Deep visual-infrared fusion for multimodal person re-identification in the wild

by

Arthur JOSI

MANUSCRIPT-BASED THESIS PRESENTED TO ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
IN PARTIAL FULFILLMENT OF A MASTER'S DEGREE
WITH THESIS
M.A.Sc.

MONTREAL, AUGUST 16, 2023

ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
UNIVERSITÉ DU QUÉBEC

BOARD OF EXAMINERS

THIS THESIS HAS BEEN EVALUATED

BY THE FOLLOWING BOARD OF EXAMINERS

Mr. Eric Granger, Thesis supervisor
Department of Systems Engineering, École de technologie supérieure

Mr. Rafael Menelau-Cruz, Thesis Co-Supervisor
Department of Software and Information Technology Engineering, École de technologie supérieure

Mr. Stephane Coulombe, Chair, Board of Examiners
Department of Software and Information Technology Engineering, École de technologie supérieure

Mr. Jose Dolz, Member of the Jury
Department of Software and Information Technology Engineering, École de technologie supérieure

THIS THESIS WAS PRESENTED AND DEFENDED

IN THE PRESENCE OF A BOARD OF EXAMINERS AND THE PUBLIC

ON AUGUST 14, 2023

AT ÉCOLE DE TECHNOLOGIE SUPÉRIEURE

ACKNOWLEDGEMENTS

To begin with, I would like to express my deepest gratitude to my advisor Eric Granger, for the trust he placed in me to discover the domain through this thesis, for his guidance in the realization of this project, and his expertise. I would also like to present a special thanks to Rafael M. O. Cruz, co-advisor of this thesis. Always available, fully supportive, very knowledgeable, and accurate.

My warmest gratitude to my colleagues at the Imaging, Vision and Artificial Intelligence Laboratory (LIVIA), and especially to Mahdi Alehdaghi, Felix Remigereau, Madhu Kiran, and Sajjad Abdoli for their support and caring.

My sincerest thanks go to my family and closest friends for their constant encouragement and support during my graduate studies, particularly to Laurence, Ilona, Nicolas, Felix, Jades, Gustave, Guilhem, Nam and Corentin, without whom this work would not exist.

Finally, I would like to thank the Natural Sciences and Engineering Research Council of Canada (NSERC) and the École de Technologie Supérieure (ÉTS) for their financial support.

Fusion profonde visible-infrarouge pour la ré-identification multimodale de personnes dans des conditions réelles

Arthur JOSI

RÉSUMÉ

La ré-identification (ReID) de personnes est une tâche cruciale de vidéo-surveillance permettant de faire correspondre des images d'individus entre elles, ces images provenant de caméras observant des scènes distinctes. Cette tâche pose d'importants défis en raison de facteurs tels que les différentes positions des caméras, les conditions de capture (l'éclairage, les conditions météorologiques, l'arrière plan), les formes corporelles complexes et les différents styles vestimentaires. Ces facteurs entraînent un large éventail de conditions de capture potentielles, conduisant à des bases de données ne couvrant qu'une fraction des scénarios éventuels. Par extension, cela conduit à des données d'apprentissage et d'évaluation de modèles ne convenant pas à la conception d'un système robuste. Sous ces contraintes, un modèle doit être construit pour capturer les caractéristiques personnelles complexes et discriminantes tout en permettant d'effectuer un traitement des données en temps réel.

Parmi les aspects précédents, la modalité visible, couramment utilisée dans les approches traditionnelles, dépend fortement de la luminosité ambiante. Les conditions de faibles luminosités peuvent avoir un impact important sur la qualité des scènes capturées, ce qui se traduit par une ReID imprécise. Cet aspect s'ajoute à la présence potentielle de captures bruitées ou floues, introduisant un obstacle supplémentaire dans la tâche de ré-identification de personnes. Les caméras infrarouges règlent les problèmes liés aux conditions d'éclairage en ne dépendant pas de celles-ci pour encoder la scène, mais ne capturent pas l'information de couleurs et sont affectées de la même manière par différentes corruptions lors de l'encodage. Par conséquent, les capteurs visibles et infrarouges apparaissent comme polyvalents et pertinents dans le contexte de la ReID, mais le fait de s'appuyer uniquement sur l'une de ces modalités compromet l'efficacité de l'approche en extérieur ou sous des conditions de capture complexes.

Dans ce mémoire, la fusion des modalités visible et infrarouge (V-I) est proposée pour relever ces défis. En encodant la scène de manière indépendante et en capturant des vidéos des mêmes individus, les caméras V-I permettent des captures corrélées tout en limitant l'effet des eventuels problèmes d'encodage.

Le chapitre 1 fournit des informations générales sur les modèles d'apprentissage profonds et les techniques pour l'identification des personnes. Ensuite, une étude des techniques de fusion et des techniques relatives à l'évaluation de modèles sous des conditions réalistes est proposée chapitre 2, permettant de soulever les challenges clés du domaine. À ce titre, multiples aspects doivent être pris en compte. Tout d'abord, les modèles de fusion multimodale sont présentés comme négligeant les caractéristiques propre à chaque modalité, se concentrant principalement sur celles partagées entre elles et manquant alors une partie importante de l'information discriminante. De plus, de récentes approchent soulignent la nécessité d'approfondir les protocoles d'évaluation en corrompant artificiellement les ensembles de données tout en mettant en œuvre des stratégies

d'apprentissage spécifiques à cet égard. Toutefois, bien que ces approches aient été explorées et se soient révélées efficaces dans le cadre unimodal, cela reste à explorer pour les algorithmes de fusion, également affectés par l'imprévisibilité du monde réel.

Ce travail présente une nouvelle architecture de modèle multimodal dans le Chapitre 3 afin d'exploiter pleinement les modalités quelles que soient les conditions de captures. Le modèle est composé de trois réseaux de neurones convolutifs, deux se concentrant sur l'extraction de caractéristiques spécifiques à chaque modalité, tandis que le troisième exploite les caractéristiques partagées via une représentation fusionnée. De plus, des approches basées sur le principe d'attention sont étudiées pour permettre une sélection dynamique des caractéristiques, prometteur dans la fusion de données multimodales sous des conditions opérationnelles difficiles. Ces conditions difficiles sont par ailleurs reproduites grâce aux ensembles de données corrompues V-I proposés, reproduisant des conditions réalistes adaptées aux scénarios de caméras co-localisées ou non, et permettant une évaluation approfondie des modèles. Pour les caméras co-localisées, d'éventuelles corrélations de corruptions sont prises en compte, ce qui n'est pas attendu et n'est donc pas appliqué pour les caméras non co-localisées, chaque caméra V et I se trouvant à des positions distinctes. Enfin, nous proposons une approche d'augmentation de données multimodales qui renforce la capacité de généralisation du modèle multimodal en favorisant la collaboration entre les modalités et en préparant le modèle à faire face à des corruptions locales ou globales spécifiques à chaque modalité.

L'utilisation de trois bases de données et de deux scénarios d'évaluations avec des données corrompues comptabilisant vingt différentes corruptions visibles et infrarouges nous permettent de montrer que la configuration multimodale V-I de ReID est une excellente stratégie pour améliorer la précision de la ReID tout en conservant une complexité compétitive. En particulier, avec un apprentissage approprié, le modèle multimodal proposé peut surpasser les systèmes à l'état de l'art sous des conditions idéales tout comme bruitées et difficiles.

