Gradual Domain Generalization for Visible-Infrared Person Re-Identification

by

Mahdi ALEHDAGHI

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

MONTREAL, JANUARY 12, 2026

ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
UNIVERSITÉ DU QUÉBEC

BOARD OF EXAMINERS

THIS THESIS HAS BEEN EVALUATED

BY THE FOLLOWING BOARD OF EXAMINERS

Prof. Éric Granger, Thesis supervisor
Department of Systems Engineering, École de technologie supérieure

Prof. Rafael M.O Cruz, Thesis Co-Supervisor
Department of Software Engineering and IT, École de technologie supérieure

Prof. Stéphane Coulombe, Chair, Board of Examiners
Department of Software Engineering and IT, École de technologie supérieure

Prof. Mohammadhadi Shateri, Member of the Jury
Department of Systems Engineering, École de technologie supérieure

Prof. Yang Wang, External Independent Examiner
Department of Computer Science and Software Engineering, Concordia University

THIS THESIS WAS PRESENTED AND DEFENDED

IN THE PRESENCE OF A BOARD OF EXAMINERS AND THE PUBLIC

ON DECEMBER 18, 2025

AT ÉCOLE DE TECHNOLOGIE SUPÉRIEURE

FOREWORD

This doctoral thesis is presented in the form of a collection of scientific articles. The research reported here was conducted between 2021 and 2025 at the Laboratoire d'imagerie, de vision et d'intelligence artificielle (LIVIA), École de technologie supérieure (ÉTS), Montréal, under the supervision of Professor Eric Granger and co-supervision of Professor Rafael M. O. Cruz.

The thesis consists of three main articles, each addressing a complementary aspect of the Visible-Infrared Person Re-Identification (VI-ReID) problem, and an additional article presented in the Appendix on interpretable deep learning. These contributions are the result of collaborative work, but the candidate was the lead author and principal investigator in all cases.

The articles are as follows:

- Chapter 2: *Adaptive Generation of Privileged Intermediate Information for Visible-Infrared Person Re-Identification*, published in **IEEE Transactions on Information Forensics and Security** (2024).
- Chapter 3: *Bidirectional Multi-step Domain Generalization for Visible-Infrared Person Re-Identification*, published in the **IEEE/CVF Winter Conference on Applications of Computer Vision** (2025).
- Chapter 4: *MixER: From Cross-Modal to Mixed-Modal Matching*, submitted to the **IEEE/CVF Winter Conference on Applications of Computer Vision** (2026).
- Chapter 5: *Beyond Patches: Mining Interpretable Part-Prototypes for Explainable AI*, submitted to the **AAAI Conference on Artificial Intelligence** (2026).

Except for minor formatting and stylistic adjustments to maintain consistency across chapters, the articles are reproduced as they appeared (or will appear) in their respective venues. The introduction and conclusion chapters provide a unifying framework, positioning the individual contributions within the broader context of gradual domain generalization for VI-ReID.

The work presented here represents the candidate's original contributions to the field, with co-authors providing guidance, feedback, and collaboration.

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my supervisor, **Professor Eric Granger**, for his exceptional guidance, continuous support, and encouragement throughout my doctoral studies. His insightful feedback and commitment to academic excellence have been invaluable to the completion of this work. I am equally grateful to my co-supervisor, **Professor Rafael M. O. Cruz**, whose expertise, constructive advice, and collaboration have greatly contributed to the quality and depth of this research. I would also like to thank **Professor Pourya Shamsolmoali** for his valuable guidance and assistance during the course of this work.

I am thankful to my colleagues at the **LIVIA laboratory**—including Arthur Josi, and Rajarshi Bhattacharya—for their collaboration, stimulating discussions, and friendship. I also extend my sincere appreciation to the members of my jury and examiners for their time, constructive feedback, and valuable insights that helped refine this dissertation.

This research was made possible through the support of **ÉTS**, the **Université du Québec**, and the funding agencies that provided essential financial assistance.

Finally, I wish to express my heartfelt gratitude to my beloved wife, **Somaye**, for her unwavering love, patience, and encouragement. Her constant support has been a source of strength and inspiration throughout this journey.

Généralisation progressive de domaine pour la ré-identification de personnes en mode visible–infrarouge

Mahdi ALEHDAGHI

RÉSUMÉ

L'identification de personnes en mode visible-infrarouge (Visible-infrared person re-identification, V-I ReID) est devenue une tâche essentielle pour les systèmes de surveillance modernes, puisqu'elle vise à associer correctement des individus capturés par des caméras de modalités différentes — visible (V) et infrarouge (I). Cependant, cette tâche reste difficile en raison de l'important **décalage de domaine** entre les modalités, se traduisant par de fortes divergences d'apparence, de texture et d'illumination. Ce décalage limite gravement la capacité de généralisation des modèles d'apprentissage profond et réduit la performance en recherche cross-modale. De plus, les méthodes classiques d'extraction de caractéristiques manquent souvent de discriminabilité car elles reposent sur des représentations globales, négligeant des attributs fins et spécifiques à l'identité, pourtant essentiels pour un appariement robuste. Les protocoles d'évaluation existants renforcent ces limites en supposant, lors des tests, une galerie uni-modale, ce qui ne correspond pas aux conditions réelles des systèmes de surveillance 24h/24, qui doivent naturellement traiter des galeries multi-modales.

Cette thèse, intitulée *Gradual Domain Generalization for Visible-Infrared Person Re-Identification*, aborde ces défis de manière systématique à travers un cadre progressif et structuré de généralisation de domaine. Trois contributions majeures y sont proposées :

1. **AGPI2 (Adaptive Generation of Privileged Intermediate Information)**, qui introduit des modalités intermédiaires dynamiques afin de guider l'apprentissage de caractéristiques invariantes aux modalités ;
2. **BMDG (Bidirectional Multi-step Domain Generalization)**, qui affine progressivement l'alignement des caractéristiques locales pour réduire plus efficacement l'écart entre domaines ;
3. **MixReID**, qui sépare les informations spécifiques à la modalité de celles propres à l'identité, afin d'améliorer la performance dans des conditions mixtes et réalistes.

Des expérimentations approfondies sur les bases de données SYSU-MM01, RegDB et LLCM démontrent que les approches proposées atteignent des résultats à l'état de l'art dans divers scénarios difficiles, incluant l'occlusion, les variations extrêmes de luminosité et les contextes multi-modaux.

Au-delà de l'identification visible-infrarouge, cette thèse s'intéresse également à un problème complémentaire et actuel : renforcer l'**interprétabilité** et l'**interopérabilité** des modèles de vision profonds. Inspirée par les mécanismes de découverte de prototypes développés pour la ReID, elle propose un cadre d'explicabilité permettant la découverte automatique de concepts primitifs et prototypiques, sans annotation manuelle. Cette approche améliore la transparence et la robustesse des explications produites, même dans des conditions difficiles. En comblant le

fossé entre performance et interprétabilité, cette thèse contribue au développement de systèmes d'intelligence artificielle plus fiables et déployables dans des applications réelles.

Mots-clés: Ré-identification visible-infrarouge de personnes, adaptation au domaine, généralisation au domaine

Gradual Domain Generalization for Visible-Infrared Person Re-Identification

Mahdi ALEHDAGHI

ABSTRACT

Visible-infrared person re-identification (V-I ReID) has emerged as a crucial task for modern surveillance systems, requiring the accurate matching of individuals captured across different modalities—visible (V) and infrared (I) cameras. However, significant challenges arise from the substantial domain gap between modalities, resulting in pronounced discrepancies in appearance, texture, and illumination. This gap severely hampers the generalization ability of deep learning models, leading to degraded cross-modal retrieval performance. Furthermore, conventional feature extraction approaches often lack discriminability due to their reliance on global representations, overlooking fine-grained, identity-specific attributes critical for robust person matching. Existing evaluation protocols exacerbate these limitations by assuming a unimodal gallery during testing, which is unrealistic for real-world 24-hour surveillance systems that naturally operate under mixed-modality conditions.

This thesis, titled Gradual domain generalization for visible-infrared person re-identification, systematically addresses these challenges through a gradual and structured domain generalization framework. Three major contributions are proposed: (1) Adaptive generation of privileged intermediate information (AGPI^2), introducing dynamic intermediate modalities to guide the learning of modality-invariant features; (2) Bidirectional multi-step domain generalization (BMDG), progressively refining part-based feature alignments to bridge domain gaps more effectively; and (3) Mixed-modality re-identification (MixER), disentangling modality-specific and identity-specific features to enhance performance under realistic mixed-gallery conditions. Extensive experiments conducted on benchmark datasets such as SYSU-MM01, RegDB, and LLCM demonstrate that the proposed approaches achieve state-of-the-art results across a variety of challenging scenarios, including occlusion, extreme lighting variations, and mixed-modal settings.

Beyond cross-modal person re-identification, this thesis also investigates a complementary and timely problem: enhancing the interpretability and interoperability of deep vision models. Inspired by the prototype discovery mechanisms introduced for ReID, a novel framework for interpretable classification is developed, enabling the automatic discovery of part-prototypical primitive concepts without requiring manual annotations. This approach not only improves the transparency and faithfulness of deep models but also ensures robust and stable explanations even under challenging conditions. By bridging the gap between performance and interpretability, this thesis contributes toward the development of more trustworthy and practically deployable deep learning systems for real-world applications.

Keywords: Visible-Infrared Person Re-Identification, Gradual Domain Adaptation, Domain Generalization

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Page

LIST OF ABBREVIATIONS AND ACRONYMS

ÉTS	École de Technologie Supérieure
ASC	Agence Spatiale Canadienne
V-I ReID	Visible-Infrared Person Re-Identification
DA	Domain Adaptation
DG	Domain Generalization
AGPI^2	Adaptive Generation of Privileged Intermediate Information
BMDG	Bidirectional Multi-step Domain Generalization
CNN	Convolutional Neural Network

LIST OF SYMBOLS AND UNITS OF MEASUREMENTS

x_v	visible input image
x_i	infrared input image
$\mathbf{f_v}$	visible features vector
$\mathbf{f_i}$	infrared features vector

INTRODUCTION

Visible-infrared person re-identification (V-I ReID) is a crucial task in surveillance and security applications, aiming to match individuals captured across different modalities—visible (V) and infrared (I) cameras. While standard visible cameras provide rich color information essential for daytime monitoring, their performance degrades significantly in low-light or nighttime environments. Infrared cameras address this blind spot by capturing thermal radiation, which is independent of external illumination. Consequently, the integration of both modalities is strictly necessary to ensure continuous, 24-hour surveillance capabilities in real-world security systems.

The key challenge in V-I ReID lies in bridging the significant domain gap between these modalities, as RGB and infrared images exhibit substantial differences in color, texture, and illumination (Ye *et al.*, 2020b). In the context of machine learning, a 'domain' refers to the specific statistical distribution of the data. Here, the visible and infrared modalities represent two distinct domains: one rich in color and texture, the other consisting solely of thermal intensity patterns. The 'domain gap' is the discrepancy between these feature distributions. Because standard models trained on color images cannot naturally interpret thermal data, bridging this gap is the fundamental challenge in cross-modality ReID. While deep learning-based approaches have shown remarkable progress in unimodal ReID, cross-modal matching remains an open problem due to these domain discrepancies. Addressing this challenge is essential for applications such as nighttime surveillance, autonomous security systems, and forensic investigations, where visibility conditions may vary significantly. Moreover, as intelligent video analysis systems are increasingly deployed in public spaces, ensuring accurate person retrieval across different lighting conditions and imaging technologies becomes imperative.

In V-I ReID, a fundamental challenge arises from the substantial domain gap between images captured in the visible (V) spectrum and those captured in the infrared (I) spectrum. This gap is primarily caused by the differences in sensing modalities, which lead to significant variations

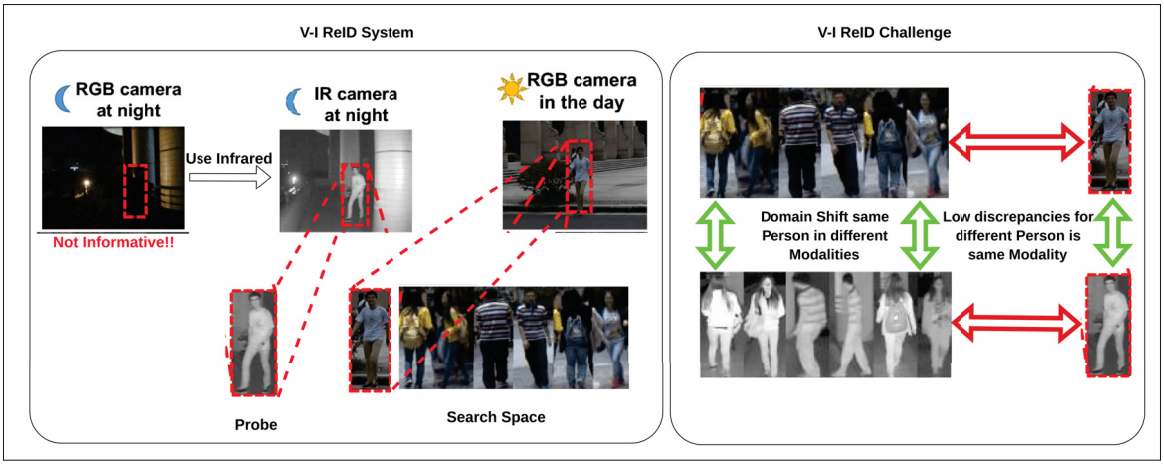


Figure 0.1 Overview of an automated V-I ReID system. The RGB cameras at night do not provide enough information for the system, and it should rely on an infrared camera. The primary challenge of V-I ReID is addressing the significant domain gap between images of the same person, captured in different modalities.

in appearance, texture, color, and structural details. Visible images capture rich color and fine texture information under adequate illumination, while infrared images primarily represent the thermal properties of objects, often lacking fine-grained visual cues. Consequently, even images of the same person can appear drastically different across these modalities. These modality-induced discrepancies severely hinder the performance of conventional re-identification systems, which are typically designed for intra-modal (same-modality) matching. Without addressing this domain gap, feature embeddings extracted from visible and infrared images remain poorly aligned, making it challenging for models to effectively associate cross-modal instances of the same identity.

If the domain gap between visible and infrared modalities is not adequately addressed in the training process, the learned feature embedding space becomes modality-biased, clustering samples based on sensor type rather than identity (Zhou, Liu, Qiao, Xiang & Loy, 2022; Wang *et al.*, 2019c; Wang & Deng, 2018), thereby defeating the objective of person re-identification. In practical surveillance systems, this could cause critical errors in identifying individuals across

day-night cycles or different camera types, undermining the reliability and robustness of security applications. Thus, learning modality-invariant or modality-bridging features is indispensable for V-I ReID to ensure accurate identity matching, reduce modality-induced bias, and achieve satisfactory real-world performance.

This thesis presents a systematic study of V-I ReID and introduces a multi-step and gradual method to enhance cross-modal matching by reducing the effect of domain gap on the model. The core contributions of this work are (a) the gradual domain generalization framework, which progressively refines feature representations by leveraging intermediate modalities and learning robust cross-modal embeddings, and (b) introducing new and practical mixed-modality matching challenges and a mathematical learning mechanism to address them. The first framework aims to minimize the domain gap through a structured approach, ensuring that learned features retain discriminative identity information while reducing modality bias. Unlike traditional domain adaptation methods that rely on a single-step transformation, this thesis explores a progressive adaptation strategy that enhances the model's generalization ability across varying conditions. The second approach is using mutual information analysis to jointly address the inter-modal and intra-modal matching at the same time, with disentangling modality-erased and modality-related information from images.

Another challenge faced by deep feature extraction models in V-I ReID is the lack of feature discriminability, primarily due to insufficient emphasis on local attributes and distinctive body parts. Conventional deep models often learn global representations that aggregate holistic appearance cues; however, in cross-modal scenarios, global features may be less reliable due to modality-induced appearance changes. As a result, the models may overlook fine-grained, identity-specific details such as clothing textures, accessories, or unique body structures, which are crucial for distinguishing between individuals. Moreover, the absence or misalignment of local features—such as the occlusion of limbs or differences in viewpoint—can further

exacerbate feature confusion, leading to reduced matching accuracy. Failure to explicitly capture and leverage local parts not only diminishes the robustness of feature embeddings but also increases intra-class variance and decreases inter-class separability, ultimately impairing cross-modal generalization performance. Therefore, enhancing local feature modeling, either through part-based learning, attribute-guided supervision, or attention mechanisms, is essential to achieve highly discriminative and modality-resilient embeddings for effective V-I ReID.

A critical limitation of existing evaluation protocols for cross-modal person re-identification (ReID) is the unrealistic assumption that the gallery set contains images from only one modality—typically either visible or infrared—during testing. While this setting simplifies evaluation, it does not accurately reflect real-world surveillance scenarios, where a practical system must operate seamlessly across day and night, and the gallery naturally comprises a mixture of images from both modalities. In actual deployments, surveillance systems must retrieve a person of interest from a gallery that includes visible images captured during daytime and infrared images captured at night, resulting in a heterogeneous and mixed-modal gallery. Current evaluations fail to address this complexity, leading to overly optimistic performance estimates and hindering the development of truly robust cross-modal retrieval systems. Recognizing this gap, recent efforts, such as the mixed-modal scenarios proposed in this thesis, advocate for more realistic evaluation protocols that involve mixed-modality galleries, thus posing new challenges for feature alignment, discriminative embedding learning, and robust matching across diverse conditions. Addressing the mixed-modal gallery challenge is essential for building reliable V-I ReID systems capable of consistent and accurate performance across a full 24-hour surveillance cycle.

To address these challenges, this thesis is structured around four research articles organized as four main chapters:

- Adaptive Generation of Privileged Intermediate Information (AGPI^2) for V-I ReID (Chapter 2, which was published in **IEEE Transactions on Information Forensics and Security, 2024**). This method introduces an intermediate modality generated through adversarial training, serving as privileged information during training to guide the model in learning modality-invariant features. By dynamically generating an intermediate space, AGPI^2 provides a bridge that reduces the distributional discrepancy between V and I images, enhancing the feature extraction process. Unlike conventional generative approaches that synthesize infrared images from visible ones, AGPI^2 aims to create a privileged space where identity-preserving attributes remain intact while modality-dependent factors are minimized.
- Bidirectional Multi-step Domain Generalization (BMDG) for V-I Person ReID (Chapter 3, which was published in **IEEE/CVF Winter Conference on Applications of Computer Vision, 2025**): A multi-step domain generalization framework that progressively reduces the modality gap by aligning part-based features across modalities, thus improving representation learning in a structured manner. BMDG refines the standard domain adaptation process by incorporating a gradual alignment technique, making it more robust to significant domain shifts. By leveraging bidirectional learning, this approach ensures that feature adaptation occurs symmetrically across visible and infrared modalities, mitigating overfitting to either domain.
- MixER: From Cross-Modal to Mixed-Modal Visible-Infrared Person Re-Identification (Chapter 4, which was sent in **The IEEE/CVF Winter Conference on Applications of Computer Vision, 2026**): A novel approach that disentangles modality-specific and modality-agnostic identity features to improve mixed-modal ReID performance in real-world surveillance settings. MixReID focuses on enhancing discriminability by separating modality-dependent features from identity-related representations, thus improving both intra- and inter-modality matching. This method also introduces novel feature disentanglement

losses, ensuring that the extracted representations are robust to variations in modality and effectively capture identity-specific cues.

- Beyond Patches: Mining Interpretable Part-Prototypes for Explainable AI (Chapter 5, which was sent in **The 40th Annual AAAI Conference on Artificial Intelligence, 2026**): The interpretability of deep models remains a challenge in computer vision tasks. Post-hoc explainability methods, such as GradCAM, provide visual interpretation based on heatmaps but lack conceptual clarity and explainability. To address these limitations, a part-prototypical concept mining network (PCMNet) is proposed that dynamically learns interpretable prototypes from meaningful regions. PCMNet clusters prototypes into concept groups, creating semantically grounded explanations without requiring additional annotations. Through a joint process of unsupervised part discovery and concept activation vector extraction, PCMNet effectively captures discriminative concepts and makes interpretable classification decisions.

The gradual domain generalization framework and mixed-modality learning mechanism proposed in this thesis enable robust cross-modal person re-identification by progressively refining feature representations and minimizing modality discrepancies, and separating and mixing discriminative modality-related and modality-agnostic features for cross-modal and mixed-modal matching. The proposed approaches are evaluated on benchmark datasets, including SYSU-MM01, RegDB, and LLCM, demonstrating state-of-the-art performance in V-I ReID. The experimental results indicate that these methods significantly outperform traditional feature learning and domain adaptation techniques, establishing a new standard for cross-modal ReID. Moreover, this thesis examines the robustness of these methods under challenging conditions, including occlusions, viewpoint variations, and extreme illumination changes, ensuring their applicability in real-world scenarios.

Beyond person re-identification, this thesis also explores the growing need for improving the interpretability and interoperability of deep models in vision tasks. Despite the tremendous success of deep learning in classification, current models largely operate as black boxes, providing little insight into the underlying decision-making process. Traditional post-hoc explanation methods, such as heatmaps, offer limited semantic understanding and often lack robustness. Motivated by the part-prototype discovery techniques initially developed for cross-modal ReID, this thesis proposes a novel framework for interpretable classification by mining part-prototypical, primitive concepts without requiring additional annotations. By dynamically discovering semantically meaningful components and structuring them into intuitive prototypes, this approach enhances model transparency, faithfulness, and stability even under challenging conditions like occlusion. This contribution, although distinct from cross-modal matching, addresses a critical and underexplored aspect of deep vision systems: making model decisions understandable, robust, and accessible for high-stakes and real-world applications.

CHAPTER 1

LITERATURE REVIEW

1.1 Person Re-Identification

Person re-identification (ReID) is a critical component in various computer vision applications, ranging from sports analytics to video surveillance. Given a query image of an individual, ReID is formulated as a ranking problem, where the goal is to retrieve and rank the most visually similar images from a gallery of previously observed individuals. The general objective is to recognize the same individual across a network of distributed, non-overlapping cameras under the assumption that the person's overall appearance (e.g., clothing) remains consistent across viewpoints (Wang *et al.*, 2019b; Ye *et al.*, 2020b; Wang *et al.*, 2019c).

Classical person re-identification approaches can be broadly categorized into two main groups (Zheng, Yang & Hauptmann, 2016a). The first category focuses on designing discriminative features and combining them to build robust descriptors (or signatures) for each individual, regardless of the scene. The second category focuses on learning a discriminative distance metric from data, aiming to minimize intra-class variance across camera views. Like other traditional computer vision systems that rely on handcrafted features, the first group suffers from poor generalization. In contrast, metric learning-based methods, although more adaptive, often face scalability issues as the number of individuals or cameras grows, increasing the computational complexity.

Deep learning (DL) techniques have significantly advanced the state of the art in person ReID, outperforming traditional methods (Ye *et al.*, 2020b). Most modern ReID systems utilize deep convolutional neural networks (CNNs) that integrate feature extraction (Dai, Ji, Wang, Wu & Huang, 2018), representation learning (Choi, Lee, Kim, Kim & Kim, 2020a; Wang *et al.*, 2019b; Fang *et al.*, 2023), and metric learning (Lu *et al.*, 2020; Jia, Zhai, Lu, Ma & Zhang, 2020) into a unified end-to-end framework (Wang *et al.*, 2019c). These systems are inspired by the cognitive process of humans, who re-identify individuals by recalling distinctive features

such as facial attributes, clothing color, or inferred traits like age or gender (Ye *et al.*, 2020b). When the visual similarity is ambiguous, humans subconsciously focus on finer-grained or more contextual features. Analogously, in deep ReID systems, a CNN learns to extract a hidden feature vector representing identity-relevant characteristics. These features are optimized during training to capture discriminative and invariant representations suitable for person matching.

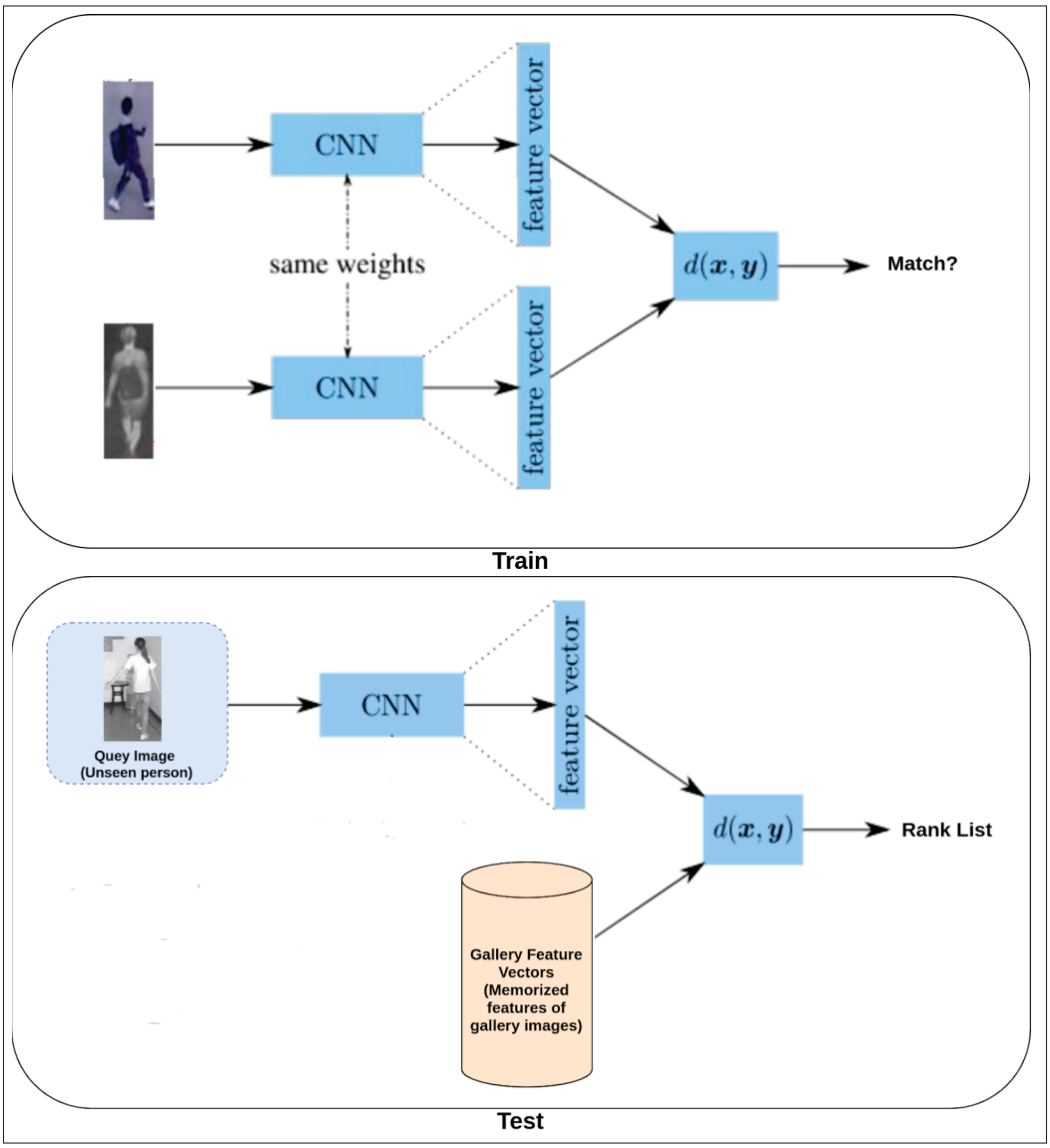


Figure 1.1 Deep learning models for training and testing person ReID.

As illustrated in Figure 1.1, automated ReID systems typically apply person detection to extract bounding boxes from video frames. These boxes are then compared against gallery images

captured from multiple cameras over time. State-of-the-art methods usually rely on global appearance features extracted from bounding boxes using CNN backbones trained with metric learning objectives. Siamese networks and triplet-loss-based CNNs are frequently employed to embed images into a feature space where similar identities are close, and dissimilar ones are far apart. During inference, ReID involves (1) extracting feature representations for a query image and (2) matching the query against a gallery by computing similarity scores between their embeddings.

Outdoor ReID systems must also be robust in challenging conditions such as nighttime surveillance, where RGB images are affected by noise and poor lighting. To address this, additional sensing modalities like infrared (IR) or thermal imaging are often used (Nguyen, Hong, Kim & Park, 2017). However, matching individuals across different modalities introduces new challenges, particularly the modality gap between RGB and IR images. The ReID model that learns from multiple modalities must account for the domain shift between them. Cross-modal ReID requires domain adaptation techniques to bridge this gap and find a shared representation space that aligns source and target domains. Domain adaptation methods are therefore essential for improving model generalization and matching accuracy in cross-modality ReID systems.

1.2 Cross-Modal Person Re-Identification

While person ReID generally refers to the task of matching individuals across different camera views within the same modality (typically RGB), many real-world applications—such as 24-hour surveillance—require recognition across different modalities. This need has given rise to Visible-Infrared Person Re-Identification (V-I ReID), a subfield of ReID in which the query image (often visible during the day) must be matched to gallery images captured in the infrared spectrum (typically at night), or vice versa. The primary challenge in V-I ReID lies in the significant modality gap—RGB and infrared images differ greatly in terms of color, texture, and illumination, which severely hinders conventional matching techniques.

The problem of V-I ReID has been actively studied in recent years, with solutions typically falling into three major categories (Wang *et al.*, 2019c): (a) image-level generation, (b) feature-level representation learning, and (c) intermediate modality-based domain adaptation.

(a) Image-Level Generation: Generative adversarial networks (GANs) (Goodfellow *et al.*, 2014) and other image synthesis approaches (Isola, Zhu, Zhou & Efros, 2017b; Zhu, Park, Isola & Efros, 2017) have been widely explored to mitigate the modality gap by generating synthetic images that approximate the target modality. Methods such as ThermalGAN (Kniaz, Knyaz, Hladuvka, Kropatsch & Mizginov, 2018) and AlignGAN (Wang *et al.*, 2019b) attempt to generate visible-style images from infrared inputs (or vice versa), enabling more effective feature matching. While these models have shown promising results in reducing visual discrepancies between modalities, they often suffer from limitations such as image artifacts, identity information loss, and high computational costs during inference. Moreover, since these approaches treat image generation as a preprocessing step, they may not be optimized end-to-end with the downstream ReID objective. Recent findings suggest that optimizing directly in the feature space—rather than transforming pixel-level representations—may yield more robust and generalizable embeddings, especially under unseen conditions.

(b) Feature-Level Representation Learning: A dominant line of work in V-I ReID focuses on learning discriminative feature embeddings that are invariant across modalities. These methods (Dai *et al.*, 2018; Fang *et al.*, 2023; Ye *et al.*, 2020b; Ye, Lan, Wang & Yuen, 2020; Ye, Ruan, Du & Shou, 2021b; Wu, Liu, Su, Shi & Tang, 2023a) often employ shared convolutional backbones to extract embeddings from both visible and infrared inputs, while using loss functions (e.g., contrastive, triplet, or classification-based) to align representations. Some approaches explicitly disentangle features into identity-relevant and modality-specific components, aiming to suppress modality bias. Additionally, part-based models enhance discriminability by aligning and learning from spatially localized regions (e.g., body parts) rather than global appearances. However, these methods may still be skewed towards the more dominant modality in training data and can struggle with extreme illumination changes or partial occlusions. The incorporation

of attention mechanisms and contrastive learning techniques has improved feature robustness, but effectively aligning multi-modal embeddings remains a fundamental challenge.

(c) Intermediate Modality-Based Domain Adaptation: To further bridge the domain gap, another class of methods introduces an intermediate representation space between visible and infrared domains. Earlier works attempted to use fixed transformations, such as grayscale conversion (Zhu *et al.*, 2020b) or spectrum histogram equalization (Alehdaghi, Josi, Cruz & Granger, 2022), to approximate this intermediate modality. However, such handcrafted solutions often fail to capture identity-preserving features (Zhang, Yan, Lu & Wang, 2021c). In contrast, the AGPI^2 approach proposed in this thesis generates dynamic, data-driven intermediate modalities tailored to individual identities using adversarial learning. This allows the model to bridge visible and infrared domains more effectively during training. Building on this idea, the Bidirectional Multi-step Domain Generalization (BMDG) framework gradually aligns feature representations across modalities through structured learning stages and bidirectional feature adaptation. Unlike traditional domain adaptation that performs a single-step alignment, these progressive strategies ensure better feature stability and more robust cross-modal generalization.

A key limitation of prior work in V-I ReID is the reliance on static transformations (Alehdaghi *et al.*, 2022; Zhu *et al.*, 2020b) or fixed feature extraction pipelines (Zhang, Cisse, Dauphin & Lopez-Paz, 2018), which often fail under real-world conditions involving diverse lighting, clothing changes, and occlusions. This thesis addresses these challenges by proposing a series of novel frameworks that enhance domain generalization through intermediate modality learning, part-aware representation alignment, and feature disentanglement for mixed-modal matching. These contributions collectively improve V-I ReID performance under a variety of conditions, ensuring that learned representations remain robust, discriminative, and modality-resilient.

1.3 Metric Learning

Deep metric learning focuses on learning an embedding space where semantically similar samples are close together while dissimilar samples are pushed apart. It has been widely used in tasks such as face recognition (Schroff, Kalenichenko & Philbin, 2015), person re-identification (Hermans, Beyer & Leibe, 2017), and cross-modal retrieval (Ye, Shen, Lin, Shao & Xie, 2018b). Deep metric learning aims to map input samples into an embedding space where similar samples are closer together, while dissimilar samples are pushed apart. This technique is crucial for tasks such as person ReID, face recognition, and few-shot learning (Schroff *et al.*, 2015). Traditional metric learning approaches, such as contrastive loss (Hadsell, Chopra & LeCun, 2006) and triplet loss (Schroff *et al.*, 2015), have been widely adopted in deep learning frameworks. The contrastive loss minimizes the distance between positive pairs while maximizing the margin for negative pairs, whereas triplet loss refines this by considering a triplet of anchor, positive, and negative samples.

1.3.1 Loss Functions

Various loss functions and training strategies have been developed to enhance metric learning, ensuring better generalization and robustness. For example, the hard sample mining (Hermans *et al.*, 2017), adaptive margin losses (Deng, Guo, Xue & Zafeiriou, 2019), and self-supervised metric learning (Chen, Kornblith, Norouzi & Hinton, 2020b). In cross-modal retrieval, deep metric learning plays a key role in aligning visible and infrared features in the VI-ReID context (Ye *et al.*, 2018b). These methods often integrate center loss and angular loss to enhance feature separability, making them highly effective in reducing intra-class variance and improving retrieval performance (Wen, Zhang, Li & Qiao, 2016).

Contrastive Loss (Hadsell *et al.*, 2006) operates on pairs of samples and penalizes the model if similar pairs are too distant or dissimilar pairs are too close in the embedding space. It is one of

the earliest formulations used with Siamese networks. Mathematically, it is defined as:

$$\mathcal{L}_{\text{contrastive}} = h \cdot D(\mathbf{f}_i, \mathbf{f}_v)^2 + (1 - h) \cdot \max(0, m - D(\mathbf{f}_i, \mathbf{f}_v))^2, \tag{1.1}$$

where $h \in \{0, 1\}$ indicates same person or not ($h = (y_i == y_v)$, D is the Euclidean distance between embeddings, and m is a predefined margin.

Triplet Loss (Schroff *et al.*, 2015) extends this concept by introducing a triplet of samples: an anchor (a), a positive (p), and a negative (n). It encourages the distance between the anchor and positive to be smaller than the distance between the anchor and negative by at least a margin α:

$$\mathcal{L}_{\text{triplet}} = \max(0, D(\mathbf{f}_a, \mathbf{f}_p) - D(\mathbf{f}_a, \mathbf{f}_n) + \alpha). \tag{1.2}$$

Center Triplet Loss (Wu *et al.*, 2021) extends the triplet loss by pushing the embedding features to the center of each class:

$$\mathcal{L}_{\text{c-triplet}} = \max(0, D(\mathbf{f}_j, \mathbf{c}_{y_j}) - \min_{l \neq y_j} D(\mathbf{f}_j, \mathbf{c}_l) + \alpha), \tag{1.3}$$

where $\mathbf{c}_{y_j}$ is the center of all embeddings belonging to the y_j.

NT-Xent Loss (Normalized Temperature-scaled Cross Entropy) (Chen *et al.*, 2020b), popularized by SimCLR, uses cosine similarity and temperature scaling, τ, to contrast positive pairs against a large number of negatives in a mini-batch. It has shown excellent results in self-supervised learning settings:

$$\mathcal{L}_{j,k} = -\log \frac{\exp(-D(\mathbf{f}_j, \mathbf{f}_k)/\tau)}{\sum_{l=1}^{N} \mathbf{1}_{[l \neq j]} \exp(-D(\mathbf{f}_j, \mathbf{f}_l)/\tau)}. \tag{1.4}$$

These loss functions form the foundation of modern deep metric learning systems and are widely used in applications such as face recognition, person re-identification, and visual retrieval.

1.4 Domain Adaptation and Generalization

Domain adaptation (DA) is a fundamental problem in machine learning, particularly in cross-modal tasks such as Visible-Infrared Person Re-Identification (Ye *et al.*, 2020b). The goal of DA is to transfer knowledge from a source domain $\mathcal{D}_s$ (where labeled data is available) to a target domain $\mathcal{D}_t$ (where data distributions differ and labeled samples are scarce or absent). Traditional supervised learning approaches often suffer from a domain shift, where models trained on one domain fail to generalize well to another. In the case of V-I ReID, this challenge is exacerbated by the large modality gap between visible (RGB) and infrared (IR) images.

Domain adaptation strategies can be broadly categorized into three approaches: (1) **Discrepancy-based DA** (Tzeng, Hoffman, Darrell & Saenko, 2015; Ge & Yu, 2017; Peng & Saenko, 2018; Huang & Belongie, 2017; Isola, Zhu, Zhou & Efros, 2017a), which minimizes statistical differences between domains; (2) **Adversarial DA** (Peng, Wu & Ernst, 2018; Tzeng *et al.*, 2020; Isola *et al.*, 2017a; Bousmalis, Silberman, Dohan, Erhan & Krishnan, 2017), which employs generative adversarial networks (GANs) or domain discriminators; and (3) **Self-supervised DA**(Kim, Cha, Kim, Lee & Kim, 2017; Bousmalis, Trigeorgis, Silberman, Krishnan & Erhan, 2016; Zhuang, Cheng, Luo, Pan & He, 2015; Ghifary, Kleijn, Zhang, Balduzzi & Li, 2016), which leverages pseudo-labeling and contrastive learning to enhance domain transferability.

1.4.1 Discrepancy-Based Domain Adaptation

A key approach in domain adaptation is to align feature distributions between the source and target domains. Discrepancy-based methods aim to reduce divergence measures such as **Maximum Mean Discrepancy (MMD)** (Gretton, Borgwardt, Rasch, Schölkopf & Smola, 2012b) and **Correlation Alignment (CORAL)** (Sun & Saenko, 2016). These techniques operate by constraining the feature representations $\mathcal{F}(x)$ such that:

$$\min_{\theta} \mathcal{L}_{\text{task}} + \lambda D_{\text{div}}(P_s, P_t) \tag{1.5}$$

where $\mathcal{L}_{\text{task}}$ is the classification or identification loss, and D_{div} is a discrepancy measure between the source P_s and target P_t distributions.

1.4.2 Adversarial-Based Domain Adaptation

A more recent and powerful approach involves adversarial learning, where a domain discriminator is introduced to guide the feature extractor towards modality-invariant representations. The objective follows a min-max optimization:

$$\min_{\mathcal{F}} \max_{\mathcal{D}} \mathbb{E}_{x_s \sim P_s}[\log \mathcal{D}(\mathcal{F}(x_s))] + \mathbb{E}_{x_t \sim P_t}[\log(1 - \mathcal{D}(\mathcal{F}(x_t)))] \tag{1.6}$$

where $\mathcal{F}$ represents the feature extractor, and $\mathcal{D}$ is the domain discriminator trained to distinguish between source and target domain samples.

State-of-the-art works such as **HiCMD** (Choi *et al.*, 2020a) and **AlignGAN** (Wang *et al.*, 2019b) have demonstrated the effectiveness of adversarial domain adaptation in V-I ReID.

1.4.3 Self-Supervised and Pseudo-Label-Based Adaptation

More recent approaches in domain adaptation leverage **self-supervised learning (SSL)** and **pseudo-labeling techniques** to improve generalization without requiring explicit domain alignment. SSL methods, such as **contrastive learning** (Chen, Fan, Girshick & He, 2020c), train models by maximizing agreement between differently augmented views of the same instance:

$$\mathcal{L}_{\text{sspl}} = -\log \frac{\exp(\mathcal{F}(x)^\top \mathcal{F}(x^+))}{\sum_{x^-} \exp(\mathcal{F}(x)^\top \mathcal{F}(x^-))} \tag{1.7}$$

where x^+ is a positive sample (same identity), x^- is a negative sample (different identity) and $\cdot^\top$ is transpose operator.

1.4.4 Gradual Domain Adaptation

Gradual domain adaptation is an emerging approach that seeks to mitigate domain gaps by progressively aligning feature distributions through intermediate transitions ((Wang, Li & Zhao, 2022a; Zhang, Deng, Jia & Zhang, 2021a; Chen & Chao, 2021)). Unlike traditional domain adaptation techniques that attempt direct alignment between a source and target domain, gradual adaptation introduces intermediate domains that bridge the distributional discrepancy in a structured manner. This strategy is particularly effective in Visible-Infrared Person Re-Identification (VI-ReID), where spectral differences between visible and infrared images present a significant challenge for cross-modal retrieval. By incorporating intermediate feature spaces, models learn smoother transitions across modalities, leading to improved generalization and reduced modality bias. Some studies (Gong, Liu, Li & Tao, 2019; Wang, Wang & Liu, 2022b) have demonstrated that progressive feature refinement can enhance robustness in cross-domain learning by avoiding abrupt distribution shifts that commonly occur in direct-domain adaptation strategies. Moreover, this gradual adaptation process can be facilitated by self-supervised learning mechanisms, contrastive feature refinement, and adversarial intermediate feature generation, all of which contribute to more stable and transferable feature representations (Pei, Cao, Long & Wang, 2018; You, Pan, Wang, Long & Jordan, 2019).

Extending this idea, multi-step domain adaptation incorporates a hierarchical approach where adaptation is performed iteratively across multiple levels, from low-level pixel alignments to high-level semantic representations. In the context of VI-ReID, this strategy ensures that domain alignment occurs in a structured manner, progressively refining both modality-invariant and identity-discriminative features. Our methods, as Bidirectional Multi-Step Domain Generalization (BMDG), propose a stepwise feature refinement process where models adapt to intermediate feature distributions before fully generalizing to the target domain. This approach mitigates the risk of negative transfer, which occurs when direct adaptation forces alignment between highly dissimilar distributions. Studies by (Zhang, Deng & Liu, 2021b) and (Pan, Yao & Sun, 2020) emphasize that multi-step adaptation leads to improved feature robustness by ensuring incremental learning, thereby preventing overfitting to either the source or target

domain. Moreover, multi-step strategies often integrate bidirectional learning mechanisms that allow simultaneous adaptation in both source-to-target and target-to-source directions, further enhancing the model's stability across heterogeneous domains (Luo, Wang & Li, 2020).

1.5 Learning Under Privileged Information

Standard supervised machine learning aims to identify a decision function that minimizes generalization error on unseen test data, relying solely on a labeled training set. For simpler tasks, the learner converges to an optimal value relatively quickly as the training set size increases. However, when the problem is complex, and the space of potential decision functions is vast, the convergence rate slows significantly (Pechyony & Vapnik, 2010b). To mitigate this, Vapnik and Vashist (Vapnik & Vashist, 2009) introduced the concept of Learning Under Privileged Information (LUPI).

The LUPI paradigm mimics the teacher-student learning dynamic, which is highly effective in human cognition. When a student learns a novel concept, a teacher provides comments, explanations, and examples—information that aids the learning process but is not available (or necessary) during the final examination. Formally, this framework builds a classifier to predict labels y based on an input space $\mathcal{X}$. However, during the training stage, the model is granted access to additional "privileged information" from a space $\mathcal{X}^\star$, which is strictly absent during the inference stage.

Consider a machine learning problem defined by a compact input space $\mathcal{X}$, a label space $\mathcal{Y}$, and a loss function $\mathcal{L}(\cdot,\cdot)$ that evaluates the discrepancy between a prediction and the ground truth. In the standard setting, given a function class $\mathcal{F}(\cdot;\mathbf{w})$ parameterized by $\mathbf{w}$ and a data distribution $x, y \sim p(x,y)$, the objective is to solve:

$$\min_{\mathbf{w}} \mathbb{E}_{x,y\sim p(x,y)}[\mathcal{L}(\mathcal{F}(x;\mathbf{w}),y)]. \tag{1.8}$$

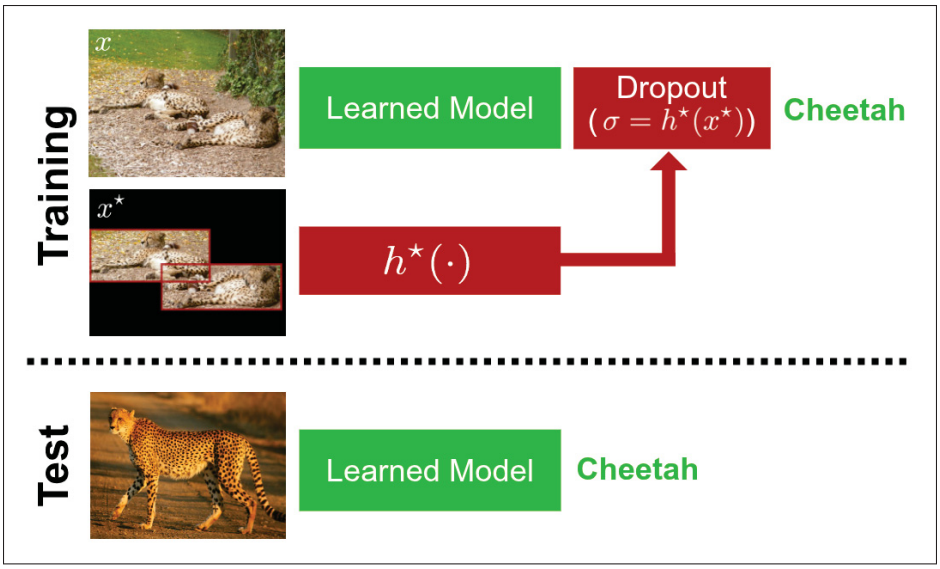


Figure 1.2 In the Learning Under Privileged Information (LUPI) paradigm, a teacher provides additional information during training
Taken from (Lambert *et al.*, 2018)

In the LUPI framework, we incorporate the privileged information available during training by defining a parametric function class for multi-view data, $\mathcal{F}^+ : \mathcal{X} \times \mathcal{X}^\star \rightarrow \mathcal{Y}$. The optimization problem then becomes:

$$\min_{\mathbf{w}} \mathbb{E}_{x,x^\star,y\sim p(x,x^\star,y)}[\mathcal{L}(\mathcal{F}^+(x,x^\star;\mathbf{w}),y)]. \tag{1.9}$$

During inference, since the privileged information $x^\star$ is unavailable, the decision function $h(x;\mathbf{w})$ must be derived by marginalizing out $x^\star$:

$$h(x;\mathbf{w}) \equiv \mathbb{E}_{x^\star\sim p(x^\star|x)}[\mathcal{F}^+(x,x^\star;\mathbf{w})]. \tag{1.10}$$

However, because privileged information often contains features orthogonal to the main modality (i.e., information not present in the visible training data), the conditional distribution $p(x^\star|x)$ is generally unknown. This makes the direct calculation of the expectation in Eq. (1.3) intractable. To address this, Lambert *et al.* (2018) proposed restricting the class of functions such that the

expectation becomes computable. They introduced a parametric family where the standard information controls the mean of the prediction, while the privileged information controls the variance:

$$\mathcal{F}^{+}(x, x^{\star}; \mathbf{w}) = \mathcal{F}^{\circ}(x; \mathbf{w}^{\circ}) \odot \mathcal{N}(\mathbf{1}, \mathcal{F}^{\star}(x^{\star}; \mathbf{w}^{\star})), \tag{1.11}$$

where $\odot$ denotes the Hadamard product. The stochastic term $\mathcal{N}(\mathbf{1}, \mathcal{F}^{\star}(x^{\star}; \mathbf{w}^{\star}))$ represents a normal random variable with a constant mean and a covariance parameterized by the privileged input $x^{\star}$ and weights $\mathbf{w}^{\star}$. Consequently, the parameters of the training function h^{+} are the concatenation of the main and privileged parameters, $\mathbf{w} = [\mathbf{w}^{\circ}, \mathbf{w}^{\star}]$.

CHAPTER 2

ADAPTIVE GENERATION OF PRIVILEGED INTERMEDIATE INFORMATION FOR VISIBLE-INFRARED PERSON RE-IDENTIFICATION

Mahdi Alehdaghi[1] , Arthur Josi[1] , Pourya Shamsolmoali[2] , Rafael M. O. Cruz[1] , Eric Granger[1]

[1] LIVIA, ILLS, Dept. of Systems Engineering, École de Technologie Supérieure,
1100 Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3
[2] University of York, U.K.

The article was published in IEEE Transactions on Information Forensics and Security,
February 2025

2.1 Introduction

Person re-identification (ReID) involves matching images or videos of individuals of interest captured across multiple non-overlapping cameras. Current methods for person ReID, such as those based on deep Siamese networks (Fu *et al.*, 2021b; Sharma, Kapil & Chapman, 2021; Somers, De Vleeschouwer & Alahi, 2023), mainly use visible images to train a feature embedding backbone for similarity matching metric learning losses.

Although deep learning (DL) models have made significant progress in recent years, person ReID remains a challenging task due to the non-rigid structure of the human body and variations in capture conditions in real-world scenarios. These variations can include illumination variations, blurring of motion, resolution, pose, viewpoint, occlusion, and background clutter (Bhuiyan *et al.*, 2020; Mekhazni, Bhuiyan, Ekladious & Granger, 2020). Moreover, visible ReID systems cannot perform well in low-light environments (e.g., at night), thereby limiting their practical application.

Visible-infrared ReID (V-I ReID) is a variant of person ReID that involves matching individuals across RGB and IR cameras. V-I ReID is challenging since it requires matching individuals across different modalities with significant differences in appearance (Lu, Lin & Hu, 2023; Du, Lei, Zhao, Dong & Su, 2024; Zhang, Kang, Zhao & Shen, 2023). For robust cross-modal matching, a V-I ReID system must train a feature extractor that encodes pedestrians' images

into feature information, which is interchangeable between V and I modalities. Then, during inference, images of people are matched across RGB and IR cameras based on these features. State-of-the-art methods for V-I ReID mainly focus on image-level (Wang *et al.*, 2019a; Kniaz *et al.*, 2018) or feature-level alignment (Ye *et al.*, 2020b; Ye *et al.*, 2020; Liu & Cheng, 2019; Zhu *et al.*, 2020b; Park, Lee, Lee & Ham, 2021). Through the use of generative modules, the former aims to bridge the modality discrepancy by encoding images into feature space and then decoding and translating from different modalities to others. The latter obtains a modality-invariant embedding space by learning a shared model to represent images in feature space in which the discrepancy is minimized. Since V and I modalities have distinct data distributions, which results in a large discrepancy between V and I images in appearance and statistics (Wu, Zheng, Yu, Gong & Lai, 2017b), the features extracted from such models tend to be biased on modality-specific information and not be proper for cross-modal matching.

Addressing this challenge involves considering an intermediate space that provides meaningful information about the primary modalities, guiding the model to focus on shared knowledge. However, the availability of such intermediate spaces is not guaranteed, and their creation or identification is not a straightforward task. Additionally, determining how to effectively utilize bridging information during both the training and testing stages poses a key aspect of this solution.

The Learning Under Privileged Information (LUPI) paradigm (Vapnik & Vashist, 2009) emerges as a solution aiming to leverage additional data available only during training. This approach has been adopted by various V-I ReID methodologies to bridge the gap between V and I images. For instance, works such as (Alehdaghi *et al.*, 2022; Fan, Luo, Zhang & Jiang, 2020; Li, Wei, Hong & Gong, 2020a; Wei, Yang, Wang & Gao, 2021; Wu, Zheng, Yu, Gong & Lai, 2017a; Ye, Shen & Shao, 2020c; Zhang *et al.*, 2021c) concentrate on exploring subspaces between the IR spectrum and RGB channels, creating intermediate domains to address the modality gap. However, existing intermediate generative approaches fail to effectively adapt or generalize the intermediate information across V and I domains. The absence of specific objectives for generalizing intermediate images results in a limited ability to adapt to significant V-I

domain shifts. To better capitalize on modality-invariant attributes, V-I ReID models necessitate appropriate intermediate images capable of bridging the substantial domain gap between I and V appearances. While utilizing intermediate images during testing enhances matching accuracy (Kniaz *et al.*, 2018; Wang *et al.*, 2019a), it introduces complexity and may not be ideal for real-time applications.

