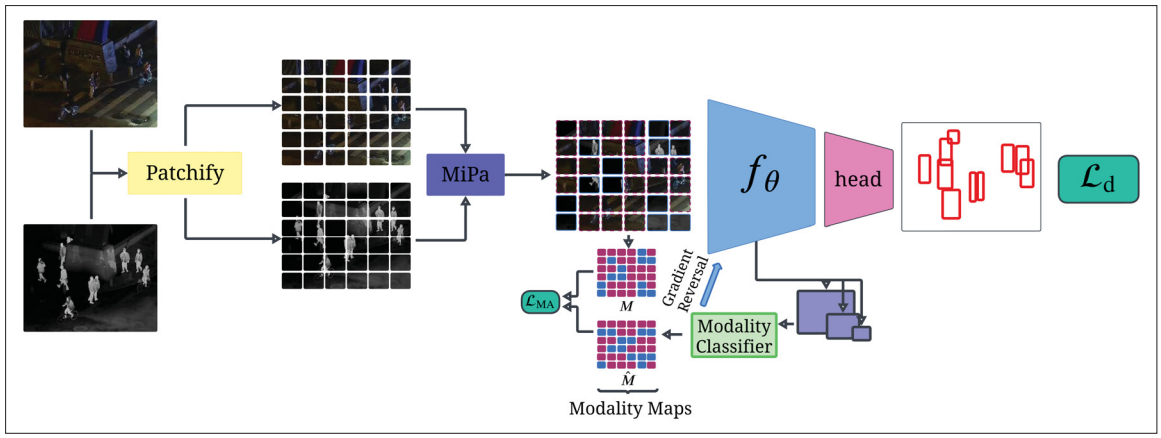


Figure 2.2 Mixed Patches (MiPa) with Modality Agnostic (MA) module. In yellow is the patchify function. In purple is the MiPa module, followed by the feature extractor (encoder). In green is the modality classifier, and in pink is the detection head

2.3 Proposed Method

While the naive way to create a multimodal vision encoder for an OD is to blend both modalities during training, we empirically show, in Section 2.4, that this approach leads to an imbalanced performance across modalities. In this section, we present our proposed solution.

2.3.1 Preliminary definitions

Let us consider a set of training samples $\mathcal{D} = \{(x_i, B_i)\}$ where $x_i \in \mathbb{R}^{W \times H \times C}$ is the image i with spatial resolution $W \times H$ and C channels. Here, a set of bounding boxes is represented by $B_i = \{b_0, b_1, ..., b_N\}$ with $b = (c_x, c_y, w, h)$ being c_x and c_y coordinates of the center of the bounding box with size $w \times h$. During the training process of a neural network-based detector, we aim to learn a parameterized function $f_\theta : \mathbb{R}^{W \times H \times C} \rightarrow \mathcal{B}$, being $\mathcal{B}$ the family of sets B_i and θ the parameters vector. For such, the optimization is guided by a loss function, which is a combination of a regression $\mathcal{L}_r$ and a classification $\mathcal{L}_c$ term, i.e., l_2 loss and binary

cross-entropy, respectively. The following Equation (2.1) defines a general loss function ($\mathcal{L}_d$) for object detection:

$$\mathcal{L}_d(\theta) = \frac{1}{|\mathcal{D}|} \sum_{(x,B)\in\mathcal{D}} \mathcal{L}_c(f_\theta(x), B) + \lambda \mathcal{L}_\text{r}(f_\theta(x), B). \tag{2.1}$$

2.3.2 Mixed Patches (MiPa)

The MiPa training method is a training technique that leverages the patch input channel from transformer-based feature extractors to build a powerful common representation between RGB/IR modalities for the unique vision encoder, which can be used in different transformer-based detectors. In short, it consists of a single encoder that receives sampling complementary patches from each modality and rearranges the input into a sort of mosaic image as shown in Figure 2.2. Such a mechanism forces the model to see both modalities for each inference without being forced to have parameters specialized on a specific one. Depending on how the nature of the patches are sampled, the technique can act as a way to gather the union of information between both modalities or as a regularization for the strongest modality (the easiest modality that tends to drive the learning process). Throughout this paper, we will reference the sampling ratio of the patches as ρ. There are several ways to pick the sampling ratio ρ; the naive way of selecting ρ is to use a fixed ratio during the training of 50%. Then, we can randomly generate a ρ value for each inference. If we have an intuition of which modality needs to be sampled more, we can manually move ρ during the training with a certain curriculum. Finally, we can let the model learn the optimal ratio by itself. In this work, we have explored all these variations to see which one is the most suitable for MiPa.

Theoretical explanation behind the MiPa approach. Here, we detail our theoretical understanding of the MiPa method. We refer to Table 2.1 for all definitions. The variable $\mathcal{X}$ can be thought of as a scene where you would see individuals walking in the street, for instance, and the functions f and g are camera lenses capturing the information of the scene via IR and RGB, respectively. The goal of MiPa ($\mathcal{M}$) is to enhance learning efficiency by merging information

Table 2.1 Definition of the random variables and information measures used to explain MiPa

General	
Input scene in patches	$\mathcal{X}_n$
Number of patches	$n \in \mathbb{N}$
Patch id	$i \in \mathbb{N}$
Random variables (RVs)	
Patch ratio	$\rho \sim U(0,1)$
Patch channel f	$m \sim \binom{n \cdot \rho}{p}$, p = $\frac{1}{2}$
Patch channel g	$l \sim n$ - m
Functions	
MiPa	$\mathcal{M}$
Self-Attention	SA
Modality channels	f, g
Information measures	
Entropy of V	$\mathcal{H}(V) \coloneqq \mathbb{E}_{p_V}[-\log p_V(V)]$
Information of X	Q, P where $Q = \mathcal{H}(q)$, $P = \mathcal{H}(p)$
Noise modality channels	η
Mutual information between P and Q	$I(Q,P) = Q + P$ - $Q \cap P$
Approximation of mutual information between P and Q	$I_a \approx I$

from both modalities, eliminating redundancy, and filtering out noise, all in a **single inference**. *Thus, say we have:*

$$f(\mathcal{X}) = P + \eta_f; g(\mathcal{X}) = Q + \eta_g, \tag{2.2}$$

where Equation (2.2) represents the visualization of the scene, which is composed of the information captured by the sensor (P or Q), P is information captured from the sensor of one modality and Q for the other modality, and some noise (η). Then the application of MiPa ($\mathcal{M}$)

can be summarized as the following Equation (2.3):

$$\mathcal{M}(f(\mathcal{X}), g(\mathcal{X})) = \begin{cases} f(\mathcal{X}_i), & i \in m \\ g(\mathcal{X}_i), & i \in l, \end{cases} \tag{2.3}$$

where $f(\mathcal{X}_i)$ represents the mapping of the patch $\mathcal{X}$ with id i using f (IR lens) and $g(\mathcal{X}_i)$ using RGB lens. Then, the combination of the individual patches of each modality is given by Equation (2.4):

$$\mathcal{M} = (P_0 + P_1 + Q_2 + ... + Q_{n-1} + P_n) + (m \cdot \eta_f + l \cdot \eta_g). \tag{2.4}$$

As RGB and IR patches do not encode the same information in the same patch visualization $\mathcal{X}_i$, the additional information of one modality improves, for instance, IR on the night, the other one. Also, this variation in the sense of information for both modalities is responsible for regularizing the training when the patches are mixed. The following Equation (2.5) represents the approximation of the real mutual information I by $\mathcal{M}$ using Equation (2.4) and approximating the noise from the scene to be similar for both sensors:

$$\mathcal{M} = I_a + \eta. \tag{2.5}$$

This approximation means that the encoded information on MiPa represents the total scene composed by both sensors, which are compressed on the vision encoder while removing the redundancy information and noise by the training process.

2.3.3 Patch-Wise Modality Agnostic Training

As previously mentioned, modality imbalances can potentially cause the model to rely mostly on one modality. Since the objective of this work is to preserve the original architecture of the model for inference, we opted for an approach where the backbone would be responsible for mediating the modalities. To do so, we designed an adaption of the GRL technique Ganin & Lempitsky

(2015) called ***patch-wise modality agnostic*** (MA) module. The key idea is to prevent the detector from relying too much on the strongest modality, the easiest modality driven by the learning process, by making the features from each modality indistinguishable, therefore sharing the same encoding. Considering that the input has a different modality for each patch, a modality that we pick during the patch mixing process, we build what we call a *modality map*, denoted as M, that specifies which modality each patch belongs to for each inference during training. Then, we use a modality classifier to predict the modality map of the features coming from the backbone. Finally, we compute the loss between the target and outputted modality maps and back-propagate the opposite gradients to the backbone encoder. To reduce the noise coming from the classifier at the beginning of the training, we slowly increase the weight (λ) of the gradients propagated to the backbone as the training goes on. We use the Binary Cross-Entropy (BCE) to compute the loss between the predicted and target modality maps, as described by the following Equation (2.6):

$$\mathcal{L}_{\text{MA}} = \frac{1}{n}\sum_{i=1}^{n} -M\log(\hat{M}) - (1-M)\log(1-\hat{M}), \tag{2.6}$$

where M is the modality map generated from ρ. The aforementioned approach for the full training pipeline can be seen in Figure 2.2. We use the following Equation (2.7) to increment the factor λ.

$$\lambda = \frac{2}{1+exp(-\gamma s)} - 1, \tag{2.7}$$

where s is the speed to which λ increases based on training epoch and γ is a hyperparameter to adjust this speed. The modality classifier can be used at any stage of the backbone; we have found empirically that using it on the features from the stage 1 works well. Finally, MiPa loss ($\mathcal{L}_{\text{MiPa}}$) can be defined as the following Equation (2.8):

$$\mathcal{L}_{\text{MiPa}} = \mathcal{L}_d + \lambda\mathcal{L}_{\text{MA}}. \tag{2.8}$$

2.4 Results and Discussion

2.4.1 Experimental Methodology

(a) Datasets: During our experiments, we explored two different RGB/IR benchmarking datasets: LLVIP and FLIR. **LLVIP:** The LLVIP dataset is a surveillance dataset composed of $12,025$ RGB/IR pairs of images for training and $3,463$ pairs for testing. The original resolution is 1280 by 1024 pixels but was resized to 640 by 512 to accelerate the training. The sole annotated class of this dataset is pedestrians. **FLIR ALIGNED:** For the FLIR dataset, we used the sanitized and aligned paired sets provided by Zhang et al. (Zhang *et al.*, 2020b), which has $4,129$ aligned pairs for training and $1,013$ pairs for testing. The FLIR images are taken from the perspective of a camera in front of a car, and the resolution is 640 by 512. It contains annotations of bicycles, dogs, cars, and people. It has been found that for the case of FLIR, the “dog" objects are inadequate for training (Cao *et al.*, 2023a), but since our objective is to evaluate if our method can make a detector modality agnostic and not beat any prior benchmark, we have decided to keep it during our evaluations.

(b) Implementation Details: All detectors were trained on an A100 NVIDIA GPU and were implemented using PyTorch. We use AdamW (Loshchilov & Hutter, 2019) as an optimizer with a learning rate of $1e^{-4}$, a batch size of 6, and for a total of 12 epochs for the case of the DINO (Zhang *et al.*, 2023b) OD. For Swin, we start with the pre-trained weights from ImageNet (Russakovsky *et al.*, 2015). The models are evaluated in terms of performance AP_{50}, and we additionally reported the AP_{75} and AP in the supplementary material. The evaluation is also performed in terms of RGB performance, IR performance, and our target metric, the average of both, because our setup requires a model that is equally good on both modalities during test time. In this work, we replicate the 1-channel IR to have 3-channel input for further use with 3-channel RGB data.

(c) Baseline Methods: In the course of this work, we considered different baselines to compare to our proposed method (MiPa). Firstly, we measure the performance of the detector trained

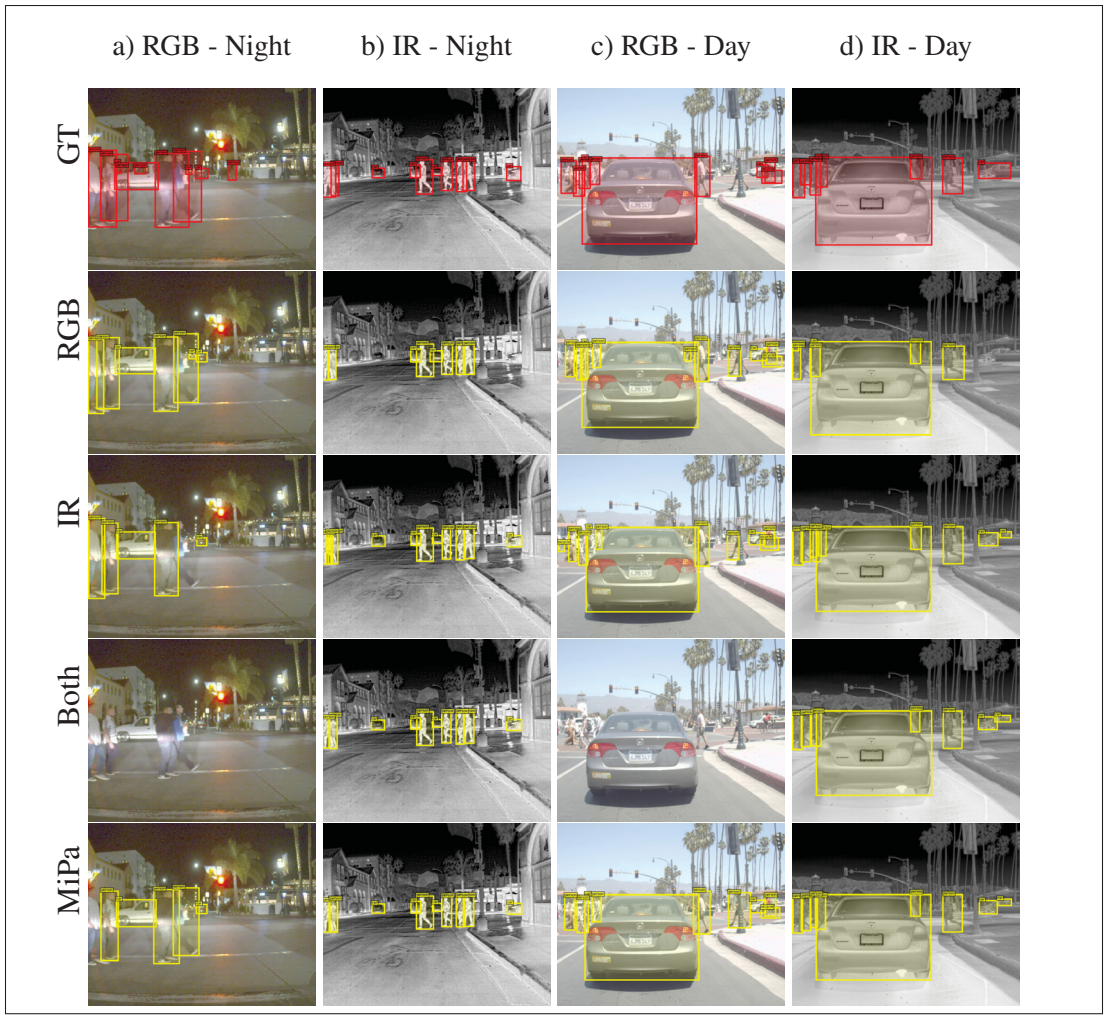


Figure 2.3 Detection over different methods for two different daytimes: Night and Day and two different modalities: RGB and IR. Detectors trained on *RGB* work better in the daytime. Detectors trained on *IR* work better at nighttime. Detectors trained on *Both* modalities in a naive way cannot work only on the dominant modality. Our *MiPa* manages to work well in all conditions

on one modality, unimodal setup, to gain a reference of the expected detection coming from each modality. Secondly, we evaluate the naive solution of simply using a dataset comprised of both modalities during training (multimodal setting), which we call *Both*. To account for the modality imbalances and further increase the fairness of our comparisons, we balanced the datasets with 25%, 50%, and 75% of one modality and the rest of the other. All models were

evaluated separately on RGB and IR. Additionally, the mean of the modalities, which represents how well the model is balanced for the two desired modalities, is calculated.

Table 2.2 Comparison of different ratio ρ sampling methods on LLVIP. Using DINO with Swin backbone

Model	Dataset: LLVIP ($AP_{50} \uparrow$)		
	RGB	**IR**	**Average**
Fixed [ρ=0.25]	78.9	**98.2**	88.55
Fixed [ρ=0.50]	73.0	97.6	85.30
Fixed [ρ=0.75]	77.4	97.5	87.45
Curriculum (ρ=0.25 / 4 epochs)	76.6	97.8	87.20
Curriculum (ρ=0.25 / 8 epochs)	80.1	97.8	88.95
Variable	**88.5**	97.5	**93.00**

2.4.2 Towards the optimal ρ

Since the way of selecting the ideal ρ was unclear, we designed different experimental settings to study the influence of ρ on learning the best way to balance the amount of RGB/IR information during the training. Let us start with a few definitions:

- Fixed ρ. In this setting, we selected a fixed proportion of RGB/IR samples, such as 0%, 25%, 50%, 75% and 100%, in which 0% correspond to no IR images in the training batch, and 100% correspond to only IR in the batch.

- Curriculum ρ. For this strategy, we analyzed which modalities were easy to learn; in this case, it was IR. Then, during the initial epochs over the training, the model focuses on the easier-to-learn modality (IR modality tends to drive the learning process when a balanced jointly dataset is given), providing between 0% to 25% of ratio for IR, which means that the model for the initial epochs is going to see more RGB data, which is harder. Then, over the rest of the training epochs, it samples from the uniform distribution such as variable ρ.

- Variable ρ. In the variable ρ, the ratio of mixed patches per batch is drawn from a uniform distribution. For each batch, a different ρ is redrawn.

We tested all the different configurations of ρ on LLVIP (see Table 2.2). For this experiment, we have made two findings. First, using an I_a following a uniform distribution gives us a better approximation of the range of information from $IR \cup RGB$ as the results from the variable give us a better balance between both modalities. Second, using less of the weaker modality (hard to learn) strengthens the learning of the strongest one (easier to learn modality), as it can be seen in Sec. 2.4.4 (Table 2.5), that we were actually able to beat the state-of-the-art by sampling 25% of RGB images and 75% of IR.

2.4.3 Patch-wise Modality Agnostic Training

The subsequent ablation shows the efficacy of the patch-wise modality agnostic method towards obtaining a single model capable of dealing with both modalities while keeping the performance stable, see Table 2.3. Additionally, we studied the sensibility of the model performances influenced by different γ hyperparameters (see Table 2.4), seen in Equation (2.7), which tunes the speed that the λ factor increases at each step the weight of gradients propagated to the encoder. We empirically demonstrate that the optimal γ varies between datasets and detectors due to the number of epochs required for each one, whereas if the model requires more training epochs, the γ should be higher. Additionally, MiPa was designed for computational efficiency during test time, so it does not increase the computational cost when the model is deployed in real-world scenarios.

2.4.4 Comparison with RGB/IR Competitors

In this section, we compare our approach in terms of detection performance with other strong methods in the literature that use RGB/IR modalities. Table 2.5 shows that MiPa is a competitive method under RGB/IR benchmarks. For instance, on FLIR, MiPa has 81.3 AP_{50}, while CSSA (Cao *et al.*, 2023a) has 79.2, ProbEn (Chen *et al.*, 2022) has 75.5, GAFF (ZHANG *et al.*,

Table 2.3 Comparison of detection performance over different baselines and MiPa for different models on Swin backbone for DINO and Deformable DETR. The evaluation is done for RGB, IR, and the average of the modalities

Detector	Model	Dataset: LLVIP (AP_{50} ↑)		
		RGB	**IR**	**Average**
DINO	RGB	90.87 ± 0.84	94.23 ± 0.57	92.55
	IR	66.87 ± 0.90	96.87 ± 0.12	81.87
	Both [ρ = 0.25]	79.73 ± 1.03	97.40 ± 0.22	88.57
	Both [ρ = 0.50]	82.40 ± 1.50	96.50 ± 0.29	89.45
	Both [ρ = 0.75]	81.23 ± 2.89	97.07 ± 0.25	89.15
	MiPa (Ours)	88.70 ± 0.45	96.97 ± 0.26	**92.83**
	MiPa + MA (Ours)	89.10 ± 0.28	96.83 ± 0.09	**92.90**
Def.DETR	RGB	80.00 ± 1.50	90.03 ± 00.87	85.02
	IR	56.10 ± 2.50	94.20 ± 00.08	75.15
	Both [ρ = 0.25]	51.20 ± 3.47	83.73 ± 16.57	67.47
	Both [ρ = 0.50]	53.57 ± 4.17	83.87 ± 16.17	68.72
	Both [ρ = 0.75]	53.53 ± 4.55	82.33 ± 18.48	67.93
	MiPa (Ours)	78.60 ± 0.42	95.20 ± 0.16	**86.90**
	MiPa + MA (Ours)	79.02 ± 0.21	95.36 ± 0.25	**87.19**
Detector	**Model**	**Dataset: FLIR (AP_{50} ↑)**		
		RGB	**IR**	**Average**
DINO	RGB	66.07 ± 0.98	56.60 ± 0.80	61.33
	IR	56.47 ± 0.79	70.40 ± 0.38	63.43
	Both [ρ = 0.25]	56.53 ± 0.76	67.57 ± 1.73	62.05
	Both [ρ = 0.50]	60.50 ± 0.66	68.93 ± 0.60	64.72
	Both [ρ = 0.75]	58.53 ± 0.92	70.43 ± 0.65	64.48
	MiPa (Ours)	63.53 ± 1.94	69.50 ± 1.84	**66.52**
	MiPa + MA (Ours)	64.80 ± 2.30	70.43 ± 0.53	**67.62**
Def.DETR	RGB	49.33 ± 1.39	43.77 ± 00.56	46.55
	IR	39.17 ± 1.48	59.20 ± 00.29	49.18
	Both [ρ = 0.25]	35.73 ± 4.95	43.00 ± 13.54	39.37
	Both [ρ = 0.50]	33.93 ± 5.15	43.33 ± 14.14	38.63
	Both [ρ = 0.75]	32.90 ± 3.54	44.13 ± 14.85	38.52
	MiPa (Ours)	48.00 ± 0.57	54.97 ± 00.90	**51.48**
	MiPa + MA (Ours)	48.27 ± 1.76	55.80 ± 00.22	**52.03**

2021) 74.6 and Halfway Fusion (Zhang *et al.*, 2020b) 71.5, RSDet (Zhao *et al.*, 2024a) 81.1 and CrossFormer (Lee *et al.*, 2024) 79.3. Furthermore, we report competitive results on LLVIP, which can be seen in the table as the people detection performance over different methods inclusively; for competitors, both modalities are used during training and inference, which is not

Table 2.4 MiPa ablation on γ and comparison with different baselines for DINO Swin. The evaluation is done for RGB, IR, and the average of the modalities in terms of AP_{50} performance

Modality	Dataset: LLVIP (AP_{50} ↑)		
	RGB	**IR**	**Average**
RGB	90.87 ± 0.84	94.23 ± 0.57	92.55
IR	66.87 ± 0.90	96.87 ± 0.12	81.87
Both [ρ = 0.25]	79.73 ± 1.03	97.40 ± 0.22	88.57
Both [ρ = 0.50]	82.40 ± 1.50	96.50 ± 0.29	89.45
Both [ρ = 0.75]	81.23 ± 2.89	97.07 ± 0.25	89.15
MiPa	88.70 ± 0.45	96.97 ± 0.26	92.83
MiPa [γ = 0.05]	89.20 ± 0.43	96.57 ± 0.39	92.88
MiPa [γ = 0.10]	89.43 ± 0.25	96.57 ± 0.31	**93.00**
MiPa [γ = 0.15]	89.10 ± 0.28	96.83 ± 0.09	92.97
Modality	**Dataset: FLIR (AP_{50} ↑)**		
	RGB	**IR**	**Average**
RGB	66.07 ± 0.98	56.60 ± 0.80	61.33
IR	56.47 ± 0.79	70.40 ± 0.38	63.43
Both [ρ = 0.25]	56.53 ± 0.76	67.57 ± 1.73	62.05
Both [ρ = 0.50]	60.50 ± 0.66	68.93 ± 0.60	64.72
Both [ρ = 0.75]	58.53 ± 0.92	70.43 ± 0.65	64.48
MiPa	63.53 ± 1.94	69.50 ± 1.84	66.52
MiPa [γ = 0.05]	64.80 ± 2.30	70.43 ± 0.53	**67.62**
MiPa [γ = 0.10]	64.03 ± 2.11	69.63 ± 1.45	66.83
MiPa [γ = 0.15]	64.27 ± 0.47	69.93 ± 1.02	67.10

our case (as we just use the IR modality for inference, in Table 2.5). For example, in LLVIP, MiPa reached 98.8 AP_{50}, and the second best was CFT with 97.5.

2.5 Conclusion

In this work, we have introduced a novel training method leveraging a patch-based strategy using a single vision encoder for OD to consolidate the mutual information between different modalities. This method, named MiPa, has enabled two different object detectors, DINO (Zhang *et al.*, 2023b) and Deformable DETR (Zhu *et al.*, 2021), to achieve *modality invariance* on LLVIP and FLIR datasets without having to make any specific changes for each modality, for example, additional encoding parameters for each modality, to their architecture or increase the

Table 2.5 Comparison with different multimodal works on RGB/IR benchmarks

Method	**Dataset**					
	FLIR			**LLVIP**		
	AP_{50}	**AP_{75}**	**AP**	**AP_{50}**	**AP_{75}**	**AP**
Halfway F. (Zhang *et al.*, 2020b)	71.5	31.1	35.8	91.4	60.1	55.1
GAFF (ZHANG *et al.*, 2021)	74.6	31.3	37.4	94.0	60.2	55.8
ProbEn (Chen *et al.*, 2022)	75.5	31.8	37.9	93.4	50.2	51.5
CSSA (Cao *et al.*, 2023a)	79.2	37.4	41.3	94.3	66.6	59.2
CFT (Qingyun, Dapeng & Zhaokui, 2021)	78.7	35.5	40.2	97.5	72.9	63.6
DIVFusion (Tang, Xiang, Zhang, Gong & Ma, 2023a)	-	-	-	89.8	-	52.0
RSDet (Zhao *et al.*, 2024a)	81.1	-	41.4	95.8	-	61.3
CrossFormer (Lee *et al.*, 2024)	79.3	38.5	42.1	97.4	75.4	65.1
MiPa (Ours)	**81.3**	**41.8**	**44.8**	**98.2**	**78.1**	**66.5**

testing inference time. Additionally, our method outperformed competitors on both datasets. Furthermore, we provide a definition from information theory regarding the knowledge captured by the MiPa method.

CHAPTER 3

HALLUCINATING RGB MODALITY FOR PERSON DETECTION THROUGH PRIVILEGED INFORMATION

Heitor R. Medeiros[a] , Fidel A. Guerrero Peña[a] , Masih Aminbeidokhti[a] , Thomas Dubail[a], Eric Granger[a], Marco Pedersoli[a]

[a] Department of Systems Engineering, École de technologie supérieure, 1100 Notre-Dame West, Montreal, Quebec, Canada H3C 1K3

Paper published in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), January 2024*

Abstract

A powerful way to adapt a visual recognition model to a new domain is through image translation. However, common image translation approaches only focus on generating data from the same distribution as the target domain. Given a cross-modal application, such as pedestrian detection from aerial images, with a considerable shift in data distribution between infrared (IR) to visible (RGB) images, a translation focused on generation might lead to poor performance as the loss focuses on irrelevant details for the task. In this paper, we propose HalluciDet, an IR-RGB image translation model for object detection. Instead of focusing on reconstructing the original image on the IR modality, it seeks to reduce the detection loss of an RGB detector, and therefore avoids the need to access RGB data. This model produces a new image representation that enhances objects of interest in the scene and greatly improves detection performance. We empirically compare our approach against state-of-the-art methods for image translation and for fine-tuning on IR, and show that our HalluciDet improves detection accuracy in most cases by exploiting the privileged information encoded in a pre-trained RGB detector. Code: https://github.com/heitorrapela/HalluciDet.

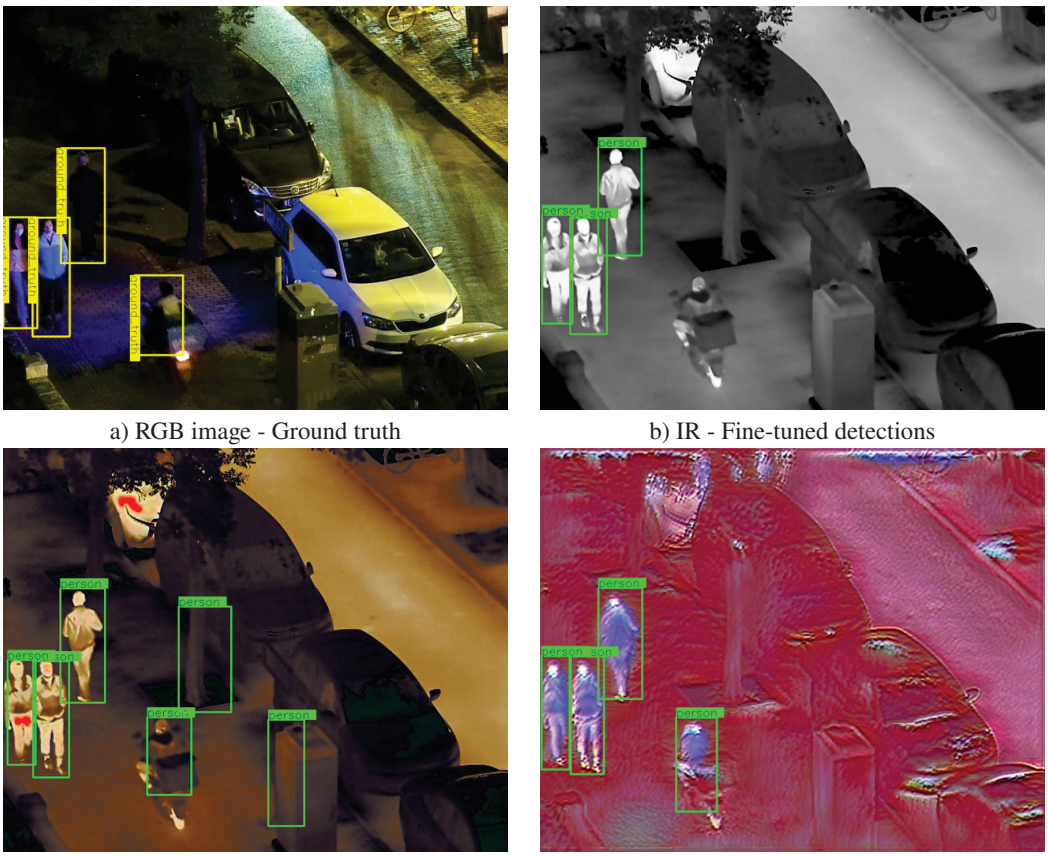


Figure 3.1 Example of detections using baseline and HalluciDet methods on LLVIP data. (a) Original RGB image with ground truth annotations (yellow). (b) IR image with corresponding detections of a fine-tuned model (green). (c) Translated image from IR to RGB produced by FastCUT and corresponding RGB detections (green). (d) Hallucinated image produced by our method and RGB detections (green); HalluciDet does not seek to reconstruct all image details but only to enhance the objects of interest

3.1 Introduction

The proliferation of hardware sensors has greatly advanced the collection of large-scale datasets. Such datasets have significantly improved the performance of deep learning (DL) algorithms across various fields, including surveillance (Chen *et al.*, 2019a), industrial monitoring (Kong & Ge, 2021), self-driving cars (Stilgoe, 2018), and robotics (Pierson & Gashler,

2017). By providing high-resolution data, these sensors offer additional observations of common environmental phenomena to aid in the effectiveness of DL algorithms (Ramachandram & Taylor, 2017).

The additional information from different sensors has been employed in diverse settings (Samaras *et al.*, 2019; Huang & Chen, 2020). In computer vision applications, combining sensors with distinct environmental sensing perspectives, such as varying points of view and modality sensing information, can increase model performance, enabling possibilities that were previously unavailable. Furthermore, in the context of self-driving cars and intelligent building applications, two modalities commonly used are visible (RGB) and infrared (IR) (Takumi *et al.*, 2017). In particular, the RGB modality offers valuable information for tasks like object detection, which generates bounding boxes for target objects within colored images. These colored images are known to have more diverse information due to their characteristics on the RGB light spectrum, especially in the presence of light. Thus, these RGB sensors are preferred to be used in daily activities where there is the presence of sunlight. On the other hand, the IR spectrum provides additional information for the visible modality when the light is low, especially during the night (Jia *et al.*, 2021), and also complementary information, primarily related to thermal sensing. Furthermore, IR is vastly applied in surveillance applications (Zhang *et al.*, 2018a), which require the device to capture information in light-restricted environments. IR object detection is known to detect objects using IR radiation emitted from the object, which varies depending on the object's material.

Despite the impressive performance of DL models, their effectiveness can significantly deteriorate when applied to modalities that were not present during the training (Dai & Van Gool, 2018; Torralba & Efros, 2011). For example, a model trained on RGB images may not perform well on IR images during testing (Yang, Yu, Zhao & Wang, 2020). To address the issue, some studies utilize image-to-image translation techniques to narrow the gap between modalities distributions. Typically, these methods employ classical pixel manipulation techniques or deep neural networks to generate intermediate representations, which are then fed into a detector trained on the source modality. However, transitioning from IR to RGB has proven challenging due to generating

color information while filtering out non-meaningful data associated with diverse heat sources. This challenge is particularly pronounced when the target category is also a heat-emitting source, such as a person.

In this work, we argue that achieving a robust intermediate representation for a given task needs guiding the image-to-image translation using a task-specific loss function. Here, we introduce HalluciDet, a novel approach for image translation focusing on detection tasks. Inspired by the learning using privileged information (LUPI) paradigm (Vapnik & Vashist, 2009), we utilize a robust people detection network previously trained on an RGB dataset to guide our translation process from IR to RGB. Our translation approach relies on an annotated IR dataset and an RGB detector to identify the appropriate representation space. The ultimate goal is to find a translation model, hereafter referred to as the Hallucination network, capable of translating IR images into meaningful representation to achieve accurate detections with an RGB detector.

Our main contributions can be summarized as follows:

(1) We propose HalluciDet, a novel approach that leverages privileged information from pre-trained detectors in the RGB modality to guide end-to-end image-to-image translation for the IR modality.

(2) Given that our model focuses on the IR detection task, HalluciDet uses a straightforward yet powerful image translation network to reduce the domain gap between IR-RGB modalities, guided by the proposed hallucination loss function incorporating standard object detection terms.

(3) Through experiments conducted on two challenging IR-RGB datasets (LLVIP and FLIR ADAS), we compare HalluciDet against various image-to-image translation and traditional pixel manipulation methods. Our approach is seen to improve detection accuracy on the IR modality by incorporating privileged information from RGB.

3.2 Related Work

3.2.1 Object detection.

Different from classification tasks, in which we want only to classify the object category, in object detection, additionally, the task is to know specific positions of the objects (Zhang *et al.*, 2023a). Deep learning object detection methods are categorized as two-stage and one-stage detectors. The two-stage detector extracts regions of interest or proposals for a second-stage classifier. Then, the second stage is responsible for classifying if there is an object in that region. One commonly used two-stage detector is the Faster R-CNN proposed by (Ren *et al.*, 2015). It is the first end-to-end DL object detector to reach real-time speed. The speedup was achieved by introducing the Region Proposal Network (RPN), a network responsible for the region proposals without impacting the computational performance compared with previous region proposals algorithms (Ren *et al.*, 2016). The one-stage detectors mainly focus on end-to-end training and real-time inference speed of the object detectors. In this scenario, the object detector has a single neural network to extract the features for the regression of the bounding box and give the class probabilities without an auxiliary network for the region proposals. Recently, there are detectors that were developed to remove the requirement of defining anchor boxes during training. For instance, the Fully Convolutional One-Stage Object Detection (FCOS) is one of these models that, due to its nature, reduces all complicated computation related to anchor boxes, which can lead to an increase in inference time.

3.2.2 Learning using Privileged Information (LUPI).

In human learning, the role of a teacher is crucial, guiding the students with additional information, such as explanations, comparisons, and so on (Vapnik & Vashist, 2009). In the LUPI setting, during the training, we have additional information provided by a teacher to help the learning procedure. Since the additional information is available at the training stage but not during the test time, we call it privileged information (Vapnik & Vashist, 2009). Recently, (Lambert *et al.*, 2018) proposed the usage of privileged information to guide the variance of a Gaussian dropout.

In a classification scenario, additional localization information is used, and its results show that it improves the generalization, requiring fewer samples for the learning process (Lambert *et al.*, 2018). (Motiian *et al.*, 2016) designed a large-margin classifier using information bottleneck learning with privileged information for visual recognition tasks. In the object detection problem, (Hoffman *et al.*, 2016) was the first work to present a modality hallucination framework, which incorporates the training RGB and Depth images, and during test time, RGB images are processed through the multi-modal framework to improve the performance of the detection. The modality hallucination network is responsible for mimicking depth mid-level features using RGB as input during the test phase. Liu *et al.* (2021a) used depth as privileged information for object detection with a Depth-Enhanced Deformable Convolution Network. In this work, we use the privileged information coming from a pre-trained RGB detector to improve the performance of the infrared detection. In practice, instead of destroying the information of the RGB detector by fine-tuning, we use the RGB detector as a guide for translating the IR input image into a new representation, which can help the RGB detector boost performance by enhancing the objects of interest.

3.2.3 Image Translation.

The objective of image translation is to learn a mapping between two given domains such that images from the source domain can be translated to the target domain. In other words, the aim is to find a function $h_\vartheta : \mathcal{X}_\mathrm{s} \rightarrow \mathcal{X}_\mathrm{t}$ such that the distribution of images $h_\vartheta(\mathcal{X}_\mathrm{s})$ in the translated domain is close to the distribution of images $\mathcal{X}_\mathrm{t}$ in the target domain. Early methods rely on autoencoders (AEs) (Hinton & Zemel, 1993) and generative adversarial networks (GANs) (Goodfellow *et al.*, 2014) to learn cross-domain mapping. Unsupervised AE methods aim to learn a representation of the data by reconstructing the input data. GANs are a type of generative model that can learn to generate new data that is similar to the training data. More recently, diffusion models have gained popularity. They are capable of generating high-quality images but lack some properties for domain translation, like on CycleGANs. For improving models such as CycleGAN, techniques such as Contrastive Unpaired Translation (CUT) (Park,

Efros, Zhang & Zhu, 2020a) and FastCUT (Park *et al.*, 2020a) were developed. CUT is an image translation model based on maximizing mutual information of patches, which is faster than previous methods while providing results as good as others. On RGB/IR modalities, the InfraGAN (Özkanoğlu & Ozer, 2022) proposes an image-level adaptation using a model based on GANs, but for RGB to IR adaptation, with a focus on the quality of the generated images, thus optimizing image quality losses. Additionally, using image translation for object detection on RGB/IR using pre-train models, Herrmann et al. (Herrmann, Ruf & Beyerer, 2018) used RGB object detectors without changing their parameters. The IR images are adapted to the RGB images using traditional computer vision pre-processing at the image level before applying it as input to the RGB object detector.

None of these methods provides an end-to-end way to directly train the image translation methods for detection applications. Furthermore, traditionally, they require more than one kind of data set composed of the original domain and the target domain. For instance, CycleGAN is based on adversarial loss, and U-net is based on reconstruction loss. Thus, if we have access to the already trained detector on the original domain, this knowledge can possibly be used during the learning of the translation network.

3.3 Proposed Method

3.3.1 Preliminary definitions.

Let $\mathbf{x}_\mathrm{i}$ be a given image with spatial resolution $W \times H$ and C channels. An object detector aims to output a set of N_reg object proposals, each represented as a bounding box $\mathbf{b}_{\mathrm{i,j}} = (c, d, e, w, h)$, where (d, e) is the location of the top-left pixel of the bounding box for the j-th object, and w and h are the width and height of the object, respectively. Additionally, a classification label $c \in \{1, 2, ..., N_\mathrm{cls}\}$ is assigned to each object of interest representing the region's class. In terms of optimization, such a task aims to maximize the detection accuracy, which typically is approximated through the average precision (AP) metric over all classes. Then, to train a

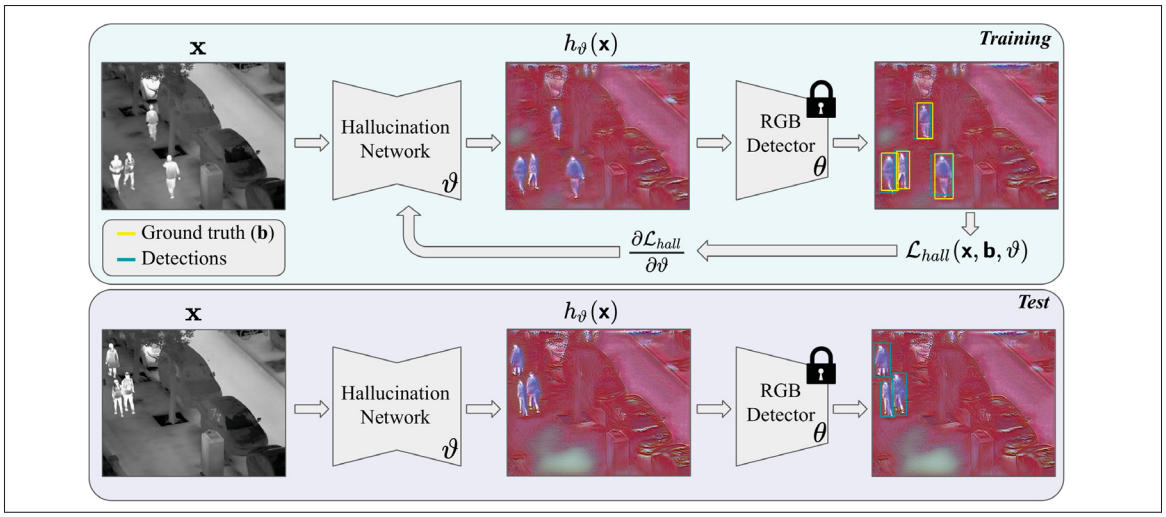


Figure 3.2 HalluciDet leverages privileged information for modality hallucination with pre-trained detectors. During training, the hallucination network learns how to use the privileged information encoded by the RGB detector to translate the IR image into a new hallucination modality representation. Then, during inference, the model provides better IR detection using the translated modality

detector, formally defined as the mapping $f_\theta : \mathbf{x}_\mathrm{i} \rightarrow \hat{\mathbf{b}}_\mathrm{i}$, a differentiable surrogate for AP metric is used, also known as the detection loss function, $\mathcal{L}_\mathrm{det}(\mathbf{b}, \mathbf{x}; \theta)$.

The detection loss can be divided into two terms. The first one is the classification loss $\mathcal{L}_\mathrm{cls}(\hat{y}_\mathrm{c}, y_\mathrm{c})$ responsible for learning the class label c. In this work, we use the cross-entropy loss function to assess the matching of bounding boxes categories $\mathcal{L}_\mathrm{ce}(\hat{y}_\mathrm{c}, y_\mathrm{c}) = -\frac{1}{N_\mathrm{cls}} \sum_{j=1}^{N_\mathrm{cls}} y_{\mathrm{c_j}} \log(p_\mathrm{j})$, where N_cls is the total number of classes, and $y_{\mathrm{c_j}}$ is the class indicator function, i.e., $y_{\mathrm{c_j}} = 1$ if c_j is the true class of the object, or $y_{\mathrm{c_j}} = 0$ otherwise. The probability provided by the detector for each category j is p_j. To ensure the right positioning of the object, a second regression term $\mathcal{L}_\mathrm{reg}(\hat{\mathbf{y}}_\mathrm{b}, \mathbf{y}_\mathrm{b})$ is used, being the $\mathcal{L}_\mathrm{L1}(\hat{\mathbf{y}}_{\mathrm{b_i}}, \mathbf{y}_{\mathrm{b_i}}) = \sum_{i=1}^{N_\mathrm{reg}} \left| \mathbf{y}_{\mathrm{b_i}} - \hat{\mathbf{y}}_{\mathrm{b_i}} \right|$ and $\mathcal{L}_\mathrm{L2}(\hat{\mathbf{y}}_{\mathrm{b_i}}, \mathbf{y}_{\mathrm{b_i}}) = \sum_{i=1}^{N_\mathrm{reg}} (\mathbf{y}_{\mathrm{b_i}} - \hat{\mathbf{y}}_{\mathrm{b_i}})^2$ losses the most commonly employed in the literature. Here N_reg is the number of bounding boxes on the image $\mathbf{x}_\mathrm{i}$. Then, the final detection loss function can be defined in general terms as:

$$\mathcal{L}_\mathrm{det}(\mathbf{x}, \mathbf{b}; \theta) = \mathcal{L}_\mathrm{cls}(f_\theta(\mathbf{x}), c) + \lambda \cdot \mathcal{L}_\mathrm{reg}(f_\theta(\mathbf{x}), \mathbf{b}), \tag{3.1}$$

where λ is a hyperparameter that controls the balance between the two terms, and θ is a vector containing the detector learnable parameters. The detectors used in this work use this general objective during their optimization process. However, they adapt each term to their specific architecture.

3.3.2 HalluciDet.

Our goal is to generate a representation from an IR image that a given RGB detector can effectively process. Let $\mathcal{X} \subset \mathbb{R}^{W \times H}$ be the set of IR data containing N images. During the learning phase, a training dataset $S = \{(\mathbf{x}_i, \mathbf{b}_i)\}$ is given such that $\mathbf{x}_i \in \mathcal{X}$ is an IR image and $\mathbf{b}_i$ is a set of bounding boxes as defined in the previous section. In addition, an RGB detector f_θ is also available. Then, a representation mapping is here defined as $h_\vartheta : \mathcal{X} \rightarrow \mathcal{R}$, where $\mathcal{R}$ is the representation space and ϑ are the learnable parameters of the translation model. Such a representation space, $\mathcal{R} \subset \mathbb{R}^{W \times H \times 3}$, is conditioned to the subset of plausible RGB images that are sufficient to obtain a proper response from the RGB detector f_θ. To find such a mapping we solve the optimization problem $\vartheta^* = \arg\min_\vartheta \mathcal{L}_{\text{hall}}(\mathbf{x}, \mathbf{b}; \vartheta)$ which implicitly uses the composition $(h_\vartheta \circ f_\theta)(\mathbf{x})$ to guide the intermediate representation.