Mots-clés: Réseaux neuronaux profonds, Fusion multimodale, Images corrompues, Augmentation de données, Ré-identification visible-infrarouge de personnes

Deep visual-infrared fusion for multimodal person re-identification in the wild

Arthur JOSI

ABSTRACT

Person re-identification (ReID) is a crucial video surveillance task, allowing one to match images of individuals captured by non-overlapping cameras. This task poses significant challenges due to factors such as varying camera positions, capture conditions (e.g., illumination, weather, background), complex body shapes, and diverse clothing styles. These factors result in a wide range of potential capture conditions leading to datasets that only cover a small fraction of the potential scenarios, and consequently to models' learning and evaluation data unsuited to the design of a robust framework. Under these constraints, a cost-effective model must be built to capture the complex and discriminative personal features while allowing to perform real-time data processing.

Among the previous aspects, the visible modality commonly used in traditional ReID frameworks is highly dependent on the prevailing luminosity. Low-light conditions can severely affect the quality of captured scenes, resulting in inaccurate ReID. This introduces an additional obstacle in the person ReID task, compounded by the potential presence of noisy or blurry captures. Infrared cameras can mitigate the issues caused by lighting conditions because they do not rely on light information for scene encoding, but do not capture color information and are similarly affected by sensor encoding issues. Therefore, visible and infrared sensors are versatile and relevant sensors in the context of person ReID, but relying solely on one of these modalities compromises the effectiveness of the framework in outdoor conditions or under complex capture conditions.

In this thesis, the multimodal setting and especially the fusion of visible and infrared (V-I) modalities are considered to address these challenges. By having a distinct encoding process while capturing videos from the same individuals, V-I cameras allow for correlated captures while limiting the effect of eventual encoding issues.

Chapter 1 provides some background on deep learning models and techniques for person ReID. Then, a review of multimodal fusion techniques and real-world data for person ReID is provided in Chapter 2, allowing us to assess the key challenges in the area. As such, multiple aspects must be considered. First, multimodal fusion models are seen to neglect modality-specific features, mostly focusing on shared knowledge instead, and consequently missing an important part of discriminant information. In addition, recent approaches emphasize the need to deepen evaluation protocols by artificially corrupting datasets but also to implement specific learning strategies in this regard. However, while such approaches have been explored and shown to be effective in the unimodal setting, this has yet to be done for fusion algorithms, similarly affected by real-world unpredictability.

This work presents a novel multimodal model architecture Chapter 3 to fully exploit modality knowledge and handle real-world data. The model is composed of three backbones, two

concentrate on extracting modality-specific features, while the third leverage shared knowledge from a fused modality representation. Furthermore, attention-based approaches are investigated to enable dynamic feature selection, which is likely suitable for multimodal feature fusion under challenging operational conditions. These conditions are reproduced through the proposed V-I corrupted datasets that replicate realistic and highly challenging conditions for both co-located and not co-located camera scenarios, allowing an in-depth model evaluation. For co-located cameras, eventual corruptions correlations are considered, not expected, and consequently not applied for not co-located cameras since each V and I cameras are at distinct locations. Finally, a multimodal data augmentation that enhances the multimodal model's capacity for generalization is proposed and works at promoting collaboration among modalities and priming the model to face modality-specific local or global corruptions.

The use of three datasets and two corrupted evaluation scenarios through twenty V and I corruptions allowed us to show that the multimodal ReID strategy can improve ReID accuracy while conserving moderate system complexity. Specifically, with the appropriate learning approach, the proposed multimodal model can outperform related state-of-the-art systems under ideal and challenging noisy real-world conditions.

Keywords: Deep Neural Networks, Multimodal Fusion, Corrupted Images, Data Augmentation, Visible-Infrared Person Re-identification

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AI	Artificial Intelligence
CAM	Class Activation Map
CCD	Correlated Corrupted Dataset
CIM	Cross-Modal Interacting Module
CIL	Consistent identity loss with Inference before bottleneck and Local-based augmentation
CL	Co-Located cameras
CMC	Cumulative Matching Characteristic
CNN	Convolutional Neural Network
CRA	Corruption Robustness Analysis
CV	Computer Vision
DA	Data Augmentation
DL	Deep Learning
DNN	Deep Neural Network
FLOPs	Floating Point Operations per second
I and IR	Infrared representation of Image
KNN	K-Nearest Neighbours
LOOQ	Leave-One-Out Query
LUPI	Learning Using Privileged Information
mAP	mean Average Precision

MCAT	Multimodal Cross-Attention Transformer
MDA	Multimodal Data Augmentation
M-PATCH	Multimodal Patch Mixing
mINP	mean Inverse Negative Penalty
ML	Machine Learning
MLP	Multi-Layer Perceptron
MMSF	Multimodal Middle Stream Fusion
MMTM	MultiModal Transfer Module
MS-REA	Multimodal Soft Random Erasing
MSA	Multi-headed Self-Attention
MSE	Mean Squared Error loss function
MSAF	Multimodal Split Attention Fusion
MSAT	Multimodal Self-Attention Transformer
NCL	Not Co-Located cameras
NNs	Neural Networks
P-R	Precision-Recall
PP	Percentile Point
PR	Pattern Recognition
ReID	Re-Identification
RGB-D	RGB-Depth (Visual and Depth representation of Image)

RNN	Recurrent Neural Networks
ROC	Receiver Operating Characteristic
ROI	Region of Interest
S-PATCH	Self Patch Mixing
S-REA	Soft Random Erasing
SGD	Stochastic Gradient Descent
SNN	Siamese Neural Network
SVM	Support Vector Machines
UCD	Uncorrelated Corrupted Dataset
V and RGB	Red Green Blue (Visual Representation of Image)
V-I	Visible - Infrared
ViT	Vision Transformer
WSOL	Weakly Supervised Object Localization

LIST OF SYMBOLS AND UNITS OF MEASUREMENTS

$\mathbb{R}$	Set of real numbers
$\mathbb{N}$	Set of natural numbers
$\mathbf{F}$	Input data (Image data or feature map)
$\mathbf{F}^k$	k-th channel of the feature maps $\mathbf{F}$
$\mathbf{F}_{(x,y)}$	C-dimensional slice of the feature maps $\mathbf{F}$ at spacial location (x, y), C being $\mathbf{F}$ channel size
$\mathbf{F}^k_{(x,y)}$	k-th channel of the feature maps $\mathbf{F}$ at spacial location (x, y)
$\mathbf{F_V}$	Visible feature map
$\mathbf{F_I}$	Infrared feature map
$\mathbf{F}^\ell_\mathbf{V}$	Visible feature map before layer ℓ
$\mathbf{F}^\ell_\mathbf{I}$	Infrared feature map before layer ℓ
$w\mathbf{F}^\ell_\mathbf{V}$	Weighted visible feature map before
$w\mathbf{F}^\ell_\mathbf{I}$	Weighted infrared feature map before layer ℓ
$\mathbf{S}^\ell_\mathbf{V}$	Visible feature map split before layer ℓ
$\mathbf{S}^\ell_\mathbf{I}$	Infrared feature map split before layer ℓ
$w\mathbf{S}^\ell_\mathbf{V}$	Weighted visible feature map split before layer ℓ
$w\mathbf{S}^\ell_\mathbf{I}$	Weighted infrared feature map split before layer ℓ
y_i	Target value for the i-th input sample
$\hat{y}_i$	Predicted target value for the i-th input sample
ℓ	Position of the convolution layer considered in the backbone

R_q	Number of persons to retrieve in the gallery for a query q
R_q^H	Position of the hardest match in the ranking list for a query q
$\mathbf{Z}$	Transformer input sequence
$\mathbf{Z_i}$	i-th transformer input sequence
Q	Number of query image
W	Input width
C	Channel size
H	Input height
$\mathbf{Q}$	Set of queries matrix
$\mathbf{K}$	Set of keys matrix
$\mathbf{V}$	Set of values matrix
$\mathbf{p}_i$	i-th transformer patch
$\mathbf{p_i^r}$	i-th refactored transformer patch
$\mathbf{p_{cls}}$	Classification token
$\mathbf{p_{cls}^r}$	Refactored classification token
$\mathbf{P}$	Learnable positional embedding
L_c	Number of cross-attention transformer modules
L_s	Number of self-attention transformer modules
$\mathbf{f}$	Feature vector
$\mathbf{f_V}$	Visible feature vector