This paper introduces an approach for the Adaptive Generation of Privileged Intermediate Information (AGPI^2 approach) that trains a backbone model for V-I person ReID by adapting a new intermediate space that connects V and I modalities. Our proposed AGPI^2 training approach comprises three key modules: a feature embedding backbone, a generator, and an ID-modality discriminator. Since each modality can be seen as a domain with its data distribution, we propose treating each modality as a separate domain. Then, our AGPI^2 aim is to generate an intermediate domain with similar attributes to the source V and I domains. This intermediate space enables the model to learn and extract common attributes between the two modalities. At inference time, the model can match the same person across modalities by representing them in a non-modality-specific feature space without relying on intermediate information. To achieve this, our framework adapts intermediate images to have the same content (body pose, background, and texture) as V images while sharing the style information (heat or temperature information) with I images. The AGPI^2 component provides an intermediate step for the feature embedding backbone, helping to reduce modality-specific information in the extracted features.

Given the absence of supervised information to generate the intermediate domain containing ID-aware information and bridging modality attributes, we propose an end-to-end approach with adversarial objectives w.r.t. the ultimate goal of V-I ReID, which is extracting unique and modality-free descriptors for each person. To this end, the generative module is encouraged to transform the V images to an intermediate domain to minimize the domain gap between such virtual and I images. Consequently, the generative module must focus on ID-aware attributes to avoid the intermediate images missing the discriminative information about the person in the V images. To reach this goal, the ID-modality discriminator is proposed, which seeks to reinforce the generator by discriminating modalities from ID-related features. Such property in

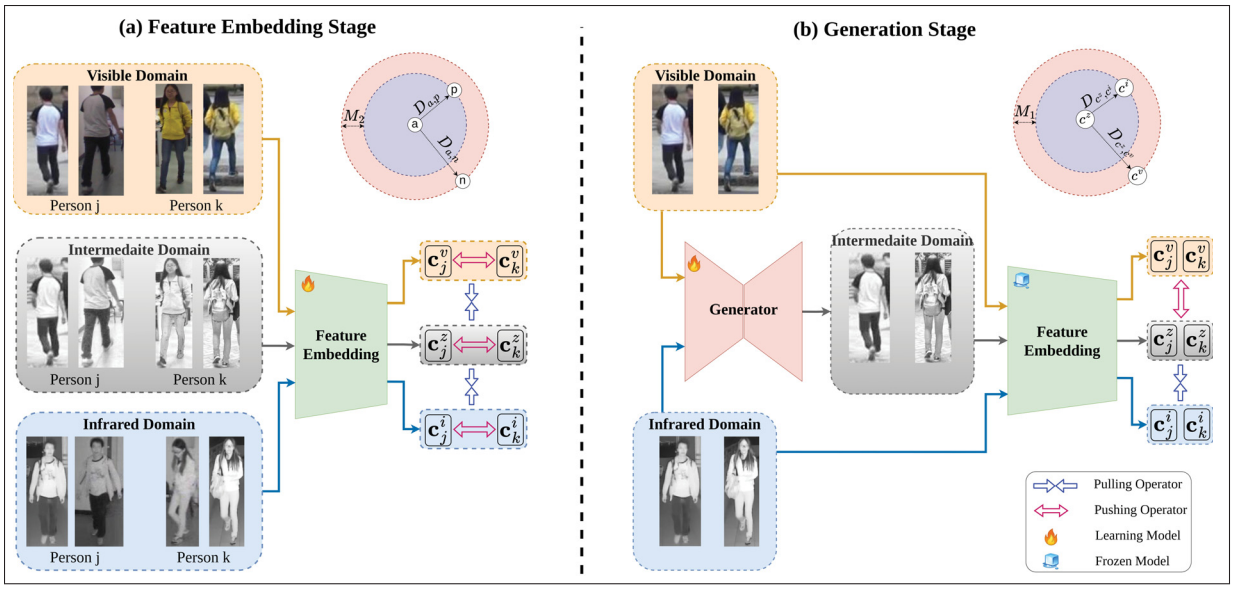


Figure 2.1 A illustration of the training strategy with our AGPI^2 approach. The generated images are learned to be similar to the infrared modality by using adversarial objectives. (a) The feature embedding stage pushes the extracted features to approach the intermediate domain, while (b) the generation stage transforms V images to an intermediate domain that approaches I images. The objective of feature embedding is to minimize the distance from the anchor to the positive ($D_{a,p}$) and maximize its distance from the negative samples ($D_{a,n}$). Let c^v, c^z, and c^i be the center of the feature distributions for each class of V, intermediate, and I images. The generator tries to create a space in which the ID-aware features of each person are close to the infrared and visible at the same time. (To reach this goal and make the transferred images not be biased toward visible, we push them more toward infrared: $D_{c^z,c^i} < D_{c^z,c^v} + M_1$)

the discriminator forces the generator to focus on ID-aware information. The feature embedding module simultaneously works to minimize the gap between the generated and V images, as shown in the left side of Fig. 2.1, while the generator's objective is to maximize this distance (right side of 2.1). This collective approach ensures the effective generation of intermediate domain images that capture both ID-aware information and modality attributes.

In other words, AGPI^2 tackles the significant gap between modalities by generating and utilizing bridging information that links the two domains. Since such information is not available all the time or hard to benefit from, our AGPI^2 training approach is on two stages: (a) Finding a privileged linking domain between two modalities and (b) making the model robust against

modality change by leveraging the intermediate domain as a clue for containing shared attributes among two modalities. In comparison with (Yu *et al.*, 2023; Dai *et al.*, 2021) that are creating an intermediate domain in the latent feature space, our intermediate domain is in the input space. Therefore, our generative module can be used in other applications to generate intermediate images between visible and infrared and even in other V-I person ReID. In Fig. 2.1, the generative module is depicted as trained to generate images with features closer to I and farther from V, enforcing a margin M_1 between the distances from the center of the feature vector representing images of each person in the intermediate space to V (D_{c^z,c^v}) and I (D_{c^z,c^i}). Simultaneously, the feature embedding stage ensures close similarity between the feature vectors of each person in the visible, intermediate, and infrared modalities.

Our main contributions are summarized as follows: (1) An adversarial training strategy called AGPI^2 is proposed to reduce the V and I domain gaps by adopting a generalized intermediate domain as privileged information.

(2) By analyzing the mutual information between intermediate space and V-I person ReID objectives, adversarial losses are proposed between the ID-modality discrimination, feature embedding, and generation modules to encourage the transformed images to incorporate discriminative ID information while reducing the influence of domain-specific attributes. Such losses make the modules play a min-max game between each other, resulting in a balanced and informative intermediate space.

(3) An extensive set of experiments on the challenging SYSU-MM01 (Wu, Zheng, Gong & Lai, 2020) and RegDB (Nguyen *et al.*, 2017) datasets indicate that our proposed AGPI^2 not only outperforms state-of-art methods for cross-modal V-I person ReID but can also be seamlessly integrated into other methods to enhance their accuracy without imposing additional computational overhead during inference. In other words, our framework is agnostic and can be applied to other image retrieval applications to improve their performance.

2.2 Related Work

(a) Visible-Infrared Person ReID:

This paper focuses on V-I ReID, where deep backbone models are trained to match individuals between V and I cameras using a labeled set of V and I images captured using RGB and IR cameras, respectively. Then, during inference, different individuals are matched across camera modalities, that is, matching query I images to gallery V images, or vice versa (Chen, Wan, Li, Jing & Sun, 2021; Dai *et al.*, 2018; Fu *et al.*, 2021a; Hao, Li, Wang & Gao, 2020; Park *et al.*, 2021; Xu, Wu, Liu & Xiao, 2021; Ye, Lan, Li & Yuen, 2018a; Zhu *et al.*, 2020b). Existing DL models used for V-I ReID can be divided into Generative and Representative models. The former seeks to map raw data between two modalities to reduce modality discrepancies at the image level, while the latter focuses on finding a suitable discriminant representation by reducing the cross-modality gap at the feature level.

Some generative models, like ThermalGAN (Kniaz *et al.*, 2018) or AlignGAN (Wang *et al.*, 2019a) employed autoencoders to translate V and I images and found a common feature space through the generation process. D^2RL (Wang, Wang, Zheng, Chuang & Satoh, 2019d) generated fake-real paired images from synthetic V and I images and encoded them into a common feature space for feature representation and matching. Other models (Choi *et al.*, 2020a; Wang *et al.*, 2020a; Yang *et al.*, 2020) relied on GANs to create synthetic images conditioned by a modality-invariant representation. However, these methods are complex at inference time, requiring image generation at test time. Moreover, they do not necessarily produce high-quality images that are capable of bridging the gap between modalities.

Representation approaches focus on training a backbone model to learn a discriminant representation that captures the essential features of each modality while exploiting their shared information (Ye *et al.*, 2018a; Dai *et al.*, 2018; Ye *et al.*, 2020; Wu *et al.*, 2017a; Liu & Cheng, 2019; Ye *et al.*, 2020b) or exploiting self-similarity (Wu *et al.*, 2022). In (Ye *et al.*, 2020b,c, 2021b; Jiang *et al.*, 2020; Park *et al.*, 2021; Chen *et al.*, 2020a), authors only used global features combined through multimodal fusion. The absence of local information limits

their accuracy in more challenging cases. To improve robustness, recent methods combined global information with local part-based features (Ye, Shen, J. Crandall, Shao & Luo, 2020a; Lu *et al.*, 2020; Huang, Liu, Li, Zheng & Zha, 2022; Wu *et al.*, 2021; Fang *et al.*, 2023; Kim, Kim, Park, Park & Sohn, 2023) or fused features of different resolutions similar to LRAR (Wu *et al.*, 2023b), which used two resolution-adaptive mechanisms. For example, (Ye *et al.*, 2020a; Lu *et al.*, 2020) divided the spatial feature map into fixed horizontal sections and applied a weighted-part aggregation. Using the weighted mechanism, these methods could dynamically pay attention to the relative importance of different parts to improve their discrimination power. (Fang *et al.*, 2023; Wu *et al.*, 2021, 2023a) dynamically detected regions in the spatial feature map to address the misalignment of fixed horizontally divided body parts.

In contrast, the proposed AGPI^2 uses a generative model only during the training phase to generate the privileged information to help the training process by bridging the modality gap. This makes our proposed method computationally efficient, relying only on the feature embedding backbone during inference.

(b) Intermediate Modality in V-I ReID: Given the significant gap between V and I distributions, cross-modal matching is a challenging problem. One possible approach is to use an intermediate modality to help the model find common semantics between these two modalities. For example, the model (Wu *et al.*, 2017a) relied on gray images instead of colored images. However, using grayscale alone limited the amount of valuable information from V images by not containing the V spectrum information. Therefore, Fan et al. (Fan *et al.*, 2020) combined the information from the three channels (V, I, and grayscale) and (Ye *et al.*, 2020c) used grayscale images as an intermediate modality during the training process, while in (Ye *et al.*, 2021b; Liu, Ma, Xia & Li, 2021; Ye, Chen, Shen & Shao, 2021a) augmented images were created by randomly selecting one value from the RGB color as the intermediate modality. Alehdaghi et al. (Alehdaghi *et al.*, 2022) introduced a novel approach involving the use of random linear combinations of RGB colors. This technique served as an effective bridge, facilitating the extraction of color-independent features and ensuring that the derived attributes are not only devoid of color bias but also contain characteristics commonly found in infrared images. (Ye *et al.*, 2021a) performed the

channel-level interaction by randomly determining whether to keep the original RGB image or perform a random channel selection. However, these methods made limited use of intermediate domains by searching for them in a small search space.

To address this limitation, methods in (Li *et al.*, 2020a; Wu *et al.*, 2017b; Zhang *et al.*, 2021c; Lu *et al.*, 2023) used generative models to create the wider search space for intermediate modality. For example, (Li *et al.*, 2020a) used a single convolution layer to transform V images into a new modality without using the I or identity information. (Wu *et al.*, 2017b) proposed a pixel-to-pixel feature fusion operation on V and I images to build the synthetic images, and (Zhang *et al.*, 2021c) introduced a shallow auto-encoder to generate intermediate images from both modalities. The images generated by these methods were biased toward the source modality. Lu et al. (Lu *et al.*, 2023) designed a channel interactive generator to generate confused modality images to address this problem. While such methods enhance performance, they ignore the intricate V-I image relationship and fail to take into account domain gaps between id-discriminative information during the generation of intermediate domains. In contrast, our approach focuses on minimizing domain shifts between visible and infrared to identify the most informative intermediate space. It also employs an ID-modality discriminator to focus on the person's identity, allowing the generative model to robustly create persons' images from one modality into an intermediate while preserving the discriminative attributes of that person. Thus, the quality of the intermediate image and the id-aware information is significantly improved.

(c) Learning Under Privileged Information: Supervised learning is based on labeled data to optimize an appropriate decision function, minimizing the generalization error. The learner's function typically converges rapidly to the optimal solution for simple learning tasks with a large amount of training data. However, the convergence rate becomes slow for more challenging tasks and larger optimization spaces (Pechyony & Vapnik, 2010a). The LUPI approach was initially proposed to accelerate the convergence of SVMs by incorporating prior knowledge into the learning process. In (Hoffman, Gupta & Darrell, 2016) used a hallucination network to introduce prior knowledge into an additional stream in CNN to extract depth-related attributes. Beyond seeking faster convergence, LUPI has been successfully applied in various applications to improve

performance (Choi, Kim & Ramani, 2017; Kampffmeyer, Salberg & Jenssen, 2018; Kumar, Banerjee & Chaudhuri, 2021; Lezama, Qiu & Sapiro, 2017; Saputra *et al.*, 2020). In (Pande, Banerjee, Kumar, Banerjee & Chaudhuri, 2019), a GAN-based model is proposed to generate prior knowledge at the testing time, and Crasto et al. (Crasto, Weinzaepfel, Alahari & Schmid, 2019) distilled this knowledge from privileged information to the V branch. Although the use of PI improves model performance, such proper and supervised information is unavailable or difficult to leverage. In this study, our focus lies in implementing the LUPI paradigm for V-I ReID, aiming to identify suitable and informative Privileged Information (PI) during training. To achieve this objective, we propose a generative model designed to dynamically transform modalities into an intermediate space, serving as privileged information exclusively during the training phase through the use of adversarial losses.

2.3 Proposed Method

Our AGPI^2 approach generates a virtual PI domain that bridges V and I modalities during training and then leverages it to learn modality-invariant features. Fig. 2.2(a) shows the overall AGPI^2 training architecture, comprising three modules: (1) The generator module that transforms V images into intermediate ones by incorporating infrared features. These intermediate images help the model simulate an infrared version of the visible images, allowing it to focus on non-color-specific attributes from the visible modality. Since such images have exact pose information of individuals and only lose the colors, the model accesses one intermediate step before seeing infrared images with larger shifts (different pose, background, and modality). (2) ID-modality discriminator that identifies modality from ID-aware features to avoid the intermediate images being over-focused on the background, which is likely to be dominant for modality transformation, and (3) A feature embedding module that extracts shared features from V and I images, guided by the proposed bridging losses to effectively minimize modality discrepancies.

During training, the ID-modality discriminator is taught to classify individuals in the intermediate and V images, considering them as belonging to the same class but distinct from the I images.

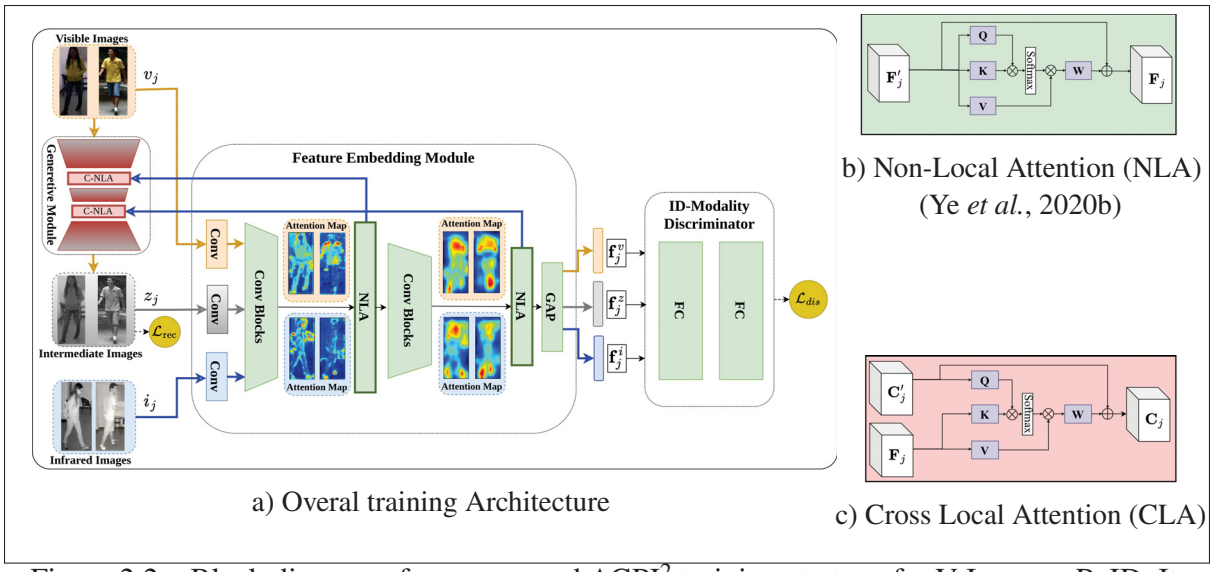


Figure 2.2 Block diagram of our proposed AGPI^2 training strategy for V-I person ReID. It takes advantage of the adaptive intermediate domain to reduce the distribution gap between the V and I domains. The feature embedding module seeks to minimize the distance between the two original (I and V) modalities by using the auxiliary generated intermediate domain. The ID-modality discriminator learns to detect the modality from ID-aware features, and the generation process aims to transform V images to infrared through adversarial training with the discriminator

This is achieved through the use of identity discriminative features extracted by the backbone. The primary goal is to guide the generative module in prioritizing the capture of personal attributes aligned with the characteristics associated with modality-specific information. Concurrently, in the generation learning stage, the generator aims to produce intermediate images that the discriminator would adversarially classify as belonging to the I modality. This iterative process leads to the generated images becoming more similar to the I images, thereby enhancing the model's ability to extract common features from both V and I modalities.

Preliminaries: Let us assume a multimodal dataset for V-I ReID problems that is represented by $\mathcal{S} = \{\mathcal{V}, \mathcal{I}\}$, where $\mathcal{V} = \{v_j\}_1^n$ are V and $\mathcal{I} = \{i_j\}_{n+1}^{n+m}$ are I images of N different persons, and $\mathcal{Y} = \{y_j\}_1^{n+m}$ contains their identity for the training process. Metric losses $\mathcal{L}$ are typically defined based on distances D between feature vectors $\mathbf{f}_j^v$ of V images and feature vectors $\mathbf{f}_j^i$ of I

images for every two distinct individuals, as:

$$\mathcal{L} = \sum_{j=1}^{n+m} C(D(\mathbf{f}_j^v, \mathbf{f}_j^i), D(\mathbf{f}_j^v, \mathbf{f}_k^v), y_j, y_k), \tag{2.1}$$

where C is metric loss objectives which can be triplet-loss (Chopra, Hadsell & LeCun, 2005) or contrastive loss (Schroff *et al.*, 2015), y_j and y_k represent the identity of person in images j and k, and $\mathbf{f}_j^v, \mathbf{f}_k^i \in \mathbb{R}^d$, where d the features dimension.

2.3.1 LUPI Framework

Given that the generated images are unavailable at test time, we formulate our proposed learning strategy according to the LUPI paradigm. With LUPI (Lambert *et al.*, 2018), the DL model leverages privileged information (PI) exclusively during the training phase, while the main information remains accessible in both the training and testing phases. For example, in cross-modal matching, the model typically relies on the V and I modalities during the training phase to leverage shared discriminative features between inputs. However, during the inference phase, the model only inputs the query image from one modality (V or I). Also, during the training phase, the model can use privileged bridging information between the two modalities to address the significant domain discrepancies between V and I images (Pechyony & Vapnik, 2010a). By incorporating the intermediate modality (Z), AGPI^2 reduces the modality gap between V and I, enabling the model to generalize more effectively and improve its ability to transfer knowledge across both modalities.

2.3.2 Mutual Information Analysis

We formulate the bridging characteristics objectives of the privileged intermediate information from the mutual information perspective, and demonstrate that jointly learning id-aware and modality-free objectives achieves a conditional mutual information maximization between

intermediate images and identity space while discarding the modality style information:

$$\max_{Z} \quad \mathrm{MI}(Z; Y|M), \tag{2.2}$$

where Y is the identity of person input images, M is the modality label, and $\mathrm{MI}(.;.)$ is the mutual information between two random variable. For addressing Eq. (2.2), we can expand it as follows:

$$\mathrm{MI}(Z; Y|M) = \overbrace{\mathrm{MI}(Z; Y)}^{\textit{ID-Aware}} - \overbrace{\mathrm{MI}(Z; Y; M)}^{\textit{ID-Modality-Aware}} . \tag{2.3}$$

To maximize this, the first term (ID-aware) should be maximized and the second term (ID-Modality-Aware) minimized. To minimize the ID-Modality-Aware part for intermediate images, an adversarial training approach is proposed between the ID-Modality Discriminator and the generative module, in which the former maximizes that part by looking for id-related information specific to the modality.

To effectively capture the distinct roles of the latent representation, we propose decomposing the latent space Z into two subspaces: Z_1, which captures the primary predictive information regarding the target Y, and $Z_2 = \mathcal{D}(Z)$, which captures the synergistic information arising from the interaction between Y and the auxiliary variable M.

We formulate this disentanglement as a dual-objective optimization problem. Z_1 is optimized to maximize the mutual information with the target Y, ensuring high predictive performance. Conversely, Z_2 is optimized to minimize the Interaction Information $I(Z_2; Y; M)$. This objective encourages the extraction of synergistic features—patterns that are only visible when considering the joint distribution of $\{Y, M\}$—while suppressing redundant marginal information.

The optimization problem is formally defined as:

$$\max_{Z_1} \mathrm{MI}(Z_1; Y) \quad \text{and} \quad \min_{Z_2} \mathrm{MI}(Z_2; Y; M) \tag{2.4}$$

2.3.2.1 Optimization of the Predictive Component (Z_1)

For the first component $Z_1 = Z$, the objective follows standard supervised learning. We maximize the mutual information $I(Z_1; Y)$ by minimizing the cross-entropy loss $\mathcal{L}_{CE}$ of a classifier f_1 trained to predict Y:

$$\mathcal{L}_{Z_1} = \mathcal{L}_{CE}(f_1(Z_1), Y) \approx -\mathbb{E}_{Z_1,Y}[\log p(Y|Z_1)] \tag{2.5}$$

2.3.2.2 Optimization of the Synergistic Component (Z_2)

For the second component Z_2, we minimize the Interaction Information to isolate unique joint dependencies. Expanding the interaction term for fixed labels Y and M yields:

$$\text{MI}(Z_2; Y; M) = \text{MI}(Z_2; Y) + \text{MI}(Z_2; M) - \text{MI}(Z_2; \{Y, M\}) \tag{2.6}$$

Minimizing this quantity requires a min-max adversarial approach. We aim to maximize the joint information $I(Z_2; \{Y, M\})$ (predicting the pair) while simultaneously minimizing the marginal information terms $\text{MI}(Z_2; Y)$ and $\text{MI}(Z_2; M)$ (unlearning the individual labels).

Using the variational bound of cross-entropy, we construct a loss function $\mathcal{L}_{Z_2}$ composed of three classification heads: a joint head f_{joint} and two adversarial marginal heads f_Y and f_M.

$$\mathcal{L}_{Z_2} = \mathcal{L}_{CE}(f_{joint}(Z_2), \{Y, M\}) - \lambda \left[\mathcal{L}_{CE}(f_Y(Z_2), Y) + \mathcal{L}_{CE}(f_M(Z_2), M)\right] \tag{2.7}$$

where λ is a hyperparameter balancing synergy extraction and redundancy removal. In this formulation, minimizing $\mathcal{L}_{CE}(f_{joint}, \cdot)$ maximizes $\text{MI}(Z_2; \{Y, M\})$, while maximizing the negated terms (via adversarial training, such as a Gradient Reversal Layer) minimizes $I(Z_2; Y)$ and $\text{MI}(Z_2; M)$. This forces Z_2 to encode information that predicts the joint configuration of Y and M without being able to predict either variable independently.

At the same time, the latter tries to create Z images that fool the discriminator, thereby minimizing this term.

2.3.3 ID-Modality Discriminator

For the generator to create intermediate images that contain id-related information, we propose a module to discriminate modality along with the identities of each individual in images. To achieve this, we employ an MLP network as our ID-modality Discriminator to classify modality types (Infrared or Visible) from features extracted from the shared embedding backbone. The generator's objective is to deceive the discriminator by creating a modality that appears false for a specific identity.

To implement this idea, we doubled the label space of each identity to train the discriminator to differentiate between the same identity in the V and I modalities:

$$y'_j = \begin{cases} 2y_j & \text{V and Intermediate modalities} \\ 2y_j + 1 & \text{I modality} \end{cases} \tag{2.8}$$

where y_j is the original label.

For training the ID-modality discriminator, the objective is defined as:

$$\mathcal{L}_{\text{dis}} = \sum_{l=1}^{2N} -y'_{j,l} \log(p'_{j,l}), \tag{2.9}$$

where $p'_{j,c}$ is the predicted probability observation j of class c by the discriminator $\mathcal{D}$.

Fig. 2.3 shows the main difference between the general modality discriminator and our ID-modality discriminator. Each training sample is a shape labeled with the target label. Other SOTA intermediate generative approaches use ordinary binary discriminators that are trained to identify only modality (V and intermediate as the same), disregarding the person's identity. In

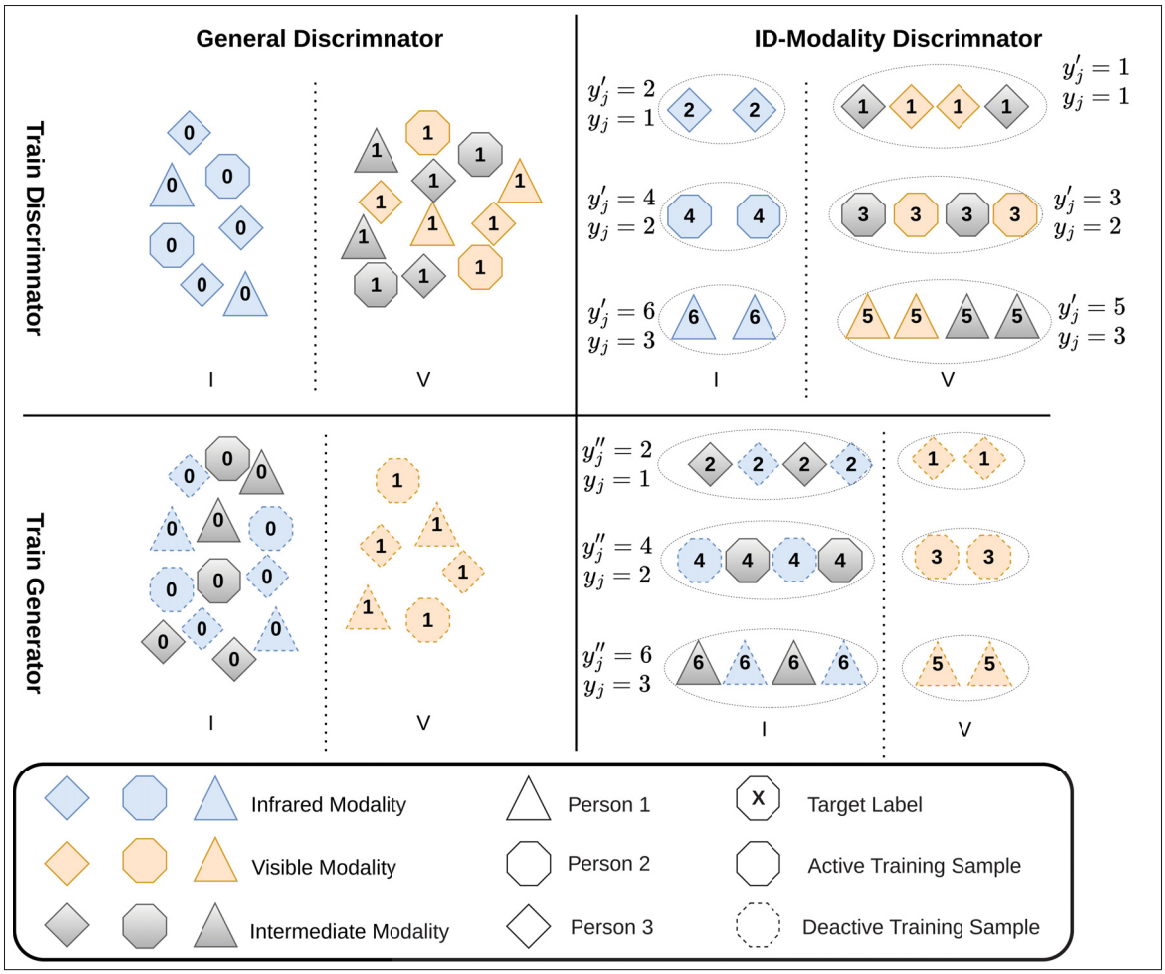


Figure 2.3 ID-Modality discriminator vs. general discriminator. Our label space is doubled to account for identity (differentiate between individuals in each modality)

contrast, by proposing new label spaces that are related to the identities of the discriminator, it would be able to detect V/I modalities from the ID-aware attributes of samples during training.

2.3.4 Generative Module

To generate the intermediate images, an adaptive and unsupervised intermediate generative model $\mathcal{G}$ is proposed to transfer visible images to intermediate modality in an end-to-end fashion regarding Eq. (4.2). The reason behind using visible images is that the visible images contain more modality-specific id-discriminative information (Jia *et al.*, 2020) and need to be filtered to

remove this biased information. To have information from both modalities in the generation process, the V images, at first, are encoded as features vectors and fused with I features F_j^i by the cross non-local attention (Wang, Girshick, Gupta & He, 2018), and then the decoder reconstructs the image, $z_j = \mathcal{G}(v_i, F_j^i)$. To allow this module to focus on ID-aware information in intermediate images and maximize the $\mathrm{MI}(Z;Y)$, the cross-entropy between identity space and intermediate images must be minimized (we proved this in Appendix I.A.1). So the $\mathcal{L}_{\text{id}}$ which is a cross-entropy loss function between the features extracted from intermediate images ($\mathbf{f}_j^z$) and identity label y_j is proposed to be minimized. Also, to filter them from any modality-specific information, these images should maximize the $\mathcal{L}_{\text{dis}}$ when they are fed to the ID-modality discriminator. To make this, we need a new label to fool the discriminator, and a label that specifies the same identity but in a different modality is chosen:

$$y''_j = 2y_j + 1. \tag{2.10}$$

The reason is that since this discriminator is not binary, every label except the ground truth with which the discriminator is trained can be selected as an opposite label. So, the proper label should be selected to avoid interrupting the id-discriminative information in the created images.

Also, since the Z images are created by transferring from V to intermediate space, they tend to be biased toward visible space. To tackle this issue, we pushed their feature more toward infrared than visible features. This adversarial objective can be defined as:

$$\mathcal{L}_{\text{adv}} = \sum_{l=1}^{2N} -y''_{j,l} \log(p''_{j,l}) + \max(0, M_1 + D_{\mathbf{c}_j^z, \mathbf{c}_j^i} - D_{\mathbf{c}_j^z, \mathbf{c}_j^v}), \tag{2.11}$$

where p''_j is the predicted label j from the discriminator, and $\mathbf{c}_j^z$, $\mathbf{c}_j^i$,$\mathbf{c}_j^v$ are the centers of feature embedding of generated I, and V images, respectively, for the person with identity y_j. The function D measures the distance between two vectors.

To generate more realistic images, we also transfer I images into the intermediate domain, which provides a supervised reconstruction loss:

$$\mathcal{L}_{\text{rec}} = |\mathcal{G}(i_j, F_j^i) - i_j|, \tag{2.12}$$

and the total loss for the generative modules is:

$$\mathcal{L}_{\text{gan}} = \mathcal{L}_{\text{rec}} + \mathcal{L}_{\text{idz}} + \lambda_{\text{adv}}\mathcal{L}_{\text{adv}}. \tag{2.13}$$

2.3.5 Feature Embedding Module

The feature embedding module, $\mathcal{F}$, has three backbones with shared layers to extract features from V, intermediate, and I images. Modified versions of the triplet and color-free losses are proposed to encourage modality-invariant descriptor features and train the V-I ReID model while considering intermediate images.

(1) Color-Free Loss: Previous studies (Pechyony & Vapnik, 2010a; Ye *et al.*, 2020c) have shown the advantages of feature invariance when extracting features from both original V and generated modality images Z. For feature representations that are robust to modality variations, we used the same color-independent loss over virtual and V images as defined in (Pechyony & Vapnik, 2010a) for training the V and intermediate backbone:

$$\mathcal{L}_{\text{cf}} = \|\mathbf{f}_j^v - \mathbf{f}_j^z\|, \tag{2.14}$$

where $\mathbf{f}_j^z$ is the feature vector of the intermediate images $z_j^\star$.

(2) Intermediate Dual Triplet Loss: Since the role of intermediate space is bridging between modalities, we use extracted features from intermediate images as the middle between them and apply two separate triplet losses in dual mode:

$$\mathcal{L}_{\text{dual}} = \mathcal{L}_{\text{tri}}(\mathbf{f}_{a\in\mathcal{V}}, \mathbf{f}_{p\in\mathcal{I}}, \mathbf{f}_{n\in\mathcal{Z}}) + \mathcal{L}_{\text{tri}}(\mathbf{f}_{a\in\mathcal{I}}, \mathbf{f}_{p\in\mathcal{Z}}, \mathbf{f}_{n\in\mathcal{V}}), \tag{2.15}$$

Algorithm 2.1 Joint AGPI^2 Training Strategy of $\mathcal{F}$, $\mathcal{G}$ and $\mathcal{D}$

Input: Multi-modal training data $\mathcal{S} = \{\mathcal{V}, \mathcal{I}\}$
Output: Model parameters $\Theta_{\mathcal{F}}$

while *all batches are not selected* **do**
  $v_j, i_j \leftarrow$ **batchSampler**(b, p);
  $i_j^{\star} \leftarrow \mathcal{G}(i_j, F_j^i)$; /* reconstruct I */
  $z_j^{\star} \leftarrow \mathcal{G}(v_j, F_j^i)$; /* generate intermediate with detached gradient style feature F_j^i */
  unfreeze $\mathcal{F}$ and $\mathcal{D}$ update $\Theta_{\mathcal{F}}$ by optimizing Eq. (2.16) ; /* detach gradient from z_j to not modify $\mathcal{G}$ */
  update $\Theta_{\mathcal{D}}$ by optimizing Eq. (2.9) while freezing $\mathcal{F}$;
  update $\Theta_{\mathcal{G}}$ by optimizing Eq. (2.13) while freezing $\mathcal{D}$;
end while

where $\mathcal{L}_{\text{tri}}$) (Ye *et al.*, 2020a,b) is common soft-margin triplet loss(Ye *et al.*, 2020a,b) with margin M_2, a, p, and n indicate the anchor, positive, and negative samples, respectively. The superscripts i, v, and z select V, I, and intermediate features, respectively.

The total loss proposed to train the feature embedding backbone is:

$$\mathcal{L}_{\text{f}} = \mathcal{L}_{\text{id}} + \mathcal{L}_{\text{dual}} + \lambda_{\text{cf}}\mathcal{L}_{\text{cf}}. \tag{2.16}$$

where $\mathcal{L}_{\text{id}}$ is the cross-entropy loss to allow identity-discriminative features (Ye *et al.*, 2020b).

Joint Learning: Each batch of data contains b person with p positive images from the V and I domains. After extracting the embedding features, the model generates intermediate images as described in Algorithm 2.1. Our joint training strategy optimizes the generator $\mathcal{G}$ by freezing the $\mathcal{F}$ and $\mathcal{D}$ parameters and also detaches the gradients from the generated images when the parameters $\mathcal{F}$ are updated.

Table 2.1 Accuracy of the proposed AGPI^2 and state-of-the-art methods on the SYSU-MM01 (single-shot setting) and RegDB datasets

Family			SYSU-MM01						RegDB					
			All Search			Indoor Search			Visible → Infrared			Infrared → Visible		
	Method	Venue	R1	R10	mAP	R1	R10	mAP	R1	R10	mAP	R1	R10	mAP
Generative	AGAN (Wang *et al.*, 2019a)	ICCV19	42.4	85.0	40.7	45.9	92.7	45.3	57.9	-	53.6	56.3	-	53.4
	D^{2}RL (Wang *et al.*, 2019d)	CVPR19	28.9	70.6	29.2	-	-	-	43.4	66.1	44.1	-	-	-
	JSIA (Wang *et al.*, 2020a)	AAAI20	45.01	85.7	29.5	43.8	86.2	52.9	48.5	-	49.3	48.1	-	48.9
	Hi-CMD (Choi *et al.*, 2020a)	CVPR20	34.94	77.58	35.94	-	-	-	86.39	-	66.04	-	-	-
	TS-GAN (Zhang, Jiang, Huang, Li & Da Xu, 2021d)	PRL21	58.3	87.8	55.1	62.1	90.8	71.3	68.2	-	69.4	-	-	-
Representation	AGW(Ye *et al.*, 2020b)	TPAMI20	47.50	-	47.65	54.17	-	62.97	70.05	-	50.19	70.49	87.21	65.90
	DDAG (Ye *et al.*, 2020a)	ECCV20	54.74	90.39	53.02	61.02	94.06	67.98	69.34	86.19	63.46	68.06	85.15	61.80
	SSFT (Lu *et al.*, 2020)	CVPR20	63.4	91.2	62.0	70.50	94.90	72.60	71.0	-	71.7	-	-	-
	SIM (Jia *et al.*, 2020)	IJCAI20	56.93	-	60.88	-	-	-	74.47	-	75.29	75.24	-	78.30
	MPANet (Wu *et al.*, 2021)	CVPR22	70.58	96.21	68.24	76.74	98.21	80.95	83.70	-	80.9	82.8	-	80.7
	DTRM (Ye *et al.*, 2021a)	TIFS22	63.03	93.82	58.63	66.35	95.58	71.76	79.09	92.25	70.09	78.02	91.75	69.56
	FTMI (Sun, Li, Chen, Peng & Zhu, 2023)	MVA23	60.5	90.5	57.3	-	-	-	79.00	91.10	73.60	78.8	91.3	73.7
	RAPV-T (Zeng, Long, Tian & Xiao, 2023)	ESA23	63.97	95.30	62.33	69.00	97.39	75.41	86.81	95.81	81.02	86.60	96.14	80.52
Intermediate	HAT(Ye *et al.*, 2020c)	TIFS20	55.29	92.14	53.89	-	-	-	71.83	87.16	67.56	70.02	86.45	66.30
	CAJ (Ye *et al.*, 2021b)	ICCV21	69.88	-	66.89	76.26	97.88	80.37	85.03	95.49	79.14	84.75	95.33	77.82
	RPIG (Alehdaghi *et al.*, 2022)	ECCVw22	71.08	96.42	67.56	82.35	98.30	82.73	87.95	98.3	82.73	86.80	96.02	81.26
	MMN (Zhang *et al.*, 2021c)	ICM21	70.60	96.20	66.90	76.20	**99.30**	79.60	**91.60**	97.70	**84.10**	87.50	96.00	80.50
	SMCL(Wei *et al.*, 2021)	ICCV21	67.39	92.84	61.78	68.84	96.55	75.56	83.93	-	79.83	83.05	-	78.57
	G2DA (Wan *et al.*, 2023)	PR23	63.94	93.34	60.73	71.06	97.31	76.01	-	-	-	-	-	-
	MCBD (Lu *et al.*, 2023)	TIFS23	71.63	94.98	67.28	79.44	98.32	79.85	90.10	97.09	82.73	87.66	96.68	81.52
AGPI^2 (ours)		-	**72.23**	**97.04**	**70.58**	**83.45**	98.62	**84.25**	89.03	**98.19**	83.89	**87.91**	**97.15**	**83.04**

Table 2.2 Accuracy of the proposed AGPI^2 and state-of-the-art methods on the Large LLCM Dataset

Family		LLCM			
		V → I		I → V	
Method	Venue	R1	mAP	R1	mAP
DDAG(Ye *et al.*, 2020a)	ECCV20	40.3	48.4	48.0	52.3
CAJ(Ye *et al.*, 2021b)	ICCV21	56.5	59.8	48.8	56.6
RPIG(Alehdaghi *et al.*, 2022)	ECCVw22	57.8	61.1	50.5	58.2
AGPI^2 (ours)	-	**61.78**	**65.08**	**56.47**	**62.79**

2.4 Results and Discussion

2.4.1 Experimental Methodology

Datasets: Cross-modal research has made extensive use of the challenging SYSU-MM01 (Wu *et al.*, 2020), RegDB (Nguyen *et al.*, 2017) and recently published LLCM (Zhang & Wang, 2023b) datasets. SYSU-MM01 is a large-scale dataset that contains more than 22K and 11K V, and I images that are not co-located belong to 491 individuals, respectively. These images were captured from four RGB and two near-infrared cameras, with 395 identities used for training and

96 for testing. The dataset has two evaluation modes, single-shot and multi-shot, based on the number of images in the gallery. The former considers only one random image per identity in the gallery, while the latter considers ten images per identity. The RegDB dataset comprises 4,120 co-located V-I images from 412 identities captured from a single camera. Ten trial configurations randomly divide the dataset into two identical sets (206 identities) for training and testing. Tests are conducted in two different ways, comparing I to V (query) and vice versa. An LLCM dataset consists of a large, low-light, cross-modality dataset that is divided into training and testing sets at a 2:1 ratio.

Implementation Details: Pre-trained ResNet50 (He, Zhang, Ren & Sun, 2016b) with two non-local attention modules is used as our feature extractor, the same as the AGW model(Ye *et al.*, 2020b). Also, concerning the analysis of different models and inputs in II.D.3, we choose VQ-VAE (Van Den Oord, Vinyals *et al.*, 2017) as the adaptor for V images into the intermediate domain. All inputs are resized to 288 by 144, then cropped randomly and filled with zero padding or mean pixels. SGD and ADAM optimizers were used for the optimization process of re-identification and generative modules, respectively. To perform the triplet loss, a minimum of two distinct individuals must be included in each mini-batch. So, we set $p = 4$ different images of $b = 8$ distinct persons from the dataset for each batch. $M_1 = 0.1$, $M_2 = 0.3$, $\lambda_{\text{adv}} = 0.1$, and $\lambda_{cf} = 10$ were set based on the analysis conducted in I.B.1.

Performance Measures: We use Cumulative Matching Characteristics (CMC), and Mean Average Precision (mAP) as evaluation metrics. The CMC is rank-k accuracy, which determines the probability of a correct cross-modality person image appearing in the top-k retrieved results. In contrast, when multiple matching images are found in a gallery, mAP measures the image retrieval performance.

2.4.2 Comparison with State-of-the-Art Methods

Table 2.1 provides a comparison of our proposed AGPI^2 with state-of-the-art V-I ReID approaches. Our experiments show AGPI^2 outperforms these methods in various situations. Although our

approach requires additional image generation during training, this process is not required during inference, making it suitable for real-time applications. Additionally, the model learns to robustly represent features across modalities. In particular, our model improves the R1 accuracy and mAP score on the large-scale SYSU-MM01 dataset by 24.73% and 22.93%, over AGW's (Ye *et al.*, 2020b) (our baseline). Compared to the second-best approach for the "indoor search" scenario, AGPI^2 outperforms by a margin of 1.15% R1 and 2.34% mAP without adding complexity to the feature backbone. For the RegDB dataset, AGPI^2 outperforms by a margin of R1 and mAP by 1.08% and 1.16%, for "visible to thermal" ReID scenarios. Similar improvements are noticeable in the "thermal to visible" scenario. In addition, our approach has the advantage that it can be easily used in different cross-modal ReID models to mine PI only in the training phase without incurring overhead during testing.

The AGPI^2 model achieves improvements on the large and complex LLCM dataset as shown in Table 2.2. Since this dataset was introduced recently, few papers reported their results, so we ran other competitors' approaches with published code on the LLCM dataset.

2.4.3 Incorporating to Other State-of-the-arts Models

The effectiveness of AGPI^2 is evident when integrated into state-of-the-art V-I ReID models (Ye *et al.*, 2020a; Wu *et al.*, 2021; Feng, Wu & Zheng, 2023). We executed the authors' provided code for each method, using their specified hyperparameters, both with and without AGPI^2. Table 2.3 highlights that AGPI^2 significantly improves performance when incorporated into these baseline models. For instance, integrating AGPI^2 into DDAG (Ye *et al.*, 2020a) and SEFL (Feng *et al.*, 2023) boosts mAP accuracy by 4% and 1%, respectively. Notably, SEFL uses supervised body-shape information from modalities and also utilizes gray images as a bridging modality alongside visible and infrared images in training, while DDAG is trained solely on visible and infrared images. Therefore, our AGPI^2 framework showcases its bridging ability more effectively in DDAG compared to SEFL.

Table 2.3 Accuracy of SOTA methods with AGPI^2 on the SYSU-MM01, testing under single-shot setting. Results were obtained by executing the author's code on PyTorch v1.12 on Ubuntu 22.04 with 40 GB NVIDIA A100 GPU

Method	**R1 (%)**	**mAP (%)**
DDAG (Ye *et al.*, 2020a)	52.74	51.92
DDAG with AGPI^2	**57.38**	**55.27**
MPANet* (Wu *et al.*, 2021)	70.58	68.24
MPANet* with AGPI^2	**72.85**	**70.66**
SEFL (Feng *et al.*, 2023)	74.05	69.21
SEFL with AGPI^2	**75.17**	**71.43**

Table 2.4 MMD between V, I, and different intermediate generation images on the SYSU-MM01 and RegDB datasets. For the I column, lower is better, and for the V, higher is better

Modality	**SYSU-MM01**		**RegDB**	
	I	V	I	V
Infrared	0	0.86	0	1.05
Visual	0.86	0	1.05	0
Grayscale (HAT (Ye *et al.*, 2020c))	0.68	0.31	0.93	0.36
Random Comb. (RPIG (Alehdaghi *et al.*, 2022))	0.64	0.34	0.92	0.39
MMN (Zhang *et al.*, 2021c)	0.61	0.33	0.87	0.43
AGPI^2 (ours)	**0.58**	**0.43**	**0.79**	**0.52**

2.4.4 Domain Shift Between Modalities

As we discussed earlier, the privileged intermediate domain bridges the main domains by sharing discriminative attributes that are common to both. Crucially, this intermediate domain must remain unbiased towards any particular domain and lie at a balanced point. To validate that our created virtual domain meets this criterion, we measured the MMD distance between the intermediate domain and both V and I domains that reflect a balanced intermediary position, ensuring an unbiased connection.

The kernel-based Maximum Mean Discrepancy (MMD) (Gretton *et al.*, 2012a) is used to measure the distance between deep features extracted by a ResNet50 model trained on ImageNet(Deng *et al.*, 2009b). Table 2.4 shows this distance for V, I, and intermediate images for the

SYSU-MM01 dataset. As described in Section 5.3, the intermediate domain must bridge the main domains to each other. Hence, it needs to have a smaller discrepancy to V and I, i.e., $\text{MMD}(V, Z) \leq \text{MMD}(I, V)$ and $\text{MMD}(I, Z) \leq \text{MMD}(I, V)$. For example, on the SYSU-MM01 dataset, the discrepancy between the intermediate images generated using our AGPI^2 method and I is 0.58, which is lower than the V-I discrepancy of 0.86. When compared to fixed gray images (Ye *et al.*, 2020c), which remain static and fail to adapt to the infrared spectrum, random gray images (Alehdaghi *et al.*, 2022) exhibit superior bridging capabilities. This advantage stems from their lack of bias towards specific color spectrums. In contrast, AGPI^2 images strike a balance between I and V modalities, leveraging adaptive information across color spectrums during generation.

This characteristic ensures that the intermediate space provides optimal information for the feature backbone to leverage common discriminative features between V and I images. Also, for visual comparison, we show examples of images created by our approach in Fig. 2.4. The third column shows intermediate images that share the same pose as the first column (V), seek to incorporate the style of the second column (I), and lose the color information.

2.4.5 Ablation Studies

A thorough ablation study is conducted to assess the impact of each module on AGPI^2 performance through various experiments on the SYSU-MM01 dataset. Initially, we compare our intermediate domain approach with other methods that use fixed and generated intermediates. Subsequently, to validate the roles of loss objectives, the generative module, and the ID-modality discriminator, we experiment with different settings and combinations with other methods.

Effect of Intermediate Image Generation Table 2.5 shows the performance of cross-modality ReID with and without using an intermediate module. With the adaptive generated images (I-V-Z training), we achieve 56.35%, while using the baseline model (I-V), we get 41.57% mAP. Additionally, the adaptive version shows comparable performance in both V and I testing settings, making it an effective alternative for a robust person ReID. Table 2.6 provides different

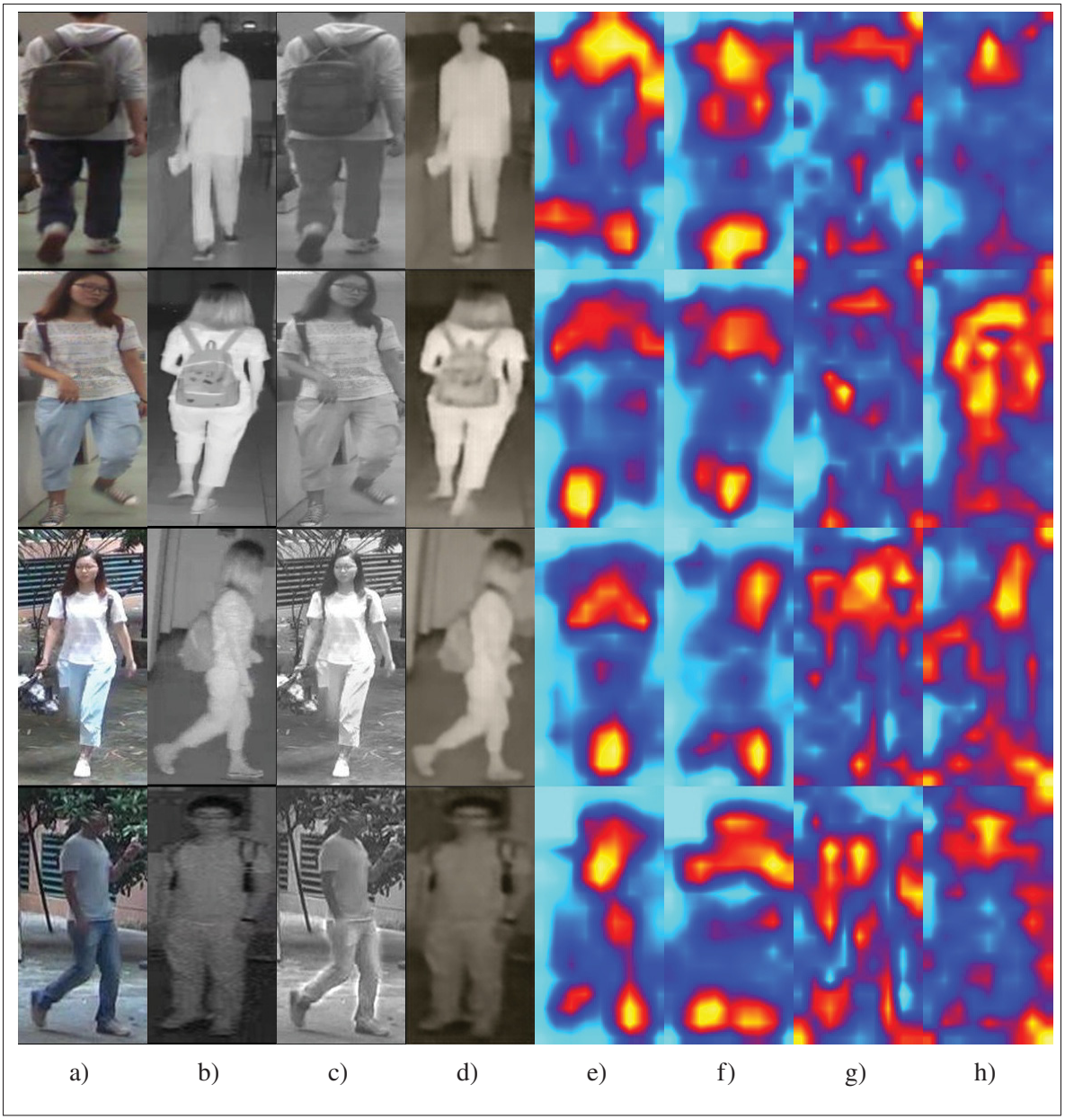


Figure 2.4 Examples of images generated with our AGPI^2 for SYSU-MM01 dataset. Columns (a) and (b) are V and I images. The intermediate and reconstructed I images are shown in (c) and (d), respectively. The GradCAM (Selvaraju *et al.*, 2017c) of our ID-Modality vs binary discriminator is shown in columns (e) to (h). Columns (e) and (f) for AGPI^2 discriminator for V and I images, respectively, and (g),(h) for Binary discriminator

options for intermediate images and it is shown that, in comparison to other approaches using the intermediate domain, AGPI^2 has better performance.