Our proposed model, HalluciDet, comprises two modules: a hallucination network responsible for the image-to-image harmonization and a detector. The Hallucination network is based on U-net (Ronneberger *et al.*, 2015), but modified with attention blocks which are more robust for image translation tasks (Li, Xiong, An & Wang, 2018b; Fan, Wang, Li & Wang, 2020). For training the HalluciDet, we train the hallucination module and condition it with the detection loss, which is the only supervision necessary for guiding the hallucination training with respect to the privileged information of the pre-trained RGB detector. This phase is responsible for translating the hallucinated image to a new representation close to the RGB modality. Please note that this strategy helps the final model to perform well on the IR modality without changing the knowledge from the detector. Under this framework, the RGB detection performance remains the same since the detector's parameters θ are not updated during the adaptation learning. On the other hand, detections over IR images are obtained by adapting the input using the Hallucination

network, followed by the evaluation over the RGB detector. As a side advantage, our model allows evaluating both modalities by providing the appropriate modality identifier during the forward pass, i.e., RGB or IR. Figure 3.2 depicts the training and evaluation process of an IR image using privileged information from the RGB detector.

The detector f_θ layers are frozen, thus preserving the prior knowledge, but the weights ϑ of the hallucination network h_ϑ are updated during the backward pass. The input minibatch is created with images from $\mathcal{X}$ set, leading to the hallucinated minibatch, which is then evaluated on f_θ to obtain the associated detections. To find the appropriate representation space, the hallucination loss $\mathcal{L}_{\text{hall}}(\mathbf{x}, \mathbf{b}, \vartheta)$ drives the optimization by updating only the hallucination network parameters. The representation space $\mathcal{R}$ is guided by $\mathcal{L}_{\text{hall}}$ to be closer enough to the RGB modality, which allows the detector to make successful predictions. As the representation is being learned with feedback from the frozen detector, it extracts the previous knowledge so that this new intermediate representation is tuned for the final detection task. The proposed hallucination loss shares some similarities with the aforementioned detection loss but with the distinction of only updating the modality adaptation parameters:

$$\mathcal{L}_{\text{hall}}(\mathbf{x}, \mathbf{b}, \vartheta) = \mathcal{L}_{\text{cls}}(f_\theta(h_\vartheta(\mathbf{x})), c) + \lambda \cdot \mathcal{L}_{\text{reg}}(f_\theta(h_\vartheta(\mathbf{x})), \mathbf{b}) \tag{3.2}$$

Equation 3.2 is optimized w.r.t ϑ. We added the hyperparameter λ to weigh the contribution of each term and for numerical stability purposes.

3.4 Experimental results and analysis

3.4.1 Experimental Methodology.

Hallucidet is evaluated on two different popular IR/RGB datasets, the LLVIP (Jia *et al.*, 2021), and FLIR ADAS (Flir, 2021). The LLVIP dataset is composed of $30,976$ images, in which $24,050$ ($12,025$ IR and $12,025$ RGB paired images) are used for training and $6,926$ for testing ($3,463$ IR and $3,463$ RGB paired images). For the FLIR, we used the sanitized and aligned

Table 3.1 Performance comparison of models on IR images using LLVIP dataset (Jia *et al.*, 2021). The table showcases the impact of different approaches, including pixel manipulation techniques, U-Net, CycleGAN, CUT, FastCUT, and HalluciDet. The detectors were trained with RGB data and evaluated on IR. To make a fair comparison with our models, we decided to start with models that do not have strong data augmentation that could benefit one modality over the other

Image-to-image translation	**Learning strategy**	**AP@50↑**		
		Test Set (Dataset: LLVIP)		
		FCOS	RetinaNet	Faster R-CNN
Blur (Herrmann *et al.*, 2018)	-	42.59 ± 4.17	47.06 ± 1.99	63.05 ± 1.96
Histogram Equalization (Herrmann *et al.*, 2018)	-	33.10 ± 4.64	36.45 ± 2.02	51.47 ± 4.03
Histogram Stretching (Herrmann *et al.*, 2018)	-	38.55 ± 4.25	41.97 ± 1.39	57.69 ± 2.78
Invert (Herrmann *et al.*, 2018)	-	53.62 ± 2.07	55.43 ± 2.03	71.83 ± 3.04
Invert + Equalization (Herrmann *et al.*, 2018)	-	50.03 ± 2.44	52.57 ± 1.50	68.69 ± 2.73
Invert + Equalization + Blur (Herrmann *et al.*, 2018)	-	50.58 ± 2.41	52.62 ± 1.36	68.91 ± 2.74
Invert + Stretching (Herrmann *et al.*, 2018)	-	51.48 ± 2.17	52.87 ± 1.80	69.34 ± 3.07
Invert + Stretching + Blur (Herrmann *et al.*, 2018)	-	51.54 ± 1.92	52.96 ± 1.80	69.59 ± 2.90
Parallel Combination (Herrmann *et al.*, 2018)	-	50.18 ± 2.25	52.52 ± 1.39	68.14 ± 2.98
U-Net (Ronneberger *et al.*, 2015)	Reconstruction	42.94 ± 4.14	47.35 ± 1.92	63.23 ± 2.03
CycleGAN (Zhu, Park, Isola & Efros, 2017b)	Adversarial	22.76 ± 1.94	27.04 ± 4.23	38.92 ± 5.09
CUT (Park, Efros, Zhang & Zhu, 2020b)	Contrastive learning	19.16 ± 2.10	21.61 ± 2.09	35.17 ± 0.32
FastCUT (Park *et al.*, 2020b)	Contrastive learning	46.87 ± 2.28	52.39 ± 2.31	67.73 ± 2.14
HalluciDet (ours)	Detection	**63.28 ± 3.49**	**56.48 ± 3.39**	**88.34 ± 1.50**

paired sets provided by Zhang et al. (Zhang *et al.*, 2020b), which have 10, 284 images, being 8, 258 for training (4, 129 IRs and 4, 129 RGBs) and 2, 026 (1, 013 IRs and 1, 013 RGBs) for test. We chose to utilize these paired IR/RGB datasets to ensure a fair comparison with other image-to-image translation techniques that employ reconstruction losses. In our experiments, we use 80% of the training set for training and the rest for validation. All results reported are on the test set. As for the FLIR dataset, we only used the person category. Initially, we have the RGB detector trained on the datasets using 5 different seeds. It's worth noting that this model starts with pre-trained weights from COCO (Lin *et al.*, 2014). Then with the RGB model trained, we use the model to perform the Hallucidet training. We tried ResNet_{50} as the backbone for the detectors and ResNet_{34} as the backbone for the Hallucination network. To ensure fairness we trained the detectors under the same conditions, i.e., data order, augmentations, etc. All the code is available at GitHub for the reproducibility of the experiments. To develop the code, we used Torchvision models for the detectors and PyTorch Segmentation Models (Iakubovskii, 2019) for the U-Net architecture of the hallucination network. Additionally, we trained with PyTorch

Lightning (Falcon & The PyTorch Lightning team, 2019) training framework, evaluated the AP with TorchMetrics (Detlefsen *et al.*, 2022), and logged all experiments with WandB (Biewald, 2020) logging tool.

3.4.2 Main Comparative Results.

In Table 3.1, we investigate how our model behaved in comparison with standard image-to-image approaches and classical computer vision approaches that are normally used to reduce the distribution gap between IR and RGB. Furthermore, we highlight the impact of using the proposed $\mathcal{L}_{\text{hall}}$ loss to guide the representation. This is accomplished by comparing our approach with a U-Net that shares the same backbone as ours but employs a standard $\mathcal{L}_{\text{L1}}$ reconstruction loss. To guarantee comparability, we reproduce the experimental setting of Herrmann *et al.* (2018) on our pipeline. We included basic pre-processing techniques that were shown to enhance IR performance on RGB models by Hermann et al. (Herrmann *et al.*, 2018). These techniques include a combination of blurring, histogram equalization, stretching, and inverting pixels. Furthermore, we included CycleGAN, which is a more powerful generative model compared with UNet. It is important to mention that training the CycleGAN is computationally more demanding than the Hallucidet. Additionally, due to the adversarial nature of the method, it does not ensure reliable convergence for the subsequent detection task. The CycleGAN was diverging with the same hyperparameters as (Jia *et al.*, 2021) on the test set, so we tuned the hyperparameters and trained until the images became good qualitatively. Because CycleGAN introduces significant noise to the images as a result of its adversarial training, the detector's performance has notably decreased. This is particularly evident due to the increase in false positives. Given that our final goal is object detection, we selected FCOS, RetinaNet, and Faster R-CNN, each representing distinct categories within the universe of detection networks. We can see that straightforward approaches like inverting pixels for the IR and expanding it to three channels significantly enhance the initial performance of IR inputs on RGB detectors. As indicated in the table, our results demonstrate a significant improvement over previous image-to-image translation techniques in terms of detection performance. The most significant

enhancement was observed in Faster R-CNN, where our proposal exhibited a remarkable 17% improvement compared to pixel inversion.

Hallucidet Visual Output. In Figure 3.1, we present a Hallucination image and compare it with both RGB and IR. The Hallucination emphasizes the person while smoothing the background, helping the detector to distinguish the regions of interest. In contrast to RGB, our method allows for easy person detection even in low-light conditions. However, IR images may introduce additional non-person-related information that could bias the detector. A visual comparison with FastCUT is also provided, revealing a correlation between the method's low performance and the high number of False Positives detected. It is important to note that while we show the Hallucination for representation demonstration, our main goal is on detection metrics. In the figure, the ground truth bounding box annotations are shown in yellow on the RGB images. The corresponding detections obtained from the IR data are presented in the following lines. It is important to note that we display the predicted detections on top of the intermediate representation for convenience. However, the actual inputs for HalluciDet approaches and FastCUT are IR images. A significant number of False Positives can be observed for FastCUT, while HalluciDet (FCOS) and HalluciDet (RetinaNet) exhibit a high number of False Negatives. The most accurate detection results are achieved with HalluciDet (Faster R-CNN), which demonstrates superior performance to the IR fine-tuned model in cases where the person's heat signature is not clearly evident, as seen in the last column. Additional figures can be found in the supplementary material.

Comparison with fine-tuning. For this experiment, we performed an evaluation of both RGB and fine-tuned IR detectors that were trained on the LLVIP and FLIR datasets. All methods from Table 3.2 were trained under the same experimental protocol using 3 different seeds.

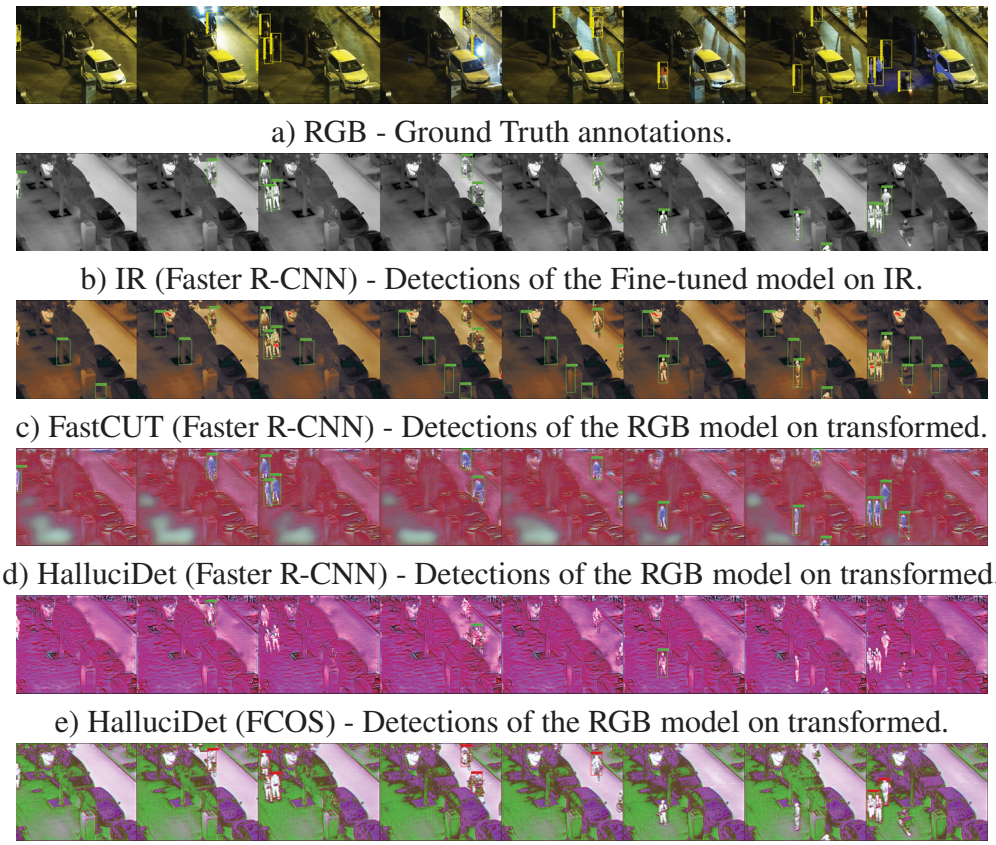
Similar to the previous experiment, we utilized a detector from each family of methods, namely FCOS, RetinaNet, and Faster R-CNN. The provided results include the mean and standard deviation of the AP on the test set. In this experiment, we compare three different approaches to adapt a model trained on RGB images to IR. As baseline we consider the case of No Adaptation,

in which the model is used directly on IR images. Then, we consider the case in which a model is adapted to the IR data with normal fine-tuning, which is the most common way of adaptation when annotations are available. Finally, we train our HalluciDet to generate a new representation of the image for the RGB detector.

As seen in Table 3.2, in all cases, the fine-tuned IR model outperformed the RGB detector over the IR modality, as expected. In the tables, we also observe a significant improvement in the performance of HalluciDet compared to the performance achieved through fine-tuning for Faster R-CNN. This improvement aligns with the quality of the representation observed in Figure 3.3, where confusing factors, such as car heat, have been removed from the image. A marginal improvement was observed with center point-based architectures like FCOS for the LLVIP dataset, although a higher difference in AP could be observed for the FLIR dataset. On the other hand, the results using RetinaNet didn't exhibit much consistency; the AP was significantly worse than that achieved through fine-tuning for the LLVIP dataset. Once again, this is consistent with the observed representation lacking the necessary discriminative information to detect people in the image.

Hallucidet with different backbones. In Table 3.3, we investigated various encoder backbones for the Hallucination network. The presented results include two MobileNet and two ResNets with different widths. Additional outcomes for alternative backbones are included in the supplementary material. In all cases, the model consistently improves upon the performance of the fine-tuned IR model. Notably, even in models with a reduced number of parameters, such as MobileNet$_{v2}$ with less than 7 million additional parameters, the gain remains consistent at nearly 5%.

Hallucidet with a different number of training samples. For the LLVIP dataset, in Figure 3.4, we explored various quantities of training samples for our method, ranging from 1% to 100%. Notably, only 30% of the data was sufficient for HalluciDet to achieve comparable performance to the fine-tuned Faster R-CNN with the complete dataset. For the FLIR dataset, in Figure 3.5, the trend to reduce the number of training samples and improve over the fine-tuning is still true,



a) RGB - Ground Truth annotations.

b) IR (Faster R-CNN) - Detections of the Fine-tuned model on IR.

c) FastCUT (Faster R-CNN) - Detections of the RGB model on transformed.

d) HalluciDet (Faster R-CNN) - Detections of the RGB model on transformed.

e) HalluciDet (FCOS) - Detections of the RGB model on transformed.

f) HalluciDet (RetinaNet) - Detections of the RGB model on transformed.

Figure 3.3 Illustration of a sequence of 8 images of LLVIP dataset. The first row is the RGB modality, then the IR modality, followed by FastCUT and different representations created by HalluciDet over various detectors

but in this case, around 70% of the training samples. The different characteristics related to the exact number of training samples with respect to the dataset are due to the number of different environment changes on the datasets. For the LLVIP, we do not have a big shift in the images because the cameras are fixed in a surveillance context. In the case of FLIR, the variance of the images is higher due to the different capture settings; with the focus on autonomous driving, the camera moves inside a car, which changes the background consistency and introduces more variance to the dataset.

Table 3.2 AP performance for various models following distinct training approaches on two datasets of LLVIP (Jia *et al.*, 2021) (top half) and FLIR (Group *et al.*, 2018) (bottom half). Starting from COCO pre-training and fine-tuning on the RGB data shown as (No Adaptation) and fine-tuning on the IR data shown as (Fine-tuning). In the case of HalluciDet, the trained RGB detector serves as the initial point, with the subsequent optimization of the Hallucination network using the IR data. The reported performance is exclusive to the person category

Method	**AP@50↑**		
	Test Set IR (Dataset: LLVIP)		
	No Adaptation	Fine-tuning	HalluciDet
FCOS	47.12 ± 4.32	63.79 ± 0.48	**64.85 ± 1.46**
RetinaNet	50.63 ± 3.22	**76.26 ± 0.75**	56.78 ± 3.85
Faster R-CNN	71.51 ± 1.16	84.94 ± 0.15	**90.92 ± 0.20**
	Test Set IR (Dataset: FLIR)		
	No Adaptation	Fine-tuning	HalluciDet
FCOS	38.52 ± 0.79	42.22 ± 1.04	**49.18 ± 0.99**
RetinaNet	44.13 ± 2.01	47.87 ± 2.21	**49.01 ± 4.08**
Faster R-CNN	55.85 ± 1.19	61.48 ± 1.55	**70.90 ± 1.35**

3.5 Conclusion

In this work, we provided a framework that uses privileged information of an RGB detector to perform the image-to-image translation from IR. The approach involves utilizing a Hallucination network to generate intermediate representations from IR data, which are then directly input into an RGB detector. An appropriate loss function was also proposed to lead the representation into a space that allows for the enhancement of the target category's importance.

In our experiments, we demonstrate that hallucination networks can be helpful for modality adaptation by obtaining an intermediate representation that effectively supports accurate responses

Table 3.3 Comparison of the number of parameters for different Hallucination Network backbones vs. AP@50 on the LLVIP dataset with the Faster R-CNN detector

Method		Params.	AP@50↑
Faster R-CNN		41.3 M	84.83
HalluciDet	MobileNet_{v3s}	+ 3.1 M	85.20
	MobileNet_{v2}	+ 6.6 M	89.73
	ResNet_{18}	+ 14.3 M	90.42
	ResNet_{34}	+ 24.4 M	90.65

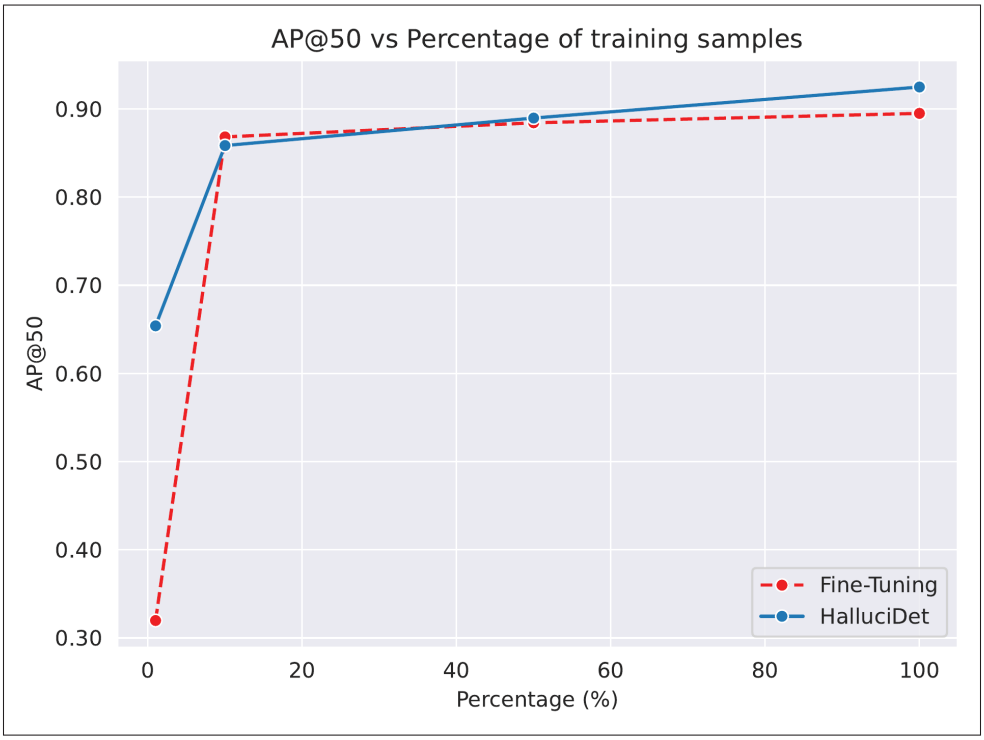


Figure 3.4 AP@50 vs. training samples percentages. The figure shows the AP@50 over the LLVIP test set using various amounts of training samples for the HalluciDet Faster R-CNN

in the object detection task. The proposed approach showed particular effectiveness for the two-stage detector Faster R-CNN, resulting in a reduction of non-person-related information.

Figure 3.5 AP@50 vs. training samples percentages. The figure shows the AP@50 over the FLIR test set using various amounts of training samples for the HalluciDet Faster R-CNN. Notably, 70% of the data was sufficient for HalluciDet to achieve comparable performance to the fine-tuned Faster R-CNN with the complete dataset

This reduction in background clutter had a positive effect on minimizing the number of False Positives, surpassing the performance of standard fine-tuning on IR data. The comparison with methods from the literature for image-to-image translation highlighted the significance of guiding the representation to achieve successful detections. Our Hallucidet demonstrated a significant performance improvement compared to the other methods. Finally, the proposed framework offers the additional advantage of maintaining performance in the RGB task, which is beneficial for applications requiring accurate responses in both modalities.

CHAPTER 4

MODALITY TRANSLATION FOR OBJECT DETECTION ADAPTATION WITHOUT FORGETTING PRIOR KNOWLEDGE

Heitor R. Medeiros[a] , Masih Aminbeidokhti[a] , Fidel A. Guerrero Peña[a] , David Latortue[a], Eric Granger[a], Marco Pedersoli[a]

[a] Department of Systems Engineering, École de technologie supérieure,
1100 Notre-Dame West, Montreal, Quebec, Canada H3C 1K3

Paper published in *Proceedings of European Conference on Computer Vision. Cham: Springer Nature Switzerland, November 2024*

Abstract

A common practice in deep learning involves training large neural networks on massive datasets to achieve high accuracy across various domains and tasks. While this approach works well in many application areas, it often fails drastically when processing data from a new modality with a significant distribution shift from the data used to pre-train the model. This paper focuses on adapting a large object detection model trained on RGB images to new data extracted from IR images with a substantial modality shift. We propose Modality Translator (ModTr) as an alternative to the common approach of fine-tuning a large model to the new modality. ModTr adapts the IR input image with a small transformation network trained to directly minimize the detection loss. The original RGB model can then work on the translated inputs without any further changes or fine-tuning to its parameters. Experimental results on translating from IR to RGB images on two well-known datasets show that our simple approach provides detectors that perform comparably or better than standard fine-tuning, without forgetting the knowledge of the original model. This opens the door to a more flexible and efficient service-based detection pipeline, where a unique and unaltered server, such as an RGB detector, runs constantly while being queried by different modalities, such as IR with the corresponding translations model. Our code is available at: https://github.com/heitorrapela/ModTr.

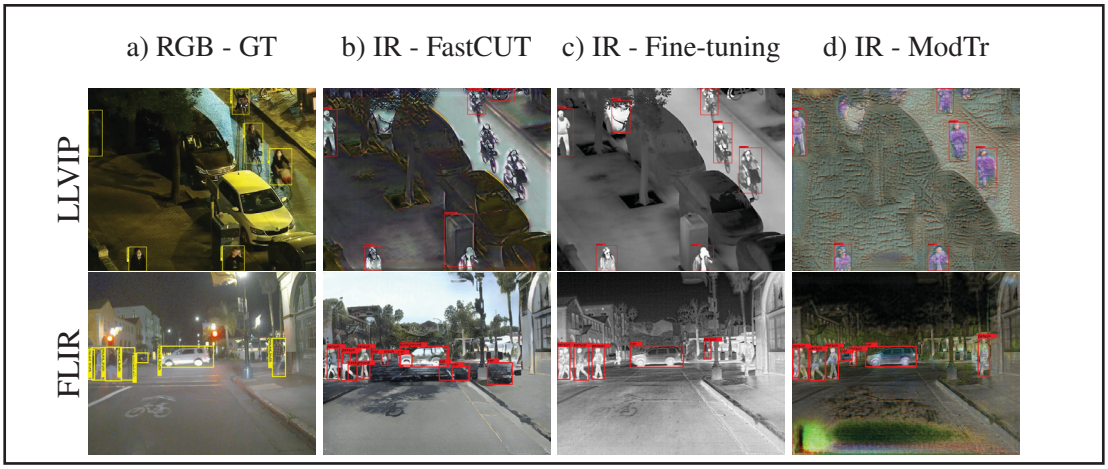


Figure 4.1 Bounding box predictions over different adaptations of the RGB detector (Faster R-CNN) for IR images on two benchmarks: LLVIP and FLIR. Yellow and red boxes show the ground truth and predicted detections, respectively. In a) we see the RGB data. In b) FastCUT is an unsupervised image translation approach that takes as input infrared images (IR) and produces pseudo-RGB images. It does not focus on detection and requires both modalities for training. In c) we have fine-tuning, which is the standard approach to adapting the detector to the new modality. It requires only IR data but forgets the original knowledge of the original RGB detector. Finally, in d) is the ModTr, which focuses the translation on detection, requires only IR data and does not forget the original knowledge so that it can be reused for other tasks. Bounding box predictions for other detectors are provided in the supplementary material

4.1 Introduction

Powerful pre-trained models have become essential in the field of computer vision, particularly in object detection (OD) tasks (Minderer *et al.*, 2022; Minderer, Gritsenko & Houlsby, 2023). These OD models are typically pre-trained on extensive natural-image RGB datasets, such as COCO (Lin *et al.*, 2014). Moreover, the knowledge encoded by these models can be leveraged for various tasks in a zero-shot way or with additional fine-tuning for downstream tasks (Vasconcelos, Birodkar & Dumoulin, 2022). However, adding new modalities to these models, such as infrared (IR), without losing the intrinsic knowledge of the detector remains a challenge (Medeiros *et al.*, 2024c).

These additional modalities, though not as common as RGB images, are still important in various tasks, like surveillance (Chen *et al.*, 2019a; Dubail *et al.*, 2022), autonomous driving (Stilgoe, 2018; Natan & Miura, 2022), and robotics (Pierson & Gashler, 2017; Jing, Potgieter, Noble & Wang, 2017), which strive to achieve robust performance in real-world environments, where capture conditions change, such as different illumination conditions (Bustos, Mashhadi, Lai-Yuen, Sarkar & Das, 2023). The dominant way to adapt pre-trained detectors to these novel conditions is by fine-tuning the model (Medeiros *et al.*, 2024c). However, fine-tuning often results in catastrophic forgetting and can destroy the intrinsic knowledge of the detector (Kirkpatrick *et al.*, 2017). Ideally, we would like to adapt the detector to new modalities without changing the original model. This is most useful for server-side applications, where a single model runs uninterrupted and can be queried by different inputs, ideally on different modalities. The main challenge lies in the significant distribution shift introduced by the new modality. This shift occurs because the pre-trained knowledge, such as the visual information in RGB images, differs markedly from the thermal data in IR images. This shift can degrade model performance when applied directly as input to the model, since the features learned from one modality may not be relevant or present in another. This can ultimately impact the resulting OD performance (Wang *et al.*, 2022).

Image translation methods (Pang, Lin, Qin & Chen, 2021; Park *et al.*, 2020a) have emerged as powerful tools to overcome the downsides of fine-tuning and narrowing the gap between source and target modalities (Hsu *et al.*, 2020). These methods do not directly work on the weight space of the original detector but rather adapt the input values to reduce the discrepancy between the source and target modalities. However, such methods often require access to source data or some statistics about it during training. Furthermore, their primary focus is on image reconstruction quality rather than the final OD task, which can cause a significant drop in performance. For instance, Figure 4.1 shows different ways to adapt the RGB detector (see the caption for more details).

Our work aims to improve the image translation paradigm while addressing its limitations. Our proposed approach, Modality Translation for OD (ModTr), incorporates the detector's knowledge

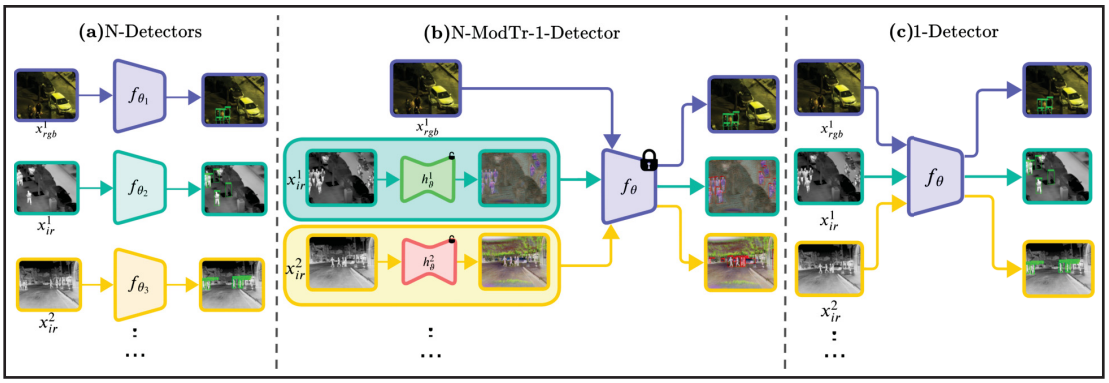


Figure 4.2 Different approaches to deal with multiple modalities and/or domains. (a) The simplest approach is to use a different detector adapted to each modality. This can lead to a high level of accuracy but requires storing several models in memory. (b) Our proposed solution uses a single pre-trained model normally trained on the more abundant data (RGB) and then adapts the input through our ModTr model. (c) A single detector is jointly trained on all modalities. This allows using of a single model but requires access to all modalities jointly, which is often impossible, especially when dealing with large pre-trained models

into the translation module by training directly for the final detection task. Unlike traditional image translation methods, ModTr does not require any source data. It is a conceptually simple approach that can be easily integrated with any detector, be it a one-stage or two-stage detector. A notable application of ModTr is using a pre-trained RGB detector as a server that incorporates different ModTr blocks as input translators for new modalities such as IR. This new detector generates the desired output with performance comparable to full fine-tuning without losing the original knowledge of the pre-trained model. In Fig. 4.2, we present several options for integrating IR modalities into an RGB system. Fig. 4.2a illustrates the N-Detectors approach, where each detector is trained for a specific case. This method effectively demands more memory and forgets previously learned information. Fig. 4.2c shows a single detector trained on combined modalities. This method does not incur additional memory, yet it requires simultaneous access to all modalities, which may not always be feasible. Fig. 4.2b illustrates our proposed approach, which involves training a specialized translator for each condition without altering the parameters of the original detector. The N-ModTr-1-Detector strikes a balance between the previous methods, addressing their shortcomings by requiring only a single detector.

Importantly, it retains the original pre-training knowledge, as it leaves the detector unchanged. In this work, we focus on the effectiveness of our approach for the IR modality, commonly used in surveillance and robotics, and the incremental modality detector server-based application, which is crucial for many settings that require uninterrupted detection predictions.

Our main contributions can be summarized as follows.

(1) We present ModTr, a method for adapting pre-trained ODs from large RGB datasets to new scarce modalities like IR, without requiring access to any source dataset, by translating the input signal.

(2) In contrast to standard fine-tuning, our approach does not modify the original detector weights. This allows the detector to retain the knowledge of the source data while adapting to a new modality. As a result, a single model can be used to handle multiple modalities across various translators. For instance, the same model can be used to process RGB during the daytime and IR at nighttime.

(3) An extensive empirical evaluation of ModTr in several scenarios, showcasing its advantages and flexibility. In particular, with our different proposed fusion strategies, ModTr achieves OD accuracy that is competitive when compared with image translation methods on two challenging RGB/IR datasets (LLVIP and FLIR).

4.2 Related Work

(a) Object Detection. OD is a computer vision task that aims to provide labels and localization for the objects in the image (Zhang *et al.*, 2023a). Two-stage detectors, exemplified by Faster R-CNN (Ren *et al.*, 2015), generate regions of interest and then use a second classifier to confirm object presence within those regions. On the other hand, one-stage detectors streamline the detection process by eliminating the proposal generation stage, aiming for end-to-end training and real-time inference speeds. RetinaNet (Lin *et al.*, 2017b) is a one-stage OD model that utilizes a focal loss function to address class imbalance during training. Also, models like FCOS (Tian, Shen, Chen & He, 2019) have emerged in this category, eliminating predefined anchor boxes to potentially enhance inference efficiency. The proposed work investigates these

three traditional and powerful detectors: Faster R-CNN, RetinaNet, and FCOS. The choice of such detectors was due to the simplicity in implementation and integration among other methods, as well as a different range of pre-trained backbone weights, such as ResNet (He *et al.*, 2016) and MobileNet (Howard *et al.*, 2017).

(b) Image Translation. Image translation is a pivotal task in computer vision, aiming to map images from a source domain to a target domain while preserving inherent content (Pang *et al.*, 2021). The goal is to discover a transformation function such that the distribution of images in the translated domain is aligned with the distribution of images in the target domain. The commonly used approaches for image translation are based on variational autoencoders (VAEs) (Kingma & Welling, 2022) and generative adversarial network (GANs) (Goodfellow *et al.*, 2014; Pang *et al.*, 2021). Isola et al. developed the Pix2Pix (Isola, Zhu, Zhou & Efros, 2017a), a method that consists of a generator (based on U-Net) and a discriminator (based on GANs architecture) that work together to generate images based on input data and labels. Then, Zhu et al. proposed a method called CycleGAN (Zhu *et al.*, 2017a), which is based on GANs, with the objective of unsupervised domain translation. Even though CycleGAN can produce quite visual results, it's hard to optimize due to the adversarial mechanism and memory footprint needed. In contrast, VAEs are easier to train than GANs but require more constraints in the optimization to produce images of good quality than GAN-based approaches. Recent advancements include diffusion models known for their high-quality image generation, although they may not inherently suit domain translation tasks. To enhance models such as CycleGAN, novel methods like Contrastive Unpaired Translation (CUT) (Park *et al.*, 2020a) and FastCUT (Park *et al.*, 2020a) have been introduced. CUT, in particular, accelerates the image translation process by maximizing mutual information between image patches, achieving competitive results quickly. In the context of RGB/IR modality, InfraGAN presents an image-level adaptation for RGB to IR conversion, prioritizing image quality (Özkanoğlu & Ozer, 2022). This approach is distinct in its focus on optimizing image quality losses. Moreover, Herrmann et al. have explored OD in RGB/IR modality by adapting IR images to RGB using traditional image preprocessing techniques, allowing the use of RGB object detectors

without parameter modification (Herrmann *et al.*, 2018). Despite significant advances in image translation, these techniques do not specifically address OD tasks. In our previous work, we introduced HalluciDet (Medeiros *et al.*, 2024c), which employs an image translation mechanism for OD. However, this approach requires prior access to the source RGB data from the same domain as the target for pre-training the detector.

(c) Adapting Without Forgetting. Catastrophic forgetting (CF) is the idea that a neural network tends to forget knowledge when sequentially trained on a different task and replaces it with knowledge tailored to the new objective (Wang, Yang, Shen & Huang, 2023b). CF can be harmful or beneficial. Researchers identified harmful learning as situations where retaining the original knowledge while adapting to a different task is necessary. In that case, it is imperative to mitigate the risk of CF. However, some CF can also be beneficial, for instance, to prevent privacy leakage from large pre-trained models, to enhance the generalization, or to remove noisy information from the originally, acquired knowledge that is negatively affecting the new tasks. In our case, knowledge-forgetting is harmful. There are different ways to address this issue including simple techniques like decreasing the learning rate (Howard & Ruder, 2018), use weight decay (Chelba & Acero, 2006; Zhang, Wu, Katiyar, Weinberger & Artzi, 2021) or mixout regularization (Lee, Cho & Kang, 2020) during fine-tuning or more complex approaches like Recall and learn (Chen *et al.*, 2020b), Robust Information Fine-tuning (Wortsman *et al.*, 2022b) or CoSDA (Feng *et al.*, 2023). Some adaptation methods use techniques based on replay of the source data or even using the weights of the initial model to keep some prior information (Menezes, de Moura, Alves & de Carvalho, 2023). Some of these works focus on adding continually different tasks in an incremental learning setting. However, these methods may still produce a loss of knowledge since the original parameters are not frozen. Furthermore, in adapting without forgetting, an adapter, which adopts a frozen pre-trained backbone to generate a representation followed by a different classifier for each downstream task (Wang *et al.*, 2023b), can be seen as a powerful method to preserve knowledge. Even though our ModTr shares some similarities, we work in the input space to adapt to the new modalities, and address this incremental modality adaptation, optimizing the translation directly for the final OD task.

4.3 Proposed Method

(a) Preliminary Definitions. The training set for OD is denoted as $\mathcal{D} = \{(x, Y)\}$, where $x \in \mathbb{R}^{W \times H \times C}$ represents an image in the dataset, with dimensions $W \times H$ and C channels. Subsequently, the OD model aims to identify N regions of interest within these images, denoted as $Y = \{(b_i, c_i)\}_{i=1}^{N}$. The top-left corner coordinates and the width and height of the object define each region of interest b_i. Additionally, a classification label c_i is assigned to each detected object, indicating its corresponding class within the dataset. In this study, the number of input channels for the detector is fixed at three, corresponding to RGB-like inputs. In terms of optimization, the primary goal of this task is to maximize detection accuracy, often measured using the average precision (AP) metric across all classes. An OD is formally represented as the mapping $f_\theta : \mathbb{R}^{W \times H \times C} \rightarrow \hat{Y}$, where θ denotes the parameter vector. To effectively train a detector, a differentiable surrogate for the AP metric, referred to as the detection cost function, $C_{det}(\theta)$, is employed. The typical structure of such a cost function involves computing the average detection loss over dataset $\mathcal{D}$, denoted as $\mathcal{L}_{det}$, described as:

$$C_{det}(\theta) = \frac{1}{|\mathcal{D}|} \sum_{(x,Y) \in \mathcal{D}} \mathcal{L}_{det}(f_\theta(x), Y). \tag{4.1}$$

(b) Modality Translation Module. Our approach primarily consists of an image-to-image translation network responsible for converting the input modality into an RGB-like space intelligible to the detector. These networks typically adopt an encoder-decoder structure to synthesize and reconstruct knowledge in a pixel-wise manner. While we employ U-Net (Ronneberger *et al.*, 2015) as the translation network, with parameters ϑ, in this work, our framework is general and not limited by the translation architecture. In general terms, this mapping is denoted as $h_\vartheta^d : \mathbb{R}^{W \times H \times C} \rightarrow \mathbb{R}^{W \times H \times 3}$, with a translation network assigned to each available input modality d. Unlike the detection network, the number of input channels varies depending on the modality, for instance, $C = 1$ for IR and depth images. It's important to note

that, being a pixel-level architecture, the output of such a network retains the spatial resolution of the input. However, the number of output channels is consistently fixed at three, corresponding to RGB-like images ($C = 3$).

Unlike other image-to-image translation approaches, we drive the process using the aforementioned detection cost (Equation (4.1)). Thus, the underlying optimization problem is formulated as $\vartheta^* = \arg\min \mathcal{L}_{det}(\vartheta)$, incorporating the output of the composition $(f_\theta \circ h^d_\vartheta)(x)$ at the loss function level. To streamline the learning process, we utilize a residual learning strategy in which the function h^d_ϑ focuses on capturing the small variations in the input that are necessary to solve the task. This approach is similar to the one employed on diffusion models, which inspired our work. For the sake of simplicity, we separate the fusion step from the translation mapping in our notation, as various types of fusion are investigated. Consequently, the proposed image-to-image translation loss function is defined as:

$$\mathcal{L}_{\text{ModTr}}(x, Y; \vartheta) = \mathcal{L}_{det}(f_\theta \left(\Phi(h^d_\vartheta(x), x)\right), Y), \tag{4.2}$$

where $\Phi(.,.)$ is a non-parametric fusion function. Note that the output of $h^d_\vartheta(x)$ is an RGB-like image, whereas x may only consist of a single channel, depending on the input modality. We have chosen this definition to simplify the notation, but appropriate reshaping should be performed during implementation to ensure compatibility.

In addition, note that, while a detection loss is employed to update the translation network, the weight vector θ remains constant. This constraint is consistent with the premise of this study, where a pre-trained detector is solely available on the server side and remains unaltered. An overview of the proposed approach can be seen in Fig. 4.2 b).

(c) Fusion strategy. As previously mentioned, we utilize a non-parametric fusion of the intermediate representation $h^d_\vartheta(x)$ and the original input x to simplify the learning process of the translation network. In this context, we employ an element-wise product, also known as

the Hadamard product, which is particularly interesting for attention mechanisms and has been explored previously for re-calibrating feature maps based on their importance (Hu, Shen & Sun, 2018). Although we investigated various fusion mechanisms, the element-wise product yielded the best results. For more details on different fusion strategies, please refer to supplementary materials.

ModTr$_{\odot}$: The Hadamard product-based fusion serves as a gating mechanism to filter or highlight information from the input image. In this approach, the output of the translation network acts as a weight map for the input, and they are fused using pixel-wise multiplication, $\odot$. Consequently, the translation network tends to highlight information from the input when the pixel value tends toward 1 or discard it when it approaches 0. Additionally, the output translation modality can be interpreted as an attention map, as described by the following Equation (4.3):

$$\mathcal{L}_{\text{ModTr}_{\odot}}(x, Y; \vartheta) = \mathcal{L}_{det}(f_{\theta}\left(h^{d}_{\vartheta}(x) \odot x\right), Y). \tag{4.3}$$

In our design choices, we opt to utilize these straightforward non-parametric functions to assist in optimization while maintaining low inference costs.

4.4 Results and Discussion

4.4.1 Experimental Methodology

Datasets LLVIP: LLVIP is a surveillance dataset composed of 30, 976 images, in which 24, 050 (12, 025 IR and 12, 025 RGB paired images) are used for training and 6, 926 for testing (3, 463 IR and 3, 463 RGB paired images) with only pedestrians annotated. **FLIR ALIGNED:** We used the sanitized and aligned paired sets provided by Zhang et al. (Zhang *et al.*, 2020b). It has 10, 284 images, that is 8, 258 for training (4, 129 IRs and 4, 129 RGBs) and 2, 026 (1, 013 IRs and 1, 013 RGBs) for test. FLIR images are captured from the perspective of a camera in the front of a car, with a resolution of 640 by 512. It contains the bicycles, dogs, cars, and people

classes. It has been found that with FLIR, the "dog" objects are inadequate for training (Cao *et al.*, 2023b), thus we decided to remove them.

Implementation details In our experiments, we randomly selected 80% of the training set for training and the rest for validation. All results reported are on the test set. As starting pre-trained weights for the detectors, we used Torchvision models with COCO (Lin *et al.*, 2014) weights and for the U-Net translation network, we used PyTorch Segmentation Models (Iakubovskii, 2019) and we changed the last layer for 3-channel (RGB-like) with a Sigmoid function, to be closer to an image with values between 0 and 1, to perform translation instead of traditional segmentation. For the translation network backbones, we explored our default ResNet_{34}, and for subsequent studies on reducing parameters, we dive into ResNet and MobileNet-family. All the code is available on GitHub for reproducibility in the experiments. To ensure fairness, we trained the detectors under the library version and the same experimental design, i.e., data order, augmentations, etc. Furthermore, we trained with PyTorch Lightning (Falcon & The PyTorch Lightning team, 2019) training framework, evaluated the APs with TorchMetrics (Detlefsen *et al.*, 2022), and logged all experiments with WandB (Biewald, 2020) logging tool. The different performance measures (e.g., APs) can be found in suppl. materials.

4.4.2 Comparison with Translation Approaches

In this section, ModTr is compared with different image-to-image translation methods employing different learning strategies. These include basic image processing strategy (Herrmann *et al.*, 2018), reconstruction strategies such as CycleGAN (Zhu *et al.*, 2017b), CUT (Park *et al.*, 2020b), and FastCUT (Park *et al.*, 2020b), which employs a contrastive learning approach, as well as HalluciDet (Medeiros *et al.*, 2024c), which utilizes a detection-based loss. As outlined in Table 4.1, we evaluated the methods based on their final detection performance across three commonly used detectors: FCOS, RetinaNet, and Faster R-CNN. The reported results are derived from the IR test set and are averaged over three different seeds, which helps mitigate the impact of randomness across runs and splits of the training and validation datasets.

For each method, we also consider its dependency on the prior knowledge data (RGB) and ground truth bounding boxes (bboxes) on the IR images. Methods that rely on reconstruction techniques do not require bbox annotations on IR images but cannot provide accurate translations for detection purposes. However, HalluciDet and ModTr require bbox annotations to adjust the input image in a discriminative manner. The main difference between HalluciDet and ModTr is the use of source images. HalluciDet requires RGB images for an initial fine-tuning of the model, while our approach can work without that fine-tuning by reusing the detector's zero-shot knowledge.

The proposed ModTr displays robustness across the three detectors and consistently exhibits improvement on two different datasets: LLVIP (Jia *et al.*, 2021) and FLIR aligned (Group *et al.*, 2018). Note that each algorithm described in Table 4.1 employs different training supervisions. For instance, CycleGAN employs an adversarial mechanism with both RGB and infrared modalities in an unpaired setting. Similarly, CUT and FastCUT operate with positive and negative patches in an unpaired setting. In contrast, HalluciDet doesn't require the presence of both modalities during training but employs a detection mechanism during training similar to ours. In our approach, we solely require examples from the target modality. In this section, we present the performance of our best approach $\text{ModTr}_{\odot}$. For additional results, refer to suppl. materials.