$\mathbf{f_I}$	Infrared feature vector
$w\mathbf{f_V}$	Weighted visible feature vector
$w\mathbf{f_I}$	Weighted infrared feature vector
S	Similarity measure
d	Vector dimension
$d_{euclidean}$	Similarity measure with Euclidean distance
d_{cosine}	Similarity measure with cosine distance
$\mathcal{L}_{\text{CE}}$	Cross-entropy loss
$\mathcal{L}_{\text{CE\_ls}}$	Cross-entropy loss with regularization via label smoothing
$\mathcal{L}_{\text{tri}}$	Triplet loss
$\mathcal{L}_{\text{BH\_tri}}$	Batch-hard triplet loss
$\hat{\mathbf{y}}$	Predicted output feature vector
AvgPool	Average pooling layer
ReLU	Rectified Linear Unit
σ	Sigmoid function
$\odot$	Element-wise product
$\mathbf{J}^{\ell}$	Global modality shared channel descriptor before layer ℓ
$\mathbf{J}_{\mathbf{V}}^{\ell}$	Global visible channel descriptor before layer ℓ
$\mathbf{J}_{\mathbf{I}}^{\ell}$	Global infrared channel descriptor before layer ℓ

INTRODUCTION

0.1 Multimodal Person Re-Identification

Person re-identification (ReID) refers to the task of matching images of individuals captured through a non-overlapping set of cameras. This image retrieval task is essential for many fields, such as sports analytics (Penate-Sanchez *et al.*, 2020; Giancola *et al.*, 2022) and video surveillance (Ye *et al.*, 2021) in sensitive locations like airports and train stations. To build a person ReID system (Fig. 0.1), the scene is initially captured by a set of distributed cameras that encode the scene through different clips that are short segments of a full video. These clips are then analyzed using an object detection and multi-object tracking algorithm, enabling the extraction of a set of consecutive bounding boxes (or Regions Of Interest (ROIs)) for each individual present in the clip, the sets being referred as a tracklets. Utilizing ROIs allows the system to focus on one individual at a time and avoids incorporating additional noise from surrounding and unrelated elements.

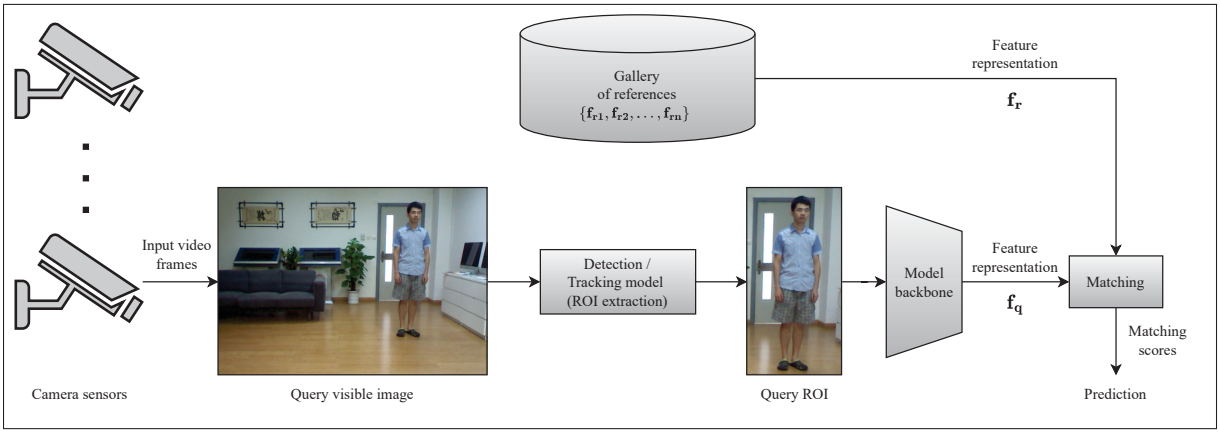

Figure 0.1 Block diagram of image-based person ReID system. The gallery contains feature representation of ROIs from distinct users, each appearing one or more times. Given a query individual capture, its ROI is first extracted and then processed to extract a feature embedding or representation $\mathbf{f_q}$, matched with the reference embedding $\mathbf{f_r}$ to produce a ranking list, that is used for final prediction.

The typical video-based system directly focuses on matching the tracklets from distinct cameras or scenes, aiming to identify similarities and determine if an individual appearing in the tracklets of different cameras is the same. In the case of image-based systems, on which we focus in our approach, a single individual's ROI is used for matching. In practice, the ROIs are usually assumed to be already extracted, and the emphasis is on analyzing and matching them. To match individuals, reference shots (queries) are compared to all other shots in the gallery, the gallery shots being pre-processed and stored in the system from previous records. This comparison process generates a ranking list for each query, with the top-ranked matches likely representing the same individuals.

Over the last decade, Deep Learning (DL) and especially Convolutional Neural Networks (CNNs) (LeCun *et al.*, 1998) and Vision Transformers (ViTs) (Dosovitskiy *et al.*, 2020) have allowed for remarkable advancements in Computer Vision (CV) tasks, such as classification (Sen *et al.*, 2020), object detection (Zou *et al.*, 2023), and segmentation (Minaee *et al.*, 2021). Indeed, DL models have achieved impressive accuracy through various improvements of their layers structure and through end-to-end learning. However, even though person ReID benefits from these model refinements (Ming *et al.*, 2022), the task remains highly challenging in real-world applications. In particular, the human body shape varies much from one individual to another and especially from one angle to another. Moreover, the capture conditions and the eventual variations in image data are numerous. To cite just a few of those challenges, one can think of the varying pose, scale, resolution, weather, luminosity, and of the eventual blur or occlusions that may occur.

The declining cost of cameras over time and the advancement in sensor technologies has contributed to the growing interest and rapid progress in the fusion of discriminant, diverse, and complementary modalities. Multimodal fusion, which involves combining knowledge captured from different sensors, is a well-established approach for enhancing the accuracy and robustness

of pattern recognition (PR) systems for various tasks (Atrey *et al.*, 2010; Baltrušaitis *et al.*, 2018; Ma *et al.*, 2019; Wang, 2021). While the fusion of multiple modalities allows leveraging the complementarity of diverse sensor data, often leading to improved accuracy, it increases the system's complexity. In this case, systems design is more challenging for real-time with resource-limited systems. However, most of the time, each modality data can be processed in parallel through modality-specific systems, ensuring comparable encoding effectiveness despite the higher computational needs.

Multimodal systems are also relevant in the context of person ReID and have been explored through the fusion of different modalities. Although limited, some literature focuses on the fusion of the [1]RGB and Depth (D) modalities (Uddin *et al.*, 2023). The D modality allows encoding the scene even under low-light or dark conditions, which is not feasible for the RGB modality. However, the D sensors only allow for short-distance captures and are consequently less valuable in outdoor environments. In contrast, the [2]IR modality allows for night-time encoding and for long-distance captures. For this reason, the infrared modality has been widely used to perform Visible-Infrared (V-I) cross-modal person ReID (Ye *et al.*, 2021), for which the idea is to match individuals from one modality to another. Still, the RGB-IR setting remains even less explored than the RGB-D setting in the context of a multimodal person ReID (Uddin *et al.*, 2023), despite its potential. Indeed, beyond the potential for using IR for long-distance and night capture, the numerous studies on V-I cross-modal person ReID confirm the capacity to discriminate using IR. The complementary information shared among modalities can reinforce our confidence in multimodal systems using V-I features. Moreover, the distinct modality encoding should allow for complementary knowledge and lead to improved accuracy.