Table 2.5 The impact on the accuracy of the proposed intermediate domain in SYSU-MM01 under the multi-shot setting

Training	Testing		R1 (%)	mAP(%)
	Query	Gallery		
I-V	V	V	97.40	91.82
	I	I	95.96	80.49
	I	V	58.69	41.57
I-V-Z	V	V	97.68	90.19
	I	I	95.34	81.62
	I	V	73.17	56.35

Table 2.6 Accuracy using different intermediate images on SYSU-MM01 using the **S**ingle and **M**ulti-shot gallery. "*" indicates that our model is trained on V images while tested with I. In "$," we use I samples to train and test the model. "+" means that we only reported the MUN(Yu *et al.*, 2023) model performance of using auxiliary bringing part

Model	R1(%)		mAP(%)	
	S	M	S	M
Pre-trained (test: I→V)	2.26	3.31	3.61	1.71
Lower (train: V, test: I→V) *	7.41	9.86	8.83	6.83
Upper (train: I, test: I→I) $	90.34	94.90	87.45	80.38
Baseline (test: I→V)	49.41	58.69	41.55	38.17
Grayscale (Ye *et al.*, 2020c) (test: I→V)	61.45	65.16	59.46	46.94
RandG (Alehdaghi *et al.*, 2022) (test: I→V)	64.50	69.01	61.05	50.84
MUN+ (Yu *et al.*, 2023) (test: I→V)	71.66	-	68.35	-
AGPI^2 (test: I→V)	**72.23**	**78.19**	**70.58**	**64.13**

Effect of Losses In the following, we discuss the effect of each loss in AGPI^2.

I. Effect of Intermediate Dual Triplet Loss: Table 2.7 shows that when used in conjunction with the dual triplet loss, this loss provides supervision to improve performance at test time. Although the ordinary triplet loss improves performance when it is integrated with $\mathcal{L}_{cf}$, it ignores the privileged information provided by the intermediate domain. To show that fact, we compared the effect of our $\mathcal{L}_{dual}$ to $\mathcal{L}_{tri}$ by using each of them alongside other losses in the training process. As can be seen, using such regularization increases mAP and rank-1 by 3% and 7%, respectively, over ordinary triplet loss.

II. Effect of Color-Free Loss: Table 2.7 illustrates the impact of $\mathcal{L}_{\text{cf}}$. The results indicate that incorporating this regularizing loss leads to an increase of 3% in mAP and 5% in rank-1 accuracy when it is used without using $\mathcal{L}_{\text{tri}}$ or $\mathcal{L}_{\text{dual}}$. When the model is trained with $\mathcal{L}_{\text{cf}}$ and $\mathcal{L}_{\text{tri}}$ the performance significantly improves (19% in mAP and 12% in rank-1) and alongside with $\mathcal{L}_{\text{dual}}$ mAP and rank-1 improved over 6% and 7% respectively. As the intermediate images are derived from visible images with color information removed and infrared style information incorporated, $\mathcal{L}_{\text{cf}}$ plays a crucial role. This ensures that the extracted features from the visible modality are color-independent, aligning them more closely with the attributes of infrared.

III.Effect of Adversarial Loss: As a result of the adversarial loss, the generator must adapt the reconstruction process, enabling the stylistic transformation of images from V to I. Without using this objective, the generated images are similar to the gray-scale channel images and do help the model adjust for differences between I and V images. As you can see in the 4th row of Table. 2.7, the performance is close to the model trained based on grey intermediate images in (Ye *et al.*, 2020c).

IV.Effect of Reconstruction Loss: To improve the quality of the reconstructed I images, supervised reconstruction loss is applied to the encoder and decoder parts. In the absence of this loss, the generated images lack sufficient important information to support the feature module and make the performance drop dramatically. As the color-free loss pushes the V features to be the same as the features of such not meaningful intermediate images, the mAP accuracy of matching between V and I decreases (from 41.55% to 5.51%) as seen in the 5th row of Table 2.7.

Generation Model

To find the highest ReID performance with respect to the quality of the generated intermediate images, we tested different generative models by adapting V to I modality or vice versa. Since V images have more information to discriminate between individuals in comparison to I (Jia *et al.*, 2020), the transferred V images play a better role in providing privileged information in the training process. In Table 2.8, the model in (Van Den Oord *et al.*, 2017) is deeper (more layers) and has a better capacity to stylize the V images from the I feature vectors. Also, we tested a

Table 2.7 Impact on the accuracy of using different losses in our AGPI^2 on SYSU-MM01 in single-shot

Losses						R1 (%)	mAP (%)
$\mathcal{L}_{id}$	$\mathcal{L}_{tri}$	$\mathcal{L}_{dual}$	$\mathcal{L}_{cf}$	$\mathcal{L}_{adv}$	$\mathcal{L}_{rec}$		
✓	✓					49.41	41.55
✓		✓		✓	✓	60.64	56.15
✓				✓	✓	40.16	35.89
✓			✓	✓	✓	45.56	41.33
✓		✓	✓		✓	60.67	58.35
✓		✓	✓	✓		9.32	5.51
✓	✓		✓	✓	✓	62.76	60.02
✓	✓	✓	✓	✓	✓	64.80	61.92
✓		✓	✓	✓	✓	67.29	63.86

large state-of-the-art model (Choi, Uh, Yoo & Ha, 2020b), which is used for multiple domain adaptation as our generator.

Table 2.8 Accuracy with different options for the generation of intermediate images

Model	V → I		I → V		# params
	R1	mAP	R1	mAP	
VQ-VAE2	65.75	61.96	61.81	58.34	20M
deep VQ-VAE2	**67.29**	**63.86**	**63.45**	**60.09**	28M
StarGAN	66.21	62.70	62.41	59.12	50M

ID-modality Discriminator vs General Discriminator The ID-modality discriminator aims to discriminate modalities based on ID-aware features for each modality to help the generative module transform such features from V and I to intermediate. To show the effectiveness of such a module on ReID performance, we compared its results with those of a general binary discriminator that uses whole features. Table 2.9 demonstrates a 4% improvement in rank-1 accuracy and a 5% enhancement in mAP measurement by introducing our ID-modality discriminator. The discriminator effectively identifies modality-specific, ID-related information, enabling the generative model to transform visible images into infrared while preserving the individuals' discriminative attributes. Consequently, the transformed images provide intermediate privileged information to the feature backbone, leading to a more robust reduction of modality discrepancies in the feature space. Additionally, Fig. 2.4 highlights the key regions of the input images that

the discriminator focuses on when making its decisions, demonstrating the relevance of the identified features.

Table 2.9 Accuracy with different options for the discriminator and generation of intermediate images

Model	V → I		I → V	
	R1	mAP	R1	mAP
Binary Discriminator	63.23	58.84	60.18	57.32
Our Discriminator	**67.29**	**63.86**	**63.45**	**60.09**

2.4.6 Inference Model Efficiency

Our AGPI^2 proposed method creates a privileged intermediate domain only in training to improve the ReID performance without adding extra computation overhead in the inference time. In Table II-2, we compared our model's size and time complexity in the inference time with the baseline(Ye *et al.*, 2020b), MPANet(Wu *et al.*, 2021), MMN(Zhang *et al.*, 2021c), DTRM (Ye *et al.*, 2021a), RPIG (Alehdaghi *et al.*, 2022) and thGAN (Kniaz *et al.*, 2018).

Table 2.10 Number parameters and floating-point operations at inference time for AGPI^2 and state-of-the-art models

Model	# of Para. (M)	Flops (G)
baseline (Ye *et al.*, 2020b)	**24.8**	**5.2**
AGPI^2	**24.8**	**5.2**
MPANet (Wu *et al.*, 2021)	30.1	6.7
MMN (Zhang *et al.*, 2021c)	26.6	5.8
DTRM (Ye *et al.*, 2021a)	26.6	5.4
RPIG (Alehdaghi *et al.*, 2022)	24.9	**5.2**
thGAN (Kniaz *et al.*, 2018)	35	8.9

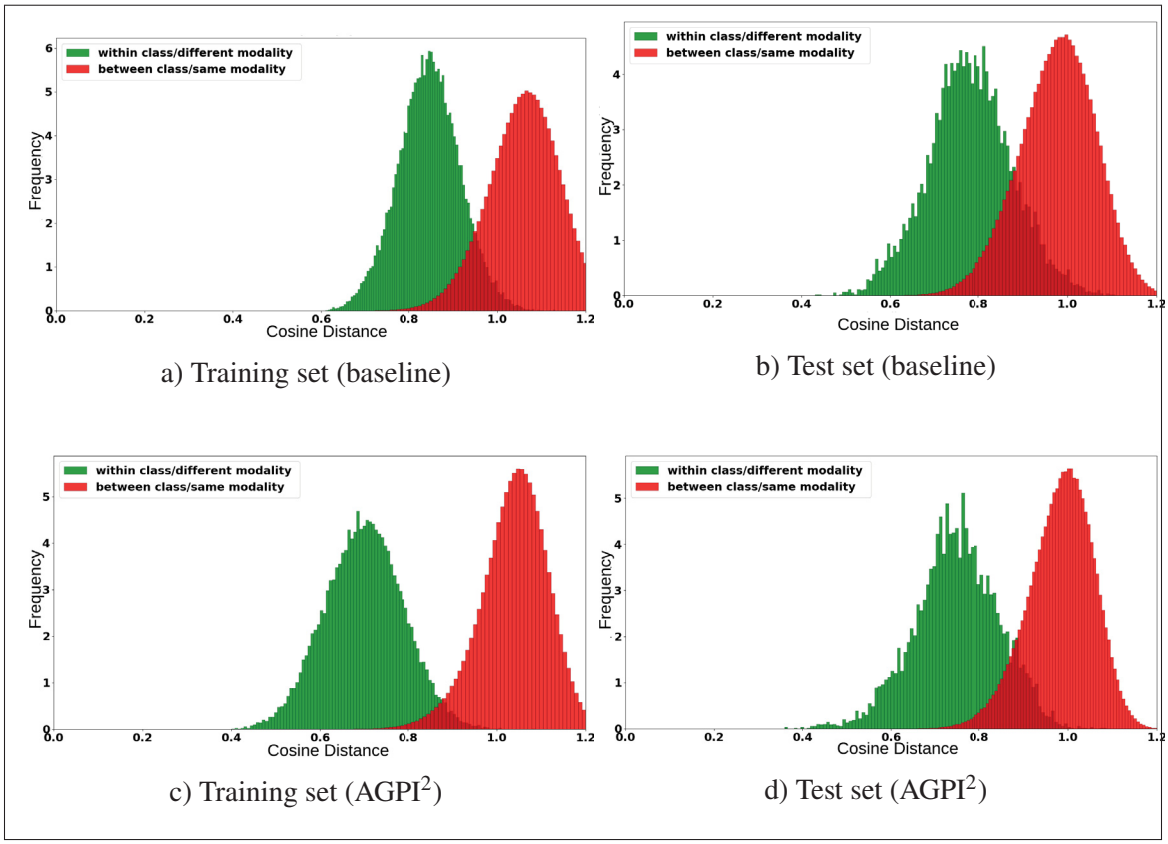
2.4.7 Data Augmentation

Random erasing augmentation (Zhong, Zheng, Kang, Li & Yang, 2020) was assessed with different settings for the feature embedding and the generation modules. Table 2.11 shows the accuracy of our proposed approach using different scenarios (Zhong *et al.*, 2020). Despite

each augmentation configuration improving performance, the best results are achieved when the augmentation probability is gradually increased during training.

Table 2.11 Matching accuracy with random erasing augmentation. s is the size of the rectangle and p is the probability of erasing. Note: "$a \rightarrow b$" means that training started from a and finished with b

Configurations	p	s	**R1 (%)**	**mAP(%)**
AGPI^2	0	-	67.29	63.86
AGPI^2 + Aug1	0.40	0.20	69.81	65.19
AGPI^2 + Aug2	0.50	0.25	70.93	66.15
AGPI^2 + Aug3	0.50	0.30	69.39	64.95
AGPI^2 + Aug4	0.30 $\rightarrow$ 0.50	0.20 $\rightarrow$ 0.30	**72.23**	**70.58**



a) Training set (baseline)

b) Test set (baseline)

c) Training set (AGPI^2)

d) Test set (AGPI^2)

Figure 2.5 Distribution of between- and within-class distances on the SYSU-MM01 dataset for (a,c) train and (b,d) test sets with the baseline model (Ye *et al.*, 2020b) and our AGPI^2 model

2.4.8 Qualitative Evaluation

A visual representation of the cosine distance distributions for positive and negative cross-modality matching pairs is provided in Fig. 2.5 for both training and testing sets. The proposed AGPI^2 approach is compared to a strong baseline (AGW model in (Ye *et al.*, 2020b)). After training, such relationships should be reversed so that the two strategies work effectively to decouple I from V. Despite this, AGPI^2 is more efficient than the baseline on the test set when combined with the adaptively generated images. By leveraging privileged intermediate information during training, AGPI^2 improves the model's ability to discriminate between V and I images at test time.

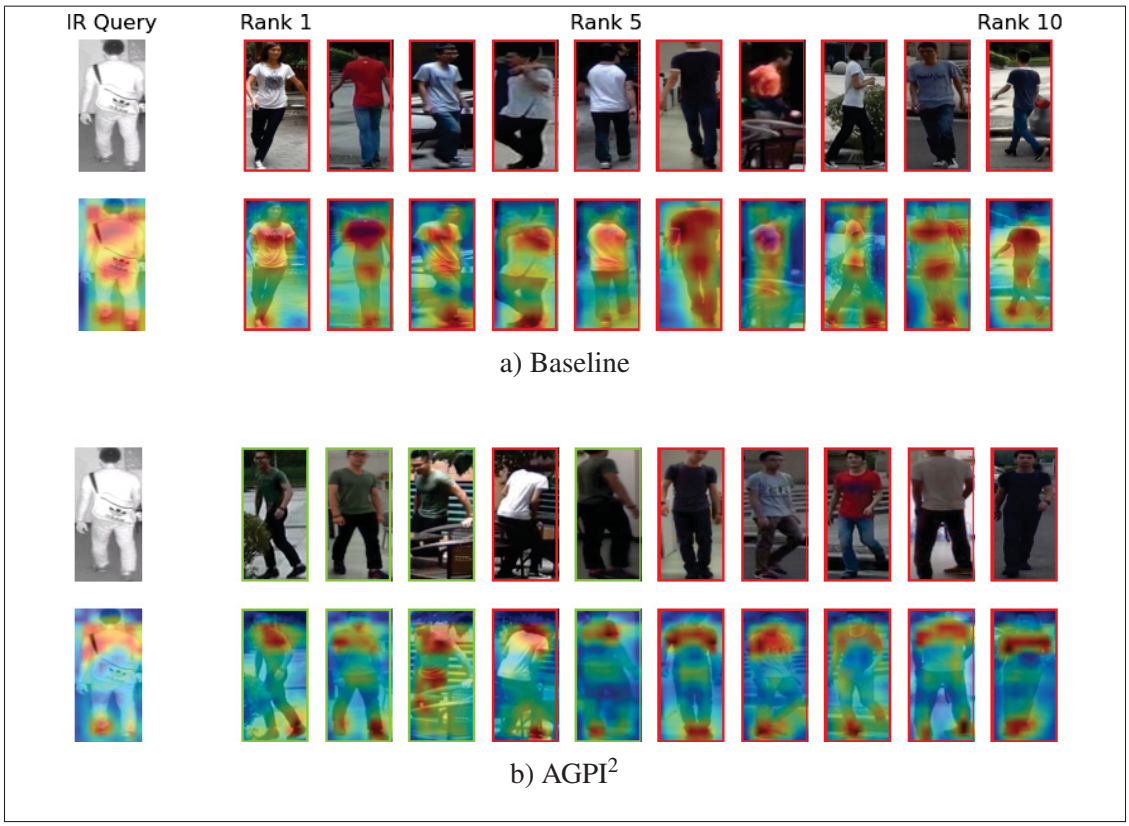


Figure 2.6 Rank-10 visualizations of the (a) baseline and (b) AGPI^2 strategies on the SYSU-MM01 dataset, along with their corresponding similarity CAM (Stylianou *et al.*, 2019). Green or red boxes respectively surround correct and incorrect matches. There are only four visible cameras, and only one image from one of them is in the gallery

Fig. 2.6 shows the rank-10 for a given class of the SYSU-MM01 dataset given by the AGW (Ye *et al.*, 2020b) (the baseline) and our proposed strategy, along with the corresponding similarity CAMs (Stylianou *et al.*, 2019). The query CAM is produced from the similarity of the best match. Compared to the baseline method, most discriminative regions and attention between them are refined and more focused. CAMs seem to focus on shoulders, hips, and feet for the baseline method, while the hip's attention almost fully disappears for our model. Interestingly, this behavior is consistent from one sample to another. Perhaps the hip focus at the baseline comes from the important color changes between the lower and upper bodies, which are probably discriminant for the visible part of the model. However, regarding the IR modality, this color marker is much less reliable and probably appears as a source of confusion for cross-modal ReID. This phenomenon tends to diminish thanks to the adaptive and informative intermediate privileged domain that is created through the training.

To demonstrate the effectiveness of our proposed ID-Modality discriminator, inspired by (Wu *et al.*, 2023b, 2022), we compared its class activation map to that of a general binary discriminator, as shown in Fig. 2.4. The results indicate that the ID-Modality discriminator is more focused on the foreground, whereas the binary discriminator primarily detects the modality based on the background.

2.5 Conclusion

In this paper, a new framework is introduced for adapting privileged intermediate information called AGPI^2. This method generates a virtual intermediate domain that bridges the gap between V and I images during training, thereby reducing the modality discrepancy problem. To create privileged intermediate information, an adversarial approach is also introduced for the generator, feature embedding, and ID-modality discriminator modules in the training process. AG outperforms state-of-the-art methods for V-I person ReID on the SYSU-MM01 and RegDB datasets while incurring no computational overhead during inference.

The AGPI^2 training approach can be integrated into other SOTA methods for ReID to improve performance and generality. Compared to other intermediate generation methods, we show AGPI^2 approach can provide more balance between V and I information for the model to diminish the modality-specific attributes in the representation feature space. In future research endeavours, we aim to extend the application of the AGPI^2 approach to multi-step and gradual intermediate privileged learning, particularly to address more substantial domain shift challenges. This will necessitate the AGPI^2 model's ability to quantify the incorporation of I domain information during the transformation of V images into the intermediate space, facilitating a gradual adaptation process.

CHAPTER 3

BIDIRECTIONAL MULTI-STEP DOMAIN GENERALIZATION FOR VISIBLE-INFRARED PERSON RE-IDENTIFICATION

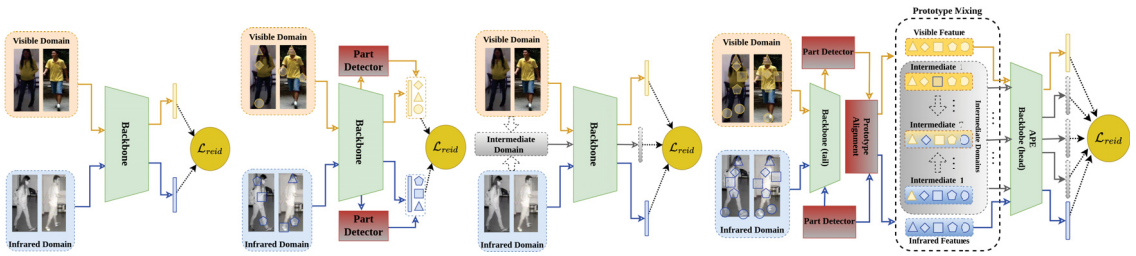
Mahdi Alehdaghi[1] , Pourya Shamsolmoali[2] , Rafael M. O. Cruz[1] , Eric Granger[1]

[1] LIVIA, ILLS, Dept. of Systems Engineering, École de Technologie Supérieure,
1100 Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3
[2] University of York, U.K.

The article was published in IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), February 2025

3.1 Introduction

V-I ReID is a variant of person ReID that involves matching individuals across RGB and IR cameras. V-I ReID is challenging since it requires matching individuals with significant differences in appearance between V and I modality images. In this context, V-I ReID systems must train a discriminant feature extraction backbone to encode consistent and identifiable attributes in RGB and IR images. Most state-of-the-art methods seek to learn global representations from the whole image by alignment at the image-level (Wang *et al.*, 2019a; Kniaz *et al.*, 2018) or feature-level (Ye *et al.*, 2020b; Ye *et al.*, 2020; Zhu *et al.*, 2020b; Park *et al.*, 2021) (see Fig. 3.1(a)). Other methods extract global modality-invariant features by disentangling them from modality-specific information (Choi *et al.*, 2020a; Yang *et al.*, 2020). However, global feature-based approaches cannot compare local discriminative attributes, resulting in the loss of important discriminant cues about the person. To address this issue and focus on the unique information in different body regions, part-based approaches (see Fig. 3.1(b)) extract local fine-grained feature representations through horizontal stripes, clustering, or attention mechanisms of spatially extracted features (Ye *et al.*, 2020a; Lu *et al.*, 2020; Huang *et al.*, 2022; Wu *et al.*, 2021). However, given the domain discrepancies between the V and I modalities and the lack of matching cross-modal parts during training, these methods often learn modality-specific attributes for each part. This typically results in feature representations that are less modality-invariant, and thus provide limited effectiveness for cross-modal matching tasks.



a) Global representation. b) Part-based representation. c) Intermediate modality. d) Our bidirectional multi-step domain generalization.

Figure 3.1 A comparison of training architectures for V-I ReID. Approaches based on (a) a global features representation, and (b) a local part-based representation to preserve locality and global features. (c) Generation of an intermediate bridging domain to guide training. (d) Our BMGD extracts and combines prototypes from modalities in each step to gradually create multiple intermediate bridging domains

Some methods leverage an intermediate modality to guide the training process such that the cross-modal domain gap is reduced (Li *et al.*, 2020a; Ye *et al.*, 2020c; Alehdaghi *et al.*, 2022; Wu *et al.*, 2017b; Zhang *et al.*, 2021c) (see Fig. 3.1(c)). Intermediate modalities may be static images (e.g., gray-scale) (Ye *et al.*, 2020c; Alehdaghi *et al.*, 2022) or created images from a generative module learned by I and V modalities (Li *et al.*, 2020a; Wu *et al.*, 2017b; Zhang *et al.*, 2021c). The resulting intermediate representations can leverage shared relations between I and V images and improve robustness to modality changes. However, they generate images that tend to contain shared non-discriminative information, resulting in the loss of ID-aware information in the intermediate domain. Moreover, one-step training strategies only create one intermediate space between V and I. This cannot provide sufficient bridging information when modalities have significant discrepancies. Single-step approaches cannot effectively deal with these discrepancies during training. Therefore, we advocate for linking a series of intermediate domains with progressively smaller gaps.

To effectively bridge the large cross-modal domain gap while preserving local ID-discriminative information, we introduce Bidirectional Multi-step Domain Generalization (BMDG), a gradual, multi-step, part-based training strategy for person V-I ReID (see Fig. 3.1(d)). At first, this strategy

identifies similar ID-aware part prototypes[1] for each modality, and then it progressively creates intermediate steps by controlling the number of the part prototypes to be mixed. Generated images can become degraded, as discriminative details are lost over consecutive generation steps, making these images less useful for training. Therefore, our proposed BMDG creates multiple intermediate domains in the feature space by aligning, and then combining ID-informative subsets of features from both modalities.

Our BMGD training strategy is comprised of two parts. The prototype alignment module learns to identify and align part-prototype representations that contain discriminative information linked to body parts in I and V images. Then, the bidirectional multi-step learning progressively extracts auxiliary intermediate feature representations by mixing prototypes from both modalities at each step. For ID-aware and robust alignment between modalities, the extracted part-prototypes must meet three conditions: (1) they should be *complementary* to other parts, (2) they should be *interchangeable* and represent consistent parts, and (3) they should be *discriminant* for person matching. Hierarchical contrastive learning is proposed to train the prototype alignment module such that these properties are respected without affecting ReID accuracy. Our novel idea is to progressively create intermediate bridging domains by mixing a growing number of body part prototypes extracted from I and V modalities. This allows leveraging common semantic information between modalities, along with discriminative information among individuals.

Bidirectional multi-step learning relies on auxiliary intermediate feature spaces that are created at each step, by progressively mixing semantically consistent prototypes from each modality. While prototype-based approaches (Ye *et al.*, 2020a; Liu *et al.*, 2021; Fang *et al.*, 2023; Zhu, Guo, Liu, Tang & Wang, 2020a) concentrate solely on enhancing discriminatory representation using local descriptors, BMDG aligns and mixes local part information in a gradual training process to mitigate domain discrepancies within the representation space. Unlike the Part-Mix method (Kim *et al.*, 2023), which utilizes a fixed proportion of mixed parts for auxiliary features, our approach employs bidirectional multi-step generation for multiple intermediate steps. This

[1] We define a *part prototype* as a feature representation corresponding to a specific body part in a cropped image of a person.

bolsters the robustness of V-I ReID training for samples with large discrepancies, especially at the beginning of training when the parts are inconsistent. Our strategy enables effective bridging of cross-modality gaps and allows gradual training by adjusting the proportion of information to be exchanged. We also propose a hierarchical prototype learning strategy, which is crucial to defining meaningful bridging steps.

Our main contributions are summarized as follows.

(1) A BMDG method is proposed to train a backbone V-I ReID model that learns a common feature embedding by gradually reducing the gap between I and V modalities through multiple auxiliary intermediate steps.

(2) A novel hierarchical prototype learning module is introduced to align discriminative and interchangeable part prototypes between modalities using two-level contrastive learning. These prototypes maintain semantic consistency to facilitate the exchange of V and I information, thus enabling the generation of intermediate bridging domains.

(3) A bidirectional multi-step learning model is proposed to progressively combine prototypes from both modalities and create intermediate feature spaces at each step. It integrates attentive prototype embedding (APE), which refines these prototypes to extract attributes that are shared between modalities, thereby improving ReID accuracy. Indeed, BMDG can find better modality-shared attributes by optimizing the model on two levels: prototype alignment and bidirectional multi-step embedding modules.

(4) Extensive experiments on the challenging SYSU-MM01 (Wu *et al.*, 2020), RegDB (Nguyen *et al.*, 2017) and LLCM(Zhang & Wang, 2023b) datasets indicate that BMDG outperforms state-of-the-art methods for V-I person ReID. They also show that our framework can be integrated into any part-based V-I ReID method and can improve the performance of cross-modal retrieval applications.

3.2 Related Work

In V-I ReID, deep backbone models are trained with labeled V and I images captured using RGB and IR cameras. During inference, cropped images of people are matched across camera modalities. That is, query I images are matched against V gallery images, or vice versa (Chen *et al.*, 2021; Dai *et al.*, 2018; Fu *et al.*, 2021a; Hao *et al.*, 2020; Park *et al.*, 2021; Xu *et al.*, 2021; Ye *et al.*, 2018a; Zhu *et al.*, 2020b). State-of-the-art methods can be classified into approaches for global or part-based representation, or that leverage an intermediate modality.

(a) Global and Part-Based Representation Methods: Representation approaches focus on training a backbone model to learn a discriminant representation that captures the essential features of each modality while exploiting their shared information (Ye *et al.*, 2018a; Dai *et al.*, 2018; Ye *et al.*, 2020; Wu *et al.*, 2017a; Liu & Cheng, 2019; Ye *et al.*, 2020b). In (Ye *et al.*, 2020b,c, 2021b; Park *et al.*, 2021; Chen *et al.*, 2020a), authors only use global features through multimodal fusion. The absence of local information limits their accuracy in more challenging cases. To improve robustness, recent methods combine global information with local part-base features (Ye *et al.*, 2020a; Lu *et al.*, 2020; Wu *et al.*, 2021; Fang *et al.*, 2023; Kim *et al.*, 2023). For example, (Ye *et al.*, 2020a; Lu *et al.*, 2020) divides the spatial feature map into fixed horizontal sections and applies a weighted-part aggregation. (Fang *et al.*, 2023; Wu *et al.*, 2021, 2023a; Zhu *et al.*, 2020a) dynamically detects regions in the spatial feature map to alleviate the misalignment of fixed horizontally divided body parts.

However, these part detectors might over-focus on specific parts, varying based on the modality or person. (Kim *et al.*, 2023) addresses this issue using Part-Mix data augmentation to regularize model part discovery. Additionally, parts may not correspond to semantically similar regions across different images. To tackle this, a bipartite graph-based correlation model between part regions has been proposed to maximize similarities (Wu *et al.*, 2023a). While part-based methods enhance representation ability, they often overlook shared information between similar semantic parts across different individuals or modalities, leading to inconsistencies in extracted part features. This hinders effective comparison or exchange. Similar to the human perception

system, distinguishing individuals by comparing attributes of similar body parts (e.g., head, torso, leg) learned from a diverse set of individuals, our approach seeks to capture and leverage this shared information for accurate matching. To achieve this, BMDG discovers different body parts as prototypes, disregarding the person's identity using hierarchical contrastive learning, and then leverages their ID-discriminative attributes for each part through part-based cross-entropy loss.

(b) Methods Based on an Intermediate Modality: In recent years, V-I ReID methods have relied on an intermediate modality to address the significant gap between V and I modalities (Fan *et al.*, 2020; Ye *et al.*, 2020c; Alehdaghi *et al.*, 2022; Li *et al.*, 2020a; Wu *et al.*, 2017b; Zhang *et al.*, 2021c; Feng *et al.*, 2023; Alehdaghi, Josi, Shamsolmoali, Cruz & Granger, 2023). Using fixed intermediate modality like grayscale (Fan *et al.*, 2020; Ye *et al.*, 2020c) or random channels (Ye *et al.*, 2021b; Alehdaghi *et al.*, 2022) can improve accuracy. (Li *et al.*, 2020a) transform V images into a new modality to reduce the gap with I images. (Wu *et al.*, 2017b) conducted a pixel-to-pixel feature fusion operation on V and I images to build the synthetic images, and (Zhang *et al.*, 2021c) used a shallow auto-encoder to generate intermediate images from both modalities. While these generated images bridge the cross-modal gap, details of the image are lost. Accuracy is degraded in the context of multi-step learning. To address this, we extract body part prototypes and gradually use them to define multiple intermediate feature spaces.

(c) Domain Generalization and Adaptation: Data augmentation is an effective strategy to learn domain-invariant features and improve model generalization (Zhou *et al.*, 2022). DG methods focus on manipulating the inputs to assist in learning general representations. (Zhou, Yang, Hospedales & Xiang, 2020) relies on domain information to create additive noise that increases the diversity of training data while preserving its semantic information. (Zhang *et al.*, 2018; Zhou, Yang, Qiao & Xiang, 2021) uses Mixup to increase data diversity by blending examples from the different training distributions.

Some domain adaptation (DA) methods allow for bridging significant shifts between source and target data distributions. Gradual and multi-step (or transitive) methods (Wang & Deng, 2018) define the number and location of intermediate domains. Intermediate Domain Labeler (Chen & Chao, 2021) also creates multiple intermediate domains, where a coarse domain discriminator sorts them such that the cycle consistency of self-training is minimized. IDM(Dai *et al.*, 2021) mixes entire latent features to create an intermediate domain that lies on the shortest path between the target and source. In fact, by utilizing appropriate intermediate domains, the source knowledge can be better transferred to the target domain. By mixing the entire hidden features, this domain may lose the local ID-discriminative information and reduce the diversity of part features. This paper reduces the cross-modality gap by exchanging features between domains and creating intermediate virtual domains through gradual mixing during training.

3.3 Proposed Method

Our BMDG approach relies on multiple virtual intermediate domains created from V and I images to learn a common representation space for V-I person ReID. Fig. 3.2(a) shows the overall BMDG training architecture comprising two modules. (1) Our **part prototype alignment learning** module extracts semantically aligned and discriminative part prototypes from I and V modalities through hierarchical contrastive learning. Each prototype represents a specific part feature in a cropped person image. Exchanging aligned part prototypes allows for the gradual creation of ID-informative intermediate spaces. This module discovers and aligns part features that are distinct and informative about the person to transform the representation feature of one person between modalities robustly. (2) Our **bidirectional multi-step learning** module creates the auxiliary intermediate feature spaces at each step by progressively mixing more learned part prototypes from each modality. This gradually reduces the modality-specific information in the final feature representation. In other words, two intermediate feature spaces are generated, at each step, by combining proportions of aligned sub-features from each modality. The model gradually learns to reduce modality-specific information and to leverage more common information from such intermediate domains by increasing the portions and making more complex training cases.

BMDG trains the feature backbone by learning from easy samples with a lower modality gap to more complex ones with a higher gap.

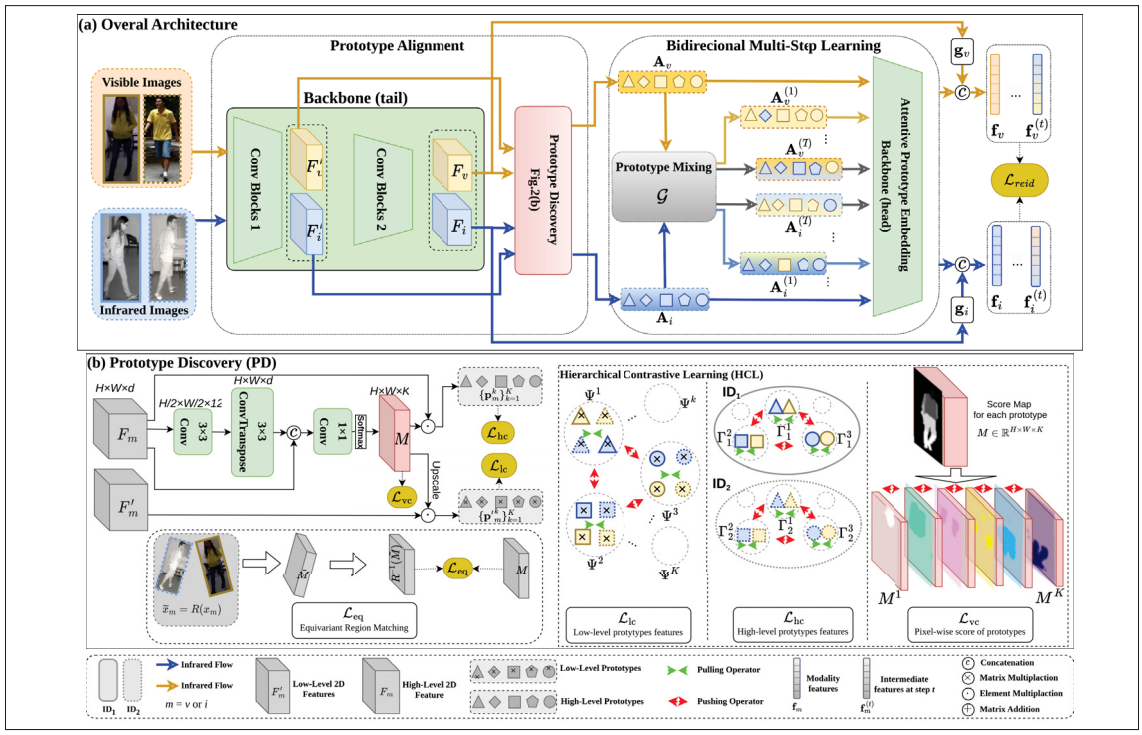


Figure 3.2 (a) Overall BMDG training architecture comprises two parts. The prototype alignment module (left) extracts body part prototype representations from V and I images. The bidirectional multi-step learning module (right) extracts discriminant features using multiple intermediate domains created by mixing prototype information. (b) Prototype discovery (PD) architecture mines prototypes from spatial features, and hierarchical contrastive learning (HCL) encourages the prototypes to focus on similar semantics for all individuals without losing ID-discriminative information

Preliminaries: For the training process, let us assume a multimodal dataset for V-I ReID, represented by $\mathcal{V} = \{x_v^j, y_v^j\}_{j=1}^{N_v}$ and $\mathcal{I} = \{x_i^j, y_i^j\}_{j=1}^{N_i}$ which contains sets of V and I images from C_y distinct individuals, with their ID labels. V-I ReID systems seek to match images captured from V and I cameras by training a deep backbone model to encode modality-invariant person embedding, denoted by $\mathbf{f}_v$ and $\mathbf{f}_i$. Given V and I images, the objective is to:

$$\max\{S(\mathbf{f}_v^n, \mathbf{f}_i^l) \cdot o\} \text{ where } o = 1 \text{ if } y_v^n = y_i^l \text{ else } o = \text{-}1, \tag{3.1}$$

where $\mathbf{f}_v^n, \mathbf{f}_i^l \in \mathbb{R}^d$, d is the dimension of the representation space, and $S(.,.)$ is similarity matching function.

3.3.1 Prototype Learning Module

To align persons' sub-features between different images in I and V, this module is used to discover different body-part attributes from diverse regions and learn to align these prototypes by making them complementary, interchangeable, and discriminative, as discussed in Section 3.1.

$$\mathcal{H}: \mathbb{R}^{W\times H\times 3} \rightarrow \mathbb{R}^{K\times d} \tag{3.2}$$

$$x \sim X \mapsto \mathcal{H}(x) \sim P \tag{3.3}$$

(a) Prototype Discovery (PD). This module is proposed to extract the prototypes and discover several discriminative regions on the backbone feature maps. At first, the PD predicts a pixel-wise masking score (probability) $M \in \mathbb{R}^{H\cdot W\times K}$ for each prototype from the features map $\mathbf{F}_m \in \mathbb{R}^{H\cdot W\times d}$ where H, W, d are the feature sizes, and K is the total number of prototypes. m is modality index which indicates v as visible or i as infrared domain. Then, it computes each prototype vector $\mathbf{p}_m^k$ by weighted aggregating of features in all pixels as:

$$\mathbf{p}_m^k = \frac{M_m^k}{|M_m^k|}\sum_{u\in U}[\mathbf{F}_m]_u\ , \quad |M_m^k| = \sum_{u\in U}[M_m^k]_u, \tag{3.4}$$

where for each pixel $u \in U = \{1, \ldots, H\cdot W\}$, we have $\sum_{k=1}^{K}[M_m^k]_u = 1$. This module (illustrated in Fig. 3.2(b) uses a shallow U-Net(Ronneberger, Fischer & Brox, 2022) architecture to create a region mask for each prototype. In the down-sample block, a convolution receptive field is applied on masks to make them lose locality by leveraging their neighbor pixels. At the same time, the up-sample propagates these high-level, coarsely localized masks into each original size. Skip connections are used to re-inject the local details lost in the down-sample phase.

(b) Hierarchical Contrastive Learning: To ensure the part prototypes represent the diverse attributes of each person, they must be non-redundant, and be independent of the others. This

independence can be formulated that the mutual information (MI) between the distribution of random variables describing parts as P^k should be zero as:

$$\mathrm{MI}(P^k; P^q) = 0 \ \forall k, q \in \{1, \ldots, K\}, \ k \neq q, \tag{3.5}$$

where P^k is the distribution of the random variable of part k. Since the estimation of mutual information is complex and time-consuming, $MI(P^k, P^q)$ is minimized by directly reducing the cosine similarity between the extracted features of different parts as:

$$\min S\{(\mathbf{p}^k, \mathbf{p}^q)\} \ \forall k, q \in \{1, \ldots, K\}, \ k \neq q, \tag{3.6}$$

where $\mathbf{p}^k$ and $\mathbf{p}^q$ are features for part k and q, respectively. To ensure that the prototype features are semantically consistent around body parts that can be aligned in different samples, we maximize the similarity of $\mathbf{p}^k$ and all other prototypes with the same index in the training batch as $\hat{\mathbf{p}}^k$.

To select samples in each batch with a positive prototype index, there are two options: the same parts belonging to the same or different persons (inter-person) as Ψ^k, and the same parts belonging to the same person with identity y (intra-person) as Γ_y^k (see Fig. 3.2.b). Since the goal is to extract dissimilar features in the representation space for two different persons, pulling apart the same prototypes from these two persons may disrupt this objective. Therefore, two contrastive losses as $\mathcal{L}_{\mathrm{lc}}$ and $\mathcal{L}_{\mathrm{hc}}$ are employed to provide inter-person information at a low-level and intra-person at a high-level:

$$\mathcal{L}_{\mathrm{lc}} = -\sum_{k=1}^{K} \sum_{\hat{\mathbf{p}}' \in \Psi^k} \log \frac{e^{\mathbf{p}'^k \cdot \hat{\mathbf{p}}'/\tau}}{e^{\mathbf{p}'^k \cdot \hat{\mathbf{p}}'/\tau} + \sum_{q \neq k} e^{\mathbf{p}'^k \cdot \mathbf{p}'^q/\tau}}, \tag{3.7}$$

$$\mathcal{L}_{\mathrm{hc}} = -\sum_{y=1}^{C_y} \sum_{k=1}^{K} \sum_{\hat{\mathbf{p}} \in \Gamma_y^k} \log \frac{e^{\mathbf{p}^k \cdot \hat{\mathbf{p}}/\tau}}{e^{\mathbf{p}^k \cdot \hat{\mathbf{p}}/\tau} + \sum_{q \neq k} e^{\mathbf{p}^k \cdot \mathbf{p}^q/\tau}}, \tag{3.8}$$

where $\mathbf{p}'^k_m$ are low-level part prototype features vector computed same as Eq. (3.4) but extracted from lower-level layers of the feature backbone. Unlike contrastive learning introduced in (Li *et al.*, 2021b; Choudhury, Laina, Rupprecht & Vedaldi, 2021) that are used only for final features,

the HCL optimizes at two levels as low-level and high-level features to force the model to cluster inputs based on parts and then cluster inside identities, as illustrated in Fig. 3.2(b). For example, $\mathcal{L}_{\text{lc}}$ encourages the low-level prototype of legs for two persons to be similar and differ from other prototypes such as the torso. While $\mathcal{L}_{\text{hc}}$ encourages the high-level legs prototypes only differ from other prototypes in that person. Also, minimizing the distance between features of pixels belonging to the same part occurrence encourages the model to extract more similar features for each prototype:

$$\mathcal{L}_{\text{c}} = \sum_{k=1}^{K} \sum_{u \in U} [M_m^k \| \mathbf{p}_m^k - \mathbf{F}_m \|]_u. \tag{3.9}$$

Describing part occurrence to a single feature vector enables us to exchange these features for the corresponding prototype in different modalities and creates the intermediate step. Apart from making part features independent of each other by feature contrastive losses, for pushing prototypes to focus on different semantic body parts in different image locations, the visual contrastive is used as:

$$\mathcal{L}_{\text{vc}} = \sum_{k=1}^{K} \sum_{q=k+1}^{K} \sum_{u \in U} [\| M_m^k - M_m^q \|]_u, \tag{3.10}$$

to extract parts from diverse regions of backbone features with the minimum intersection.

Also, for equivariant matching of regions, a random rigid transformation (R) is applied on the input images (x_m) to create the transformed images ($\tilde{x}_m$). Both are then processed by the model, and the transformation on the masking score ($\tilde{M}_m$) is inverted from the transformed images. The objective is to minimize the distance between the transformed and original maps:

$$\mathcal{L}_{\text{eq}} = \sum_{k=1}^{K} \| M_m^k - R^{-1}(\tilde{M}_m^k) \|. \tag{3.11}$$

(c) Part Discrimination. The information encoded by prototypes must describe the object in the foreground. That is, the mutual information between the joint probability $\{P^k\}_{k=1}^K$ and identity Y should be maximized:

$$\text{MI}(P^1, \ldots, P^K; Y). \tag{3.12}$$

In the supplementary material, we proofed that for maximizing Eq. (3.12), we can minimize cross-entropy loss between each prototypes features and identities. In fact, by doing this, the model makes the part prototype regions lie on the object and have enough information about the foreground. In other words, each part prototype should be able to recover the identity of the person in the images. So part ID-discriminative loss is defined as:

$$\mathcal{L}_{\mathrm{p}} = \frac{-1}{K} \sum_{k=1}^{K} \sum_{c=1}^{C_y} y_c \log(W_c^k(\mathbf{p}_m^k)), \tag{3.13}$$

where W^k is linear layer predicting the probability of identity y_c from $\mathbf{p}_m^k$. To ensure prototype diversity and discrimination, we do not share parameters of W^k and W^q and randomly drop out a proportion of them during training, preventing any single part from dominating.

3.3.2 Bidirectional Multi-step Learning

After identifying prototypes, multiple auxiliary intermediate domains are created, and features are extracted using the Attentive Prototype Embedding (APE) module.

(a) Attentive Prototype Embedding. Instead of directly concatenating learned prototypes as the final feature, we process all other prototypes in input images. The motivation of this design is that each prototype may discriminate the part-attributes shared for all individuals rather than ID attributes. The Attentive Prototype Embedding module (which is detailed in the supply. materials), $\mathcal{F}$, leverages relevant person information between prototypes by applying an attention mechanism to achieve person-aware final features. To aggregate the ID-discriminative features of prototypes, the APE module first uses a fully connected layer to score and emphasize important channels in their feature map. It weights each prototype feature by computing the similarities between them.

$$\begin{aligned} \mathcal{F}(\mathbf{A}_m) = \mathbf{W}_{\mathrm{mlp}}(\mathbf{C}_m), \text{ where } \mathbf{C}_m = \mathbf{W}_v(\mathbf{A}_m) \otimes \mathbf{B}_m, \\ \mathbf{B}_m = \sigma(\mathbf{W}_q(\mathbf{A}_m) \otimes \mathbf{W}_k(\mathbf{A}_m)), \end{aligned} \tag{3.14}$$

and

$$\mathbf{A}_m = [\mathbf{p}_m^1; \mathbf{p}_m^2; \ldots; \mathbf{p}_m^K], \tag{3.15}$$

where $\mathbf{A}_m \in \mathbb{R}^{K\times d}$, $\otimes$ is matrix multiplication and σ is the Sigmund function. $\mathbf{W}_{\text{mlp}}$, $\mathbf{W}_v$, $\mathbf{W}_q$ and $\mathbf{W}_k$ are linear layer. Also, to increase the discriminative ability of the final feature embedding, the global features, $\mathbf{g}_m$, extracted by the backbone head are concatenated to the output of APE:

$$\mathbf{f}_m = [\mathcal{F}(\mathbf{A}_m); \mathbf{g}_m], \tag{3.16}$$

where $\mathbf{g}_m$ is denoted by $\frac{1}{H\cdot W}\sum_{u\in U}[\mathbf{F}_m]_u$.

Our APE has two advantages: (1) it allows adjusting modality-agnostic attention between prototypes regardless of the modality by applying cross-modality prototype features, and (2) it improves the representation ability of final embedding by emphasizing most discriminative prototypes.

(b) Bidirectional Multi-Step Learning. To deal with the significant shift between modalities in the feature space, the model is trained via multiple intermediate steps that gradually bridge this domain gap. Using intermediates with less domain shift, the model gradually learns to leverage cross-modality discriminating clues at each step(Wang *et al.*, 2022a; Chen & Chao, 2021; Abnar *et al.*, 2021; Zhang *et al.*, 2021a). Initially, the discrepancies between modalities are small, letting the model learn from the easier samples first, then converge on more complex cases with larger shifts. One solution to achieve these intermediates is to start from one modality as the source and transform gradually to the other as the target. However, this makes the model forget the learned knowledge of the source and be biased on the target.

To address this issue, modalities are transformed bidirectionally to create an intermediate domain at each step. By gradually transforming domains to each other, the domain gap vanishes over multiple steps. Each intermediate auxiliary step provides discriminative information about the person across both modalities, enabling the training process to transform inputs from one to the other. By creating these intermediate domains, our model aims to neither lose nor duplicate information from the main domains. For gradual transformation in each step, we exchange the features of the same prototypes from both modalities to create a mixed and virtual representation space with less domain discrepancies. These prototype-mixed intermediate

feature spaces are used alongside V and I prototypes to provide APE along with the ability to extract common features from the modalities. For step t, which $t \in \{1, \ldots, T\}$ and T is the number of intermediate steps, the intermediate features for the same person are generated using a mixing function $\mathcal{G}(.,.,.)$ as:

$$\mathbf{A}_m^{(t)} = \mathcal{G}(\mathbf{A}_m, \mathbf{A}_{\tilde{m}}, t), \tag{3.17}$$

where $\tilde{m} \neq m$ which mixes the prototypes from two modalities. For example, we use a simple but yet effective random prototype mixing strategy for $\mathcal{G}(.,.,.)$:

$$\mathcal{G}(\mathbf{A}_m, \mathbf{A}_{\tilde{m}}, t) = [\mathbf{p}^1_{r(m,\tilde{m},t)}, \ldots, \mathbf{p}^K_{r(m,\tilde{m},t)}], \tag{3.18}$$

where $r(.,.,.)$ is a weighted random selector:

$$r(m, \tilde{m}, t) = \begin{cases} m & t/T \leq \mathcal{U}(0,1) \\ \tilde{m} & \text{else} \end{cases} \tag{3.19}$$

where $\mathcal{U}(0,1)$ is a uniform random generator. For intermediate step ($t < T$), we have $\mathbf{f}_m^{(t)} = [\mathcal{F}(\mathbf{A}_m^{(t)}); \mathbf{g}_m]$ and in the last step $\mathbf{f}_m^{(T)} = [\mathcal{F}(\mathbf{A}_m^{(T)}); \mathbf{g}_{\tilde{m}}]$ is used. The module $\mathcal{F}$ learns to gradually reduce the discrepancies in representation space by applying metric learning objectives bidirectionally between $\mathbf{f}_i$ and $\mathbf{f}_i^{(t)}$ and between $\mathbf{f}_v$ and $\mathbf{f}_v^{(t)}$ as :

$$\begin{aligned} \mathcal{L}_{\text{bcc}} &= \mathcal{L}_{\text{cc}}(\mathbf{f}_v, \mathbf{f}_v^{(t)}) + \mathcal{L}_{\text{cc}}(\mathbf{f}_i, \mathbf{f}_i^{(t)}) \\ \mathcal{L}_{\text{bce}} &= \mathcal{L}_{\text{ce}}(\mathbf{f}_v) + \mathcal{L}_{\text{ce}}(\mathbf{f}_v^{(t)}) + \mathcal{L}_{\text{ce}}(\mathbf{f}_i) + \mathcal{L}_{\text{ce}}(\mathbf{f}_i^{(t)}) \\ \mathcal{L}_{\text{re}} &= \mathcal{L}_{\text{bce}} + \mathcal{L}_{\text{bcc}}, \end{aligned} \tag{3.20}$$

that $\mathcal{L}_{\text{cc}}$, $\mathcal{L}_{\text{ce}}$ are center cluster and cross-entropy losses (Wu *et al.*, 2021). Since $\mathbf{f}_i^{(t)}$ begins from $\mathbf{f}_i$ at first steps and approaches to $\mathbf{f}_v$ at ending steps, the model is gradually trained to extract more robust modality invariant features by seeing samples with low modality gap to the harder ones with higher gap(Wang *et al.*, 2022a).

3.3.3 Training and Inference

In each step, the model tries to represent the same feature space for two input domains, V/I and the intermediate ones, using the overall loss expressed as:

$$\mathcal{L} = \mathcal{L}_{\text{re}} + \lambda_{\text{f}}(\mathcal{L}_{\text{lc}} + \mathcal{L}_{\text{hc}}) + \lambda_{\text{v}}\mathcal{L}_{\text{vc}} + \lambda_{\text{c}}\mathcal{L}_{\text{c}} + \lambda_{\text{i}}\mathcal{L}_{\text{p}} + \lambda_{\text{e}}\mathcal{L}_{\text{eq}} \tag{3.21}$$

where λ are hyper-parameters for weighting losses. During inference, the tail and head backbones are combined to extract the prototypes and global features as $\mathbf{A}_i$, $\mathbf{A}_v$, $\mathbf{g}_i$ and $\mathbf{g}_v$ from given input image x_i and x_v. Using Eq. (3.16), the $\mathbf{f}_i$ and $\mathbf{f}_v$ are computed by the APE. The matching score for a query input image is computed by cosine similarly.