As reported in Table 4.1, the detection performance of ModTr over the LLVIP dataset exhibited significant improvements. Specifically, it surpassed HalluciDet, the second best, by more than 29.0 AP with both FCOS and RetinaNet architectures, while obtaining comparable results with Faster R-CNN. Such disparity with the previous technique can be attributed to the loss of previous knowledge inherent in HalluciDet, which necessitates a pre-fine-tuning strategy on the source modality. Although the performance of the FLIR dataset also improved, the dataset's inherent challenges, such as changing the background from a moving car setup, make detection more difficult. Nonetheless, our proposal consistently enhances results, with improvements of more than 11 AP for FCOS and RetinaNet, and over 7 AP for Faster R-CNN. We also observed improvements on the AP_{50} and AP_{75}. Because of the space constraint, we include these in

Table 4.1 Detection performance (AP) of ModTr versus baseline image-to-image methods to translate the IR to RGB-like images, using three different detectors (FCOS, RetinaNet, and Faster R-CNN). The methods were evaluated on IR test set of LLVIP and FLIR datasets. The RGB column indicates if the method required access to RGB images during training, and Box refers to the use of ground truth boxes during training

Image translation	**RGB**	**Box**	**Test Set IR (Dataset: LLVIP)**		
			FCOS	**RetinaNet**	**Faster R-CNN**
Histogram Equal. (Herrmann *et al.*, 2018)			31.69 ± 0.00	33.16 ± 0.00	38.33 ± 0.02
CycleGAN (Zhu *et al.*, 2017b)	✓		23.85 ± 0.76	23.34 ± 0.53	26.54 ± 1.20
CUT (Park *et al.*, 2020b)	✓		14.30 ± 2.25	13.12 ± 2.07	14.78 ± 1.82
FastCUT (Park *et al.*, 2020b)	✓		19.39 ± 1.52	18.11 ± 0.79	22.91 ± 1.68
HalluciDet (Medeiros *et al.*, 2024c)	✓	✓	28.00 ± 0.92	19.95 ± 2.01	57.78 ± 0.97
ModTr$_{\odot}$ (ours)		✓	**57.63 ± 0.66**	**54.83 ± 0.61**	**57.97 ± 0.85**
Image translation	**RGB**	**Box**	**Test Set IR (Dataset: FLIR)**		
			FCOS	**RetinaNet**	**Faster R-CNN**
Histogram Equal. (Herrmann *et al.*, 2018)			22.76 ± 0.00	23.06 ± 0.00	24.61 ± 0.01
CycleGAN (Zhu *et al.*, 2017b)	✓		23.92 ± 0.97	23.71 ± 0.70	26.85 ± 1.23
CUT (Park *et al.*, 2020b)	✓		18.16 ± 0.75	17.84 ± 0.75	20.29 ± 0.48
FastCUT (Park *et al.*, 2020b)	✓		24.02 ± 2.37	22.00 ± 2.73	26.68 ± 2.59
HalluciDet (Medeiros *et al.*, 2024c)	✓	✓	23.74 ± 2.09	22.29 ± 0.45	29.91 ± 1.18
ModTr$_{\odot}$ (ours)		✓	**35.49 ± 0.94**	**34.27 ± 0.27**	**37.21 ± 0.46**

supplementary materials. These promising results indicate that our proposal can effectively translate images from the original IR modality to an RGB-like representation, sufficiently close to the source data to be usable by the detector.

4.4.3 Translation vs. Fine-tuning

In this section, we further show that the proposed approach can be trained jointly with both translation and detector, which preserves the detector's knowledge. Here, ModTr is compared to three baselines fully fine-tuning (FT), FT of the head and LoRA (Hu *et al.*, 2022), and our best ModTr fusion strategy, as shown in Tab. 4.2.

We conduct LoRA fine-tuning using two settings. In the first, we apply LoRA across all layers; in the second, only to the last layer of detectors. The latter results in superior performance, so we have adopted it as our default LoRA setting. The Tab. 4.2 shows AP for the LLVIP

and FLIR datasets, with a consistent trend across all detectors (FCOS, RetinaNet, and Faster R-CNN). Furthermore, in the case of the FLIR dataset, we observed enhancements of ModTr over the standard detector FT. As demonstrated, our approach surpasses standard fine-tuning while maintaining the detector's performance in the original modality. It is worth noting that our method also improves performance in terms of localization metrics such as AP_{50} and AP_{75} compared to fine-tuning alone, and we provide detailed results in the supplementary materials.

Table 4.2 Detection performance (AP) of ModTr versus baseline fine-tuning (FT) of the detector, FT of the head and LoRA (Hu *et al.*, 2022), using three different detectors (FCOS, RetinaNet, and Faster R-CNN). The methods were evaluated on IR test set of LLVIP and FLIR datasets. Results with "-" diverged from the optimization

Method	**Test Set IR (Dataset: LLVIP)**		
	FCOS	**RetinaNet**	**Faster R-CNN**
Fine-Tuning (FT)	57.37 ± 2.19	53.79 ± 1.79	**59.62 ± 1.23**
FT Head	49.11 ± 0.70	44.00 ± 0.28	59.33 ± 2.17
LoRA (Hu *et al.*, 2022)	47.72 ± 0.58	-	54.83 ± 1.30
$ModTr_{\odot}$ (ours)	**57.63 ± 0.66**	**54.83 ± 0.61**	57.97 ± 0.85
Method	**Test Set IR (Dataset: FLIR)**		
	FCOS	**RetinaNet**	**Faster R-CNN**
Fine-Tuning (FT)	27.97 ± 0.59	28.46 ± 0.50	30.93 ± 0.46
FT Head	27.40 ± 0.12	26.78 ± 0.70	33.53 ± 0.36
LoRA (Hu *et al.*, 2022)	-	-	29.44 ± 0.61
$ModTr_{\odot}$ (ours)	**35.49 ± 0.94**	**34.27 ± 0.27**	**37.21 ± 0.46**

4.4.4 Different Backbones for ModTr

In this context, we evaluate ModTr and examine the trade-off between performance and parameter cost. It is widely recognized that increasing the number of parameters can enhance performance, but this relationship is not strictly linear. We demonstrated that models with fewer parameters can still achieve good performance; for example, $MobileNet_{v2}$, with fewer parameters than

ResNet_{18}, sometimes outperformed it. This trade-off highlights the versatility of the model, which can be deployed with MobileNet-based architectures and utilized in low-cost devices. In Table 4.3, the default number of parameters is successfully reduced from 24.4M (ResNet_{34}) to 6.6M using MobileNet_{v2} while maintaining similar performance. For instance, on LLVIP, MobileNet_{v2} achieved a mean AP of 56.15, comparable to 56.35 AP_{50} from ResNet_{34} (others APs and detectors are reported in the supplementary material).

This approach opens up new possibilities, particularly in scenarios where using one translation network and one detector (e.g., one ModTr and one detector for RGB/IR) proves advantageous. This setup requires a total of 44.9M parameters, compared to 83.6M parameters, when employing two detectors—one for each modality (for example, for Faster R-CNN). Similar reductions in parameter costs were observed for FCOS (from 66.4M to 36.3M) and RetinaNet (from 68M to 37.1M) when using one detector for both modalities while preserving the knowledge of the previous modality and incorporating a new one. These numbers are based on MobileNet_{v3s}, which strikes a balance between performance and the number of parameters, making it suitable for memory-restricted systems. The complete evaluations for FCOS and RetinaNet are included in the supplementary material.

4.4.5 Knowledge Preservation through Input Modality Translation

ModTr is designed to prevent catastrophic forgetting by keeping the weights of the pre-trained detector fixed. In this section, we demonstrate how various adaptation paradigms, shown in Figure 4.2, effectively solve the final task while preserving intrinsic knowledge. We compare our proposed method, ModTr, with two fine-tuning baseline methods. The first baseline method involves N-detectors, each fine-tuning the target modality individually. The second baseline method employs a single detector trained on the joint modality using balanced sampling. Note that while a copy of the original detector can be used in the N-detectors paradigm, it is unavailable in the 1-detector paradigm because the original modality is assumed to be inaccessible during training.

Table 4.3 Detection performance (AP) of ModTr with different backbones for the translation networks with different numbers of parameters, using three different detectors (FCOS, RetinaNet, and Faster R-CNN). The methods were evaluated on IR test set of LLVIP and FLIR datasets

Test Set IR (Dataset: LLVIP)		
Method	**Parameters**	**AP ↑**
Faster R-CNN	41.8 M	
MobileNet_{v3s}	+ 3.1 M	54.51 ± 0.28
MobileNet_{v2}	+ 6.6 M	56.15 ± 0.51
ResNet_{18}	+ 14.3 M	55.53 ± 1.14
ResNet_{34}	+ 24.4 M	56.35 ± 0.65
Test Set IR (Dataset: FLIR)		
Faster R-CNN	41.8 M	
MobileNet_{v3s}	+ 3.1 M	32.06 ± 0.75
MobileNet_{v2}	+ 6.6 M	36.77 ± 0.67
ResNet_{18}	+ 14.3 M	36.68 ± 0.22
ResNet_{34}	+ 24.4 M	37.21 ± 0.46

In all scenarios, we use COCO as the pre-training dataset and LLVIP and FLIR as target domains. Specifically, in the N-detectors scenario (Fig.4.2a), we fine-tune one detector on each dataset and use a copy of the original detector for the RGB modality. In the 1-detector scenario (Fig.4.2c), we fine-tune one detector on the combined FLIR and LLVIP datasets. In the N-ModTr-1-Detector scenario (Fig.4.2b), two translators are trained, one per dataset. To assess catastrophic forgetting, we re-evaluate each scenario on COCO-val.

Table 4.4 shows the final performance. While all adaptation paradigms achieve relatively similar performance, the 1-detector method completely fails in the zero-shot scenario. The N-detectors method mitigates this by duplicating the detector three times. In contrast, ModTr preserves knowledge using a single detector and three efficient translators, demonstrating its practicality

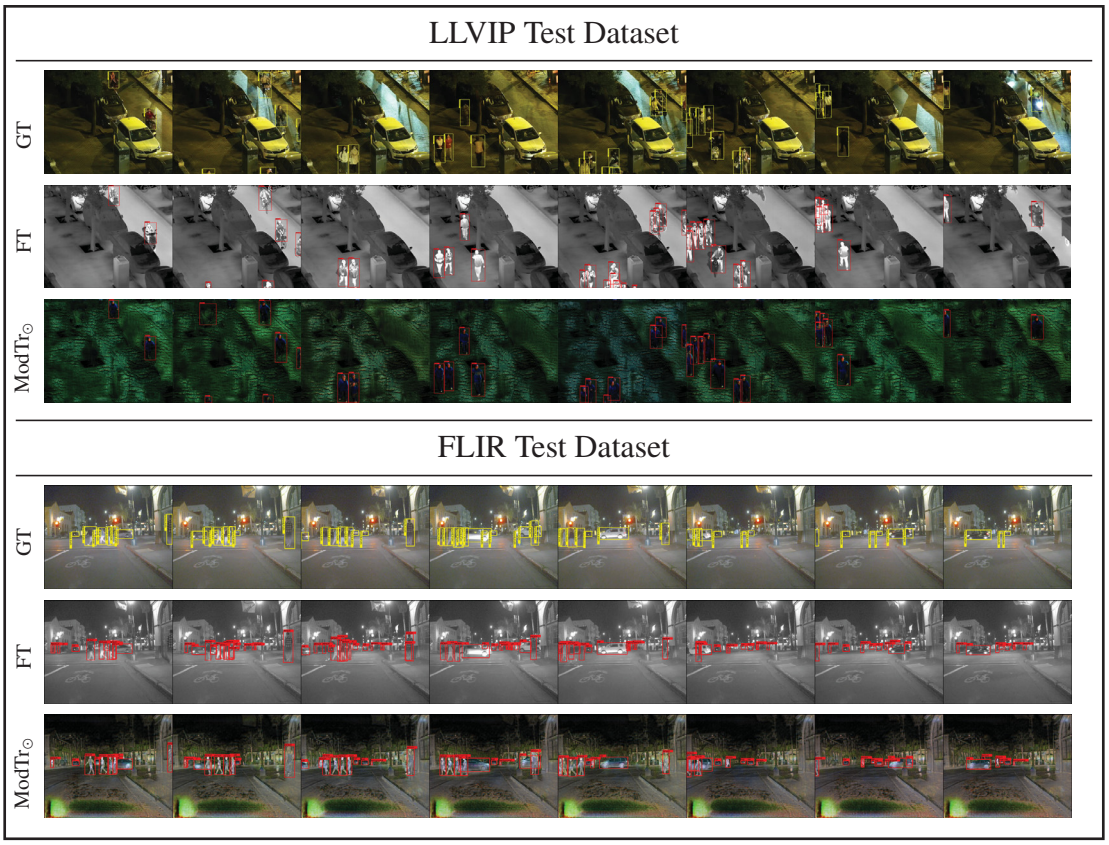


Figure 4.3 Illustration of a sequence of 8 images of LLVIP and FLIR dataset for Faster R-CNN. For each dataset, the first row is the RGB modality, followed by the IR modality and different representations created by ModTr. For visualizations of other detectors and variants of ModTr, please refer to the supplementary materials

for embedded devices, as it requires less memory. Based on the average performance on all datasets, ModTr obtains the best results.

4.4.6 Visualization of ModTr Translated Images

In Figure 4.3, we present qualitative results for LLVIP and FLIR, alongside a comparison with fine-tuning. Each dataset section includes three rows: the first row displays the ground-truth RGB images, the second row showcases the results of fine-tuning using IR, and the last row features ModTr with a Hadamard product-based fusion over the Faster R-CNN detector. Due to space constraints, additional visualizations for other detectors and fusion strategies are provided

Table 4.4 Detection performance (AP) of knowledge preserving techniques N-Detectors, 1-Detector, and N-ModTr-1-Detector, using three different detectors (FCOS, RetinaNet, and Faster R-CNN). The methods were evaluated on COCO and IR test sets of LLVIP and FLIR datasets

Detector	Dataset	N-Detectors	1-Detector	N-ModTr-1-Det.
FCOS	LLVIP	57.37 ± 2.19	**58.55 ± 0.89**	57.63 ± 0.66
	FLIR	27.97 ± 0.59	26.70 ± 0.48	**35.49 ± 0.94**
	COCO	**38.41 ± 0.00**	00.33 ± 0.04	**38.41 ± 0.00**
	AVG.	41.25 ± 0.92	28.52 ± 0.47	**43.84 ± 0.53**
RetinaNet	LLVIP	53.79 ± 1.79	53.26 ± 3.02	**54.83 ± 0.61**
	FLIR	28.46 ± 0.50	25.19 ± 0.72	**34.27 ± 0.27**
	COCO	**35.48 ± 0.00**	00.29 ± 0.01	**35.48 ± 0.00**
	AVG.	39.24 ± 0.76	26.24 ± 1.28	**41.52 ± 0.29**
Faster R-CNN	LLVIP	59.62 ± 1.23	**62.50 ± 1.29**	57.97 ± 0.85
	FLIR	30.93 ± 0.46	28.90 ± 0.33	**37.21 ± 0.46**
	COCO	**39.78 ± 0.00**	00.40 ± 0.00	**39.78 ± 0.00**
	AVG.	43.44 ± 0.56	30.60 ± 0.54	**44.98 ± 0.43**

in the supplementary materials. Notably, the IR results exhibit some false positives, particularly when detected objects overlap. Our method mitigates this issue effectively. Further insights, provided in the supplementary materials, reveal how our method effectively blurs or removes objects that do not belong to the target classes, thereby enhancing detection accuracy. Although the obtained intermediate representations are not visually pleasant, they prove more efficient for incorporating the knowledge necessary for the OD. Additionally, we conducted experiments with loss function terms aimed at enhancing the visual effects of the image, but they were not conclusive in terms of helping the detection performance.

4.4.7 Fine-tuning of ModTr and the Detector

The main reason to use ModTr is to avoid fine-tuning the detector for a specific task so that it can preserve its knowledge and be used for multiple modalities. However, in this section, we consider what would happen if we learn jointly ModTr and the detector weights. Results

are reported in Figure 4.4. We see that fine-tuning the detector can further boost performance. Thus, another application of ModTr could be used to improve the fine-tuning of a detector with a reduced additional computational cost.

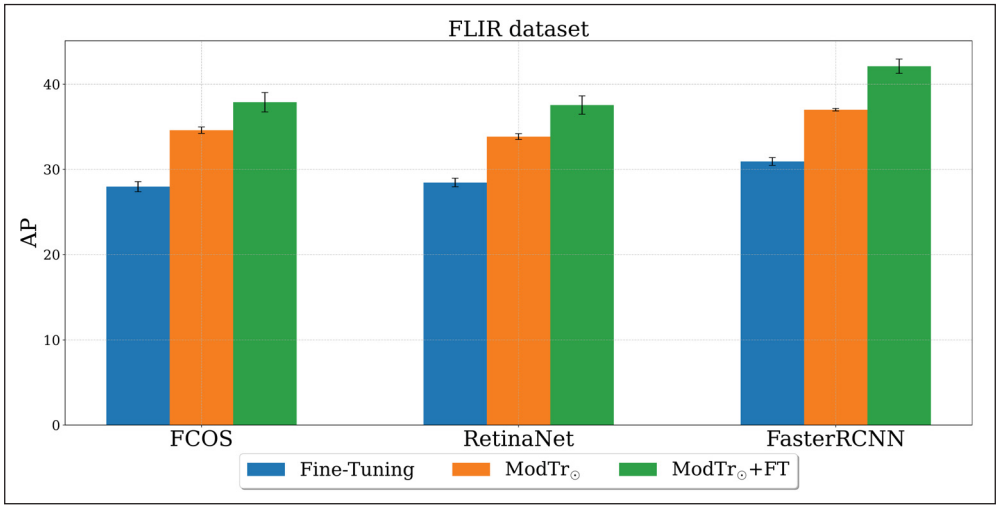


Figure 4.4 Comparison of the performance of fine-tuning the ModTr and normal fine-tuning on the FLIR dataset for the three different detectors (FCOS, RetinaNet, and Faster R-CNN). In blue, the Fine-tuning; in orange, the ModTr$_{\odot}$, and in green, ModTr$_{\odot}$ + FT

4.5 Conclusion

In this paper, a novel method called ModTr is proposed for adapting RGB object detectors (ODs) for IR modality without changing their parameters. A key advantage of our approach is that it preserves the full knowledge of the detector, allowing the translation network to act as a node that changes the modality for an unaltered detector. This is much more flexible and computationally efficient than having a specialized OD for each modality. Our approach performs well in various settings, outperforming powerful image-to-image models and previous competitors. We evaluated ModTr for different tasks, including detection based on image translation, comparison with traditional fine-tuning, and incremental IR modality application. Experimental results show the high performance and versatility of our method in all these settings.

Additionally, to explore integrating modalities beyond IR, we applied ModTr to Canny edges extracted from IR images as detailed in the supplementary material. While ModTr significantly enhances the performance of zero-shot RGB OD on edges, it still does not match the effectiveness of full fine-tuning on this other modality. We believe this shortfall arises from the limited information provided by edges compared to the richer data in the IR modality, leading to lower initial zero-shot OD performance. A potential solution is to replace the deterministic translator module within ModTr with a generative model. This substitution could enrich modality information by generating the missing data, potentially improving the zero-shot detector's performance. This promising direction will be explored in future research.

CHAPTER 5

VISUAL MODALITY PROMPT FOR ADAPTING VISION-LANGUAGE OBJECT DETECTORS

Heitor R. Medeiros[a] , Atif Belal[a] , Srikanth Muralidharan[a],
Eric Granger[a], Marco Pedersoli[a]

[a]Department of Systems Engineering, École de technologie supérieure,
1100 Notre-Dame West, Montreal, Quebec, Canada H3C 1K3

Paper published in *Proceedings of International Conference on Computer Vision, October 2025*

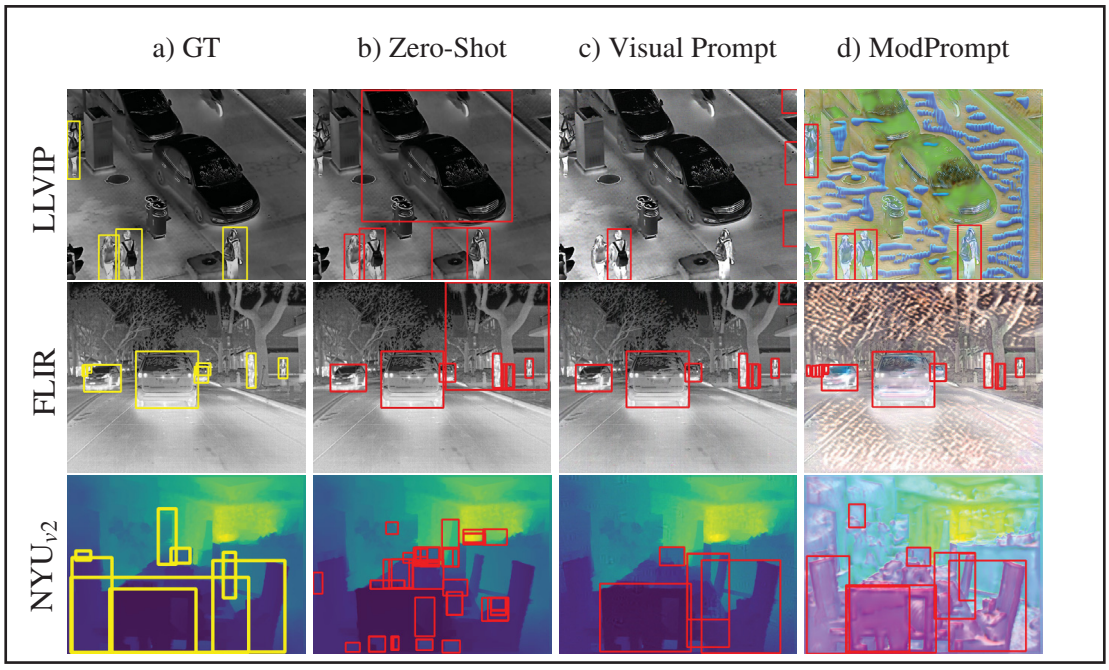


Figure 5.1 YOLO-World detections of different approaches across modalities: LLVIP and FLIR datasets (infrared) and NYU$_{v2}$ (depth). Each column corresponds to a different approach: (a) Ground Truth (GT): Shows in yellow the GT bounding boxes for objects. (b) Zero-Shot (ZS): Displays detections (in red) from a zero-shot model. It has missed several detections and some inaccurate boxes without tuning. (c) Visual Prompt: Illustrates detections with a visual prompt added to the image. It shows improvements over ZS, with more accurate detection in certain areas, but still misses some objects. (d) ModPrompt (Ours): Detections from our model. ModPrompt generates artifacts on the image to enhance objects and suppress the background, facilitating detection

Abstract

The zero-shot performance of object detectors degrades when tested on different modalities, such as infrared and depth. While recent work has explored image translation techniques to adapt detectors to new modalities, these methods are limited to a single modality and traditional detectors. Recently, vision-language detectors (VLDs), such as YOLO-World and Grounding DINO, have shown promising zero-shot capabilities; however, they have not yet been adapted for other visual modalities. Traditional fine-tuning approaches compromise the zero-shot capabilities of the detectors. The visual prompt strategies commonly used for classification with vision-language models apply the same linear prompt translation to each image, making them less effective. To address these limitations, we propose ModPrompt, a visual prompt strategy to adapt VLDs to new modalities without degrading zero-shot performance. In particular, an encoder-decoder visual prompt strategy is proposed, further enhanced by the integration of inference-friendly modality prompt decoupled residual, facilitating a more robust adaptation. Empirical benchmarking results show our method for modality adaptation on YOLO-World and Grounding DINO for challenging infrared (LLVIP, FLIR) and depth (NYU_{v2}) datasets, achieving performance comparable to full fine-tuning while preserving the model's zero-shot capability. Our code is available at https://github.com/heitorrapela/ModPrompt.

5.1 Introduction

Object detection (OD) is a key challenge in computer vision, aiming to localize and classify objects in images (Zou *et al.*, 2023). Recent OD advancements, driven by applications like autonomous driving (Stilgoe, 2018; Michaelis *et al.*, 2019), surveillance (Ramachandran & Sangaiah, 2021; Dubail *et al.*, 2022), and robotics (Eitel *et al.*, 2015; Pierson & Gashler, 2017; Ivorra *et al.*, 2018), have improved real-world systems significantly. Although traditional detectors like YOLO (Jocher, Chaurasia & Qiu, 2023) or DINO (Zhang *et al.*, 2023b) perform well, they are limited to fixed vocabularies, such as the 80 categories in the COCO (Lin *et al.*, 2014) dataset, and exhibit poor zero-shot performance (Bansal, Sikka, Sharma, Chellappa & Divakaran, 2018). These detectors are trained on predefined categories, restricting their adaptability to varying scenarios.

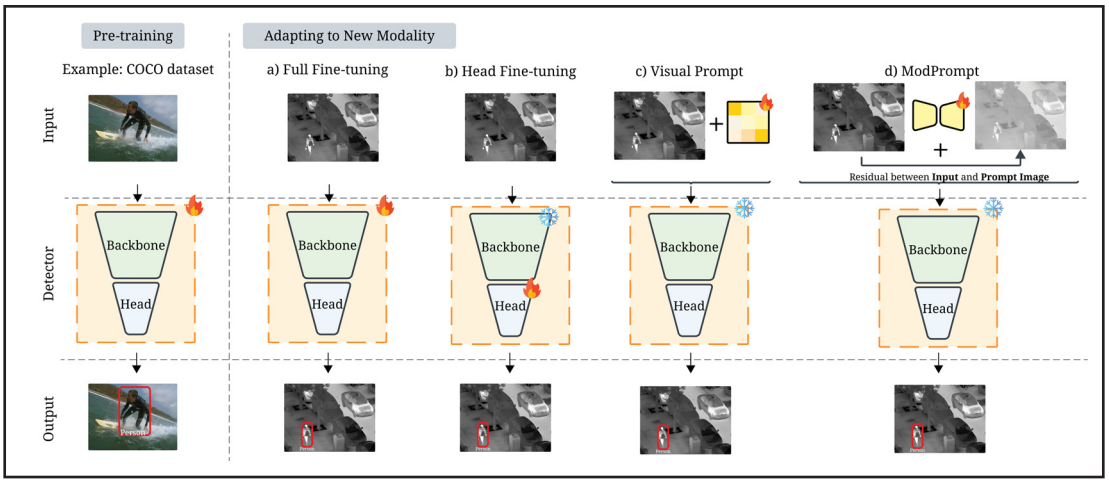


Figure 5.2 Strategies to adapt object detectors to new modalities: (a) Full Fine-tuning: Both the backbone (responsible for feature extraction) and the head (responsible for the final output, like object detection) are updated with new training data. (b) Head Fine-tuning: Only the head is fine-tuned while the backbone remains frozen. (c) Visual Prompt: Uses a visual prompt added to the input. The backbone and head remain unchanged, but the visual prompt guides the model to better interpret the new modality. (d) Our Modality Prompt: Similar to a visual prompt, the input image is combined with a prompt. The main difference is that here the prompt is not static; it is a transformation of the input image

In contrast, Vision-Language Models (VLMs) combine visual representations and semantic text embeddings, allowing them to dynamically understand pixel-level features in context Radford *et al.* (2021). VLMs enhance accuracy and flexibility by recognizing unseen objects through the integration of language, paving the way for open-vocabulary detection (Zang, Li, Zhou, Huang & Loy, 2022), such as Grounding DINO (Liu *et al.*, 2024a) and YOLO-World (Cheng *et al.*, 2024).

Generally, these open-vocabulary detectors are pre-trained on large-scale RGB datasets, making them capable of zero-shot detection Zhang *et al.* (2022); Liu *et al.* (2024a). However, their performance degrades if the domain shift is large, as in a different modality, such as infrared or depth. A common approach to mitigate this is by fine-tuning the VLMs on the downstream modality dataset. However, it makes the VLMs lose their zero-shot capabilities due to catastrophic forgetting (Lester, Al-Rfou & Constant, 2021). Some techniques have been explored to adapt

the VLMs to downstream tasks for image classification. These include text prompt tuning (TPT) (Zhou, Yang, Loy & Liu, 2022b,a) and visual prompt tuning (VPT) (Jia *et al.*, 2022; Bahng, Jahanian, Sankaranarayanan & Isola, 2022). Unlike traditional fine-tuning, prompt tuning involves adding learnable prompts to the model's input, allowing it to remain unchanged and preserve its zero-shot capability. VPT methods for classification learn visual prompt tokens to adapt the classifier of transformer-based models to the downstream tasks. Visual prompts are effective when adapting VLMs to a downstream task of the same modality (normally RGB). However, these methods are also less effective if the downstream data is from another visual modality with a large modality shift between pre-trained and downstream data, as shown in Table 5.1.

In addition to incorporating language into detectors, another major driver of advancements in real-world applications is the integration of diverse visual modalities, such as infrared (Medeiros, Latortue, Granger & Pedersoli, 2025b; Wang *et al.*, 2022; Li, Li, Li, Li & Xu, 2021) and depth (Li *et al.*, 2023b; Xu, Li, Wu & Luo, 2017b; Yang *et al.*, 2022b). These modalities provide additional spatial information, enhancing visibility in low-light and obstructed conditions, resulting in more robust and accurate object detectors (Bustos *et al.*, 2023). Adapting detectors for different modalities unlocks their potential in specialized contexts by combining the efficiency of transfer learning with modality-specific enhancements. This approach bridges the gap between general-purpose detectors capable of zero-shot detection and the nuanced demands of novel modalities, where labels are still scarce compared to RGB data.

Therefore, to overcome limitations in modality adaptation, image translation (Herrmann *et al.*, 2018; Özkanoğlu & Ozer, 2022) methods have been explored for traditional detector-based models. However, these approaches primarily optimize the image translation loss rather than directly focusing on improving detection performance. More recent approaches, such as HalluciDet (Medeiros *et al.*, 2024c) and ModTr (Medeiros *et al.*, 2024b), have demonstrated improvement by emphasizing detection loss optimization during adaptation. However, these methods are limited to traditional detectors without language mechanisms, and they did not explore more challenging modalities such as depth. For HalluciDet, the prior pre-trained

knowledge is lost during the adaptation phase. In ModTr, the COCO pre-trained knowledge is preserved, but it lacks the zero-shot powerful capability of VLMs. In contrast, our work leverages recent vision-language detectors, incorporating textual information and allowing for adaptation to different modalities, exploring the benefits of both visual and text information for improving modality adaptation.

In this paper, we propose ModPrompt, a visual prompt strategy to adapt vision-language ODs in the input space to new visual modalities without sacrificing zero-shot knowledge of the vision-language detector. ModPrompt is a detector-agnostic encoder-decoder visual prompt strategy that can easily be integrated with any VLM detector, regardless of the detector's backbone type. We also introduce an inference-friendly residual mechanism named Modality Prompt Decoupled Residual (MPDR) capable of adapting the learnable text embeddings with the ModPrompt loss, while preserving full knowledge of the model due to the decoupled embedding parameters. We demonstrate that the combination of MPDR with ModPrompt improves the performance of the detector across different modalities while retaining the original zero-shot knowledge, both in the vision and language components. To the best of our knowledge, this is the first work that focuses on adapting state-of-the-art VLM-based OD models to new modalities. Furthermore, ModPrompt is validated on two state-of-the-art VLM ODs, Grounding DINO and YOLO-World, and across four widely used visual modalities, infrared, depth, event-based, and LiDAR.

Our main contributions can be summarized as follows.

(1) We introduce ModPrompt, a novel pixel-level prompt adaptation strategy based on input translation. Our method is backbone-agnostic, leading to improved adaptability of vision-language detectors across different modalities and preserving the vision encoder knowledge. Additionally, we introduce MPDR, a text-embedding modality prompt decoupled residual mechanism, which preserves full knowledge of the text-embedding while increasing the performance of the model for the novel modality.

(2) We conduct a detailed investigation of various pixel-level visual prompt techniques, offering insights into why traditional prompt methods struggle with adapting object detectors to new modalities. Our analysis demonstrates how ModPrompt overcomes these limitations to significantly enhance detection performance.

(3) We provide a comprehensive benchmark on the visual prompts for modality adaptation across two primary open-vocabulary OD models, YOLO-World and Grounding DINO. We evaluate our method across infrared, depth, even-based, and LiDAR datasets, achieving in some cases a level of performance comparable to full fine-tuning while preserving the model's zero-shot capability.

Table 5.1 Detection performance (APs) for YOLO-World and Grounding DINO for the two main datasets evaluated: LLVIP-IR and NYU_{v2}-Depth. The different visual prompt adaptation techniques are compared with our ModPrompt, and the zero-shot (ZS), head finetuning (HFT), and full finetuning (FT) are also reported, where the full finetuning is the upper bound

Dataset	Method	YOLO-World			Grounding DINO		
		AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP
LLVIP - IR	Zero-Shot (ZS)	81.00 ± 0.00	57.80 ± 0.00	53.20 ± 0.00	85.50 ± 0.00	62.70 ± 0.00	56.50 ± 0.00
	Head Finetuning (HFT)	93.57 ± 0.05	73.83 ± 0.19	64.80 ± 0.08	87.53 ± 0.06	65.57 ± 0.23	58.10 ± 0.20
	Full Finetuning (FT)	97.43 ± 0.05	77.93 ± 0.21	67.73 ± 0.09	97.17 ± 0.31	79.93 ± 0.83	67.83 ± 0.96
	Visual Prompt (Fixed)	70.30 ± 7.89	45.67 ± 6.97	43.53 ± 5.79	83.83 ± 0.06	61.53 ± 0.23	55.13 ± 0.15
	Visual Prompt (Random)	60.13 ± 0.29	38.73 ± 0.17	36.87 ± 0.12	83.87 ± 0.06	61.37 ± 0.06	55.03 ± 0.06
	Visual Prompt (Padding)	79.87 ± 1.00	51.77 ± 0.90	49.30 ± 0.83	82.73 ± 0.31	60.00 ± 0.35	55.13 ± 0.15
	Visual Prompt (WM)	82.00 ± 1.59	53.90 ± 1.06	50.90 ± 0.94	69.57 ± 0.93	41.37 ± 1.27	40.77 ± 0.87
	Visual Prompt (WM_{v2})	74.10 ± 0.43	46.47 ± 0.62	44.70 ± 0.22	69.87 ± 1.12	41.77 ± 1.30	41.13 ± 0.96
	ModPrompt (Ours)	**92.80 ± 0.29**	**70.73 ± 1.02**	**62.87 ± 0.63**	**93.13 ± 0.15**	**67.17 ± 0.78**	**60.10 ± 0.50**
NYU_{v2} - Depth	Zero-Shot (ZS)	04.80 ± 0.00	03.10 ± 0.00	03.00 ± 0.00	08.30 ± 0.00	05.60 ± 0.00	05.30 ± 0.00
	Head Finetuning (HFT)	21.03 ± 0.12	12.03 ± 0.37	12.00 ± 0.14	18.24 ± 0.85	13.23 ± 0.46	12.10 ± 0.21
	Full Finetuning (FT)	49.90 ± 0.08	36.40 ± 0.29	33.57 ± 0.26	51.60 ± 2.09	39.17 ± 1.88	35.77 ± 1.70
	Visual Prompt (Fixed)	04.67 ± 0.05	03.07 ± 0.05	02.90 ± 0.00	08.27 ± 0.06	05.57 ± 0.06	05.27 ± 0.06
	Visual Prompt (Random)	04.23 ± 0.12	02.63 ± 0.05	02.53 ± 0.05	08.33 ± 0.06	05.53 ± 0.06	05.27 ± 0.06
	Visual Prompt (Padding)	03.97 ± 0.05	02.50 ± 0.00	02.43 ± 0.05	07.63 ± 0.06	05.17 ± 0.06	04.63 ± 0.06
	Visual Prompt (WM)	10.67 ± 0.12	06.90 ± 0.22	06.57 ± 0.05	04.87 ± 0.06	02.97 ± 0.06	02.97 ± 0.06
	Visual Prompt (WM_{v2})	10.63 ± 0.12	06.67 ± 0.12	06.50 ± 0.08	05.00 ± 0.10	03.07 ± 0.06	03.03 ± 0.06
	ModPrompt (Ours)	**37.17 ± 0.57**	**27.50 ± 0.64**	**24.93 ± 0.50**	**21.70 ± 0.20**	**15.03 ± 0.29**	**14.13 ± 0.21**

5.2 Related Works

Open-Vocabulary Object Detection. Recent advances in open-vocabulary OD exploit vision-language models (VLMs) with region-level multimodal alignment to recognize novel concepts. RegionCLIP (Zhong *et al.*, 2022) uses pseudo labels for contrastive pretraining, while Gu et

al. (Gu *et al.*, 2022) employ knowledge distillation with CLIP-based region embeddings. Lin et al. (Lin *et al.*, 2023) leverage bipartite matching for image-text alignment. GLIP (Li *et al.*, 2022b) and GLIPv2 (Zhang *et al.*, 2022) enhance zero-shot detection via multimodal fusion and region-grounded pretraining. Grounding DINO (Liu *et al.*, 2024a) introduces multiphase image-text fusion, and YOLO-World (Cheng *et al.*, 2024) enables efficient open-vocabulary detection with reparametrizable fusion.

Downstream Adaptation of Large Vision Models. Recent work has focused on parameter-efficient adaptation of large vision models without erasing prior knowledge. Co-op (Zhou *et al.*, 2022b) learns continuous text prompts for downstream tasks. Visual Prompt Tuning (Jia *et al.*, 2022) adds tunable parameters at various network stages. Bahng et al. (Bahng *et al.*, 2022) propose image-space linear probing. CLIP-Adapter (Gao *et al.*, 2024) introduces feature-space adaptation via bottlenecks and residuals. Yu et al. (Yu, Lu, Jin, Chen & Wang, 2023) use task residuals to decouple new parameters for classification. While these methods seem effective, they are not suitable for detection.

Image Translation for Detection. Image translation methods adapt images across domains by modifying domain-specific characteristics while preserving content (Pang *et al.*, 2021), often using generative models (Goodfellow *et al.*, 2014; Kingma & Welling, 2022). Pix2Pix (Isola *et al.*, 2017a) leverages paired supervision, while CycleGAN (Zhu *et al.*, 2017a) enables unpaired translation. For OD, HalluciDet (Medeiros *et al.*, 2024c) performs a two-step adaptation using RGB fine-tuning and translation, whereas ModTr (Medeiros *et al.*, 2024b) employs zero-shot COCO-trained detectors with fusion strategies to preserve prior knowledge. In contrast, we adapt large VLM-based ODs (e.g., YOLO-World, Grounding DINO) to novel modalities (e.g., IR, depth) while retaining their zero-shot capabilities.

5.3 Proposed Method

5.3.1 Preliminary Definitions

Vision Language Object Detection. The training dataset for traditional object detectors can be represented as $\mathcal{D} = \{(X, Y)\}_{j=1}^{M}$. Here, $X \in \mathbb{R}^{W \times H \times C}$ represents an image with dimensions $W \times H$ and C channels, and $Y = \{(b_i, c_i)\}_{i=1}^{N}$ consist of bounding boxes b_i and object category c_i. For the visual language model, the annotations are reformulated as a region-text pair $Y = \{(b_i, t_i)\}_{i=1}^{N}$, where t_i corresponds to the text for the region b_i. Specifically, t_i is the name of the object category in b_i. The objective of the object detection task is to accurately identify all objects in a given image. The average precision (AP) metric across classes is used as the standard evaluation protocol. An object detector is formally represented as the mapping $f_\theta : \mathbb{R}^{W \times H \times C} \rightarrow \hat{Y}$, where θ denotes the parameter vector. The detection cost function $C_{det}(\theta)$, a differentiable proxy for the AP metric, is computed as an average loss of the detection loss $\mathcal{L}_{det}$ over the entire dataset $\mathcal{D}$, described mathematically as:

$$C_{det}(\theta) = \frac{1}{|\mathcal{D}|} \sum_{(x,Y)\in\mathcal{D}} \mathcal{L}_{det}(f_\theta(x), Y). \tag{5.1}$$

Visual Prompt for OD. Given a frozen pre-trained detector, the objective of the Visual Prompt method is to learn a task-specific visual prompt v_ϕ, parametrized by ϕ, which can be combined with the input x to improve the final detection performance. During test time, this visual prompt is added to the test images to adapt them for the desired detection task. As described in the following equation:

$$C_{\text{vp}}(\phi) = \frac{1}{|\mathcal{D}|} \sum_{(x,Y)\in\mathcal{D}} \mathcal{L}_{det}(f_\theta(x + v_\phi), Y). \tag{5.2}$$

5.3.2 ModPrompt

The visual prompt methods discussed above learn simple linear transformations, like adding a fixed patch to each image. These transformations are learned during the training process and are not conditioned on the input image at inference, making them less effective. In ModPrompt, we incorporate a function h_ϑ, an encoder-decoder inspired by U-Net (Ronneberger *et al.*, 2015), dependent on the input image x, which is trained conditioned on labels Y. Our encoder-decoder is flexible in terms of the encoder backbone initialized from pre-trained weights. We adapt the last layer of the decoder to constrain it to be 3-channels with values ranging between 0 and 1 to simulate a pseudo-RGB image. Note that this mechanism is completely different from traditional U-Nets, which are trained to reconstruct and generate c-channels for semantic segmentation tasks, requiring both input and ground truth segmentation map during training. However, ModPrompt requires access only to the pre-trained RGB detector and final target modality data, so it is based on guidance detection loss instead of reconstruction. The ModPrompt training cost is defined by the following equation:

$$C_{\text{mp}}(\vartheta) = \frac{1}{|\mathcal{D}|} \sum_{(x,Y)\in\mathcal{D}} \mathcal{L}_{det}(f_\theta(x + h_\vartheta(x)), Y). \tag{5.3}$$

This translation function (f_θ) is responsible for adapting the input image to a modality representation optimally suited to address detection in the target modality. Our idea of a visual ModPrompt is inspired by text conditional prompt (Zhou *et al.*, 2022a) work, which incorporates a small network for the text encoder conditional adaptation. In our approach, the visual prompt depends on the input image.

5.3.3 Modality Prompt Decoupled Residual with Knowledge Preservation (MPDR)

The vision-language models take both text and image as input. ModPrompt aims to transform the input image to adapt to the new modality. For the adaptation of the textual part, text prompt tuning

methods (Zhou *et al.*, 2022b,a) have been explored for classification. But these methods require back-propagating through the whole text backbone, making them computationally expensive. Recently, task residual Yu *et al.* (2023) was proposed that tunes the precomputed text embedding of the vision-language classifiers. Inspired by this, we design a strategy to efficiently tune the embedding of the vision-language detectors. As illustrated in Figure 5.3, offline text embeddings are generated for each object category. Then, the MPDR, which are learnable embedding tokens, are tuned to adapt to a specific modality end-to-end with ModPrompt loss. The MPDR is crucial for decoupling the knowledge from the original embedding, which preserves prior text-embedding knowledge and is able to incorporate new knowledge during the adaptation stage. During the training, the loss of ModPrompt and MPDR work in synergy to adapt to any new modality. In the test phase, the MPDR can be deactivated easily by zero-masking its parameters, which is the same as full zero-shot embedding knowledge, or enabled, which is the adapted embedding for this modality (that's why decoupling the knowledge is important) and doesn't introduce any inference overhead. The following equation describes the final training cost:

$$C_{\text{mp-tp}}(\vartheta, \phi) = C_{\text{mp}}(\vartheta) + C_{\text{tp}}(\phi), \tag{5.4}$$

where $C_{\text{mp}}(\vartheta)$ is the optimization of parameters ϑ of the encoder-decoder function, and $C_{\text{tp}}(\phi)$ is the text-prompt adaptation on the embedding space by training new parameters ϕ for the class embeddings in the offline vocabulary.

5.3.4 Training Summary

For training the ModPrompt with MPDR, we first pre-compute the text embeddings for the target classes, using the text encoder. Then, we sum the learnable embedding residuals with the frozen embeddings and train them together with the ModPrompt. Note that, different from other embedding adaptation approaches, our MPDR can be used with ModPrompt for full zero-shot knowledge preservation due to the residual decoupled embedding learning strategy. The cost

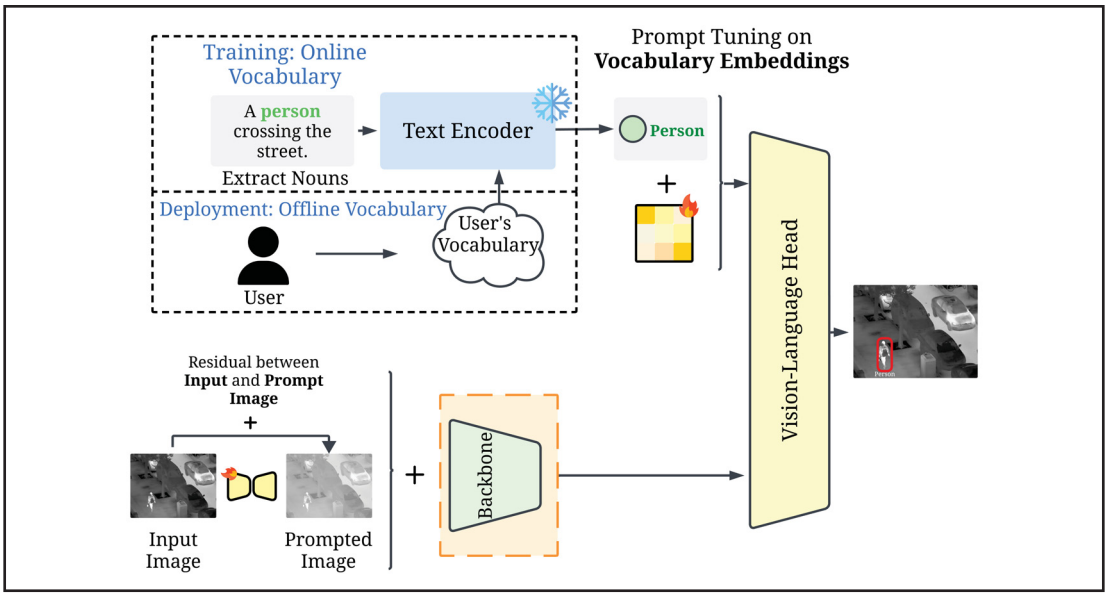


Figure 5.3 Our proposed strategy for text-prompt tuning: an inference-friendly and knowledge-preserving decoupled embedding tuning method. An offline embedding is generated for each object category, and then a novel decoupled residual trainable parameters and the ModPrompt are integrated into the detector to adapt it to new modalities

function is based on Equation 5.4, but instead of the online cost function for text $C_{\text{tp}}(\phi)$, we work on the text embedding space, which is better suited for efficient training and inference.