---

[1] The RGB and V acronyms are used interchangeably in this thesis for the visible modality.

[2] The IR and I acronyms are used interchangeably in this thesis for the infrared modality.

0.2 Problem Statement

In conventional video surveillance systems, the ability to provide accurate person ReID solely relies on the RGB modality. RGB sensors capture discriminant person visualizations from a given scene under good lighting and visibility, but can easily lack precision in low light conditions (e.g., weather change and night-time), or when the sensors are subject to encoding data corruption such as blur and noise. While these events are expected to occur in real-world environments, their intensity is somewhat unpredictable, as is the case with weather or encoding issues. In any case, such events can strongly affect the ReID accuracy (Chen *et al.*, 2021) since the models only rely on the corrupted RGB images. In this thesis, the challenges of person ReID are addressed by combining the V and I modalities. Although several methods have been proposed in the literature for the fusion of I and V modalities, it is unclear how to develop a cost-effective fusion for V-I Re-ID in real-world conditions.

Deterministic fusion refers to methods in which predefined rules determine the fusion outcome, and do not involve learning parameters for the fusion model. These approaches are usually easy to develop, either by directly combining signal sources at a sensor level (Lohweg & Mönks, 2010), or through feature-level or scores-level fusion, using for example feature concatenation or majority voting decision rules (Snoek *et al.*, 2005; Nojavanasghari *et al.*, 2016; Wörtwein & Scherer, 2017), respectively. Unlike deterministic fusion, learning-based fusion involves models that automatically learn the fusion rules from the data. Such fusion allows for the models' consideration of subtle and implicit modality correlations and can be performed at different stages of the feature extraction process (named intermediate or model-level fusion) (Ramachandram & Taylor, 2017; Gandhi *et al.*, 2022), often enhancing the fusion capacity.

The state-of-art multimodal models mainly focus on attention-based fusion (Joze *et al.*, 2020; Su *et al.*, 2020; Mocanu & Tapu, 2022), where attention mechanisms allow for a dynamic feature selection between modalities (Guo *et al.*, 2022). In this case, the fusion often follows

a model-level fusion. Recently, Baltrušaitis *et al.* (2018) observed how model-level fusion models were mainly focusing on modality-shared information, losing partial knowledge through modality-specific information. While using attention comes up as an intuitive strategy when dealing with real-world data as it supposedly selects the relevant information dynamically, the models may over-focus on the modality-shared information, biased by its recurrence among modalities. Several transformer-based models avoid such issues and ensure the exploitation of the modality-specific knowledge by using distinct streams for the specific and shared knowledge (Wei *et al.*, 2020b; Sun *et al.*, 2021; Lian *et al.*, 2021), but require large models trained with large-scale data that are too complex for most real-world applications.

Despite the recent advancements regarding multimodal fusion and the potential of the V-I setting for person ReID, multimodal approaches using solely the RGB and IR modalities are limited to the deterministic model proposed by Nguyen *et al.* (2017). Therefore, it is essential to develop V-I person ReID methods that consider the recent state-of-art multimodal fusion techniques and real-world requirements, while being assured of their ability to leverage the shared and specific modality information and address.

The learning and evaluation data often fail to capture the complexity and variability of the real world, resulting in models that do not generalize well. This lack of data realism comes from controlled data collection environments but also from the simple fact that the range of real-world events is too wide to get fully captured. Recent works have introduced corrupted evaluation protocols to address this issue, emphasizing the importance of considering such data during model evaluation (Hendrycks & Dietterich, 2019; Michaelis *et al.*, 2019). While models may perform well on the originally collected data, their robustness and adaptation may be lacking when faced with corrupted datasets. This issue has also been highlighted in the context of a person ReID, where Chen *et al.* (2021) provided corrupted datasets and conducted an extensive evaluation of state-of-the-art models. However, these evaluation sets are limited to the visible

modality, making challenging the design of a robust multimodal framework with the current literature.

To handle real-world condition data, data augmentation (Ciregan *et al.*, 2012) is a powerful approach that does not bring supplementary complexity to the pipeline. Data augmentation up-samples the learning data quantity through various data transformations. This way, a framework encounters a wider variety of cases and usually benefits from a higher generalization power. Various ways to augment data were proposed for unimodal approaches, including image distortions (Hendrycks *et al.*, 2019), noise augmentation (Rusak *et al.*, 2020), and local occlusions (Zhong *et al.*, 2020; Chen *et al.*, 2021). When two or more modalities are considered, the data augmentation can likely be tailored to make the model learn intra-modality recurrent and reliable features, but also learn how to well select the multimodal knowledge when the modalities discriminant power is punctually imbalanced. This aspect has yet to be explored since state-of-art multimodal data augmentation (MDA) techniques have been proposed to either increase the multimodal learning data quantity (Xu *et al.*, 2020) or to reduce the modality domain gap for cross-modal approaches (Nakamura *et al.*, 2022). Nevertheless, considering the value of these approaches in the unimodal setting, there is significant potential to make a substantial difference in how multimodal models handle data.

0.3 Proposed Approach

This research introduces a novel V-I multimodal person ReID model, derived from existing transformer architectures (Wei *et al.*, 2020a; Sun *et al.*, 2021; Lian *et al.*, 2021), that utilizes three CNN backbones for feature extraction. The objective of this model is to leverage both the V and I modality-specific knowledge, using backbones dedicated to each individual modality, and the modality-shared knowledge obtained from a middle backbone that exploits the correlations between modalities. By employing a modality-specific backbone, the model ensures that a corrupted modality does not affect the entire feature representation. The benefits

of such architecture are explored while being compared to state-of-art attention-based CNN and transformer models from the literature that we adapted for V-I person ReID. These alternative models are promising because they provide distinct ways of utilizing knowledge by involving dynamic feature extraction mechanisms.

To evaluate the performance of the V-I ReID model effectively, corrupted multimodal datasets are designed in this thesis. These datasets are carefully constructed to respect the definition of each modality in real-world scenarios, which may involve co-located or non-co-located cameras. Co-located cameras imply that corruptions may be correlated across modalities, such as occlusion corruption appearing on both V and I cameras since cameras are in the same location. On the other hand, non-co-located cameras capture different scenes from the V to the I camera. Consequently, each modality is independently corrupted in the related dataset. The designed datasets prove to be highly relevant in the analysis of multimodal models in real-world V-I person ReID applications.

Addressing real-world data uncertainties and enhancing the generalization capability of models are crucial tasks. To accomplish this, one powerful approach is by using data augmentation, which offers an effective approach that does not increase the complexity of the inference model. In this thesis, we propose a specific augmentation strategy called Masking and Local Multimodal Data Augmentation (ML-MDA), which is tailored to learning multimodal data with real-world corruptions. The impact of ML-MDA on the accuracy of multimodal person ReID using both clean and corrupted data is evaluated. Furthermore, this approach is compared to state-of-art models trained with unimodal data augmentation strategies. The results show that our ML-MDA is more advantageous due to its multimodal nature and the exploration opportunities it enables.

0.4 Organization and Contributions

This thesis is manuscript based and is structured into three chapters and three appendices.

Chapter 1 provides the background knowledge needed to understand the research work. As the thesis focuses on multimodal fusion, specifically in the context of V-I person ReID, it defines CV and Machine Learning (ML) algorithms along with the related concepts.

Chapter 2 presents a comprehensive analysis of the state-of-the-art literature on multimodal fusion, multimodal person ReID, and techniques for evaluating and dealing with real-world data. Research gaps are identified through a critical analysis of the state-of-art literature.