3.4 Results and Discussion

In this section, we compare the BMDG with SOTA V-I ReID methods, including global, part-based, and intermediate approaches. Extensive ablation studies are conducted to show the effectiveness of the proposed bi-directional learning, the impact of the number of parts-prototypes, and intermediate steps during learning. The supplementary materials provide detailed information on our experimental methodology, including the implementation of baseline models, performance measures, and descriptions of the SYSU-MM01 (Wu *et al.*, 2020), RegDB (Nguyen *et al.*, 2017), and LLCM (Zhang & Wang, 2023b) datasets.

3.4.1 Comparison with State-of-Art Methods:

Table 3.1 compares the accuracy of BMDG with state-of-the-art V-I ReID approaches. Our experiments on the SYSU-MM01 and RegDB datasets show that BMDG outperforms the SOTA methods in the majority of situations. Compared to (Kim *et al.*, 2023), BMDG has lower R1 accuracy (-1%) in the All Search settings on the SYSU-MM01 dataset. However, it achieves higher performance in Indoor settings (+2%) and on the RegDB dataset (+10%). These results can be attributed to BMDG's ability to effectively capture more ID-related knowledge across modalities. This is achieved by learning discriminative part-prototypes and gradually reducing modality-specific information in the extracted final features. Additionally, the prototype

Table 3.1 Accuracy of the proposed BMDG and state-of-the-art methods on the SYSU-MM01 (single-shot setting) and RegDB datasets. All numbers are percent. Results for methods were obtained from the original papers. AIM (Fang *et al.*, 2023) means re-ranking method

Family			SYSU-MM01						RegDB					
			All Search			Indoor Search			Visible → Infrared			Infrared → Visible		
	Method	**Venue**	**R1**	**R10**	**mAP**	**R1**	**R10**	**mAP**	**R1**	**R10**	**mAP**	**R1**	**R10**	**mAP**
Global	AGW (Ye *et al.*, 2020b)	TPAMI'20	47.50	-	47.65	54.17	-	62.97	70.05	-	50.19	70.49	87.21	65.90
	CAJ (Ye *et al.*, 2021b)	ICCV'21	69.88	-	66.89	76.26	97.88	80.37	85.03	95.49	79.14	84.75	95.33	77.82
	RAPV-T (Zeng *et al.*, 2023)	ESA'23	63.97	95.30	62.33	69.00	97.39	75.41	86.81	95.81	81.02	86.60	96.14	80.52
	G2DA (Wan *et al.*, 2023)	PR'23	63.94	93.34	60.73	71.06	97.31	76.01	-	-	-	-	-	-
Part-based	DDAG (Ye *et al.*, 2020a)	ECCV'20	54.74	90.39	53.02	61.02	94.06	67.98	69.34	86.19	63.46	68.06	85.15	90.31
	SSFT (Lu *et al.*, 2020)	CVPR'20	63.4	91.2	62.0	70.50	94.90	72.60	71.0	-	71.7	-	-	-
	MPANet (Wu *et al.*, 2021)	CVPR'22	70.58	96.21	68.24	76.74	98.21	80.95	83.70	-	80.9	82.8	-	80.7
	SAAI (Fang *et al.*, 2023)[a]	ICCV'23	75.03	-	71.69	-		-	-	-	-	-	-	-
	SAAI (Fang *et al.*, 2023) [b]	ICCV'23	75.90	-	77.03	82.20	-	80.01	91.07	-	91.45	92.09	-	92.01
	CAL (Wu *et al.*, 2023a)	ICCV'23	74.66	96.47	71.73	79.69	98.93	83.68	94.51	**99.70**	88.67	93.64	**99.46**	87.61
	PartMix (Kim *et al.*, 2023)	CVPR'23	77.78	-	74.62	81.52	-	84.83	84.93	-	82.52	85.66	-	82.27
Intermediate	SMCL(Wei *et al.*, 2021)	ICCV'21	67.39	92.84	61.78	68.84	96.55	75.56	83.93	-	79.83	83.05	-	78.57
	MMN (Zhang *et al.*, 2021c)	ICM'21	70.60	96.20	66.90	76.20	**99.30**	79.60	91.60	97.70	84.10	87.50	96.00	80.50
	RPIG (Alehdaghi *et al.*, 2022)	ECCVw'22	71.08	96.42	67.56	82.35	98.30	82.73	87.95	98.3	82.73	86.80	96.02	81.26
	FTMI (Sun *et al.*, 2023)	MVA'23	60.5	90.5	57.3	-	-	-	79.00	91.10	73.60	78.8	91.3	73.7
	G2DA (Wan *et al.*, 2023)	PR'23	63.94	93.34	60.73	71.06	97.31	76.01	-	-	-	-	-	-
	SEFL (Feng *et al.*, 2023)	CVPR'23	75.18	96.87	70.12	78.40	97.46	81.20	91.07	-	85.23	92.18	-	86.59
	BMDG (ours)[a]	-	76.15	97.42	73.21	83.53	98.02	84.92	93.92	98.11	89.18	94.08	97.0	88.67
	BMDG (ours)[b]	-	**78.08**	**97.90**	**78.22**	**83.59**	98.96	**86.35**	**94.76**	98.91	**92.21**	**94.56**	98.31	**93.07**

[a] without AIM [b] with AIM

alignment module creates gradual and bidirectional intermediate spaces without sacrificing discriminative ability, allowing BMDG to surpass intermediate methods. Moreover, BMDG can be easily integrated into different part-based V-I ReID models, enhancing generality through gradual training. To show its effectiveness, we integrated our approach into state-of-the-art part-based or prototype-based baseline models (Ye *et al.*, 2020a; Wu *et al.*, 2021; Fang *et al.*,

Table 3.2 Accuracy of the proposed BMDG and state-of-the-art methods on the Large LLCM Dataset

Family		**LLCM**			
		V → I		I → V	
Method	**Venue**	**R1**	**mAP**	**R1**	**mAP**
DDAG(Ye *et al.*, 2020a)	ECCV'20	40.3	48.4	48.0	52.3
CAJ(Ye *et al.*, 2021b)	ICCV'21	56.5	59.8	48.8	56.6
RPIG(Alehdaghi *et al.*, 2022)	ECCVw'22	57.8	61.1	50.5	58.2
DEEN(Zhang & Wang, 2023a)	CVPR'23	62.5	65.8	54.9	62.9
BMDG (ours)	-	**63.4**	**66.3**	**56.4**	**63.5**

Table 3.3 Accuracy of part-based ReID methods with BMDG on the SYSU-MM01, under single-shot setting. Results were obtained by executing the author's code on our servers

Method	**R1 (%)**	**mAP (%)**
DDAG (Ye *et al.*, 2020a)	53.62	52.71
DDAG with BMDG	**55.36** (+1.74↑)	**54.05** (+1.34↑)
MPANet (Wu *et al.*, 2021)	66.24	62.89
MPANet with BMDG	**68.74** (+2.50↑)	**64.25** (+1.36↑)
SAAI (Fang *et al.*, 2023)	71.87	68.16
SAAI with BMDG	**73.69** (+1.82↑)	**70.08** (+1.92↑)

2023). We executed the original code published by authors using the hyper-parameters and other configurations they provided with and without BMDG. Table 3.3 shows that BMDG improves performance for mAP and rank-1 accuracy for all methods by an average of 2.02% and 1.54 %, respectively. Similar performance improvements are observed on the large and complex LLCM dataset (see Table 3.2). Since LLCM was introduced recently, few papers reported their results, so we executed other approaches with published code on the LLCM dataset.

3.4.2 Ablation Studies:

(a) Step size and number of part prototypes. We evaluate the best option for the number of parts and steps. Table 3.4 (top) shows that using the BMDG approach increases the R1 accuracy when increasing the number of steps T for a specific number of prototypes K. Also, it increases performance in multi-step with K parts, letting the model learn from the lower gap sample and converge on harder cases.

Another option for each step, instead of exchanging some parts, is to use the Mixup(Zhang *et al.*, 2018) or IDM (Dai *et al.*, 2021) strategy for all parts from two modalities with normalized weight: $\mathbf{A}_m^{(t)} = \alpha^{(t)}\mathbf{A}_m + (1-\alpha^{(t)})\mathbf{A}_{m'}$, where α starts from 0 to 1 w.r.t. step t and number of steps. Results in Table 3.4 (bottom) show that using multi-step increases performance marginally since it does not utilize the ability of prototypes to create meaningful intermediate steps. This is explained by the fact that prototypes contain some modality-specific information, which is crucial for transforming domains in intermediate steps and may be lost in the averaging process.

(c) Loss functions. Contrastive objectives between parts ($\mathcal{L}_{\text{lc}}$ and $\mathcal{L}_{\text{hc}}$) are essential for the model to detect similar semantics between objects, thereby creating an intermediate step by swiping corresponding parts. Table 3.5 shows this necessity when used in conjunction with $\mathcal{L}_{\text{c}}$. Using such regularization in the multi-step approach increases mAP by 3% and R1 by 4%. Also, results of part separation loss, $\mathcal{L}_{\text{vc}}$, indicate that using this loss pushes the parts to be scattered on the body regions and gives the part features a more robust representation. Additionally, it enables the intermediate steps to generalize the feature space better by increasing the performance over the single-step with multi-step setting. The part identity loss $\mathcal{L}_{\text{p}}$ lets the model extract ID-related prototype features that contain discriminative information about the object and generate an id-informative intermediate step. Without this objective, parts features are likely to focus on background information. As shown in the two last rows, performance increases when using this objective.

(d) Bidirectional domain generalization. To show the impact of the BMGD approach, we conduct experiments under the following settings: (1) single-step, (2) one-directional multi-step

Table 3.4 R1 accuracy of BMDG using (a) our prototype mixing and (b) IDM (Dai *et al.*, 2021) for different numbers of part prototypes (K) and intermediate steps (T)

T	**Number of part prototypes (K)**					
	3	4	5	6	7	10
Prototype exchanging (Our)						
0	68.25	69.24	69.40	70.27	70.03	68.12
1	69.97	70.81	72.07	71.97	71.31	69.28
2	71.20	72.35	73.98	73.61	72.45	71.25
3	73.32	73.94	74.11	74.98	73.22	71.67
4	-	74.08	74.15	**75.43**	73.51	71.99
6	-	-	-	75.37	73.52	72.15
10	-	-	-	-	-	72.07
Entire mixup (IDM (Dai *et al.*, 2021))						
0	68.25	69.24	69.4	70.27	70.03	68.12
1	68.53	69.19	69.56	70.53	70.3	68.41
4	68.70	70.03	69.75	70.84	70.77	68.53
6	69.13	70.81	70.1	**71.16**	70.29	68.90
10	69.21	70.15	69.52	70.95	70.90	68.13

Table 3.5 Impact on the accuracy of using different BMDG losses, using SYSU-MM01 for single-step ($T = 0$) and multi-step ($T = K$) settings. The baseline $\mathcal{L}_{\text{re}}$ loss is used in all cases

Losses						**R1 (%)**		**mAP (%)**	
$\mathcal{L}_{\text{p}}$	$\mathcal{L}_{\text{lc}}$	$\mathcal{L}_{\text{hc}}$	$\mathcal{L}_{\text{vc}}$	$\mathcal{L}_{\text{c}}$	$\mathcal{L}_{\text{eq}}$	$T = 0$	$T = K$	$T = 0$	$T = K$
baseline						51.09	-	49.68	-
	✓	✓				62.19	65.28	58.91	60.73
	✓	✓	✓			66.35	69.50	63.01	66.48
			✓	✓		66.67	67.31	61.04	61.68
✓				✓		59.78	60.66	55.41	56.05
✓	✓	✓	✓			68.30	69.04	66.87	66.99
	✓	✓	✓	✓		69.68	71.20	67.08	68.13
✓	✓	✓	✓	✓		69.90	72.67	68.11	69.75
✓	✓	✓	✓	✓	✓	**70.27**	**75.37**	**68.45**	**72.71**

(from V $\rightarrow$ I and I $\rightarrow$ V), and (3) bidirectional for creating the intermediate series. The results are shown in Table 3.6. Although using multiple intermediate steps from one modality to another improves the model's accuracy by allowing it to gradually learn and utilize common discriminative features, this process can lead to losing information about the source modality. The bi-directional approach addresses these issues and achieves the best performance, as it is not biased towards any specific modality.

Table 3.6 Impact on rank-1 and mAP accuracy using different settings for multi-step domain generalization. All numbers are percent. $V \rightarrow I$ means that the intermediates are created only from V, by replacing prototypes from I features

Settings	**Number of part prototypes** (K)							
	4		5		6		7	
	R1	mAP	R1	mAP	R1	mAP	R1	mAP
Single step	69.2	65.9	69.4	66.1	70.2	66.3	70.4	66.5
One (V $\rightarrow$ I)	71.0	66.5	71.5	67.1	72.3	67.5	71.6	66.8
One (I $\rightarrow$ V)	70.8	66.3	71.2	67.0	72.6	67.6	71.0	66.7
Bidirectional	**74.0**	**71.8**	**74.1**	**71.9**	**75.4**	**72.8**	**73.5**	**70.2**

3.5 Conclusion

This paper introduces the BMDG framework for V-I ReID. It uses a novel prototype learning approach to learn the most discriminative and complementary set of prototypes from both modalities. These prototypes generate multiple intermediate domains between modalities by progressively mixing prototypes from each modality, reducing the domain gap. The effectiveness of BMDG is shown experimentally on several datasets, where it outperforms state-of-the-art methods for V-I person ReID. Results also highlight the impact of considering multiple intermediate steps and bidirectional training to improve accuracy. Moreover, BMDG can be integrated into other part-based V-I ReID methods, significantly improving their accuracy. A limitation of BMDG is its dependency on discriminant body parts.

CHAPTER 4

MIXER: FROM CROSS-MODAL TO MIXED-MODAL VISIBLE-INFRARED RE-IDENTIFICATION

Mahdi Alehdaghi[1] , Rajarshi Bhattacharya[1] , Pourya Shamsolmoali[2] , Rafael M. O. Cruz[1] , Eric Granger[1]

[1] LIVIA, ILLS, Dept. of Systems Engineering, École de Technologie Supérieure, 1100 Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3
[2] University of York, U.K.

The article was submitted in IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), September 2025

4.1 Introduction

VI-ReID is essential in surveillance, enabling the matching of individuals across different camera views and lighting conditions. Current VI-ReID methods typically focus on cross-modality matching, where a V or I image serves as the query against a gallery of images from the other modality (Fig.4.1a, top), like comparing day-time V images to night-time I images. While effective for cross-modality matching, real-world surveillance captures both modalities around the clock, leading to a gallery with mixed V and I images (Fig.4.1a, bottom). This mixed gallery more accurately reflects real-world conditions, where a considerable challenge arises — same-modality images in the gallery (e.g., V-to-V) belonging to different identities exhibit lower domain gaps than cross-modality images (e.g., I-to-V). Future VI-ReID methods must address these mixed gallery scenarios, balancing cross- and same-modality matching for reliable person ReID in real-world applications.

VI-ReID methods are generally expected to perform well in mixed-modality settings, which can be interpreted in various ways. We identify four distinct interpretations based on the identity and camera relationships between query and gallery images. In each setting, specific images are removed to assess the robustness of the matching model. The mixed-modal setting introduced in (Liu *et al.*, 2025), which considers only same-modality images of the same identity as the query, lacks informativeness and realism. To address this, we propose more realistic and challenging

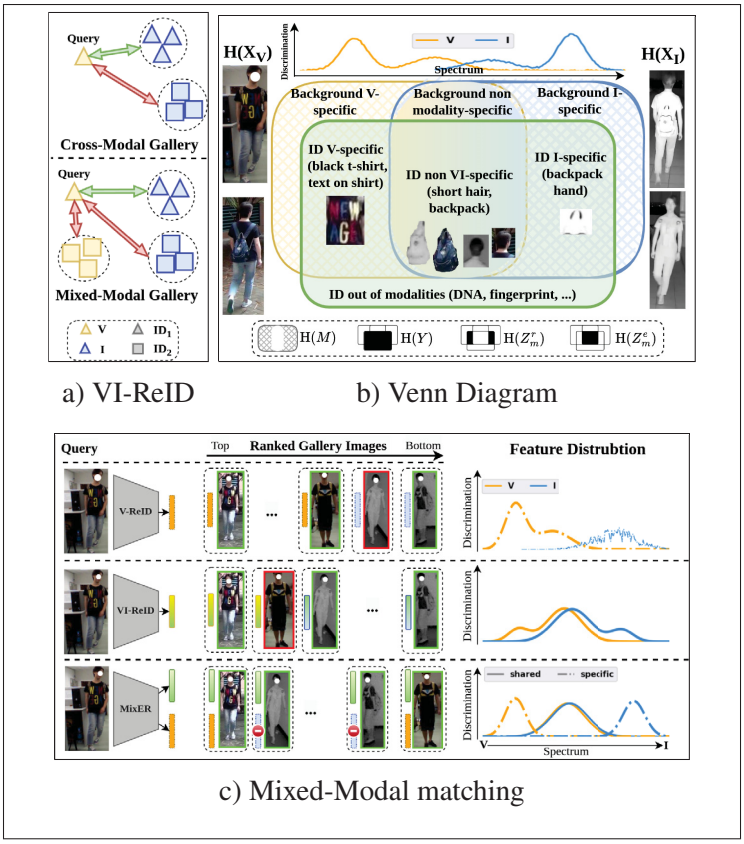


Figure 4.1 (a) VI-ReID of a query image matched against a cross-modal and mixed-modal gallery. (b) With VI-ReID methods, V, I, and ID information reveal modality-specific and modality-shared ID features. (c) Mixed-modal approaches leverage these features for matching, while uni-modal methods are limited to intra-modality matching due to a large modality gap, and cross-modal methods enable inter-modal matching by extracting shared features. Our MixER method disentangles these features to reduce the gap in shared features and enhance the discrimination in modality-specific features ($H(\cdot)$ measures the entropy of a random variable)

evaluation settings that include both same-modality images of the query identity and different individuals. This design tests model robustness against low-modality gaps, ensuring effective identity discrimination within the same modality while also handling cross-modal variations. However, we show that two specific settings pose significant challenges, where state-of-the-art uni-modal and cross-modal ReID methods struggle to achieve high performance. A cost-effective VI-ReID solution in these settings should rely on a single backbone model to extract discriminative features and dynamically adapt them based on intra-modality or inter-modality matching requirements. Fig. 4.1b illustrates the identity-discriminative information present in the

two modalities, which can be categorized into modality-specific and modality-shared attributes. While V-ReID models primarily extract modality-specific features, VI-ReID methods aim to capture modality-invariant identity-discriminative features for effective cross-modal matching.

Given the substantial discrepancy between I and V modality images, uni-modal models typically struggle to remain discriminative across modalities (as illustrated in the first row of Fig.4.1c). To address this challenge, VI-ReID methods often focus on minimizing this discrepancy between their extracted feature representations. For example, GAN methods (Wang *et al.*, 2019a; Kniaz *et al.*, 2018) are used to translate the modalities and shared-backbones (Ye *et al.*, 2020b; Cui, Zhou & Peng, 2024; Park *et al.*, 2021; Alehdaghi *et al.*, 2022; Ren & Zhang, 2024; Feng *et al.*, 2023; Zhang & Wang, 2023b) are used to map images into a modality-invariant feature space, creating a shared representation that minimizes modality discrepancies. Despite the significant improvements in recent years, the features extracted by these methods often contain some modality-specific information, which weakens the effectiveness of inter-modal matching and limits the minimization of modality gap in the feature space.

A limitation of current VI-ReID methods in mixed matching tasks is their lack of attention to identity features that exist only in one modality, as shown in the second row of Fig.4.1c. Some attributes, such as shirt patterns that are only relevant in the V modality or material textures unique to I, are modality-specific and, therefore, unavailable in the other modality (see Fig.4.1b). For accurate cross-modal matching, this modality-specific information should be erased from identity representations, only allowing access to modality-shared information across both V and I. However, effectively separating modality-specific from modality-invariant is challenging, as it requires effective unsupervised disentanglement to identify them.

To address the limitations of VI-ReID in mixed-modal scenarios, this paper introduces a **Mix**ed Modality-**E**rased and Modality-**R**elated (MixER) ID-discriminative feature learning approach. By isolating discriminative modality-specific features within one subspace, our approach promotes modality-erased features to capture robust discriminative semantic concepts across modality variations within the other subspace. Implicit Discriminative Knowledge Learning

(IDKL) (Ren & Zhang, 2024) uses modality-specific features to enhance modality-shared ones through knowledge distillation and implicit similarity. In contrast, our approach enforces an orthogonal complement structure. This constraint ensures that modality-related features remain independent of shared features. This allows discovering a diverse embedding space to ensure both modality-specific and -shared are effectively used. Building on these assumptions and inspired by (Feng *et al.*, 2023), we formulate our objectives from a mutual information perspective, demonstrating that joint learning of modality-erased and modality-related features optimizes mutual information with identity, producing effective feature representations for mixed-modal matching.

As illustrated in Fig.4.1c, MixER leverages modality-erased features for inter-modality matching while refining intra-modality matching by mixing them with modality-related features via ID-modality-aware, modality-confusion, and feature-fusion losses. Furthermore, our framework can be integrated into state-of-the-art approaches to enhance their performance in mixed-modal and cross-modal matching settings.

The main contributions of this paper are summarized as follows. **(1)** We motivate and formalize more comprehensive and challenging settings for mixed-modal V-I ReID evaluation, where galleries may have data from both I and V modalities. **(2)** A mixed modality-erased and -related (MixER) feature learning paradigm is introduced for enhancing mixed- and cross-modal VI-ReID on a single feature embedding backbone. It separates modality-erased features from modality-related ones through orthogonal decomposition and gradient reversal. Modality-erased features capture modality-shared discriminative semantics, while modality-related features are designed to extract additional discriminative attributes unique to each modality, thereby improving the diversity of learned representations. **(3)** Our experiments on SYSU-MM01, RegDB, and LLCM show that MixER strategy outperforms state-of-the-art VI-ReID methods in cross-modal and mixed-modal settings. Additionally, it can be integrated into any VI-ReID method, improving its performance for retrieval applications.

4.2 Related Work

(a) Cross-modal Person Re-Identification. Person ReID aims to identify individuals by matching distinctive characteristics in query images within a larger gallery (Chen *et al.*, 2018; Zheng, Yang & Hauptmann, 2016b; Zheng *et al.*, 2020; Zhou, Zhong, Cheng, Liang & Ma, 2023). VI-ReID extends this task to low-light conditions, but modality-specific discrepancies pose challenges. Existing methods extract global (Wang *et al.*, 2019a; Kniaz *et al.*, 2018; Alehdaghi *et al.*, 2022; Ye *et al.*, 2020b; Ye *et al.*, 2020; Zhu *et al.*, 2020b; Park *et al.*, 2021) or part-based features (Ye *et al.*, 2020a; Lu *et al.*, 2020; Huang *et al.*, 2022; Wu *et al.*, 2021; Alehdaghi, Shamsolmoali, Cruz & Granger, 2024) using striping or attention mechanisms. While effective for ID discrimination, these methods struggle to learn modality-invariant representations. Recent approaches focus on modality-invariant feature learning by disentangling identity features from modality-specific attributes (Choi *et al.*, 2020a; Yang *et al.*, 2020; Lu *et al.*, 2020; Zhu *et al.*, 2023; Zheng, Li, Han, Gao & Sang, 2019; Lu, Lin & Hu, 2024; Feng *et al.*, 2023). GAN methods (Choi *et al.*, 2020a; Yang *et al.*, 2020; Lu *et al.*, 2020; Alehdaghi *et al.*, 2023) attempt to separate content from modality style but often lose identity-related details. Other techniques enforce orthogonality between modality-specific and invariant features (Zhu *et al.*, 2023; Zheng *et al.*, 2019), though ensuring complete modality suppression remains a challenge. Additionally, shape-based approaches (Lu *et al.*, 2024; Feng *et al.*, 2023) leverage body structure to refine identity features while minimizing modality bias.

(b) Multi-modal Learning. Different data modalities offer complementary features for representation, an area explored in multi-modal learning (Liang, Zadeh & Morency, 2022). Recent transformer-based methods (Gabeur, Sun, Alahari & Schmid, 2020; Kim, Son & Kim, 2021; Li *et al.*, 2021a) fuse multi-modal inputs directly as tokens rather than extracting separate modality-specific features. Real-world applications, however, often lack complete modality availability, presenting challenges addressed by approaches that handle missing modalities (Wang *et al.*, 2023; Pan, Liu, Xia & Shen, 2021; Cai, Wang, Gao, Shen & Ji, 2018; Lee, Tsai, Chiu & Lee, 2023). The ImageBind (Girdhar *et al.*, 2023) exemplifies large-scale multi-modal unification, mapping diverse modalities into a shared representation space. In ReID, (Li, Ye,

Zhang & Du, 2024) presents a unified model across RGB, infrared, sketch, and text modalities, achieving strong cross-domain performance. Inter-ReID (Pang, Zhao, Liu, Sharma & Wang, 2024) focuses on multi-modal ReID. Given an image of a person captured in one modality, the objective is to rank similar gallery images in the other modalities. Instruct-ReID (He *et al.*, 2024) further unifies ReID tasks by leveraging a multi-modal backbone for text and image prompts, achieving SOTA performance across various ReID settings. In contrast, we formalize new settings for a comprehensive and realistic mixed-modal ReID evaluation as the general use case for cross-modal person ReID. In particular, it allows us to evaluate the challenge posed by a low modality gap, where the primary difficulty lies in distinguishing individuals of the same modality but different identities.

(c) Critical Analysis. While significant focus has been placed on cross-modal V-I person ReID and its associated challenges, the mixed-modality query-gallery scenario remains largely unaddressed. Despite achieving SOTA results in standard settings, existing VI-ReID methods do not report their performance in mixed-modality settings. (Liu *et al.*, 2025) proposed a specific mixed-modal setting and proposes a ReID approach that enhances the performance with this setting by applying a re-ranking search on same-modality samples in the gallery. It does not explicitly extract modality-specific information to improve either same- or cross-modal matching. Furthermore, while some methods (Ren & Zhang, 2024; Zhu *et al.*, 2023) attempt to disentangle modality-specific features to enhance the modality-shared component in cross-modal contexts, we show that directly using modality-specific information does not improve inter-modal matching, where such information is unavailable and must instead be erased from shared features. These modality-specific features are, however, beneficial for intra-modal matching within mixed-modality settings. Our findings are further supported by mutual information analysis and experimental results across proposed mixed-modal scenarios.

4.3 The Proposed MixER Method

To address the limitation of ReID methods in mixed-modal settings, we propose the MixER learning paradigm to combine ID-discriminative modality-erased and modality-related features. Fig.4.2 provides an overview of MixER learning approach. It relies on a shared backbone with three sub-modules to extract independent modality-related and -erased features by applying the orthogonal decomposition, modality-confusion, and modality-related losses.

Problem Definition. A multimodal dataset for VI-ReID is composed of visible $\mathcal{V} = \{x_v^{(j)}, y_v^{(j)}\}_{j=1}^{N_v}$ and infrared $\mathcal{I} = \{x_i^{(j)}, y_i^{(j)}\}_{j=1}^{N_i}$ sets of images from C_y distinct individuals, with their ID labels. Our proposed Mixed VI-ReID system seeks to match images captured from V and I cameras by using one deep backbone model that encodes modality-invariant person embeddings, denoted by $\mathbf{z}_v$ and $\mathbf{z}_i$. Given query images (V or I), the objective is to retrieve images with the same identity over the gallery set containing both V and I modalities, by computing and sorting the distance value $D(\cdot, \cdot)$ for each gallery image:

$$D(\mathbf{z}_m^{(j)}, \mathbf{z}_{m'}^{(p)}) < D(\mathbf{z}_m^{(j)}, \mathbf{z}_{m''}^{(n)}), \tag{4.1}$$

where $y_m^{(j)} = y_{m'}^{(p)} \neq y_{m''}^{(n)}$, m, m', m'' are modalities that could be v or i independently, and superscripts p and n indicating indices of positive and negative samples, respectively. To learn these features, we decompose them into two independent components, each meeting specific constraints suitable for inter-modal and intra-modal matching. This is achieved by maximizing the mutual information (MI) between these features and the ID labels.

4.3.1 Mutual Information Analysis

To extract ID-discriminative features from images, the model needs to maximize the MI between these extracted features[1], Z_m, and label spaces, Y:

$$\max_{Z_m} \quad \mathrm{MI}(Z_m; Y), \tag{4.2}$$

[1] Uppercase is used as random variables and lowercase for samples.

where Y represents the identity of the individuals in the input images. To ensure that the learned features are effective in mix-modality scenarios, they are decomposed into two independent components: **(a) modality-erased** (Z_m^e), which should not contain any modality information to be proper for inter-modal matching (e.g., short-hair attributes in Fig.4.1b), and **(b) modality-related** (Z_m^r) component to refine the modality-erased for improving the intra-modal matching. This component should contain ID information that is also relevant to the modality (e.g., text on a t-shirt in Fig.4.1b). To have two independent components, the MI between them must be zero. Thus, the optimization becomes:

$$\begin{aligned} \max_{Z_m^e, Z_m^r} \quad & \{\mathrm{MI}(Z_m^e, Z_m^r; Y)\} \text{ s.t. } \mathrm{MI}(Z_m^e; Z_m^r) = 0, \\ & \mathrm{MI}(Z_m^e; M) = 0 \text{ and } \mathrm{MI}(Z_m^r; Y|M) = 0, \end{aligned} \tag{4.3}$$

where M is the modality label space. Constraint $\mathrm{MI}(Z_m^e; M)$=0 ensures that modality information is erased from Z_m^e and $\mathrm{MI}(Z_m^r; Y|M)$=0 ensures that Z_m^r does not contain ID-aware information, which disregards the modality label. Also, Z_m^e and Z_m^r should be independent.

Theorem 4.1. *If Z_m^e and Z_m^r are independent, then $MI(Z_m^e, Z_m^r; Y) \geq MI(Z_m^e; Y) + MI(Z_m^r; Y)$.*

Proof. Can be found in the supplementary materials. □

So we can lower-bound $\mathrm{MI}(Z_m^e, Z_m^r; Y)$ by $\mathrm{MI}(Z_m^e; Y) + \mathrm{MI}(Z_m^r; Y)$. By incorporating M into Eq. (4.3) and using Theorem 4.1:

$$\begin{aligned} \mathrm{MI}(Z_m^e, Z_m^r; Y) \geq & \mathrm{MI}(Z_m^e; Y) + \mathrm{MI}(Z_m^r; Y) \\ = & \mathrm{MI}(Z_m^e; Y|M) + \mathrm{MI}(Z_m^e; Y; M) \\ & + \mathrm{MI}(Z_m^r; Y|M) + \mathrm{MI}(Z_m^r; Y; M). \end{aligned} \tag{4.4}$$

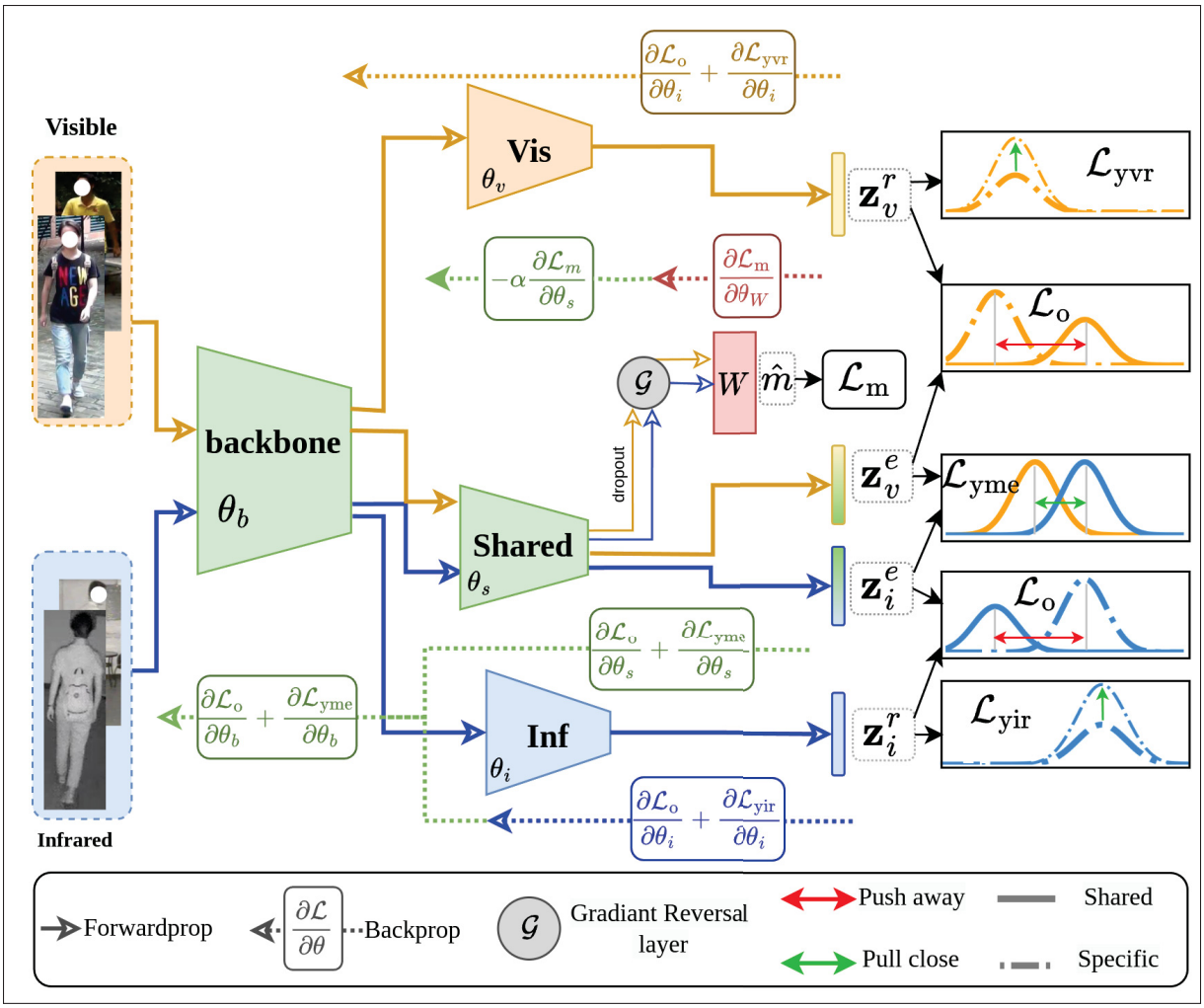


Figure 4.2 The overall architecture of our proposed MixER method. It extracts two independent ID-discriminative feature vectors by orthogonal decomposition, modality-confusion, and modality-aware losses for learning modality-erased and modality-related feature representation

Since the MI(.;.) is non-negative, MI($Z_m^e; Y; M$)=0, and MI($Z_m^r; Y|M$)=0, Eq. (4.3) can be formulated as the maximization of the following Lagrangian:

$$\max_{Z_m^e, Z_m^r} \{ \overbrace{\mathrm{MI}(Z_m^e; Y|M) - \lambda_1 \mathrm{MI}(Z_m^e; M)}^{\textit{Modality-Erased Learning}} + \overbrace{\mathrm{MI}(Z_m^r; Y; M)}^{\textit{Modality-Related Learning}} - \overbrace{\lambda_2 \mathrm{MI}(Z_m^e; Z_m^r)}^{\text{Orthogonal Feature Learning}} \}. \quad (4.5)$$

4.3.2 Modality Erased and Related Learning

(a) Orthogonal Feature Decomposition. To decompose the extracted features into modality-erased and -related feature vectors, two modality-specific and one shared sub-modules are proposed to project the $\mathbf{z}_m$ to them as:

$$\mathbf{z}_m^e = \mathcal{F}_s(\mathbf{z}_m), \quad \mathbf{z}_m^r = \mathcal{F}_m(\mathbf{z}_m), \tag{4.6}$$

where $\mathbf{z}_m = \mathcal{F}_b(x_m)$ is extracted by a shared backbone. One reason behind using specific sub-modules for modality-related features is to prevent them from sharing information through model parameters.

Minimizing $\mathrm{MI}(Z_m^e; Z_m^r)$: When minimizing the MI between Z_m^r and Z_m^e, they must be independent to avoid affecting each other through the learning. Since the MI estimation is complex and time-consuming (Belghazi *et al.*, 2018; Feng *et al.*, 2023), we estimate the independence as minimizing the orthogonal loss:

$$\mathcal{L}_{\mathrm{o}} = \mathbb{E}_{(\mathbf{z}_m^r, \mathbf{z}_m^e) \sim (Z_m^r, Z_m^e)} \frac{\mathbf{z}_m^r \cdot \mathbf{z}_m^e}{\|\mathbf{z}_m^r\|_2 \, \|\mathbf{z}_m^e\|_2}. \tag{4.7}$$

(b) Learning Modality Erased Features. The goal of modality-erased learning is extracting features, $\mathbf{z}_m^r$, that discriminate the ID without using modality information to make them appropriate for cross-modal matching when the modality-specific information is absent.

Maximizing $\mathrm{MI}(Z_m^e; Y|M)$: given the MI attributes, we expand it as (The proof is in the suppl. materials):

$$\mathrm{MI}(Z_m^e; Y|M) = \mathrm{MI}(Z_m^e; Y) - \mathrm{MI}(Z_m^e; Y; M) = \mathrm{MI}(Z_m^e; Y). \tag{4.8}$$

To maximize Eq. (4.8), the cross-entropy loss $\mathcal{L}_{\mathrm{ce}}$ must be minimized as (see the property III.1 in supply. materials):

$$\mathcal{L}_{\mathrm{meid}}(Z_m^e, Y) = \mathbb{E}_{(\mathbf{z}_m^e, y_m) \sim (Z_m^e, Y)} \mathcal{L}_{\mathrm{ce}}(\mathbf{z}_m^e, y_m). \tag{4.9}$$

To obtain distinct features for each person, the center-cluster loss($\mathcal{L}_{\text{cc}}$) (Wu *et al.*, 2021) is also minimized. So, our modality-erased identity loss is:

$$\mathcal{L}_{\text{yme}} = \mathcal{L}_{\text{meid}}(Z_m^e, Y) + \mathcal{L}_{\text{cc}}(Z_m^e, Y). \tag{4.10}$$

Minimizing $\text{MI}(Z_m^e; M)$: Modality-erased features should not distinguish the origin modality of input samples. Therefore, an adversarial modality classifier based on the Gradient Reversed Layer (GRL)(Ganin & Lempitsky, 2015) is used to propagate the reverse gradient onto model parameters. The modality-confusion loss for this objective is:

$$\mathcal{L}_{\text{m}} = \mathbb{E}_{(\mathbf{z}_m^e, m)\sim(Z_m^e, M)}\mathcal{L}_{\text{ce}}(\mathcal{G}(W^T\mathbf{z}_m^e), m), \tag{4.11}$$

where $\mathcal{G}$ is the GRL and $W \min \mathbb{R}^{d\times 2}$ is a linear layer.

(c) Learning Modality Related Features. To learn features that leverage modality-related ID-discriminating information simultaneously, a new doubled label space is proposed to account for identity alongside the modality by separating the same person in each modality:

$$y'_m \sim Y' = \begin{cases} 2y_m & m = \text{v} \\ 2y_m + 1 & m = \text{i}, \end{cases} \tag{4.12}$$

and minimizing the cross-entropy loss between $\mathbf{z}_m^r$ and y':

$$\mathcal{L}_{\text{mrid}}(Z_m^r, Y') = \mathbb{E}_{(\mathbf{z}_m^r, y'_m)\sim(Z_m^r, Y')}\mathcal{L}_{\text{ce}}(\mathbf{z}_m^r, y'_m). \tag{4.13}$$

The modality-aware loss function for Z_m^r is :

$$\mathcal{L}_{\text{ymr}} = \mathcal{L}_{\text{mrid}}(Z_m^r, Y') + \mathcal{L}_{\text{cc}}(Z_m^r, Y'), \tag{4.14}$$

where the $\mathcal{L}_{\text{cc}}$ is the center-cluster loss (Wu *et al.*, 2021).

(d) Mixed-Modal Matching Fusion. To balance modality-erased and modality-related features for the matching process at the inference time, we propose a mixed cross-modal triplet loss to

avoid having one component be dominant or useless. The feature fusion loss is:

$$\begin{aligned}\mathcal{L}_{\mathrm{f}} = &\max\{D(\mathbf{z}_m^{f,(j)}, \mathbf{z}_m^{f,(p)}) - D(\mathbf{z}_m^{e,(j)}, \mathbf{z}_{\tilde{m}}^{e,(n)}) + \alpha, 0\} \\ &+ \max\{D(\mathbf{z}_m^{e,(j)}, \mathbf{z}_{\tilde{m}}^{e,(p)}) - D(\mathbf{z}_m^{f,(j)}, \mathbf{z}_m^{f,(n)}) + \alpha, 0\},\end{aligned} \tag{4.15}$$

where $D(.,.)$ is the distance between two embeddings, $\tilde{m} \neq m$, n and p are positive and negative samples to the j instance. $\mathbf{z}_m^f$ is formed by concatenating $\mathbf{z}_m^e$ and $\mathbf{z}_m^r$.

(e) Overall Training. We jointly optimize the network in an end-to-end manner by using the overall loss:

$$\mathcal{L} = \mathcal{L}_{\mathrm{yme}} + \mathcal{L}_{\mathrm{ymr}} + \lambda_{\mathrm{m}}\mathcal{L}_{\mathrm{m}} + \lambda_{\mathrm{o}}\mathcal{L}_{\mathrm{o}} + \lambda_{\mathrm{f}}\mathcal{L}_{\mathrm{f}}, \tag{4.16}$$

and λ_{m}, λ_{o} and λ_{f} are hyperparameters for weighting losses.

(f) Inference.

During inference, MixER performs mixed-modal matching using a single backbone with three small projection heads rather than using three different backbones. For a given image x_m, it extracts $\mathbf{z}_m^r$, $\mathbf{z}_m^e$ and then $\mathbf{z}_f^r$. Matching scores are computed using cosine similarity between feature vectors. Fused features ($\mathbf{z}_f^r$), are used for images within the same modality, while cross-modal matching uses only modality-erased features ($\mathbf{z}_m^e$).

4.4 Results and Discussion

4.4.1 Experimental Methodology

(a) Datasets. Research on cross-modal V-I ReID has mainly used the SYSU-MM01 (Wu *et al.*, 2020), RegDB (Nguyen *et al.*, 2017) datasets, and LLCM (Zhang & Wang, 2023b) datasets. SYSU-MM01 is a large dataset containing more than 22K V and 11K I images of 491 individuals captured with 4 RGB and 2 NIR cameras. RegDB contains 4K co-located V-I images of 412 individuals. Randomly divide the dataset into two sets of the same size for training and testing.

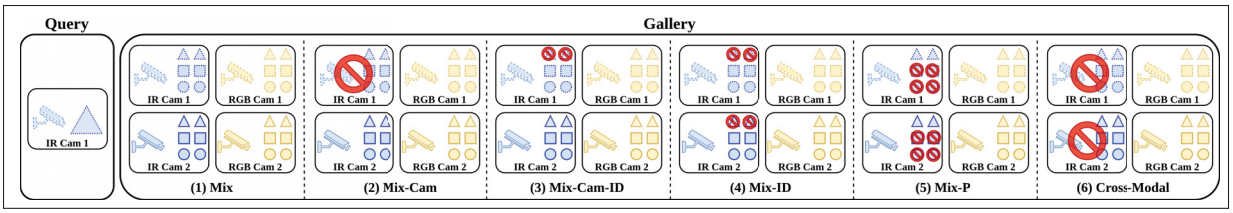


Figure 4.3 Different settings for forming a gallery based on modality, camera, and query identity. The gallery set images are: (a) all images from both modalities; (b) and all images except ones captured with the same camera; (c) same camera and same identity; (d) same identity in the same modality as the query; (e) is mixed settings introduced in (Liu *et al.*, 2025) and (f) all images from another modality

The LLCM dataset consists of a low-light cross-modality dataset comprising 1K identities and is divided into training and testing sets at a 2:1 ratio.

(b) Mixed-Modal Evaluations. Since mixed-modal evaluation has not been performed on V-I datasets, we introduce multiple settings of such assessment based on real-world scenarios where the query and gallery images can be either I or V. Unlike cross-modality settings, where the query and gallery sets strictly contain images of opposite modalities, the mixed-modal setting may contain images from both modalities in the query and gallery sets. We structured the mixed-modality settings into 5 cases as illustrated in Fig.5.2, each one defined by the exclusion of images from the gallery based on the query identity and camera: **(1) Mix**: Images of all individuals and cameras are included in the gallery, regardless of the query camera type. **(2) Mix-Camera**: The gallery excludes images from the query's camera and identity. **(3) Mix-Camera-ID**: The gallery excludes images taken by the same camera as the query and belonging to the query person. **(4) Mix-ID**: Excludes all images of the query person captured by cameras of the same modality as the query. **(5) Mix-P**: Includes the images of the same person ID, seen in the same modality in the query and gallery sets (Liu *et al.*, 2025). We argue that this is a relatively easy and less realistic setting compared to the first 4 in the list, which diminishes its relevance as a research objective.

Since ReID methods have not been evaluated in these settings, we retrained several open-sourced state-of-the-art (SOTA) VI-ReID methods to assess their performance across these configurations. Rank-1 (R1) accuracy and mean average precision (mAP) were measured for each setting in line

with the dataset evaluation criteria. Note that for the RegDB and LLCM datasets, there is only one camera per modality. Therefore, the "Mix-Camera" and "Mix-Camera-Identity" settings are not applicable.

(c) Implementation Details. To extract modality-erased features, the SAAI model(Fang *et al.*, 2023) was used without prototype learning as a baseline with ResNet50 (He *et al.*, 2016b) as the backbone. For modality-related learning modules, `layer4` is cloned from the backbone for each modality. Each image input is resized to 288 by 144, then cropped and erased randomly, and filled with zero padding or mean pixels. We used an ADAM optimizer (Kingma & Ba, 2014) with a linear warm-up strategy for the optimization process. Each training batch contains 8 V and 8 I images from 10 randomly selected identities. The model was trained for 180 epochs, following (Fang *et al.*, 2023), the initial learning rate is set to 0.0004 and decreased by factors of 0.1 and 0.01 at 80 and 120 epochs, respectively.

(d) Benchmark methods. To assess effectiveness across different settings, we benchmarked MixER against several open-source SOTA VI-ReID methods—DDAG (Ye *et al.*, 2020a), MPANet (Wu *et al.*, 2021), DEEN (Zhang & Wang, 2023b), SGEIL (Feng *et al.*, 2023), SAAI (Fang *et al.*, 2023), and IDKL (Ren & Zhang, 2024). Each method was retrained from scratch using the authors' provided parameters, with implementation details available in the supplementary materials.

4.4.2 Comparison with State-of-the-Art Methods

(a) Mixed-Modal Results. Tables 4.1, 4.2a, and 4.2b show the performance of SOTA alongside our MixER across various mixed-modal settings with an I query (V query results are in suppl. materials) on SYSU-MM01(single-shot), RegDB, and LLCM, respectively. MixER(E) and MixER(R) use only $\mathbf{z}_m^e$ and $\mathbf{z}_m^r$ features, respectively, while MixER in a cross-modal setting uses $\mathbf{z}_m^e$ and $\mathbf{z}_m^f$ for the mixed-modal cases. These cross-dataset results highlight the effectiveness of our approach across different VI-ReID methods. In these mixed settings, the difficulty level increases from setting 1 to 4 due to the progressive reduction in positive samples within the

Table 4.1 Performance of SOTA VI-ReID techniques in mixed, cross, and uni-modal settings on the SYSU-MM01 dataset. AIM (Affinity Inference) and KR (k-reciprocal) are re-ranking processes originally applied in the SAAI (Fang *et al.*, 2023) and IDKL (Ren & Zhang, 2024) GitHub projects. Notably, the Mix-P setting (Liu *et al.*, 2025) yields highly accurate results, suggesting its lower difficulty and limited relevance as a realistic evaluation

Method	Venue	Mixed-Modal										Cross-Modal				Uni-Modal			
		Mix		**Mix-Cam**		**Mix-Cam-ID**		**Mix-ID**		Mix-P(Liu *et al.*, 2025)		**All**		**Indoor**		**I→I**		**V→V**	
		R1	**mAP**	**R1**	**mAP**	**R1**	**mAP**	**R1**	**mAP**	**R1**	**mAP**	**R1**	**mAP**	**R1**	**mAP**	**R1**	**mAP**	**R1**	**mAP**
DDAG(Ye *et al.*, 2020a)	ECCV20	91.91	57.60	80.81	53.54	70.41	45.14	23.92	29.09	98.41	45.87	53.29	50.61	61.02	67.98	80.15	86.80	97.89	91.90
MPANet (Wu *et al.*, 2021)	CVPR21	95.46	72.82	89.45	70.34	81.95	63.81	49.06	51.08	99.16	72.87	70.58	68.24	76.54	80.95	88.53	92.85	98.31	94.08
DEEN(Zhang & Wang, 2023b) + Flip	CVPR23	-	-	-	-	-	-	-	-	-	-	74.7	71.8	80.3	83.3	-	-	-	-
DEEN(Zhang & Wang, 2023b)	CVPR23	94.07	74.06	87.50	72.09	79.50	65.66	56.77	55.30	98.51	74.06	72.55	68.59	84.21	82.14	86.14	90.98	96.83	89.93
SGEIL(Feng *et al.*, 2023)	CVPR23	94.82	71.03	89.22	70.16	79.33	61.34	46.60	48.37	99.76	70.13	73.34	67.44	84.09	81.65	86.52	91.36	97.29	90.24
SAAI(Fang *et al.*, 2023) +AIM	ICCV23	-	-	-	-	-	-	-	-	-	-	75.90	77.03	83.20	88.01	-	-	-	-
SAAI(Fang *et al.*, 2023)	ICCV23	96.01	74.59	90.63	72.51	84.30	65.94	52.49	53.30	99.17	72.55	73.87	69.71	84.19	82.59	89.29	93.06	98.24	93.46
IDKL(Ren & Zhang, 2024) + KR	CVPR24	-	-	-	-	-	-	-	-	-	-	81.42	79.85	87.14	89.37	-	-	-	-
IDKL(Ren & Zhang, 2024)	CVPR24	95.27	72.81	90.15	71.87	80.73	63.54	47.34	50.86	99.53	72.79	72.05	69.67	84.54	83.76	89.55	93.18	98.55	95.13
MixER(R)		96.22	38.94	88.32	25.37	82.58	23.65	0.87	0.99	99.99	38.94	1.19	3.61	1.94	5.2	90.02	93.93	98.63	94.81
MixER(E)		96.46	75.21	90.24	72.89	84.55	66.02	52.89	54.86	98.24	76.22	75.04	71.22	85.82	84.06	90.18	93.65	97.95	93.38
MixER	Ours	**96.63**	**79.64**	**91.77**	**76.35**	**87.56**	**72.70**	**65.14**	**62.76**	98.89	80.39	75.04	71.22	85.82	84.06	**93.06**	**95.82**	**99.14**	**95.30**
MixER +AIM		-	-	-	-	-	-	-	-	-	-	76.12	75.45	85.30	85.54	-	-	-	-
MixER +KR		-	-	-	-	-	-	-	-	-	-	**85.96**	**83.54**	**92.48**	**91.49**	-	-	-	-

same modality as the query, challenging each method's robustness in mixed scenarios. Notably, this analysis also enables us to examine each method's strengths and weaknesses in novel contexts; for instance, SGEIL (Feng *et al.*, 2023) performs best in the "MIX" but exhibits a significant performance drop in the "Mix-ID" on the SYSU-MM01 dataset compared to other methods. In contrast to Mix-P (Liu *et al.*, 2025), where the R1 lacks discriminative power for comparing different VI-ReID approaches due to their uniformly high performance, our experimental settings offer a more comprehensive and realistic framework for evaluating VI-ReID methods. The proposed MixER consistently outperforms existing methods across nearly all mixed settings, highlighting its capacity to bridge the gap between V and I images by learning modality-erased features while enhancing same-modality matching through modality-related information. Importantly, modality-related features (MixER(R)) perform less effectively in Mix-ID and Cross-modal settings, where the same person's modality does not appear in the gallery. However, they help correct erroneous intra-modal matches by adjusting the similarity scores produced by modality-erased features. This supports our approach: modality-related information should be excluded from modality-erased features for robust inter-modal matching.

To further evaluate the adaptability of our learning paradigm, we applied it to several SOTA VI-ReID methods as a baseline on the SYSU dataset. We tested it in two variations: (a) stopping

gradient backpropagation of our proposed modality-related and erased learning losses and (b) enabling end-to-end training with combined gradients from the existing and our proposed losses to strengthen cross-modal matching by filtering out modality-related features from the backbone.