5.4 Results and Discussion

5.4.1 Experimental Methodology

(a) Datasets: LLVIP: LLVIP is a surveillance dataset composed of 12,025 IR images in the training set and 3,463 IR images in the test set with only pedestrians annotated. **FLIR ALIGNED:** We used the sanitized and aligned paired sets provided by Zhang et al. Zhang *et al.* (2020b). It has 4,129 training IR images and 1,013 IR test images. FLIR images are captured from the perspective of a camera in the front of a car, with a resolution of 640 by 512. It contains the bicycles, dogs, cars, and people classes. Following Cao *et al.* (2023b), we decided to remove “dog" objects because the number of annotations is inadequate for training

Table 5.2 Detection performance (APs) for YOLO-World and Grounding DINO on LLVIP-IR and NYU$_{v2}$-Depth datasets. Each visual prompt adaptation strategy is compared with the learnable MPDR (gains in parentheses), which updates the new modality embeddings while preserving the original embedding knowledge

Detector	Method	LLVIP-IR			NYU$_{v2}$-DEPTH		
		AP$_{50}$	AP$_{75}$	AP	AP$_{50}$	AP$_{75}$	AP
YOLO-World	Fixed	86.60 ± 0.00 (+16.3)	63.53 ± 0.05 (+17.8)	57.67 ± 0.05 (+14.1)	12.47 ± 0.05 (+7.80)	07.50 ± 0.00 (+4.43)	07.37 ± 0.05 (+4.47)
	Random	86.43 ± 0.05 (+26.3)	63.40 ± 0.08 (+24.6)	57.50 ± 0.00 (+20.6)	12.23 ± 0.05 (+8.00)	07.30 ± 0.08 (+4.67)	07.20 ± 0.00 (+4.67)
	Padding	83.57 ± 0.12 (+3.70)	59.03 ± 0.05 (+7.26)	54.23 ± 0.05 (+4.93)	10.63 ± 0.09 (+6.66)	05.60 ± 0.08 (+3.10)	05.97 ± 0.05 (+3.54)
	WeightMap	87.47 ± 0.17 (+5.47)	62.23 ± 0.31 (+8.33)	57.43 ± 0.05 (+6.53)	13.43 ± 0.09 (+2.76)	08.07 ± 0.12 (+1.17)	07.90 ± 0.08 (+1.33)
	ModPrompt	**96.60 ± 0.37 (+3.80)**	**77.27 ± 0.33 (+6.54)**	**67.37 ± 0.05 (+4.50)**	**44.67 ± 0.17 (+7.50)**	**32.53 ± 0.66 (+5.03)**	**29.93 ± 0.12 (+5.00)**
Grounding DINO	Fixed	88.70 ± 0.66 (+4.87)	67.73 ± 0.50 (+6.20)	60.13 ± 0.45 (+5.00)	10.00 ± 0.00 (+1.73)	06.40 ± 0.28 (+0.83)	06.40 ± 0.28 (+1.13)
	Random	88.80 ± 0.95 (+4.93)	68.00 ± 1.21 (+6.63)	60.37 ± 0.90 (+5.33)	09.90 ± 0.14 (+1.57)	06.50 ± 0.00 (+0.97)	06.15 ± 0.07 (+0.88)
	Padding	86.93 ± 0.55 (+4.20)	65.23 ± 1.24 (+5.23)	58.07 ± 1.10 (+4.17)	09.27 ± 0.15 (+1.64)	06.00 ± 0.17 (+0.83)	05.57 ± 0.15 (+0.87)
	WeightMap	74.90 ± 0.92 (+5.33)	46.80 ± 0.70 (+5.43)	45.77 ± 0.46 (+5.00)	06.17 ± 0.12 (+1.30)	03.53 ± 0.15 (+0.56)	03.60 ± 0.10 (+0.63)
	ModPrompt	**93.73 ± 0.25 (+0.60)**	**68.27 ± 0.40 (+1.10)**	**60.87 ± 0.50 (+0.77)**	**26.63 ± 0.38 (+4.93)**	**18.53 ± 0.32 (+3.50)**	**17.27 ± 0.46 (+3.14)**

with the FLIR dataset; results of FLIR can be seen in the supp. material due to space constraints.

NYU$_{v2}$:. The NYU-Depth V2 (Silberman *et al.*, 2012) dataset provides a valuable collection of indoor video sequences captured with a Microsoft Kinect camera. In this dataset, we used the depth information, which is composed of 795 training images and 654 test images with a resolution of 640 by 480. The dataset comprised 19 different classes, such as bathtubs, beds, bookshelves, boxes, and chairs. In all our experiments, only IR and depth images were used.

(b) Implementation Details: The YOLO-World models were trained on an A100 NVIDIA GPU, and Grounding DINO on V100, and were implemented using PyTorch. For YOLO-World, we used the original work (Cheng *et al.*, 2024), which is based on MMDET (Chen *et al.*, 2019b), and for Grounding DINO, we used the open-source version from the MMDET library. For the YOLO-World, we use both YOLO-World-S-v1 and YOLO-World-S-v2, with the YOLOv8 backbone and CLIP (Radford *et al.*, 2021) as the language model. For Grounding DINO, we use Swin-Transformer Tiny (Liu *et al.*, 2021b) as the detector backbone and BERT (Devlin, Chang, Lee & Toutanova, 2019) for the language encoder. Our models are evaluated in terms of AP performance using COCO API (Lin *et al.*, 2014). The batch size for YOLO-World models was 8, and the models were trained with a single GPU. For Grounding DINO, the batch size was 16, and 2 GPUs were used for training using DDP. Yolo-World and Grounding Dino were trained with the maximum training epoch of 80 and 60, respectively. For offline embedding extraction,

we used CLIP-ViT-base-patch32 and Bert-base-uncased for YOLO-World and Grounding DINO, respectively. Additional details are provided in the supp. materials.

(c) Baseline Methods:

- Zero-shot Model. The zero-shot model refers to performing the inference with the original pre-trained open-vocabulary detector model, but on our final detection modality. Normally, these models are trained with large RGB datasets, such as Object365 (Shao *et al.*, 2019a) or GoldG (Kamath *et al.*, 2021).

- Finetuning. The finetuning strategy refers to adapting a pre-trained model by changing its parameters directly for the final task, but it destroys the pre-training knowledge of the model. The most common strategies for adapting detectors are the *head finetuning*, which adapts only the head of the detector, and *full finetuning*, which adapts all the weights of the detector.

- Visual Prompt. For this baseline, we define a family of different visual prompt strategies explored in our work. Specifically, we have *fixed*, *random*, *padding*, and *weight map*. In *fixed*, we add a fixed patch in a fixed position for all training. In *random*, we have a random padding that is sampled per image. In *padding*, we add a padding prompt and combine it with the image. In *weight map*, we consider the learnable weights to be a mask with the dimensions of the input image.

- Modality Prompt. Here, we refer to our proposed encoder-decoder visual prompt, in which we learn how to translate the input, guided by the final detection loss, and then we combine its output with the original input.

5.4.2 Visual Modality Adaptation

This section explores two main open-vocabulary detectors, YOLO-World and Grounding DINO, for the input space visual prompt adaptation on IR and depth modalities. In the Table 5.1, we tabulate the results using YOLO-World and Grounding DINO. We compare zero-shot (ZS), full-finetuning (FT) of the visual backbone, visual prompt with fixed patch, random patch, padding patch, weight map (WM) on all pixels, and weight map with scale shift (WM_{v2}), and our ModPrompt. Every visual prompt strategy adapts at the input level, so we keep all other

Table 5.3 Detection performance (APs) for YOLO-World under the two main datasets evaluated: LLVIP-IR and NYU_{v2}-Depth. We compared the main visual prompt strategies *fixed*, *random*, *padding*, and ModPrompt. The variations consist of the number of prompt pixels (p_s = 30, 200, or 300) and for ModPrompt, the MobileNet (MB) or ResNet (RES)

Method	**Variation**	**LLVIP - IR**		
		AP_{50}	**AP_{75}**	**AP**
Fixed	30	61.60 ± 0.75	39.93 ± 0.52	37.97 ± 0.56
	300	70.30 ± 7.89	45.67 ± 6.97	43.53 ± 5.79
Random	30	60.13 ± 0.29	38.73 ± 0.17	36.87 ± 0.12
	300	56.27 ± 0.46	33.73 ± 0.62	33.13 ± 0.42
Padding	30	79.87 ± 1.00	51.77 ± 0.90	49.30 ± 0.83
	300	39.53 ± 2.36	15.90 ± 1.02	19.07 ± 1.18
ModPrompt	MB	**92.80 ± 0.29**	**70.73 ± 1.02**	**62.87 ± 0.63**
	RES	91.03 ± 0.12	68.40 ± 1.10	61.43 ± 0.58
Method	**Variation**	**NYU_{v2} - Depth**		
		AP_{50}	**AP_{75}**	**AP**
Fixed	30	04.67 ± 0.05	03.07 ± 0.05	02.90 ± 0.00
	300	03.43 ± 0.05	02.00 ± 0.08	02.10 ± 0.00
Random	30	04.23 ± 0.12	02.63 ± 0.05	02.53 ± 0.05
	300	01.53 ± 0.17	00.77 ± 0.12	00.87 ± 0.12
Padding	30	03.97 ± 0.05	02.50 ± 0.00	02.43 ± 0.05
	200	00.37 ± 0.12	00.10 ± 0.08	00.17 ± 0.05
ModPrompt	MB	35.37 ± 0.12	25.20 ± 0.24	23.27 ± 0.17
	RES	**37.17 ± 0.57**	**27.50 ± 0.64**	**24.93 ± 0.50**

parameters frozen to avoid catastrophic forgetting. In this setup, the YOLO-World with LLVIP-IR dataset, we observe that ModPrompt outperforms the other prompt strategies with AP_{50} 92.80 and the second best was WM with 82.00, similar trends for the AP_{75} with ModPrompt reaching 70.73 and WM with 53.90. Additionally, in terms of AP, our method has 62.87 AP, while the second-best approach has 50.90 AP. For Grounding DINO, in LLVIP-IR, our ModPrompt had

an 93.13 AP_{50}, 67.17 AP_{75}, and 60.10 AP, and the second best (fixed patch) had an AP_{50} of 83.87, but close to random and padding prompts. We observe that ModPrompt performs better when objects are well-defined in the image and when objects are not too small, otherwise, like all other input-level pixel strategies it struggles, especially on refined bounding-box localization, which can be seen with AP_{75} and AP, whereas in AP_{50} it shows good results. Performance on the FLIR dataset is reported in the supp. material.

Comparison with SOTA Modality Translation OD methods. Our ModPrompt technique is compared with recent works on modality translators for ODs: HalluciDet (Medeiros *et al.*, 2024c) and ModTr (Medeiros *et al.*, 2024b). In Fig. 5.4, we observe that our results are better in all APs for the LLVIP dataset. Additional results for the FLIR-IR dataset are presented in the supplementary material.

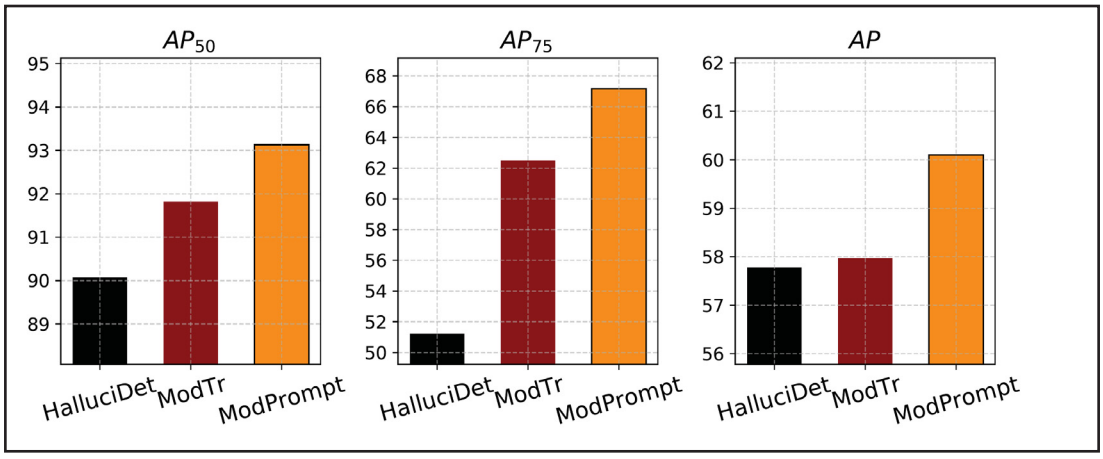


Figure 5.4 Detection performance on LLVIP dataset of different SOTA Modality Translation OD methods in terms of APs

Comparison of the number of Train Parameters and Knowledge Preservation. In Table 5.4 we compare the number of training parameters and show the catastrophic forgetting in the compared baseline methods. As we can see, ModPrompt preserves the zero-shot performance on the COCO dataset, while the performance drops significantly in HFT and FT. Other visual

prompt strategies, such as WM and WM_2, can preserve the zero-shot performance, but our average performance is much better than theirs while requiring fewer trainable parameters.

Table 5.4 AP_{50} of YOLO-World on LLVIP-IR and COCO data. We compare the number of trainable parameters and show the catastrophic forgetting in HFT and FT baselines

Method	**Params (M)**	**LLVIP**	**COCO**	**Avg.**
ZS	0.00	81.00 ± 0.00	51.90 ± 0.00	66.45 (00.00)
HFT	2.31	93.57 ± 0.05	00.66 ± 0.04	47.12 (-19.33)
FT	76.81	97.43 ± 0.05	00.10 ± 0.00	48.77 (-17.68)
WM	3.93	87.47 ± 0.17	51.90 ± 0.00	69.69 (+3.24)
WM_{v2}	7.86	85.00 ± 0.08	51.90 ± 0.00	68.45 (+2.00)
ModPrompt	3.08	95.63 ± 0.04	51.90 ± 0.00	**73.77 (+7.32)**

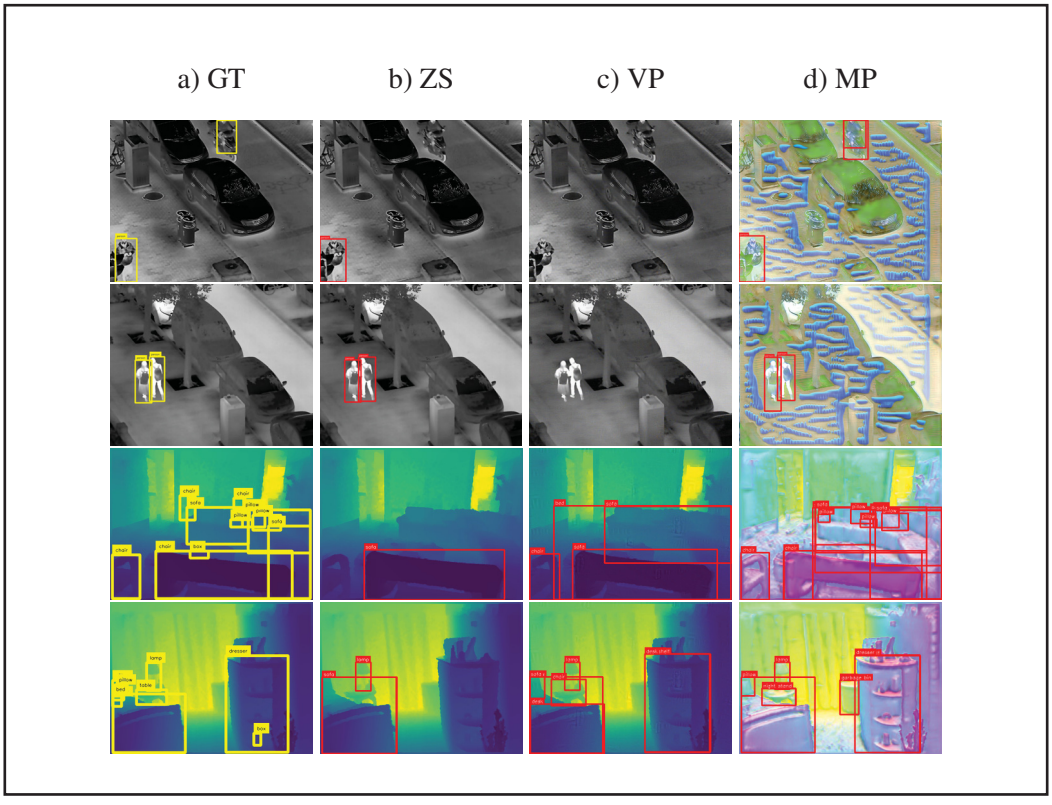


Figure 5.5 Detections for YOLO-World for the different approaches: First two rows for LLVIP (infrared), and last two rows for NYU_{v2} (depth). Each column corresponds to a different approach: (a) GT (Ground Truth): yellow boxes. (b) ZS: zero-shot detections (red). (c) VP: detections with visual prompts. (d) MP (Ours): detections with ModPrompt

5.4.3 Ablation Studies

Visual Prompts. We evaluate different variations of the visual prompt adaptation methods introduced in our paper. Specifically, we compare the performance when different input patch sizes are used; for instance, $p_s = 30$ refers to a patch size of 30 pixels. In this study, we tested various patch sizes for each of the visual prompt methods and reported the performance in Table 5.3. We evaluate ModPrompt using two different translators with U-Net based backbones-MobileNet (MB) (Howard *et al.*, 2017) and ResNet (RES) (He *et al.*, 2016). Empirically, we observe minimal variation in results when increasing the prompt patch size. In some cases, a larger prompt could even degrade the performance of the detector on the new modality, as seen with $p_s = 200$ on NYU_{v2} datasets. A similar trend can be observed with ModPrompt, where sometimes the MB version, with far fewer parameters, can outperform the RES version. This finding suggests that a lighter MB translator can be a viable choice over a heavier RES translator, making it better suited for real-time performance, with minimal degradation in detection metrics.

MPDR Knowledge Preservation mechanism. Table 5.2 shows that MDPR can be beneficial for improving the majority of the visual prompt strategies while preserving the original knowledge of the detector. For the text adaptation setting, we pre-compute the embeddings for the target classes using our language model, avoiding forward/backpropagation on it. In this setting, we also tried performing adaptations directly on the embeddings without using MPDR parameters, similar to text/embedding prompt tuning; in this case, the performance on the final target modality was close to MDPR. However, the zero-shot knowledge of the model was lost; thus, we opted to report only the MDPR in the main manuscript.

Broader Modality Generalization. To further demonstrate the effectiveness of ModPrompt, we extended our evaluation to include Event-based (PEDRo) (Boretti *et al.*, 2023) and LiDAR (STCrowd) (Cong *et al.*, 2022) modalities. As shown in Table 5.5, ModPrompt achieves strong performance, outperforming HFT and the visual prompt baseline. These results demonstrate

that our method generalizes effectively across four distinct modalities (IR, Depth, Event-based, and LiDAR), confirming its robustness beyond the original experimental scope.

Table 5.5 Results on PEDRo (event-based) and STCrowd (LiDAR) datasets on YOLO-World (YW)

Method	Event-based (PEDRo)			LiDAR (STCrowd)		
	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP
ZS	23.9	08.9	11.4	00.0	00.0	00.0
HFT	62.1	27.7	31.4	56.8	21.8	27.0
FFT	94.1	78.7	66.9	88.5	61.2	55.7
WM	49.9	21.3	24.7	34.8	10.0	14.6
ModPrompt	92.9	74.2	63.8	85.7	54.8	51.2

Study on Computational Cost. Our MPDR module is explicitly designed for inference-friendliness. Since text embeddings are precomputed, no forward pass through the text encoder is required either during inference or training. As shown in Figure 5.6, the MPDR module reduces the training time and increases the inference FPS. Although MPDR and MPDR WM require less computation, our MPDR ModPrompt achieves substantially higher detection performance while maintaining reasonable efficiency.

5.4.4 Qualitative Results

Figure 5.5 shows the visualization of ModPrompt on YOLO-World compared with zero-shot and the best visual prompt baseline. Thus, we can see that visual prompt-based methods, in general, struggle a bit to provide attached bounding boxes, but they do a good job of adapting. The VP method could not detect people in the LLVIP image well, while ModPrompt did a good job. Additional visualizations are provided in the supplementary material.

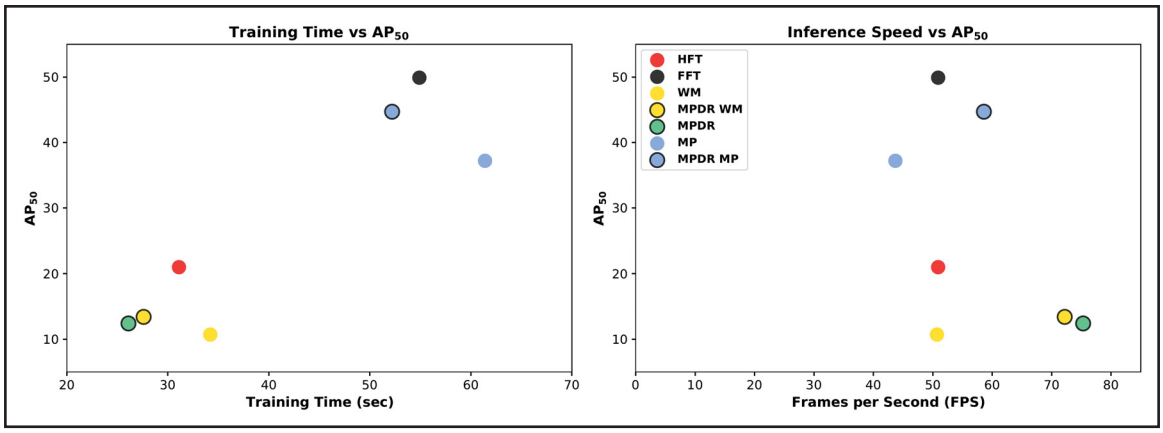


Figure 5.6 Comparison of Training Time and Inference Speed vs. Detection Performance on NYU_{v2}. Left: shows training time per epoch (in seconds) vs. AP_{50}. Right: presents inference time (FPS) vs. AP_{50}. Each point is a method, with HFT, FFT, WM, Modality Prompt (MP), and MP with Decoupled Residuals (MPDR) variants shown in distinct colors. Higher values in both FPS and AP_{50} indicate better runtime and accuracy trade-offs

5.5 Conclusion

In this work, we presented ModPrompt, a visual prompt strategy that adapts open-vocabulary object detectors to new modalities (e.g., infrared, depth, event-based, and LiDAR) while preserving zero-shot capabilities. Our encoder-decoder design yields strong performance across challenging benchmarks. We also introduced MPDR, a decoupled text-prompt mechanism that retains language knowledge during adaptation and can match full fine-tuning performance. ModPrompt outperforms prior pixel-level prompts and generalizes across both CNN and transformer backbones, offering a robust and efficient solution for modality-adaptive object detection. By expanding the potential of vision-language models in different applications, ModPrompt offers a compelling solution for future advancements in adaptable object detection frameworks, especially in settings that require full knowledge preservation.

Acknowledgments. This work was supported by Distech Controls Inc., the Natural Sciences and Engineering Research Council of Canada, the Digital Research Alliance of Canada, and MITACS.

CONCLUSION AND RECOMMENDATIONS

This thesis investigated the challenge of adapting deep object detectors to new sensing modalities, addressing the practical and scientific limitations of RGB-centric models in multimodal environments. Building on the motivation of enabling robust perception under diverse sensing conditions, we proposed contributions that progressively explore adaptation from data-level to model-level strategies. Through four complementary contributions—MiPa, HalluciDet, ModTr, and ModPrompt—we demonstrated that it is possible to achieve robust cross-modality generalization without retraining, large paired datasets, or architecture modification, ultimately bridging the gap between pre-trained RGB-based detectors and real-world multimodal perception systems.

Each contribution was designed to address a specific gap in the adaptation pipeline. MiPa (Mixed Patches) introduced a modality-agnostic training strategy for transformer-based detectors, using patch-level mixing and a dedicated alignment loss to enable robust cross-modal generalization from RGB and IR data—without requiring access to both modalities during inference. HalluciDet proposed a hallucination-based translation framework guided by detection losses, leveraging privileged information from RGB detectors to synthesize representations that enhance IR object detection. ModTr advanced this direction by jointly optimizing translation and detection objectives in an end-to-end framework, improving semantic consistency and generalization. Finally, ModPrompt extended modality adaptation to open-vocabulary vision-language models, introducing a conditioned visual prompt and residual embedded prompts to adapt models like YOLO-World and Grounding DINO without retraining or losing zero-shot capabilities.

Taken together, these contributions advance our understanding of modality adaptation as a modular, scalable, and detection-driven process. They show that adaptation can be achieved efficiently—through selective translation, representation regularization, or prompt-based conditioning—while preserving the detector's pre-trained knowledge. This unified view

highlights the feasibility of parameter-efficient adaptation as a practical path toward generalizable multimodal detection systems.

Beyond the individual methods, this thesis also contributes a conceptual foundation for future multimodal learning frameworks that can seamlessly adapt across sensors, environments, and downstream tasks. While this thesis lays a strong foundation for modality adaptation in object detection, several avenues remain open for exploration:

- **Generalization beyond IR and depth:** Future works could extend the proposed methods to more diverse modalities such as hyperspectral, LiDAR, or event-based sensors, each of which poses unique adaptation challenges.
- **Unified multimodal fusion:** Combining information from multiple modalities simultaneously, rather than adapting from one to another, offers a promising direction. Incorporating fusion strategies into detection-aware training could improve performance and robustness in multi-sensor environments.
- **Continual and online adaptation:** Developing systems that can continually adapt to evolving modalities or environments without catastrophic forgetting would increase real-world applicability. Techniques from continual learning and memory-aware optimization could be explored in this context.
- **Foundation models for multimodality:** The rise of large-scale vision-language and multimodal foundation models opens opportunities for universal detection backbones. Investigating how prompt-based and adapter-based techniques can plug into these models for modality-specific refinement is a promising direction.
- **Benchmarking and standardization:** As modality adaptation gains traction, there is a need for standardized evaluation protocols and datasets tailored to different sensors and use cases. Contributing to the design of such benchmarks would help align the research community.

In summary, this work moves object detection toward the next generation of adaptable and knowledge-preserving visual systems. The insights gained here open the door to future exploration of continual adaptation, foundation models, and unified multimodal perception architectures capable of operating robustly across any sensing modality in real-world conditions.

APPENDIX I

SUPPLEMENTARY MATERIAL FOR THE PAPER TITLED MIXED PATCH VISIBLE-INFRARED MODALITY AGNOSTIC OBJECT DETECTION

In this supplementary material, we provide additional information to reproduce our work. This supplementary material is divided into the following sections: Detailed diagrams (Section 1), Towards the optimal ρ (Section 2), Ablation on γ (Section 3) and MiPa on different detectors (Section 4).

1. Detailed diagrams

In this section, we provide additional diagrams aimed at enhancing the comprehension of both the baselines and our method in more detail. In Figure I-1, we show the simple strategy for constructing a multimodal model utilizing patches; this is our *Both* model in the main manuscript. First, the framework divides the images from both modalities (RGB and IR) into patches (yellow block). Subsequently, the extracted patches are fed into the backbone of the model (depicted in blue) and the head in pink.

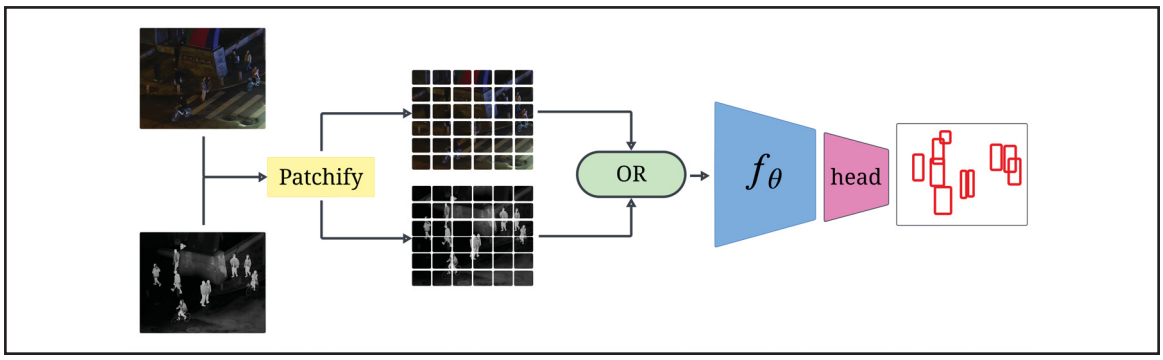


Figure-A I-1 Our *Both* baseline for multimodal object detection learning with patches.
The yellow block is the patchify function.
In green, we have the block representing one or the other patch modality to use.
In blue is the backbone, and in pink is the head of the detector

In Figure I-2, we present the proposed mix patches diagram. Similar to the previous diagram, we initially apply the patchify function (in yellow), followed by the mix patches function (in purple).

This function receives the patches and performs a mix patches operation, such as sampling the patches from both modalities according to a uniform distribution. Finally, the backbone is illustrated in blue, and the head in pink.

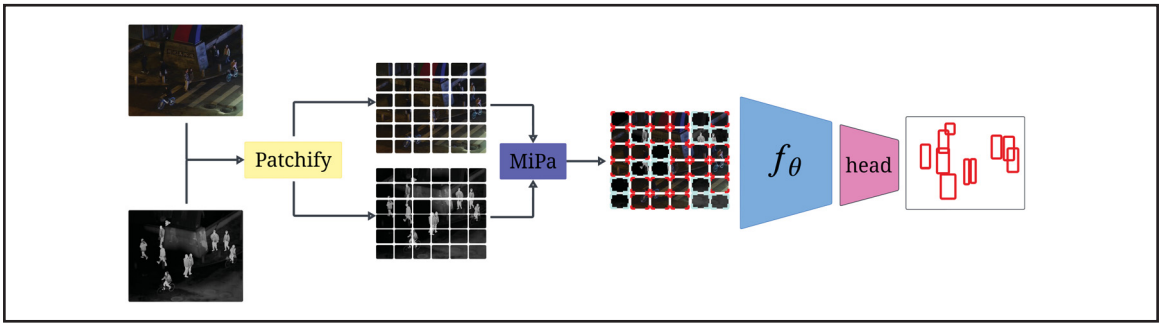


Figure-A I-2 Mix Patches diagram. First, in yellow, is the patchify function, which is responsible for providing the patches. Second, in purple, is the mix patches function, which is responsible for mixing the patches based on a pre-defined policy, e.g., uniform distribution of both modalities. Then, in blue is the backbone, and in pink is the detection head

Lastly, we provide an overview of an implementation of MiPa with DINO in Figure I-3. While the image is similar to the previous one, we offer additional visualizations showcasing the Swin backbone alongside the modality classifier. For the sake of simplicity and to emphasize the MiPa's modality classifier and the patchify/mix patches components, we omit the detection head in the figure.

2. Towards the optimal ρ

In this section, similar to the main manuscript, we provide the study of various strategies devised within this work to find the optimal approach to select the parameter ρ. This parameter represents the proportion of one modality, IR in our context, sampled during the training to facilitate optimal learning. As shown in Table I-1, the variable strategy yields the most favorable results in terms of providing the optimal ρ. This effectiveness is attributed to the inherent characteristics of MiPa to act as a regularizer for the weaker modality, which is the RGB in our setup. Thus, as described, the variable strategy is the method that reached the best average

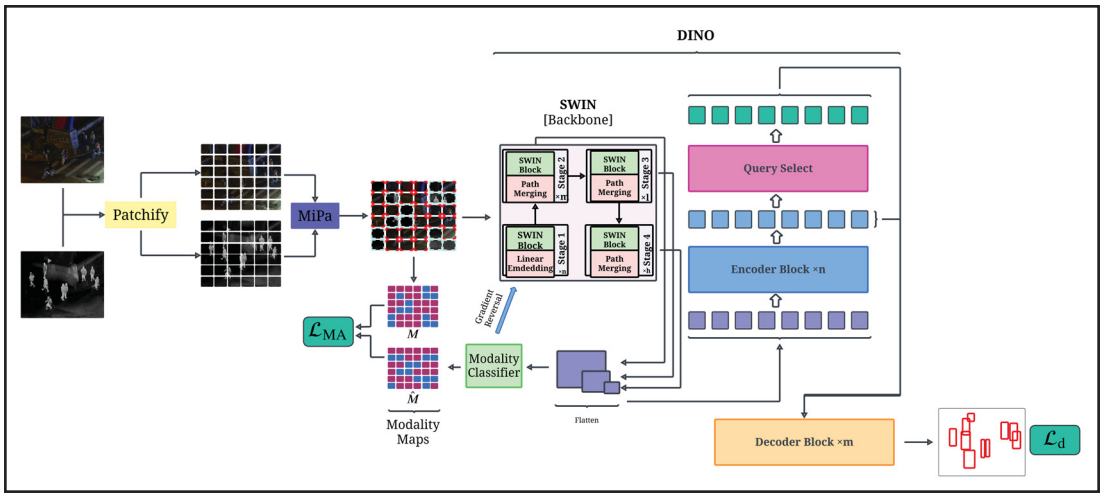


Figure-A I-3 MiPa with DINO. First, in yellow, is the patchify function, which is responsible for providing the patches. Second, bold purple is the mixing patches function, which is responsible for mixing the patches based on a pre-defined policy, e.g., uniform distribution of both modalities. Then, we have the DINO alongside the modality classifier head for the GRL (MA module)

across all the different APs. For example, the variable strategy was able to reach 88.5 AP_{50} in RGB, outperforming other strategies. Although its performance in IR was slightly lower than that of the Fixed strategy [ρ = 0.25] (achieving 97.5 AP_{50}), the variable strategy's overall mean performance was superior with 93.00 AP_{50}. This trend is similar to the other AP metrics, in which the RGB was improved, and the mean performance was better with the variable strategy.

Table-A I-1 Comparison of different ratio ρ sampling methods on LLVIP. Using DINO with Swin backbone

Model	Dataset: LLVIP								
	AP_{50}			AP_{75}			AP		
	RGB	IR	AVG.	RGB	IR	AVG.	RGB	IR	AVG.
Fixed [ρ=0.25]	78.9	**98.2**	88.55	41.5	78.1	59.80	42.5	66.5	54.50
Fixed [ρ=0.50]	73.0	97.6	85.30	31.1	78.1	54.60	36.0	67.0	51.50
Fixed [ρ=0.75]	77.4	97.5	87.45	40.5	76.5	58.50	42.0	65.2	53.60
Curriculum (ρ=0.25 for 4 epochs; then variable)	76.6	97.8	87.20	38.0	77.0	57.50	40.7	65.7	53.20
Curriculum (ρ=0.25 for 8 epochs; then variable)	80.1	97.8	88.95	40.9	**79.1**	60.00	43.0	**67.6**	55.30
Variable	**88.5**	97.5	**93.00**	**48.9**	77.4	**63.15**	**48.9**	66.6	**57.75**

3. Ablation on γ

In this section, we expand our comparison for different γ, in which we provide the full study on different AP metrics. The parameter γ governs the rate at which the modality invariance loss influences training. Thus, for FLIR, the best γ value was 0.05. As shown in the Table I-2, we study various values of γ with steps of 0.05, selected following the GRL equation (MA module) described in our manuscript and inspired by previous works Ganin & Lempitsky (2015). In this study, the values vary between 0.05 and 0.40, but the values may vary depending on the necessary number of epochs for training, as this function is step-dependent during training. Models that require more epochs may have larger values for γ. On FLIR, MiPa [$\gamma = 0.05$] was able to outperform the other baselines with an average of 67.62 AP_{50}, which is an increase from normal MiPa with 66.52 and the best baseline with 64.72 (Both [$\rho = 0.50$]). Moreover, MiPa [$\gamma = 0.05$] reached 29.77 in terms of AP_{75}, which is an average increase from 29.25 of normal MiPa, and 27.45 from the best baseline (Both [$\rho = 0.75$]). Note that for such a case, Both [$\rho = 0.75$] was better in terms of localization (AP_{75}) in comparison with Both [$\rho = 0.50$], even though it is worse than normal MiPa and MiPa with modality agnostic layer. Finally, in terms of AP, the trend is similar, so on average, we outperform all baselines and normal MiPa, which means that we are better in terms of localization and classification in each modality simultaneously. Thus, in this section, our goal of reaching a better balance between modalities while creating a robust model is successfully achieved.

Table-A I-2 Comparison of detection performance over different baselines and MiPa for DINO with Swin. The evaluation is done for RGB, IR, and the average of the modalities

Model	Backbone	Modality	Test Set (Dataset: FLIR)								
			RGB			IR			Average		
			AP_{50} ↑	AP_{75} ↑	AP ↑	AP_{50} ↑	AP_{75} ↑	AP ↑	AP_{50} ↑	AP_{75} ↑	AP ↑
DINO	Swin	RGB	66.07 ± 0.98	27.97 ± 0.22	32.33 ± 0.47	56.60 ± 0.80	20.87 ± 0.56	26.30 ± 0.19	61.33	24.42	29.32
		IR	56.47 ± 0.79	17.00 ± 0.98	24.30 ± 0.69	70.40 ± 0.38	38.80 ± 0.66	38.97 ± 0.31	63.43	27.90	31.63
		Both [$\rho = 0.25$]	56.53 ± 0.76	18.33 ± 0.55	25.60 ± 0.33	67.57 ± 1.73	31.33 ± 2.10	34.87 ± 1.35	62.05	24.83	30.23
		Both [$\rho = 0.50$]	60.50 ± 0.66	19.60 ± 1.29	27.37 ± 0.58	68.93 ± 0.60	33.03 ± 1.32	35.90 ± 0.82	64.72	26.32	31.63
		Both [$\rho = 0.75$]	58.53 ± 0.92	19.40 ± 0.83	26.47 ± 0.75	70.43 ± 0.65	35.50 ± 1.23	37.53 ± 0.41	64.48	27.45	32.00
		MiPa	63.53 ± 1.94	22.33 ± 0.82	29.47 ± 0.92	69.50 ± 1.84	36.17 ± 0.46	37.57 ± 0.67	66.52	29.25	33.52
		MiPa [$\gamma = 0.05$]	64.80 ± 2.30	24.77 ± 1.05	30.60 ± 0.62	70.43 ± 0.53	34.77 ± 1.18	37.50 ± 0.43	**67.62**	**29.77**	**34.05**
		MiPa [$\gamma = 0.10$]	64.03 ± 2.11	24.10 ± 1.63	30.63 ± 1.22	69.63 ± 1.45	33.13 ± 1.95	36.80 ± 1.39	66.83	28.62	33.72
		MiPa [$\gamma = 0.15$]	64.27 ± 0.47	24.40 ± 0.93	30.07 ± 0.68	69.93 ± 1.02	33.83 ± 1.24	36.80 ± 0.86	67.10	29.12	33.43
		MiPa [$\gamma = 0.20$]	61.83 ± 1.39	22.83 ± 1.01	28.53 ± 0.76	69.27 ± 1.57	31.87 ± 2.02	35.73 ± 1.31	65.55	27.35	32.13
		MiPa [$\gamma = 0.30$]	62.20 ± 2.49	22.93 ± 1.35	29.10 ± 1.28	67.47 ± 2.04	32.53 ± 0.66	35.87 ± 0.69	64.83	27.73	32.48
		MiPa [$\gamma = 0.40$]	61.13 ± 2.88	22.30 ± 0.57	28.50 ± 0.99	67.93 ± 0.92	32.47 ± 0.48	35.87 ± 0.49	64.53	27.38	32.18

4. MiPa on different detectors

In this section, we present additional quantitative results, including various performance metrics measured in terms of different APs. In Table I-3, we outline the results obtained using the Swin backbone for DINO and Deformable DETR across baselines, MiPa, and MiPa with a modality invariance layer. As shown, MiPa demonstrates superior performance compared to using both modalities jointly and other baselines across different datasets.

Table-A I-3 Comparison of detection performance over different baselines and MiPa for DINO and Deformable DETR. The evaluation is done for RGB, IR, and the average of the modalities

Model	Backbone	Modality	Dataset: LLVIP								
			RGB			IR			Average		
			AP_{50} ↑	AP_{75} ↑	AP ↑	AP_{50} ↑	AP_{75} ↑	AP ↑	AP_{50} ↑	AP_{75} ↑	AP ↑
DINO	Swin	RGB	90.87 ± 0.84	54.20 ± 1.02	51.87 ± 0.79	94.23 ± 0.57	67.13 ± 0.85	59.43 ± 0.48	92.55	60.67	55.65
		IR	66.87 ± 0.90	20.27 ± 0.98	29.03 ± 0.76	96.87 ± 0.12	73.53 ± 0.40	64.27 ± 0.12	81.87	46.90	46.65
		Both [ρ = 0.25]	79.73 ± 1.03	45.70 ± 0.43	44.97 ± 0.33	97.40 ± 0.22	76.03 ± 0.83	65.87 ± 0.45	88.57	60.87	55.42
		Both [ρ = 0.50]	82.40 ± 1.50	47.27 ± 1.65	46.43 ± 1.03	96.50 ± 0.29	74.17 ± 2.10	64.83 ± 0.96	89.45	60.72	55.63
		Both [ρ = 0.75]	81.23 ± 2.89	45.60 ± 2.49	45.23 ± 2.13	97.07 ± 0.25	74.73 ± 1.41	65.27 ± 0.82	89.15	60.17	55.25
		MiPa (Ours)	88.70 ± 0.45	46.67 ± 0.86	48.00 ± 0.28	96.97 ± 0.26	73.07 ± 1.42	64.30 ± 1.10	**92.83**	59.87	**56.15**
		MiPa + MA (Ours)	89.10 ± 0.28	46.60 ± 0.86	48.10 ± 0.33	96.83 ± 0.09	71.17 ± 0.70	63.17 ± 0.58	**92.97**	58.88	55.63
Def.DETR	Swin	RGB	80.00 ± 1.50	35.50 ± 0.22	40.27 ± 0.41	90.03 ± 0.87	50.37 ± 0.85	49.67 ± 0.48	85.02	42.93	**44.97**
		IR	56.10 ± 2.50	10.77 ± 1.47	21.10 ± 1.34	94.20 ± 0.08	62.20 ± 0.86	56.73 ± 0.47	75.15	36.48	38.92
		Both [ρ = 0.25]	51.20 ± 3.47	22.57 ± 1.96	25.70 ± 1.91	83.73 ± 16.57	54.17 ± 16.62	48.30 ± 14.93	67.47	38.37	37.00
		Both [ρ = 0.50]	53.57 ± 4.17	23.13 ± 2.15	26.57 ± 2.11	83.87 ± 16.17	52.67 ± 17.17	49.37 ± 12.64	68.72	37.90	37.97
		Both [ρ = 0.75]	53.53 ± 4.55	22.83 ± 2.72	26.5 ± 2.63	82.33 ± 18.48	51.33 ± 18.56	48.13 ± 14.03	67.93	37.08	37.32
		MiPa (Ours)	78.60 ± 0.42	23.33 ± 5.85	29.20 ± 6.37	95.20 ± 0.16	62.60 ± 0.78	56.80 ± 0.45	**86.90**	**42.97**	43.00
		MiPa + MA (Ours)	79.02 ± 0.21	24.36 ± 2.85	31.25 ± 4.32	95.36 ± 0.25	63.38 ± 0.43	57.25 ± 0.43	**87.19**	**43.87**	44.25
Model	**Backbone**	**Modality**	**Dataset: FLIR**								
			RGB			IR			Average		
			AP_{50} ↑	AP_{75} ↑	AP ↑	AP_{50} ↑	AP_{75} ↑	AP ↑	AP_{50} ↑	AP_{75} ↑	AP ↑
DINO	Swin	RGB	66.07 ± 0.98	27.97 ± 0.22	32.33 ± 0.47	56.60 ± 0.80	20.87 ± 0.56	26.30 ± 0.19	61.33	24.42	29.32
		IR	56.47 ± 0.79	17.00 ± 0.98	24.30 ± 0.69	70.40 ± 0.38	38.80 ± 0.66	38.97 ± 0.31	63.43	27.90	31.63
		Both [ρ = 0.25]	56.53 ± 0.76	18.33 ± 0.55	25.60 ± 0.33	67.57 ± 1.73	31.33 ± 2.10	34.87 ± 1.35	62.05	24.83	30.23
		Both [ρ = 0.50]	60.50 ± 0.66	19.60 ± 1.29	27.37 ± 0.58	68.93 ± 0.60	33.03 ± 1.32	35.90 ± 0.82	64.72	26.32	31.63
		Both [ρ = 0.75]	58.53 ± 0.92	19.40 ± 0.83	26.47 ± 0.75	70.43 ± 0.65	35.50 ± 1.23	37.53 ± 0.41	64.48	27.45	32.00
		MiPa (Ours)	63.53 ± 1.94	22.33 ± 0.82	29.47 ± 0.92	69.50 ± 1.84	36.17 ± 0.46	37.57 ± 0.67	**66.52**	**29.25**	**33.52**
		MiPa + MA (Ours)	64.80 ± 2.30	24.77 ± 1.05	30.60 ± 0.62	70.43 ± 0.53	34.77 ± 1.18	37.50 ± 0.43	**67.62**	**29.77**	**34.05**
Def.DETR	Swin	RGB	49.33 ± 1.39	13.93 ± 0.30	20.97 ± 0.53	43.77 ± 0.56	10.13 ± 0.08	17.37 ± 0.19	46.55	12.03	19.17
		IR	39.17 ± 1.48	08.57 ± 0.24	14.90 ± 0.50	59.20 ± 0.29	20.03 ± 0.33	26.93 ± 0.62	49.18	14.30	20.92
		Both [ρ = 0.25]	35.73 ± 4.95	08.27 ± 1.51	14.00 ± 2.38	43.00 ± 13.54	14.30 ± 5.97	19.23 ± 7.01	39.37	11.28	16.62
		Both [ρ = 0.50]	33.93 ± 5.15	08.23 ± 1.43	13.60 ± 2.17	43.33 ± 14.14	14.70 ± 6.34	19.63 ± 7.43	38.63	11.47	16.62
		Both [ρ = 0.75]	32.90 ± 3.54	07.70 ± 1.20	12.97 ± 1.65	44.13 ± 14.85	14.17 ± 6.30	19.47 ± 7.37	38.52	10.93	16.22
		MiPa (Ours)	48.00 ± 0.57	15.23 ± 0.69	20.70 ± 0.45	54.97 ± 0.90	19.80 ± 0.28	25.50 ± 0.42	**51.48**	**17.52**	**23.10**
		MiPa + MA (Ours)	48.27 ± 1.76	14.57 ± 1.05	20.63 ± 0.96	55.80 ± 0.22	21.00 ± 0.67	26.33 ± 0.39	**52.03**	**17.78**	**23.48**

APPENDIX II

SUPPLEMENTARY MATERIAL FOR THE PAPER TITLED HALLUCIDET: HALLUCINATING RGB MODALITY FOR PERSON DETECTION THROUGH PRIVILEGED INFORMATION

In this supplementary material, we provide additional information to reproduce our work. The source code is publicly available: https://github.com/heitorrapela/HalluciDet. Here, we provide ablation on the hyperparameter of the HalluciDet loss, qualitative examples of the obtained detections, and additional results.