Chapter 3 introduces **a new multimodal model called Multimodal Middle Stream Fusion (MMSF)** based on insights from the existing literature and with the aim of building a robust multimodal model. Attention-based models from other tasks are also adapted to person ReID, as attention mechanisms have great potential for handling corrupted data. To conduct a realistic analysis of the models' performance and account for Not Co-Located (NCL) and Co-Located (CL) cameras, **two corrupted datasets are created for V-I ReID**. **The Uncorrelated Corrupted Dataset (UCD)** applies corruption independently on each camera pair, suited for the NCL scenario as each V and I camera is differently located. **The Correlated Corrupted Dataset (CCD)** considers eventual corruption correlations that may exist between the V and I cameras of a camera pair, especially under the CL camera scenario. To address the challenges posed by the dataset and real-world conditions, **a MDA strategy is proposed**, which involves local occlusions and modality masking. Through our fusion models, particularly the MMSF architecture, and with the proposed MDA, the relevance of the multimodal setting is shown under normal and highly challenging conditions with corrupted V-I data. The content of this chapter has been submitted to the International Journal of Computer Vision (IJCV) special issue on multimodal learning in April 2023.

Appendix I is preliminary work related to Chapter 3. It compares the performance of a basic fusion model against state-of-art person ReID models on the CCD dataset (referred to as "-C*" in this preliminary work). The appendix delves into the experiments which led to the proposed

MDA, namely the Masking and Local Multimodal Data Augmentation (ML-MDA). Results show that the multimodal approach outperforms the unimodal state-of-art models in various settings, except for the most challenging scenario involving the corruption of the V and I modalities together and at all times. Still, the performance achieved by the simple fusion model was shown to be promising for future and more developed multimodal strategies. The contents of this appendix have been published in the Winter Conference on Applications of Computer Vision (WACV) "Real-World Surveillance: Applications and Challenges" workshop in January 2023 (Josi *et al.*).

Appendix II presents three multimodal hybrid architectures (combining CNN and transformer-based models), and compares them to the best CNN-based models so far in terms of accuracy and complexity. This appendix supports the focus on CNN-based models.

Appendix III serves as supplementary material for Chapter 3, providing additional details and experiments to support the main manuscript. The importance of the MMSF fusion position regarding the CL and NCL camera scenarios is assessed. Unlike Chapter 3, which summarizes the experimental results graphically, this appendix presents a detailed numerical analysis of the model's performance and complexity. Details on the infrared corruptions are also provided, along with a qualitative analysis based on CAMs produced from clean and corrupted V-I pairs.

CHAPTER 1

BACKGROUND

1.1 Visible and Infrared Spectrum

The electromagnetic spectrum encompasses a wide range of wavelengths (Elert, 1998) that have been separated into different regions (Fig. 1.1), two significant regions being the visible and the infrared spectrum. The visible spectrum corresponds to the wavelengths the human eyes are capable of perceiving, being approximately between 400 and 700 nanometers. Traditional RGB cameras capture images within this spectrum by relying on sensors that are sensitive to these electromagnetic wavelengths.

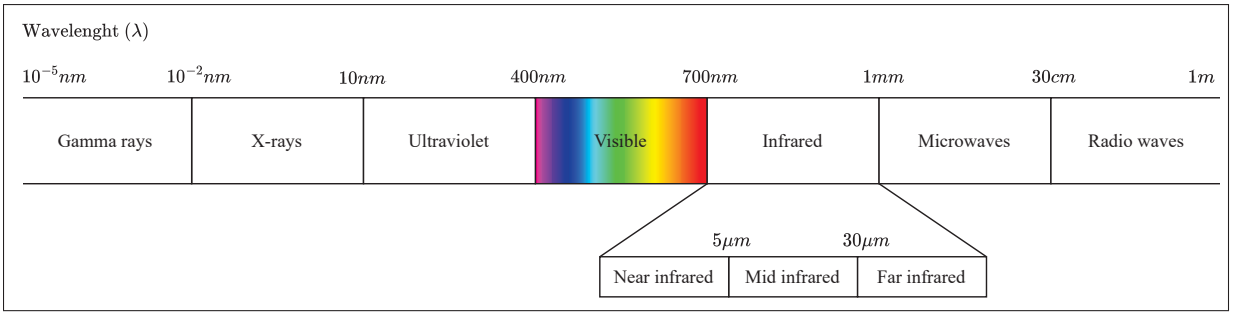

Figure 1.1 Electromagnetic spectrum regions

In contrast, the infrared spectrum extends beyond the range of human vision, typically starting around 700 nanometers and continuing up to 1-millimeter wavelength. This portion of the spectrum is usually subdivided into three regions, the near-infrared, mid-wavelength infrared, and far-infrared. One must notice that the division of these regions may vary depending on the literature and applications.

Infrared cameras are specialized devices designed to capture infrared radiation in the near-infrared or other portions of the infrared spectrum. Thermal cameras are a specific type of infrared camera that is designed to detect and display thermal energy or heat emitted by objects. Thermal cameras operate in the long-wave infrared and, consequently, in the infrared spectrum's

mid and far infrared regions. In practice, for infrared cameras, the collected data is processed, and the infrared electromagnetic wave gradient is displayed using a color gradient or grayscale. Since infrared cameras do not rely on visible electromagnetic waves, a representation can be produced even in dark environments. This aspect, combined with the ability to rely on heat emitted by objects, make them useful in various applications such as night-vision surveillance (Zhang *et al.*, 2018; Krišto *et al.*, 2020) and heat analysis for facility inspection (Wang *et al.*, 2010) or environmental monitoring (Coppola *et al.*, 2016; Valade *et al.*, 2019). The obtained representation enables the identification of patterns that are not naturally visible through the human eyes and classic visible cameras.

1.2 Deep Learning Models

1.2.1 Machine Learning

Machine learning (ML) lies in the field of artificial intelligence (AI) and focuses on developing algorithms and models that learn mappings from data to provide decisions without being explicitly programmed. In Pattern Recognition (PR) applications, ML models have been mostly developed for classification, regression, and clustering tasks. For instance, popular ML models for pattern classification include k-nearest neighbor (KNN), artificial NNs, decision trees, linear regression, or support vector machines (SVM) (Cortes & Vapnik, 1995). These models are able to learn complex mappings from real data, allowing them to generalize and make accurate predictions on unseen examples.

Inspired by the structure of the human brain, artificial NNs models are built from multiple layers of interconnected neurons. An artificial neuron is being assigned a set of weights $\mathbf{w} = (w_1, w_2, ..., w_n)$ and a bias term b. The weights and bias are learnable parameters applied to the input data to transform it and exploit it to perform the objective task. For a given input vector $\mathbf{x} = (x_1, x_2, ..., x_n)$, those parameters are applied as follows:

$$y = \sigma(\sum_{i=1}^{n} w_i x_i + b) \tag{1.1}$$

where σ is an activation function, allowing to introduce non-linear relationships.

One common type of NNs for classification and regression is the multi-layer perceptron (MLP), which consists of an input layer, one or more hidden layers, and an output layer. MLPs are considered shallow NNs, as they typically contain only a few hidden layers, but are already considered universal approximators of highly non-linear functions (Gardner & Dorling, 1998). Later, given the increasing availability of computational power, deep NNs (DNNs) with more hidden layers were developed. While other architectures usually need prior feature extraction using distinct and eventually handcrafted approaches, DNNs differ by being eventually trained end-to-end. While traditional ML models typically rely on handcrafted features or learned subspace projections for feature extraction, DNNs introduce end-to-end training to learn feature embeddings along with task-specific mappings.