Table 4.2 Performance of SOTA VI-ReID techniques in mixed, cross, and uni-modal settings on the (a) RegDB and (b) LLCM datasets

(a) RegDB dataset

Method	Mixed-Modal				Cross-Modal			
	Mix		Mix-ID		I→V		V→I	
	R1	mAP	R1	mAP	R1	mAP	R1	mAP
DDAG (Ye *et al.*, 2020a)	99.9	76.17	45.29	45.54	69.34	63.46	68.06	61.80
MPANet (Wu *et al.*, 2021)	**100**	76.46	42.18	44.66	84.27	80.20	83.20	79.82
DEEN (Zhang & Wang, 2023b)	99.95	83.06	66.21	59.61	90.29	83.98	91.21	85.13
SAAI (Fang *et al.*, 2023)	**100**	82.29	58.01	55.41	86.21	80.0	86.60	81.51
IDKL (Ren & Zhang, 2024)	**100**	83.83	61.46	60.54	87.23	83.20	87.91	85.07
SAAI +AIM	-	-	-	-	92.09	92.01	91.07	91.45
IDKL + KR	-	-	-	-	94.22	90.43	94.72	90.19
MixER	**99.9**	**89.44**	**79.95**	**73.45**	90.49	85.63	90.53	86.42
MixER + AIM	-	-	-	-	90.78	90.18	90.35	90.02
MixER + KR	-	-	-	-	**97.09**	**93.01**	**96.55**	**93.55**

(b) LLCM dataset

Method	Mixed-Modal				Cross-Modal			
	Mix		Mix-ID		I→V		V→I	
	R1	mAP	R1	mAP	R1	mAP	R1	mAP
DDAG (Ye *et al.*, 2020a)	69.34	63.46	68.06	61.80	40.14	26.88	45.17	29.94
MPANet (Wu *et al.*, 2021)	96.76	53.73	38.79	29.03	46.48	30.96	52.15	37.00
DEEN (Zhang & Wang, 2023b)	97.63	**64.69**	49.11	38.38	69.49	54.75	73.95	58.75
SAAI (Fang *et al.*, 2023)	96.61	60.05	40.86	31.27	59.37	45.65	64.37	48.60
IDKL (Ren & Zhang, 2024)	97.73	62.46	38.88	33.85	62.53	49.33	70.36	55.04
IDKL + KR	-	-	-	-	70.72	65.19	72.22	66.43
MixER	**97.80**	64.45	**57.1**	**45.71**	65.76	51.08	70.79	56.61
MixER+ AIM	-	-	-	-	66.11	62.89	74.10	**68.20**
MixER + KR	-	-	-	-	**73.72**	**65.36**	**76.14**	65.46

Table 4.3 Accuracy of the proposed method and SOTA methods as the baseline on the SYSU-MM01 (single-shot setting) as I query. "(f)" indicates that our proposed losses were not back-propagated through the baseline models

Method	Cross-modal		Mix		Mix-Cam		Mix-Cam-ID		Mix-ID	
	R1	mAP	R1	mAP	R1	mAP	R1	mAP	R1	mAP
DDAG(Ye *et al.*, 2020a)	53.29	50.61	91.91	57.60	80.81	53.54	70.41	45.14	23.92	29.09
DDAG(f)+MixER	53.21	50.86	92.04	60.29	81.27	55.07	72.05	48.68	35.71	34.81
DDAG+MixER	54.61	51.27	92.52	62.52	81.74	57.16	72.96	51.93	38.79	38.79
MPANet(Wu *et al.*, 2021)	69.34	66.42	95.46	72.82	89.45	70.34	81.95	63.81	49.06	51.08
MPANet(f)+MixER	69.66	66.51	95.02	77.01	90.43	72.63	83.98	69.80	63.96	60.75
MPANet+MixER	72.18	68.18	95.80	78.67	91.34	74.98	83.95	70.55	66.00	61.16
DEEN(Zhang & Wang, 2023b)	72.55	68.59	94.07	74.06	87.50	72.09	79.50	65.66	56.77	55.30
DEEN(f)+MixER	72.63	68.49	94.55	78.91	88.38	75.27	82.77	70.9	65.8	60.05
DEEN+MixER	74.19	69.00	95.00	79.00	89.00	76.00	83.00	71.00	66.00	61.00
SGEIL(Feng *et al.*, 2023)	73.34	67.44	94.82	71.03	89.22	70.16	79.33	61.34	46.60	48.37
SGEIL(f)+MixER	72.17	67.81	94.90	76.45	90.14	73.43	82.63	68.86	64.14	60.15
SGEIL+MixER	74.08	69.19	95.33	78.01	91.16	74.29	83.97	69.40	65.37	61.50
SAAI(Fang *et al.*, 2023)	72.19	69.71	96.01	74.59	90.63	72.51	84.30	65.94	52.49	53.30
SAAI(f)+MixER	73.64	69.51	96.21	79.12	91.7	75.03	87.0	71.22	63.72	60.91
SAAI+MixER	75.37	71.92	97.27	80.47	92.66	76.81	88.51	73.24	66.23	62.45
IDKL(Ren & Zhang, 2024)	72.05	69.67	95.27	72.81	90.15	71.87	80.73	63.54	47.34	50.86
IDKL(f)+MixER	72.65	70.2	95.58	74.24	90.51	71.78	83.77	65.41	52.63	55.57
IDKL+MixER	73.44	71.02	96.10	76.67	91.38	74.27	85.61	68.44	58.37	57.91

Results in Table 4.3 indicate that modality-related learning significantly boosts performance in mixed-modal settings. In contrast, end-to-end modality-erased learning enhances both mixed and cross-modal matching. For example, our modality-related learning improves the mAP of SAAI(Fang *et al.*, 2023) and IDKL (Ren & Zhang, 2024) in "Mix-ID" settings by more than 6% and 4%, respectively.

(b) Cross-Modal Results. To show that erasing modality-specific information from features makes them more proper for cross-modal matching, we measure the performance of MixER and compare it with SOTA VI-ReID in Tables 4.1, 4.2a, and 4.2b. Our experiments show that

Table 4.4 Performance of techniques of V-I ReID on Cross-Dataset RGB Market1501 dataset. Columns indicate the training dataset, and rows show the model's performance on the Market1501

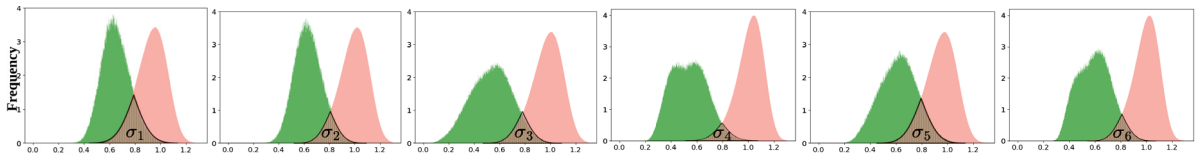
Source:	**SYSU-MM01**		**RegDB**		**LLCM**	
Target:	**Market1501(V→V)**					
	R1	**mAP**	**R1**	**mAP**	**R1**	**mAP**
DDAG(Ye *et al.*, 2020a)	82.39	55.57	11.99	3.32	53.40	20.18
MPANet(Wu *et al.*, 2021)	78.97	51.61	**18.17**	4.79	56.91	22.79
DEEN (Zhang & Wang, 2023b)	66.50	33.81	6.85	1.7	58.81	23.72
SGEIL (Feng *et al.*, 2023)	79.18	48.72	-	-	-	-
SAAI (Fang *et al.*, 2023)	84.32	57.62	14.54	3.44	55.68	22.77
IDKL(Ren & Zhang, 2024)	76.66	49.59	12.84	2.79	54.61	21.95
MixER(E)	82.90	56.56	14.31	3.97	54.45	22.52
MixER(R)	83.46	56.85	16.29	4.08	59.56	25.93
MixER(E+R)	**87.93**	**62.41**	17.7	**5.02**	**61.37**	**27.88**
Upper-bound	95.1	87.8	95.1	87.8	95.1	87.8

Table 4.5 Impact of losses on MixER performance

Settings					**Cross-Modal**		**Mix-Cam-ID**	
$\mathcal{L}_{\text{yme}}$	$\mathcal{L}_{\text{ymr}}$	$\mathcal{L}_{\text{o}}$	$\mathcal{L}_{\text{m}}$	$\mathcal{L}_{\text{f}}$	**R1**	**mAP**	**R1**	**mAP**
✓					69.74	66.38	80.57	62.51
✓	✓				69.28	66.09	83.61	64.57
✓		✓			70.25	67.11	81.07	62.99
✓			✓		71.72	67.60	78.55	60.38
✓	✓	✓	✓		73.40	70.87	84.64	65.83
✓	✓	✓	✓	✓	**73.43**	**70.92**	**87.56**	**72.70**

MixER outperforms these methods in various situations. Modality-specific submodules are not triggered during inference in cross-modal matching. For example, compared to the second-best approach for the "All Search" scenario, MixER outperforms by a margin of 2.3% R1 and 3.3% mAP without adding complexity to the model. For the "I→V" mode on RegDB, MixER achieves 91.4% R1 accuracy and 85.5% mAP. For the "V→I" mode, our method also obtains 90.4% R1 accuracy and 86.8% mAP. The results validate the effectiveness of our proposed MixER and show that it can effectively reduce the discrepancy between the V and I modalities.

In addition, our modality-erased feature learning has the advantage that it can be used in different VI-ReID models to improve their performance in cross-modal settings without incurring overhead during testing. To show this adaptability, in Table 4.3, in the first column, end-to-end training improves the performance of all methods in the SYSU-MM01 dataset. For example, MixER increases the mAP of SGEIL(Feng *et al.*, 2023) and SAAI(Fang *et al.*, 2023) by 1.75% and 1.37%.



a) Baseline (Cross) b) MixER (Cross) c) Baseline (Mix) d) MixER (Mix) e) Baseline (Mix-Cam) f) MixER (Mix-Cam)

Figure 4.4 The intra-class (green) and inter-class (pink) distances distribution of features in different gallery settings. Here, the SAAI method (Fang *et al.*, 2023) is considered as baseline

(c) Uni-Modal Results. To illustrate the effectiveness of learning modality-related alongside modality-erased in MixER for uni-modal matching, our trained model was tested on the SYSU-MM01 dataset in V→V and I→I settings, as well as on an unseen V dataset. Table 4.1 (column "Uni-Modal") presents these results for SYSU-MM01, demonstrating that our MixER approach outperforms other methods since it uses the complementary information of discriminative modality-related and modality-erased features. Table 4.4 reports the performance of various methods, while each column indicates the V-I training dataset, and each row shows the model's

performance on the Market1501 (Zheng *et al.*, 2015) dataset. Our MixER approach achieves the highest R1 and mAP scores, with values of 87.9% and 62.4%, respectively. When using only modality-related features (activating only the V branch) or only modality-erased features, the mAP scores are 56.8% and 56.5%, respectively. However, the integration of both feature types increases the mAP and R1 scores by more than 3%, demonstrating that this fusion improves generalization and provides a complementary representation of input images.

4.4.3 Ablation Studies

Losses. Table 4.5 shows the impact of each loss component on performance. Using only the $\mathcal{L}_{\text{y}}$ as a baseline achieves results of 69.7% R1 and 66.3% mAP in cross-modal matching. Adding $\mathcal{L}_{\text{ymr}}$ improves Mix-Cam-ID R1 to 83.6%, indicating $\mathcal{L}_{\text{ymr}}$'s benefit for intra-modality matching consistency, though cross-modal performance slightly decreases. Incorporating only $\mathcal{L}_{\text{o}}$ improves both cross- and mixed-modal performance, whereas $\mathcal{L}_{\text{m}}$ provides a more significant enhancement in Cross-Modal metrics. The combined effect of $\mathcal{L}_{\text{y}}$, $\mathcal{L}_{\text{yme}}$, $\mathcal{L}_{\text{o}}$, and $\mathcal{L}_{\text{m}}$ achieves 73.4% R1 in cross-modal, showing these losses' synergy. Finally, adding $\mathcal{L}_{\text{f}}$ achieves optimal results across all metrics, highlighting its important role in refining features for mixed-modal settings.

4.4.4 Qualitative Analysis

Feature distribution. To examine the effectiveness of our method in cross-modal and mixed-modal matching, we visualize inter-class and intra-class distance distributions on the SYSU dataset, as shown in Fig.4.4(a-f). Compared to the baseline, MixER reduces both the mean and variance of intra-class distances across all gallery settings, decreasing the area of overlap ($\sigma_2 < \sigma_1$, $\sigma_4 < \sigma_3$, and $\sigma_6 < \sigma_5$). This demonstrates that MixER more effectively reduces intra-class distances, thereby decreasing modality discrepancy in modality-erased features and enhancing identity discrimination in modality-related features. Furthermore, UMAP visualizations (see supplementary material) show that MixER better separates identities and reduces modality discrepancy within the mixed-modal gallery.

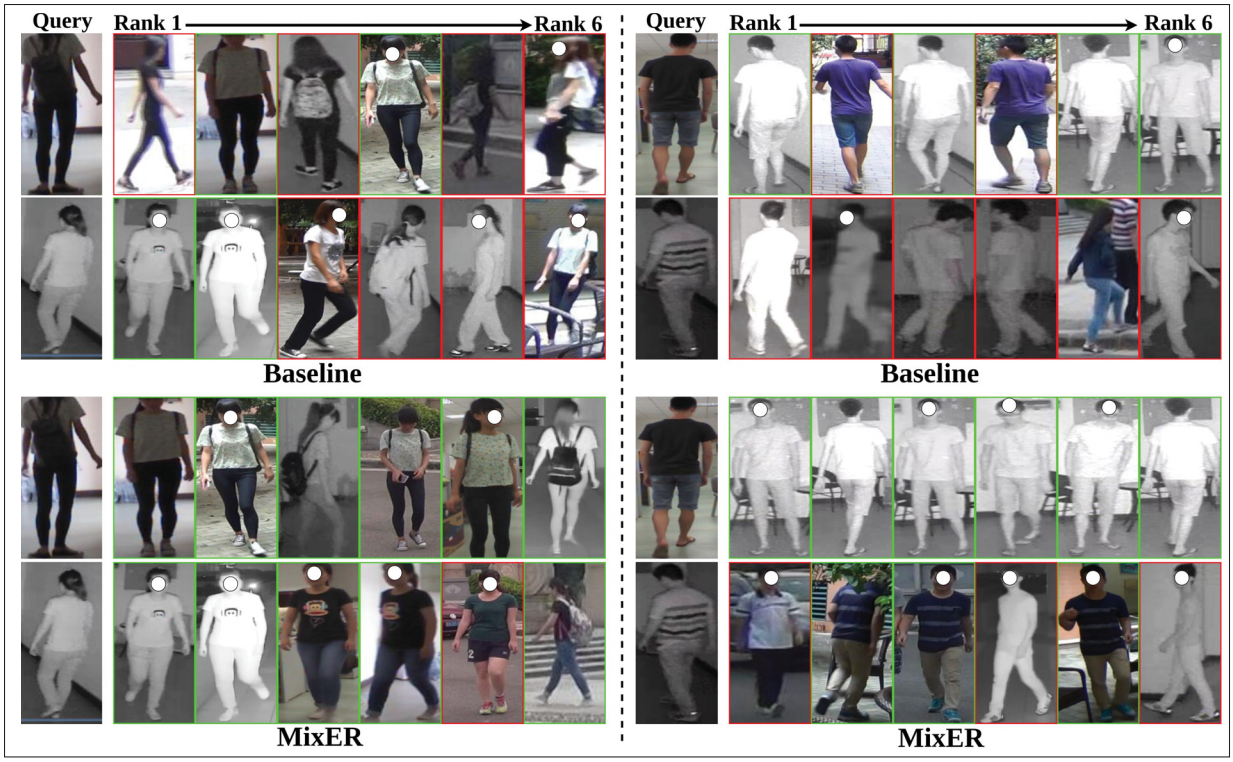


Figure 4.5 Rank-6 retrieval results obtained by the baseline (top) and the proposed MixER (bottom) on SYSU dataset under Mix-Cam-ID (left) and Mix-ID (right) settings

Retrieval results. To show the effectiveness of MixER, we present retrieval results on the SYSU-MM01 dataset in Fig. 4.5. Green boxes indicate correct matches, while red boxes represent incorrect ones. Overall, combining modality-erased and -related features at the bottom significantly improves the results, with more correct matches appearing in higher ranked positions compared to the baseline. Modality-erased learning aims to find the closest matches based solely on shared attributes between V and I images. However, it may mistakenly select incorrect images. Modality-related features refine this by penalizing incorrect matches based on attributes specific to the query modality. For example, in the second row at the top, the model ranks images of a person wearing a T-shirt and shorts, albeit in different colors, as the top matches. However, in the second row at the bottom, the model produces better matches by leveraging modality-erased and -related information.

4.5 Conclusion

In this paper, we propose Mix-Modal ReID, a new evaluation setting that better reflects real-world person re-identification challenges in mixed visible-infrared galleries. To enhance robustness across modalities, we introduce a modality-erased and modality-related feature learning approach. Experiments on SYSU-MM01, RegDB, and LLCM show that our method outperforms SOTA VI-ReID approaches in both mixed- and cross-modal settings, revealing the limitations of existing methods. Our approach provides a strong foundation for more effective and adaptable VI-ReID systems in real-world applications.

CHAPTER 5

BEYOND PATCHES: MINING INTERPRETABLE PART-PROTOTYPES FOR EXPLAINABLE AI

Mahdi Alehdaghi[1] , Rajarshi Bhattacharya[1] , Pourya Shamsolmoali[2] , Rafael M. O. Cruz[1] , Eric Granger[1]

[1] LIVIA, ILLS, Dept. of Systems Engineering, École de Technologie Supérieure, 1100 Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3
[2] University of York, U.K.

The article was submitted to The 40th Annual AAAI Conference on Artificial Intelligence, August 2025

5.1 Introduction

Deep learning (DL) has revolutionized computer vision, enabling models to achieve remarkable performance across a range of tasks, from object detection to image classification. Despite these successes, the opaque nature of DL models has raised concerns regarding their interpretability and trustworthiness. In high-stakes domains such as healthcare (Fellous, Sapiro, Rossi, Mayberg & Ferrante, 2019), autonomous driving (Kim & Joe, 2022), and security (Chen *et al.*, 2019; Nauta, Schlötterer, Van Keulen & Seifert, 2023), understanding why a model makes a certain decision is as critical as the decision itself. Interpretability not only fosters trust but also aids in identifying potential biases, debugging models, and ensuring compliance with ethical and legal standards. Consequently, there is a growing demand for deep vision models that are not only accurate but also capable of explaining their predictions in a clear and comprehensible manner (Saranya & Subhashini, 2023; Vilone & Longo, 2020).

Common post-hoc explainability techniques such as GradCAM (Selvaraju *et al.*, 2017b) or ScoreCAM(Wang *et al.*, 2020b) try to interpret the model's decision by reverse-engineering and finding the regions of the input image that contribute most to the output after the training process, as shown in Fig. 5.2a. In contrast to such black-box reverse-engineering methods, ante-hoc techniques (Nauta *et al.*, 2023; Chen *et al.*, 2019; Rymarczyk, Struski, Tabor & Zieliński, 2021; Nauta, Van Bree & Seifert, 2021; Zhu, Jin, Zhang & Chen, 2024) make the backbone

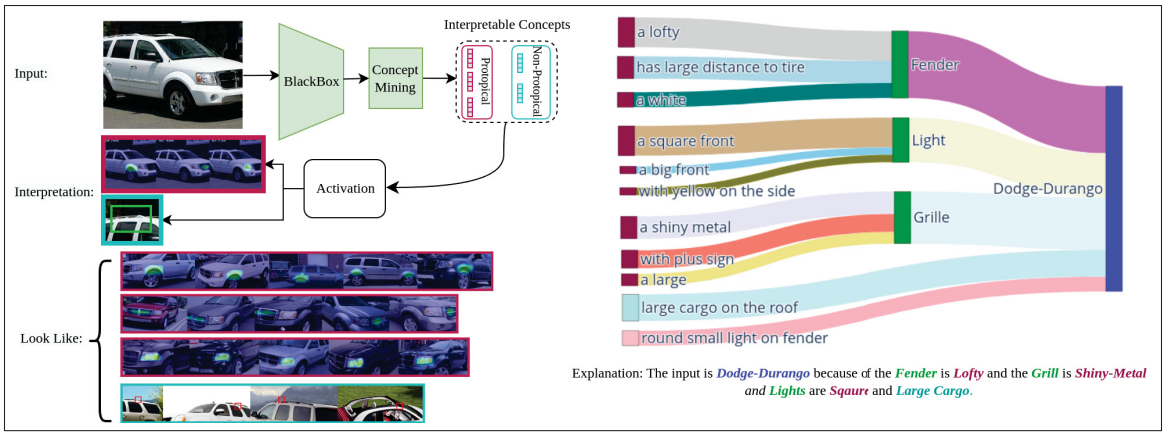
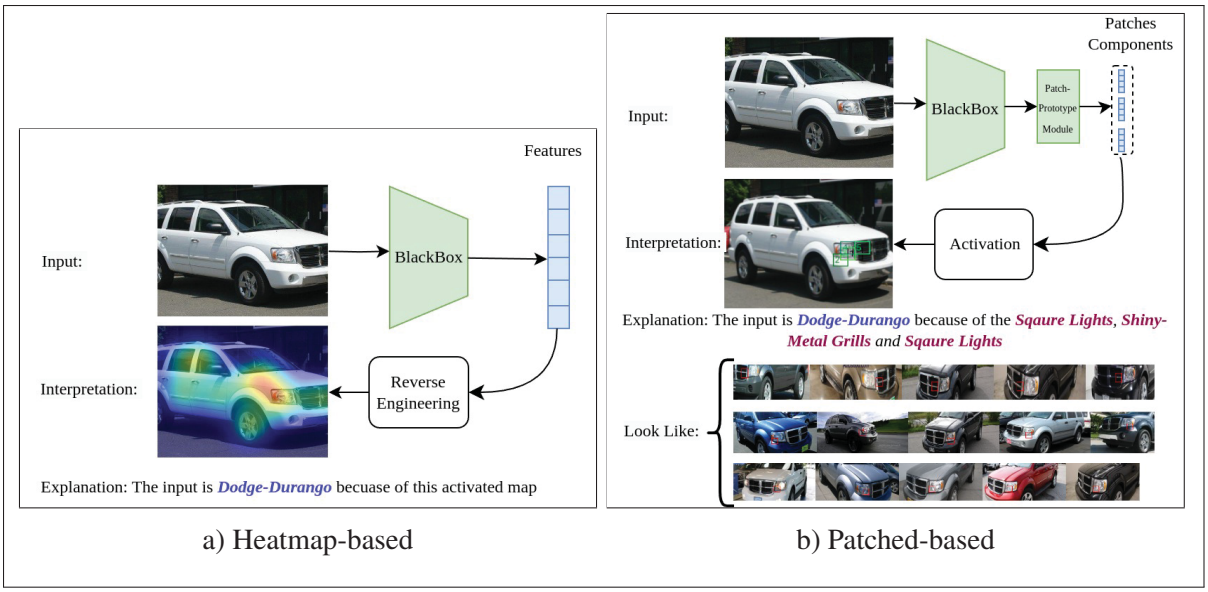


Figure 5.1 PCMNet mines interpretable prototypical and non-prototypical concepts to explain the model's decision

produce interpretable components to explain its decision. This recognition-by-components theory (Biederman, 1987) describes how humans recognize objects by segmenting them into multiple meaningful concepts. Some ante-hoc methods such as ProtoPNet (Chen *et al.*, 2019) and PIPNet (Nauta *et al.*, 2023), as shown in Fig. 5.2b, segment the spatial deep features extracted by the backbone (before the final pooling layer) into fixed, small patch components and learn a set of prototypes from them. These prototypes, activated by the patches, provide an interpretation of the model's decision and allow users to identify and compare similar prototypes across different input images. However, these methods have significant limitations. Fixed patch sizes can lead to instability in prototype activation for larger areas, which may fail to focus on primitive semantics. Smaller patches, similarly, may also lose their interpretive capacity due to the limited area they represent. Furthermore, many discriminative and explainable attributes require larger regions to capture their semantic meaning effectively to avoid duplicating the same concept in explanation examples (two similar rows at the bottom of Fig.5.2b). In other words, the regions containing the prototypes can be dynamic and require training.

Direct application of naïve unsupervised part-discovery or Slot Attention-based methods often falls short in capturing primitive, semantically meaningful concepts from images and fails to provide intuitive explanations of model decisions based on extracted prototypes (van der Klis *et al.*, 2023; Locatello *et al.*, 2020). This limitation arises because part features are



a) Heatmap-based b) Patched-based

Figure 5.2 Explaining the decision through examples. (a) Heatmap-based methods see the feature backbones as black-box models, and try to locate regions activated by the most important features. (b) With patch-based methods, the features are decomposed into components extracted by image patches, and an explanation is provided by these components

typically employed merely to identify similar parts across objects, without enforcing conceptual consistency, while Slot Attention features tend to lack explicit prototypical representations required for interpretable reasoning. The result, therefore, is a reduction in the diversity and a lack of interpretability of the prototypes, hindering the model's ability to provide meaningful and comprehensive explanations for its decisions.

To address these limitations, this paper introduces an intrinsically interpretable image classifier that mines meaningful and interpretable prototypical concepts from dynamic regions in images to explain the model's decisions. Our unsupervised concept mining network identifies semantically meaningful components, relying only on class labels without requiring additional part labels or annotations. Our part-prototype concept mining network (PCMNet) detects larger, coherent regions of pixels as meaningful prototype parts. This design allows the components to identify relationships between semantically similar prototypes across instances from different classes. Then, we leverage these prototypes to mine a set of primitive yet discriminative concepts used

to explain the model's reasoning directly. For instance, Figure 5.2 demonstrates an use case where the prototype heatmaps, derived from our pipeline, shows that the decision-making was based on the "lofty fenders", "shiny metal grille" and "square shaped lights", to conclude that the vehicle in the image, is a Dodge Durango.

In the first stage, PCMNet employs an unsupervised part discovery mechanism to learn a set of prototypes using a novel center-clustering loss. In the second stage, the model encodes the deep features into a sparse Concept Activation Vector (CAV) (Kim *et al.*, 2018), which provides a spatially and semantically interpretable explanation of the final decision. Each concept activation is computed as the cosine similarity between part features and class-aware prototype centroids. To ensure that these activated concepts are both primitive and class-discriminative, PCMNet applies a clustering algorithm to the part features corresponding to each class-specific prototype. The center of each resulting cluster then serves as a class-aware prototype centroid. Interpretation of the resulting CAV is straightforward, as it is produced by a sparse, non-negative linear layer that guarantees all concept values remain non-negative.

The contributions of this paper are summarized as follows. **(1)** PCMNet is presented as a part-prototypical concept mining framework for explaining vision classifiers with primitive, prototypical, and interpretable human-meaningful components. **(2)** Concepts extracted by PCMNet are leveraged from a broader region than single patches to contain more semantical part-prototype abstraction, similar to the human brain, by using a novel contrastive prototype extraction. **(3)** Two-level clustering mechanisms are proposed for finding interpretable parts at the pixel level and primitive and intuitive concepts at the feature level. **(4)** Our experiments show that PCMNet explanation is more robust to occlusions, outperforming state-of-the-art patched-based prototype methods according to classification and explainability metrics.

5.2 Related Work

(a) Post-hoc or heatmap-based explanation: Heatmap-based explainability methods are a widely used family of post-hoc techniques that visualize which parts of an input image contribute most to a model's decision. These techniques, often referred to as attribution methods, assign importance scores to different image regions to highlight influential features. Gradient-based methods, such as GradCAM (Selvaraju *et al.*, 2017a) and ScoreCAM (Wang *et al.*, 2020b), generate heatmaps by backpropagating gradients concerning input features, revealing class-specific activation regions. While GradCAM produces class-dependent heatmaps, other methods like FullGrad (Srinivas & Fleuret, 2019) are class-agnostic, providing a more general view of model behavior across different outputs. Despite their popularity, gradient-based methods suffer from high sensitivity to noise, which can lead to unreliable and inconsistent heatmaps. To mitigate this, gradient-free CAM techniques, such as ScoreCAM, have been proposed to generate more stable and interpretable explanation maps.

Beyond gradient-based methods, attribution propagation approaches offer an alternative way to decompose model predictions into layer-wise relevance scores. Techniques such as Layer-wise Relevance Propagation (LRP) (Bach *et al.*, 2015) recursively distribute relevance through the network, providing a structured breakdown of model decisions. While initially designed for convolutional neural networks (CNNs), recent adaptations have extended these methods to vision transformers (ViTs) (Chefer, Gur & Wolf, 2021), leveraging their self-attention mechanisms for improved interpretability. However, despite their effectiveness in highlighting important image regions, both gradient-based and attribution propagation methods remain fundamentally limited in providing high-level, human-interpretable concepts. Unlike our proposed PCMNet, which explains decisions through part-based prototypes, heatmap-based methods do not inherently capture semantic relationships between features, making them less suitable for interpretable concept-based explanations.

Additionally, since these methods are inherently post-hoc, they do not enable real-time user intervention during test-time inference. This means that once a model makes a prediction, users

cannot modify or refine the explanations interactively, limiting their practical use in applications requiring human-in-the-loop decision-making. In contrast, PCMNet introduces an ante-hoc, concept-driven interpretability approach that allows for more direct, structured, and interactive explanations by providing semantically meaningful components rather than relying solely on heatmap visualizations.

(b) Slot attention: Slot Attention has emerged as a powerful mechanism for learning object-centric representations by dynamically assigning feature regions to a fixed number of latent slots (Locatello *et al.*, 2020). Unlike traditional attention mechanisms that distribute importance weights across all features, Slot Attention iteratively refines a small set of slots, each representing a distinct object or meaningful component of an image. This allows models to decompose scenes into structured representations without explicit supervision (Kipf *et al.*, 2022). By leveraging iterative updates with attention-based routing, Slot Attention enables more interpretable feature learning, where each slot encapsulates a semantically meaningful part of the input. This property makes it particularly useful for explainable AI, as it provides a structured decomposition of an image, rather than relying on uninterpretable distributed feature maps (Greff, Van Steenkiste & Schmidhuber, 2020).

Recent works have applied Slot Attention in various domains, such as object discovery, segmentation, and interpretable classification (Singh *et al.*, 2022; Monet, Greff, Van Steenkiste, Locatello & Bachem, 2021). In the context of explainable deep learning, it has been used to learn disentangled representations where each slot corresponds to a distinct semantic concept (Karazija, Dorr, Perrot, Staniute & Ritchie, 2021). However, these methods typically focus on object-level decomposition rather than part-based representation. Our PCMNet builds on this idea by using Slot Attention-inspired mechanisms to discover meaningful part prototypes rather than full objects, ensuring better granularity in concept-based explanations. By applying slot-like iterative attention within a contrastive prototype extraction framework, PCMNet refines part-based representations, making them more interpretable and aligned with human perception and reasoning.

(c) Concept-based explanation: Contrary to traditional local feature attribution methods that assess the importance of input-level features, concept-based XAI methods (Dreyer, Achtibat, Samek & Lapuschkin, 2024; Chen *et al.*, 2019; Rymarczyk *et al.*, 2021; Bach *et al.*, 2015) analyze the role of latent representations in specific layers of a deep neural network (DNN). These methods explore concepts through various mechanisms, including individual neurons, directional representations such as Concept Activation Vectors (CAVs) (Kim *et al.*, 2018), or broader feature subspaces. Early XAI research primarily focused on understanding how these concepts contribute to global decision-making, identifying the most relevant concepts for a given output class (Dreyer *et al.*, 2024; Bach *et al.*, 2015). Some approaches enhance model interpretability by employing local feature attribution techniques to localize and quantify the importance of concepts for individual predictions, thereby enabling concept-based explanations at the instance level (Nauta *et al.*, 2023, 2021). While these methods provide faithful explanations aligned with the model's internal reasoning, they are vulnerable to performance degradation under occlusion, as their extracted concepts often rely on limited and highly localized discriminative regions of the input. PCMNet addresses this limitation by part-based prototypical concept reasoning to rely on other discriminative parts that are not occluded.

Additionally, although instance-wise concept-based explanations provide deeper insights, they can also be overwhelming and difficult for stakeholders to process, as hundreds of concepts may need to be analyzed for each instance. To address this, some works highlight the benefits of visualizing local decisions within a global embedding space (Dreyer *et al.*, 2024; Donnelly, Barnett & Chen, 2022). With PCMNet, we build upon this idea by introducing part prototypes that encapsulate local concepts. This approach reduces the cognitive load for human interpretation by enabling comparisons between individual (local) prediction and prototypical concepts, making explanations more intuitive and structured.

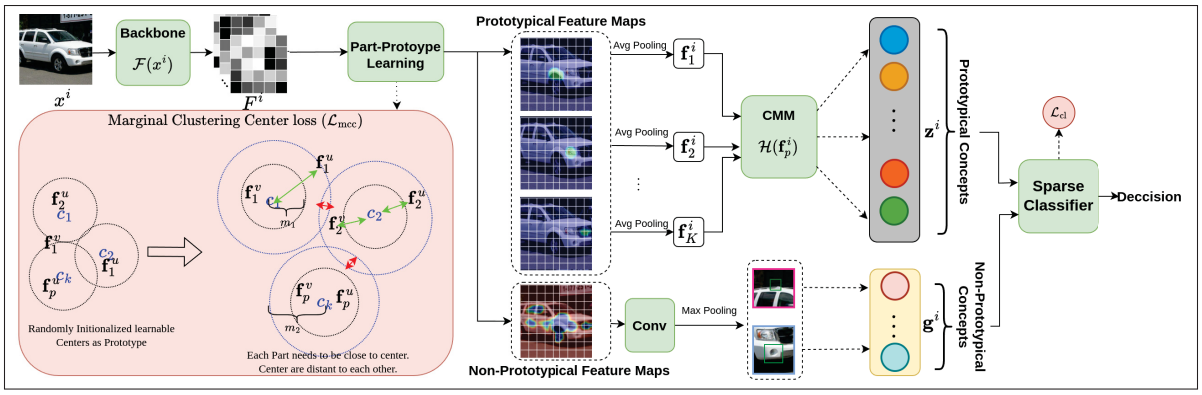


Figure 5.3 Overall architecture of PCMNet. In the first stage, part prototypes are learned from spatial features extracted by the backbone. In the second stage, prototypes are clustered within each class to identify frequently occurring concepts. The center of these clusters is computed as concept prototypes, and the distance between part prototypes and concept clusters determines the activated concepts for the model's decision-making. This approach enables interpretable and concept-driven classification

5.3 Proposed Prototypical Concept Mining Network

To explain the decision for a classification problem with a training set $\mathcal{T}$ containing N images $\{(x^1, y^1), ..., (x^N, y^N\} \in \mathcal{X} \times \mathcal{Y}$, with L classes, the PCMNet extracts primitive and sparse concepts that are activated and discriminate the object. These concepts are described in a limited set of prototypes that existed for almost all images in the datasets (e.g., lights, fenders, and grilles for cars or heads and wings for birds). In other words, the output of our model is the activated concept vector $z_i = [z_p^i]_{p=1}^K$, in which p is the prototype index. Fig. 5.3 shows the overall overview of our approach.

PCMNet transforms features extracted from the backbone model into explainable and interpretable concepts in three steps. **Step 1**: The model finds semantically independent, meaningful regions from input images in an unsupervised way by minimizing our novel Marginal Cluster Center loss to learn the prototypes that describe these parts. Then, the part features are disentangled to discriminate the class label. **Step 2**: Once the part features are learned for the training data, for each part inside all instances of each class, a set of shared elementary characteristics is found by applying a clustering approach to these features. Then, the centroid list, which represents the

prototypical concepts, is generated by calculating the center of each cluster. **Step 3**: Extract the Concept Activation Vector (CAV) for the generated centroid list by computing the similarity of extracted features and each centroid. The CAVs are used to discriminate the class of the input and explain the decision. Also, to avoid losing some discriminative concepts that are not prototypical (i.e., specific and unique for this class and not available for others), we combine the features that are located outside of prototypical maps into the CAVs for classification.

5.3.1 Step 1: Part-Prototype Discovery

Our model is shown in Fig. 5.3. The input images are given to a backbone to extract the 3D deep features as $F^i \in \mathbb{R}^{D\times w\times h}$ where D is the channel size and w, h denotes the width and height. Based on extracted deep features F^i, our model finds semantically independent meaningful regions that can be shared in different inputs. We name them "part-prototype". For determining regions, the model scores each pixel in F^i to specify their belongingness to each part-prototype class. The mapping score for each pixel in the location (q) is represented by M^i, we have $\sum_{p=1}^{K+1} M_p^i(q) = 1$. The spatial feature for each part, F_p^i, is computed by element-wise multiplication of M_p^i to F^i. The part features, $\mathbf{f}_p^i$ is computed as:

$$\mathbf{f}_p^i = \mathcal{GP}(\mathsf{conv}(F_p^i)) \in \mathbb{R}^{d_f}, \tag{5.1}$$

where $\mathcal{GP}$ is the global average pooling and d_f is the dimension of features. The last index (K+1) is used to cover the background or non-part regions of the foreground (Alehdaghi, Shamsolmoali, Cruz & Granger, 2025). We name these features non-prototypical concepts that are computed as:

$$\mathbf{g}^i = \mathcal{MP}(\mathsf{conv}(F_{K+1}^i)) \in \mathbb{R}^{d_f}, \tag{5.2}$$

where $\mathcal{MP}$ is the max-pooling. We use the PDiscoNet (van der Klis *et al.*, 2023) as the baseline for extracting parts features. Our part loss comprises their final loss ($\mathcal{L}_{\text{bl}}$) without their orthogonal component and our marginal cluster center loss ($\mathcal{L}_{\text{mcc}}$):

$$\mathcal{L}_{\text{part}} = \mathcal{L}_{\text{bl}} + \alpha\mathcal{L}_{\text{mcc}}. \tag{5.3}$$

The goal of $\mathcal{L}_{\text{mcc}}$ is to ensure the selected region for each part remains consistent across all inputs and serves as a prototype; it must capture similar semantic information across different classes. Additionally, these sub-part features should be both class-descriptive and contain unique ID-aware information.

5.3.1.1 Marginal Cluster Center Loss

We proposed learnable prototypes as features to be distant from each other and then pushed the model to discover the parts from input images to be close to these prototypes. A soft margin is used to make space for the same part from input images with different classes.

$$\mathcal{L}_{\text{mcc}} = \frac{1}{K(K-1)} \sum_{p=1}^{K} \left([||\mathbf{f}_p^i - c_p|| - m_1]_\star + \sum_{q=p+1}^{K} [m_2 - ||c_p - c_q||]_\star \right), \tag{5.4}$$

where c_p is a learnable prototype for part p and $[\cdot]_\star$ is $\max(\cdot, 0)$.

5.3.2 Step 2: Part-Centroid List Generation

During training in Step 1, the backbone learns to extract and use part-prototype features from the input data to fulfill its training task. To find all relevant spatial and class-aware concepts, as shown at the top of Fig. 5.4, we collect all extracted features from the backbone for each class and each part and then cluster them to have a minimal set of plain concepts that consistently describe the visual attributes of those parts. Multiple concepts are considered. They can be seen in each part of each class for more complex objects. To mine those concept prototypes, the centroids of each cluster are computed as $C(l, p, j) \in \mathbb{R}_f^d$ for a clustering index of l, part p, and a class of j.

5.3.3 Step 3: Concept Activation Mining

Once the centroid list is generated, each concept's activation value is defined by the similarity of extracted features to each centroid:

$$\mathbf{z}^i = [\mathcal{S}(C(l, p, j), \mathbf{f}_p^i) \; \forall p \in \{1..K\} \text{ and } j \in \{1..L\}] \in \mathbb{R}^{d_c} \tag{5.5}$$

where $\mathcal{S}$ is cosine distance and $[.]$ represents the vector concatenation. d_c is the number of total concept centroids. We aim to explain each output class of our model using a small set of interpretable concepts. We, therefore, train a sparse classifier on top of z^i to obtain the final classification scores $o^i = W_1^T \mathbf{z}^i + W_2^T \mathbf{g}^i + b$ and the predicted class $\hat{y}^i = \arg\max(o^i)$. Here, $W_1 \in \mathbb{R}^{d_c \times L}$, $W_2 \in \mathbb{R}^{d_f \times L}$ and $b \in \mathbb{R}^L$ denote the classification weights and bias term, respectively. This layer is trained with the following classification loss, using the GLM-SAGA optimizer (Wong, Santurkar & Madry, 2021):

$$\mathcal{L}_{\text{cl}} = \mathcal{L}_{ce}(W_1^T \mathbf{z}^i + W_2^T \mathbf{g}^i + b, y^i) + \lambda \mathcal{R}(W_1), \tag{5.6}$$

where $\mathcal{L}_{\text{ce}}$ is the cross-entropy loss, y^i is the label, λ is the sparsity regularization strength and $\mathcal{R}(W) = (1-\gamma)\frac{1}{2}\|W\|_F + \gamma\|W\|_1$, here $\|W\|_F$ is the Forbenius norm and $\|W\|_1$ is the element-wise matrix.

5.3.4 Overall Training

We jointly optimize the network in an end-to-end manner by loss:

$$\mathcal{L} = \mathcal{L}_{\text{part}} + \beta \mathcal{L}_{\text{cl}}, \tag{5.7}$$

where β is hyperparameter. For the first step of training, we set $\beta = 0$; then, once the part-prototypical concepts are learned, we set $\beta = 2$ based on our hyperparameter analysis.

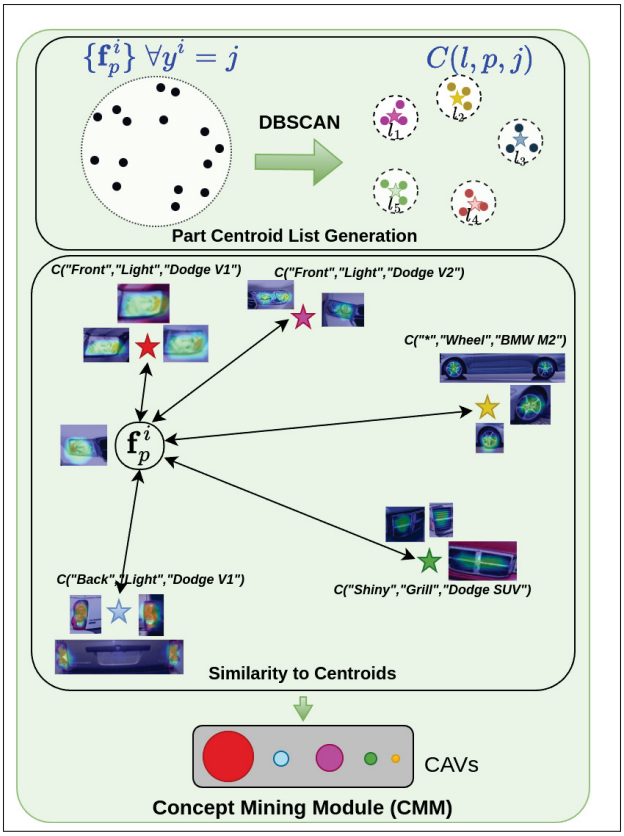


Figure 5.4 Concept Mining Module. First, a part centroid list is generated by applying DBSCAN(Ester *et al.*, 1996) clustering within each class and computing the centers of the resulting clusters. These centroids serve as frequently occurring concept prototypes. Next, to activate the most relevant concept for a given part feature, the similarity (inverse distance) to the centroids is measured. These values are then used as CAVs to inform the model's decision-making process

5.4 Experiments and Results

PCMNet is validated on standard benchmarks in prototype literature: Stanford Cars (Krause, Stark, Deng & Fei-Fei, 2013) (196 classes of different make and models) and CUB-200-2011 (Wah, Branson, Welinder, Perona & Belongie, 2011) (200 bird species, including attribute annotation for visual features). The ResNet50 model pretrained on ImageNet (Deng *et al.*, 2009a) is used as the feature backbone. Each training batch (32 images) is resized to 488×488, randomly cropped, erased, and padded. Optimization follows ADAM (Kingma & Ba, 2014) with a linear warm-up. We set regularization parameters λ and γ as same in (Oikarinen, Das,

Table 5.1 Performance of PCMNet according to xAI metrics on the Stanford Cars and Birds datasets

	Method	Consistency (Intra) ↑	Consistency (Inter) ↓	F(1)-F(5) ↑	Sp ↑	Stability	C Acc ↑
Stanford Cars	Resnet50+Act	**83.45 ± 2.7**	27.26 ± 28.10	0.16 - 0.23	22.74	65.4	84.1
	Resnet50+W × Act	**83.45 ± 2.7**	27.26 ± 28.10	1.54 - 6.61	22.74	65.4	84.1
	ProtoPNet + Act	45.70±4.9	11.22±5.5	2.2-55.64	60.92	69.6	84.5
	ProtoPNet + W × Act	45.70 ± 4.9	11.22±5.5	9.3-80.12	60.92	69.6	84.5
	PiPNet(Covx) + Act	52.78±7.4	9.07±9.4	0.02-60.62	70.59	65.2	88.5
	PiPNet(Covx) + W × Act	52.78 ± 7.4	9.07±9.4	10.83-**95.66**	**70.59**	65.2	88.5
	PiPNet(Res50) + Act	42.40±17.5	17.4±2.4	0.18 - 1.21	63.19	60.8	86.46
	PiPNet(Res50) + W × Act	42.40±17.5	17.4±2.4	9.43 - 93.15	63.19	60.8	86.46
	Ours + Act	55.32±6.1	1.88 ± 9.37	37.94 - 63.3	57.45	**71.5**	**88.7**
	Ours + W × Act	55.32±6.1	**1.88 ± 9.37**	**59.33** - 91.73	57.45	**71.5**	**88.7**
Birds (CUB 200)	Resnet50+Act	57.38 ± 3.50	25.45 ± 23.82	0.35 - 1.12	24.14	60.3	81.2
	Resnet50+Rel × Act	57.38 ± 3.50	25.45 ± 23.82	2.43 - 5.73	24.14	60.3	81.2
	ProtoPNet + Act	51.47±5.8	10.16±9.2	1.15-51.96	79.51	63.6	81.45
	ProtoPNet + W × Act	51.47 ± 5.8	10.16±9.2	9.36-75.50	79.51	63.6	81.45
	PiPNet(Covx) + Act	54.80±8.3	7.77±8.0	0.01-54.14	80.72	64.7	**85.0**
	PiPNet(Covx) + W × Act	54.80 ± 8.3	7.77±8.0	5.13-92.83	**80.72**	64.7	85.0
	PiPNet(Res50) + Act	42.28±3.2	15.57±15.7	0.14 - 2.35	77.83	65.9	82.0
	PiPNet(Res50) + W × Act	42.28±3.2	15.57±15.7	9.40 - 89.77	77.83	65.9	82.0
	PCMNet (Ours) + Act	**58.55±7.5**	**1.22 ± 10.1**	41.82 - 75.24	57.45	67.1	84.7
	PCMNet (Ours) + W × Act	**58.55±7.5**	**1.22 ± 10.1**	**55.13 - 89.88**	57.45	**67.8**	84.7

Nguyen & Weng, 2023). $\alpha = 1.5, \beta = 2$, with $m_1 = 0.3$ and $m_2 = 1.5$ chosen via hyperparameter analysis in the suppl. material. More details on the experimental protocol are presented in the supplementary materials.

5.4.1 Explainability Metrics

To evaluate the explainability of final features, we rely on performance metrics established in the literature (Dreyer *et al.*, 2024) with some minor modifications to express better comprehension.

(a) Faithfulness:: Faithfulness(Dasgupta, Frost & Moshkovitz, 2022; Chan, Kong & Liang, 2022) evaluates how much the activated concepts influence the model's final prediction, serving as a key metric for assessing the quality of explanations. One widely used approach for measuring faithfulness is concept deletion, which involves removing selected concepts from the model's reasoning process and observing the resulting change in output confidence. Prior works, such as (Dreyer *et al.*, 2024), typically assess this by removing only the single most important concept. However, this limited scope fails to capture the broader contribution of

other activated concepts and may not reveal the full extent of their impact. To provide a more comprehensive assessment, we extend this approach by successively removing the top-k most activated concepts and measuring the resulting drop in model confidence. A larger drop indicates higher faithfulness, as it suggests that the removed concepts were indeed critical to the model's decision. This extended evaluation allows us to better quantify how much the explanation aligns with the model's internal reasoning.

(b) Stability: The extracted concepts from input images that share some semantic concepts must be stable and explainable for unseen images. To evaluate stability, we compute CAVs on k-fold subsets of the data ($k = 10$ as default) similar to (Dreyer *et al.*, 2024) for all classes at steps 2 and 3. We then map CAVs together using a Hungarian loss function and measure the cosine similarity between them.

(c) Consistency: Activated concepts from images with the same class should be similar to be consistent and explain the model's decision. To measure this consistency, we extract the CAV from images and then measure the cosine similarity between them. The mean cosine similarity is assessed between images from the same class and between images from different classes. The ratio between inter/intra-class similarity should be high for a stable explanation.

(d) Sparseness: The sparseness metric essentially describes the uniformity of the concept activation, where having some concepts activate more than others is considered easier to interpret. This is because a uniform distribution of concept activations would provide little information on the importance of specific concepts or parts of the images, i.e., a high entropy in the generated concepts.

5.4.2 Comparison to Prototype xAI Methods

To evaluate PCMNet, we report the classification accuracy and the XAI evaluation criteria discussed in Sec 4 and compare them to other ante-hoc methods (ProtoPNet(Chen *et al.*, 2019) and PIPNet(Nauta *et al.*, 2023)) in tables 5.1. The consistency metric is reported in two ways: Intra and Inter, which are cosine similarities between activated concepts of the same class and

different classes, respectively. Faithfulness is reported as **F(*n*)**, measuring accuracy drops when the top ***n*** important concepts are removed from the decision. To compute the contribution of each concept or feature to the final decision, we evaluate two strategies: (1) using raw values of each concept ("Act"), and (2) computing a weighted contribution (gradient) as the product of raw values and its corresponding weight in the classification layer ("W $\times$ Act").

5.4.3 Ablation Studies

5.4.3.1 Number of Prototypes

We assess the impact of the number of parts on both classification accuracy and faithfulness in Table 5.2. As the model extracts more prototypical concepts, its decisions become more aligned with the underlying features, enhancing interpretability. Interestingly, when the number of prototypes is smaller, the faithfulness scores for the top concepts tend to be higher. This is because the model relies on a more compact set of prototypes, concentrating its decisions on fewer, more influential concepts.

5.4.3.2 Effect of Modules

To evaluate the effectiveness of each component in PCMNet, we conduct an ablation study on the baseline PDiscoNet model, which extracts local parts without explicit concept learning.

Table 5.2 Impact on performance of the number of prototypes

K value	Cars		Birds	
	C Acc	**F(1)-F(5)**	**C Acc**	**F(1)-F(5)**
0 (Resnet50)	84.1	1.54 - 6.61	83.2	1.21-5.56
4	83.5	**79.96** - 91.01	83.5	**71.42**-85.64
5	84.1	70.11 – 91.33	84.9	66.7 - 85.9
6	85.2	66.02 – 90.65	85.7	59.9 - 86.7
7	86.7	62.47 – 91.25	**86.0**	55.7 - **87.3**
8	**87.5**	59.33 – **91.73**	85.1	53.0 - 86.4
9	86.4	55.10 – 91.52	84.4	52.5-86.1

Figure 5.5 Visualizing activated concepts in the test set and examples from the training set for Stanford Cars and CUB datasets

As shown in Table 5.3, we analyze classification accuracy (C Acc) and feature faithfulness under occlusion (F(3)) on both the Cars and Birds datasets, alongside the model size and FLOPs. Introducing the Marginal Clustering Center (MCC) loss into the baseline improves classification performance (from 84.2% to 86.1% on Cars) and more than doubles faithfulness scores, indicating that enforcing clustering consistency across part features supports more explainability. The inclusion of the Concept Mining Module (CMM)—responsible for extracting sparse and interpretable concepts—dramatically boosts faithfulness (from 4.9% to 55.6% on Birds), even with minimal impact on accuracy. When both MCC and CMM are combined, PCMNet achieves the best results across all metrics, including 87.5% classification accuracy and 67.4% faithfulness on Cars, with only a marginal increase in parameter count and computational cost. These findings underscore the complementary roles of MCC in enforcing cluster-level consistency and CMM in mining high-quality, human-aligned concepts.