1. Ablation of hyperparameters λ for HalluciDet

In this section, we show the sensitivity of HalluciDet to the hyperparameters during training. For these experiments, as we did not want to have the influence of data augmentation on the pipeline, we removed the data augmentations that could benefit the starting detector and also the HalluciDet. Thus, the results in the main manuscript are the results with the detector using transformations such as color jitter and horizontal flip, and the same transformations were used for training the HalluciDet and respective baselines. In this ablation, we focused on the balancing of λ, so for this case, we kept both detectors and HalluciDet without data augmentations.

The cost function of the hallucination network $\mathcal{L}_{hall}$ (Equation A II-1) contains three terms: regression loss, classification loss, and other losses. Here, the other loss terms are dependent on the detection method used, e.g., for Faster-RCNN $\mathcal{L}_{*} = \mathcal{L}_{rpn} + \mathcal{L}_{obj}$, where the regression loss $\mathcal{L}_{rpn}$ is applied to the region proposal network, and $\mathcal{L}_{obj}$ is the object/background classification loss.

$$\mathcal{L}_{\text{hall}} = \lambda_{cls} \cdot \mathcal{L}_{\text{cls}} + \lambda_{reg} \cdot \mathcal{L}_{\text{reg}} + \lambda_{*} \cdot \mathcal{L}_{*} \tag{A II-1}$$

As shown on Table II-1 of this supplementary materials, the HalluciDet ablation study was divided into different ways of balancing the regression and classification parts of the loss. In practice, it is better to use both components (regression and classification), but we recommend prioritizing the regression part for optimal balance.

Table-A II-1 Comparison between different weights on the losses terms. In this table, the models are started frozen from RGB, the same as reported in the paper. Then, the hallucination network is trained with different lambda values to see its impacts on the model's performance.
Results over LLVIP test set

Method	**Ablation (Loss Weight)**	**AP@0.5↑** Test Set (Dataset: LLVIP) χ^{IR}
HalluciDet (RetinaNet)	$\lambda_{cls} = 0.0, \lambda_{reg} = 1.0$	65.81
	$\lambda_{cls} = 1.0, \lambda_{reg} = 0.0$	60.77
	$\lambda_{cls} = 0.01, \lambda_{reg} = 0.1$	68.58
	$\lambda_{cls} = 0.1, \lambda_{reg} = 0.01$	60.03
HalluciDet (FCOS)	$\lambda_{cls} = 0.0, \lambda_{reg} = 1.0, \lambda_{box_{cnt}} = 1.0$	63.01
	$\lambda_{cls} = 1.0, \lambda_{reg} = 0.0, \lambda_{box_{cnt}} = 0.0$	60.92
	$\lambda_{cls} = 0.01, \lambda_{reg} = 0.1, \lambda_{box_{cnt}} = 0.1$	65.02
	$\lambda_{cls} = 0.1, \lambda_{reg} = 0.01, \lambda_{box_{cnt}} = 0.01$	64.59
HalluciDet (Faster R-CNN)	$\lambda_{cls} = 0.1, \lambda_{obj} = 0.1, \lambda_{reg} = 0.01, \lambda_{RPNbox_{reg}} = 0.01$	85.35
	$\lambda_{cls} = 0.01, \lambda_{obj} = 0.01, \lambda_{reg} = 0.1, \lambda_{RPNbox_{reg}} = 0.1$	88.72
	$\lambda_{cls} = 1.0, \lambda_{obj} = 1.0, \lambda_{reg} = 0.0, \lambda_{RPNbox_{reg}} = 0.0$	83.97
	$\lambda_{cls} = 0.0, \lambda_{obj} = 0.0, \lambda_{reg} = 1.0, \lambda_{RPNbox_{reg}} = 1.0$	84.08

2. Hallucidet and additional results on FLIR

Similar to the main manuscript, we added additional ablations with respect to the FLIR dataset.

2.0.0.1 Hallucidet with a different encoder.

Similar to the main manuscript, we provided a study on the different backbones of the hallucination network encoder but focused on the FLIR dataset, as shown in Table II-2. The results show a similar trend, in which models with more capacity in terms of parameters can learn more robust representations for the test set distribution, thus increasing the AP@50.

Table-A II-2 Comparison of the number of parameters for different Hallucination Network backbones vs. AP@50 on the FLIR dataset with the Faster R-CNN detector

Method		Params.	AP@50↑
Faster R-CNN		41.3 M	61.48
HalluciDet	MobileNet_{v3s}	+ 3.1 M	53.62
	MobileNet_{v2}	+ 6.6 M	67.74
	ResNet_{18}	+ 14.3 M	68.56
	ResNet_{34}	+ 24.4 M	71.58

3. Qualitative analysis of Hallucidet Detections

In this section, we provided an additional sequence of batch images, similar to the main manuscript. Here, we can find more than one batch of 8 images for the LLVIP dataset (Figure II-1), and then two batches of 8 images each for the FLIR dataset (Figure II-2, Figure II-3). Thus, the trend and explanations for detections remain the same as those described in the main manuscript.

Processing time comparison: In terms of trade between more parameters that can increase the speed for processing and the performance of the detection, we highlight some important discussion about it. For the models classified as nonlearning methods, there is no increase in the inference speed and the training part, but they have lower detection performance. For the models that are learning in the input space, such as image translation methods like CycleGAN or FastCUT, given the same backbone network, HalluciDet has faster training and equal inference time to the deep learning baselines, and we can improve the detection performance.

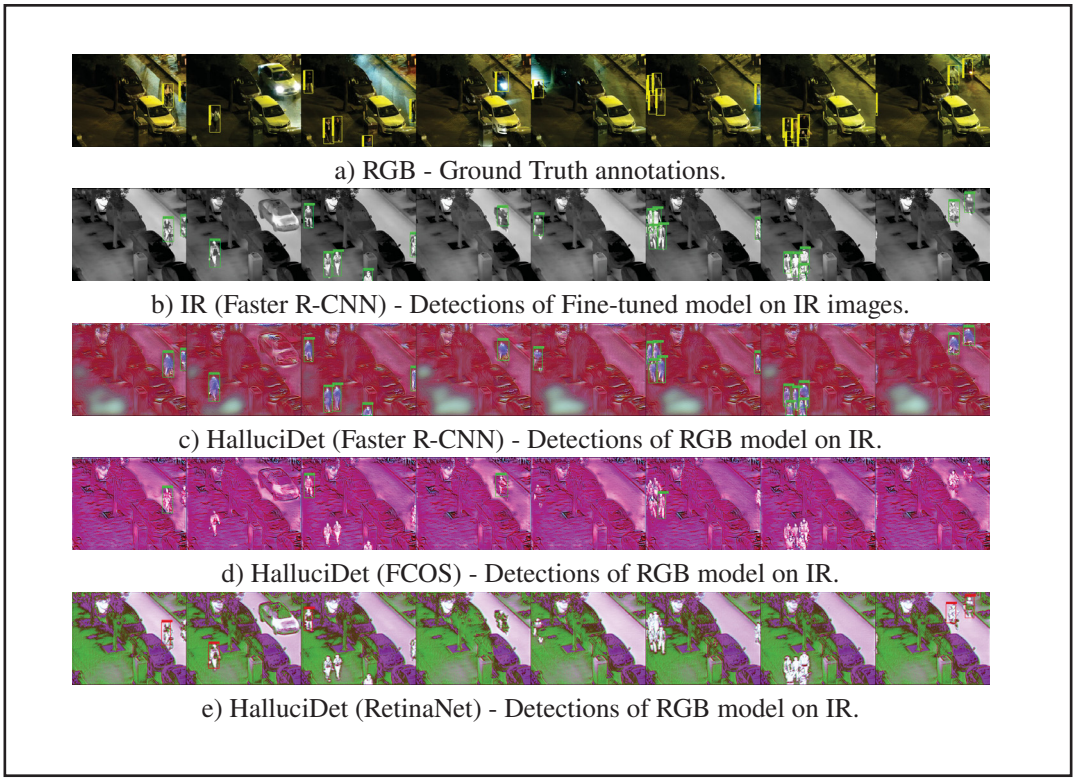
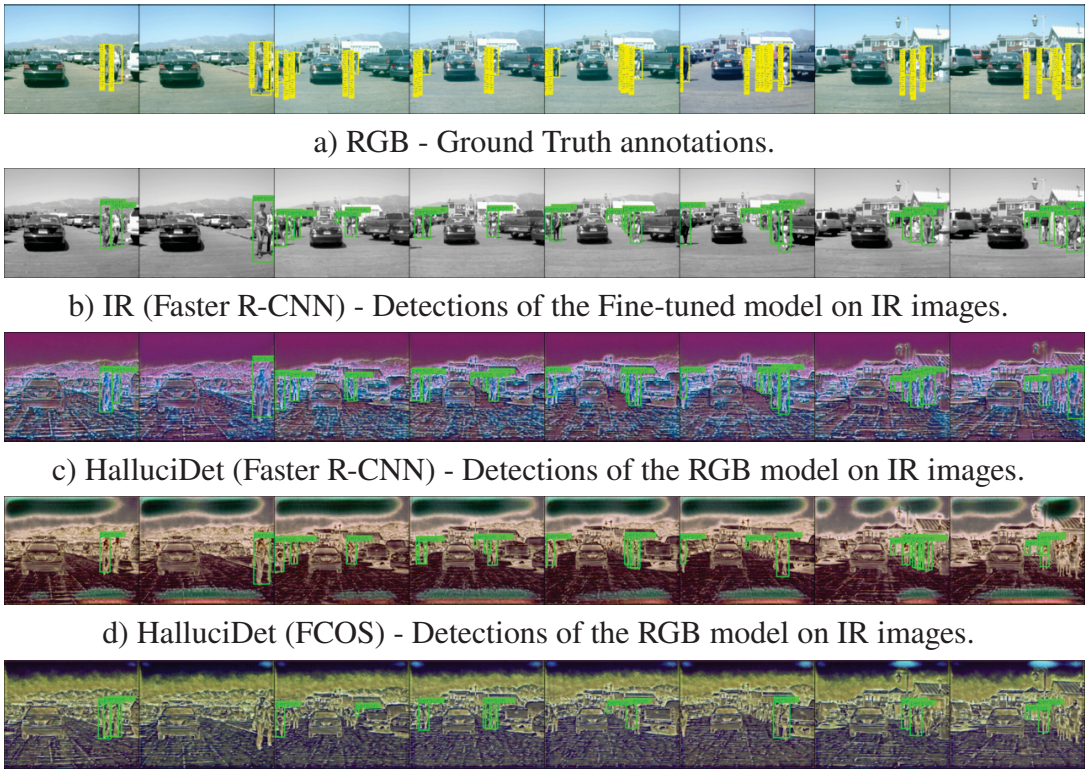


Figure-A II-1 Illustration of a sequence of 8 images of the LLVIP dataset. The first row is the RGB modality, then the IR modality, followed by different representations created by HalluciDet over various detectors



a) RGB - Ground Truth annotations.

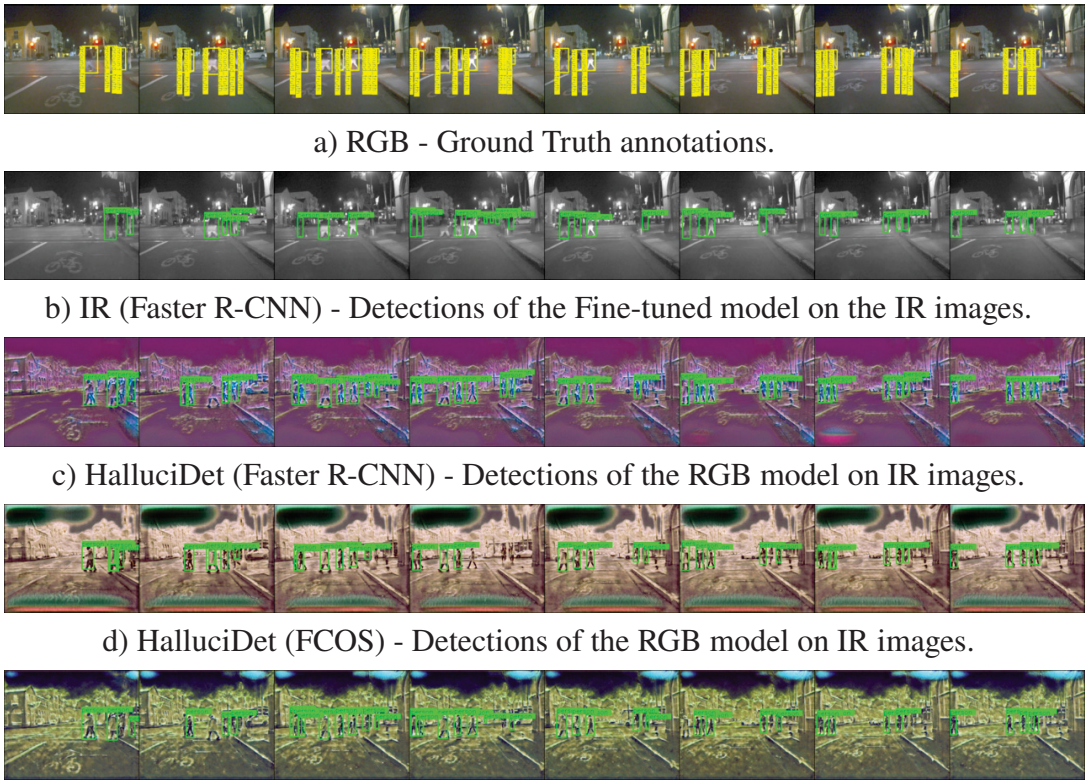
b) IR (Faster R-CNN) - Detections of the Fine-tuned model on IR images.

c) HalluciDet (Faster R-CNN) - Detections of the RGB model on IR images.

d) HalluciDet (FCOS) - Detections of the RGB model on IR images.

e) HalluciDet (RetinaNet) - Detections of the RGB model on IR images.

Figure-A II-2 Illustration of a sequence of 8 images of the FLIR dataset. The first row is the RGB modality, then the IR modality, followed by different representations created by HalluciDet over various detectors



a) RGB - Ground Truth annotations.

b) IR (Faster R-CNN) - Detections of the Fine-tuned model on the IR images.

c) HalluciDet (Faster R-CNN) - Detections of the RGB model on IR images.

d) HalluciDet (FCOS) - Detections of the RGB model on IR images.

e) HalluciDet (RetinaNet) - Detections of the RGB model on IR images.

Figure-A II-3 Another sequence of 8 images of the FLIR dataset. The first row is the RGB modality, then the IR modality, followed by different representations created by HalluciDet over various detectors

APPENDIX III

SUPPLEMENTARY MATERIAL FOR THE PAPER TITLED MODALITY TRANSLATION FOR OBJECT DETECTION ADAPTATION WITHOUT FORGETTING PRIOR KNOWLEDGE

In this supplementary material, we provide additional information to reproduce our work. The code is available at https://github.com/heitorrapela/ModTr. This supplementary material is divided into the following sections: Detailed diagrams (Section 1), Different fusion strategies (Section 2), Quantitative Results (Section 3), in which we provide additional numerical results in terms of APs, Qualitative Results (Section 4), in which we provide additional visualizations and Additional Modality: Canny Edges (Section 5), in which we show experiments with edges extracted from IR images using Canny edge detector.

1. Detailed Diagrams

In this section, we expand and detail the diagrams provided in the main manuscript. In Figure III-1, we describe the traditional approach of employing specialized detectors for individual modalities. For example, we depict an RGB detector (highlighted in purple) and two IR detectors (highlighted in green and yellow), with each trained over a different dataset.

In Figure III-2, we illustrate the proposed ModTr. This method involves using a single pre-trained detector model typically trained on the more prevalent data, i.e., RGB, and additional input adaptation network. For clarity, the RGB modality (along with the RGB detector) is depicted in purple, while an adaptation block of ModTr is shown in green for one IR modality and in red for the other IR modality with different distribution.

Lastly, we present a final diagram (Figure III-3), depicting a detector trained on the joint distribution of all modalities. The detector, shown in purple, undergoes training with all available modalities. While this approach enables the model to learn shared features, it may not be optimal. Nevertheless, it incurs a lower memory cost compared to employing one detector for each modality.

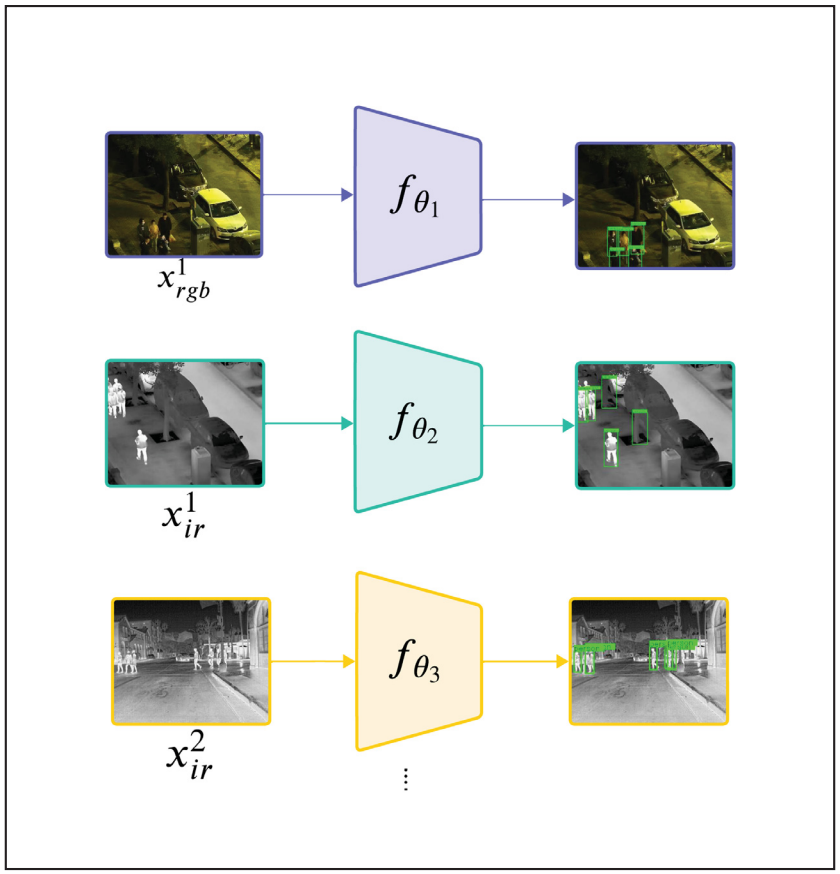


Figure-A III-1 The simplest approach is to use a different detector adapted to each modality. This can lead to a high level of accuracy but requires storing models in memory multiple times. In purple is the RGB detector, in green is one IR detector for one dataset, and in yellow is another detector for another IR dataset

2. Different fusion strategies

Here, we present a comprehensive overview of the alternative fusion strategies we explored. In addition to the element-wise product fusion, which produced the best results for our study and was reported in the main manuscript, we also tested strategies **ModTr**$_{+}$, and **ModTr**$_{\oplus}$, detailed below:

ModTr$_{+}$: The addition mechanism involves forwarding the input modality and summing it with the output of the translation network, as detailed in Equation A III-1. This residual connection

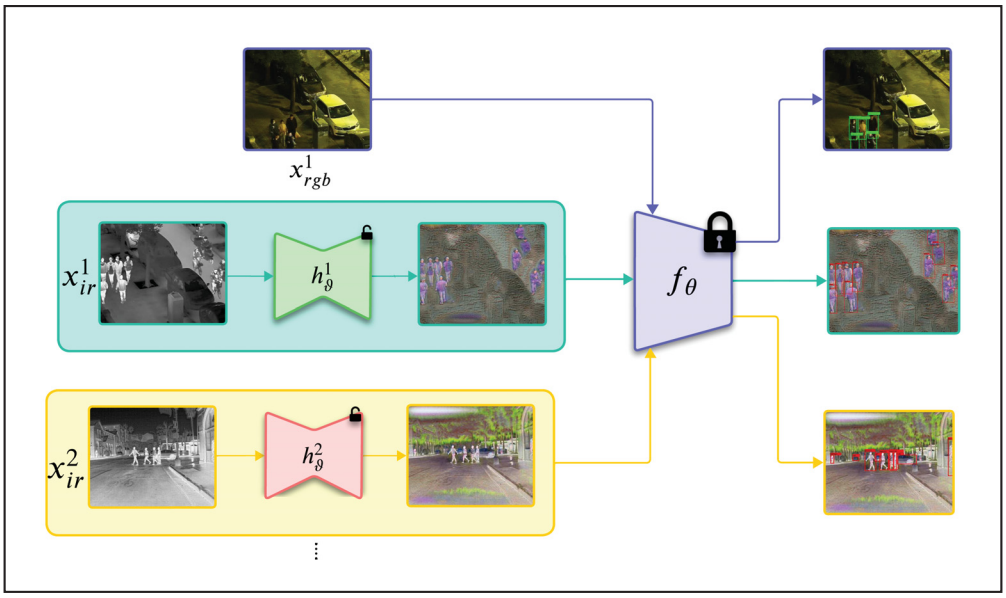


Figure-A III-2 Our proposed solution is based on using a single pre-trained detector model normally trained on the more abundant data (RGB) and then adapting the input through our ModTr block

of the input serves as regularization for the model, aiding the network in learning the missing information necessary for detector operation. This operator learns the new representation by amplifying pixel values when the weights of the translation representation tend toward 1, or preserving the original information when they tend toward 0. Such a range of values is due to our modification on the U-Net (Ronneberger *et al.*, 2015) to generate, in which we changed the last layer to a Sigmoid function layer so we can better control the generated image to be closer to a real RGB-like image:

$$\mathcal{L}_{\text{ModTr}_+}(x, Y; \vartheta) = \mathcal{L}_{det}(f_\theta \left(h^d_\vartheta(x) + x \right), Y). \tag{A III-1}$$

ModTr$_\oplus$: The subsequent fusion mechanism draws inspiration from DenseFuse Li & Wu (2018), which employs the relative importance of pixels as an attention mechanism for both the translation network's output and input. This attention mechanism operates by providing a

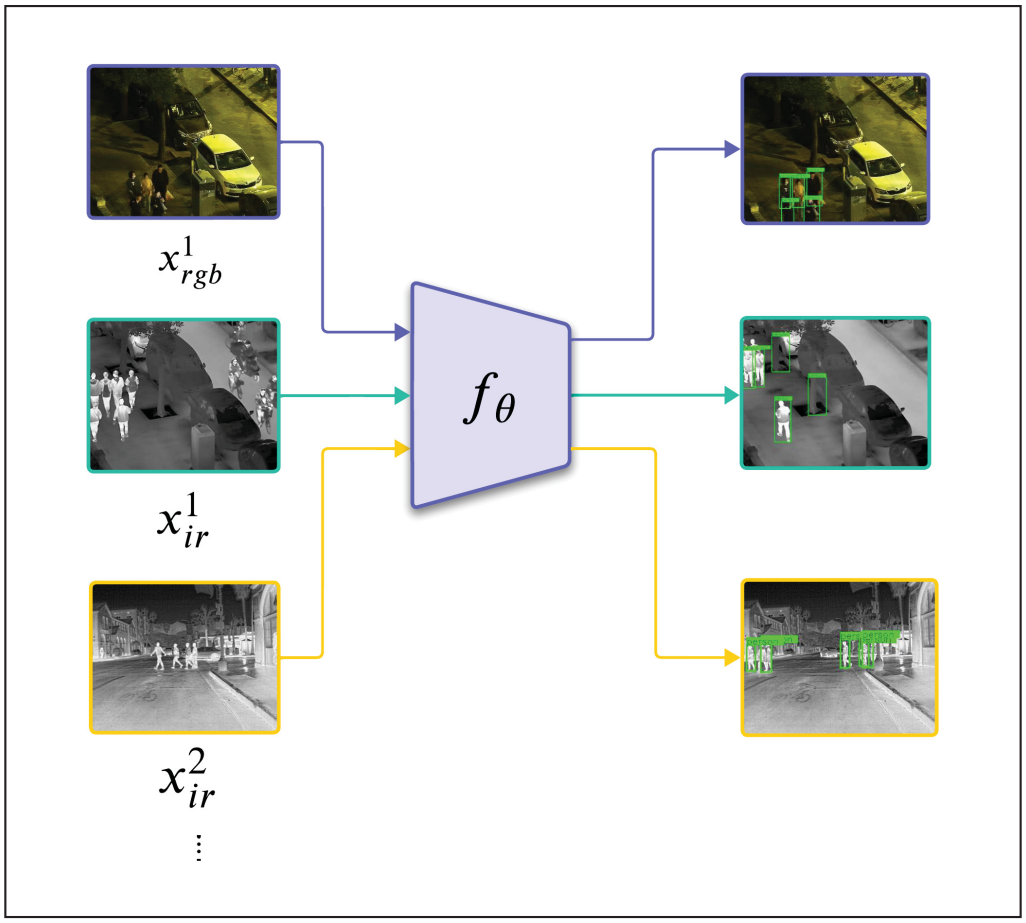


Figure-A III-3 A single detector is trained on all modalities jointly. This allows the use of a single model but requires access to all modalities jointly, which is often not possible, especially when dealing with large pre-trained models

weighted average of the channels. The implementation details for such an operator can be found on Li & Wu (2018). Then, the proposed loss function is given by the following Equation A III-2:

$$\mathcal{L}_{\text{ModTr}_\oplus}(x, Y; \vartheta) = \mathcal{L}_{det}(f_\theta\left(\Phi(h^d_\vartheta(x), x)\right), Y), \tag{A III-2}$$

with

$$\Phi(x, \hat{x}) = \frac{x \odot e^x + \hat{x} \odot e^{\hat{x}}}{e^x + e^{\hat{x}}}.$$

3. Quantitative Results

In this section, we provide further details for the different AP metrics of the experiments in the main manuscript. In Table III-1, we compare the best ModTr with different image translation strategies. Some of the competitors use RGB data for the translation, while others utilize bounding box annotations. Here, we can see that ModTr outperforms the competitors in terms of AP metrics across different detectors, demonstrating its superiority in terms of classification and localization. For instance, when compared with HalluciDet (Medeiros *et al.*, 2024c) using FCOS, which was the second best, our method shows an improvement of approximately 28 better in terms of AP_{50}, with a similar trend observed for RetinaNet on the LLVIP dataset. The gap is narrower with Faster R-CNN, with an improvement of only around 2 AP_{50}. While the gap is smaller for the FLIR dataset, it remains consistent across all APs and detectors; for example, an improvement of 6 AP_{50} for Faster R-CNN, approximately 14 AP_{50} for RetinaNet, and 11 AP_{50} for FCOS. In terms of localization for the FLIR dataset, significant improvements are observed, such as around 7 AP for Faster R-CNN, 11 AP for RetinaNet, and 11 AP for FCOS when compared with HalluciDet. For the other methods, which do not rely on bounding boxes, substantial improvements are evident; for instance, our method exhibits an improvement of more than 11 AP over FastCUT (Park *et al.*, 2020a), with similar trends observed for other competitors.

In Table III-2, we show that compared with fine-tuning (FT), the ModTr is better even without modifying the parameters of the detector. Thus, it preserves the detector's knowledge for further tasks while improving performance. For instance, in terms of localization with AP_{75} and AP for FCOS, RetinaNet on LLVIP, we reached the performance of the FT, while with AP_{50}, we were comparable. For the FLIR dataset, we outperform the FT in all detectors for the different APs and also in all different fusion strategies.

In Table III-3, we explore the potential of achieving comparable performance in detection tasks while significantly reducing the number of parameters by employing a smaller backbone for the translation network. The MobileNet_{v2} with only 6.6 million additional parameters, achieves

Table-A III-1 Detection performance of ModTr versus baseline image-to-image methods to translate the IR to RGB-like images. Here, we used three different detectors (FCOS, RetinaNet, and Faster R-CNN). The methods were evaluated on IR test set of LLVIP and FLIR datasets. The RGB column indicates if the method required access to RGB images during training, and Box refers to the use of ground truth bboxes during training

Image translation	RGB	Box	Test Set IR (Dataset: LLVIP)								
			FCOS			RetinaNet			Faster R-CNN		
			AP_{50} ↑	AP_{75} ↑	AP ↑	AP_{50} ↑	AP_{75} ↑	AP ↑	AP_{50} ↑	AP_{75} ↑	AP ↑
Histogram Equal. (Herrmann *et al.*, 2018)			53.74 ± 0.00	32.57 ± 0.00	31.69 ± 0.00	59.93 ± 0.00	33.04 ± 0.00	33.16 ± 0.00	65.70 ± 0.04	39.02 ± 0.11	38.33 ± 0.02
CycleGAN (Zhu *et al.*, 2017b)	✓		41.72 ± 1.63	23.83 ± 0.78	23.85 ± 0.76	43.17 ± 1.52	22.34 ± 0.88	23.34 ± 0.53	45.44 ± 1.89	26.82 ± 1.59	26.54 ± 1.20
CUT (Park *et al.*, 2020b)	✓		26.48 ± 2.88	13.68 ± 2.72	14.30 ± 2.25	25.64 ± 3.77	11.74 ± 2.33	13.12 ± 2.07	27.96 ± 1.70	13.59 ± 2.77	14.78 ± 1.82
FastCUT (Park *et al.*, 2020b)	✓		34.92 ± 3.63	19.07 ± 1.33	19.39 ± 1.52	35.73 ± 2.53	16.36 ± 0.44	18.11 ± 0.79	42.09 ± 3.51	21.44 ± 1.57	22.91 ± 1.68
HalluciDet (Medeiros *et al.*, 2024c)	✓	✓	64.17 ± 0.61	18.80 ± 1.45	28.00 ± 0.92	60.38 ± 3.59	06.75 ± 1.38	19.95 ± 2.01	90.07 ± 0.72	51.23 ± 1.81	57.78 ± 0.97
$ModTr_{\odot}$ (ours)		✓	**92.04 ± 0.47**	**63.84 ± 0.93**	**57.63 ± 0.66**	**91.56 ± 0.64**	**59.49 ± 1.11**	**54.83 ± 0.61**	**91.82 ± 0.49**	**62.51 ± 0.87**	**57.97 ± 0.85**
Image translation	**RGB**	**Box**	**Test Set IR (Dataset: FLIR)**								
			FCOS			**RetinaNet**			**Faster R-CNN**		
			AP_{50} ↑	AP_{75} ↑	AP ↑	AP_{50} ↑	AP_{75} ↑	AP ↑	AP_{50} ↑	AP_{75} ↑	AP ↑
Histogram Equal. (Herrmann *et al.*, 2018)			52.09 ± 0.00	16.44 ± 0.00	22.76 ± 0.00	53.13 ± 0.00	16.50 ± 0.00	23.06 ± 0.00	56.50 ± 0.10	17.62 ± 0.04	24.61 ± 0.01
CycleGAN (Zhu *et al.*, 2017b)	✓		49.01 ± 1.28	21.16 ± 0.71	23.92 ± 0.97	49.04 ± 1.71	19.93 ± 0.54	23.71 ± 0.70	54.48 ± 2.09	23.08 ± 1.54	26.85 ± 1.23
CUT (Park *et al.*, 2020b)	✓		38.70 ± 1.05	14.85 ± 0.49	18.16 ± 0.75	39.08 ± 1.42	13.69 ± 0.61	17.84 ± 0.75	43.34 ± 1.53	16.09 ± 0.38	20.29 ± 0.48
FastCUT (Park *et al.*, 2020b)	✓		45.19 ± 4.46	22.93 ± 2.09	24.02 ± 2.37	43.04 ± 4.95	19.82 ± 2.78	22.00 ± 2.73	49.98 ± 4.57	25.52 ± 2.85	26.68 ± 2.59
HalluciDet (Medeiros *et al.*, 2024c)	✓	✓	54.20 ± 2.50	17.36 ± 2.23	23.74 ± 2.09	52.06 ± 1.47	16.21 ± 0.31	22.29 ± 0.45	63.11 ± 1.54	23.91 ± 1.10	29.91 ± 1.18
$ModTr_{\odot}$ (ours)		✓	**65.99 ± 0.78**	**33.73 ± 1.74**	**35.49 ± 0.94**	**66.31 ± 0.93**	**31.22 ± 0.69**	**34.27 ± 0.27**	**69.20 ± 0.36**	**34.58 ± 0.56**	**37.21 ± 0.46**

Table-A III-2 Detection performance (AP) of ModTr with different fusion strategies versus baseline fine-tuning (FT) of the detector. Here, we used three different detectors (FCOS, RetinaNet, and Faster R-CNN. The methods were evaluated on IR test set of LLVIP and FLIR datasets

Method	Test Set IR (Dataset: LLVIP)								
	FCOS			RetinaNet			Faster R-CNN		
	AP_{50} ↑	AP_{75} ↑	AP ↑	AP_{50} ↑	AP_{75} ↑	AP ↑	AP_{50} ↑	AP_{75} ↑	AP ↑
Fine-Tuning	**93.14 ± 1.06**	62.08 ± 3.37	57.37 ± 2.19	**93.61 ± 0.59**	56.20 ± 3.78	53.79 ± 1.79	**94.64 ± 0.84**	62.50 ± 0.79	**59.62 ± 1.23**
$ModTr_{+}$	90.57 ± 1.46	62.38 ± 0.31	56.44 ± 0.75	91.09 ± 0.73	55.06 ± 1.81	53.18 ± 1.03	91.89 ± 0.39	61.44 ± 0.72	57.14 ± 0.50
$ModTr_{\oplus}$	91.11 ± 0.84	62.69 ± 1.53	57.01 ± 0.71	90.49 ± 1.11	58.73 ± 0.55	54.43 ± 0.35	91.20 ± 0.46	61.31 ± 0.73	56.95 ± 0.37
$ModTr_{\odot}$	92.04 ± 0.47	**63.84 ± 0.93**	**57.63 ± 0.66**	91.56 ± 0.64	**59.49 ± 1.11**	**54.83 ± 0.61**	91.82 ± 0.49	**62.51 ± 0.87**	57.97 ± 0.85
Method	**Test Set IR (Dataset: FLIR)**								
	FCOS			**RetinaNet**			**Faster R-CNN**		
	AP_{50} ↑	AP_{75} ↑	AP ↑	AP_{50} ↑	AP_{75} ↑	AP ↑	AP_{50} ↑	AP_{75} ↑	AP ↑
Fine-Tuning	60.22 ± 0.97	21.94 ± 0.42	27.97 ± 0.59	61.77 ± 1.02	22.37 ± 0.45	28.46 ± 0.50	66.15 ± 0.94	24.48 ± 0.71	30.93 ± 0.46
$ModTr_{+}$	64.90 ± 0.48	32.78 ± 0.27	34.63 ± 0.24	65.30 ± 0.66	30.00 ± 0.81	33.70 ± 0.59	68.64 ± 0.77	34.96 ± 0.90	37.09 ± 0.74
$ModTr_{\oplus}$	**65.46 ± 0.61**	**33.21 ± 0.55**	34.94 ± 0.52	63.87 ± 0.51	**30.93 ± 0.38**	33.72 ± 0.22	68.64 ± 1.29	**35.48 ± 0.33**	37.16 ± 0.47
$ModTr_{\odot}$	65.25 ± 0.33	32.24 ± 0.95	**35.49 ± 0.94**	**64.96 ± 0.68**	**30.93 ± 0.50**	**34.27 ± 0.27**	**68.84 ± 0.40**	34.77 ± 0.22	**37.21 ± 0.46**

performance on par with $ResNet_{34}$ which have 24.4 million parameters. This reduction in parameters is even more pronounced when compared with a new detector on the desired modality. For example, a new FCOS detector would require more 33.2 million parameters. This trend is similar over the various detectors and remains consistent for the FLIR dataset (Table III-4).

Table-A III-3 Detection performance of ModTr with different backbones for the translation networks with different numbers of parameters. Here, we used three different detectors (FCOS, RetinaNet, and Faster R-CNN).
The methods were evaluated on IR test set of LLVIP dataset

Test Set IR (Dataset: LLVIP)				
Method	**Parameters**	**AP$_{50}$ ↑**	**AP$_{75}$ ↑**	**AP ↑**
FCOS	33.2 M			
MobileNet$_{v3s}$	+ 3.1 M	83.34 ± 1.76	53.34 ± 0.75	50.05 ± 1.01
MobileNet$_{v2}$	+ 6.6 M	90.36 ± 0.54	60.17 ± 0.56	55.33 ± 0.62
ResNet$_{18}$	+ 14.3 M	89.53 ± 1.03	58.54 ± 1.57	54.25 ± 0.90
ResNet$_{34}$	+ 24.4 M	90.90 ± 1.29	62.96 ± 2.44	56.93 ± 1.44
RetinaNet	34.0 M			
MobileNet$_{v3s}$	+ 3.1 M	87.67 ± 0.18	49.99 ± 0.57	49.65 ± 0.07
MobileNet$_{v2}$	+ 6.6 M	90.21 ± 0.82	54.60 ± 2.62	52.42 ± 1.33
ResNet$_{18}$	+ 14.3 M	89.53 ± 1.82	52.68 ± 2.06	51.40 ± 1.40
ResNet$_{34}$	+ 24.4 M	90.35 ± 0.60	57.18 ± 0.29	53.60 ± 0.41
Faster R-CNN	41.8 M			
MobileNet$_{v3s}$	+ 3.1 M	89.14 ± 0.63	56.85 ± 0.69	54.51 ± 0.28
MobileNet$_{v2}$	+ 6.6 M	91.32 ± 0.73	60.34 ± 0.66	56.15 ± 0.51
ResNet$_{18}$	+ 14.3 M	90.81 ± 0.46	59.39 ± 2.06	55.53 ± 1.14
ResNet$_{34}$	+ 24.4 M	91.04 ± 0.29	60.53 ± 2.16	56.35 ± 0.65

4. Qualitative Results

In this section, we provide additional visualizations for all the methods, including images generated by our proposal and its detection results. First, in Figure III-4, we present the detections in more detail, highlighting issues of those methods relying solely on translation, such as FastCUT, and some false positives when there is only FT. Subsequently, we provide additional visualizations for a batch of images processed by various detectors, i.e., in Figure III-5 for FCOS, in Figure III-6 for RetinaNet, and Figure III-7 for Faster R-CNN for both datasets.

Table-A III-4 Detection performance (AP) of ModTr with different backbones for the translation networks with different numbers of parameters. Here, we used three different detectors (FCOS, RetinaNet, and Faster R-CNN). The methods were evaluated on IR test set of FLIR dataset

Test Set IR (Dataset: FLIR)				
Method	**Parameters**	**AP_{50} ↑**	**AP_{75} ↑**	**AP ↑**
FCOS	33.2 M			
$MobileNet_{v3s}$	+ 3.1 M	56.73 ± 0.34	27.57 ± 0.65	29.66 ± 0.14
$MobileNet_{v2}$	+ 6.6 M	64.49 ± 0.99	32.17 ± 0.38	32.17 ± 0.38
$ResNet_{18}$	+ 14.3 M	64.39 ± 1.68	32.72 ± 1.50	34.44 ± 1.13
$ResNet_{34}$	+ 24.4 M	65.99 ± 0.78	33.73 ± 1.74	35.49 ± 0.94
RetinaNet	34.0 M			
$MobileNet_{v3s}$	+ 3.1 M	47.30 ± 0.54	18.63 ± 0.10	22.67 ± 0.18
$MobileNet_{v2}$	+ 6.6 M	64.01 ± 1.51	29.70 ± 0.62	33.12 ± 0.68
$ResNet_{18}$	+ 14.3 M	64.20 ± 0.58	30.84 ± 0.47	33.44 ± 0.47
$ResNet_{34}$	+ 24.4 M	66.31 ± 0.93	31.22 ± 0.69	34.27 ± 0.27
Faster R-CNN	41.8 M			
$MobileNet_{v3s}$	+ 3.1 M	61.03 ± 1.26	29.87 ± 0.86	32.06 ± 0.75
$MobileNet_{v2}$	+ 6.6 M	68.64 ± 0.56	34.76 ± 1.27	36.77 ± 0.67
$ResNet_{18}$	+ 14.3 M	68.49 ± 0.53	34.52 ± 0.23	36.68 ± 0.22
$ResNet_{34}$	+ 24.4 M	69.20 ± 0.36	34.58 ± 0.56	37.21 ± 0.46

5. Additional Modality: Canny Edges

As described in Bachmann *et al.* (2024), edges extracted with a Canny Edge detector from IR images can be used as an additional modality. Thus, in our work, we provide some qualitative results with it as well, as illustrated by Figure III-8 and quantitative results in Table III-5.

Here, we clarify that our approach is dependent on the zero-shot capability of the model to incorporate knowledge of the translation network. So, we can incorporate other modalities as well as the IR, but if the new modality does not have a good zero-shot like the Canny Edges, our method can incorporate the new knowledge but not surpass the fully fine-tuning (FT). But it is also important to mention that even without surpassing for edges, we were able to keep prior

Table-A III-5 Detection performance of Faster R-CNN Zero-Shot, Fine-Tuning, and ModTr on Edge Modality for LLVIP and FLIR Datasets. Edges from IR images were extracted using the Canny Edge Detector

Method	**Test Set Canny Edges (Dataset: LLVIP)**		
	Faster R-CNN		
	AP_{50} ↑	**AP_{75} ↑**	**AP ↑**
Zero-Shot	14.33 ± 0.00	08.15 ± 0.00	08.01 ± 0.00
Fine-Tuning	87.39 ± 1.74	55.50 ± 2.61	54.48 ± 2.40
ModTr	81.23 ± 1.53	48.66 ± 1.57	47.06 ± 1.18
Method	**Test Set Canny Edges (Dataset: FLIR)**		
	Faster R-CNN		
	AP_{50} ↑	**AP_{75} ↑**	**AP ↑**
Zero-Shot	18.43 ± 0.00	06.68 ± 0.00	08.44 ± 0.00
Fine-Tuning	60.43 ± 2.05	26.38 ± 0.28	30.88 ± 0.88
ModTr	51.81 ± 1.41	22.14 ± 0.83	25.59 ± 0.83

knowledge and reach good performance even though we still have a gap for the edges version when compared with FT.