In recent years, DL which encompasses a wide range of ML models, has provided state-of-art models for many tasks and applications. One of the key advantages of DL models is their scalability to a wide range of tasks and domains through the extraction of rich feature representations from the data. Human experts cannot easily craft such discriminant and abstract features through engineering. Hence, the development of DL has revolutionized many fields, being highly successful in various applications and achieving state-of-art results in benchmark tasks such as image classification (Sen *et al.*, 2020), object detection (Zou *et al.*, 2023), and sentiment analysis (Birjali *et al.*, 2021).

1.2.2 Supervised learning

This thesis focuses on DL models that perform supervised learning using labeled image datasets. Supervised learning involved labeled datasets that can be defined as $\mathcal{D}_s = \{(x_i, y_i)|i \in \{1, 2, ..., N_s\}\}$, where x_i represents an input data, y_i its corresponding target, and N_s the number of dataset pair samples. This labeled information is utilized to evaluate and improve the model's prediction capacity through cost functions, also called loss functions. Indeed, once computed,

the cost function values can be used by the model to optimize itself, updating the weights and biases of its neuron layers with back-propagation (Rumelhart *et al.*, 1986).

Classification is one of the main tasks in supervised learning (Sen *et al.*, 2020; Xie *et al.*, 2020), for which the models are expected to predict the targets or labels for unseen input data points once optimized. The optimization of models for classification often relies on the cross-entropy loss, noted $\mathcal{L}_{\text{CE}}$ and defined as follows for a given sample:

$$\mathcal{L}_{\text{CE}} = -\sum_{i=1}^{N_c} q_i \cdot log(p_i) \tag{1.2}$$

where i is the index for each class, $\mathbf{p} = (p_1, ..., p_{Nc})$ the output vector of a classifier containing per-class probability values, $N_c \in \mathbb{N}$ the number of classes, and $\mathbf{q} = (q_1, ..., q_{Nc})$ ground truth label vector, containing values equal to 1 if it is the true class and 0 otherwise. Using the cross-entropy with regularization via Label smoothing loss (Szegedy *et al.*, 2016), a cross-entropy loss variation, can eventually allow for better model generalization (Müller *et al.*, 2019). In this case, the true label vector $\mathbf{q}$ is replaced by a smoothed label vector $\mathbf{q}' = (q'_1, ..., q'_{Nc})$ to encourage the model to be less confident in its prediction, resulting in a better model generalization. The smoothed label vector $\mathbf{q}'$ can be defined as follows:

$$q'_i = (1 - \epsilon)q_i + \frac{\epsilon}{N_c} \tag{1.3}$$

where ϵ is the label smoothing parameter.

Regression is another well-known task (Mendes-Moreira *et al.*, 2012; Čížek & Sadıkoğlu, 2020) for which the objective is to optimize a function that maps input data to continuous outputs. For regression, the used loss function is usually different, the mean squared error (MSE) loss being a classic cost function in this context, defined as follows:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{1.4}$$

with N being the number of samples, $\{y_i\}_{i=1}^{N}$ the target values for each sample, and $\{\hat{y}_i\}_{i=1}^{N}$ the predicted target values each sample.

Well-known ML and DL models that perform supervised learning for classification and regression include CNNs for image processing, recurrent NNs (RNNs) for sequential or time series data such as natural language, and transformer networks for these two applications. Since person ReID is part of the image processing tasks, CNNs and Transformers models will be detailed in the following sections. Before presenting those models, the concept of attention is detailed since the attention principle is a key concept for transformer models and recent CNN models.

1.2.3 Attention mechanisms

Observing a specific part of an environment or giving more attention to one or another cue of any object we seek to identify is a natural process (Rensink, 2000). So it is for language processing, for which more or less importance is instinctively given to some words in a sentence to quickly understand it (Dabre *et al.*, 2020). Similarly, while translating from one language to another, some words require specific attention. For example, French language nouns are assigned a gender (masculine or feminine) indicated by the use of specific indefinite articles such as "un" (masculine) and "une" (feminine). Therefore, giving special attention to nouns and their gender when translating from English to French is essential. These are well-oiled mechanics, instinctively developed while learning and growing. Unsurprisingly, multiple researchers working on DL models took inspiration from those natural behaviors, mimicking human attention mechanisms through the design of distinct modules allowing the models to dynamically select the meaningful knowledge within each input (Guo *et al.*, 2022).

In the latest DL models, attention mechanisms become omnipresent, as it is the case in CNN models' architectures (Jaderberg *et al.*, 2015; Hu *et al.*, 2018; Li *et al.*, 2019; Zhang *et al.*, 2020; Niu *et al.*, 2021). This also allowed for the design of transformer models, mainly relying on this concept through their self-attention modules (Vaswani *et al.*, 2017; Dosovitskiy *et al.*, 2020; Han *et al.*, 2020). While attention allows for better selecting the knowledge within a single modality,

one can notice how the dynamic knowledge selection transfers really well to the multimodal setting and its challenges (Gu *et al.*, 2018; Joze *et al.*, 2020; Su *et al.*, 2020). For any multimodal framework, each modality may be differently informative from one input to another, making the dynamic selection of the relevant knowledge across modalities highly suitable.

The rest of this section provides a summary of the CNN and transformer models.

1.2.4 Convolutional Neural Networks

Convolutional Neural Networks (CNNs) (LeCun *et al.*, 1998) are a powerful class of neural networks that have revolutionized the field of Computer Vision (CV). In fact, this architecture is specifically designed to process data with a spatial or temporal structure, including images, but also some types of audio and video data, as it allows for 2D or 3D data structure (Tran *et al.*, 2015). This model specificity has enabled significant progress in solving challenging tasks previously hardly or only partially tackled, like for classification with AlexNet (Krizhevsky *et al.*, 2017), or for image segmentation with the fully convolutional network model proposed by Long *et al.* (2015). The central idea behind CNNs is to learn local and translation-invariant features from raw data automatically. Their architecture consists of multiple layers, typically including convolutional, activation, pooling, and fully connected layers. each of these layers is detailed in the following part.

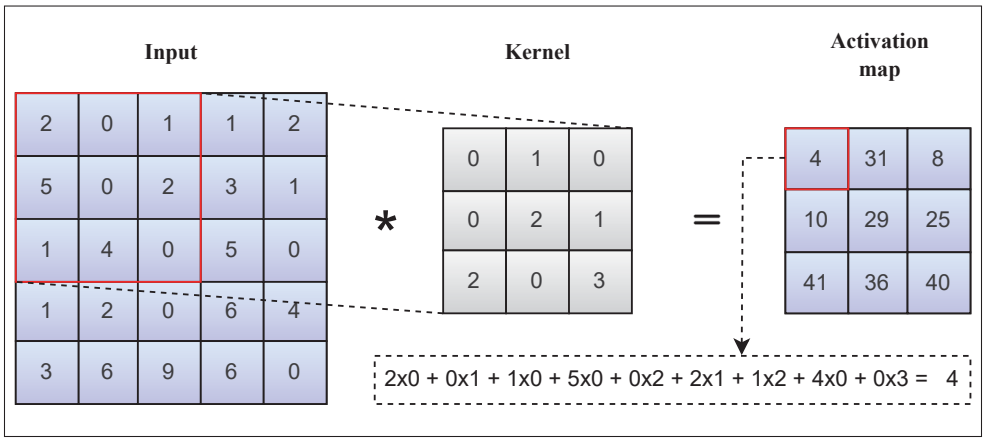

Figure 1.2 Convolution operation representation with a filter size of 3, a stride of (1, 1), and a padding of 0.