Table 5.3 Classification accuracy and faithfulness using different PCMNet modules

Settings	**Cars**		**Birds**		**#Params**	**FLOPs**
	C Acc	**F(3)**	**C Acc**	**F(3)**		
Baseline(van der Klis *et al.*, 2023)	81.2	5.4	82.3	4.9	27.1M	17.2G
Base + MCC	86.1	9.6	84.5	9.1	27.2M	17.2G
Base + CMM	84.3	59.2	82.1	55.6	27.4M	17.5G
Base + MCC + CMM	**87.5**	**67.4**	84.7	**64.9**	27.5M	17.5G
PipPNet(Resnet)	86.4	62.7	82.0	58.6	25.1M	27.6G
PipPNet(Covx)	88.5	60.6	**85.0**	57.5	29.3M	32.3G
ProtoPNet	84.5	38.6	81.4	31.7	28.9M	18.4G

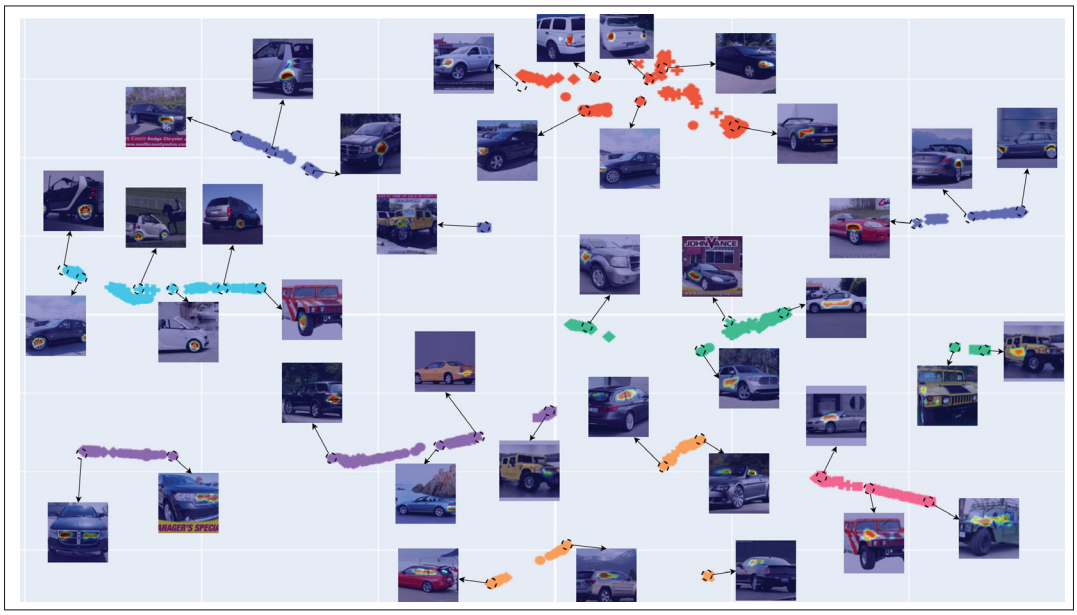


Figure 5.6 t-SNE visualization of concept-level part features extracted by PCMNet. Colors represent different prototypes, while shapes indicate object classes

5.4.4 Robustness to Occlusion

To evaluate the robustness of PCMNet's decision-making under occlusion, we conducted controlled experiments in which regions corresponding to the most activated concepts were systematically masked, and the resulting images were re-evaluated by the model. This experiment aims to assess the stability of model predictions when key interpretable regions are removed. Specifically, we identify the center of the most activated concept—based on the model's response

to clean images—using the concept representation (patch for PIPNet and ProtoPNet, mask for PCMNet), and then occlude a rectangular region centered around this point. We consider three levels of occlusion, masking 10%, 20%, and 30% of the image area. PCMNet's performance under these conditions is compared against two other interpretable baselines: ProtoPNet (Chen *et al.*, 2019) and PIPNet (Nauta *et al.*, 2023).

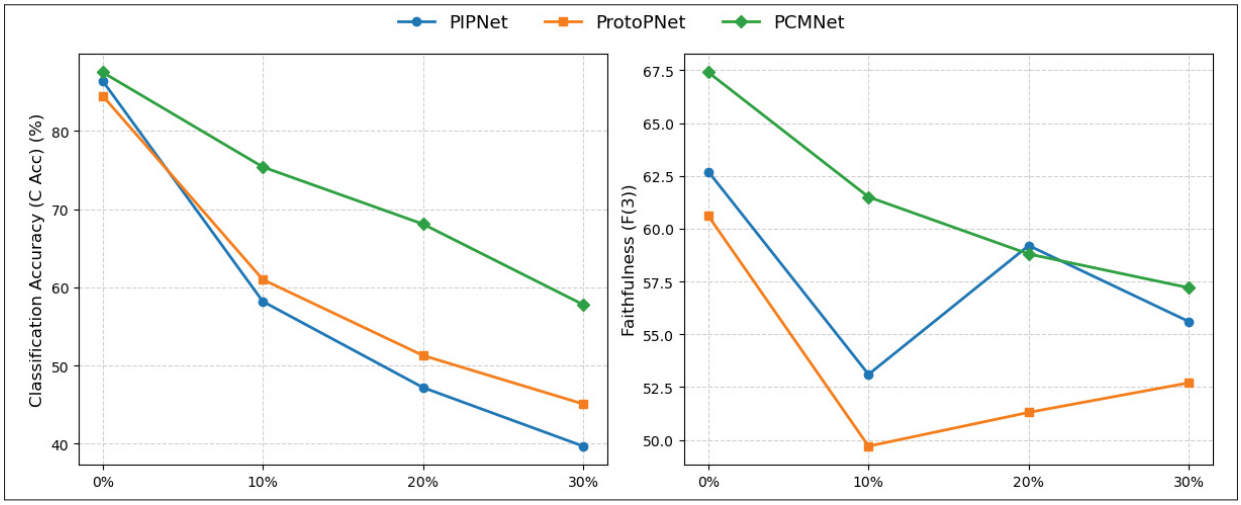


Figure 5.7 Classification Accuracy and Faithfulness under Different Occlusion Levels

Fig. 5.7 shows that PCMNet maintains stable accuracy at lower occlusion levels and degrades more gradually than ProtoPNet and PIPNet. This resilience stems from PCMNet's broader part-based concepts, which introduce redundancy and enable accurate predictions even when key components are missing. Unlike patch-based models that rely on fixed, small regions, it extracts prototypes from semantically meaningful, wider regions. As a result, such methods experience sharp performance drops under occlusion due to their dependence on localized patterns, whereas PCMNet remains more robust by leveraging multiple related part concepts, preventing performance from relying too heavily on a single visual cue.

5.4.5 Qualitative Results

5.4.5.1 Concept Activation and Occlusion Robustness

To further illustrate the interpretability and robustness of PCMNet, we compare the activated concepts of our method with PIPNet (Nauta *et al.*, 2023) under clean and occluded conditions, as visualized in Fig. 5.5. For each method, we select two activated concepts on test images and retrieve training set examples with the highest activation values for those concepts. This comparison shows several key advantages of PCMNet:

Broader Concept Coverage: PCMNet extracts more diverse and semantically rich concepts compared to PIPNet. Our method leverages attention-based masks instead of fixed patches to capture part-level semantics over a broader spatial context, leading to more varied and interpretable regions across different object parts.

More Informative Examples: The training examples retrieved by PCMNet for each concept exhibit more clearly activated and semantically aligned regions. This suggests that the discovered concepts better reflect the underlying factors influencing the model's decision, thereby enhancing explanation quality.

Robustness to Occlusion: Under occlusion, PIPNet often fails to highlight meaningful or consistent regions due to its reliance on limited regions. In contrast, PCMNet maintains interpretability by activating alternative, unoccluded parts that belong to the same or related concepts. This demonstrates the model's ability to generalize concept reasoning even when certain visual cues are missing.

5.4.5.2 Visualization of Concept Prototypes

To better understand the behavior of the learned prototypes in PCMNet, we visualize the extracted part-level features using t-SNE (van der Maaten & Hinton, 2008), as shown in Figure 5.6. Each

point corresponds to a feature from a localized part, colors indicating the prototype (i.e., concept) and marker shapes denoting the object class. This visualization represents several insights:

Disentanglement of Prototypes: Features assigned to different prototypes form well-separated clusters, suggesting that PCMNet effectively disentangles the latent interpretable part-based concept space. This separation implies that the Marginal Clustering Center loss and Concept Mining Module (CMM) encourages diverse and semantically distinct part concepts, which supports both interpretability of model decisions and robustness on occlusion.

Multi-Class Concept Coverage: Many prototype clusters contain the same semantic parts from multiple object classes. This indicates PCMNet learns class-agnostic concepts such as lights, wheels, or windows, which are shared across different categories. Such generalization is crucial for explainability, as it facilitates human-aligned, semantically meaningful concepts that transcend dataset labels.

Intra-Class Concept Diversity: Interestingly, features from the same class are distributed across multiple prototype clusters. This highlights PCMNet's ability to capture *intra-class variation* through multiple part-based concepts. For instance, different views or structural variations of a car class may be assigned to distinct concepts (Backlight vs headlight), further enhancing the model's fine-grained interpretability based on human perception and explanation.

5.5 Conclusion

We presented PCMNet, a novel framework designed to bridge the gap between accuracy and interpretability in deep learning models. By mining part-prototypes from dynamic image regions, PCMNet improves upon existing patch-based prototype models by ensuring semantic coherence across instances. Our results show that PCMNet enhances faithfulness, concept stability, and explainability, outperforming existing interpretable models such as ProtoPNet and PIPNet. Moreover, our occlusion-based robustness analysis confirmed that PCMNet remains stable under missing or distorted input regions, demonstrating its reliability in real-world scenarios.

Unlike conventional post-hoc explainability methods, PCMNet allows for more human-aligned, concept-based reasoning, making deep learning decisions more transparent and interpretable.

5.6 Additional Experiment and Results

5.6.1 Implementation Details

We used Resnet50 as the backbone, which extracts 2048 features. We project these features into 512 features before global pooling, i.e., $d_f = 512$. Based on our experiment and set of hyper-parameters, for the CUBs dataset, the $d_c = 2954$ and Stanford Cars $d_c = 3092$ are chosen. As the DBSCAN is used for clustering inside of each class, for each part and class, the number of clusters is not fixed.

The $\mathcal{L}_{\text{bl}}$(van der Klis *et al.*, 2023) loss also is

$$\mathcal{L}_{\text{bl}} = \mathcal{L}_{\text{cond}} + \mathcal{L}_{\text{equi}} + \mathcal{L}_{\text{pres}}, \tag{5.8}$$

where the objectives for each loss are:

- $\mathcal{L}_{cond}$: Each detected part should consist of a condensed and contiguous image region.
- $\mathcal{L}_{equi}$: encourages the equivariance of the attention maps in for input images and their augmented version (translation, rotation or scaling).
- $\mathcal{L}_{pres}$: All parts should be present at least in some of the images in the batch to avoid finding non-relevant parts.

5.6.2 Additional Ablation Studies

5.6.2.1 Effect of Hyper-parameters

We investigate the sensitivity of our model to the hyper-parameters α and β, which control the strength of the concept consistency and interpretability regularization terms, respectively. As shown in Fig. 5.8, our model remains robust across a wide range of values. Specifically,

the classification accuracy (C Acc) and faithfulness score **F(*n*)** exhibit stable trends, with peak performance achieved around $\alpha = 1.5$ and $\beta = 2.0$. This indicates that PCMNet effectively balances accuracy and interpretability without requiring extensive hyper-parameter tuning. The consistency of the scores across different settings highlights the reliability of the proposed regularization components.

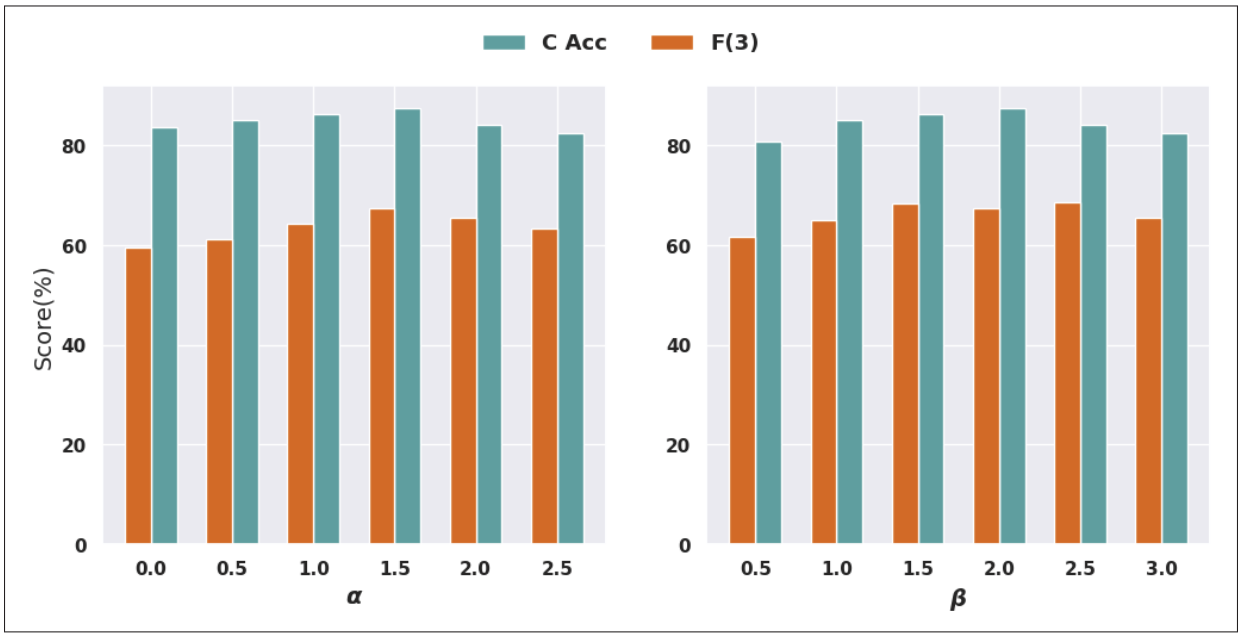
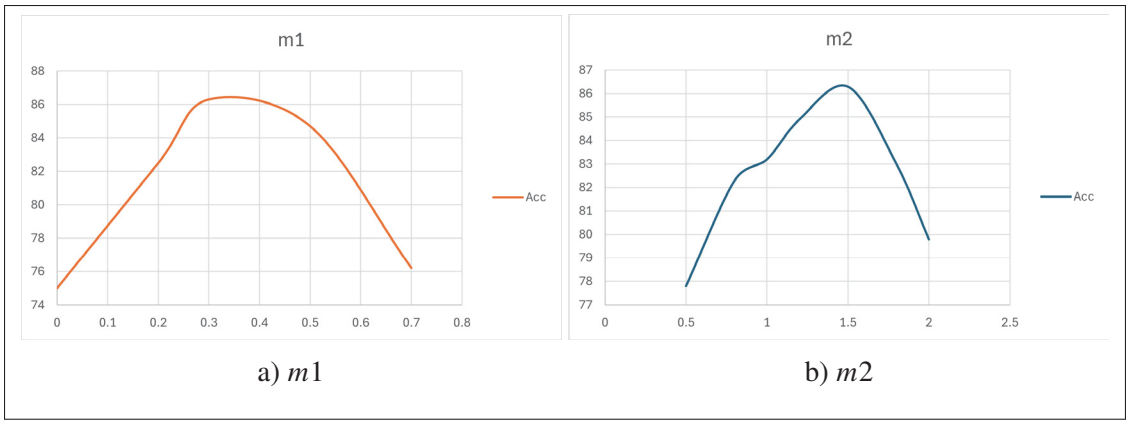


Figure 5.8 Impact of hyper-parameters α and β on classification accuracy (C Acc) and faithfulness score **F(*3*)**. Our method maintains strong performance across a wide range of values, showing robustness and stability in both predictive accuracy and interpretability

Also, in Fig. III-2, we evaluate the impact of hyperparameters in Eq.5.4 on classification accuracy. At first, we set $m_2 = 1$ and measure the effect of m_1 by gradually increasing the values. Once it is found, we search for m_2. The best performance is achieved when m_1 is set to 0.3 and m_2 to 1.5, respectively. The upward trend of the bars demonstrates the effectiveness of each loss.

5.6.2.2 Effect of Non-Prototypical Concepts

To evaluate the impact of different concept vector types on model performance and interpretability, we conduct an ablation study using prototypical ($\mathbf{z^i}$) and non-prototypical ($\mathbf{g^i}$) concept vectors. As



a) $m1$ b) $m2$

Figure 5.9 The model's classification accuracy in different values of $m1$ and $m2$

shown in Table 5.4, using only non-prototypical vectors significantly reduces both classification accuracy and faithfulness scores. In contrast, employing only prototypical vectors retains higher classification performance and strong concept-based faithfulness. The full PCMNet, which combines both concept types, achieves the best results across both datasets, suggesting that the interplay between prototypical and non-prototypical representations enhances both prediction accuracy and interpretability.

Table 5.4 Impact of prototypical ($\mathbf{z}^i$) and non-prototypical ($\mathbf{g}^i$) concept vectors on PCMNet's classification accuracy (C Acc) and faithfulness scores (F(1)-F(5)) on Cars and Birds datasets. The full model achieves superior performance, highlighting the benefit of combining both concept types

CAVs	Cars		Birds	
	C Acc	**F(1)-F(5)**	**C Acc**	**F(1)-F(5)**
Baseline	84.1	1.54 - 6.61	83.2	1.21-5.56
PCMNet + Prototypical	86.2	61.8 – 92.63	84.5	56.8 - 87.4
PCMNet + Non-Prototypical	79.4	42.80 – 75.73	78.15	35.9 - 67.7
PCMNet (Full)	**87.5**	**59.33 – 91.73**	**86.0**	**55.7 - 87.3**

5.6.3 Additional Qualitative Results

More results of the PCMNet prototypical concept are shown in Figs. 5.10 and 5.11. 5.12 illustrates examples of non-prototypical concepts discovered by our model for the Cars (left) and

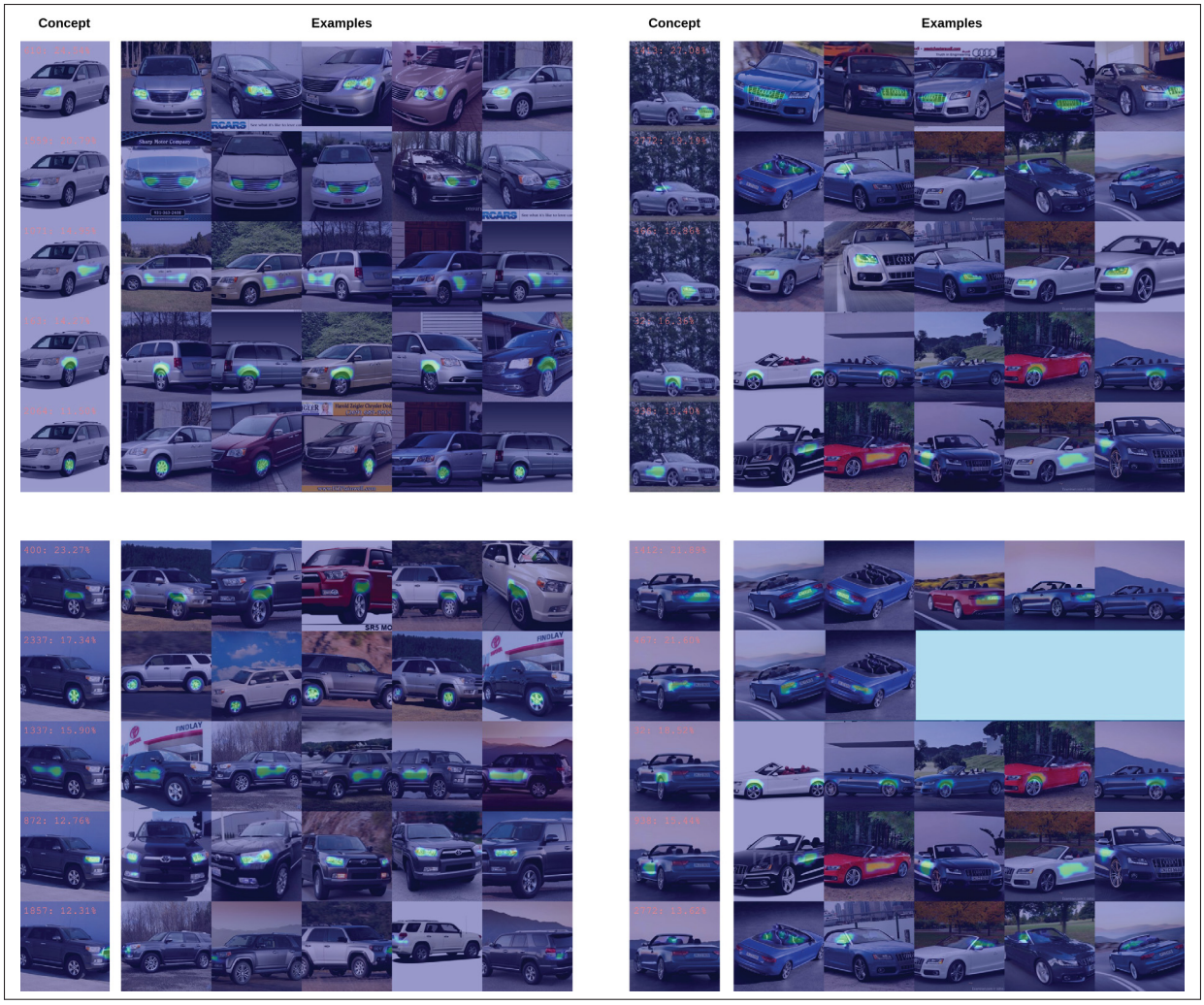


Figure 5.10 Visualization of activated concepts in test data and showing some examples in the training set for StanfordCars datasets. Each concept is shown with an index and percentage of explanation for the model decision

CUBs (right) datasets. For each dataset, the concept location on a test image is highlighted with a green bounding box, while corresponding activations in training examples are shown on the right using red bounding boxes. These concepts provide complementary cues to prototypical ones. For example, in the first row of Cars, the "cargo" attributes are not well represented in prototypical concepts since they appear in a small subset of vehicles, yet they are highly discriminative for decision-making. Similarly, the last row in CUBs shows the "webbed feet" concept, which is especially relevant and explainable for Albatross species, while foot prototypes are not

sufficiently distinctive across the dataset. These findings highlight the value of incorporating non-prototypical concepts in improving both interpretability and robustness of the model.

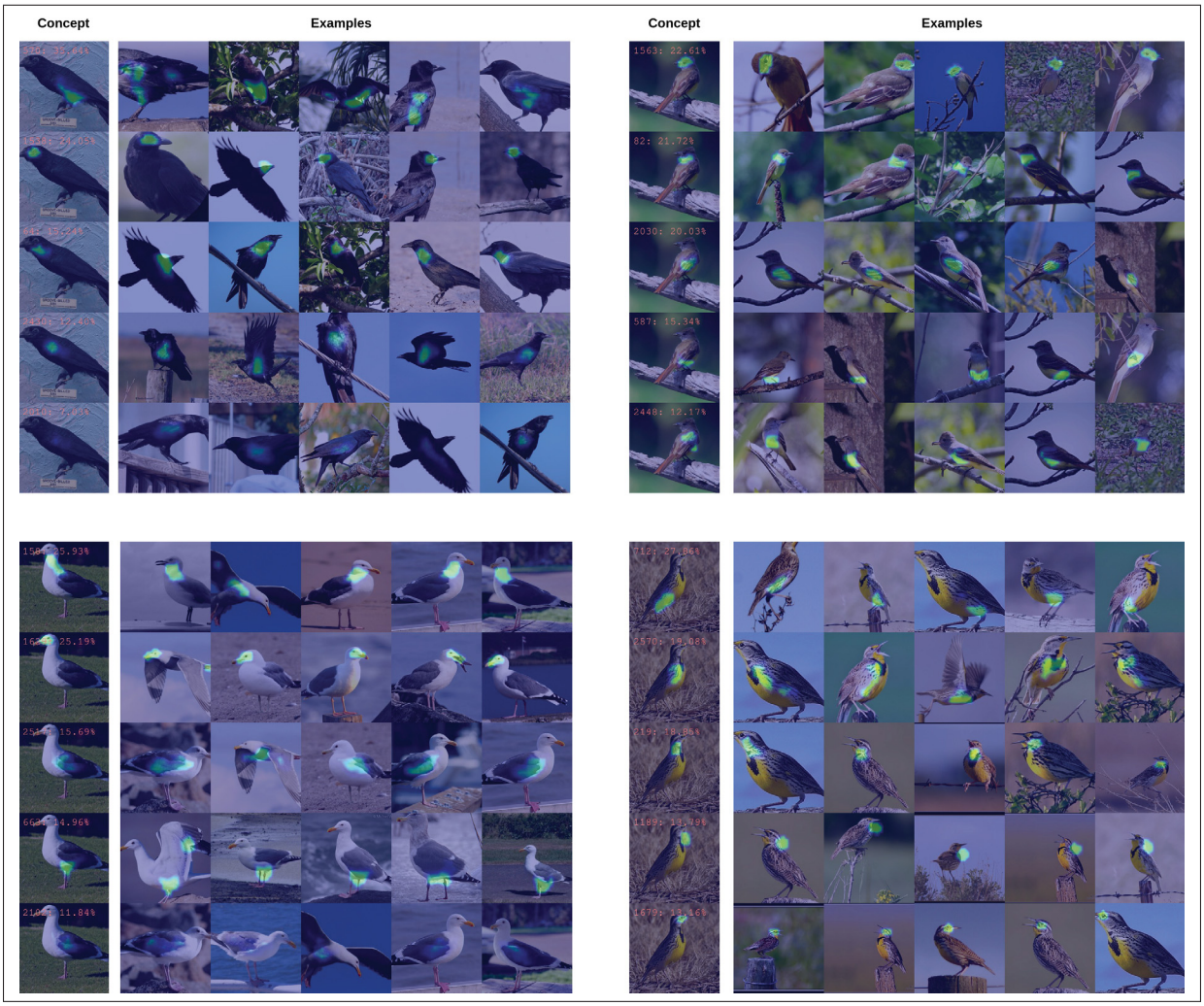


Figure 5.11 Visualization of activated concepts in test data and showing some examples in the training set for CUB-200 Birds datasets. Each concept is shown with an index and percentage of explanation for the model decision

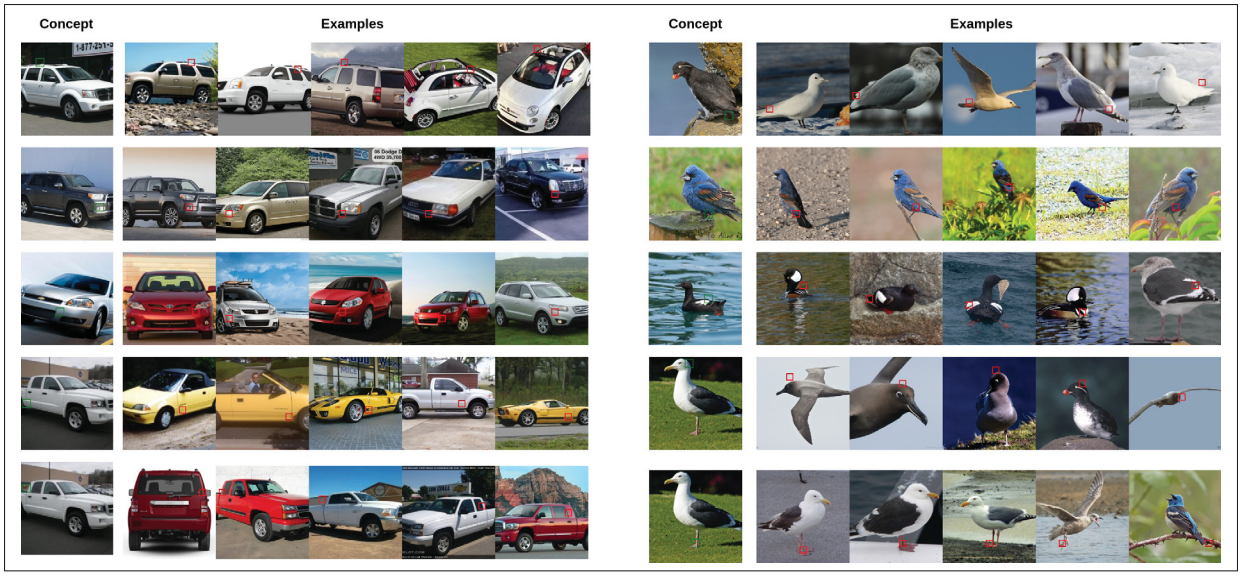


Figure 5.12 Visualization of non-prototypical activated concepts in test data and showing some examples in the training set for (left) Stanford Cars and (right) CUB-200 Birds datasets. Similar to ProtoPNet and PIPNet, we select the patch for the region of the activated concept

CONCLUSION AND RECOMMENDATIONS

6.1 Introduction

This thesis investigated the problem of *Visible-Infrared Person Re-Identification (VI-ReID)*, a challenging task in modern surveillance and security applications where the goal is to match individuals across visible (RGB) and infrared (IR) cameras. Unlike conventional uni-modal person re-identification, VI-ReID must overcome a **substantial domain gap** caused by differences in sensing modalities, lighting conditions, and environmental factors.

Despite recent advances, most existing VI-ReID methods are designed and evaluated under simplified conditions, often assuming uni-modal galleries or restricted matching protocols. Such assumptions limit the applicability of these models in **real-world 24-hour surveillance**, where visible and infrared data naturally coexist and where occlusions, viewpoint changes, and noise are common.

To address these challenges, this thesis introduced a **gradual domain generalization framework** that systematically improves cross-modal alignment, enhances robustness, and reflects real-world conditions through novel evaluation protocols. The contributions are presented through three core articles and one complementary study in explainable AI, which collectively advance both the methodological and practical aspects of VI-ReID.

6.2 Summary of Contributions

6.2.1 Adaptive Generation of Privileged Intermediate Information (AGPI^2)

The first contribution introduced **AGPI**^{2}, a framework that generates adaptive intermediate modalities to act as privileged information during training. Instead of directly synthesizing one

modality from another, AGPI2 creates a **bridging representation** that reduces the discrepancy between visible and infrared domains while preserving identity-discriminative information.

Key outcomes include:

- Demonstrated that **privileged intermediate spaces** can effectively regularize training and reduce modality bias.
- Achieved **state-of-the-art results** on SYSU-MM01, RegDB, and LLCM datasets.
- Provided extensive **ablation studies** showing the importance of loss functions, discriminators, and augmentation strategies in shaping modality-invariant embeddings.

6.2.2 Bidirectional Multi-step Domain Generalization (BMDG)

The second contribution proposed **BMDG**, a structured framework that gradually aligns part-based prototypes across modalities. Unlike single-step adaptation methods, BMDG leverages **multi-step refinement** to progressively close the domain gap while ensuring bidirectional consistency.

Key outcomes include:

- Introduced a **prototype learning module** that captures semantically consistent body parts across modalities.
- Showed that gradual alignment improves **generalization under severe domain shifts**.
- Outperformed state-of-the-art methods, particularly in scenarios with significant cross-modal discrepancies.

6.2.3 Mixed-Modality Re-Identification (MixER / Mixed-ReID)

The third contribution addressed the gap between existing evaluation protocols and real-world surveillance needs. Current benchmarks assume a gallery from a single modality, whereas real systems encounter **mixed galleries** containing both visible and infrared images.

To solve this, we proposed:

1. **New evaluation protocols** (Mix-all, Mix-camera, Mix-identity, Mix-open, etc.), which more faithfully model real-world conditions.
2. **MixER**, a method that disentangles identity features into **modality-related** and **modality-erased** components, enabling improvements in both intra- and inter-modality matching.

Key outcomes include:

- Exposed the limitations of existing state-of-the-art models under realistic mixed conditions.
- Demonstrated that **identity-aware disentanglement** significantly improves both cross- and mixed-modal retrieval accuracy.
- Established MixER as a **new baseline framework** for future research in VI-ReID.

6.2.4 Interpretable Concept Discovery (PCMNet)

Finally, beyond improving performance, this thesis explored the **interpretability** of deep ReID models. We introduced **PCMNet**, a framework that mines interpretable **part-prototypes** from images without manual annotation. These prototypes represent semantically meaningful concepts (e.g., clothing textures, shapes, accessories) and allow the model's decisions to be **explained in human-understandable terms**.

Key outcomes include:

- Proposed a new approach to **faithful and robust explanations** for deep vision models.
- Demonstrated strong interpretability without sacrificing classification performance.
- Paved the way for **trustworthy deployment** of ReID models in high-stakes applications.

6.3 Key Findings and Insights

From the above contributions, several broad insights emerge:

1. **Gradual adaptation outperforms single-step methods.** Progressive refinement of intermediate domains, as shown in AGPI2 and BMDG, yields more stable and robust feature alignments.
2. **Mixed-modality evaluation is essential.** Existing benchmarks overestimate real-world performance; realistic protocols reveal critical weaknesses and inspire stronger models.
3. **Disentanglement of features is powerful.** Separating modality-related from modality-erased information leads to more generalizable embeddings and provides a path for tackling unseen domains.
4. **Interpretability strengthens trust.** Performance alone is insufficient for deployment—explainable mechanisms such as PCMNet improve transparency, reliability, and acceptance in real-world systems.

6.4 Future Directions

6.4.1 Dataset Expansion and Real-World Scenarios

- Collecting **larger-scale mixed-modality datasets** with diverse sensors, occlusions, and environments.
- Incorporating **temporal cues** from video streams for sequential reasoning.

6.4.2 Integration with Foundation and Multimodal Models

- Leveraging **vision-language models** (e.g., CLIP) to enhance cross-modal alignment using textual supervision.
- Exploring **large multimodal pretraining** for transfer to low-resource VI-ReID tasks.

6.4.3 Efficient Deployment on Edge Devices

- Developing **lightweight architectures** for real-time inference in surveillance systems.
- Investigating **heterogeneous computing strategies** for efficient deployment across CPU, GPU, and NPU hardware.

6.4.4 Extending Beyond VI-ReID

- Applying the gradual domain generalization framework to other **cross-modal tasks** such as thermal-visible tracking or multimodal face recognition.
- Studying **privacy-preserving ReID** by combining interpretability with secure feature representations.

6.5 Closing Remarks

This thesis has demonstrated that **gradual domain generalization** provides a powerful paradigm for bridging modality gaps and building robust VI-ReID systems. By progressively refining embeddings, introducing realistic evaluation protocols, and enhancing interpretability, we moved closer to **trustworthy, generalizable, and deployable** solutions for real-world surveillance.

The findings not only push the boundaries of VI-ReID research but also open opportunities for broader cross-modal and interpretable AI. As intelligent systems continue to play a growing role in society, the principles established here—**gradual adaptation, disentanglement, and transparency**—can serve as guiding foundations for the next generation of multimodal AI.

APPENDIX I

SUPPLEMENTARY OF AGPI^2

I.A. Proofs

I.A.1 Maximizing MI$(Z;Y)$

Proposition I.1. *Let Z and Y be random variables with domains $\mathcal{Z}$ and $\mathcal{Y}$, respectively. Minimizing the conditional cross-entropy loss of predicted label $\hat{Y}$, denoted by $\mathcal{H}(Y;\hat{Y}|Z)$, is equivalent to maximizing the MI$(Z;Y)$*

Proof. Let us define the MI as entropy,

$$\mathrm{MI}(Z;Y) = \underbrace{\mathcal{H}(Y)}_{\delta} - \underbrace{\mathcal{H}(Y|Z)}_{\xi} \tag{A I-1}$$

Since the domain $\mathcal{Y}$ does not change, the entropy of the identity δ term is a constant and can therefore be ignored. Maximizing $\mathrm{MI}(Z;Y)$ can only be achieved by minimizing the ξ term. By expanding its relation to the cross-entropy (Boudiaf *et al.*, 2020):

$$\mathcal{H}(Y,\hat{Y}|Z) = \mathcal{H}(Y|Z) + \underbrace{\mathcal{D}_{\mathrm{KL}}(Y||\hat{Y}|Z)}_{\geq 0}, \tag{A I-2}$$

where:

$$\mathcal{H}(Y|Z) \leq \mathcal{H}(Y,\hat{Y}|Z). \tag{A I-3}$$

So, $\mathcal{H}(Y|Z)$ is upper-bounded by our cross-entropy loss, and minimizing such loss results in minimizing the ξ term.

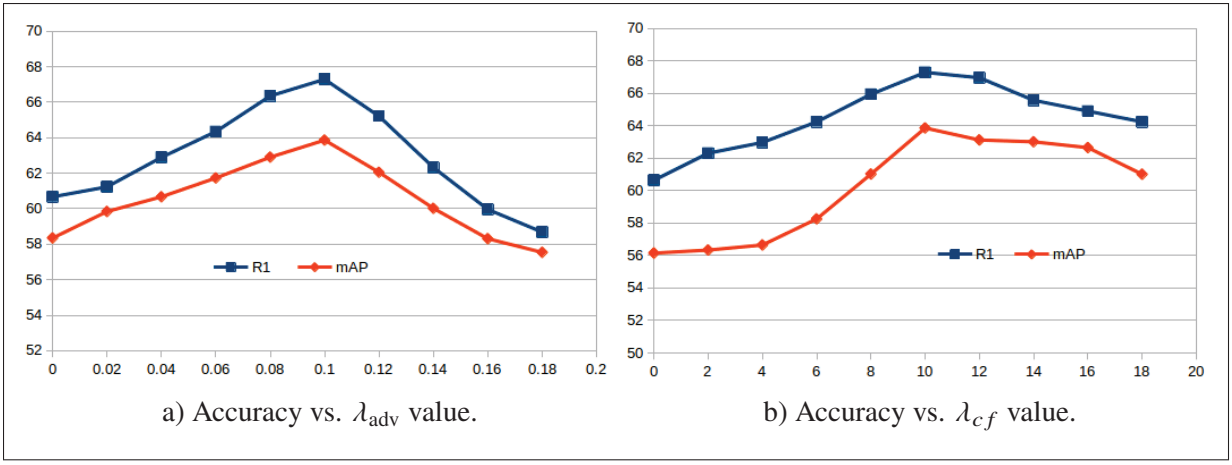
Through the maximization ID-Aware part of Eq. 2.3, training can naturally be decoupled in 2 steps. First, weights of the generative and feature backbone modules are fixed, and only the classifier parameters (i.e., the weight of the fully connected layer at the feature backbone) are

minimized w.r.t. Eq. A III-13. Through this step, $\mathcal{D}_{\text{KL}}(Y||\hat{Y}|Z)$ is minimized by adjusting $\hat{Y}$ while the $\mathcal{H}(Y|Z)$ does not change. In the second step, the prototype module's weights are minimized w.r.t. $\mathcal{H}(Y|Z)$, while the classifier parameters are fixed. □

I.B. Additional Results

I.B.1 Hyperparameter Values

This subsection provides an analysis of the impact on ReID accuracy of the λ_a and λ_{cf} hyperparameters. λ_a balances how much the model preserves the content of input images and the information of the style image, whereas λ_{cf} pushes the similarity of the intermediate feature on V extracted features. Fig. I-1 a) shows that accuracy improves when λ_a increases until it reaches a value of 0.1. Then, the performance begins to decline. Similar results can be observed when λ_{cf} varies from 1 to 10 (Fig. I-1 b)). The performance of our proposed AGPI^2 strategies varies depending on these hyperparameter values, but most results exceed the state-of-the-art methods.



a) Accuracy vs. λ_{adv} value. b) Accuracy vs. λ_{cf} value.

Figure-A I-1 Impact on accuracy of λ_{adv} and λ_{cf} hyperparameters on the SYSU-MM01 dataset under the single-shot setting

Table-A I-1 Impact of using generated images as intermediate versus using as visible in training or testing for SYSU dataset. AGPI^2 method uses them as intermediate in $\mathcal{L}_{dual}$ and CycleGAN(Zhu *et al.*, 2017) uses as visible modality and applies the $\mathcal{L}_{tri}$

Method	**Use**		**Performance**		**Complexity**	
	Train	**Test**	**R1(%)**	**mAP(%)**	**# of Para.**	Flops
Using generative images as intermediate in the training						
AGPI^2	✓	✗	72.23	70.58	24.8M	5.2G
AGPI^2	✓	✓	65.76	62.23	52.8M	8.3G
Using generative images as other modality in the t raining						
CycleGAN	✓	✗	42.05	41.17	24.8M	5.2G
CycleGAN	✓	✓	26.91	28.75	42.5M	7.1G

I.B.2 Comparison to CycleGAN

we conducted a new experiment to show the comparison between using intermediate and directly generated modalities in the training and inference phases separately. In Table. I-1, in the first two rows, the results for AGPI^2 are reported in which the generated images are used as an intermediate. The $\mathcal{L}_{dual}$ has been applied in the training process. In comparison, in the two last rows, the generated images are used as other modality (Visible images are replaced with generated images in training) and the $\mathcal{L}_{\text{tri}}$ is applied between I and generated images. In the second and fourth rows, the generated images are used as the gallery instead of the visible images. AGPI^2, which generates and uses images as an intermediate, performs better than CycleGAN(Zhu *et al.*, 2017) using generated images directly as another modality. We can conclude that although the transformed images try to lose style information from the source and achieve them from the target domain, they are likely to have some content information from the target and style from the source, so the provided privileged information by them is not enough for the model to compensate the other modality in the training.

Additionally, generating the transformed images in testing makes the model trigger the GAN module and increases the model complexity. Also, such generated images may lose some information compared to the original images and degrade the matching performance. These images are shown in Fig. I-2.

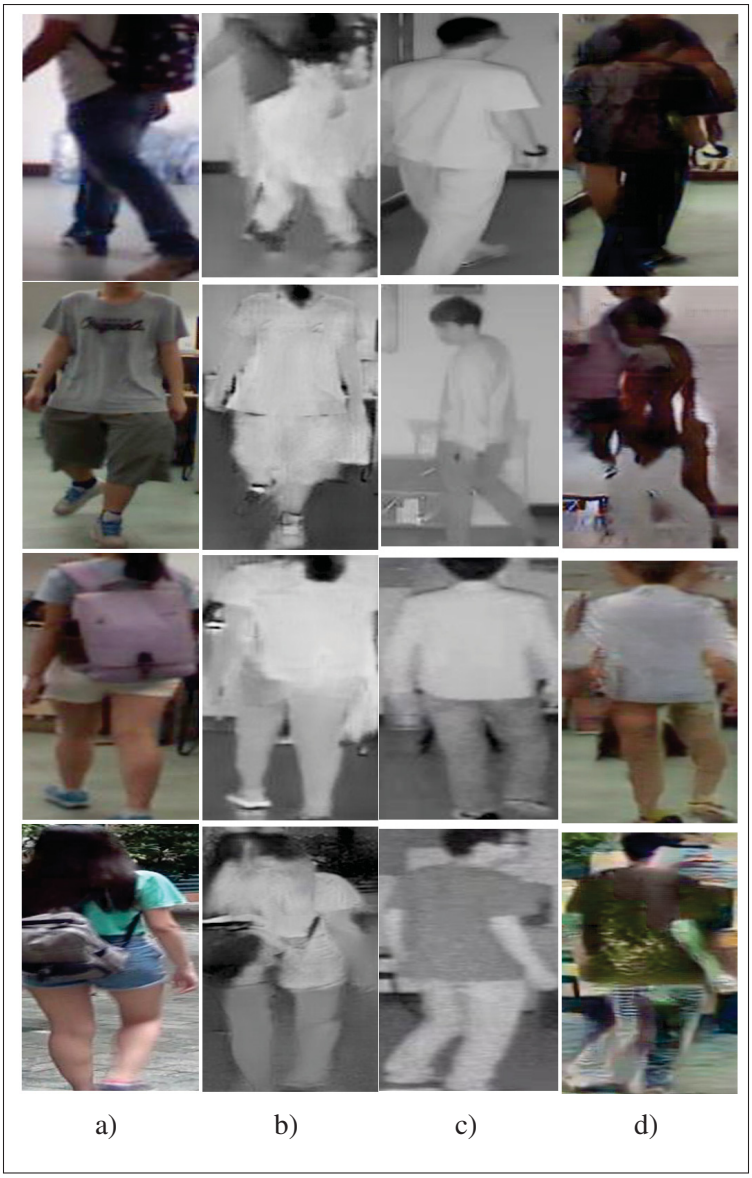


Figure-A I-2 Examples of images generated with CycleGAN(Zhu *et al.*, 2017) for SYSU-MM01 dataset. Column (a) contains the real V, and column (b) is the translated visible image to infrared. Columns (c) and (d) are the real I images and translated versions of them to the V domain

I.B.3 Feature Visualization

To demonstrate that the intermediate images in AGPI^2 training approach bridge between visible and infrared extracted features, we visualized the feature distributions of the intermediate,

Table-A I-2 MMD between V, I, and different intermediate learned features for 100 and 50 different persons from train and test sets, respectively, on the SYSU-MM01. Z images are the intermediate images reconstructed by $\mathcal{G}$

Modality	**Train**			**Test**		
	I-Z	**V-Z**	**I-V** (↓)	**I-Z**	**V-Z**	**I-V** (↓)
baseline	-	-	0.12	-	-	0.53
Grayscale (HAT)	0.07	0.03	0.08	0.46	0.09	0.45
Random(RPIG)	0.05	0.04	0.06	0.41	0.11	0.38
AGPI^2 (ours)	0.03	0.05	**0.05**	0.38	0.16	**0.33**

visible, and infrared representations during the training process. As shown in Fig. I-3, the intermediate features are positioned between the visible and infrared features. The visible features gradually align with the intermediate feature space when we apply $\mathcal{L}$cf and the first part of $\mathcal{L}$dual. As the model optimizes the second part of $\mathcal{L}_{\text{dual}}$, the infrared features progressively discard modality-specific information and converge toward the intermediate feature space. By the end of the training, the visible and infrared features had converged to a small region of the feature space, indicating that they had lost their modality-discriminative characteristics.

I.B.4 MMD Distance on extracted learned features:

The objective of the intermediate feature is to serve as a bridge between visible and infrared information by providing privileged identity-discriminative information while eliminating modality-specific style information. To demonstrate the reduction in the domain gap achieved through the use of intermediate features, we applied the kernel Maximum Mean Discrepancy (MMD) function to the extracted features from visible, infrared, and intermediate images. To measure MMD between modalities, we compute the center feature vector for each person within each modality and apply the MMD kernel to these vectors. Since, during the inference phase, infrared images are used as queries to the model in order to identify the matched person from the gallery, the distance between infrared (I) and intermediate (Z) features is crucial. As shown in Table I-2, the use of intermediate features significantly reduces this distance compared to the baseline.

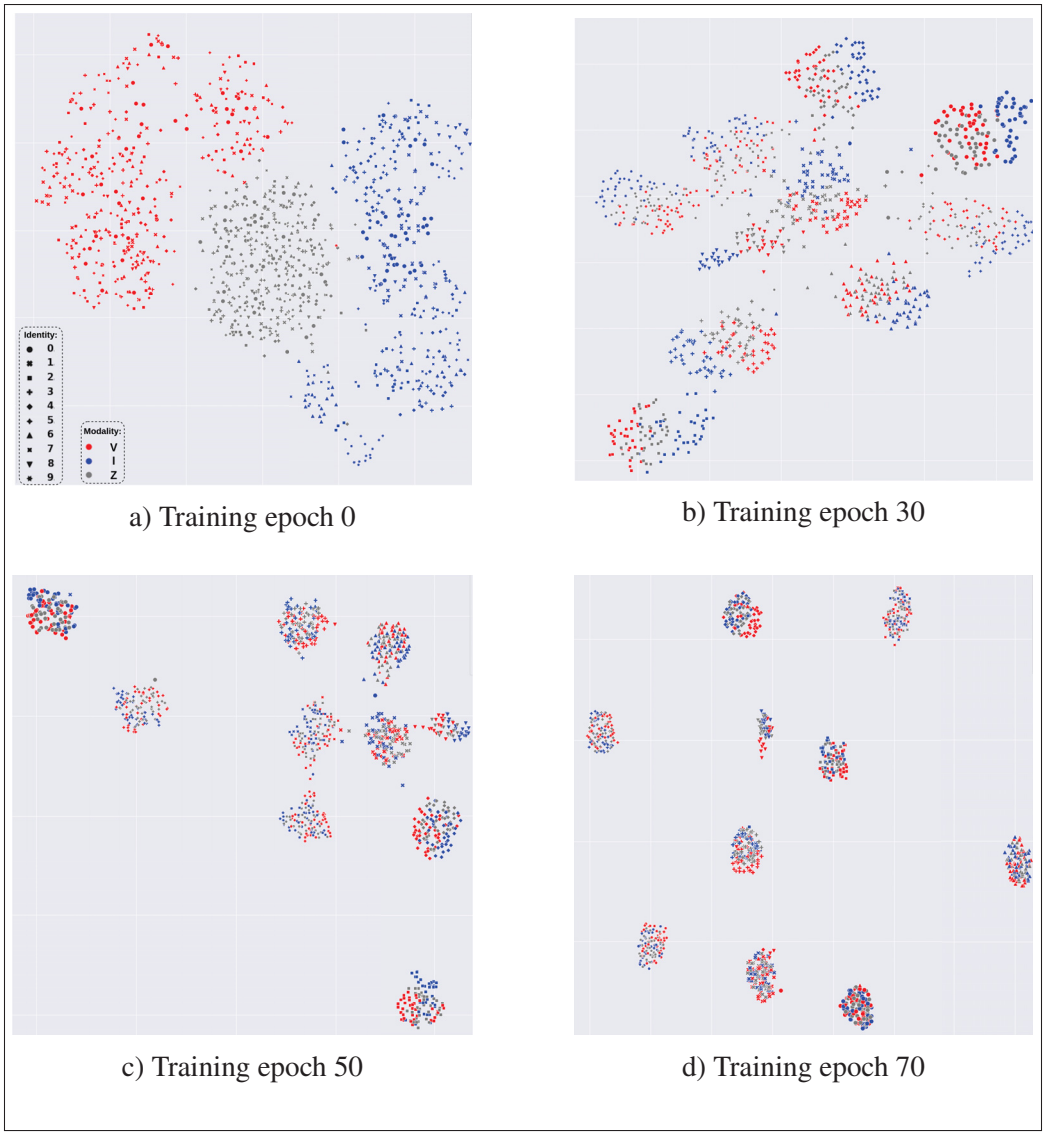


Figure-A I-3 UMAP (McInnes *et al.*, 2018) visual representation of ten randomly selected identities from the SYSU-MM01 dataset through training process. Identities are shown with different shapes, and modalities are shown with different colors (red and blue for visible and infrared and intermediate features are shown in gray color)

I.B.5 Comparing to Image-based Intermediate method in VI-ReID

We also conducted a new experiment showing the qualitative and quantitative comparison between CycleGAN(Zhu *et al.*, 2017), ThermalGAN(Kniaz *et al.*, 2018), D^{2}RL(Wang *et al.*, 2019d), HI-CMD(Choi *et al.*, 2020a) and MMN(Zhang *et al.*, 2021c) and our AGPI^2. We

used the official code published by the authors of these methods and trained them with default hyperparameters.

Table-A I-3 Impact of using generated images as intermediate versus using as visible in training or testing for SYSU dataset. AGPI^2 method uses them as intermediate in $\mathcal{L}_{dual}$ and CycleGAN(Zhu *et al.*, 2017) uses as visible modality and applies the $\mathcal{L}_{tri}$

Method	**Use**		**Performance**		**Complexity**	
	Train	**Test**	**R1(%)**	**mAP(%)**	**#Para.**	Flops
CycleGAN(Zhu *et al.*, 2017)	✓	✗	42.05	41.17	24.8M	5.2G
CycleGAN	✓	✓	26.91	28.75	42.5M	7.1G
thGAN(Kniaz *et al.*, 2018)	✓	✗	-	-	-	-
thGAN	✓	✓	17.77	18.13	70.0M	13.7G
AlignGAN(Wang *et al.*, 2019a)	✓	✗	24.81	22.05	23.5M	5.2G
AlignGAN	✓	✓	42.4	40.7	50.4M	9.6G
HiCMD(Choi *et al.*, 2020a)	✓	✓	34.94	35.94	40.1M	7.3G
HiCMD	✓	✓	15.57	17.38	100.6M	13.5G
MMN(Zhang *et al.*, 2021c)	✓	✗	65.14	61.57	72.9M	6.1G
MMN(Only Z)	✓	✓	62.96	58.34	35.4M	6.2G
MMN (Z,I,V)	✓	✓	70.1	66.2	72.9M	12.1G
AGPI^2	✓	✗	72.23	70.58	24.8M	5.2G
AGPI^2	✓	✓	65.76	62.23	52.8M	8.3G

Table I-3 presents a comparison of open-sourced generative image-based methods for Visible-Infrared Person Re-Identification (VI-ReID) on the SYSU-MM01 dataset. For each method, we report the performance in terms of Rank-1 accuracy and mAP, alongside the computational complexity measured by the number of parameters and FLOPs. The table highlights the differences between using generated images as an intermediate representation during training versus as a matching modality during testing.