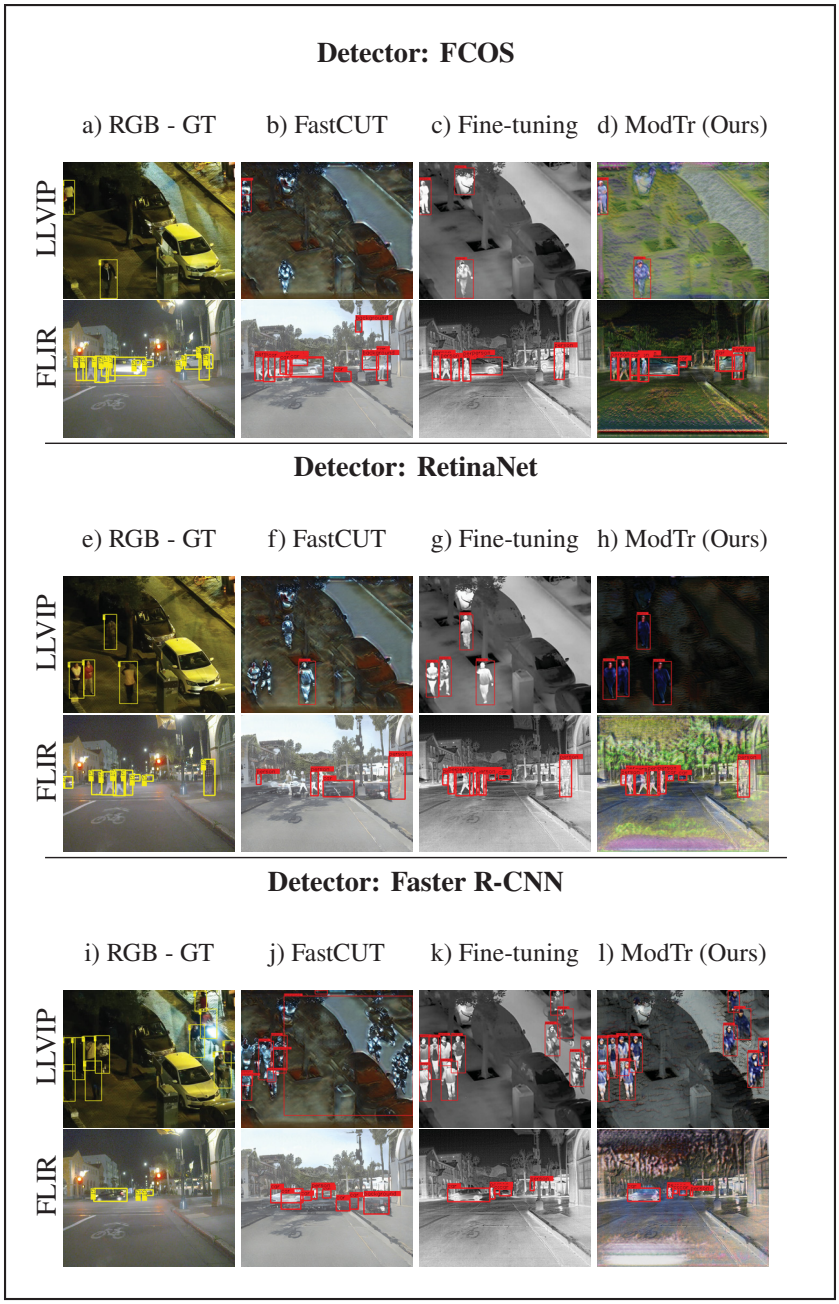


Figure-A III-4 Bounding box predictions for IR images from LLVIP and FLIR using different OD methods. Yellow and red boxes indicate ground truth and predictions, respectively. FastCUT (Park *et al.*, 2020a) translates IR to pseudo-RGB but is not detection-focused and requires both modalities. Fine-tuning adapts to IR using only IR data but forgets pre-trained knowledge. In contrast, ModTr adapts for detection using only IR and preserves the original model knowledge

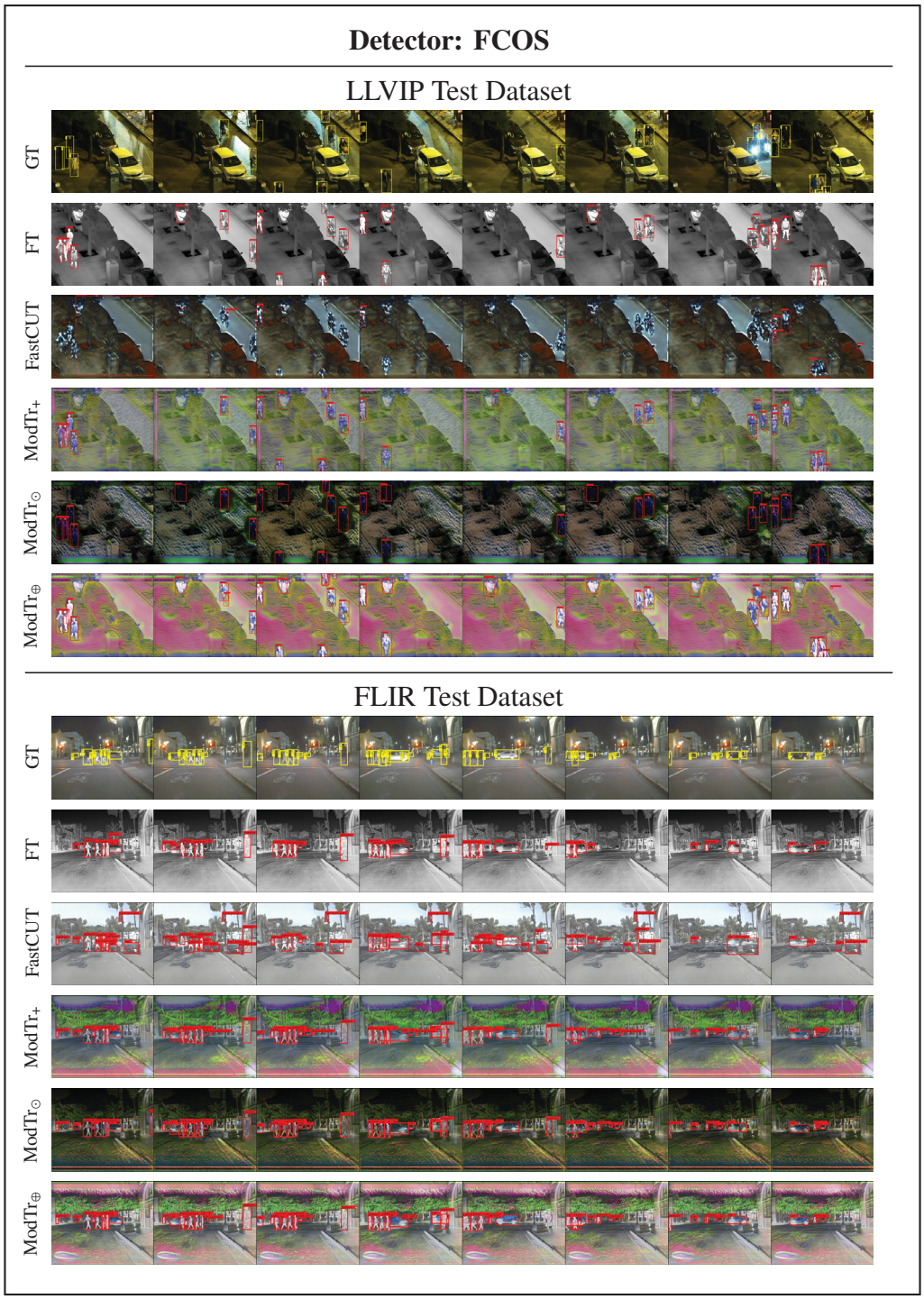


Figure-A III-5 Illustration of a sequence of 8 images of LLVIP and FLIR dataset for FCOS. For each dataset, the first row is the RGB modality, followed by the IR modality, followed by FastCUT (Park *et al.*, 2020a), and different representations created by ModTr and its variations

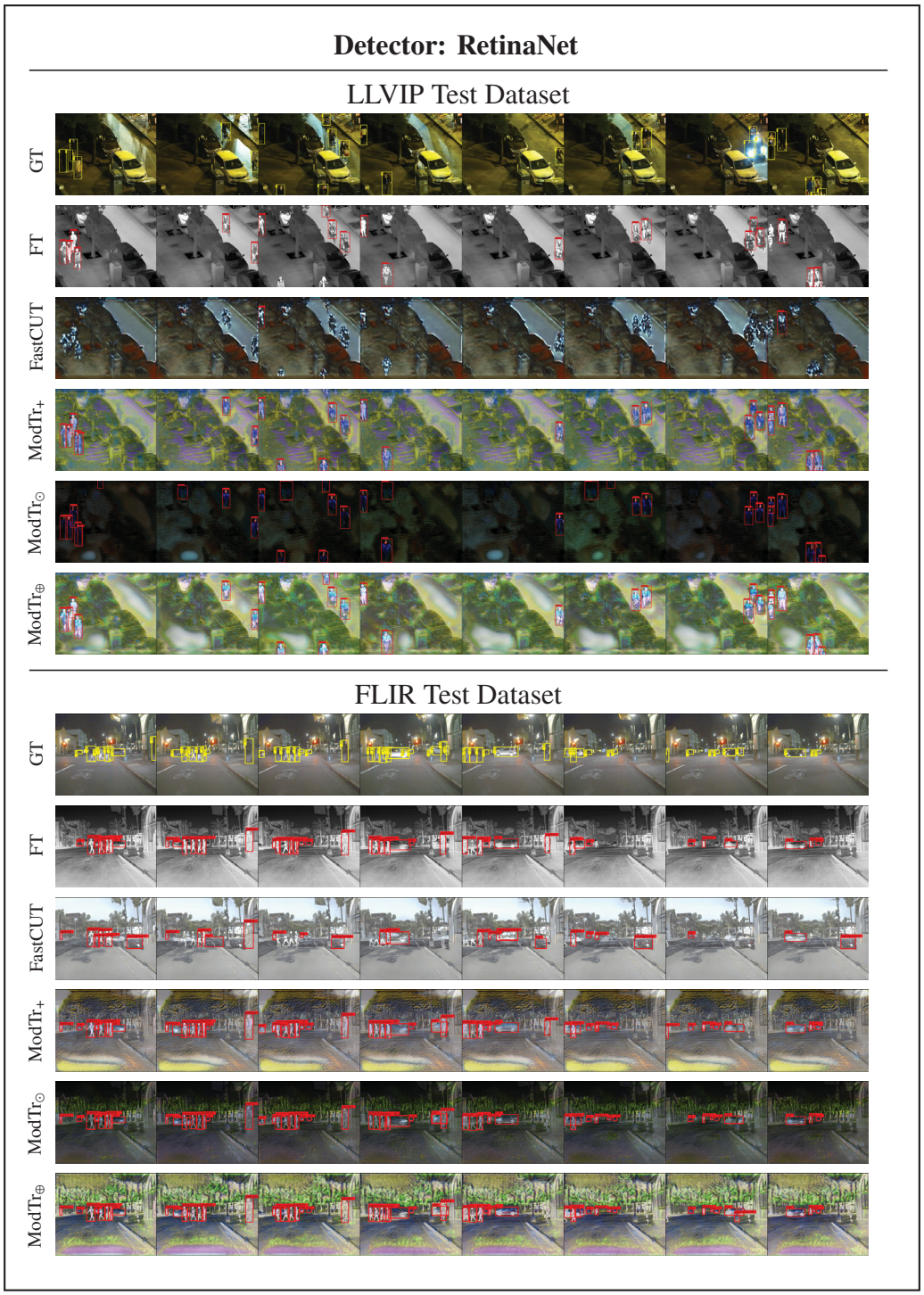


Figure-A III-6 Illustration of a sequence of 8 images of LLVIP and FLIR dataset for RetinaNet. For each dataset, the first row is the RGB modality, followed by the IR modality, followed by FastCUT (Park *et al.*, 2020a), and different representations created by ModTr and their variations

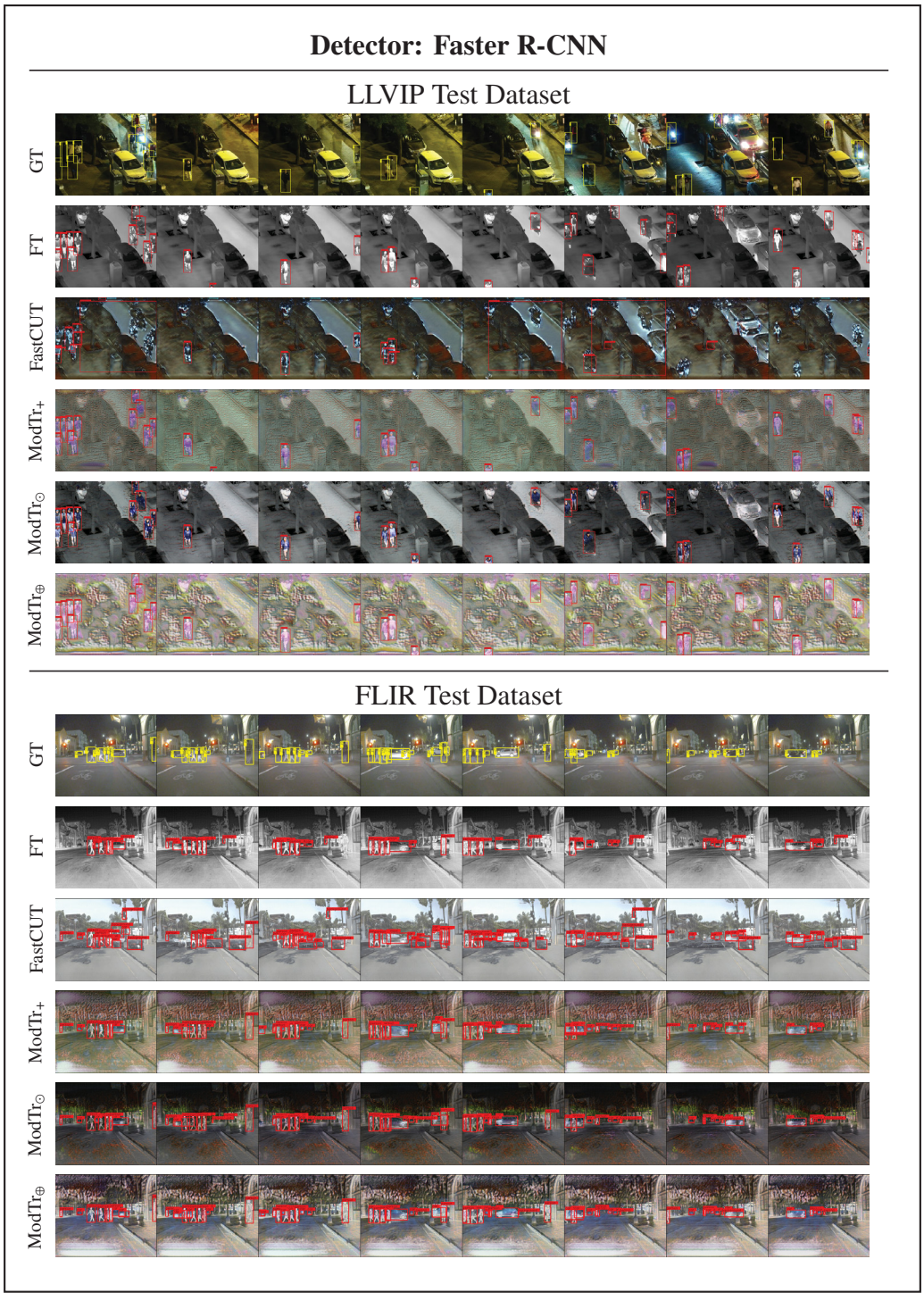


Figure-A III-7 Illustration of a sequence of 8 images of LLVIP and FLIR dataset for Faster R-CNN. For each dataset, the first row is the RGB modality, followed by the IR modality, followed by FastCUT (Park *et al.*, 2020a), and different representations created by ModTr and their variations

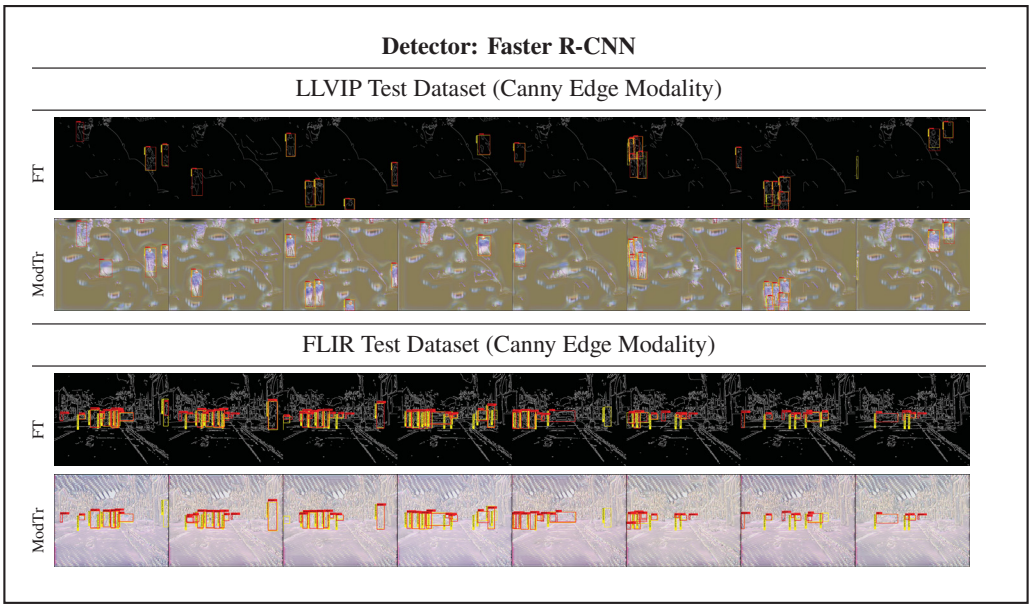


Figure-A III-8 Illustration of a sequence of 8 images from LLVIP and FLIR Canny edges test set for Faster R-CNN. The first row is the FT on the Canny edges modality, and the second row is the result of ModTr for the LLVIP. The third row is FT on the Canny edges for FLIR, and the fourth row is the result of ModTr

APPENDIX IV

SUPPLEMENTARY MATERIAL FOR THE PAPER TITLED VISUAL MODALITY PROMPT FOR ADAPTING VISION-LANGUAGE OBJECT DETECTORS

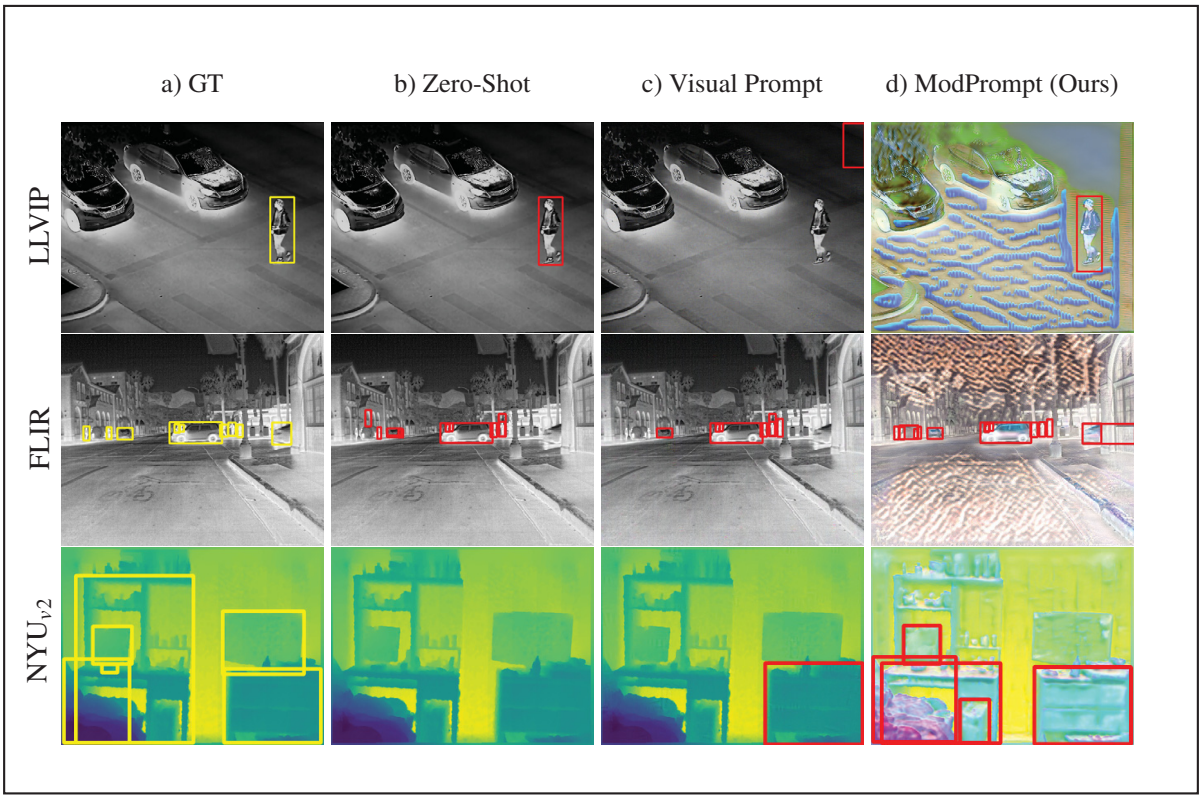


Figure-A IV-1 Detections of different approaches across modalities: LLVIP and FLIR datasets (infrared) and NYU$_{v2}$ (depth). Each column corresponds to a different approach: (a) GT (Ground Truth): Shows in yellow the ground-truth bounding boxes for objects. (b) Zero-Shot: Displays detections (in red) from a zero-shot model. This model misses several detections and predicts inaccurate boxes without specific tuning. (c) Visual Prompt: Illustrates detections with a weight map visual prompt added to the image. It shows improvements over zero-shot, with more accurate detection, but still misses some objects. (d) ModPrompt (Ours): Detections from our proposed model. ModPrompt generates artifacts on the image to enhance objects and suppress the background, facilitating the detector

In the supplementary material, we provide additional information to reproduce our work. The source code is available at https://github.com/heitorrapela/ModPrompt. The supplementary material is divided into the following sections: Section 1 with additional details regarding the

implementation and architecture of the vision-language object detectors used in our work. Then in Section 2 we formally define text-prompt tuning in more detail. And in Section 3 we provide some additional results. Specifically, in Section 3.1 we provide additional main results on the FLIR-IR dataset. Then, we provide results with different backbones, YOLO-World-Large and Grounding DINO-B in Section 3.2 on FLIR-IR and NYU_{v2}-DEPTH. Further, in Section 3.3 we provide the results for the FLIR-IR dataset with the learnable MPDR. In Section 3.4 we provide results of ablations using different visual prompt strategies. Then, in Section 3.5 we compare our method with state-of-the-art modality translation OD methods, and in Section 3.6 we show additional visualizations. Finally, in Section 4 we provide some limitations of our work and possible future directions.

Table-A IV-1 Detection performance (APs) for YOLO-World and Grounding DINO for the FLIR-IR dataset. The different visual prompt adaptation techniques are compared with our ModPrompt, and the zero-shot (ZS), head finetuning (HFT), and full finetuning (FT) are also reported, where the full finetuning is the upper bound

Dataset	Method	YOLO-World			Grounding DINO		
		AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP
FLIR - IR	Zero-Shot (ZS)	64.70 ± 0.00	32.10 ± 0.00	34.90 ± 0.00	64.30 ± 0.00	29.30 ± 0.00	32.90 ± 0.00
	Head Finetuning (HFT)	73.90 ± 0.14	36.47 ± 0.12	40.00 ± 0.00	67.27 ± 0.15	33.60 ± 0.26	36.20 ± 0.10
	Full Finetuning (FT)	80.47 ± 1.11	41.13 ± 0.41	44.17 ± 0.39	80.73 ± 0.06	42.17 ± 0.78	44.17 ± 0.25
	Visual Prompt (Fixed)	45.60 ± 0.08	21.90 ± 0.08	23.77 ± 0.05	64.27 ± 0.06	29.47 ± 0.06	32.83 ± 0.12
	Visual Prompt (Random)	43.27 ± 0.29	20.63 ± 0.17	22.47 ± 0.12	64.03 ± 0.08	**29.50 ± 0.00**	32.77 ± 0.15
	Visual Prompt (Padding)	45.63 ± 1.44	22.13 ± 1.08	23.87 ± 0.91	61.73 ± 0.23	27.60 ± 0.17	31.20 ± 0.20
	Visual Prompt (WM)	54.43 ± 0.78	26.37 ± 0.54	28.67 ± 0.40	54.73 ± 0.08	23.20 ± 0.10	27.20 ± 0.10
	Visual Prompt (WM_{v2})	52.43 ± 0.50	25.10 ± 0.22	27.50 ± 0.22	54.80 ± 0.10	23.13 ± 0.21	27.27 ± 0.06
	ModPrompt (Ours)	**69.03 ± 1.06**	**34.87 ± 0.40**	**37.33 ± 0.12**	**65.03 ± 0.05**	29.23 ± 0.55	**32.90 ± 0.26**

1. Additional Details of Vision-Language ODs

For the YOLO-World, we use AdamW optimizer with a learning rate $2e^{-4}$, weight decay 0.05, and batch size 8. And for the Grounding-DINO, we use AdamW optimizer with a learning rate $1e^{-4}$, weight decay $1e^{-4}$, and batch size 8. For the main manuscript, we used YOLO-World Small and Grounding-DINO Tiny. For the experiments with text, we extract the embeddings and optimize them without the text encoder for efficient adaptation of the embedding space. Additionally, we provide results with bigger backbones to further corroborate our findings.

Table-A IV-2 Detection performance (APs) for YOLO-World-Large and Grounding DINO-B on FLIR-IR and NYU_{v2}-Depth datasets. Each visual prompt adaptation strategy is compared with our ModPrompt

Detector	Method	FLIR-IR			NYU_{v2}-DEPTH		
		AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP
YOLO-World-Large	Zero-Shot (ZS)	71.60 ± 0.00	37.60 ± 0.00	39.30 ± 0.00	05.30 ± 0.00	03.70 ± 0.00	03.50 ± 0.00
	Head Finetuning (HFT)	82.27 ± 0.21	43.53 ± 0.12	45.93 ± 0.05	24.43 ± 0.17	14.63 ± 0.29	14.53 ± 0.17
	Full Finetuning (FT)	84.33 ± 0.45	44.00 ± 0.29	46.70 ± 0.22	54.13 ± 0.05	40.43 ± 0.09	37.33 ± 0.17
	Visual Prompt (Fixed)	71.83 ± 0.09	37.70 ± 0.00	39.40 ± 0.00	05.20 ± 0.00	03.60 ± 0.00	03.40 ± 0.00
	Visual Prompt (Random)	71.40 ± 0.14	37.40 ± 0.08	39.23 ± 0.05	04.67 ± 0.12	03.17 ± 0.09	02.97 ± 0.09
	Visual Prompt (Padding)	66.27 ± 0.12	33.83 ± 0.05	35.97 ± 0.05	02.87 ± 0.12	01.80 ± 0.08	01.80 ± 0.08
	Visual Prompt (WM)	71.70 ± 0.14	37.53 ± 0.25	39.43 ± 0.05	15.10 ± 0.45	09.83 ± 0.26	09.50 ± 0.24
	Visual Prompt (WM_{v2})	70.90 ± 0.08	36.73 ± 0.19	38.73 ± 0.05	14.53 ± 0.41	09.57 ± 0.33	09.13 ± 0.31
	ModPrompt (Ours)	**77.13 ± 0.29**	**41.37 ± 0.69**	**43.23 ± 0.24**	**39.27 ± 0.53**	**28.93 ± 0.17**	**26.73 ± 0.09**
Grounding DINO-Big	Zero-Shot (ZS)	64.80 ± 0.00	29.10 ± 0.00	32.90 ± 0.00	08.50 ± 0.00	05.70 ± 0.00	05.40 ± 0.00
	Head Finetuning (HFT)	68.40 ± 0.14	34.60 ± 0.28	36.95 ± 0.21	08.50 ± 0.00	06.10 ± 0.10	05.67 ± 0.06
	Full Finetuning (FT)	81.97 ± 0.25	34.60 ± 0.28	45.85 ± 0.07	53.93 ± 0.40	40.53 ± 0.15	37.50 ± 0.26
	Visual Prompt (Fixed)	64.87 ± 0.06	29.17 ± 0.06	33.00 ± 0.00	08.53 ± 0.06	05.73 ± 0.06	05.53 ± 0.06
	Visual Prompt (Random)	64.83 ± 0.06	29.13 ± 0.06	32.93 ± 0.06	08.53 ± 0.06	05.77 ± 0.15	05.47 ± 0.06
	Visual Prompt (Padding)	62.63 ± 0.06	27.27 ± 0.06	31.53 ± 0.06	07.93 ± 0.06	05.13 ± 0.06	04.83 ± 0.06
	Visual Prompt (WM)	56.77 ± 0.31	21.93 ± 0.55	27.00 ± 0.36	05.57 ± 0.06	03.23 ± 0.06	03.17 ± 0.06
	Visual Prompt (WM_{v2})	57.03 ± 0.40	22.20 ± 0.40	27.20 ± 0.30	05.73 ± 0.06	03.33 ± 0.06	03.33 ± 0.06
	ModPrompt (Ours)	**65.73 ± 0.15**	**30.07 ± 0.12**	**33.47 ± 0.12**	**25.30 ± 0.53**	**17.60 ± 0.20**	**16.57 ± 0.35**

2. Formal definition of Text-Prompt Tuning

Following our definitions of visual prompts for object detection in the main manuscript, we define the text-prompt tuning using our notation in this section. YOLO-World follows the YOLOv8 loss from Jocher *et al.* (2023), with a text contrastive head to provide the object-text similarities; for more details about YOLO-World loss check (Cheng *et al.*, 2024). Here, we provide a generic definition, independent of the model. Thus, we define the text-prompt cost function ($C_{\text{tp}}(\phi)$), with the following Equation:

$$C_{\text{tp}}(\phi) = \frac{1}{|\mathcal{D}|} \sum_{(x_t, Y) \in \mathcal{D}} \mathcal{L}_{text}(g_\psi(x_t + h_\phi), Y), \tag{A IV-1}$$

where x_t is the input text, g_ψ is the text-encoder, h_ϕ is the additional prompt parametrized by ϕ. In the case of YOLO-World, $\mathcal{L}_{text}$ can be seen as label assignment of Feng, Zhong, Gao, Scott & Huang (2021) to match the predictions with ground-truth annotations, with Binary Cross Entropy (BCE), and assign each positive prediction with a text index as the classification label.

3. Additional Results

3.1 Main Results on FLIR-IR data:

In Table IV-1 we compare the performance of our method against the baselines on the FLIR-IR dataset. It can be observed that our ModPrompt achieves the highest performance in terms of APs for YOLO-World, and for Grounding DINO, the AP_{50} and AP results were the highest, while the AP_{75} is equally good as the random prompt. The FLIR-IR dataset is a more challenging dataset composed of 3 classes, some small bounding boxes, and missing annotations, which make the problem more difficult when the input image is being changed only with detector feedback and without the availability of the RGB groundtruth images for image-to-image translation during training. We observe that ModPrompt performs better when objects are well-defined in the image and when objects are not too small, otherwise, like all other input-level pixel strategies it faces challenges, especially on refined bounding-box localization, which can be seen with AP_{75} and AP, whereas in AP_{50} it always shows good results.

3.2 Results with Different Detection Backbones:

In this section, we provide results for the YOLO-World-Large and Grounding DINO-Big models with different visual prompt strategies and our ModPrompt. In Table IV-2, we show that ModPrompt is better than all visual prompt methods for FLIR and NYU_{v2}.

3.3 MPDR with FLIR-IR data:

In Table IV-3, we provide additional results for FLIR-IR with our MPDR module. We emphasize that the knowledge preservation strategy improves performance in many cases. However, this dataset is too noisy, which compromises translation methods such as ModPrompt, resulting in degradation of performance in some cases.

3.4 Ablation Studies on Visual Prompts:

We evaluate different variations of the visual prompt adaptation methods. Specifically, we compare the performance when different input patch sizes are used; for instance, $p_s = 30$ refers

Table-A IV-3 Detection performance (APs) for YOLO-World and Grounding DINO on FLIR-IR data. Each visual prompt adaptation strategy is compared with the learnable MPDR (results in parentheses are the gain with the MPDR module), which is responsible for updating the new modality embeddings and not changing the original embedding knowledge

Detector	Method	FLIR-IR		
		AP_{50}	AP_{75}	AP
YOLO-World	**Fixed**	63.93 ± 0.26 (+0.63)	31.90 ± 0.08 (-0.60)	34.63 ± 0.05 (+0.33)
	Random	63.30 ± 0.14 (+0.43)	31.73 ± 0.09 (-0.44)	34.27 ± 0.12 (+0.17)
	Padding	59.23 ± 0.12 (+0.66)	29.10 ± 0.08 (-0.10)	31.50 ± 0.08 (+0.17)
	WeightMap	63.10 ± 0.28 (+0.33)	31.60 ± 0.29 (-0.10)	34.20 ± 0.08 (+0.33)
	ModPrompt	**72.73 ± 0.00 (-1.74)**	**37.70 ± 0.00 (-0.60)**	**40.37 ± 0.00 (-0.73)**
Grounding DINO	**Fixed**	66.53 ± 0.98 (+2.26)	**32.83 ± 2.50 (+3.36)**	**34.83 ± 0.90** (+2.00)
	Random	66.10 ± 1.08 (+2.07)	31.40 ± 1.31 (+1.90)	34.53 ± 0.90 (+1.76)
	Padding	63.33 ± 1.10 (+1.60)	29.73 ± 1.27 (+2.13)	32.93 ± 0.90 (+1.73)
	WeightMap	55.60 ± 0.92 (+0.87)	24.37 ± 0.65 (+1.17)	28.30 ± 0.66 (+1.10)
	ModPrompt	**67.80 ± 0.14 (+2.77)**	31.10 ± 0.31 (+1.87)	33.70 ± 0.09 (+0.80)

to a patch size of 30 pixels. In this study, we test multiple patch sizes for each of the visual prompt methods and report the performance in Table IV-4. We evaluate ModPrompt using two different translators with U-Net based backbones, MobileNet (MB) (Howard *et al.*, 2017) and ResNet (RES) (He *et al.*, 2016). Here, we provide the additional results on the FLIR-IR dataset.

3.5 Comparison with SOTA Modality Translation OD methods:

Our ModPrompt technique is compared with recent state-of-the-art modality translation methods for ODs: HalluciDet (Medeiros *et al.*, 2024c) and ModTr (Medeiros *et al.*, 2024b). In Fig. IV-2, we observe that our results are better in all APs for the FLIR dataset.

3.6 Qualitative Results:

In this section, we provide more visual results for the methods compared, where the performance of each model can be shown by the bounding box predictions. For instance, in Figure IV-1, we

Table-A IV-4 Detection performance (APs) for YOLO-World on FLIR-IR data. We compared the main visual prompt strategies *fixed*, *random*, *padding*, and ModPrompt. The variations consist of the number of prompt pixels (p_s = 30, 200 or 300) and for ModPrompt, the MobileNet (MB) or ResNet (RES)

Method	**Variation**	**FLIR - IR**		
		AP_{50}	**AP_{75}**	**AP**
Fixed	30	45.60 ± 0.08	21.90 ± 0.08	23.77 ± 0.05
	300	29.30 ± 0.37	13.50 ± 0.54	15.00 ± 0.37
Random	30	43.27 ± 0.29	20.63 ± 0.17	22.47 ± 0.12
	300	19.13 ± 0.33	09.00 ± 0.42	09.80 ± 0.33
Padding	30	45.63 ± 1.44	22.13 ± 1.08	23.87 ± 0.91
	200	00.53 ± 0.12	00.17 ± 0.17	00.27 ± 0.09
ModPrompt	MB	66.80 ± 0.29	**35.23 ± 0.38**	36.53 ± 0.12
	RES	**69.03 ± 1.06**	34.87 ± 0.40	**37.33 ± 0.12**

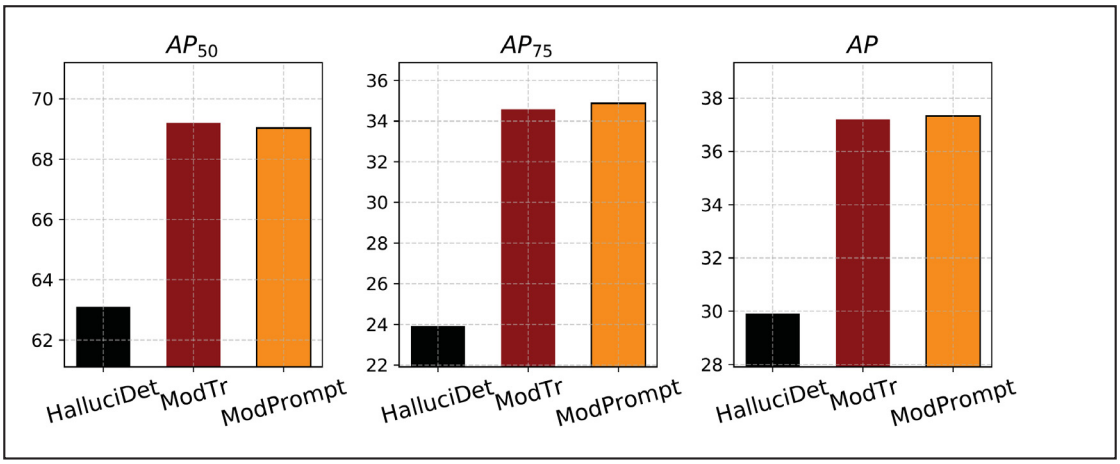


Figure-A IV-2 Detection performance on FLIR-IR dataset of different modality translators for OD in terms of APs

can see that Zero-Shot (ZS) is performing well on the FLIR-IR dataset, apart from some false positives. However, the ZS doesn't perform well on the modality that is too different from the

pre-training weight, such as depth (NYU_{v2} dataset). For the Visual Prompt (weight map), we have some false positives and missing detections (e.g., the person in the first row for LLVIP and a wrong bounding box). For ModPrompt, we see that we have some good overall detections because the encoder-decoder architecture tends to suppress a little bit of background, as we can see in the first row or second row, but the bounding boxes are not as precise as the right ones in the ZS (see first-row comparison between ZS and ModPrompt). Additionally, in Figure IV-3, we provide a batch of 8 images from the test, and we can observe a similar trend for depth and IR modalities as discussed above. Surprisingly, in some cases of the FLIR dataset (challenging dataset with really hard small bounding boxes and some missing labels), our method tends to detect objects that are not labeled in the ground truth (for instance a small person in the first row and first column, behind the two cars on the left, which are not labeled, but our method get its right). This shows the effectiveness of our method and exemplifies the ability of visual language detectors to detect unseen objects.

4. Limitations and future works

Limitations: We believe our work still has some limitations, which we believe can be further improved in subsequent works. For instance, adaptation strategies still require target labels, which could be explored in other tasks, such as unsupervised or semi-supervised approaches. Additionally, we argue that our method is still not perfect for small objects, and it incorporates some noise, which can be further minimized by other loss constraints if we have access to additional source data (which we didn't during training). Some of the limitations were already discussed in the qualitative results, which can be summarized as difficulties with small objects and duplications of bounding box predictions.

Future works: Future works could improve the conditioning on both text and vision, and exploit more label-efficient adaptation strategies such as test-time adaptation or few-shot learning.

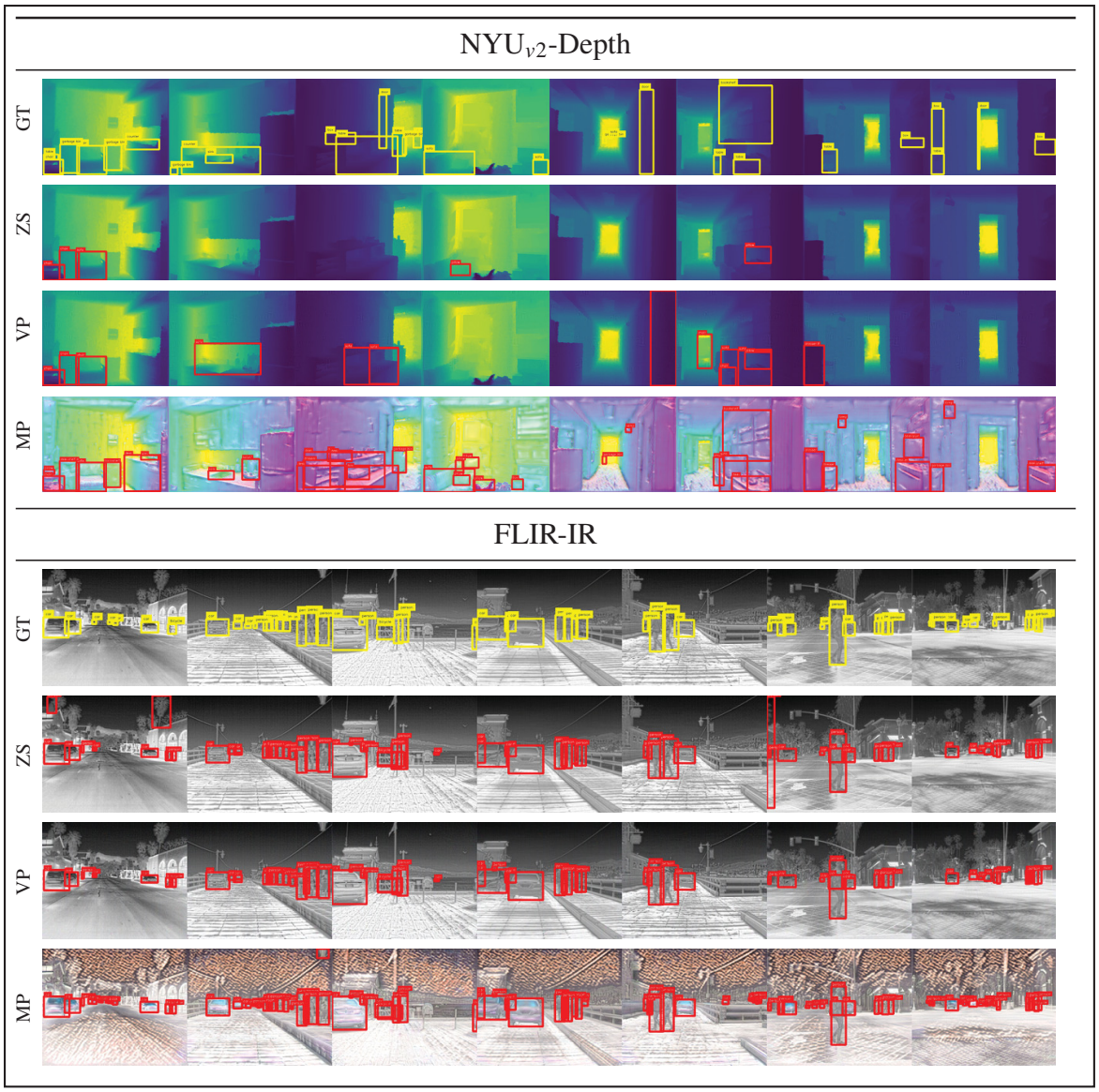


Figure-A IV-3 Detections of different approaches across modalities for YOLO-World: NYU$_{v2}$ (depth) and FLIR (infrared). Each row corresponds to a different approach: GT (Ground Truth): Shows in yellow the ground-truth bounding boxes for objects. ZS (Zero-Shot): Displays detections (in red) from a zero-shot model YOLO-World-s. VP (Visual Prompt): Illustrates detections with weight map visual prompt added to the image. MP (ModPrompt): Detections from our proposed model

APPENDIX V

PAPER AND SUPPLEMENTARY MATERIAL FOR WISE-OD: BENCHMARKING ROBUSTNESS IN INFRARED OBJECT DETECTION

Heitor R. Medeiros[a], Atif Belal[a], Srikanth Muralidharan[a]
Eric Granger[a], Marco Pedersoli[a]

[a] Department of Systems Engineering, École de technologie supérieure,
1100 Notre-Dame West, Montreal, Quebec, Canada H3C 1K3

Paper submitted to *IEEE/CVF Winter Conference on Applications of Computer Vision, September 2025*

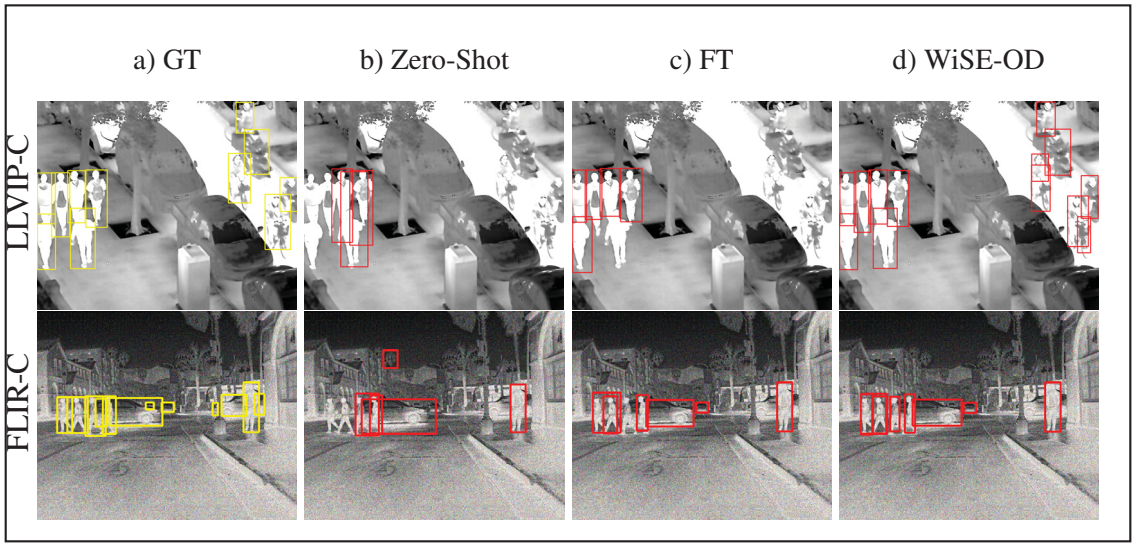


Figure-A V-1 Robustness in Infrared Object Detection on LLVIP-C and FLIR-C datasets. In the first row, LLVIP-C has a brightness corruption severity level of 5; in the second row, FLIR-C shot noise corruption has a severity level of 2. In a) ground truth in yellow, b) zero-shot COCO detection, c) fine-tuning (FT), and d) WiSE-OD for Faster R-CNN detector

Abstract: Object detection (OD) in infrared (IR) imagery is critical for low-light and nighttime applications. However, the scarcity of large-scale IR datasets forces models to rely on weights pre-trained on RGB images. While fine-tuning on IR improves accuracy, it often compromises robustness under distribution shifts due to the inherent modality gap between RGB and IR.

To address this, we introduce LLVIP-C and FLIR-C, two cross-modality out-of-distribution (OOD) benchmarks built by applying corruption to standard IR datasets. Additionally, to fully leverage the complementary knowledge from RGB and infrared trained models, we propose WiSE-OD, a weight-space ensembling method with two variants: WiSE-OD$_{ZS}$, which combines RGB zero-shot and IR fine-tuned weights, and WiSE-OD$_{LP}$, which blends zero-shot and linear probing. Evaluated across three RGB-pretrained detectors and two robust baselines, WiSE-OD improves both cross-modality and corruption robustness without any additional training or inference cost.

1. Introduction

In recent years, deep learning (DL) has achieved significant success across various computer vision tasks, including object detection (OD) (Zou *et al.*, 2023) using thermal infrared (IR) imaging (Medeiros *et al.*, 2024c; Medeiros, Belal, Muralidharan, Granger & Pedersoli, 2025a). Unlike visible spectrum imaging (RGB), which relies on reflected light, thermal IR imaging captures the heat emitted by objects, allowing it to function independently of lighting conditions. This makes IR-based OD highly effective in challenging environments with limited or absent visible light, such as night-time surveillance and autonomous driving cars (Teutsch, Sappa & Hammoud, 2021). Despite these advantages, IR OD models must maintain consistent performance and reliable predictions despite variations in input, due to occlusions, viewpoint shifts, or image degradation. Ensuring such robustness is therefore essential for real-world applications in surveillance (Ramachandran & Sangaiah, 2021; Dubail *et al.*, 2022), autonomous driving (Michaelis *et al.*, 2019), and defense (Liu, Feng, Hu, Yuan & Fan, 2024b), where fluctuating environmental conditions can impact sensor inputs and compromise system reliability.