In the convolutional layers, filters (also referred to as kernels) are applied to the input. Each filter is learned and updated during the training phase (using back-propagation (Rumelhart *et al.*, 1986)) to recognize specific and discriminant data patterns. The filters allow the extraction of local features by sliding over the input and performing the convolution operation, as represented in Fig. 1.2 and defined Eq. 1.5. In practice, three distinct parameters must be selected: the size, the stride, and the padding. The size fixes the size of the filters. The stride influences how the filter slide over the input by fixing the pixel number by which the filter shifts from one operation to another. The padding controls the spatial dimension of the output by adding borders to the initial input with zero pixel values.

$$y = \sigma(\mathbf{W} * \mathbf{X} + b) \tag{1.5}$$

With σ being a non-linear activation function, $\mathbf{X}$ an input matrix, $\mathbf{W}$ a weight matrix, $*$ the convolution operation, and b a bias term.

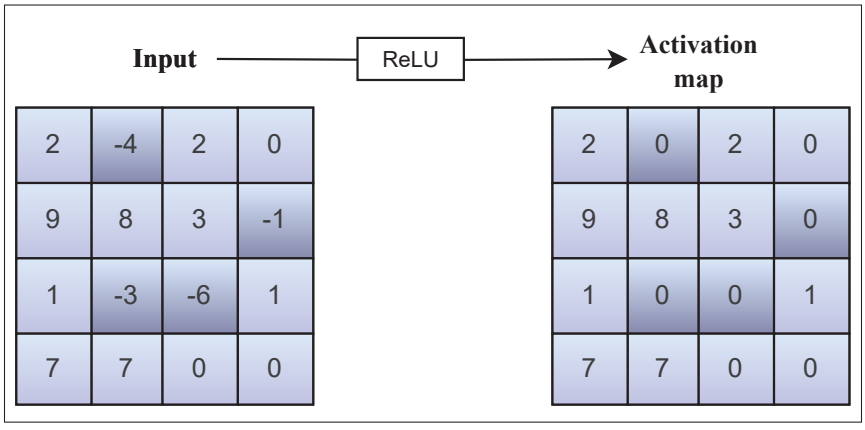

Figure 1.3 Application of the ReLU activation function.

Activation functions introduce non-linearities into the CNN, allowing it to model complex relationships and capture non-linear features in the data. While many activation functions exist, the ReLU activation function (Nair & Hinton, 2010) is the most used in literature as it allows for efficiency through a simple computation and since it allows for conserving a good range

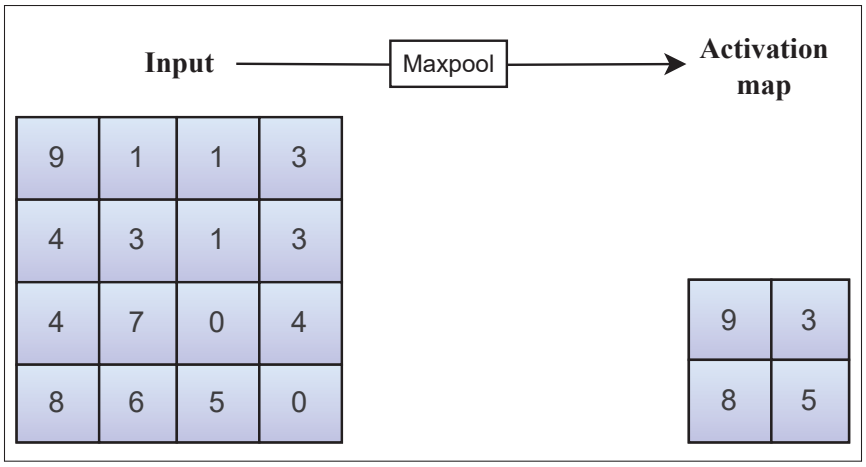

Figure 1.4 Max pooling operation with a stride of 2.

of values for the output. Indeed, this function replaces negative activation values by 0 while it conserves the others (Fig. 1.3).

The pooling layers (Fig. 1.4) downsample the feature maps obtained from the convolutional layers, reducing the spatial dimensionality of the data and making the network more robust to variations in the input. Pooling is typically performed by taking either the maximum value (max pooling) or the average value (average pooling) within a small neighborhood (filter size). The stride is a parameter to set as well for the pooling layers.

Finally, the fully connected layers use the high-level features extracted by the previously presented layers to make a prediction about the initial input. These layers learn a set of weights, which is applied to the input by matrix multiplication.

In summary, the architecture of a typical CNN involves applying convolutional and pooling layers to extract and down-sample features, activation functions to introduce non-linearity, and fully connected layers to make predictions. These operations and their effect on the input image are provided in Fig. 1.5. One can observe the evolution of feature map dimensions through convolution and pooling operations along with the feature vector creation using a flattening operation and its final processing using a MLP.

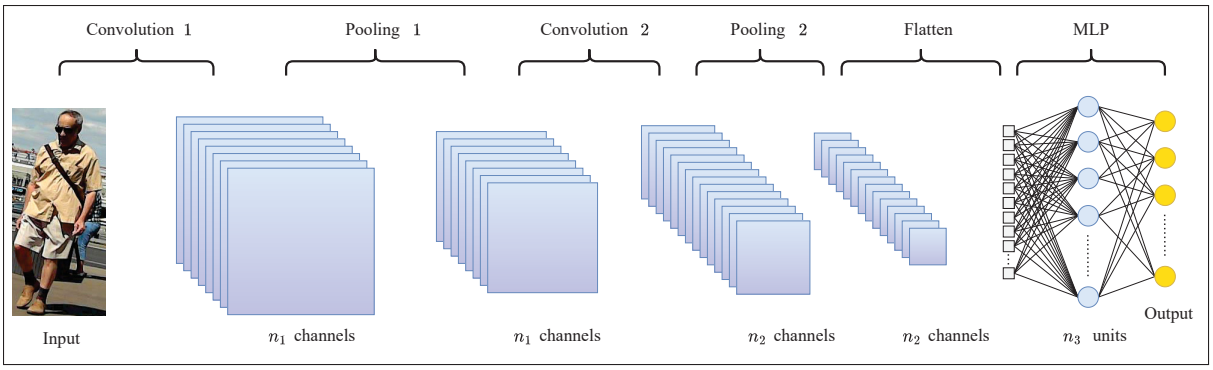

Figure 1.5 Block diagram of a CNN for image classification, and its main operations to predict a class label given an input image. CNNs are generally composed of several convolution, pooling, and ReLU layers, followed by fully connected layers (like an MLP). The channel size of activation maps increases after convolutions and their spatial dimension reduces after pooling operations. The activation maps are finally flattened and passed through a MLP for classification.

Even if the CNN's basic operations and layers are well known, the design of a CNN model remains delicate. Among CNNs, various models architectures have been provided by researchers over the years, such as VGG (Simonyan & Zisserman, 2014), ResNet (He *et al.*, 2016), and AlexNet (Krizhevsky *et al.*, 2017). In practice, each architecture may have its own advantages depending on the task considered and the available data. Consequently, the models must be chosen wisely based on datasets and empirical studies.

1.2.5 Vision Transformers (ViTs)

Transformers are specific DL model architectures originally proposed for natural language processing tasks by Vaswani *et al.* (2017). These models gained popularity for their strong ability to handle sequenced data. From this success, ViT (Fig. 1.6) was developed by Dosovitskiy *et al.* (2020), adapting the original transformer architecture to the CV field.