The MMN (Zhang *et al.*, 2021c) approach utilizes generated intermediate images during both training and testing. Specifically, MMN extracts features from visible, infrared, and generated (Z) images, combining them for improved matching. While this strategy boosts performance, as evidenced by the 70.1% Rank-1 accuracy and 66.2% mAP in the (Z, I, V) setup, it significantly increases computational overhead, with a FLOPs count of 12.1G—almost double that of single-modality models.

In contrast, our proposed AGPI^2 method treats generated images as privileged information during training, incorporating them into the dual loss $\mathcal{L}_{dual}$, without altering the backbone for testing. This leads to significantly lower computational overhead, with only 8.3G FLOPs in the test phase, while maintaining competitive performance. Notably, when the generated images (Z) are used as a query instead of infrared images (I), AGPI^2 achieves a Rank-1 accuracy of 65.76%, outperforming MMN (62.96%), which demonstrates the informativeness and bridging capability of the generated images.

Furthermore, AGPI^2 not only delivers superior matching performance but also minimizes computational costs compared to methods like CycleGAN (Zhu *et al.*, 2017) and thermalGAN (Kniaz *et al.*, 2018), which require substantial resources due to their use of generated images during both training and testing. Overall, AGPI^2 balances effectiveness and efficiency, making it a compelling solution for VI-ReID tasks with constrained computational budgets.

Also, we generated some samples from the SYSU-MM01 dataset in Figure I-4 for each method. Our AGPI^2 produces more realistic images with more details compared to others.

I.B.6 Incorporating to Extra SOTA Models

Moreover, to compare with the requested method (PartMix (Kim *et al.*, 2023) and DEEN(Zhang & Wang, 2023b)), we apply our AGPI^2 approach to them and report the performance and efficacy. The PartMIX method (Kim *et al.*, 2023) utilizes part-based information by introducing an additional sub-module for part detection. To ensure a fair comparison, we integrated our AGPI^2 training approach into the SAAI method (Fang *et al.*, 2023), a publicly available part-based VI-ReID method, and retrained it from scratch. Similarly, AGPI^2 was added to the DEEN method (Zhang & Wang, 2023b) to evaluate its effectiveness on a non-part-based baseline. The results of these integrations on the SYSU-MM01 and RegDB datasets are shown in Table I-4.

For the part-based baseline, integrating AGPI^2 with part information leads to a substantial improvement in performance. Specifically, the combined AGPI^2 + Part approach achieves a notable increase in mAP for SYSU-MM01 (All Search) from 74.62% to 78.56% while

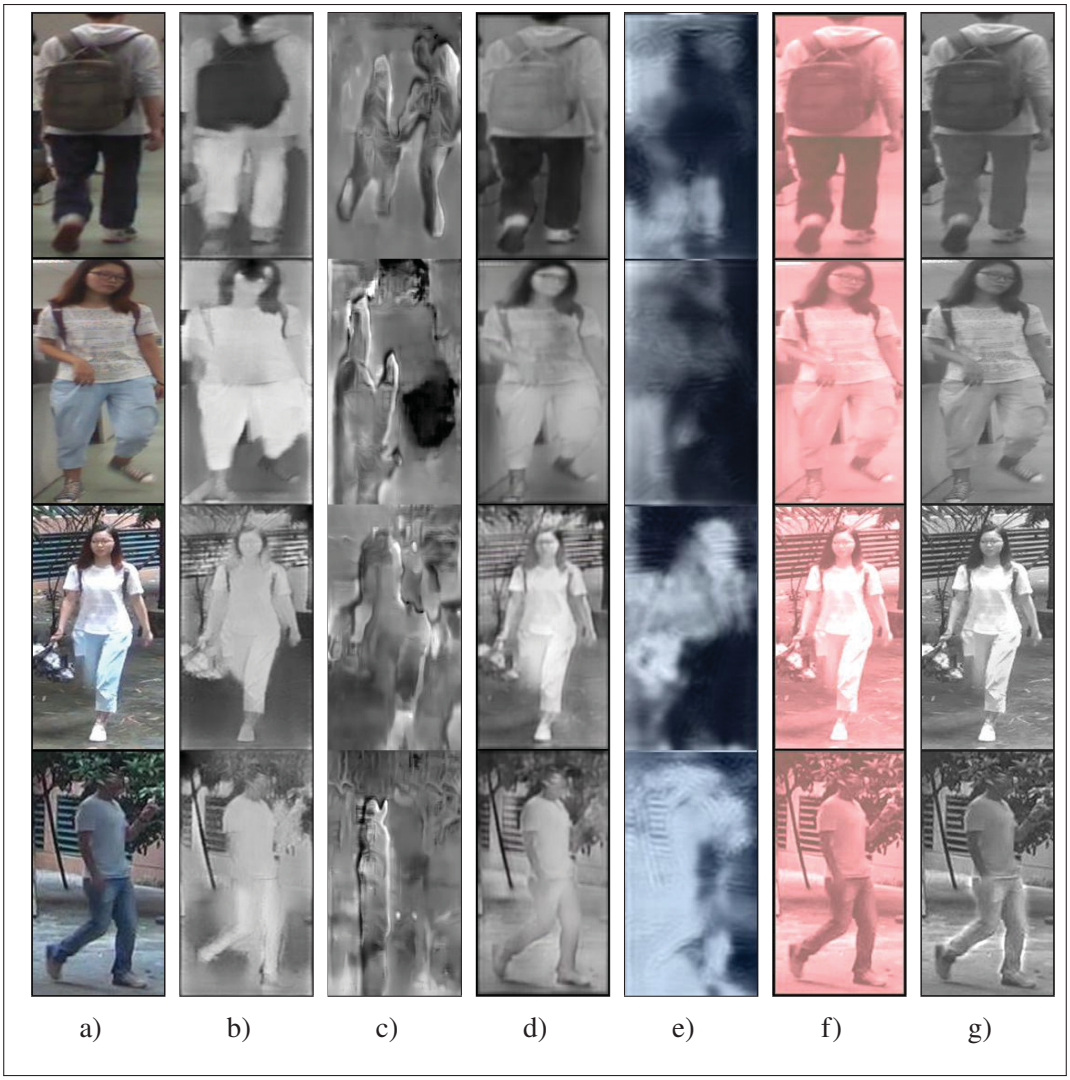


Figure-A I-4 Examples of 4 intermediate images generated from (a) Visible by (b)CycleGAN(Zhu *et al.*, 2017),(c) ThermalGAN (Kniaz *et al.*, 2018), (d) AlignGAN (Wang *et al.*, 2019a), (e) HiCMD (Choi *et al.*, 2020a), (f) MMN(Zhang *et al.*, 2021c) and (g) our AGPI^2 on the SYSU-MM01 dataset

maintaining competitive Rank-1 accuracy. On RegDB, the combined approach further improves cross-modal performance, with a significant boost in mAP for the I→V setting (from 82.52% to 91.15%). These results highlight the complementary nature of AGPI^2 when combined with part-based strategies, effectively enhancing both identity discrimination and modality bridging.

Table-A I-4 Accuracy of integrating the proposed AGPI^2 with different baselines on the SYSU-MM01 (single-shot setting) and RegDB datasets

Method	SYSU-MM01				RegDB				Complexity	
	All Search		Indoor Search		V → I		I → V		#Params	FLOPs
	R1	**mAP**	**R1**	**mAP**	**R1**	**mAP**	**R1**	**mAP**		
PartMix	77.78	74.62	81.52	84.34	84.93	82.52	85.66	82.27	52.5M	7.4G
AGPI^2 (ours)	72.23	70.58	83.45	84.25	89.03	83.89	87.91	83.04	24.8M	5.2G
AGPI^2 + Part	**77.35**	**78.56**	**84.60**	**88.91**	**91.22**	**91.15**	**91.3**	**92.5**	52.5M	7.4G
DEEN	74.7	71.8	80.3	83.3	91.1	85.8	89.5	83.4	89.0M	39.8G
AGPI^2 (ours)	72.23	70.58	**83.45**	84.25	89.03	83.89	87.91	83.04	24.8M	5.2G
DEEN + AGPI^2	**75.36**	**72.7**	83.19	**84.91**	**92.5**	**86.41**	**90.38**	**84.7**	89.0M	39.8G

For the DEEN baseline, AGPI^2 integration results in consistent improvements across all settings. For example, in SYSU-MM01 (All Search), the DEEN + AGPI^2 combination improves mAP from 71.8% to 72.7%, and on RegDB (V→I), mAP increases from 85.8% to 86.41%. These gains demonstrate the adaptability of AGPI^2 in improving feature representations across different baseline architectures. Notably, AGPI^2 achieves these enhancements with significantly lower complexity compared to other methods, as evidenced by its reduced parameter count and FLOPs.

These improvements demonstrate that AGPI^2 acts as a powerful enhancement tool when integrated with other models, further boosting their accuracy. Moreover, this integration does not introduce additional computational overhead during inference, making it a highly efficient solution.

APPENDIX II

SUPPLEMENTARY OF BMDG

II.A. Proofs

In Section 3.3 of the manuscript, our model seeks to represent input images using discriminative prototypes that are complementary. To ensure their complementarity, we minimize the mutual information (MI) between the data distributions of each pair of prototypes using a contrastive loss between feature representations of the different prototypes. To improve the discrimination, the prototype representations encode ID-related information by maximizing the MI between the joint distribution among all prototypes in the label distribution space. In this section, we prove that by minimizing the cross-entropy loss between features of each prototype class and the person ID in images, we can learn discriminative prototype representations.

II.A.1 Maximizing MI$(P^1, \ldots, P^K; Y)$

This section provides a proof that MI$(P^1, \ldots, P^K; Y)$ (Eq. 3.12) can be lower-bounded by :

$$\text{MI}(P^1; Y) + \cdots + \text{MI}(P^K; Y), \tag{A II-1}$$

following the properties of mutual information.

P.1 (Nonnegativity) For every pair of random variables X and Y:

$$\text{MI}(X; Y) \geq 0 \tag{A II-2}$$

P.2 For every pair random variables X, Y that are independent:

$$\text{MI}(X; Y) = 0. \tag{A II-3}$$

P.3 (Monotonicity) For every three random variables X, Y and Z:

$$\mathrm{MI}(X;Y;Z) \leq \mathrm{MI}(X;Y) \tag{A II-4}$$

P.4 For every three random variables X, Y and Z, the mutual information of joint distortions X and Z to Y is:

$$\mathrm{MI}(X,Z;Y) = \mathrm{MI}(X;Y) + \mathrm{MI}(Z;Y) - \mathrm{MI}(X;Z;Y) \tag{A II-5}$$

$$MI(P^1,\ldots,P^K;Y). \tag{A II-6}$$

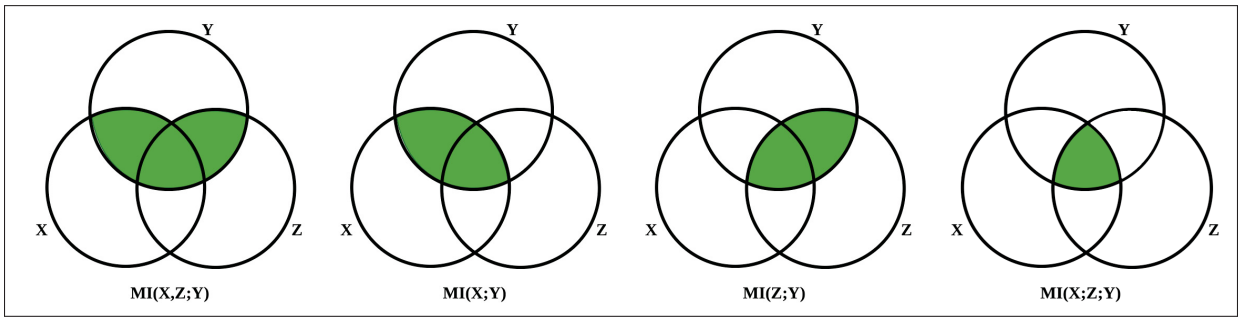


Figure-A II-1 Venn diagram of theoretic measures for three variables X, Y, and Z, represented by the lower left, upper, and lower right circles, respectively

Theorem II.1. *Let $P^1,\ldots,P^K$ and Y be random variables with domains $\mathcal{P}^1,\ldots,\mathcal{P}^K$ and $\mathcal{Y}$, respectively. Let every pair P^k and P^q $(k \neq q)$ be independent. Then, maximizing $MI(P^1,\ldots,P^K;Y)$ can be approximated by maximizing the sum of MI between each of P^k to Y, $\sum_{k=1}^{K} MI(P^k;Y)$.*

Proof. First, we define $\tilde{P}^k$ as the joint distribution of $P^{k+1},\ldots,P^K$. Using the **P.4** we have:

$$\mathrm{MI}(P^1,\tilde{P}^1;Y) = \underbrace{\mathrm{MI}(P^1;Y)}_{\alpha} + \underbrace{\mathrm{MI}(\tilde{P}^1;Y)}_{\beta} - \underbrace{\mathrm{MI}(P^1;\tilde{P}^1;Y)}_{\gamma}, \tag{A II-7}$$

To maximize this, α and β should be maximized and γ minimized. Given **P.3**, $\min \mathrm{MI}(P^1; \tilde{P}^1; Y)$ can be upper-bounded by $\min \mathrm{MI}(P^1; \tilde{P}^1)$:

$$\mathrm{MI}(P^1; \tilde{P}^1; Y) \leq \mathrm{MI}(P^1; \tilde{P}^1), \tag{A II-8}$$

So by minimizing the right term of Eq. A II-8, γ is also minimized. We show that $\mathrm{MI}(P^1; \tilde{P}^1) = 0$ by expanding $\tilde{P}^1$ to $(P^2, \tilde{P}^2)$:

$$\mathrm{MI}(P^1; P^2, \tilde{P}^2) = \mathrm{MI}(P^1; P^2) + \mathrm{MI}(P^1; \tilde{P}^2) - \mathrm{MI}(P^1; P^2; \tilde{P}^2). \tag{A II-9}$$

Given **P.1** and Eq. 3.5, $\mathrm{MI}(P^1; P^2) = 0$ and $\mathrm{MI}(P^1; P^2; \tilde{P}^2) \leq \mathrm{MI}(P^1; P^2) = 0$ so that:

$$\mathrm{MI}(P^1; P^2, \tilde{P}^2) = \mathrm{MI}(P^1; \tilde{P}^2). \tag{A II-10}$$

After recursively expanding $\tilde{P}^2$:

$$\mathrm{MI}(P^1; P^2, \tilde{P}^2) = 0. \tag{A II-11}$$

To maximize β, we can rewrite and expand recursively in Eq. A II-7 :

$$\mathrm{MI}(\tilde{P}^1; Y) = \mathrm{MI}(P^2, \tilde{P}^2; Y) = \underbrace{\mathrm{MI}(P^2; Y)}_{\text{maximizing}} + \underbrace{\mathrm{MI}(\tilde{P}^2; Y)}_{\text{expanding}} - \underbrace{\mathrm{MI}(P^2; \tilde{P}^2; Y)}_{0}. \tag{A II-12}$$

Therefore, it can be shown that:

$$\mathrm{MI}(\tilde{P}^k; Y) = \mathrm{MI}(P^{k+1}, \tilde{P}^{k+1}; Y) = \mathrm{MI}(P^k; Y) + \cdots + \mathrm{MI}(P^K; Y) \quad \forall k \in \{0, \ldots, K-1\}. \tag{A II-13}$$

and for $k = 0$, we have:

$$\mathrm{MI}(P^1, \ldots, P^K; Y) = \sum_{k=1}^{K} \mathrm{MI}(P^k; Y), \tag{A II-14}$$

Finally, for maximizing $\mathrm{MI}(P^1, \ldots, P^K; Y)$, we need to maximize each $\mathrm{MI}(P^k; Y)$ so that each prototype feature contains Id-related information and is complemented. In other words, each prototype seeks to describe the input images from different aspects. □

II.A.2 Maximizing MI$(P^k; Y)$

In Section 3.3, the MI between the representation of each prototype and the label of persons are maximized by minimizing cross-entropy loss (see Eq: 3.13 of the manuscript). This approximation is formulated as **Proposition III.1**.

Proposition II.1. *Let P^k and Y be random variables with domains $\mathcal{P}^k$ and $\mathcal{Y}$, respectively. Minimizing the conditional cross-entropy loss of predicted label $\hat{Y}$, denoted by $\mathcal{H}(Y; \hat{Y}|P^k)$, is equivalent to maximizing the MI$(P^k; Y)$*

Proof. Let us define the MI as entropy,

$$\mathrm{MI}(P^k; Y) = \underbrace{\mathcal{H}(Y)}_{\delta} - \underbrace{\mathcal{H}(Y|P^k)}_{\xi} \tag{A II-15}$$

Since the domain $\mathcal{Y}$ does not change, the entropy of the identity δ term is a constant and can therefore be ignored. Maximizing $\mathrm{MI}(P^k, Y)$ can only be achieved by minimizing the ξ term. We show that $\mathcal{H}(Y|P^k)$ is upper-bounded by our cross-entropy loss (Eq. 3.13), and minimizing such loss results in minimizing the ξ term. By expanding its relation to the cross-entropy (Boudiaf *et al.*, 2020):

$$\mathcal{H}(Y; \hat{Y}|P^k) = \mathcal{H}(Y|P^k) + \underbrace{\mathcal{D}_{\mathrm{KL}}(Y||\hat{Y}|P^k)}_{\geq 0}, \tag{A II-16}$$

where:

$$\mathcal{H}(Y|P^k) \leq \mathcal{H}(Y; \hat{Y}|P^k). \tag{A II-17}$$

Through the minimization of Eq. 3.13, training can naturally be decoupled in 2 steps. First, weights of the prototype module are fixed, and only the classifier parameters (i.e., weight W^k of the fully connected layer) are minimized w.r.t. Eq. A III-13. Through this step, $\mathcal{D}_{\mathrm{KL}}(Y||\hat{Y}|P^k)$ is minimized by adjusting $\hat{Y}$ while the $\mathcal{H}(Y|P^k)$ does not change. In the second step, the

prototype module's weights are minimized w.r.t. $\mathcal{H}(Y|P^k)$, while the classifier parameters W^k are fixed. □

II.B. Additional Details on the proposed method

II.B.1 Training Algorithm

To train the model, BMDG uses a batch of data containing N_b person with N_p positive images from the V and I modalities. Algorithm II-1 shows the details of the BMGD training strategy for optimizing the feature backbone by gradually increasing mixing prototypes.

At first, the prototype mining module extracts K prototypes from infrared and visible images in lines 4 and 5 . Then, at lines 6 and 7, $\mathcal{G}$ function mixes these prototypes from each modality to create two intermediate features. It is noted that the ratio of mixing gradually increases w.r.t the step number t to create more complex samples. To refine the final feature descriptor for input images, the attentive embedding module, $\mathcal{F}$, is applied to prototypes to leverage the attention between them. At the end of each iteration, the model's parameters will be optimized by minimizing the cross-modality ReID objectives between each modality features vector and its gradually created intermediate.

II.B.2 Attentive Prototype Embedding

Details of the Attentive Prototype Embedding module are depicted in II-2.

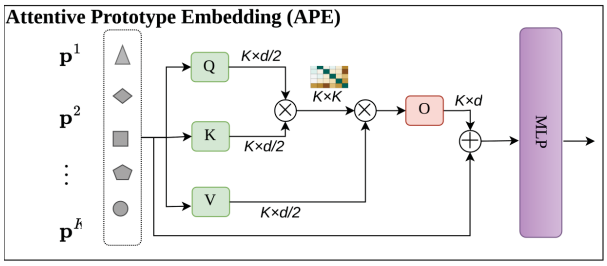


Figure-A II-2 Attentive prototype embedding (APE) architecture

Algorithm-A II-1 BMDG Training Strategy

Input: $\mathcal{S} = \{\mathcal{V}, \mathcal{I}\}$ as training data and T, K as hyper-parameters
for $t = 1, \ldots, T$ **do**
  while *all batches are not selected* **do**
    $x_v^j, x_i^j \leftarrow$ **batchSampler**(N_b, N_p) ;
    extract prototypes $\mathbf{A}_v^j$ and global features $\mathbf{g}_v^j$ from visible images v^j /* left-side of Fig. 4.2(a) */
    extract prototypes $\mathbf{A}_i^j$ and global features $\mathbf{g}_i^j$ from infrared images i^j /* left-side of Fig. 4.2(a) */
    $\mathbf{A}_v^{(t)} \leftarrow \mathcal{G}(\mathbf{A}_v^j, \mathbf{A}_i^j, t)$ /* V intermediate by gradually increasing the mixing ratio w.r.t t from I prototypes */
    $\mathbf{A}_i^{(t)} \leftarrow \mathcal{G}(\mathbf{A}_i^j, \mathbf{A}_v^j, t)$ /* I intermediate by gradually increasing the mixing ratio w.r.t t from V prototypes */
    if $t < T$ **then**
      $\mathbf{f}_v^{(t)} \leftarrow [\mathcal{F}(\mathbf{A}_v^{(t)}); \mathbf{g}_v]$ /* embedding intermediate visible features */
      $\mathbf{f}_i^{(t)} \leftarrow [\mathcal{F}(\mathbf{A}_i^{(t)}); \mathbf{g}_i]$ /* embedding intermediate infrared features */
    else
      if $t = T$ **then**
        $\mathbf{f}_v^{(t)} \leftarrow [\mathcal{F}(\mathbf{A}_v^{(t)}); \mathbf{g}_i]$;
        $\mathbf{f}_i^{(t)} \leftarrow [\mathcal{F}(\mathbf{A}_i^{(t)}); \mathbf{g}_v]$;
      end if
    end if
    update model's parameters by optimizing Eq. 5.7 ;
  end while
end for

II.C. Additional Details on the Experimental Methodology

II.C.1 Datasets:

Research on cross-modal V-I ReID has extensively used the SYSU-MM01 (Wu *et al.*, 2020), RegDB (Nguyen *et al.*, 2017), and recently published LLCM (Zhang & Wang, 2023b) datasets. SYSU-MM01 is a large dataset containing more than 22K RGB and 11K IR images of 491 individuals captured with 4 RGB and 2 near-IR cameras, respectively. Of the 491 identities,

395 were dedicated to training, and 96 were dedicated to testing. Depending on the number of images in the gallery, the dataset has two evaluation modes: single-shot and multi-shot. RegDB contains 4,120 co-located V-I images of 412 individuals. Ten trial configurations randomly divide the dataset into two sets of 206 identities for training and testing. The tests are conducted in two ways – comparing I to V (query) and vice versa. An LLCM dataset consists of a large, low-light, cross-modality dataset that is divided into training and testing sets at a 2:1 ratio.

II.C.2 Experimental protocol:

We used a pre-trained ResNet50 (He *et al.*, 2016b) as the deep backbone model. Each batch contains 8 RGB and 8 IR images from 10 randomly selected identities. Each image input is resized to 288 by 144, then cropped and erased randomly, and filled with zero padding or mean pixels. ADAM optimizer with a linear warm-up strategy was used for the optimization process. We trained the model by 180 epochs, in which the initial learning rate is set to 0.0004 and is decreased by factors of 0.1 and 0.01 at 80 and 120 epochs, respectively. $K = 6$, $T = 4$, $\lambda_f = 0.1$, $\lambda_v = 0.05$, $\lambda_p = 0.2$ and $\lambda_i = 0.4$ are set based on the analyses shown in the ablation study in the main paper and in Section II.D.1. $\lambda_{eq} = 0.5$ is for all experiments.

II.C.3 Performance measures:

We use Cumulative Matching Characteristics (CMC) and Mean Average Precision (mAP) as assessment metrics in our study. In CMC, rank-k accuracy is measured to determine how likely it is that a precise cross-modality image of the person will be present in the top-k retrieved results. As an alternative, mAP can be used as a measure of image retrieval performance when multiple matching images are found in a gallery.

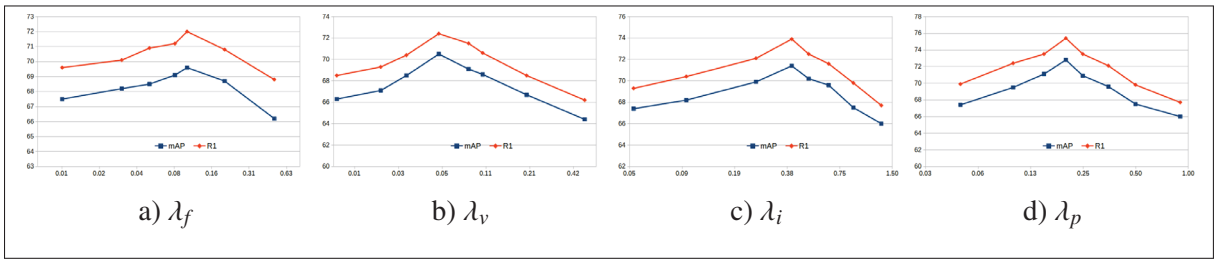
II.D. Additional Quantitative Results

This section provides more quantitative and qualitative results to validate our proposed BMDG method.

II.D.1 Hyperparameter values:

This subsection analyzes the impact of λ_f, λ_v, λ_p, and λ_i on V-I ReID accuracy. We initially set $\lambda_v = 0.01$, $\lambda_p = 0.05$, and $\lambda_i = 0.8$, experimenting with various values for λ_f. As shown in Fig. II-3, accuracy improves with increasing λ_f until it reaches 0.1. Elevated λ_f enhances prototype diversity, boosting the discriminative ability of final features in the diverse space. However, excessively high values disperse prototype features in the feature space, diminishing discriminability and hindering accurate identification.

Similar trends are observed when $\lambda_p = 0.05$ and $\lambda_i = 0.8$, varying λ_v from 0.01 to 0.05, resulting in improved performance. Higher λ_v compresses prototype regions excessively, lacking sufficient ID-related information. Conversely, λ_i enhances the discriminative capabilities of prototypes in images. Balancing these factors, we find optimal values of 0.05 and 0.4 for λ_v and λ_i, respectively. Additionally, based on experimentation, we set $\lambda_p = 0.2$ at the end of our analysis.



a) λ_f b) λ_v c) λ_i d) λ_p

Figure-A II-3 Accuracy of the proposed BMDG over λ_f, λ_v, λ_i, and λ_p values on SYSU-MM01 dataset in all-search and single-shot mode

II.D.2 Step size and number of part prototypes:

In section 3.4.2, we discussed the step size and number of part prototypes based on Rank-1 accuracy. Here we report the mAP measurement for Table 3.4 in paper in Table II-1a in paper in Table II-1b, respectively:

Table-A II-1 mAP accuracy of BMDG using (a) our prototype mixing and (b) Mixup (Zhang *et al.*, 2018) setting for different numbers of part prototypes (K) and intermediate steps (T)

(a) Prototype exchanging

T	**Number of part prototypes (K)**					
	3	4	5	6	7	10
0	65.98	67.03	67.23	68.11	67.55	65.66
1	67.42	68.72	69.28	69.46	68.32	69.28
2	69.51	70.08	70.72	71.14	69.97	71.25
3	71.44	71.69	71.82	72.02	71.06	71.67
4	-	71.98	72.15	**72.86**	71.19	69.00
6	-	-	-	72.40	71.17	69.54
10	-	-	-	-	-	69.46

(b) Mixup (Zhang *et al.*, 2018)

T	**Number of part prototypes (K)**					
	3	4	5	6	7	10
0	65.98	67.03	67.23	68.11	67.55	65.66
1	66.31	67.22	67.31	68.28	68.5	65.8
4	66.50	67.68	67.79	68.52	68.49	65.88
6	66.98	67.77	68.05	**68.73**	68.56	66.02
10	67.04	67.60	67.93	68.65	68.54	65.57

II.D.3 Model efficiency:

Our BMDG proposed a method to extract alignable part-prototypes in feature extraction and then compute an attention embedding for the final representation features. In Table II-2, we showed each component size and time complexity in the inference time compared to the baseline we used.

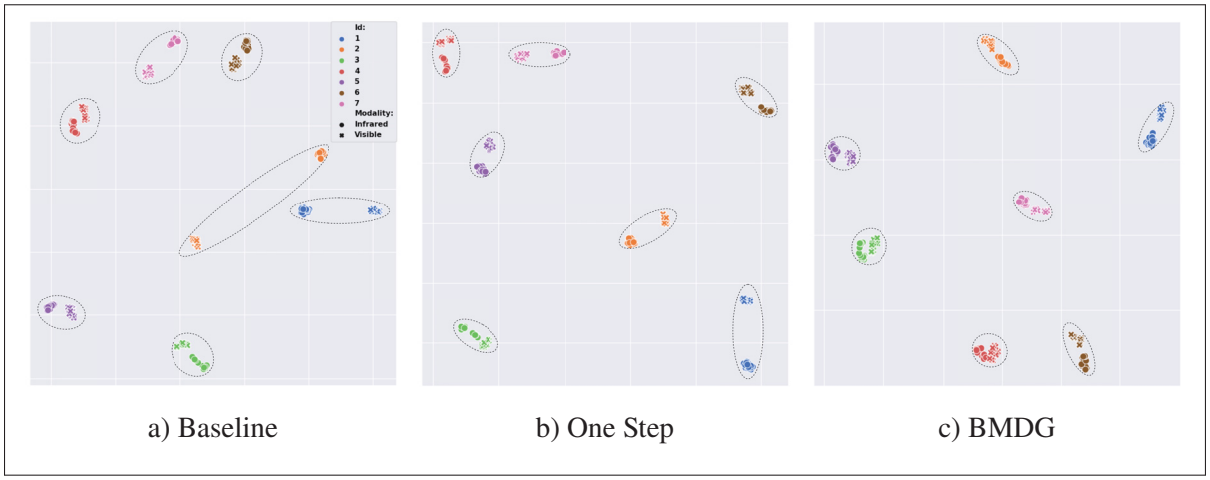
Table-A II-2 Number parameters and floating-point operations at inference time for BMDG and all its sub-modules

Model	**# of Para. (M)**	**Flops (G)**
Feature Backbone	24.8	5.2
Prototype Mining	3.1	0.2
Attentive Prototype Embedding	1.8	0.3
baseline (Ye *et al.*, 2020b)	24.9	5.2
BMDG	29.7	5.7

II.E. Visual Results

II.E.1 UMAP projections:

To show the effectiveness of BMDG, we randomly select 7 identities from the SYSU-MM01 dataset and project their feature representations using the UMAP method (McInnes *et al.*, 2018) for (a) Baseline, (b) one-step (prototypes without gradual training), and (c) BMDG. Visualization results (Figure III-3) show that compared with the baseline and one-step approach, the feature representations learned with our BMDG method are well clustered according to their respective identity, showing a strong capacity to discriminate. BMDG is effective for learning robust and identity-aware features. Our BMDG method reduces this distance across modalities for each person and provides more separation among samples from different people.



a) Baseline b) One Step c) BMDG

Figure-A II-4 Distributions of learned V and infrared features of 7 identities from SYSU-MM01 dataset for (a) the baseline, (b) one-step using part-prototypes, and (c) our BMDG method by UMAP (McInnes *et al.*, 2018). Each color shows the identity

Also, to show how the intermediate features gradually mix the modalities, we draw intermediate features for 6 steps in Figure II-5. At the beginning of training, the features are based on modality while at step 6, the features are concentrated on each identity.

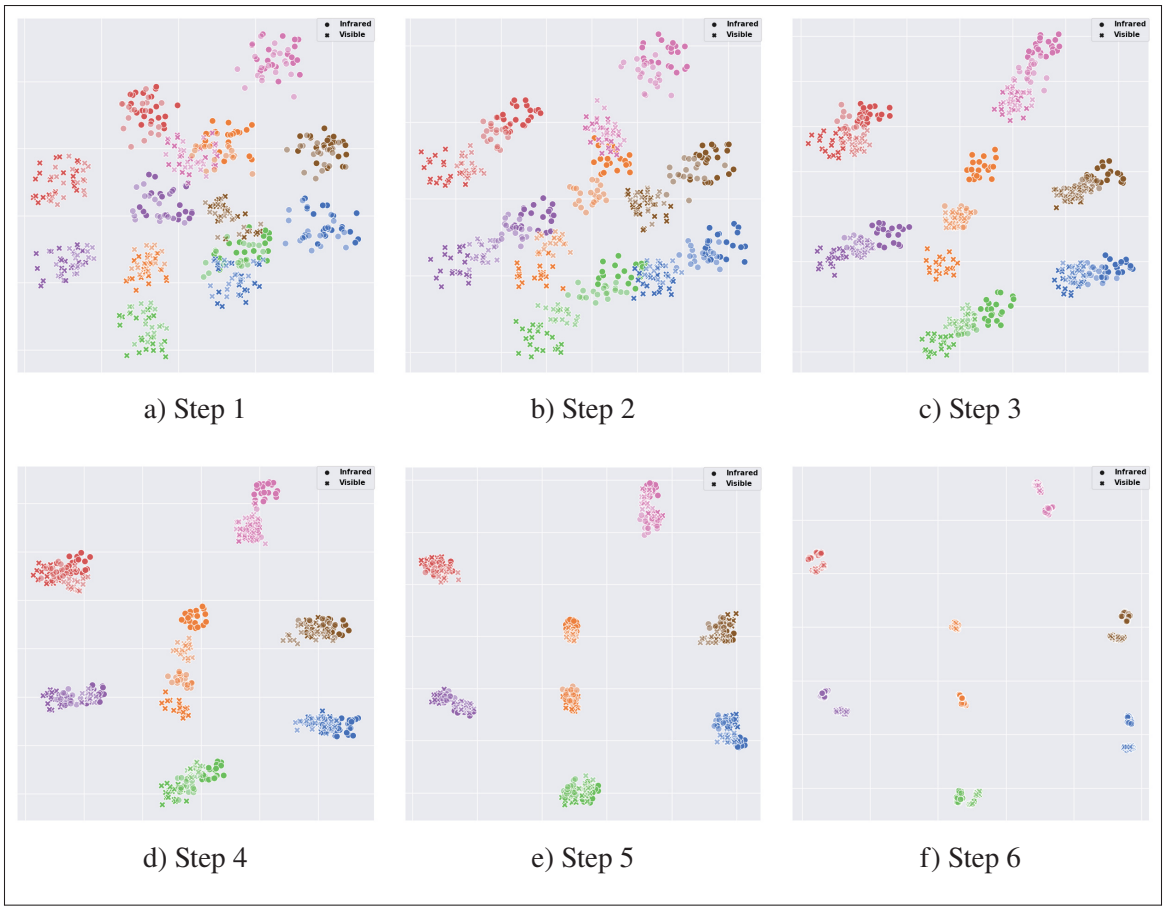
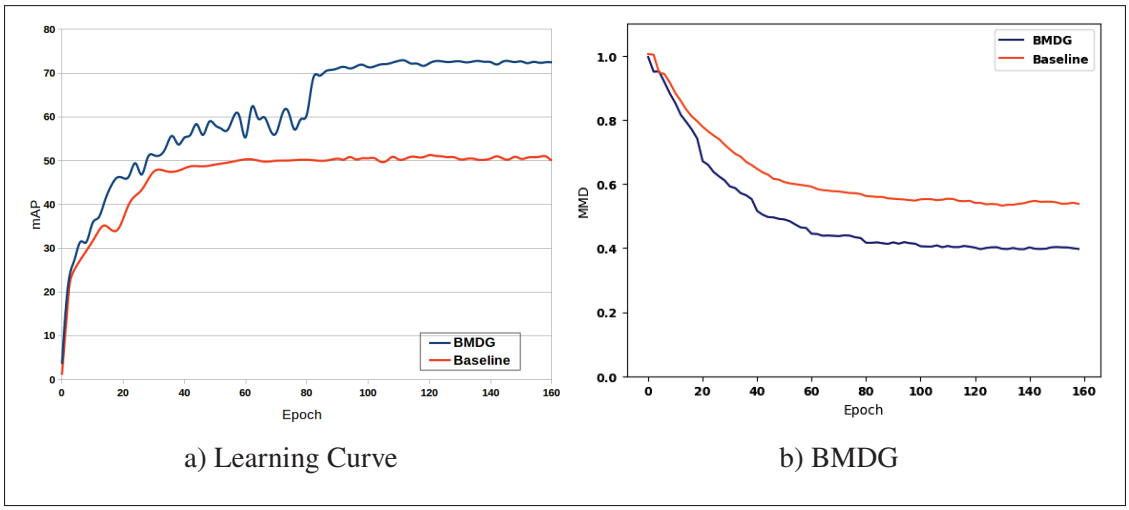


Figure-A II-5 Distributions of learned V and infrared features of 7 identities from the SYSU-MM01 dataset in training for 6 steps at epochs 10, 30, 50, 70, 90, and 160, respectively, visualized by UMAP (McInnes *et al.*, 2018). Each color shows one identity; intermediate features are drawn with lower opacity

II.E.2 Domain shift:

To estimate the level of domain shift over data from V and I modalities, we measured the MMD distance for each training epoch. To this end, for each epoch, we selected 10 random images from 50 random identities and extracted the prototype and global features, then measured the MMD distance between the centers of those features for each modality as shown in Fig. II-6(b). We report the normalized MMD distances between I and V features for our BMDG approach when compared with the baseline. Our method reduces this distance more than the y baseline.

Thus, the results show that the intermediate domains improve the model robustness to a large multi-modal domain gap by gradually increasing the mix in prototypes over multiple steps.



a) Learning Curve b) BMDG

Figure-A II-6 The mAP and domain shift (MMD distance) between I and V modalities over training epochs. (a) The learning curve of BMDG vs Baseline(Ye *et al.*, 2020b). (b) MMD distance over the center of multiple person's infrared features to visible modality

II.E.3 Part-prototype masking:

To show spatial information related to prototype features, we visualize the score map in the PM module (see Fig. II-7). Our approach encodes prototype regions linked to similar body parts without considering person identity. Our model tries to find similar regions for each class of prototypes and then extracts ID-related information for that region. Therefore, BMDG is more robust for matching the same part features.

II.E.4 Semi-supervised body part detection:

An additional benefit of our HCL module lies in its ability to detect meaningful parts in a semi-supervised manner. By forcing the model to identify semantic regions that are both informative about foreground objects and contrastive to each other, our hierarchical contrastive learning provides robust part detection, even in the absence of part labels. To assess HCL, we fine-tuned our ReID model as a student using a pre-trained part detector (Li, Xu, Wei & Yang,

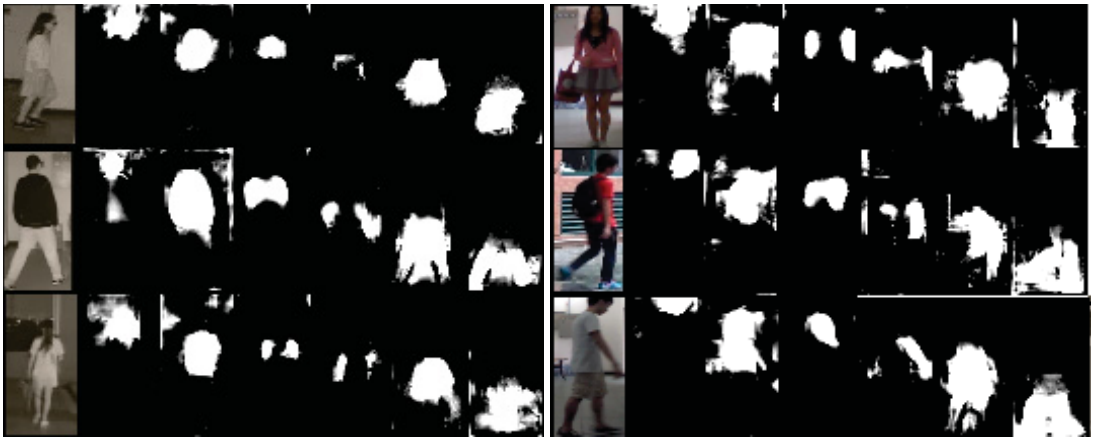


Figure-A II-7 Prototypes regions extracted by the PRM module for (a) infrared and (b) visible images. Note that the mask size is 18×9, which is then resized to fit the original input image. As shown, the mask of prototypes focuses on similar body parts without accounting for identity

2020b) on the PASCAL-Part Dataset (Chen *et al.*, 2014) as the teacher. In Fig. II-8, the results show our model's strong capacity for detecting body parts compared to its teacher.

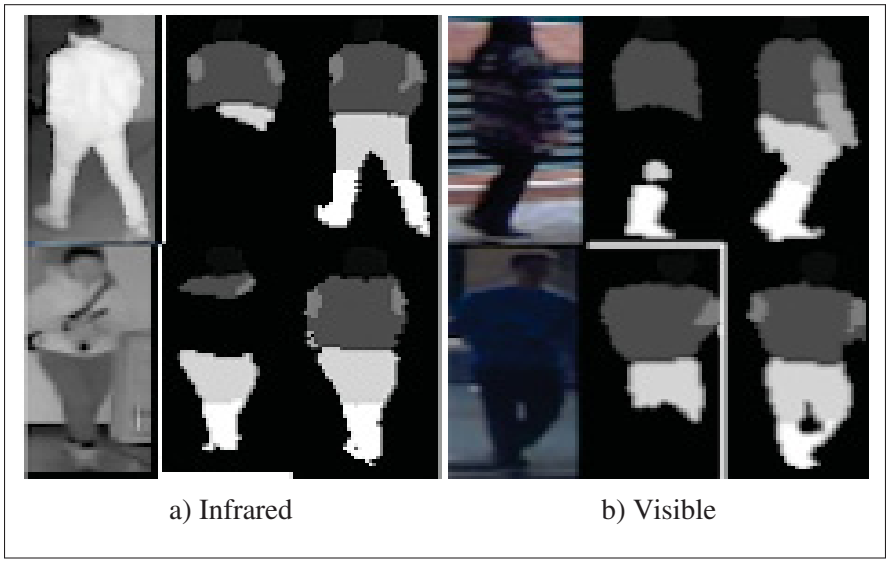


Figure-A II-8 Semi-supervised part discovery on the SYSU-MM01 dataset. The first columns are the (a) infrared and (b) visible images. The second column images are the result from (Li *et al.*, 2020b), and the last column are results with our fine-tuned model in BMDG

APPENDIX III

SUPPLEMENTARY OF MIXER

III.A. Proofs:

In Section 4.3, our model is designed to represent input images through two distinct and complementary feature types: modality-related and modality-erased representations. To ensure the independence of these representations, we minimize the mutual information (MI) between the data distributions of them. This is achieved by applying an orthogonal loss to their feature representations. The modality-related features are intended to capture ID-aware information specific to the modality, whereas the modality-erased features are explicitly designed to exclude any modality-specific information. To ensure that the extracted features satisfy these conditions, we use constrained optimization to maximize the mutual information between the joint distribution of modality-related and modality-erased features and the label distribution. In this section, we prove that our proposed loss functions meet these constraints and align with the properties of mutual information.

III.A.1 Backgrounds

The following highlights the main properties of mutual information:

P 1 (**Nonnegativity).** *For every pair of random variables X and Y:*

$$MI(X;Y) \geq 0 \tag{A III-1}$$

P 2. *For random variables X, Y that are independent:*

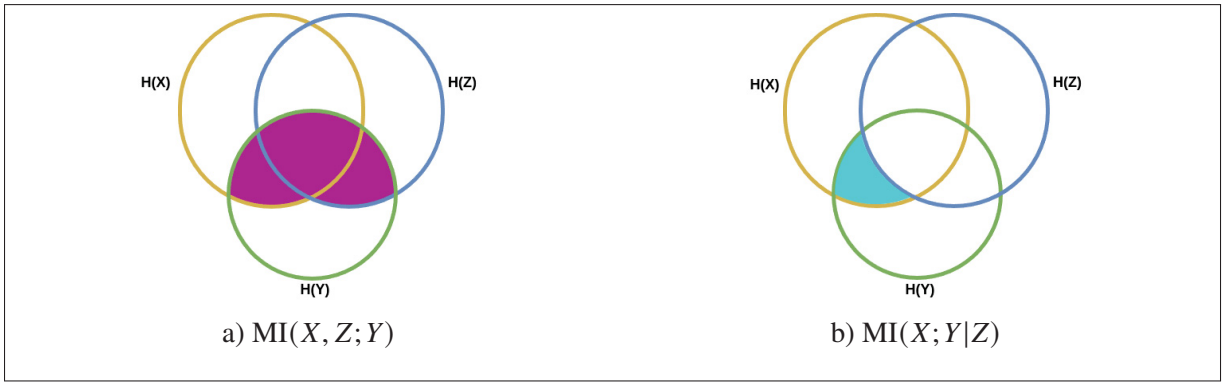
$$MI(X;Y) = 0. \tag{A III-2}$$

P 3 (**Monotonicity**). *For every three random variables X, Y and Z:*

$$MI(X;Y;Z) \leq MI(X;Y) \tag{A III-3}$$

P 4. *For every three random variables X, Y and Z, the mutual information of joint distribution X and Z to Y is (Fig. III-1a):*

$$MI(X,Z;Y) = MI(X;Y) + MI(Z;Y) - MI(X;Z;Y) \tag{A III-4}$$



a) MI$(X,Z;Y)$ b) MI$(X;Y|Z)$

Figure-A III-1 Venn diagram of theoretic measures for three variables X, Y, and Z, represented by the left, right and bottom circles, respectively

P 5 (**Conditional**). *For every three random variables X, Y and Z, the conditional mutual information is(Fig. III-1b):*

$$MI(X;Y|Z) = MI(X;Y) - MI(X;Y;Z) \tag{A III-5}$$

Also, we described some hypotheses that are used in our learning: - Following (Feng *et al.*, 2023), the orthogonality between $\mathbf{z}_m^r$ and $\mathbf{z}_m^e$ can be regarded as a relaxation of independence:

$$\forall(\mathbf{z}_m^r, \mathbf{z}_m^e) \sim (Z_m^r, Z_m^e),\ \mathbf{z}_m^r, \mathbf{z}_m^e \perp \mathbf{z}_m^e \implies \mathrm{MI}(Z_m^r; Z_m^e) \simeq 0$$

- Following (Feng *et al.*, 2023; Achille & Soatto, 2018), we posit that if Z is a representation of X, then Z is conditionally independent of all other variables in the system given X. This is formally expressed as:

$$\forall A, B \quad I(A; Z \mid X, B) = 0. \tag{A III-6}$$

Definition III.1 (**Sufficiency).** *A representation Z of X is sufficient for Y if and only if:*

$$MI(X; Y|Z) = 0 \iff MI(X; Y) = MI(Z; Y) \tag{A III-7}$$

Any model with access to a sufficiently informative representation Z must be able to predict Y with at least the same level of accuracy as if it had access to the original data X. Representation Z is considered sufficient for Y if and only if the task-relevant information remains unchanged during the encoding process. If the cross-entropy loss between Z and Y is minimized, then, as suggested in (Achille & Soatto, 2018), Z can be assumed to be a sufficient representation of X for the task Y.

Definition III.2 (**Data Processing Inequality).** *Let three random variables form the Markov chain $Y \rightarrow X \rightarrow Z$, implying that the conditional distribution of Z depends only on X and is conditionally independent of X, we have:*

$$MI(X; Y) \geq MI(Z; Y). \tag{A III-8}$$

The data processing inequality (DPI) is a fundamental inequality in information theory that states the mutual information between two random variables cannot increase through processing.

Theorem III.1. *If Z_m^e and Z_m^r are independent, then $MI(Z_m^e, Z_m^r; Y) \geq MI(Z_m^e; Y) + MI(Z_m^r; Y)$. In other words, $MI(Z_m^e; Y) + MI(Z_m^r; Y)$ is a lower bound of $MI(Z_m^e, Z_m^r; Y)$ and we can maximize it by maximizing its lower-bound.*

Proof. For independent random variables Z_m^e and Z_m^r, we have the following relationship:

$$\mathrm{MI}(Z_m^e; Z_m^r) = 0.$$

Also, w.r.t to P 4:

$$\mathrm{MI}(Z_m^r, Z_m^e; Y) = \mathrm{MI}(Z_m^r; Y) + \mathrm{MI}(Z_m^e; Y) - \mathrm{MI}(Z_m^r; Z_m^e; Y),$$

and P 3:

$$\mathrm{MI}(Z_m^r; Z_m^e; Y) \leq \mathrm{MI}(Z_m^r; Z_m^e) = 0,$$

so, we have:

$$\mathrm{MI}(Z_m^r; Z_m^e; Y) \leq 0 \tag{A III-9}$$

and,

$$\mathrm{MI}(Z_m^e, Z_m^r; Y) \geq \mathrm{MI}(Z_m^e; Y) + \mathrm{MI}(Z_m^r; Y) \tag{A III-10}$$

□

III.A.2 Minimizing MI$(Z_m^e; M)$

In order to minimizing $\mathrm{MI}(Z_m^e; M)$, we use adversarial learning approach and convert $Z_m^e \in \mathcal{Z}_m^e$ to $Z_m^o \in \mathcal{Z}_m^o$ with deterministic and learnable function $\mathcal{W} : \mathcal{Z}_m^e \to \mathcal{Z}_m^o$:

$$Z_m^o = \mathcal{W}(Z_m^e; \theta_W). \tag{A III-11}$$

Then, we maximize the $\mathrm{MI}(Z_m^o; M)$ with a standard SGD approach through the optimization of θ_W and reversed SGD through the model's parameters. Maximizing $\mathrm{MI}(Z_m^o; M)$ is achieved by minimizing cross-entropy loss between the estimated modality label from Z_m^o as $\hat{m}$ and ground truth label, m ($\mathcal{L}_\mathrm{m}$ in main manuscript). This approximation is formulated in **Proposition III.1**.

Proposition III.1. *Let Z and Y be random variables with domains $\mathcal{Z}$ and $\mathcal{Y}$, respectively. Minimizing the conditional cross-entropy loss of predicted label $\hat{Y}$, denoted by $\mathcal{H}(Y;\hat{Y}|Z)$, is equivalent to maximizing the MI$(Z;Y)$*

Proof. Let us define the MI as entropy,

$$\mathrm{MI}(Z,Y) = \underbrace{\mathcal{H}(Y)}_{\delta} - \underbrace{\mathcal{H}(Y|Z)}_{\xi} \tag{A III-12}$$

Since the domain $\mathcal{Y}$ does not change, the entropy of the identity δ term is a constant and can therefore be ignored. Maximizing $\mathrm{MI}(Z,Y)$ can only be achieved through a minimization of the ξ term. We show that $\mathcal{H}(Y|Z)$ is upper-bounded by the cross-entropy loss, and minimizing such loss results in minimizing the ξ term. By expanding its relation to the cross-entropy (Boudiaf *et al.*, 2020):

$$\mathcal{H}(Y;\hat{Y}|Z) = \mathcal{H}(Y|Z) + \underbrace{\mathcal{D}_{\mathrm{KL}}(Y||\hat{Y}|Z)}_{\geq 0}, \tag{A III-13}$$

where we have:

$$\mathcal{H}(Y|Z) \leq \mathcal{H}(Y;\hat{Y}|Z), \tag{A III-14}$$

where minimizing $\mathcal{H}(Y;\hat{Y}|Z)$ results minimizing $\mathcal{H}(Y|Z)$. □

III.A.3 Maximizing MI$(Z_m^e;Y|M)$

Eq. 4.8 of main manuscript can be rewritten w.r.t P 5 as :

$$\mathrm{MI}(Z_m^e;Y|M) = \mathrm{MI}(Z_m^e;Y) - \mathrm{MI}(Z_m^e;Y;M). \tag{A III-15}$$

For the second part RHS:

$$\mathrm{MI}(Z_m^e;Y;M) \leq \mathrm{MI}(Z_m^e;M),$$

where the $\mathcal{L}_{\text{m}}$ in main manuscript is minimized, results $\text{MI}(Z_m^e; M) \simeq 0$ and :

$$\text{MI}(Z_m^e; Y; M) \simeq 0.$$

So, the Eq. A III-15 is:

$$\text{MI}(Z_m^e; Y|M) \simeq \text{MI}(Z_m^e; Y). \tag{A III-16}$$

To maximize the $\text{MI}(Z_m^e; Y)$, the $\mathcal{L}_{\text{meid}}$ is minimized w.r.t to **Proposition III.1**.

III.A.4 Minimizing MI$(Z_m^r; Y|M)$

To enforce the modality-related features Z_m^r to leverage identity-aware information that is dependent on modality (i.e., specific identity-discriminative information), we minimize the amount of identity-aware information in these features that disregards the modality. Below, we demonstrate that $\mathrm{MI}(Z_m^r; Y \mid M)$ is upper-bounded by zero if the features Z_m^r are sufficient for both tasks Y (identity) and M (modality) simultaneously. To ensure that the modality-related features Z_m^r serve as sufficient representations of the input images X for both detecting modality and identifying identity, the loss function $\mathcal{L}_{\text{mrid}}$ (as defined in the main manuscript) is applied to these features.