Without large-scale IR pre-training datasets, OD models for IR typically initialize from powerful models pre-trained on large-scale RGB datasets (e.g., COCO (Lin *et al.*, 2014)), followed by fine-tuning on IR data. While this pipeline yields strong in-domain (ID) performance, where ID refers to test samples similar to the training data, it often compromises robustness against out-of-domain (OOD) samples (Hendrycks & Dietterich, 2019). OOD samples differ

significantly from the training data, leading to performance degradation. This deterioration stems from the fact that the fine-tuning process tends to cause the model to prioritize task-specific information at the expense of broader knowledge acquired during pre-training. As a result, the model struggles to generalize to new or diverse scenarios (Wortsman *et al.*, 2022b). This issue is further amplified by the already substantial modality shift between RGB and IR, making robust transfer learning even more difficult (Medeiros *et al.*, 2024b). In classification, several techniques have been proposed to improve robustness under distribution shifts, including linear probing (LP), LP followed by fine-tuning (LP-FT) (Kumar, Raghunathan, Jones, Ma & Liang, 2022), and weight-space ensembling (WiSE-FT) (Wortsman *et al.*, 2022b). While these methods effectively enhance robustness in classification, they either are not directly applicable to object detection or remain under-explored (Wortsman *et al.*, 2022b). This is because of the complex detection architectures and robustness, which demand both accurate localization and classification. Moreover, cross-modality adaptation from RGB to infrared (IR) introduces additional challenges due to the modality shift and the scarcity of large-scale IR data (Medeiros *et al.*, 2024c,b, 2025a).

To address these challenges, in this paper, we introduce two key components: a novel cross-modality RGB/IR corruption benchmark, and two efficient approaches to improve average performance in the IR OOD corruption setting without additional training or inference cost. Our benchmark, LLVIP-C and FLIR-C, applies common corruption transforms to the original LLVIP and FLIR datasets to evaluate cross-modality OOD performance. Using this benchmark, we assess three families of IR ODs fine-tuned from RGB pre-trained models against standard robust fine-tuning baselines. Our analysis shows that traditional methods underperform in corruption settings; therefore, we introduced WiSE-OD, a simple approach that preserves the original detection head to combine zero-shot and fine-tuned weights, yielding WiSE-OD$_{ZS}$ and its linear probing variation WiSE-OD$_{LP}$. We empirically found that these weight-space ensembling methods exhibit significant robustness. Additional analysis across various levels of corruption demonstrates that these methods improve average IR model performance by preserving the ID accuracy from fine-tuning and OOD robustness from zero-shot weights, which explains their

effectiveness.

Our main contributions can be summarized as follows:

- A new benchmark, LLVIP-C and FLIR-C, is introduced to advance the evaluation of robust cross-modality OD between RGB and IR. This benchmark is essential for measuring detector performance across diverse, real-world conditions. Within this framework, we comprehensively evaluate three widely used object detection models: Faster R-CNN, FCOS, and RetinaNet, each initialized with COCO pre-trained weights.
- WiSE-OD with two variations: WiSE-OD$_{ZS}$ and WiSE-OD$_{LP}$, is proposed for OD, an efficient technique to combine zero-shot and finetuned weights, enhancing the robustness under distribution shift.
- Extensive experiments are conducted on the proposed benchmark and our WiSE-OD technique, demonstrating a significant gain in performance over three OD frameworks.

2. Related Works

Object detection. OD is one of the most challenging computer vision tasks (Liu *et al.*, 2020), especially due to many different environmental conditions (Michaelis *et al.*, 2019). The objective of OD is to localize with a bounding box and provide labels for all objects in an image (Zhang *et al.*, 2023a). Commonly, detectors can be categorized into two different groups: one-stage and two-stage detectors. The most famous and traditional two-stage detector is the Faster R-CNN (Ren *et al.*, 2015), which first generates regions of interest and then uses a second classifier to confirm object presence within those regions. On the contrary, one-stage detectors eliminate the proposal generation stage, which focuses on real-time inference speed. On the one-stage detectors group, RetinaNet (Lin *et al.*, 2017b) utilizes a focal loss to address the class imbalance. Also, models like FCOS (Tian *et al.*, 2019) have emerged in this category, eliminating predefined anchor boxes to potentially enhance inference efficiency. Our work focuses on these three traditional and powerful detectors: Faster R-CNN, RetinaNet, and FCOS.

Robustness in Object Detection. Robustness in OD refers to the model's ability to maintain performance despite variations in input conditions. Hendrycks & Dietterich (2019) proposed diverse corruptions for classification datasets, resulting in ImageNet-C and CIFAR10-C. Michaelis *et al.* (2019) extended this to OD, proposing Pascal-C, COCO-C, and Cityscapes-C with a study on corruption severity and detector performance. Beghdadi, Mallem & Beji (2022) introduced additional local transformations for RGB OD on COCO, and Mao *et al.* (2023) proposed COCO-O with six types of natural distribution shifts. Despite growing efforts for RGB OD robustness, IR OD still lacks such benchmarks. In this direction, Josi, Alehdaghi, Cruz & Granger (2023b) applied classification corruptions to IR for person ReID. Given the widespread use of IR in surveillance and autonomous driving, a robustness benchmark for IR OD is essential.

Robust Fine-Tuning. The deep learning community has explored various fine-tuning (FT) strategies to improve robustness in classification tasks. A common approach is linear probing (LP), where the backbone is frozen and only the head is trained. Kumar *et al.* (2022) extended this with LP-FT, which first trains a linear head before unfreezing the backbone for full fine-tuning. Wortsman *et al.* (2022b) proposed WiSE-FT, which ensembles the weights of a zero-shot pretrained model and its fine-tuned counterpart in weight space, showing strong performance under distribution shifts on ImageNet.

In this work, we adapt these robustness techniques, which were originally developed for classification, to the more challenging cross-modality object detection setting, offering simple yet effective strategies to mitigate corruption effects. In the next section, we introduce our proposed IR OD benchmark and baselines.

3. Background

In this section, we introduce preliminary definitions that are necessary to understand this work, and subsequently, we define our proposed benchmark.

Object Detection. Consider a set of training samples $\mathcal{D} = \{(x_i, B_i)\}$, where $x_i \in \mathbb{R}^{W \times H \times C}$ are images with spatial resolution $W \times H$ and C channels, and $B_i = \{b_0, b_1, ..., b_N\}$ is a set of bounding boxes corresponding image x_i. Each bounding box can be represented as $b = (c_x, c_y, w, h, o)$ where c_x and c_y are the center coordinates of the bounding box with size $w \times h$ and o is the class label. During training we aim to learn a parameterized function $f_\theta : \mathbb{R}^{W \times H \times C} \rightarrow \mathcal{B}$, with $\mathcal{B}$ being the family of sets B_i and θ the model's parameters vector. The optimization of f_θ is guided by a combination of a regression $\mathcal{L}_r$ and classification $\mathcal{L}_c$ loss, i.e., l_2 loss and binary cross-entropy, respectively. The loss function for object detection can be represented as:

$$\mathcal{L}_d(\theta) = \frac{1}{|\mathcal{D}|} \sum_{(x,B)\in\mathcal{D}} \mathcal{L}_c(f_\theta(x), B) + \lambda \mathcal{L}_r(f_\theta(x), B). \tag{A V-1}$$

Robustness to Corruption. Corruption robustness measures a classifier's average performance under classifier-agnostic input distortions (Hendrycks & Dietterich, 2019). Let $f : \mathcal{X} \rightarrow \mathcal{Y}$ be a classifier trained on samples drawn from a distribution $\mathcal{Q}$, and let $C = \{ c : \mathcal{X} \rightarrow \mathcal{X}\}$ be a set of corruption functions (e.g. noise, blur, contrast) and $\mathcal{E} = \{ e : \mathcal{X} \rightarrow \mathcal{X}\}$ a set of additional perturbation functions. The classifier's clean accuracy is $\mathbb{P}_{(x,y)\sim\mathcal{Q}}\big(f(x) = y\big)$. Its corruption robustness, i.e., its expected accuracy under all compositions of one corruption and one perturbation, can be represented as:

$$\mathbb{E}_{c\sim C}\, \mathbb{E}_{e\sim\mathcal{E}}\Big[\, \mathbb{P}_{(x,y)\sim\mathcal{Q}}\big(f\big(e(c(x))\big) = y\big)\Big].$$

Weight-space Ensembling. Given a mixing coefficient $\lambda \in [0, 1]$, weight-space ensembling can be defined as the following function:

$$f_{\text{wse}}(\theta_i, \theta_j; \lambda) \;=\; (1 - \lambda)\, \theta_i \;+\; \lambda\, \theta_j, \tag{A V-2}$$

which computes the element-wise convex combination of two parameter vectors θ_i and θ_j. The resulting ensemble parameter $\theta_{\text{ens}} = f_{\text{wse}}(\theta_i, \theta_j; \lambda)$ is then used to initialize the model for prediction. A notable example of this technique is WiSE-FT (Wortsman *et al.*, 2022b). Moreover, weight-space ensembling builds on principles of output-space ensemble averaging (Izmailov, Podoprikhin, Garipov, Vetrov & Wilson, 2018) and has demonstrated improved OOD robustness on classification benchmarks (Wortsman *et al.*, 2022a).

4. OD IR Robustness Benchmark

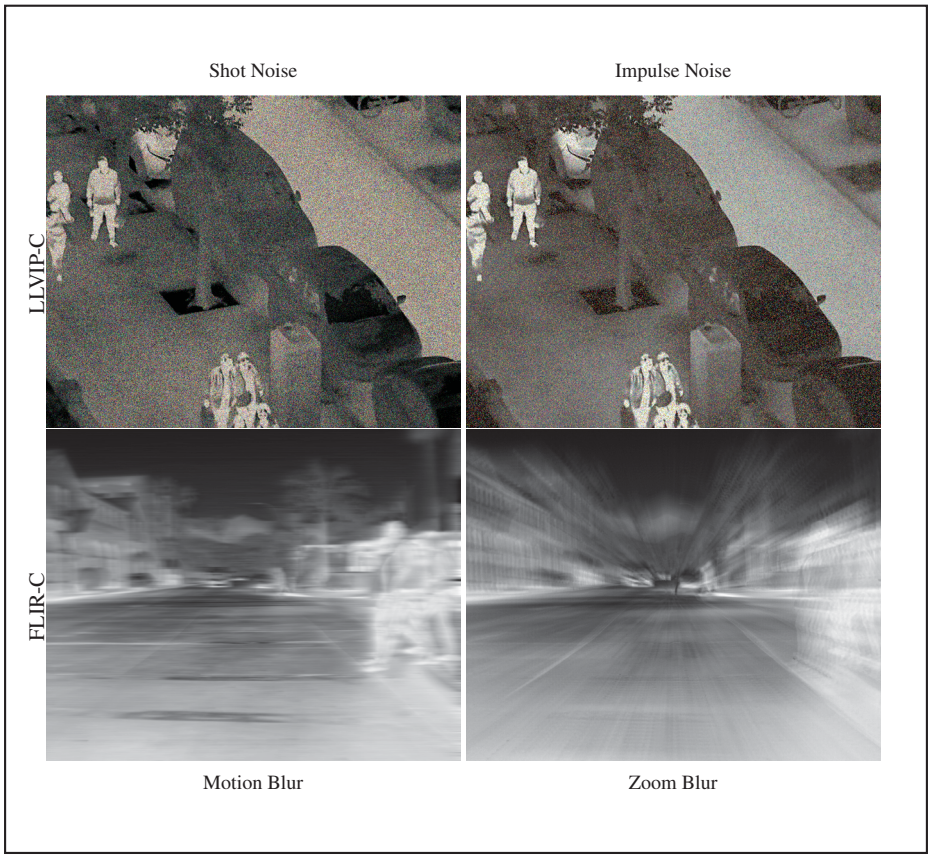


Figure-A V-2 LLVIP-C and FLIR-C examples. First row, we have one example from the LLVIP-C test set with two different corruptions: Shot Noise, and Impulse Noise with a severity level of 5. In the second row, we have one example from the FLIR-C test set with Motion Blur and Zoom Blur with a severity level of 5

4.1 Benchmark Datasets

For our proposed robust IR OD benchmark, we explore two classical datasets containing paired RGB and infrared images: LLVIP and FLIR.

LLVIP: The LLVIP dataset is a surveillance dataset composed of $12,025$ paired IR and RGB images for training and $3,463$ paired IR and RGB images for testing. The resolution of images is 1280×1024 pixels, and annotations consist of bounding boxes around pedestrians.

FLIR ALIGNED: For the FLIR dataset, we used the sanitized and aligned paired sets provided (Zhang *et al.*, 2020b), which contains $4,129$ paired IR and RGB images for training, and $1,013$ paired IR and RGB images for testing. The FLIR images are captured by a front-mounted car camera at a resolution of 640×512 pixels, and annotations contain bicycles, dogs, cars, and people.

LLVIP-C and FLIR-C: In this section, we present our two corrupted benchmarks: LLVIP-C and FLIR-C, derived from the LLVIP and FLIR datasets. In Figure V-2, the first row shows an LLVIP-C test example corrupted with Shot Noise and Impulse Noise at severity level 5. The second row shows an FLIR-C test example corrupted with Motion Blur and Zoom Blur at severity level 5. As illustrated qualitatively, severity level 5 is too strong for the FLIR images, already compressed JPEGs, and both zero-shot and fine-tuned models perform worse on FLIR-C than on LLVIP-C at this level. Therefore, we recommend a maximum corruption severity of 2 for FLIR-C based on qualitative and quantitative results. For the following experiments, we use severity level 5 for LLVIP-C and severity level 2 for FLIR-C.

5. WiSE-OD

Our proposed method, WiSE-OD (f_{wod}) in Figure V-3, extends the idea of WiSE-FT to object detection setting. Let $\theta_{\text{RGB}}^{\text{COCO}}$ and $\theta_{\text{IR}}^{\text{FT}}$ denote the parameters of the RGB pretrained COCO detector and the fully fine-tuned IR detection models, respectively. WiSE-OD constructs a new detector by interpolating these parameter vectors in weight space:

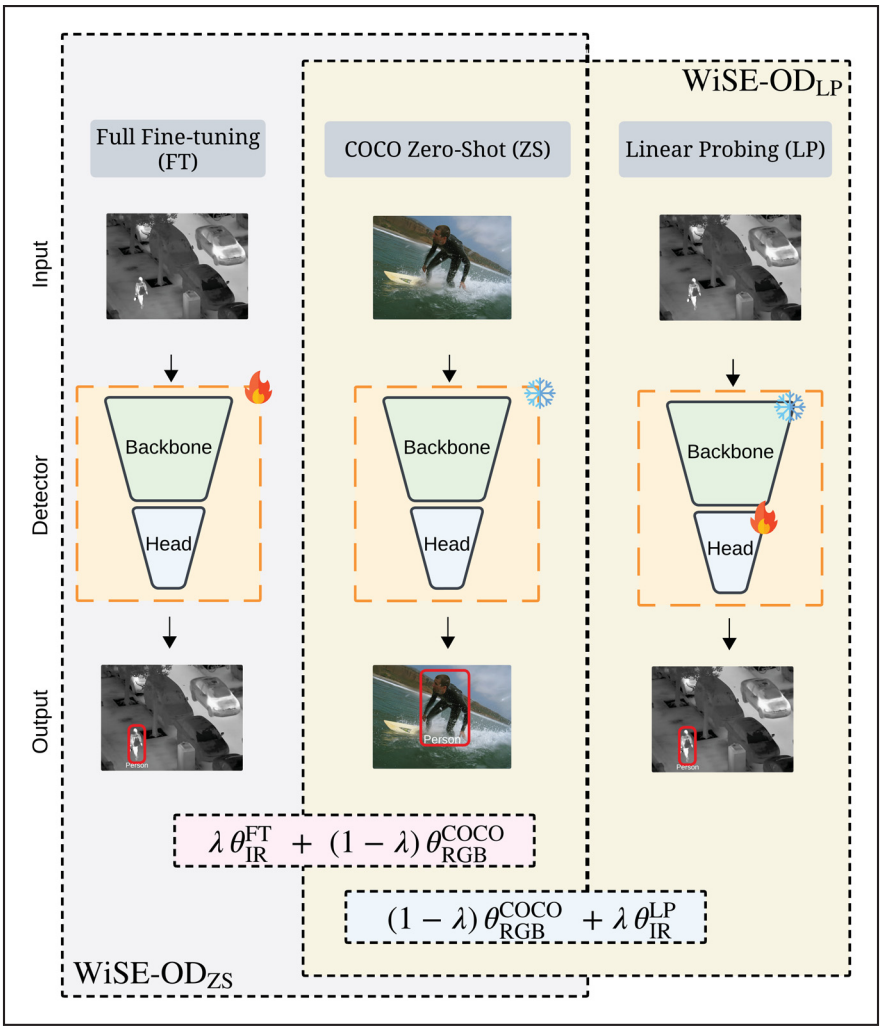


Figure-A V-3 Our proposed method: WiSE-OD and its variants. In the large grey box, we have WiSE-OD$_{ZS}$ with the equation inside the pink square, and WiSE-OD$_{LP}$ in the yellow large box with the equation inside the blue square

$$f_{\text{wod}}(\theta^{\text{COCO}}_{\text{RGB}}, \theta^{\text{FT}}_{\text{IR}}; \lambda) \;=\; (1-\lambda)\,\theta^{\text{COCO}}_{\text{RGB}} \;+\; \lambda\,\theta^{\text{FT}}_{\text{IR}}. \tag{A V-3}$$

The resulting interpolated model inherits both the broad generalization of large-scale COCO pretraining and the modality-specific accuracy of IR fine-tuning, yet requires no extra modules or change to the inference pipeline, only a one-time weight merge. We evaluate two variants: WiSE-OD$_{ZS}$ uses $\theta^{\text{FT}}_{\text{IR}}$ (full fine-tuning), and WiSE-OD$_{LP}$ uses $\theta^{\text{LP}}_{\text{IR}}$ (linear probing on the

detection head with a frozen backbone). Both variants consistently improve robustness under domain shift and common corruptions, while maintaining the same inference cost as a single detector. This weight-space ensembling is model-agnostic and can be extended to fuse multiple checkpoints or modalities by hierarchical interpolation.

Metrics: Following the methodology for benchmarking robustness in OD (Michaelis *et al.*, 2019), we select AP_{50} as our detection performance metric for both LLVIP-C and FLIR-C evaluations. We also report the dataset-specific performance (P), defined as the AP_{50} on the original target dataset (infrared), and the mean performance under corruption, mPC, defined as:

$$mPC = \frac{1}{N_c} \sum_{c=1}^{N_c} P_c, \tag{A V-4}$$

where P_c is the AP_{50} under corruption, and mPC is the average over all the corruptions. In our case, $N_c = 14$, we decided to remove the glass blur corruption because it lacks fast implementation for our benchmark.

Baseline models: In our study, we utilize three OD architectures: Faster R-CNN, FCOS, and RetinaNet, all initialized with COCO pre-trained weights. These models are trained on the COCO dataset, which contains 80 object categories, providing strong initial performance for most detection tasks and facilitating subsequent fine-tuning. We evaluate the following robust fine-tuning methods using our proposed benchmark:

1. **Zero-Shot (ZS) –** Unmodified detectors used directly for deployment without any fine-tuning.
2. **Linear probing (LP) –** Train the classification and regression heads on top of a frozen backbone by minimizing the detection loss.
3. **Full fine-tuning (FT) –** Update both the detection heads and the backbone parameters by minimizing the detection loss.
4. **LP-FT –** A two-stage process in which we first apply linear probing and then perform full fine-tuning initialized from the LP stage.

Table-A V-1 AP_{50} performance over the pertubations on different datasets. For LLVIP-C with severity level 5, and FLIR-C with severity level 2 for Faster R-CNN

	LLVIP-C					
	Zero-Shot	FT	LP	LP-FT	WiSE-OD$_{ZS}$	WiSE-OD$_{LP}$
Original	71.21 ± 0.02	93.68 ± 0.86	91.82 ± 0.15	92.18 ± 0.03	96.06 ± 0.22	96.24 ± 0.03
Gaussian Noise	59.24 ± 0.07	67.46 ± 7.45	75.12 ± 0.12	72.51 ± 0.28	86.68 ± 0.44	85.45 ± 0.76
Shot Noise	51.48 ± 0.14	64.83 ± 7.79	70.82 ± 0.27	69.89 ± 0.26	85.26 ± 0.50	85.25 ± 0.12
mPC	33.67	44.80	47.11	46.92	50.52	**51.59**

5. **Weight-ensembling –** Two variants, WiSE-OD$_{ZS}$ and WiSE-OD$_{LP}$, which interpolate parameters between the zero-shot or linear-probing models and the fully fine-tuned models, respectively.

Table-A V-2 Detection performance for the OD IR Robustness Benchmark. mPC metric for LLVIP-C with severity 5 and FLIR-C with severity level 2

Detector	**LLVIP-C**					
	Zero-Shot	FT	LP	LP-FT	WiSE-OD$_{ZS}$	WiSE-OD$_{LP}$
Faster R-CNN	40.96	56.40	70.96	70.02	75.08	**75.83**
FCOS	36.11	61.17	63.91	60.26	**76.50**	75.95
RetinaNet	37.50	61.13	60.37	61.11	73.69	**74.39**
Detector	**FLIR-C**					
	Zero-Shot	FT	LP	LP-FT	WiSE-OD$_{ZS}$	WiSE-OD$_{LP}$
Faster R-CNN	33.67	44.80	47.11	46.92	50.52	**51.59**
FCOS	28.85	41.07	38.92	38.14	**47.13**	46.76
RetinaNet	28.27	42.71	36.71	36.45	45.35	**47.53**

6. Experiments and Results

6.1 Training protocol

For this work, we split each training dataset into 80% for training and 20% for validation, reserving the original test set for final evaluation. All models were implemented in PyTorch, optimized with the Adam optimizer, and trained on an NVIDIA A100 GPU. We set a maximum training budget of 200 epochs for all detectors; in practice, fine-tuning typically converges within 10–20 epochs, depending on the model and dataset. We used a cosine annealing scheduler on the training loss and applied early stopping based on validation AP_{50}.

6.2 Benchmark Quantitative Results

In this section, we measured the mPC performance of all the proposed baselines for our benchmark on LLVIP-C with a severity level of 5 and the FLIR-C dataset with a severity level of 2. Results are shown in Table V-2 for Faster R-CNN, FCOS, and RetinaNet under zero-shot, FT, LP, LP-FT, WiSE-OD_{ZS}, and WiSE-OD_{LP}. We see in Table V-2, on average, WiSE-OD_{ZS} with λ fixed at 0.5, i.e., equal weighting of Zero-Shot and FT, outperforms all other baselines without the need to tune any hyperparameters. For instance, on LLVIP-C, WiSE-OD_{ZS} improved mPC by 18.68 over FT and by 4.12 over LP for Faster R-CNN. In most cases, our proposed variant WiSE-OD_{LP} outperformed WiSE-OD_{ZS} for Faster R-CNN and RetinaNet.

6.3 Detection performance per corruption

In this section, we evaluated the benchmark per corruption. For Table V-1, we show the in-domain performance (evaluation on infrared of the LLVIP dataset), which we named “Original”, which is the original LLVIP infrared test set. Then, we have the corruptions for the LLVIP-C and the mPC metric for the Faster R-CNN detector; the same methodology was used for FLIR and FLIR-C. The original performance is measured in terms of AP_{50} for Faster R-CNN; additional results for FCOS, RetinaNet, and all the detectors are provided in the supplementary material. As described in the Table V-1, the in-domain performance for LP (91.82) and LP-FT (92.18) is lower than the FT (93.63), but the mPC is much higher than the zero-shot and FT. For instance, we have 70.96 for LP and 70.02 mPC for LP-FT (LP-FT was lower in OOD compared to LP), but its in-domain performance for LP-FT was a bit higher than LP. The FT was able to beat LP and

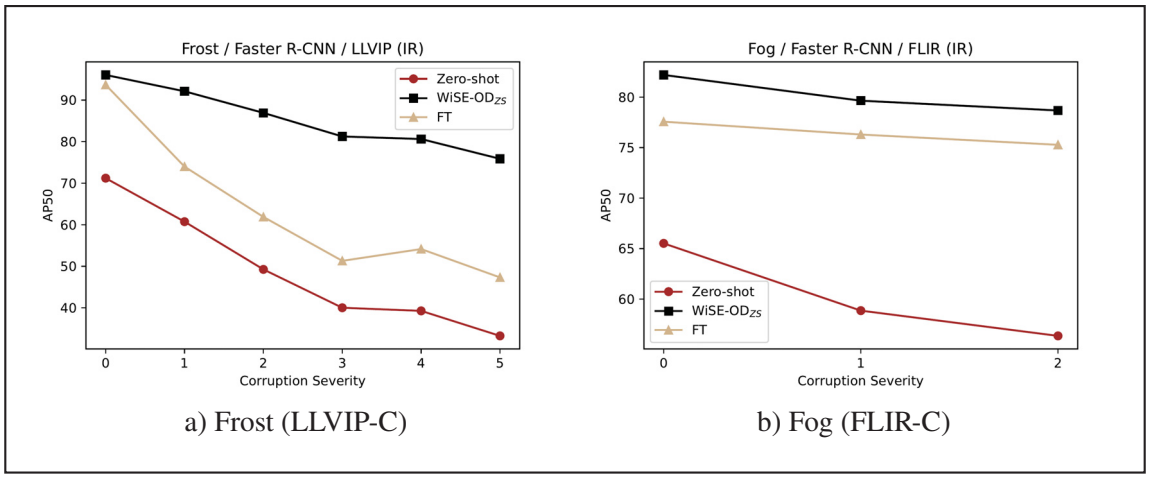
LP-FT in the in-domain performance, but the mPC for FT was 56.40, which is 14.56 lower than LP and 13.62 lower than LP-FT. The WiSE-OD$_{ZS}$ and WiSE-OD$_{LP}$ were able to outperform the others with in-domain of 96.06 and 96.24, respectively, and for the mPC, the WiSE-OD$_{LP}$ had 75.83 and the WiSE-OD$_{ZS}$ 75.08, which the WiSE-OD$_{ZS}$ is an increase of 18.68 from the original FT and 4.12 mPC over LP. For FLIR-C, we also have good improvements compared to the others. It is important to mention that the WiSE-OD$_{ZS}$ is a free-training technique, and for this table, λ is fixed at 0.5, same for WiSE-OD$_{LP}$, but this variation needs the LP model instead of the zero-shot.

6.4 Performance over different corruption levels

In this section, we measured the per AP$_{50}$ performance for Faster R-CNN, FCOS, and RetinaNet over different corruption severity levels for the benchmark. Here, in Figure V-4, we provided the Frost for LLVIP-C and Fog for FLIR-C for Faster R-CNN. When the corruption severity level increases, e.g., from 1 to 5 in LLVIP-C, we can see a large drop in the zero-shot and FT, while the WiSE-OD$_{ZS}$ is more stable and can bring more robustness to the final model. Some corruptions have more impact than others for each dataset; for instance, in FLIR-C, the noise corruption affected the performance more due to the original low-quality images. On IR modality, the contrast perturbation seems to affect the detection performance more because the images are already bad in contrast when compared to natural RGB images. A similar trend of stability of WiSE-OD$_{ZS}$ for other detectors and corruptions over zero-shot and FT is shown in the supp. materials.

6.5 Activation map analysis

In this section, we perform the qualitative analysis of the activation maps of the Faster R-CNN detector over the corruptions for the zero-shot model, WiSE-OD$_{ZS}$, and FT model over the corruptions on the LLVIP-C dataset. In Figure V-5, we can see the Grad-CAM (Selvaraju *et al.*, 2017) for impulse noise in the first row, then the zoom blur in the second row, with the ground-truth bounding box in red (additional activation maps figures for all the corruptions are



a) Frost (LLVIP-C) b) Fog (FLIR-C)

Figure-A V-4 AP_{50} performance for Faster R-CNN over different corruption severity levels for Frost (a) and Fog (b). For each perturbation, we evaluated different levels of corruption for the Zero-Shot, WiSE-OD$_{ZS}$, and FT models. Here, we provide the Frost on LLVIP-C and Fog on FLIR-C. Additional study in supp. materials

provided in supp. materials). We see that the FT model and zero-shot do not detect the person well for such images under the corruptions, while for WiSE-OD$_{ZS}$, the detector was able to activate the features for a person even with the corruption, which means that it is more robust for such cases. However, in general, the WiSE-OD$_{ZS}$ was able to activate more the person features in different corruptions than the FT; there are some corruption cases that it cannot detect as well as other ones; for such cases in which the user wants to perform better in a specific corruption for a real-world application, we recommend tuning the λ which can give more importance for the zero-shot weight or the FT depending on the corruption. Finally, we observed that WiSE-OD preserves more complete object regions in activation maps across both LLVIP-C and FLIR-C, particularly in corrupted images with low contrast or fog. This suggests that ensembling retains complementary semantic cues from both zero-shot priors and fine-tuned details, leading to more stable predictions under corruptions.

6.6 WiSE-OD$_{ZS}$: Ablation study on λ

In this section, we extensively conducted studies about the λ value to combine the zero-shot RGB COCO pre-training weights of Faster R-CNN, FCOS, and RetinaNet with the FT IR

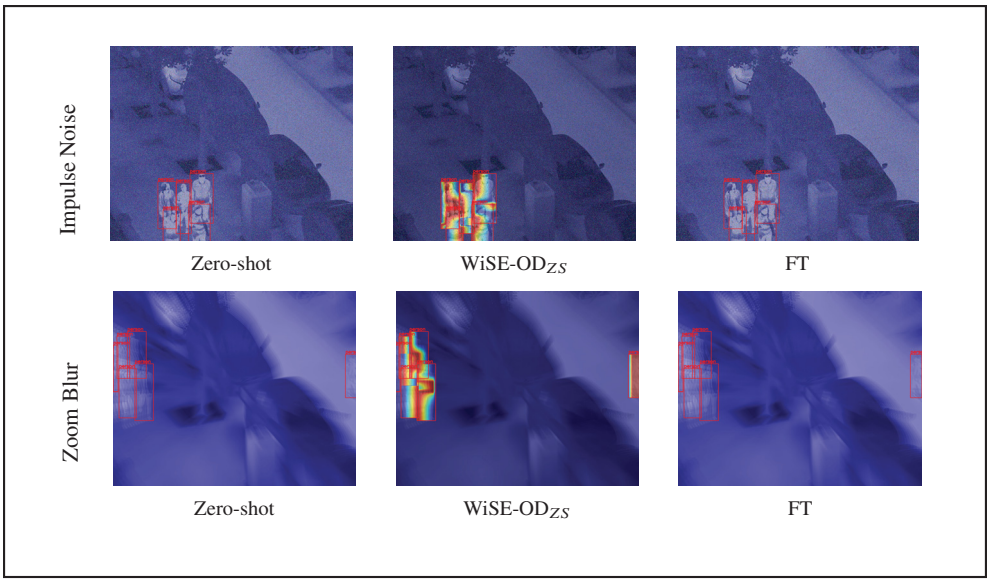


Figure-A V-5 Activation map analysis for zero-shot COCO pre-train Faster R-CNN detector, the WiSE-OD$_{ZS}$ and FT detector on IR for LLVIP-C dataset. In red are the GTs, and for WiSE-OD$_{ZS}$ the models can activate features that represent a person under corruptions. The first row is impulse noise, and the second row is zoom blur (severity 5)

under the respective datasets LLVIP-C and FLIR-C. Evaluating the performance of such weight ensembling WiSE-OD$_{ZS}$ under the different corruptions settings. Here, in the main manuscript, we show the results for Faster R-CNN in Table V-3 for some values of λ, and we provide the detailed ablation and additional results in the supplementary material. It is important to mention that in the rest of the main manuscript, λ was fixed to 0.5, while here, we wanted to further investigate the potential of the WiSE-OD$_{ZS}$ over the different corruptions. In Table V-3, the best results for in-domain performance was with $\lambda = 0.5$, while the best out-of-domain was $\lambda = 0.2$, which shows that for Faster R-CNN the zero-shot model can bring robustness for the model, but some corruptions such as pixelate are better when the λ is higher.

We observe that $\lambda = 0.5$ consistently provides a favorable trade-off, even for FLIR-C, where $\lambda = 0.5$ was the best ID. For LLVIP-C, WiSE-OD achieved the best mean performance under corruption ($\lambda = 0.2$, mPC 75.82), outperforming both the pure fine-tuned model ($\lambda = 1.0$, mPC 56.40) and the zero-shot model ($\lambda = 0.0$, mPC 40.96). Notably, $\lambda = 0.5$ performs best under

Table-A V-3 Ablation of λ over LLVIP-C and FLIR-C dataset for Faster R-CNN. Where $\lambda = 0.0$ represents the zero-shot model, $\lambda = 0.5$ represents default WiSE-OD$_{ZS}$ and $\lambda = 1.0$ represents the fine-tuning model. For LLVIP-C, the severity level is 5

	LLVIP-C				
	$\theta(\lambda = 0.0)$	$\theta(\lambda = 0.2)$	$\theta(\lambda = 0.5)$	$\theta(\lambda = 0.8)$	$\theta(\lambda = 1.0)$
Original	71.21 ± 0.02	93.88 ± 0.28	96.06 ± 0.22	95.41 ± 0.60	93.68 ± 0.86
Gaussian Noise	59.24 ± 0.07	86.52 ± 0.40	86.68 ± 0.44	78.47 ± 3.43	67.46 ± 7.45
Shot Noise	51.48 ± 0.14	83.86 ± 0.70	85.26 ± 0.50	76.42 ± 3.74	64.83 ± 7.79
Impulse Noise	56.62 ± 0.07	86.93 ± 0.65	88.54 ± 0.33	80.94 ± 2.70	71.32 ± 6.33
Defocus Blur	47.90 ± 0.08	88.41 ± 0.31	89.74 ± 0.98	85.85 ± 2.55	80.48 ± 3.60
Motion Blur	26.39 ± 0.23	81.10 ± 0.44	86.81 ± 0.71	83.24 ± 1.70	78.32 ± 3.18
Zoom Blur	02.47 ± 0.02	27.97 ± 0.62	22.83 ± 2.44	14.46 ± 1.82	11.18 ± 1.56
Snow	33.65 ± 0.01	67.67 ± 0.77	65.97 ± 1.90	38.50 ± 2.61	13.46 ± 4.45
Frost	33.25 ± 0.38	72.31 ± 0.23	75.87 ± 0.39	65.10 ± 0.57	47.32 ± 3.45
Fog	59.60 ± 0.10	89.79 ± 0.43	84.51 ± 3.80	64.47 ± 9.62	50.90 ± 10.0
Brightness	41.77 ± 0.03	82.38 ± 0.17	82.10 ± 1.20	62.81 ± 3.16	35.36 ± 6.97
Contrast	47.48 ± 0.04	50.59 ± 4.48	10.57 ± 3.82	00.77 ± 0.31	00.00 ± 0.00
Elastic transform	52.42 ± 0.18	89.92 ± 0.51	94.72 ± 0.07	94.32 ± 0.34	92.41 ± 0.93
Pixelate	03.95 ± 0.01	66.27 ± 3.14	85.06 ± 3.32	88.97 ± 2.35	87.69 ± 2.67
JPEG compression	57.22 ± 0.02	87.87 ± 0.89	92.59 ± 1.22	91.86 ± 1.38	88.93 ± 1.69
mPC	40.96	75.82	75.08	66.15	56.40
	FLIR-C				
	$\theta(\lambda = 0.0)$	$\theta(\lambda = 0.2)$	$\theta(\lambda = 0.5)$	$\theta(\lambda = 0.8)$	$\theta(\lambda = 1.0)$
Original	65.52 ± 0.07	77.49 ± 0.10	82.20 ± 0.07	80.18 ± 0.11	77.57 ± 0.24
Gaussian Noise	31.21 ± 0.29	42.62 ± 2.92	42.49 ± 4.48	34.85 ± 3.85	28.07 ± 2.91
Shot Noise	25.26 ± 0.12	33.91 ± 2.98	30.45 ± 3.96	21.88 ± 2.93	15.73 ± 2.05
Impulse Noise	17.69 ± 0.03	24.85 ± 2.26	22.51 ± 2.78	16.96 ± 2.17	13.22 ± 2.27
Defocus Blur	25.32 ± 0.22	44.57 ± 2.40	54.08 ± 1.74	55.00 ± 0.98	52.47 ± 0.99
Motion Blur	25.01 ± 0.25	40.63 ± 2.17	51.03 ± 2.16	53.85 ± 2.19	51.71 ± 2.12
Zoom Blur	08.98 ± 0.05	13.72 ± 0.97	16.93 ± 1.00	18.32 ± 1.09	17.97 ± 0.90
Snow	09.84 ± 0.14	14.55 ± 2.07	13.94 ± 2.66	10.36 ± 2.48	07.86 ± 2.01
Frost	21.96 ± 0.50	33.37 ± 2.39	37.97 ± 3.63	36.47 ± 4.00	33.87 ± 4.69
Fog	56.36 ± 0.28	72.17 ± 0.86	78.68 ± 1.24	78.11 ± 1.49	73.61 ± 0.06
Brightness	64.41 ± 0.26	75.68 ± 0.19	79.72 ± 0.35	77.79 ± 1.24	75.18 ± 0.99
Contrast	54.59 ± 0.04	71.38 ± 0.95	78.36 ± 1.06	78.02 ± 1.09	75.47 ± 1.29
Elastic transform	41.88 ± 0.24	63.89 ± 0.37	73.39 ± 0.40	72.51 ± 0.58	69.68 ± 1.15
Pixelate	38.67 ± 0.11	55.23 ± 2.37	61.12 ± 3.13	58.87 ± 4.58	54.91 ± 6.21
JPEG compression	50.24 ± 0.14	63.21 ± 0.56	66.65 ± 0.89	63.27 ± 2.36	57.55 ± 3.04
mPC	33.67	46.41	50.52	48.30	44.80

heavy corruptions such as Gaussian noise, Fog, and Brightness shifts, scenarios where both FT and ZS individually struggle. For example, under Fog and Brightness, $\lambda = 0.5$ yields 84.51 and

82.10, respectively, while $\lambda = 1.0$ achieves only 50.90 and 35.36. This highlights the benefit of WiSE-OD in preserving complementary robustness features from both models. Interestingly, $\lambda = 0.8$ performs well in several cases but shows more variability, suggesting that moderate ensembling (rather than heavily biasing toward FT) is more robust under distribution shifts. These results justify the use of $\lambda = 0.5$ as a robust default and motivate future work on adaptive λ selection strategies.

6.7 Limitations

Main Limitations. While WiSE-OD demonstrates robust performance across corruption types and detectors, it has certain limitations. First, the method assumes access to both a zero-shot model and a fine-tuned or linearly probed counterpart, which may not always be feasible in constrained deployment scenarios. Second, the fixed mixing coefficient λ, although effective at 0.5 in general, may not be optimal for all corruption types or datasets. As shown in our ablation study, tuning λ for specific perturbations can yield better results, suggesting that an adaptive λ selection mechanism could further enhance performance. Additionally, the method's effectiveness is tied to the quality of the pretrained models; for instance, if the zero-shot or fine-tuned models underperform due to poor initialization or training instability, the ensemble will inherit those weaknesses. Lastly, our study focuses on robustness to synthetic corruptions and may not fully capture all real-world distribution shifts, such as hardware-specific noise or extreme weather conditions.

Failure cases. Despite its strong average performance across corruption types, WiSE-OD is not without failure cases. We observe that its robustness can degrade under extreme conditions such as severe snow, heavy blur, or very low brightness, particularly in the FLIR-C dataset. In these settings, both the zero-shot and fine-tuned models tend to misfire, either missing objects entirely or generating fragmented predictions, causing the ensemble to inherit and even amplify these errors. Furthermore, in scenes with high thermal clutter or indistinct object boundaries, the fusion of weights may lead to uncertain activations and unstable bounding boxes. These failure

modes highlight the importance of complementary improvements such as corruption-aware ensembling or dynamic λ adjustment.

6.8 Final Discussion and Future Work

Our findings suggest that simple weight-space ensembling strategies can significantly improve robustness in cross-modality detection. However, one promising direction is to make these ensembles adaptive. Rather than using a fixed λ, one could explore data-driven mechanisms, e.g., predicting λ per image or corruption type using a lightweight auxiliary network. Another natural extension is to ensemble more than two model variants (e.g., LP, FT, multiple seeds) to form model soups for detection, which have shown improvements in classification. Additionally, while our current benchmark focuses on synthetic corruptions adapted from ImageNet-C, future benchmarks could incorporate real-world IR degradations, such as thermal blooming, lens fogging, or sensor-specific quantization artifacts. Moreover, exploring the integration of WiSE-OD with domain generalization methods or meta-learning approaches could yield more adaptive solutions in unseen IR environments. Finally, applying our method to other sensor modalities like depth, multispectral, or event-based data could reveal whether WiSE-OD generalizes beyond IR, positioning it as a lightweight alternative to domain adaptation and robustness training pipelines. Additionally, future works could explore adaptive or corruption-aware weight blending and extend the benchmark to broader domain shifts like day-night transitions or different IR sensors.

7. Conclusion

In this work, we presented a new benchmark for robustness IR OD based on the work of Hendrycks & Dietterich (2019), with the target of traditional IR datasets such as LLVIP and FLIR. Our new benchmark is a challenging setting for IR robustness with the introduction of LLVIP-C and FLIR-C. Furthermore, we did an extensive study of different robust fine-tuning strategies over our proposed benchmark. Additionally, we present the WiSE-OD method and its variations WiSE-OD$_{ZS}$ and WiSE-OD$_{LP}$, where both were able to surpass the traditional

robustness strategies while also increasing in-domain performance over different detectors, such as Faster R-CNN, FCOS, and RetinaNet. Our extensive study shows that a simple but WiSE-OD strategy can mitigate performance drop without any additional training cost.

Acknowledgments

This work was supported in part by Distech Controls Inc., the Natural Sciences and Engineering Research Council of Canada, the Digital Research Alliance of Canada, and MITACS.

BIBLIOGRAPHY

Agrawal, K. & Subramanian, A. (2019). Enhancing object detection in adverse conditions using thermal imaging. *arXiv preprint arXiv:1909.13551*.

Alehdaghi, M., Josi, A., Cruz, R. M. & Granger, E. (2022). Visible-infrared person re-identification using privileged intermediate information. *European Conference on Computer Vision*, pp. 720–737.

Alehdaghi, M., Josi, A., Cruz, R. M. & Granger, E. (2023). Visible-infrared person re-identification using privileged intermediate information. *Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part V*, pp. 720–737.

Alom, M. Z., Taha, T. M., Yakopcic, C., Westberg, S., Sidike, P., Nasrin, M. S., Van Esesn, B. C., Awwal, A. A. S. & Asari, V. K. (2018). The history began from alexnet: A comprehensive survey on deep learning approaches. *arXiv preprint arXiv:1803.01164*, .

Azad, R., Khosravi, N., Dehghanmanshadi, M., Cohen-Adad, J. & Merhof, D. (2022). Medical image segmentation on mri images with missing modalities: A review. *arXiv preprint arXiv:2203.06217*.

Bachmann, R., Mizrahi, D., Atanov, A. & Zamir, A. (2022). Multimae: Multi-modal multi-task masked autoencoders. *European Conference on Computer Vision*, pp. 348–367.

Bachmann, R., Kar, O. F., Mizrahi, D., Garjani, A., Gao, M., Griffiths, D., Hu, J., Dehghan, A. & Zamir, A. (2024). 4m-21: An any-to-any vision model for tens of tasks and modalities. *Advances in Neural Information Processing Systems*, 37, 61872–61911.

Bahng, H., Jahanian, A., Sankaranarayanan, S. & Isola, P. (2022). Exploring visual prompts for adapting large-scale models. *arXiv preprint arXiv:2203.17274*, .

Baltrušaitis, T., Ahuja, C. & Morency, L.-P. (2018). Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2), 423–443.

Bansal, A., Sikka, K., Sharma, G., Chellappa, R. & Divakaran, A. (2018). Zero-shot object detection. *Proceedings of the European conference on computer vision (ECCV)*, pp. 384–400.

Bayoudh, K., Knani, R., Hamdaoui, F. & Mtibaa, A. (2021). A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets. *The Visual Computer*, 38, 2939 - 2970. Retrieved from: https://api.semanticscholar.org/CorpusID:235410640.

Beghdadi, A., Mallem, M. & Beji, L. (2022). Benchmarking performance of object detection under image distortions in an uncontrolled environment. *2022 IEEE International Conference on Image Processing (ICIP)*, pp. 2071–2075.

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F. & Vaughan, J. W. (2010). A theory of learning from different domains. *Machine learning*, 79, 151–175.

Bhattarai, M. & Martinez-Ramon, M. (2020). A deep learning framework for detection of targets in thermal images to improve firefighting. *IEEE Access*, 8, 88308–88321.

Biewald, L. [Software available from wandb.com]. (2020). Experiment Tracking with Weights and Biases.

Boretti, C., Bich, P., Pareschi, F., Prono, L., Rovatti, R. & Setti, G. (2023, jun). PEDRo: an Event-based Dataset for Person Detection in Robotics. *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*.

Brehar, R., Vancea, F., Marita, T., Vancea, C. & Nedevschi, S. (2019). Object detection in monocular infrared images using classification–regresion deep learning architectures. *2019 IEEE 15th International Conference on Intelligent Computer Communication and Processing (ICCP)*, pp. 207–212.

Bustos, N., Mashhadi, M., Lai-Yuen, S. K., Sarkar, S. & Das, T. K. (2023). A systematic literature review on object detection using near infrared and thermal images. *Neurocomputing*, 126804.

Cai, Q., Ma, M., Wang, C. & Li, H. (2023). Image neural style transfer: A review. *Computers and Electrical Engineering*, 108, 108723.

Cao, Y., Zhou, T., Zhu, X. & Su, Y. (2019). Every feature counts: An improved one-stage detector in thermal imagery. *2019 IEEE 5th International Conference on Computer and Communications (ICCC)*, pp. 1965–1969.

Cao, Y., Bin, J., Hamari, J., Blasch, E. & Liu, Z. (2023a). Multimodal Object Detection by Channel Switching and Spatial Attention. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 403-411. doi: 10.1109/CVPRW59228.2023.00046.

Cao, Y., Bin, J., Hamari, J., Blasch, E. & Liu, Z. (2023b). Multimodal Object Detection by Channel Switching and Spatial Attention. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 403–411.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A. & Zagoruyko, S. (2020). End-to-end object detection with transformers. *European conference on computer vision*, pp. 213–229.

Chaurasia, A. & Culurciello, E. (2017). Linknet: Exploiting encoder representations for efficient semantic segmentation. *2017 IEEE Visual Communications and Image Processing (VCIP)*, pp. 1–4.