Theorem III.2. *If the representation Z_m^r of X is sufficient for both Y and M, then:*

$$MI(Z_m^r; Y \mid M) = 0. \tag{A III-17}$$

Proof. From the definition of a sufficient representation Z_m^r for a task M (Definition III.1), we have:

$$\mathrm{MI}(X; M \mid Z_m^r) = 0 \iff \mathrm{MI}(X; M) = \mathrm{MI}(Z_m^r; M) \Rightarrow \mathrm{MI}(Z_m^r; M \mid X) = 0. \tag{A III-18}$$

Similarly, for task Y, we have:

$$\mathrm{MI}(Z_m^r; Y \mid X) = 0. \tag{A III-19}$$

Expanding $\mathrm{MI}(Z_m^r; Y \mid X)$, we obtain:

$$\mathrm{MI}(Z_m^r; Y \mid X) = \mathrm{MI}(Z_m^r; Y; M \mid X) + \mathrm{MI}(Z_m^r; Y \mid X, M) = 0. \tag{A III-20}$$

For the first term on the RHS of Eq. A III-20, we have:

$$\mathrm{MI}(Z_m^r; Y; M \mid X) = \mathrm{MI}(Z_m^r; M \mid X) - \mathrm{MI}(Z_m^r; M \mid X, Y) = 0,$$

where $\mathrm{MI}(Z_m^r; M \mid X) = 0$ follows from Eq. A III-19, since Z_m^r is a sufficient representation of X for task M, and $\mathrm{MI}(Z_m^r; M \mid X, Y) = 0$ is due to Hypothesis III.A.1. For the second term on the RHS of Eq. A III-20, we have:

$$\mathrm{MI}(Z_m^r; Y \mid X, M) = \mathrm{MI}(Z_m^r; Y \mid M) - \mathrm{MI}(Z_m^r; Y; X \mid M) = 0. \tag{A III-21}$$

Since $\mathrm{MI}(Z_m^r; Y; X \mid M) \leq \mathrm{MI}(Z_m^r; Y \mid M)$, and $\mathrm{MI}(Z_m^r; Y \mid M) = \mathrm{MI}(Z_m^r; Y; X \mid M)$, it follows that:

$$\mathrm{MI}(Z_m^r; Y; X \mid M) = 0. \tag{A III-22}$$

Thus, $\mathrm{MI}(Z_m^r; Y \mid M) = 0$, completing the proof. □

III.B. Additional Experiments

III.B.1 Implementation Details of Open-Source SOTA Methods

In this section, we present implementation details for each open-source state-of-the-art (SOTA) method used in our manuscript. For each method, we use the hyperparameter based on their paper or official code in GitHub:

- **DDAG** (Ye *et al.*, 2020a): This method employs ResNet-50 as the backbone with stride one at the last layer, with input images resized to 288×144 pixels that are augmented with zero-padding and horizontal flipping. Based on the DDAG paper(Ye *et al.*, 2020a), 8 people with 4 V and 4 I images are selected in the batch, and $p = 3$ is set.
- **DEEN** (Zhang & Wang, 2023b): A modified ResNet-50 is used as the backbone, enhanced with DEE modules that introduce two additional branches to the network. Input images are resized to 344×144 pixels. During training, augmentations such as Random Erase and Random Channel augmentation are applied. At inference time, the original image and its horizontally flipped counterpart are both processed through the backbone, and the average of the extracted features is used as the final representation. For evaluation under our mixed-modal settings, we removed the flipping process and instead concatenated the features from all DEE branches to create the final representative features.

- **MPANet** (Wu *et al.*, 2021): This method also employs ResNet-50 as the backbone, with an additional convolutional layer designed to detect more discriminative regions in the feature space. The final representative feature vector is constructed by concatenating part features and global features obtained from a Global Average Pooling (GAP) layer. Input images are resized to 344×144 pixels, and augmentations such as Random Erase and Random Channel augmentation are applied during training.
- **SGEIL** (Feng *et al.*, 2023): This method employs two ResNet-50 backbones, one for visible images and the other for infrared images, along with an additional ResNet-50 backbone for shape images. Input images are resized to 288×144 pixels, with augmentations similar to those used in DEEN. During training, model weights are updated using SGD, while an exponential moving average (EMA) is simultaneously applied to update the backbone weights. At the end of training, the best performance between the SGD and EMA models is selected. Note that SGEIL requires shape images for training, and since this information is available only for the SYSU-MM01 dataset, its performance on RegDB and LLCM is not reported.
- **SAAI** (Fang *et al.*, 2023): Similar to MPANet, this method uses ResNet-50 as the backbone, with an additional convolutional layer and learnable parameters to identify part prototypes. The final representative feature vector is obtained by concatenating part features and global features from the GAP layer. Input images are resized to 288×144 pixels, and Random Erase and Random Channel augmentation are applied during training. An Affinity Inference Module is used during inference to rerank the gallery.
- **IDKL** (Ren & Zhang, 2024): This method uses ResNet-50 as the backbone, with additional branches added to layers 3 and 4 to extract modality-specific features. Input images are resized to 388×144 pixels, and Random Erase and Random Channel augmentation are applied during training. During inference, k-reciprocal encoding is applied to rerank the gallery, enhancing the retrieval process.

III.B.2 Additional Mixed-Modal Results

In the main manuscript, we reported the performance of mixed-modal settings for existing datasets with infrared query images and mixed-modal gallery images. Table III-1, presents the performance with visible query. Across all mixed settings in SYSU-MM01, MixER consistently outperforms the compared methods, achieving the highest Rank-1 accuracy (R1) and mean Average Precision (mAP) scores. Specifically, for the Mix setting, our method achieves a notable improvement in mAP (87.29%) compared to the next best method, IDKL (84.78%), showcasing its robustness in handling mixed gallery conditions. In more challenging settings like Mix-Cam and Mix-ID, MixER significantly outperforms the SOTA, demonstrating its ability to adapt to modality and identity constraints effectively.

In the RegDB dataset, which focuses on visible-infrared retrieval, MixER achieves the highest scores across both the Mix and Mix-ID settings. The proposed method achieves a remarkable mAP of 92.78% in the Mix setting, outperforming the best-performing baseline (IDKL) by over 4.4%. Similarly, for the Mix-ID setting, MixER achieves a substantial improvement in mAP (81.42%) compared to SAAI (68.82%). On the LLCM dataset, MixER maintains competitive performance, achieving the highest scores in both the Mix and Mix-ID settings. In particular, MixER improves mAP in the Mix-ID setting to 45.18%, which surpasses the previous best method, DEEN, by a notable margin (43.65%). These results highlight the generalizability and robustness of our method in addressing varying cross-modal and identity constraints.

The proposed MixER consistently achieves superior performance across all datasets and settings, highlighting its effectiveness in extracting discriminative modality-related and modality-erased features. The substantial improvements in challenging settings, such as Mix-ID, underscore the effectiveness of the disentangling strategy employed by MixER, which allows it to handle both modality and identity variations effectively. Unlike competing methods, MixER achieves a balanced improvement in both Rank-1 accuracy and mAP, demonstrating its robustness in retrieval precision and ranking performance.

Table-A III-1 Accuracy of the proposed method and open-sourced state-of-the-art methods on the SYSU-MM01 (single-shot setting), RegDB, and LLCM datasets in different mixed gallery settings. Visible images are chosen as the query

Method	SYSU-MM01								RegDB				LLCM			
	Mix		**Mix-Cam**		**Mix-Cam-ID**		**Mix-ID**		**Mix**		**Mix-ID**		**Mix**		**Mix-ID**	
	R1	**mAP**	**R1**	**mAP**	**R1**	**mAP**	**R1**	**mAP**	**R1**	**mAP**	**R1**	**mAP**	**R1**	**mAP**	**R1**	**mAP**
DDAG (Ye *et al.*, 2020a)	97.59	79.62	94.68	76.61	92.96	73.42	41.09	45.85	99.9	77.53	45.39	47.13	99.24	51.31	22.50	17.94
MPANet (Wu *et al.*, 2021)	**97.94**	83.80	95.23	80.98	94.10	78.91	54.54	57.24	**100**	84.32	60.24	61.57	99.17	59.61	25.08	18.82
DEEN (Zhang & Wang, 2023b)	95.79	82.18	92.29	80.32	89.85	77.32	59.13	61.03	99.95	88.49	75.24	71.45	99.25	73.73	57.90	43.65
SGEIL (Feng *et al.*, 2023)	96.76	80.52	94.05	78.74	91.10	74.82	48.72	52.96	-	-	-	-	-	-	-	-
SAAI (Fang *et al.*, 2023)	97.63	83.88	95.22	81.50	93.82	79.08	55.08	57.61	100	88.19	69.22	68.82	99.57	70.60	46.12	34.63
IDKL (Ren & Zhang, 2024)	98.25	84.78	96.15	82.51	94.87	79.85	54.00	57.51	99.95	88.33	71.70	70.64	99.57	70.54	39.25	32.04
MixER (ours)	97.14	**87.29**	**96.27**	**85.67**	**94.95**	**84.79**	**70.96**	**70.76**	**100**	**92.78**	**85.39**	**81.42**	**99.80**	**74.59**	**58.87**	**45.18**

III.B.3 Additional Ablation Studies

III.B.3.1 Computational Complexity

We compare the training times of several state-of-the-art methods in Cross-Modal ReID to demonstrate how much additional computational burden is added due to the augmentation of these methodologies with our method. The training times have been documented in Table III-2. Note that each method has its own variable number of training epochs, keeping in mind the optimal number of epochs suggested for these methods. We argue that the overall increase in computational time is a reasonable trade-off with a projected performance increase, thus highlighting the superiority of our method. The upper-bound model represents a best-case scenario with three separate SAAI models trained independently for visible, infrared, and VI images. Also, to compare with the upper-bound, using MixER is more efficient.

Table-A III-2 Training times for state-of-the-art Cross-Modal ReID methods (on SYSY-MM01 dataset) compared to the training times of the same methods modified with our method. Time has been reported in minutes

Method	**Training time**	**Training time with MixER**	**# epochs**	**# of param**	**# of params with MixER**	**Flops**	**Flops with MixER**
DDAG	111	137	80	40M	54M	0.5T	0.6T
DEEN	686	732	151	61M	75M	1.3T	1.5T
SGEIL	597	631	120	87M	101M	0.9T	0.97T
SAAI	78	101	160	72M	85M	0.75T	0.8T
IDKL	182	204	180	88M	106M	0.95T	1.1T
MPANet	144	179	140	74.8M	88M	0.9T	1.02T
Upper-Bound	184	-	160	216M	-	2.25T	-

Table III-3 compares the performance of the baseline SAAI method (Fang *et al.*, 2023), its modified version enhanced with our proposed MixER framework, and an upper-bound model across mixed-modal, cross-modal, and uni-modal settings on the SYSU-MM01 dataset. The results show that the SAAI+MixER model consistently outperforms the baseline SAAI in mixed-modal settings, achieving notable improvements in both Rank-1 accuracy and mAP. For instance, in the Mix setting, SAAI+MixER achieves an mAP of 80.47% compared to 74.59% for the baseline. In the challenging Mix-ID setting, it improves mAP from 53.30% to 62.45%, demonstrating the effectiveness of MixER in disentangling modality-related and modality-erased features.

In cross-modal settings, SAAI+MixER also demonstrates robustness, achieving a slight improvement in mAP for the "All" setting (71.08% vs. 69.71%) while maintaining competitive performance in uni-modal scenarios. In the I→I and V→V settings, SAAI+MixER either matches or outperforms the baseline without compromising intra-modality retrieval. Notably, in some cases, SAAI+MixER exceeds the upper-bound, such as in the Mix setting where it achieves an mAP of 80.47

Table-A III-3 Performance of SAAI ((Fang *et al.*, 2023)) VI-ReID technique in mixed, cross, and uni-modal settings on the SYSU-MM01 dataset compared to upper-bound. The upper-bound is a model that contains three separate SAAI models for V,I, and VI images

Method	Mixed-Modal								Cross-Modal				Uni-Modal			
	Mix		**Mix-Cam**		**Mix-Cam-ID**		**Mix-ID**		**All**		**Indoor**		**I→I**		**V→V**	
	R1	**mAP**	**R1**	**mAP**	**R1**	**mAP**	**R1**	**mAP**	**R1**	**mAP**	**R1**	**mAP**	**R1**	**mAP**	**R1**	**mAP**
SAAI	96.01	74.59	90.63	72.51	84.30	65.94	52.49	53.30	73.87	69.71	84.19	82.59	89.29	93.06	98.24	93.46
SAAI+MixER	97.27	80.47	92.66	76.81	88.51	73.24	66.23	62.45	74.25	71.08	-	-	92.11	94.57	98.60	94.08
Upper-bound	95.74	78.20	91.18	74.94	85.44	68.71	59.38	55.70	73.87	69.71	84.19	82.59	88.06	92.80	98.85	95.20

III.B.3.2 The influence of hyperparameters.

In this section, we present a bar chart (Fig. III-2) to examine the detailed influence of hyperparameters by gradually increasing their value. As we can see, the best performance is achieved when λ_1 is set to 0.4, λ_2 is set to 0.6, and λ_3 is set to 0.4, respectively. The upward trend of the bars demonstrates the effectiveness of each loss.

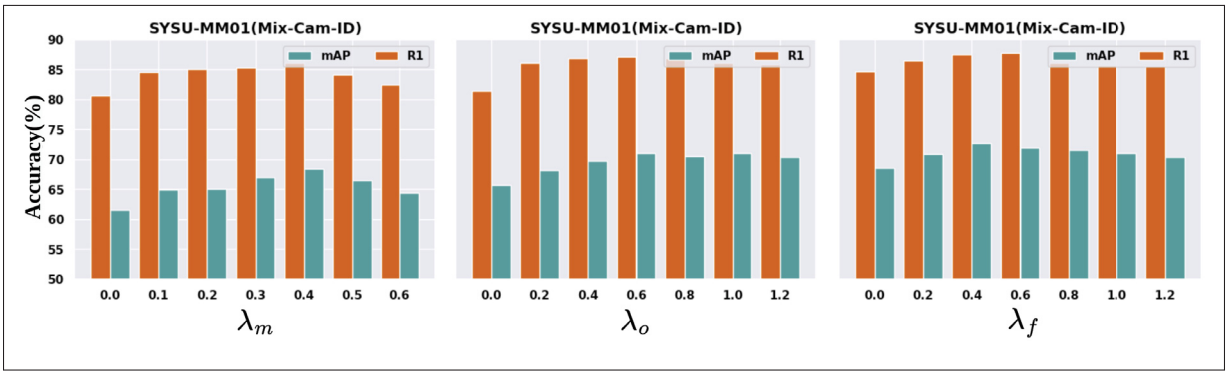


Figure-A III-2 Influence of different λ_m,λ_o and λ_f values on SYSU-MM01 in Mix-Cam-ID evaluation

III.B.3.3 Choice of Backbone.

We test the performance of our method on two main choices of backbones, ResNet (He, Zhang, Ren & Sun, 2016a), and ViT (Alexey, 2020). For vision transformer models, we resized images to 224 by 224 pixels, and we used the extracted feature from the last layer as representative features. We observe that despite its success in recent times (Islam, 2022), standard ViT models struggle to perform on par with ResNet. This is likely due to the large image resolution ViTs (224×224) are designed to train on. This size allows patches to be as large as 16×16. However, our input images were originally resized to a shape 288×144 restricts us from using a ViT architecture reliably as the ViT architecture dimensions depend on input dimensions. In this regard, ResNet is a much more flexible backbone that can work on a wide range of image resolutions and therefore delivers better performance. All models were trained following the standard implementation settings as described in Section 4.1 of the main paper, with the exception of ViT inputs being resized to 224×224, instead of 288×144.

Table-A III-4 The influence of the choice of baseline backbone on the performance of the proposed method

Backbone	**# Parameters**	**Cross-Modal**		**Mix-Cam-ID**	
		R1	***mAP***	***R1***	***mAP***
ResNet-18	29.02M	67.68	63.51	80.49	62.61
ResNet-50	58.15M	73.43	70.92	87.56	72.0
ViT-B-16	102.87M	55.47	54.97	72.23	57.51
ViT-L-16	331.55M	52.57	53.04	68.14	57.18

III.B.4 Feature Distributions Visualization

We visualized the feature distributions of the baseline and our modules using UMAP (McInnes *et al.*, 2018) in the 2D space, as shown in Fig. III-3(a-d). The results indicate that the modality-erased feature brings embeddings of the same person closer across modalities compared to the baseline (see circular features in Fig. III-3(a,b)). Meanwhile, the modality-related component pushes apart intra-modality features of different individuals. Together, MixER effectively leverages both components to better separate identities and reduce modality discrepancy within the mixed-modal gallery.

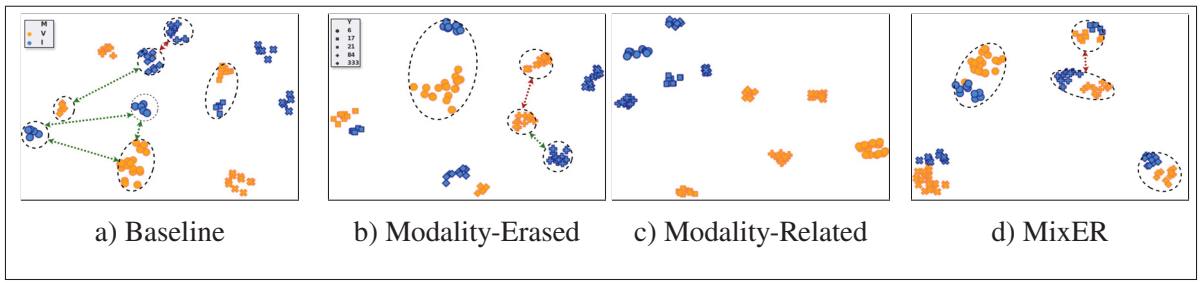


Figure-A III-3 The distribution of feature embeddings in the 2D feature space, where orange and blue colors denote the V and I. The samples with the same shape are from the same person. The green and red arrows show the distance between the same person's features and different persons, respectively

BIBLIOGRAPHY

Abnar, S., Berg, R. v. d., Ghiasi, G., Dehghani, M., Kalchbrenner, N. & Sedghi, H. (2021). Gradual domain adaptation in the wild: When intermediate distributions are absent. *arXiv preprint arXiv:2106.06080*.

Achille, A. & Soatto, S. (2018). Information dropout: Learning optimal representations through noisy computation. *IEEE transactions on pattern analysis and machine intelligence*, 40(12), 2897–2905.

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

Alehdaghi, M., Josi, A., Shamsolmoali, P., Cruz, R. M. & Granger, E. (2023). Adaptive Generation of Privileged Intermediate Information for Visible-Infrared Person Re-Identification. *arXiv preprint arXiv:2307.03240*.

Alehdaghi, M., Shamsolmoali, P., Cruz, R. M. & Granger, E. (2024). Bidirectional Multi-Step Domain Generalization for Visible-Infrared Person Re-Identification. *arXiv preprint arXiv:2403.10782*.

Alehdaghi, M., Shamsolmoali, P., Cruz, R. M. & Granger, E. (2025). Bidirectional multi-step domain generalization for visible-infrared person re-identification. *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 763–773.

Alexey, D. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv: 2010.11929*.

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R. & Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7), e0130140.

Belghazi, M. I., Baratin, A., Rajeshwar, S., Ozair, S., Bengio, Y., Courville, A. & Hjelm, D. (2018, 10–15 Jul). Mutual Information Neural Estimation. *Proceedings of the 35th International Conference on Machine Learning*, 80(Proceedings of Machine Learning Research), 531–540. Retrieved from: https://proceedings.mlr.press/v80/belghazi18a.html.

Bhuiyan, A., Liu, Y., Siva, P., Javan, M., Ayed, I. B. & Granger, E. (2020). Pose guided gated fusion for person re-identification. *WACV*.

Biederman, I. (1987). Recognition-by-components: a theory of human image understanding. *Psychological review*, 94(2), 115.

Boudiaf, M., Rony, J., Ziko, I. M., Granger, E., Pedersoli, M., Piantanida, P. & Ben Ayed, I. (2020). A unifying mutual information view of metric learning: cross-entropy vs. pairwise losses. *European Conference on Computer Vision*, pp. 548–564.

Bousmalis, K., Trigeorgis, G., Silberman, N., Krishnan, D. & Erhan, D. (2016). Domain separation networks. *Advances in neural information processing systems*, 29.

Bousmalis, K., Silberman, N., Dohan, D., Erhan, D. & Krishnan, D. (2017). Unsupervised pixel-level domain adaptation with generative adversarial networks. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3722–3731.

Cai, L., Wang, Z., Gao, H., Shen, D. & Ji, S. (2018). Deep adversarial learning for multi-modality missing data completion. *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 1158–1166.

Chan, C. S., Kong, H. & Liang, G. (2022). A comparative study of faithfulness metrics for model interpretability methods. *arXiv preprint arXiv:2204.05514*.

Chefer, H., Gur, S. & Wolf, L. (2021). Transformer interpretability beyond attention visualization. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Chen, C., Li, O., Tao, D., Barnett, A., Rudin, C. & Su, J. K. (2019). This looks like that: deep learning for interpretable image recognition. *Advances in neural information processing systems*, 32.

Chen, D., Li, H., Liu, X., Shen, Y., Shao, J., Yuan, Z. & Wang, X. (2018). Improving deep visual representation for person re-identification by global and local image-language association. *Proceedings of the European conference on computer vision (ECCV)*, pp. 54–70.

Chen, H.-Y. & Chao, W.-L. (2021). Gradual domain adaptation without indexed intermediate domains. *Advances in Neural Information Processing Systems*, 34, 8201–8214.

Chen, K., Pan, Z., Wang, J., Jiao, S., Zeng, Z. & Miao, Z. (2020a). Joint Feature Learning Network for Visible-Infrared Person Re-identification. *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pp. 652–663.

Chen, T., Kornblith, S., Norouzi, M. & Hinton, G. (2020b). A simple framework for contrastive learning of visual representations. *International conference on machine learning*, pp. 1597–1607.

Chen, X., Mottaghi, R., Liu, X., Fidler, S., Urtasun, R. & Yuille, A. (2014). Detect what you can: Detecting and representing objects using holistic models and body parts. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1971–1978.

Chen, X., Fan, H., Girshick, R. B. & He, K. (2020c). A Simple Framework for Contrastive Learning of Visual Representations. *International Conference on Machine Learning (ICML)*, 0–1.

Chen, Y., Wan, L., Li, Z., Jing, Q. & Sun, Z. (2021). Neural feature search for rgb-infrared person re-identification. *CVPR*, pp. 587–597.

Choi, C., Kim, S. & Ramani, K. (2017). Learning hand articulations by hallucinating heat distribution. *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3104–3113.

Choi, S., Lee, S., Kim, Y., Kim, T. & Kim, C. (2020a). HI-CMD: hierarchical cross-modality disentanglement for visible-infrared person re-identification. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10257–10266.

Choi, Y., Uh, Y., Yoo, J. & Ha, J.-W. (2020b). StarGAN v2: Diverse Image Synthesis for Multiple Domains. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Chopra, S., Hadsell, R. & LeCun, Y. (2005). Learning a similarity metric discriminatively, with application to face verification. *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, 1, 539–546.

Choudhury, S., Laina, I., Rupprecht, C. & Vedaldi, A. (2021). Unsupervised part discovery from contrastive reconstruction. *Advances in Neural Information Processing Systems*, 34, 28104–28118.

Crasto, N., Weinzaepfel, P., Alahari, K. & Schmid, C. (2019). Mars: Motion-augmented rgb stream for action recognition. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7882–7891.

Cui, Z., Zhou, J. & Peng, Y. (2024). DMA: Dual Modality-Aware Alignment for Visible-Infrared Person Re-Identification. *IEEE Transactions on Information Forensics and Security*.

Dai, P., Ji, R., Wang, H., Wu, Q. & Huang, Y. (2018). Cross-Modality Person Re-Identification with Generative Adversarial Training. *IJCAI*, 1, 2.

Dai, Y., Liu, J., Sun, Y., Tong, Z., Zhang, C. & Duan, L.-Y. (2021). Idm: An intermediate domain module for domain adaptive person re-id. *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11864–11874.

Dasgupta, S., Frost, N. & Moshkovitz, M. (2022). Framework for evaluating faithfulness of local explanations. *International Conference on Machine Learning*, pp. 4794–4815.

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

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

Deng, J., Guo, J., Xue, N. & Zafeiriou, S. (2019). ArcFace: Additive angular margin loss for deep face recognition. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4690–4699.

Donnelly, J., Barnett, A. J. & Chen, C. (2022). Deformable protopnet: An interpretable image classifier using deformable prototypes. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10265–10275.

Dreyer, M., Achtibat, R., Samek, W. & Lapuschkin, S. (2024). Understanding the (Extra-)Ordinary: Validating Deep Model Decisions with Prototypical Concept-based Explanations. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 3491-3501. doi: 10.1109/CVPRW63382.2024.00353.

Du, Y., Lei, C., Zhao, Z., Dong, Y. & Su, F. (2024). Video-Based Visible-Infrared Person ReID With Auxiliary Samples. *IEEE Trans. Information Forensics and Security*, 19, 1313-1325.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X. et al. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *kdd*, 96(34), 226–231.

Fan, X., Luo, H., Zhang, C. & Jiang, W. (2020). Cross-Spectrum Dual-Subspace Pairing for RGB-infrared Cross-Modality Person Re-Identification. *ArXiv*, abs/2003.00213.

Fang, X., Yang, Y. & Fu, Y. (2023). Visible-Infrared Person Re-Identification via Semantic Alignment and Affinity Inference. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11270–11279.

Fellous, J.-M., Sapiro, G., Rossi, A., Mayberg, H. & Ferrante, M. (2019). Explainable artificial intelligence for neuroscience: behavioral neurostimulation. *Frontiers in neuroscience*, 13, 1346.

Feng, J., Wu, A. & Zheng, W.-S. (2023). Shape-Erased Feature Learning for Visible-Infrared Person Re-Identification. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22752–22761.

Fu, C., Hu, Y., Wu, X., Shi, H., Mei, T. & He, R. (2021a). Cm-nas: Cross-modality neural architecture search for visible-infrared person re-identification. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11823–11832.

Fu, D., Chen, D., Bao, J., Yang, H., Yuan, L., Zhang, L., Li, H. & Chen, D. (2021b). Unsupervised pre-training for person re-identification. *CVPR*.

Gabeur, V., Sun, C., Alahari, K. & Schmid, C. (2020). Multi-modal transformer for video retrieval. *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pp. 214–229.

Ganin, Y. & Lempitsky, V. (2015). Unsupervised domain adaptation by backpropagation. *International conference on machine learning*, pp. 1180–1189.

Ge, W. & Yu, Y. (2017). Borrowing treasures from the wealthy: Deep transfer learning through selective joint fine-tuning. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1086–1095.

Ghifary, M., Kleijn, W. B., Zhang, M., Balduzzi, D. & Li, W. (2016). Deep reconstruction-classification networks for unsupervised domain adaptation. *European conference on computer vision*, pp. 597–613.

Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K. V., Joulin, A. & Misra, I. (2023). Imagebind: One embedding space to bind them all. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15180–15190.

Gong, M., Liu, K., Li, D. & Tao, D. (2019). DLOW: Domain Flow for Adaptation and Generalization. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2477–2486.

Goodfellow, I. J., 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, 0–1.

Greff, K., Van Steenkiste, S. & Schmidhuber, J. (2020). Binding in modular neural networks. *arXiv preprint arXiv:2006.01882*.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., Smola, A. & DUMMY, D. (2012a). A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1), 723–773.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B. & Smola, A. (2012b). A Kernel Two-Sample Test. *Journal of Machine Learning Research*, 13, 723–773.

Hadsell, R., Chopra, S. & LeCun, Y. (2006). Dimensionality Reduction by Learning an Invariant Mapping. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Hao, Y., Li, J., Wang, N. & Gao, X. (2020). Modality adversarial neural network for visible-thermal person re-identification. *Pattern Recognition*, 107, 107533.

He, K., Zhang, X., Ren, S. & Sun, J. (2016a). Deep Residual Learning for Image Recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

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

He, W., Deng, Y., Tang, S., Chen, Q., Xie, Q., Wang, Y., Bai, L., Zhu, F., Zhao, R., Ouyang, W. et al. (2024). Instruct-ReID: A Multi-purpose Person Re-identification Task with Instructions. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17521–17531.

Hermans, A., Beyer, L. & Leibe, B. (2017). In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 0–1.

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.

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, Z., Liu, J., Li, L., Zheng, K. & Zha, Z.-J. (2022). Modality-adaptive mixup and invariant decomposition for RGB-infrared person re-identification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(1), 1034–1042.

Islam, K. (2022). Recent advances in vision transformer: A survey and outlook of recent work. *arXiv preprint arXiv:2203.01536*.

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

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.

Jia, M., Zhai, Y., Lu, S., Ma, S. & Zhang, J. (2020). A similarity inference metric for RGB-infrared cross-modality person re-identification. *arXiv preprint arXiv:2007.01504*, 0–1.

Jiang, J., Jin, K., Qi, M., Wang, Q., Wu, J. & Chen, C. (2020). A Cross-Modal Multi-granularity Attention Network for RGB-IR Person Re-identification. *Neurocomputing*.

Kampffmeyer, M., Salberg, A.-B. & Jenssen, R. (2018). Urban land cover classification with missing data modalities using deep convolutional neural networks. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 11(6), 1758–1768.

Karazija, P., Dorr, K., Perrot, M., Staniute, M. & Ritchie, D. (2021). CLEVR-TEX: A Visual Reasoning Dataset for Object-centric Representations. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F. et al. (2018). Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). *International conference on machine learning*, pp. 2668–2677.

Kim, H.-S. & Joe, I. (2022). An XAI method for convolutional neural networks in self-driving cars. *PLoS one*, 17(8), e0267282.

Kim, M., Kim, S., Park, J., Park, S. & Sohn, K. (2023). PartMix: Regularization Strategy to Learn Part Discovery for Visible-Infrared Person Re-identification. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18621–18632.

Kim, T., Cha, M., Kim, H., Lee, J. K. & Kim, J. (2017). Learning to discover cross-domain relations with generative adversarial networks. *International conference on machine learning*, pp. 1857–1865.

Kim, W., Son, B. & Kim, I. (2021). Vilt: Vision-and-language transformer without convolution or region supervision. *International conference on machine learning*, pp. 5583–5594.

Kingma, D. P. & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kipf, T., van Steenkiste, S., Greff, K., Locatello, F., Schmidhuber, J. & Bachem, O. (2022). Conditional Object-Centric Learning from Video. *International Conference on Learning Representations (ICLR)*.

Kniaz, V. V., Knyaz, V. A., Hladuvka, J., Kropatsch, W. G. & Mizginov, V. (2018). Thermalgan: Multimodal color-to-thermal image translation for person re-identification in multispectral dataset. *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 0–0.

Krause, J., Stark, M., Deng, J. & Fei-Fei, L. (2013). 3d object representations for fine-grained categorization. *Proceedings of the IEEE international conference on computer vision workshops*, pp. 554–561.

Kumar, S., Banerjee, B. & Chaudhuri, S. (2021). Improved Landcover Classification using Online Spectral Data Hallucination. *Neurocomputing*, 439, 316–326.

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.

Lee, Y.-L., Tsai, Y.-H., Chiu, W.-C. & Lee, C.-Y. (2023). Multimodal prompting with missing modalities for visual recognition. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14943–14952.

Lezama, J., Qiu, Q. & Sapiro, G. (2017). Not afraid of the dark: Nir-vis face recognition via cross-spectral hallucination and low-rank embedding. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6628–6637.

Li, D., Wei, X., Hong, X. & Gong, Y. (2020a). Infrared-visible cross-modal person re-identification with an x modality. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04), 4610–4617.

Li, H., Ye, M., Zhang, M. & Du, B. (2024). All in One Framework for Multimodal Re-identification in the Wild. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17459–17469.

Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C. & Hoi, S. C. H. (2021a). Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34, 9694–9705.

Li, P., Xu, Y., Wei, Y. & Yang, Y. (2020b). Self-Correction for Human Parsing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. doi: 10.1109/TPAMI.2020.3048039.

Li, Y., He, J., Zhang, T., Liu, X., Zhang, Y. & Wu, F. (2021b). Diverse part discovery: Occluded person re-identification with part-aware transformer. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2898–2907.

Liang, P. P., Zadeh, A. & Morency, L.-P. (2022). Foundations and Trends in Multimodal Machine Learning: Principles, Challenges, and Open Questions. *arXiv preprint arXiv:2209.03430*.

Liu, H. & Cheng, J. (2019). Enhancing the Discriminative Feature Learning for Visible-Thermal Cross-Modality Person Re-Identification. *CoRR*, abs/1907.09659. Retrieved from: http://arxiv.org/abs/1907.09659.

Liu, H., Ma, S., Xia, D. & Li, S. (2021). SFANet: A Spectrum-aware Feature Augmentation Network for Visible-Infrared Person Re-Identification. *arXiv preprint arXiv:2102.12137*.

Liu, W., Xu, X., Chang, H., Yuan, X. & Wang, Z. (2025). Mix-Modality Person Re-Identification: A New and Practical Paradigm. *ACM Trans. Multimedia Comput. Commun. Appl.* doi: 10.1145/3715142.

Locatello, F., Weissenborn, D., Unterthiner, T., Mahendran, A., Heigold, G., Uszkoreit, J. & Kipf, T. (2020). Object-centric learning with slot attention. *Advances in Neural Information Processing Systems (NeurIPS)*.

Lu, Y., Wu, Y., Liu, B., Zhang, T., Li, B., Chu, Q. & Yu, N. (2020). Cross-modality person re-identification with shared-specific feature transfer. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13379–13389.

Lu, Z., Lin, R. & Hu, H. (2023). Modality and Camera Factors Bi-Disentanglement for NIR-VIS Object ReID. *IEEE Trans. Information Forensics and Security*, 18, 1989-2004.

Lu, Z., Lin, R. & Hu, H. (2024). Disentangling Modality and Posture Factors: Memory-Attention and Orthogonal Decomposition for Visible-Infrared Person Re-Identification. *IEEE Transactions on Neural Networks and Learning Systems*.

Luo, Y., Wang, H. & Li, Z. (2020). Progressive Learning for Adaptation and Generalization in Domain Transfer. *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 578–595.

McInnes, L., Healy, J. & Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Mekhazni, D., Bhuiyan, A., Ekladious, G. & Granger, E. (2020). Unsupervised domain adaptation in the dissimilarity space for person re-identification. *European Conference on Computer Vision*, pp. 159–174.

Monet, D., Greff, K., Van Steenkiste, S., Locatello, F. & Bachem, O. (2021). Object-centric learning with slot-based attention. *International Conference on Learning Representations (ICLR)*.

Nauta, M., Van Bree, R. & Seifert, C. (2021). Neural prototype trees for interpretable fine-grained image recognition. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14933–14943.

Nauta, M., Schlötterer, J., Van Keulen, M. & Seifert, C. (2023). Pip-net: Patch-based intuitive prototypes for interpretable image classification. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2744–2753.

Nguyen, D. T., Hong, H. G., Kim, K. W. & Park, K. R. (2017). Person recognition system based on a combination of body images from visible light and thermal cameras. *Sensors*, 17(3), 605.

Oikarinen, T., Das, S., Nguyen, L. M. & Weng, T.-W. (2023). Label-free concept bottleneck models. *arXiv preprint arXiv:2304.06129*.

Pan, X., Yao, T. & Sun, Y. (2020). Unsupervised Domain Adaptation with Hierarchical Feature Alignment. *IEEE Transactions on Neural Networks and Learning Systems*, 31(10), 4213–4225.

Pan, Y., Liu, M., Xia, Y. & Shen, D. (2021). Disease-image-specific learning for diagnosis-oriented neuroimage synthesis with incomplete multi-modality data. *IEEE transactions on pattern analysis and machine intelligence*, 44(10), 6839–6853.

Pande, S., Banerjee, A., Kumar, S., Banerjee, B. & Chaudhuri, S. (2019). An adversarial approach to discriminative modality distillation for remote sensing image classification. *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 0–0.

Pang, Z., Zhao, L., Liu, Y., Sharma, G. & Wang, C. (2024). Inter-modality similarity learning for unsupervised multi-modality person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*.

Park, H., Lee, S., Lee, J. & Ham, B. (2021). Learning by Aligning: Visible-Infrared Person Re-identification using Cross-Modal Correspondences. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12046–12055.

Pechyony, D. & Vapnik, V. (2010a). On the theory of learnining with privileged information. *Advances in neural information processing systems*, 23.

Pechyony, D. & Vapnik, V. (2010b). On the theory of learnining with privileged information. *Advances in neural information processing systems*, 23, 0–1.

Pei, Z., Cao, Z., Long, M. & Wang, J. (2018). Multi-Adversarial Domain Adaptation. *Thirty-Second AAAI Conference on Artificial Intelligence (AAAI)*.

Peng, K.-C., Wu, Z. & Ernst, J. (2018). Zero-shot deep domain adaptation. *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 764–781.

Peng, X. & Saenko, K. (2018). Synthetic to real adaptation with generative correlation alignment networks. *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1982–1991.

Ren, K. & Zhang, L. (2024). Implicit Discriminative Knowledge Learning for Visible-Infrared Person Re-Identification. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 393–402.

Ronneberger, O., Fischer, P. & Brox, T. (2022). Convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015 Conference Proceedings*.

Rymarczyk, D., Struski, Ł., Tabor, J. & Zieliński, B. (2021). Protopshare: Prototypical parts sharing for similarity discovery in interpretable image classification. *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pp. 1420–1430.

Saputra, M. R. U., de Gusmao, P. P., Lu, C. X., Almalioglu, Y., Rosa, S., Chen, C., Wahlström, J., Wang, W., Markham, A. & Trigoni, N. (2020). Deeptio: A deep thermal-inertial odometry with visual hallucination. *IEEE Robotics and Automation Letters*, 5(2), 1672–1679.

Saranya, A. & Subhashini, R. (2023). A systematic review of Explainable Artificial Intelligence models and applications: Recent developments and future trends. *Decision analytics journal*, 7, 100230.

Schroff, F., Kalenichenko, D. & Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 815–823.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D. & Batra, D. (2017a). Grad-CAM: Visual explanations from deep networks via gradient-based localization. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D. & Batra, D. (2017b). Grad-cam: Visual explanations from deep networks via gradient-based localization. *Proceedings of the IEEE international conference on computer vision*, pp. 618–626.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D. & Batra, D. (2017c). Grad-cam: Visual explanations from deep networks via gradient-based localization. *Proceedings of the IEEE international conference on computer vision*, pp. 618–626.

Sharma, C., Kapil, S. R. & Chapman, D. (2021). Person re-identification with a locally aware transformer. *arXiv:2106.03720*.

Singh, S., Li, Y., Ebrahimi, S., Darrell, T., Efros, A. A. & Yu, J. (2022). Illiterate DALL-E Learns to Compose. *European Conference on Computer Vision (ECCV)*.

Somers, V., De Vleeschouwer, C. & Alahi, A. (2023). Body Part-Based Representation Learning for Occluded Person Re-Identification. *WACV*.

Srinivas, S. & Fleuret, F. (2019). Full-Gradient Representation for Neural Network Visualization. *Advances in Neural Information Processing Systems (NeurIPS)*.

Stylianou, A., Souvenir, R. & Pless, R. (2019). Visualizing deep similarity networks. *2019 IEEE winter conference on applications of computer vision (WACV)*, pp. 2029–2037.

Sun, B. & Saenko, K. (2016). Deep CORAL: Correlation Alignment for Deep Domain Adaptation. *European Conference on Computer Vision (ECCV)*, 0–1.

Sun, J., Li, Y., Chen, H., Peng, Y. & Zhu, J. (2023). Visible-infrared person re-identification model based on feature consistency and modal indistinguishability. *Machine Vision and Applications*, 34(1), 14.

Tzeng, E., Hoffman, J., Darrell, T. & Saenko, K. (2015). Simultaneous deep transfer across domains and tasks. *Proceedings of the IEEE international conference on computer vision*, pp. 4068–4076.

Tzeng, E., Devin, C., Hoffman, J., Finn, C., Abbeel, P., Levine, S., Saenko, K. & Darrell, T. (2020). Adapting deep visuomotor representations with weak pairwise constraints. *Algorithmic Foundations of Robotics XII: Proceedings of the Twelfth Workshop on the Algorithmic Foundations of Robotics*, pp. 688–703.

Van Den Oord, A., Vinyals, O. et al. (2017). Neural discrete representation learning. *Advances in neural information processing systems*, 30.

van der Klis, R., Alaniz, S., Mancini, M., Dantas, C. F., Ienco, D., Akata, Z. & Marcos, D. (2023). PDiscoNet: Semantically consistent part discovery for fine-grained recognition. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1866–1876.

van der Maaten, L. & Hinton, G. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9(86), 2579–2605. Retrieved from: http://jmlr.org/papers/v9/vandermaaten08a.html.

Vapnik, V. & Vashist, A. (2009). A new learning paradigm: Learning using privileged information. *Neural networks*, 22(5-6), 544–557.

Vilone, G. & Longo, L. (2020). Explainable artificial intelligence: a systematic review. *arXiv preprint arXiv:2006.00093*.

Wah, C., Branson, S., Welinder, P., Perona, P. & Belongie, S. (2011). The caltech-ucsd birds-200-2011 dataset.

Wan, L., Sun, Z., Jing, Q., Chen, Y., Lu, L. & Li, Z. (2023). G2DA: Geometry-guided dual-alignment learning for RGB-infrared person re-identification. *Pattern Recognition*, 135, 109150. doi: https://doi.org/10.1016/j.patcog.2022.109150.

Wang, G.-A., Zhang, T., Yang, Y., Cheng, J., Chang, J., Liang, X. & Hou, Z.-G. (2020a). Cross-modality paired-images generation for RGB-infrared person re-identification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07), 12144–12151.

Wang, G., Zhang, T., Cheng, J., Liu, S., Yang, Y. & Hou, Z. (2019a, October). RGB-Infrared Cross-Modality Person Re-Identification via Joint Pixel and Feature Alignment. *The IEEE International Conference on Computer Vision (ICCV)*.

Wang, H., Wang, Z., Du, M., Yang, F., Zhang, Z., Ding, S., Mardziel, P. & Hu, X. (2020b). Score-CAM: Score-weighted visual explanations for convolutional neural networks. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 24–25.

Wang, H., Li, B. & Zhao, H. (2022a). Understanding gradual domain adaptation: Improved analysis, optimal path and beyond. *International Conference on Machine Learning*, pp. 22784–22801.

Wang, H., Chen, Y., Ma, C., Avery, J., Hull, L. & Carneiro, G. (2023). Multi-modal learning with missing modality via shared-specific feature modelling. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15878–15887.

Wang, J., Wang, Y. & Liu, X. (2022b). Progressive Domain Adaptation for Robust Visual Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10), 5013–5028.

Wang, M. & Deng, W. (2018). Deep visual domain adaptation: A survey. *Neurocomputing*, 312, 135–153.

Wang, P., Wei, Y., Yang, Y., Zheng, L., Ye, Q. & Jiao, J. (2019b). AlignGAN: Learning Modality-Invariant Features for Visible-Infrared Person Re-Identification. *International Journal of Computer Vision (IJCV)*, 0–1.

Wang, X., Girshick, R., Gupta, A. & He, K. (2018). Non-local neural networks. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7794–7803.

Wang, Z., Wang, Z., Zheng, Y., Wu, Y., Zeng, W. & Satoh, S. (2019c). Beyond intra-modality: A survey of heterogeneous person re-identification. *arXiv preprint arXiv:1905.10048*, 0–1.

Wang, Z., Wang, Z., Zheng, Y., Chuang, Y.-Y. & Satoh, S. (2019d). Learning to reduce dual-level discrepancy for infrared-visible person re-identification. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 618–626.

Wei, Z., Yang, X., Wang, N. & Gao, X. (2021). Syncretic modality collaborative learning for visible infrared person re-identification. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 225–234.

Wen, Y., Zhang, K., Li, Z. & Qiao, Y. (2016). A Discriminative Feature Learning Approach for Deep Face Recognition. *Proceedings of the European Conference on Computer Vision (ECCV)*, 0–1.

Wong, E., Santurkar, S. & Madry, A. (2021). Leveraging sparse linear layers for debuggable deep networks. *International Conference on Machine Learning*, pp. 11205–11216.

Wu, A., Zheng, W.-S., Yu, H.-X., Gong, S. & Lai, J. (2017a). RGB-infrared cross-modality person re-identification. *Proceedings of the IEEE international conference on computer vision*, pp. 5380–5389.

Wu, A., Zheng, W.-S., Yu, H.-X., Gong, S. & Lai, J. (2017b). RGB-infrared cross-modality person re-identification. *Proceedings of the IEEE international conference on computer vision*, pp. 5380–5389.

Wu, A., Zheng, W.-S., Gong, S. & Lai, J. (2020). RGB-IR Person Re-identification by Cross-Modality Similarity Preservation. *International journal of computer vision*, 128(6), 1765–1785.

Wu, J., Liu, H., Su, Y., Shi, W. & Tang, H. (2023a). Learning Concordant Attention via Target-aware Alignment for Visible-Infrared Person Re-identification. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11122–11131.

Wu, L., Liu, D., Zhang, W., Chen, D., Ge, Z., Boussaid, F., Bennamoun, M. & Shen, J. (2022). Pseudo-pair based self-similarity learning for unsupervised person re-identification. *IEEE Transactions on Image Processing*, 31, 4803–4816.

Wu, L. Y., Liu, L., Wang, Y., Zhang, Z., Boussaid, F., Bennamoun, M. & Xie, X. (2023b). Learning Resolution-Adaptive Representations for Cross-Resolution Person Re-Identification. *IEEE Transactions on Image Processing*.

Wu, Q., Dai, P., Chen, J., Lin, C.-W., Wu, Y., Huang, F., Zhong, B. & Ji, R. (2021). Discover Cross-Modality Nuances for Visible-Infrared Person Re-Identification. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4330–4339.

Xu, X., Wu, S., Liu, S. & Xiao, G. (2021). Cross-Modal Based Person Re-identification via Channel Exchange and Adversarial Learning. *International Conference on Neural Information Processing*, pp. 500–511.

Yang, Y., Zhang, T., Cheng, J., Hou, Z., Tiwari, P., Pandey, H. M. et al. (2020). Cross-modality paired-images generation and augmentation for RGB-infrared person re-identification. *Neural Networks*, 128, 294–304.

Ye, M., Lan, X., Wang, Z. & Yuen, P. C. (2020). Bi-Directional Center-Constrained Top-Ranking for Visible Thermal Person Re-Identification. *IEEE Transactions on Information Forensics and Security*, 15, 407-419.

Ye, M., Lan, X., Li, J. & Yuen, P. (2018a). Hierarchical discriminative learning for visible thermal person re-identification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

Ye, M., Shen, J., Lin, G., Shao, L. & Xie, X. (2018b). Hierarchical Discriminative Learning for Visible Thermal Person Re-Identification. *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*.

Ye, M., Shen, J., J. Crandall, D., Shao, L. & Luo, J. (2020a). Dynamic Dual-Attentive Aggregation Learning for Visible-Infrared Person Re-identification. *Computer Vision – ECCV 2020*, pp. 229–247.

Ye, M., Shen, J., Lin, G., Xiang, T., Shao, L. & Hoi, S. C. H. (2020b). Deep Learning for Person Re-identification: A Survey and Outlook. *arXiv preprint arXiv:2001.04193*, 0–1.

Ye, M., Shen, J. & Shao, L. (2020c). Visible-infrared person re-identification via homogeneous augmented tri-modal learning. *IEEE Transactions on Information Forensics and Security*, 16, 728–739.

Ye, M., Chen, C., Shen, J. & Shao, L. (2021a). Dynamic tri-level relation mining with attentive graph for visible infrared re-identification. *IEEE Transactions on Information Forensics and Security*, 17, 386–398.

Ye, M., Ruan, W., Du, B. & Shou, M. Z. (2021b). Channel Augmented Joint Learning for Visible-Infrared Recognition. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13567–13576.

You, K., Pan, Y., Wang, Z., Long, M. & Jordan, M. I. (2019). Towards Accurate Model Selection in Deep Unsupervised Domain Adaptation. *International Conference on Machine Learning (ICML)*.

Yu, H., Cheng, X., Peng, W., Liu, W. & Zhao, G. (2023). Modality unifying network for visible-infrared person re-identification. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11185–11195.

Zeng, X., Long, J., Tian, S. & Xiao, G. (2023). Random area pixel variation and random area transform for visible-infrared cross-modal pedestrian re-identification. *Expert Systems with Applications*, 215, 119307.

Zhang, H., Cisse, M., Dauphin, Y. N. & Lopez-Paz, D. (2018). mixup: Beyond Empirical Risk Minimization. *International Conference on Learning Representations*.

Zhang, Y., Deng, B., Jia, K. & Zhang, L. (2021a). Gradual Domain Adaptation via Self-Training of Auxiliary Models. *arXiv preprint arXiv:2106.09890*, 0–1.

Zhang, Y., Deng, W. & Liu, X. (2021b). Unsupervised Domain Adaptation with Progressive Feature Alignment. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2320–2328.

Zhang, Y., Kang, Y., Zhao, S. & Shen, J. (2023). Dual-Semantic Consistency Learning for Visible-Infrared Person ReID. *IEEE Trans. Information Forensics and Security*, 18, 1554-1565. doi: 10.1109/TIFS.2022.3224853.

Zhang, Y. & Wang, H. (2023a, June). Diverse Embedding Expansion Network and Low-Light Cross-Modality Benchmark for Visible-Infrared Person Re-Identification. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2153-2162.

Zhang, Y. & Wang, H. (2023b, June). Diverse Embedding Expansion Network and Low-Light Cross-Modality Benchmark for Visible-Infrared Person Re-Identification. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2153-2162.

Zhang, Y., Yan, Y., Lu, Y. & Wang, H. (2021c). Towards a unified middle modality learning for visible-infrared person re-identification. *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 788–796.

Zhang, Z., Jiang, S., Huang, C., Li, Y. & Da Xu, R. Y. (2021d). RGB-IR cross-modality person ReID based on teacher-student GAN model. *Pattern Recognition Letters*, 150, 155–161.

Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J. & Tian, Q. (2015). Scalable person re-identification: A benchmark. *Proceedings of the IEEE international conference on computer vision*, pp. 1116–1124.

Zheng, L., Yang, Y. & Hauptmann, A. G. (2016a). Person re-identification: Past, present and future. *arXiv preprint arXiv:1610.02984*, 0–1.

Zheng, L., Yang, Y. & Hauptmann, A. G. (2016b). Person re-identification: Past, present and future. *arXiv preprint arXiv:1610.02984*.

Zheng, R., Li, L., Han, C., Gao, C. & Sang, N. (2019). Camera Style and Identity Disentangling Network for Person Re-identification. *BMVC*, pp. 66.

Zheng, Z., Zheng, L., Garrett, M., Yang, Y., Xu, M. & Shen, Y.-D. (2020). Dual-path convolutional image-text embeddings with instance loss. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 16(2), 1–23.

Zhong, Z., Zheng, L., Kang, G., Li, S. & Yang, Y. (2020). Random Erasing Data Augmentation. *AAAI*.

Zhou, K., Yang, Y., Hospedales, T. & Xiang, T. (2020). Deep domain-adversarial image generation for domain generalization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07), 13025–13032.

Zhou, K., Yang, Y., Qiao, Y. & Xiang, T. (2021). Domain generalization with mixstyle. *arXiv preprint arXiv:2104.02008*.

Zhou, K., Liu, Z., Qiao, Y., Xiang, T. & Loy, C. C. (2022). Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zhou, X., Zhong, Y., Cheng, Z., Liang, F. & Ma, L. (2023). Adaptive sparse pairwise loss for object re-identification. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19691–19701.

Zhu, J.-Y., Park, T., Isola, P. & Efros, A. A. (2017). Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. *Computer Vision (ICCV), 2017 IEEE International Conference on*.

Zhu, K., Guo, H., Liu, Z., Tang, M. & Wang, J. (2020a). Identity-guided human semantic parsing for person re-identification. *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pp. 346–363.

Zhu, X., Zheng, M., Chen, X., Zhang, X., Yuan, C. & Zhang, F. (2023). Information disentanglement based cross-modal representation learning for visible-infrared person re-identification. *Multimedia Tools and Applications*, 82(24), 37983–38009.

Zhu, Y., Yang, Z., Wang, L., Zhao, S., Hu, X. & Tao, D. (2020b). Hetero-center loss for cross-modality person re-identification. *Neurocomputing*, 386, 97–109.

Zhu, Z., Jin, Z., Zhang, J. & Chen, H. (2024). Enhancing model interpretability with local attribution over global exploration. *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 5347–5355.

Zhuang, F., Cheng, X., Luo, P., Pan, S. J. & He, Q. (2015). Supervised representation learning: Transfer learning with deep autoencoders. *IJCAI*, 15, 4119–4125.