Chebrolu, K. N. R. & Kumar, P. (2019). Deep learning based pedestrian detection at all light conditions. *2019 International Conference on Communication and Signal Processing (ICCSP)*, pp. 0838–0842.

Chelba, C. & Acero, A. (2006). Adaptation of maximum entropy capitalizer: Little data can help a lot. *Computer Speech and Language*, 20(4), 382-399. doi: https://doi.org/10.1016/j.csl.2005.05.005.

Chen, C., Zheng, Z., Ding, X., Huang, Y. & Dou, Q. (2020a). Harmonizing transferability and discriminability for adapting object detectors. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8869–8878.

Chen, J., Li, K., Deng, Q., Li, K. & Yu, P. S. (2019a). Distributed deep learning model for intelligent video surveillance systems with edge computing. *IEEE Transactions on Industrial Informatics*, .

Chen, K., Wang, J., Pang, J., Cao, Y., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Xu, J. et al. (2019b). MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, .

Chen, L.-C., Papandreou, G., Schroff, F. & Adam, H. (2017). Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, .

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F. & Adam, H. (2018a). Encoder-decoder with atrous separable convolution for semantic image segmentation. *Proceedings of the European conference on computer vision (ECCV)*, pp. 801–818.

Chen, S., Hou, Y., Cui, Y., Che, W., Liu, T. & Yu, X. (2020b). Recall and learn: Fine-tuning deep pretrained language models with less forgetting. *arXiv preprint arXiv:2004.12651*, .

Chen, X., Xu, C., Yang, X., Song, L. & Tao, D. (2018b). Gated-gan: Adversarial gated networks for multi-collection style transfer. *IEEE Transactions on Image Processing*, 28(2), 546–560.

Chen, Y.-T., Shi, J., Ye, Z., Mertz, C., Ramanan, D. & Kong, S. (2022). Multimodal Object Detection via Probabilistic Ensembling.

Chen, Y., Li, W., Sakaridis, C., Dai, D. & Van Gool, L. (2018c). Domain adaptive faster r-cnn for object detection in the wild. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3339–3348.

Cheng, T., Song, L., Ge, Y., Liu, W., Wang, X. & Shan, Y. (2024). Yolo-world: Real-time open-vocabulary object detection. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16901–16911.

Choi, H., Kim, S., Park, K. & Sohn, K. (2016). Multi-spectral pedestrian detection based on accumulated object proposal with fully convolutional networks. *2016 23rd International Conference on Pattern Recognition (ICPR)*, pp. 621–626.

Chowdhary, K. & Chowdhary, K. (2020). Natural language processing. *Fundamentals of artificial intelligence*, 603–649.

Cong, P., Zhu, X., Qiao, F., Ren, Y., Peng, X., Hou, Y., Xu, L., Yang, R., Manocha, D. & Ma, Y. (2022). Stcrowd: A multimodal dataset for pedestrian perception in crowded scenes. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19608–19617.

Dai, D. & Van Gool, L. (2018). Dark model adaptation: Semantic image segmentation from daytime to nighttime. *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pp. 3819–3824.

Dai, Z., Cai, B., Lin, Y. & Chen, J. (2021). Up-detr: Unsupervised pre-training for object detection with transformers. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1601–1610.

Dalal, N. & Triggs, B. (2005). Histograms of oriented gradients for human detection. *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, 1, 886–893.

Danaci, K. I. & Akagunduz, E. (2022). A survey on infrared image and video sets. *arXiv preprint arXiv:2203.08581*.

Das, A., Das, S., Sistu, G., Horgan, J., Bhattacharya, U., Jones, E., Glavin, M. & Eising, C. (2023). Revisiting Modality Imbalance In Multimodal Pedestrian Detection.

Dayarathna, T., Muthukumarana, T., Rathnayaka, Y., Denman, S., de Silva, C., Pemasiri, A. & Ahmedt-Aristizabal, D. (2023). Privacy-preserving in-bed pose monitoring: A fusion and reconstruction study. *Expert Systems with Applications*, 213, 119139.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K. & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255.

Detlefsen, N. S., Borovec, J., Schock, J., Jha, A. H., Koker, T., Di Liello, L., Stancl, D., Quan, C., Grechkin, M. & Falcon, W. (2022). TorchMetrics-Measuring Reproducibility in PyTorch. *Journal of Open Source Software*, 7(70), 4101.

Devaguptapu, C., Akolekar, N., M Sharma, M. & N Balasubramanian, V. (2019). Borrow from anywhere: Pseudo multi-modal object detection in thermal imagery. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 0–0.

Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S. et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, .

Dou, Q., Liu, Q., Heng, P. A. & Glocker, B. (2020). Unpaired multi-modal segmentation via knowledge distillation. *IEEE transactions on medical imaging*, 39(7), 2415–2425.

Drenkow, N., Sani, N., Shpitser, I. & Unberath, M. (2021). A systematic review of robustness in deep learning for computer vision: Mind the gap? *arXiv preprint arXiv:2112.00639*.

Du, C., Li, T., Liu, Y., Wen, Z., Hua, T., Wang, Y. & Zhao, H. (2021). Improving Multi-Modal Learning with Uni-Modal Teachers.

Dubail, T., Guerrero Peña, F. A., Medeiros, H. R., Aminbeidokhti, M., Granger, E. & Pedersoli, M. (2022). Privacy-preserving person detection using low-resolution infrared cameras. *European Conference on Computer Vision*, pp. 689–702.

Eitel, A., Springenberg, J. T., Spinello, L., Riedmiller, M. & Burgard, W. (2015). Multimodal deep learning for robust RGB-D object recognition. *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 681–687.

Falcon, W. & The PyTorch Lightning team. (2019). Pytorch lightning.

Fan, T., Wang, G., Li, Y. & Wang, H. (2020). Ma-net: A multi-scale attention network for liver and tumor segmentation. *IEEE Access*, 8, 179656–179665.

Fang, Y., Liao, B., Wang, X., Fang, J., Qi, J., Wu, R., Niu, J. & Liu, W. (2021). You only look at one sequence: Rethinking transformer in vision through object detection. *Advances in Neural Information Processing Systems*, 34, 26183–26197.

Feichtenhofer, C., Li, Y., He, K. et al. (2022). Masked autoencoders as spatiotemporal learners. *Advances in neural information processing systems*, 35, 35946–35958.

Felzenszwalb, P., McAllester, D. & Ramanan, D. (2008). A discriminatively trained, multiscale, deformable part model. *2008 IEEE conference on computer vision and pattern recognition*, pp. 1–8.

Feng, C., Zhong, Y., Gao, Y., Scott, M. R. & Huang, W. (2021). Tood: Task-aligned one-stage object detection. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3490–3499.

Feng, H., Yang, Z., Chen, H., Pang, T., Du, C., Zhu, M., Chen, W. & Yan, S. (2023). CoSDA: Continual Source-Free Domain Adaptation.

Flir, T. (2021). Free flir thermal dataset for algorithm training. Teledyne FLIR LLC All rights reserved.

Fokkinga, E. P., Eker, T. A., van Woerden, J. E., Witon, J.-M., Stallinga, S. O., Visser, A., Schutte, K. & Heslinga, F. G. (2025). Generative AI methods for synthesis of image data to train AI for automated scene understanding in a military context: a review of opportunities. *Synthetic Data for Artificial Intelligence and Machine Learning: Tools, Techniques, and Applications III*, 13459, 9–31.

French, R. M. (1999). Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences*, 3(4), 128–135.

Ganin, Y. & Lempitsky, V. (2015). Unsupervised Domain Adaptation by Backpropagation.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., March, M. & Lempitsky, V. (2016). Domain-adversarial training of neural networks. *Journal of machine learning research*, 17(59), 1–35.

Gao, P., Geng, S., Zhang, R., Ma, T., Fang, R., Zhang, Y., Li, H. & Qiao, Y. (2024). Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, 132(2), 581–595.

Garillos-Manliguez, C. A. & Chiang, J. Y. (2021). Multimodal deep learning and visible-light and hyperspectral imaging for fruit maturity estimation. *Sensors*, 21(4), 1288.

Gatys, L. A., Ecker, A. S. & Bethge, M. (2016). Image style transfer using convolutional neural networks. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2414–2423.

Gauen, K., Dailey, R., Laiman, J., Zi, Y., Asokan, N., Lu, Y.-H., Thiruvathukal, G. K., Shyu, M.-L. & Chen, S.-C. (2017). Comparison of visual datasets for machine learning. *2017 IEEE International Conference on Information Reuse and Integration (IRI)*, pp. 346–355.

Geng, X., Liu, H., Lee, L., Schuurmans, D., Levine, S. & Abbeel, P. (2022). Multimodal masked autoencoders learn transferable representations. *arXiv preprint arXiv:2205.14204*, .

Ghenescu, V., Barnoviciu, E., Carata, S.-V., Ghenescu, M., Mihaescu, R. & Chindea, M. (2018). Object recognition on long range thermal image using state of the art dnn. *2018 Conference Grid, Cloud & High Performance Computing in Science (ROLCG)*, pp. 1–4.

Girdhar, R., Singh, M., Ravi, N., Van Der Maaten, L., Joulin, A. & Misra, I. (2022). Omnivore: A single model for many visual modalities. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16102–16112.

Girdhar, R., El-Nouby, A., Singh, M., Alwala, K. V., Joulin, A. & Misra, I. (2023). Omnimae: Single model masked pretraining on images and videos. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10406–10417.

Girshick, R. (2015). Fast r-cnn. *Proceedings of the IEEE international conference on computer vision*, pp. 1440–1448.

Girshick, R., Donahue, J., Darrell, T. & Malik, J. (2015). Region-based convolutional networks for accurate object detection and segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 38(1), 142–158.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. & Bengio, Y. (2014). Generative adversarial nets. *Advances in neural information processing systems*, 27, 1-9.

Group, F. et al. (2018). Flir thermal dataset for algorithm training.

Gu, X., Lin, T.-Y., Kuo, W. & Cui, Y. (2022). Open-vocabulary Object Detection via Vision and Language Knowledge Distillation. *International Conference on Learning Representations*.

Gupta, A., Dollar, P. & Girshick, R. (2019). Lvis: A dataset for large vocabulary instance segmentation. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5356–5364.

He, K., Zhang, X., Ren, S. & Sun, J. (2015). Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 37(9), 1904–1916.

He, K., Zhang, X., Ren, S. & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P. & Girshick, R. (2022). Masked autoencoders are scalable vision learners. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16000–16009.

Hendrycks, D. & Dietterich, T. (2019). Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. *International Conference on Learning Representations*.

Herrmann, C., Ruf, M. & Beyerer, J. (2018). CNN-based thermal infrared person detection by domain adaptation. *Autonomous Systems: Sensors, Vehicles, Security, and the Internet of Everything*, 10643, 1064308.

Hicsonmez, S., Samet, N., Akbas, E. & Duygulu, P. (2020). GANILLA: Generative adversarial networks for image to illustration translation. *Image and Vision Computing*, 95, 103886.

Hinton, G. E. & Zemel, R. (1993). Autoencoders, minimum description length and Helmholtz free energy. *Advances in neural information processing systems*, 6, .

Hoffman, J., Gupta, S. & Darrell, T. (2016). Learning with side information through modality hallucination. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 826–834.

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M. & Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, .

Howard, J. & Ruder, S. (2018). Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*, .

Hsu, H.-K., Yao, C.-H., Tsai, Y.-H., Hung, W.-C., Tseng, H.-Y., Singh, M. & Yang, M.-H. (2020). Progressive domain adaptation for object detection. *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 749–757.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L. & Chen, W. (2022). LoRA: Low-Rank Adaptation of Large Language Models. *International Conference on Learning Representations*. Retrieved from: https://openreview.net/forum?id=nZeVKeeFYf9.

Hu, J., Shen, L. & Sun, G. (2018). Squeeze-and-excitation networks. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7132–7141.

Huang, X. & Belongie, S. (2017). Arbitrary style transfer in real-time with adaptive instance normalization. *Proceedings of the IEEE international conference on computer vision*, pp. 1501–1510.

Huang, Y. & Chen, Y. (2020). Survey of state-of-art autonomous driving technologies with deep learning. *2020 IEEE 20th international conference on software quality, reliability and security companion (QRS-C)*, pp. 221–228.

Hwang, S., Park, J., Kim, N., Choi, Y. & Kweon, I. S. (2015). Multispectral Pedestrian Detection: Benchmark Dataset and Baselines. *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Iakubovskii, P. (2019). Segmentation Models Pytorch. GitHub.

Inoue, N., Furuta, R., Yamasaki, T. & Aizawa, K. (2018). Cross-domain weakly-supervised object detection through progressive domain adaptation. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5001–5009.

Isola, P., Zhu, J.-Y., Zhou, T. & Efros, A. A. (2017a). Image-to-Image Translation with Conditional Adversarial Networks. *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*.

Isola, P., Zhu, J.-Y., Zhou, T. & Efros, A. A. (2017b). Image-to-image translation with conditional adversarial networks. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1125–1134.

Ivorra, E., Ortega, M., Alcañiz, M. & Garcia-Aracil, N. (2018). Multimodal computer vision framework for human assistive robotics. *2018 Workshop on Metrology for Industry 4.0 and IoT*, pp. 1–5.

Izmailov, P., Podoprikhin, D., Garipov, T., Vetrov, D. & Wilson, A. G. (2018). Averaging weights leads to wider optima and better generalization. *arXiv preprint arXiv:1803.05407*.

Jaccard, P. (1901). Étude comparative de la distribution florale dans une portion des Alpes et des Jura. *Bull Soc Vaudoise Sci Nat*, 37, 547–579.

Janssens, O., Van de Walle, R., Loccufier, M. & Van Hoecke, S. (2017). Deep learning for infrared thermal image based machine health monitoring. *IEEE/ASME Transactions on Mechatronics*, 23(1), 151–159.

Jia, M., Tang, L., Chen, B.-C., Cardie, C., Belongie, S., Hariharan, B. & Lim, S.-N. (2022). Visual prompt tuning. *European Conference on Computer Vision*, pp. 709–727.

Jia, X., Zhu, C., Li, M., Tang, W. & Zhou, W. (2021). LLVIP: A Visible-infrared Paired Dataset for Low-light Vision. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3496–3504.

Jing, C., Potgieter, J., Noble, F. & Wang, R. (2017). A comparison and analysis of RGB-D cameras' depth performance for robotics application. *2017 24th International Conference on Mechatronics and Machine Vision in Practice (M2VIP)*, pp. 1–6.

Jing, Y., Yang, Y., Feng, Z., Ye, J., Yu, Y. & Song, M. (2019). Neural style transfer: A review. *IEEE transactions on visualization and computer graphics*, 26(11), 3365–3385.

Jocher, G., Chaurasia, A., Stoken, A., Borovec, J., Kwon, Y., Fang, J., Michael, K., Montes, D., Nadar, J., Skalski, P. et al. (2022). ultralytics/yolov5: v6. 1-tensorrt, tensorflow edge tpu and openvino export and inference.

Jocher, G., Chaurasia, A. & Qiu, J. (2023). Ultralytics YOLOv8 (Version 8.0.0). Retrieved from: https://github.com/ultralytics/ultralytics.

Johnson, J., Alahi, A. & Fei-Fei, L. (2016). Perceptual losses for real-time style transfer and super-resolution. *European conference on computer vision*, pp. 694–711.

Josi, A., Alehdaghi, M., Cruz, R. M. & Granger, E. (2023a). Fusion for visual-infrared person ReID in real-world surveillance using corrupted multimodal data. *arXiv preprint arXiv:2305.00320*.

Josi, A., Alehdaghi, M., Cruz, R. M. & Granger, E. (2023b). Multimodal data augmentation for visual-infrared person ReID with corrupted data. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 32–41.

Kamath, A., Singh, M., LeCun, Y., Synnaeve, G., Misra, I. & Carion, N. (2021). MDETR – Modulated Detection for End-to-End Multi-Modal Understanding. Retrieved from: https://arxiv.org/abs/2104.12763.

Kawaguchi, K., Kaelbling, L. P. & Bengio, Y. (2017). Generalization in deep learning. *arXiv preprint arXiv:1710.05468*, 1(8).

Kemker, R., McClure, M., Abitino, A., Hayes, T. & Kanan, C. (2018). Measuring catastrophic forgetting in neural networks. *Proceedings of the AAAI conference on artificial intelligence*, 32(1), 1-9.

Khodabandeh, M., Vahdat, A., Ranjbar, M. & Macready, W. G. (2019). A robust learning approach to domain adaptive object detection. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 480–490.

Kieu, M., Bagdanov, A. D. & Bertini, M. (2021). Bottom-up and layerwise domain adaptation for pedestrian detection in thermal images. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 17(1), 1–19.

Kim, S., Choi, J., Kim, T. & Kim, C. (2019a). Self-training and adversarial background regularization for unsupervised domain adaptive one-stage object detection. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6092–6101.

Kim, T., Jeong, M., Kim, S., Choi, S. & Kim, C. (2019b). Diversify and match: A domain adaptive representation learning paradigm for object detection. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12456–12465.

Kingma, D. P. & Welling, M. (2022). Auto-Encoding Variational Bayes.

Kini, J., Fleischer, S., Dave, I. & Shah, M. (2023). Egocentric RGB+ Depth Action Recognition in Industry-Like Settings. *arXiv preprint arXiv:2309.13962*, .

Kirillov, A., He, K., Girshick, R. & Dollár, P. (2017). A unified architecture for instance and semantic segmentation.

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A. et al. (2017). Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13), 3521–3526.

Knyaz, V. (2019). Multimodal data fusion for object recognition. *Multimodal Sensing: Technologies and Applications*, 11059, 198–209.

Kong, X. & Ge, Z. (2021). Deep learning of latent variable models for industrial process monitoring. *IEEE Transactions on Industrial Informatics*, 18(10), 6778–6788.

Krizhevsky, A., Sutskever, I. & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, .

Kumar, A., Raghunathan, A., Jones, R. M., Ma, T. & Liang, P. (2022). Fine-Tuning can Distort Pretrained Features and Underperform Out-of-Distribution. *International Conference on Learning Representations*.

Lambert, J., Sener, O. & Savarese, S. (2018). Deep learning under privileged information using heteroscedastic dropout. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8886–8895.

LeCun, Y., Bengio, Y. & Hinton, G. (2015). Deep learning. *nature*, 521(7553), 436–444.

Lee, C., Cho, K. & Kang, W. (2020). Mixout: Effective Regularization to Finetune Large-scale Pretrained Language Models.

Lee, S., Park, J. & Park, J. (2024). CrossFormer: Cross-guided attention for multi-modal object detection. *Pattern Recognition Letters*, 179, 144–150.

Lester, B., Al-Rfou, R. & Constant, N. (2021). The Power of Scale for Parameter-Efficient Prompt Tuning. Retrieved from: https://arxiv.org/abs/2104.08691.

Li, C., Song, D., Tong, R. & Tang, M. (2018a). Multispectral pedestrian detection via simultaneous detection and segmentation. *arXiv preprint arXiv:1808.04818*, .

Li, F., Zhang, H., Liu, S., Guo, J., Ni, L. M. & Zhang, L. (2022a). Dn-detr: Accelerate detr training by introducing query denoising. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13619–13627.

Li, H., Xiong, P., An, J. & Wang, L. (2018b). Pyramid attention network for semantic segmentation. *arXiv preprint arXiv:1805.10180*, .

Li, H. & Wu, X.-J. (2018). DenseFuse: A fusion approach to infrared and visible images. *IEEE Transactions on Image Processing*, 28(5), 2614–2623.

Li, L. H., Zhang, P., Zhang, H., Yang, J., Li, C., Zhong, Y., Wang, L., Yuan, L., Zhang, L., Hwang, J.-N. et al. (2022b). Grounded language-image pre-training. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10965–10975.

Li, R., Jiao, Q., Cao, W., Wong, H.-S. & Wu, S. (2020a). Model adaptation: Unsupervised domain adaptation without source data. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9641–9650.

Li, S., Li, Y., Li, Y., Li, M. & Xu, X. (2021). Yolo-firi: Improved yolov5 for infrared image object detection. *IEEE access*, 9, 141861–141875.

Li, W., Li, F., Luo, Y., Wang, P. et al. (2020b). Deep domain adaptive object detection: A survey. *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 1808–1813.

Li, W., Peng, Y., Zhang, M., Ding, L., Hu, H. & Shen, L. (2023a). Deep model fusion: A survey. *arXiv preprint arXiv:2309.15698*.

Li, Y., Ge, Z., Yu, G., Yang, J., Wang, Z., Shi, Y., Sun, J. & Li, Z. (2023b). Bevdepth: Acquisition of reliable depth for multi-view 3d object detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(2), 1477–1485.

Li, Y. & Ibanez-Guzman, J. (2020). Lidar for autonomous driving: The principles, challenges, and trends for automotive lidar and perception systems. *IEEE Signal Processing Magazine*, 37(4), 50–61.

Lin, C., Sun, P., Jiang, Y., Luo, P., Qu, L., Haffari, G., Yuan, Z. & Cai, J. (2023). Learning Object-Language Alignments for Open-Vocabulary Object Detection. *The Eleventh International Conference on Learning Representations*.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P. & Zitnick, C. L. (2014). Microsoft coco: Common objects in context. *European conference on computer vision*, pp. 740–755.

Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C. L. & Dollár, P. (2015). Microsoft COCO: Common Objects in Context.

Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B. & Belongie, S. (2017a). Feature pyramid networks for object detection. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2117–2125.

Lin, T.-Y., Goyal, P., Girshick, R., He, K. & Dollár, P. (2017b). Focal loss for dense object detection. *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988.

Liu, L., Ouyang, W., Wang, X., Fieguth, P., Chen, J., Liu, X. & Pietikäinen, M. (2020). Deep learning for generic object detection: A survey. *International journal of computer vision*, 128, 261–318.

Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H. et al. (2024a). Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *European Conference on Computer Vision*, pp. 38–55.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y. & Berg, A. C. (2016). Ssd: Single shot multibox detector. *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pp. 21–37.

Liu, X., Feng, Y., Hu, S., Yuan, X. & Fan, H. (2024b). Benchmarking the Robustness of UAV Tracking Against Common Corruptions. *2024 IEEE 7th International Conference on Multimedia Information Processing and Retrieval (MIPR)*, pp. 465–470.

Liu, Y., Zhang, Z. & Zhang, L. (2021a). Depth Privileged Object Detection with Depth-Enhanced DCN. *International Conference on Neural Information Processing*, pp. 438–446.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S. & Guo, B. (2021b). Swin transformer: Hierarchical vision transformer using shifted windows. *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022.

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T. & Xie, S. (2022). A convnet for the 2020s. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11976–11986.

Liu, Z., Xu, Z., Jin, J., Shen, Z. & Darrell, T. (2023). Dropout Reduces Underfitting. *International Conference on Machine Learning*.

Loshchilov, I. & Hutter, F. (2019). Decoupled Weight Decay Regularization. *International Conference on Learning Representations*.

Luo, Y., Remillard, J. & Hoetzer, D. (2010). Pedestrian detection in near-infrared night vision system. *2010 IEEE Intelligent Vehicles Symposium*, pp. 51–58.

Mao, X., Chen, Y., Zhu, Y., Chen, D., Su, H., Zhang, R. & Xue, H. (2023). Coco-o: A benchmark for object detectors under natural distribution shifts. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6339–6350.

McMahan, B., Moore, E., Ramage, D., Hampson, S. & y Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. *Artificial intelligence and statistics*, pp. 1273–1282.

Medeiros, H. R., Belal, A., Muralidharan, S., Granger, E. & Pedersoli, M. (2024a). Visual Modality Prompt for Adapting Vision-Language Object Detectors. *arXiv preprint arXiv:2412.00622*, .

Medeiros, H. R., Belal, A., Muralidharan, S., Granger, E. & Pedersoli, M. (2025a). Visual Modality Prompt for Adapting Vision-Language Object Detectors. *arXiv preprint arXiv:2412.00622*.

Medeiros, H. R., Latortue, D., Granger, E. & Pedersoli, M. (2025b). Mixed patch visible-infrared modality agnostic object detection. *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 9023–9032.

Medeiros, H. R., Aminbeidokhti, M., Peña, F. A. G., Latortue, D., Granger, E. & Pedersoli, M. (2024b). Modality translation for object detection adaptation without forgetting prior knowledge. *European Conference on Computer Vision*, pp. 51–68.

Medeiros, H. R., Pena, F. A. G., Aminbeidokhti, M., Dubail, T., Granger, E. & Pedersoli, M. (2024c). HalluciDet: Hallucinating RGB Modality for Person Detection Through Privileged Information. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1444–1453.

Menezes, A. G., de Moura, G., Alves, C. & de Carvalho, A. C. (2023). Continual object detection: a review of definitions, strategies, and challenges. *Neural networks*, 161, 476–493.

Michaelis, C., Mitzkus, B., Geirhos, R., Rusak, E., Bringmann, O., Ecker, A. S., Bethge, M. & Brendel, W. (2019). Benchmarking robustness in object detection: Autonomous driving when winter is coming. *arXiv preprint arXiv:1907.07484*, .

Miezianko, R. & Pokrajac, D. (2008). People detection in low resolution infrared videos. *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1–6.

Minderer, M., Gritsenko, A., Stone, A., Neumann, M., Weissenborn, D., Dosovitskiy, A., Mahendran, A., Arnab, A., Dehghani, M., Shen, Z. et al. (2022). Simple open-vocabulary object detection. *European Conference on Computer Vision*, pp. 728–755.

Minderer, M., Gritsenko, A. & Houlsby, N. (2023). Scaling open-vocabulary object detection. *Advances in Neural Information Processing Systems*, 36, 72983–73007.

Mizrahi, D., Bachmann, R., Kar, O., Yeo, T., Gao, M., Dehghan, A. & Zamir, A. (2023). 4m: Massively multimodal masked modeling. *Advances in Neural Information Processing Systems*, 36, 58363–58408.

Motiian, S., Piccirilli, M., Adjeroh, D. A. & Doretto, G. (2016). Information bottleneck learning using privileged information for visual recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1496–1505.

Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. MIT press.

Natan, O. & Miura, J. (2022). End-to-end autonomous driving with semantic depth cloud mapping and multi-agent. *IEEE Transactions on Intelligent Vehicles*, 8(1), 557–571.

Nataprawira, J., Gu, Y., Goncharenko, I. & Kamijo, S. (2021). Pedestrian detection using multispectral images and a deep neural network. *Sensors*, 21(7), 2536.

Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H. & Ng, A. Y. (2011). Multimodal deep learning. *ICML*.

Özkanoğlu, M. A. & Ozer, S. (2022). InfraGAN: A GAN architecture to transfer visible images to infrared domain. *Pattern Recognition Letters*, 155, 69–76.

O'Mahony, N., Campbell, S., Carvalho, A., Harapanahalli, S., Hernandez, G. V., Krpalkova, L., Riordan, D. & Walsh, J. (2020). Deep learning vs. traditional computer vision. *Advances in Computer Vision: Proceedings of the 2019 Computer Vision Conference (CVC), Volume 1 1*, pp. 128–144.

Padilla, R., Netto, S. L. & Da Silva, E. A. (2020). A survey on performance metrics for object-detection algorithms. *2020 international conference on systems, signals and image processing (IWSSIP)*, pp. 237–242.

Pang, Y., Lin, J., Qin, T. & Chen, Z. (2021). Image-to-Image Translation: Methods and Applications.

Park, S., Vien, A. G. & Lee, C. (2023). Cross-modal transformers for infrared and visible image fusion. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(2), 770–785.

Park, T., Efros, A. A., Zhang, R. & Zhu, J.-Y. (2020a). Contrastive learning for unpaired image-to-image translation. *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, pp. 319–345.

Park, T., Efros, A. A., Zhang, R. & Zhu, J.-Y. (2020b). Contrastive Learning for Unpaired Image-to-Image Translation. *European Conference on Computer Vision*.

Pawłowski, M., Wróblewska, A. & Sysko-Romańczuk, S. (2023). Effective techniques for multimodal data fusion: A comparative analysis. *Sensors*, 23(5), 2381.

Peng, X., Wei, Y., Deng, A., Wang, D. & Hu, D. (2022). Balanced Multimodal Learning via On-the-fly Gradient Modulation. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8228-8237. doi: 10.1109/CVPR52688.2022.00806.

Pierson, H. A. & Gashler, M. S. (2017). Deep learning in robotics: a review of recent research. *Advanced Robotics*, 31(16), 821–835.

Qingyun, F., Dapeng, H. & Zhaokui, W. (2021). Cross-modality fusion transformer for multispectral object detection. *arXiv preprint arXiv:2111.00273*, .

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J. et al. (2021). Learning transferable visual models from natural language supervision. *International conference on machine learning*, pp. 8748–8763.

Ramachandram, D. & Taylor, G. W. (2017). Deep multimodal learning: A survey on recent advances and trends. *IEEE signal processing magazine*, 34(6), 96–108.

Ramachandran, A. & Sangaiah, A. K. (2021). A review on object detection in unmanned aerial vehicle surveillance. *International Journal of Cognitive Computing in Engineering*, 2, 215–228.

Razakarivony, S. & Jurie, F. (2016). Vehicle detection in aerial imagery: A small target detection benchmark. *Journal of Visual Communication and Image Representation*, 34, 187–203.

Redmon, J. & Farhadi, A. (2017). YOLO9000: better, faster, stronger. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7263–7271.

Redmon, J. & Farhadi, A. (2018). Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, .

Redmon, J., Divvala, S., Girshick, R. & Farhadi, A. (2016). You only look once: Unified, real-time object detection. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 779–788.

Ren, S., He, K., Girshick, R. & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 91–99.

Ren, S., He, K., Girshick, R. & Sun, J. (2016). Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6), 1137–1149.

Robins, A. (1993). Catastrophic forgetting in neural networks: the role of rehearsal mechanisms. *Proceedings 1993 The First New Zealand International Two-Stream Conference on Artificial Neural Networks and Expert Systems*, pp. 65–68.

Rokach, L. (2010). Ensemble-based classifiers. *Artificial intelligence review*, 33, 1–39.

Ronneberger, O., Fischer, P. & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241.

RoyChowdhury, A., Chakrabarty, P., Singh, A., Jin, S., Jiang, H., Cao, L. & Learned-Miller, E. (2019). Automatic adaptation of object detectors to new domains using self-training. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 780–790.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M. et al. (2015). Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115, 211–252.

Ryu, J. & Kim, S. (2018). Small infrared target detection by data-driven proposal and deep learning-based classification. *Infrared Technology and Applications XLIV*, 10624, 106241J.

Sagi, O. & Rokach, L. (2018). Ensemble learning: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4), e1249.

Saito, K., Ushiku, Y., Harada, T. & Saenko, K. (2019). Strong-weak distribution alignment for adaptive object detection. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6956–6965.

Samaras, S., Diamantidou, E., Ataloglou, D., Sakellariou, N., Vafeiadis, A., Magoulianitis, V., Lalas, A., Dimou, A., Zarpalas, D., Votis, K. et al. (2019). Deep learning on multi sensor data for counter UAV applications—A systematic review. *Sensors*, 19(22), 4837.

Schapire, R. E. et al. (1999). A brief introduction to boosting. *Ijcai*, 99(999), 1401–1406.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D. & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. *Proceedings of the IEEE international conference on computer vision*, pp. 618–626.

Shao, S., Li, Z., Zhang, T., Peng, C., Yu, G., Zhang, X., Li, J. & Sun, J. (2019a, October). Objects365: A Large-Scale, High-Quality Dataset for Object Detection. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Shao, S., Li, Z., Zhang, T., Peng, C., Yu, G., Zhang, X., Li, J. & Sun, J. (2019b). Objects365: A large-scale, high-quality dataset for object detection. *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 8430–8439.

Sharma, M., Dhanaraj, M., Karnam, S., Chachlakis, D. G., Ptucha, R., Markopoulos, P. P. & Saber, E. (2020). YOLOrs: Object detection in multimodal remote sensing imagery. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14, 1497–1508.

Shen, Y. & Gao, M. (2019). Brain tumor segmentation on MRI with missing modalities. *Information Processing in Medical Imaging: 26th International Conference, IPMI 2019, Hong Kong, China, June 2–7, 2019, Proceedings 26*, pp. 417–428.

Silberman, N., Hoiem, D., Kohli, P. & Fergus, R. (2012). Indoor segmentation and support inference from rgbd images. *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part V 12*, pp. 746–760.

Simonyan, K. & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, .

St-Laurent, L., Maldague, X. & Prévost, D. (2007). Combination of colour and thermal sensors for enhanced object detection. *2007 10th International Conference on Information Fusion*, pp. 1–8.

Stilgoe, J. (2018). Machine learning, social learning and the governance of self-driving cars. *Social studies of science*, 48(1), 25–56.

Suard, F., Rakotomamonjy, A., Bensrhair, A. & Broggi, A. (2006). Pedestrian detection using infrared images and histograms of oriented gradients. *2006 IEEE Intelligent Vehicles Symposium*, pp. 206–212.

Sun, Z., Cao, S., Yang, Y. & Kitani, K. M. (2021). Rethinking transformer-based set prediction for object detection. *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3611–3620.

Takumi, K., Watanabe, K., Ha, Q., Tejero-De-Pablos, A., Ushiku, Y. & Harada, T. (2017). Multispectral object detection for autonomous vehicles. *Proceedings of the on Thematic Workshops of ACM Multimedia 2017*, pp. 35–43.

Tang, L., Xiang, X., Zhang, H., Gong, M. & Ma, J. (2023a). DIVFusion: Darkness-free infrared and visible image fusion. *Information Fusion*, 91, 477–493.

Tang, Q., Liang, J. & Zhu, F. (2023b). A comparative review on multi-modal sensors fusion based on deep learning. *Signal Processing*, 109165.

Teutsch, M., Sappa, A. D. & Hammoud, R. I. (2021). Computer vision in the infrared spectrum: challenges and approaches. *Challenges and Approaches*.

Tian, Z., Shen, C., Chen, H. & He, T. (2019). Fcos: Fully convolutional one-stage object detection. *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9627–9636.

Tian, Z., Shen, C., Chen, H. & He, T. (2020). Fcos: A simple and strong anchor-free object detector. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(4), 1922–1933.

Torralba, A. & Efros, A. A. (2011). Unbiased look at dataset bias. *CVPR 2011*, pp. 1521–1528.

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A. & Jégou, H. (2021). Training data-efficient image transformers & distillation through attention. *International conference on machine learning*, pp. 10347–10357.

Uijlings, J. R., Van De Sande, K. E., Gevers, T. & Smeulders, A. W. (2013). Selective search for object recognition. *International journal of computer vision*, 104(2), 154–171.

Ulyanov, D., Vedaldi, A. & Lempitsky, V. (2016). Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*.

Vandersteegen, M., Reusen, W., Beeck, K. V. & Goedemé, T. (2023). Person Detection Using an Ultra Low-Resolution Thermal Imager on a Low-Cost MCU. *Image and Vision Computing: 37th International Conference, IVCNZ 2022, Auckland, New Zealand, November 24–25, 2022, Revised Selected Papers*, pp. 33–47.

Vapnik, V. & Vashist, A. (2009). A new learning paradigm: Learning using privileged information. *Neural networks*, 22(5-6), 544–557.

Vasconcelos, C., Birodkar, V. & Dumoulin, V. (2022). Proper reuse of image classification features improves object detection. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13628–13637.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30, .

Wagner, J., Fischer, V., Herman, M., Behnke, S. et al. (2016a). Multispectral Pedestrian Detection using Deep Fusion Convolutional Neural Networks. *ESANN*, 587, 509–514.

Wagner, J., Fischer, V., Herman, M. & Behnke, S. (2016b, 04). Multispectral Pedestrian Detection using Deep Fusion Convolutional Neural Networks.

Wang, B., Ali, S., Fan, X. & Abuhmed, T. (2023a). Real-time human detection and behavior recognition using low-cost hardware. *2023 17th International Conference on Ubiquitous Information Management and Communication (IMCOM)*, pp. 1–8.

Wang, G., Zhang, T., Cheng, J., Liu, S., Yang, Y. & Hou, Z. (2019). Rgb-infrared cross-modality person re-identification via joint pixel and feature alignment. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3623–3632.

Wang, H., Cai, Y., Chen, X. & Chen, L. (2016). Night-time vehicle sensing in far infrared image with deep learning. *Journal of Sensors*, 2016.

Wang, Q., Chi, Y., Shen, T., Song, J., Zhang, Z. & Zhu, Y. (2022). Improving RGB-infrared object detection by reducing cross-modality redundancy. *Remote Sensing*, 14(9), 2020.

Wang, R. & Ge, T. (2016). *Advances in solar heating and cooling*. Woodhead Publishing.

Wang, W., Tran, D. & Feiszli, M. (2020). What Makes Training Multi-Modal Classification Networks Hard? *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12692-12702. doi: 10.1109/CVPR42600.2020.01271.

Wang, Z., Yang, E., Shen, L. & Huang, H. (2023b). A Comprehensive Survey of Forgetting in Deep Learning Beyond Continual Learning.

Wang, Z., Simoncelli, E. P. & Bovik, A. C. (2003). Multiscale structural similarity for image quality assessment. *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, 2, 1398–1402.

Wang, Z., Bovik, A. C., Sheikh, H. R. & Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4), 600–612.

Woo, S., Debnath, S., Hu, R., Chen, X., Liu, Z., Kweon, I. S. & Xie, S. (2023). ConvNeXt V2: Co-designing and Scaling ConvNets with Masked Autoencoders.

Wortsman, M., Ilharco, G., Gadre, S. Y., Roelofs, R., Gontijo-Lopes, R., Morcos, A. S., Namkoong, H., Farhadi, A., Carmon, Y., Kornblith, S. et al. (2022a). Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. *International conference on machine learning*, pp. 23965–23998.

Wortsman, M., Ilharco, G., Kim, J. W., Li, M., Kornblith, S., Roelofs, R., Lopes, R. G., Hajishirzi, H., Farhadi, A., Namkoong, H. et al. (2022b). Robust fine-tuning of zero-shot models. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7959–7971.

Wu, A., Zheng, W.-S., Yu, H.-X., Gong, S. & Lai, J. (2017). RGB-infrared cross-modality person re-identification. *Proceedings of the IEEE international conference on computer vision*, pp. 5380–5389.

Wu, X., Zhu, F., Zhao, R. & Li, H. (2023). Cora: Adapting clip for open-vocabulary detection with region prompting and anchor pre-matching. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7031–7040.

Xie, R., Yu, F., Wang, J., Wang, Y. & Zhang, L. (2019). Multi-level domain adaptive learning for cross-domain detection. *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 0–0.

Xiong, F., Wang, Q. & Gao, Q. (2019). Consistent embedded GAN for image-to-image translation. *IEEe Access*, 7, 126651–126661.

Xu, D., Ouyang, W., Ricci, E., Wang, X. & Sebe, N. (2017a). Learning cross-modal deep representations for robust pedestrian detection. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5363–5371.

Xu, S., Zhou, D., Fang, J., Yin, J., Bin, Z. & Zhang, L. (2021). FusionPainting: Multimodal fusion with adaptive attention for 3d object detection. *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, pp. 3047–3054.

Xu, X., Li, Y., Wu, G. & Luo, J. (2017b). Multi-modal deep feature learning for RGB-D object detection. *Pattern Recognition*, 72, 300–313.

Yakubovskiy, P. (2020). Segmentation Models Pytorch. GitHub. Retrieved from: https://github.com/qubvel/segmentation_models.pytorch.

Yang, S., Yu, S., Zhao, B. & Wang, Y. (2020). Reducing the feature divergence of RGB and near-infrared images using Switchable Normalization. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 46–47.

Yang, T., Li, Y., Zhao, C., Yao, D., Chen, G., Sun, L., Krajnik, T. & Yan, Z. (2022a). 3D ToF LiDAR in mobile robotics: A review. *arXiv preprint arXiv:2202.11025*, .

Yang, Z., Chen, J., Miao, Z., Li, W., Zhu, X. & Zhang, L. (2022b). Deepinteraction: 3d object detection via modality interaction. *Advances in Neural Information Processing Systems*, 35, 1992–2005.

Yao, L., Han, J., Wen, Y., Liang, X., Xu, D., Zhang, W., Li, Z., Xu, C. & Xu, H. (2022). Detclip: Dictionary-enriched visual-concept paralleled pre-training for open-world detection. *Advances in Neural Information Processing Systems*, 35, 9125–9138.

Yu, T., Lu, Z., Jin, X., Chen, Z. & Wang, X. (2023). Task residual for tuning vision-language models. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10899–10909.

Yuan, L., Hou, Q., Jiang, Z., Feng, J. & Yan, S. (2022). Volo: Vision outlooker for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 45(5), 6575–6586.

Zang, Y., Li, W., Zhou, K., Huang, C. & Loy, C. C. (2022). Open-vocabulary detr with conditional matching. *European Conference on Computer Vision*, pp. 106–122.

Zhang, A., Lipton, Z. C., Li, M. & Smola, A. J. (2023a). *Dive into deep learning*. Cambridge University Press.

Zhang, H., Li, F., Liu, S., Zhang, L., Su, H., Zhu, J., Ni, L. & Shum, H.-Y. (2023b). DINO: DETR with Improved DeNoising Anchor Boxes for End-to-End Object Detection. *The Eleventh International Conference on Learning Representations*.

Zhang, H., Li, F., Liu, S., Zhang, L., Su, H., Zhu, J., Ni, L. & Shum, H.-Y. (2023c). DINO: DETR with Improved DeNoising Anchor Boxes for End-to-End Object Detection. *The Eleventh International Conference on Learning Representations*.

Zhang, H., Zhang, P., Hu, X., Chen, Y.-C., Li, L., Dai, X., Wang, L., Yuan, L., Hwang, J.-N. & Gao, J. (2022). Glipv2: Unifying localization and vision-language understanding. *Advances in Neural Information Processing Systems*, 35, 36067–36080.

Zhang, H., Wang, Y., Dayoub, F. & Sünderhauf, N. (2020a). Swa object detection. *arXiv preprint arXiv:2012.12645*.

Zhang, H., Fromont, E., Lefèvre, S. & Avignon, B. (2020b). Multispectral fusion for object detection with cyclic fuse-and-refine blocks. *2020 IEEE International Conference on Image Processing (ICIP)*, pp. 276–280.

Zhang, H., Fromont, E., Lefevre, S. & Avignon, B. (2020c). Multispectral Fusion for Object Detection with Cyclic Fuse-and-Refine Blocks. *2020 IEEE International Conference on Image Processing (ICIP)*, pp. 276-280. doi: 10.1109/ICIP40778.2020.9191080.

ZHANG, H., FROMONT, E., LEFEVRE, S. & AVIGNON, B. (2021). Guided Attentive Feature Fusion for Multispectral Pedestrian Detection. *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 72-80. doi: 10.1109/WACV48630.2021.00012.

Zhang, H., Luo, C., Wang, Q., Kitchin, M., Parmley, A., Monge-Alvarez, J. & Casaseca-De-La-Higuera, P. (2018a). A novel infrared video surveillance system using deep learning based techniques. *Multimedia tools and applications*, 77, 26657–26676.

Zhang, J., Huang, J., Luo, Z., Zhang, G., Zhang, X. & Lu, S. (2023d). DA-DETR: Domain Adaptive Detection Transformer With Information Fusion. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23787–23798.

Zhang, L., Liu, Z., Zhang, S., Yang, X., Qiao, H., Huang, K. & Hussain, A. (2019a). Cross-modality interactive attention network for multispectral pedestrian detection. *Information Fusion*, 50, 20–29.

Zhang, L., Zhu, X., Chen, X., Yang, X., Lei, Z. & Liu, Z. (2019b). Weakly aligned cross-modal learning for multispectral pedestrian detection. *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5127–5137.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E. & Wang, O. (2018b). The unreasonable effectiveness of deep features as a perceptual metric. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595.

Zhang, S., Wen, L., Bian, X., Lei, Z. & Li, S. Z. (2018c). Single-shot refinement neural network for object detection. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4203–4212.

Zhang, T., Wu, F., Katiyar, A., Weinberger, K. Q. & Artzi, Y. (2021). Revisiting Few-sample BERT Fine-tuning.

Zhao, H., Shi, J., Qi, X., Wang, X. & Jia, J. (2017). Pyramid scene parsing network. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2881–2890.

Zhao, T., Yuan, M. & Wei, X. (2024a). Removal and selection: Improving rgb-infrared object detection via coarse-to-fine fusion. *arXiv preprint arXiv:2401.10731*, .

Zhao, Y., Lv, W., Xu, S., Wei, J., Wang, G., Dang, Q., Liu, Y. & Chen, J. (2024b). Detrs beat yolos on real-time object detection. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16965–16974.

Zhong, Y., Yang, J., Zhang, P., Li, C., Codella, N., Li, L. H., Zhou, L., Dai, X., Yuan, L., Li, Y. et al. (2022). Regionclip: Region-based language-image pretraining. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16793–16803.

Zhou, K., Yang, J., Loy, C. C. & Liu, Z. (2022a). Conditional prompt learning for vision-language models. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16816–16825.

Zhou, K., Yang, J., Loy, C. C. & Liu, Z. (2022b). Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9), 2337–2348.

Zhou, Z., Rahman Siddiquee, M. M., Tajbakhsh, N. & Liang, J. (2018). Unet++: A nested u-net architecture for medical image segmentation. In *Deep learning in medical image analysis and multimodal learning for clinical decision support* (pp. 3–11). Springer.

Zhu, J.-Y., Park, T., Isola, P. & Efros, A. A. (2017a). Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. *Computer Vision (ICCV), 2017 IEEE International Conference on*.

Zhu, J.-Y., Park, T., Isola, P. & Efros, A. A. (2017b). Unpaired image-to-image translation using cycle-consistent adversarial networks. *Proceedings of the IEEE international conference on computer vision*, pp. 2223–2232.

Zhu, J.-Y., Zhang, R., Pathak, D., Darrell, T., Efros, A. A., Wang, O. & Shechtman, E. (2017c). Toward multimodal image-to-image translation. *Advances in neural information processing systems*, 30.

Zhu, X., Su, W., Lu, L., Li, B., Wang, X. & Dai, J. (2021). Deformable DETR: Deformable Transformers for End-to-End Object Detection. *International Conference on Learning Representations*.

Zou, Z., Chen, K., Shi, Z., Guo, Y. & Ye, J. (2023). Object detection in 20 years: A survey. *Proceedings of the IEEE*, 111(3), 257–276.