The key component of transformer models is self-attention and lies in a block of layers called Multi-headed Self-Attention (MSA) (Vaswani *et al.*, 2017). Thanks to the self-attention mechanisms, the model can selectively attend to and weigh different parts of the input data. To do so, the input must be first subdivided into patches. Since visual data is not sequenced by

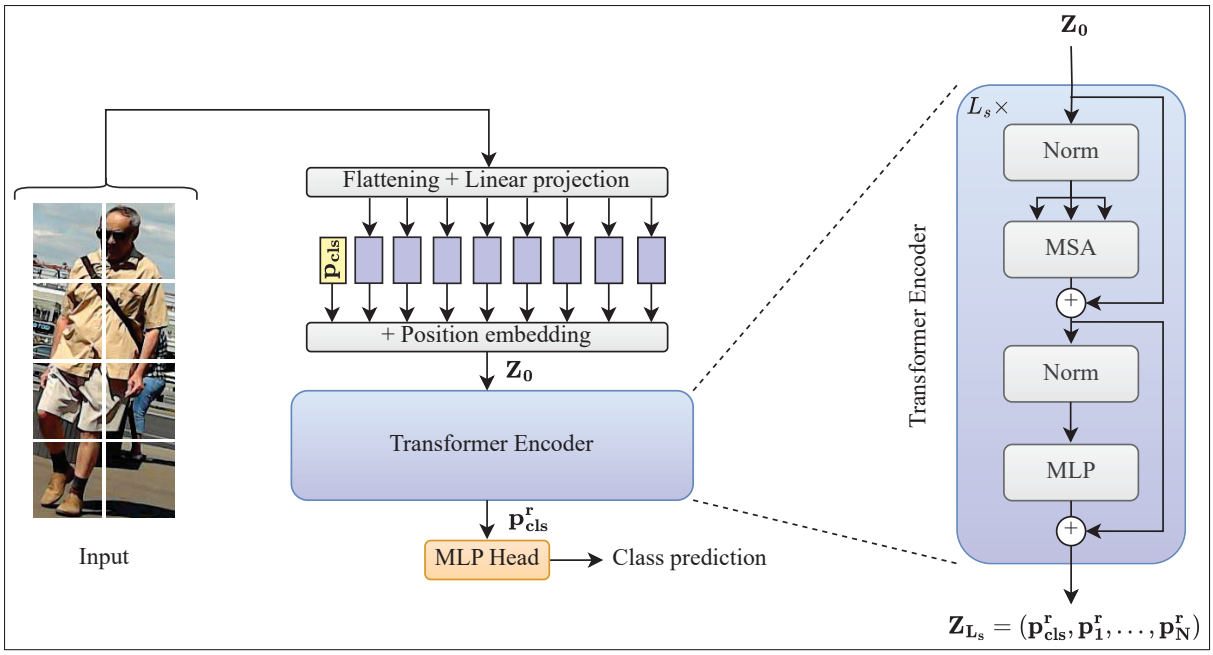

Figure 1.6 Block diagram of a ViT. The input image is first split into patches and pre-processed to form $\mathbf{Z_0}$. This input goes through the transformer encoder formed by L_s blocks, leading to the output $\mathbf{Z}_{\mathbf{L_s}}$. The refactored classification token $\mathbf{p}^{\mathbf{r}}_{\mathbf{cls}}$ contained in $\mathbf{Z}_{\mathbf{L_s}}$ is finally processed by the MLP head.

nature as text would be, Dosovitskiy *et al.* (2020) proposed to divide the input image or feature maps into a grid of fixed-size patches. Let $\mathbf{F} \in \mathbb{R}^{H\times W\times C}$ be the input, with $H \in \mathbb{N}$, $W \in \mathbb{N}$ and $C \in \mathbb{N}$ being respectively the input height, width, and channel size. Let us consider the input subdivision into $N = \frac{H}{S} \times \frac{W}{S}$ patches, noted $\mathbf{p}_i$, with $i \in \{1, ..., N\}$. The patches are then flattened regarding the spatial dimension and passed through a linear embedding layer E with an output dimension D. A classification token $\mathbf{p_{cls}} \in \mathbb{R}^D$ of the same dimension is added and used later as the global feature representation for the classification. To the representation is summed a learnable positional embedding $\mathbf{P} \in \mathbb{R}^{(N+1)\times D}$. Hence, the transformer input sequence $\mathbf{Z_0}$ is defined as follows:

$$\mathbf{Z_0} = [\mathbf{p_{cls}}, E(\mathbf{p_1}), E(\mathbf{p_2}), \ldots, E(\mathbf{p_N})] + \mathbf{P} \tag{1.6}$$

After having pre-processed the data and obtained the $\mathbf{Z_0}$ representation, this representation is passed through a sequence of $L_s \in \mathbb{N}$ alternating MSA and MLP blocks (Dosovitskiy *et al.*, 2020).

For each block, the input passes through a LayerNorm right before, and a residual connection is applied right after. After the L_s blocks, the output matrix $\mathbf{Z_{L_s}}$ contains the refactored tokens (or patches) as $\mathbf{Z_{L_s}} = (\mathbf{p^r_{cls}}, \mathbf{p^r_1}, ..., \mathbf{p^r_N})$, the refactored tokens being noted $\mathbf{p^r_i} \in \mathbb{R}^D$ with $i \in \{1, ..., n\}$ and the refactored classification token noted $\mathbf{p^r_{cls}} \in \mathbb{R}^D$. This process allows for exploiting the relationships within and among the tokens and, consequently, for extracting rich feature representations. The refactored classification token is finally used as input to an MLP for class prediction.

Let us formally define the MSA layers. Self-attention allows selectively attending and refactoring the tokens from a given input sequence $\mathbf{Z} \in \mathbb{R}^{N \times D}$, by first defining a set of queries $\mathbf{Q}$, keys $\mathbf{K}$, and values $\mathbf{V}$:

$$\mathbf{Q} = \mathbf{Z}\mathbf{W^q}, \mathbf{K} = \mathbf{Z}\mathbf{W^k}, \mathbf{V} = \mathbf{Z}\mathbf{W^v} \tag{1.7}$$

where $\mathbf{W}^q \in \mathbb{R}^{D \times D_q}$, $\mathbf{W}^k \in \mathbb{R}^{D \times D_k}$, and $\mathbf{W}^v \in \mathbb{R}^{D \times D_v}$ being weight matrices. Then, each set of queries keys and values is given as input to the MSA layer, which generates attention weights through the softmax σ function, and applies those weights to the set of values, refactoring the tokens:

$$\mathrm{MSA}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \sigma(\frac{\mathbf{QK}}{\sqrt{D_k}})\mathbf{V} \tag{1.8}$$

In comparison, the CNN architecture, and especially the convolution operations, only operate on a local neighborhood of an image, which may lead to weaker feature representations. ViTs can also handle variable-sized inputs, avoiding cropping or resizing the data as needed for the CNNs. However, transformer architectures have a high computation cost and require large datasets to get optimized (Han *et al.*, 2020). For example, the base ViT model requires 18G FLOPs to process an image and gather a total of 86M parameters against 4.1G FLOPs and 25M parameters for a ResNet-50 model (He *et al.*, 2016). Hence, ViT models may not be suited to some fields or tasks.

Many application areas, like medical imaging, are often limited in terms of data availability (Greenspan *et al.*, 2016) due to privacy concerns and complex data annotation requiring medical experts. Also, various tasks like person ReID require close to real-time applications, which

high computation requirements might compromise (Remigereau *et al.*, 2022). This aspect may be tackled using hybrid architectures (Dosovitskiy *et al.*, 2020; Prakash *et al.*, 2021), using first a CNN to encode and reduce the input dimension, followed by self-attention transformer mechanisms to enrich the feature representation. Plus, hybrid models are especially more beneficial than transformer-only models on smaller architectures (Dosovitskiy *et al.*, 2020), the hybrid models being consequently more adapted to tasks with light model requirements. For example, the ResNet-50 + ViT-B/16 model (hybrid) performs similarly as the larger model ViT-L/16 (transformer only), with pre-training computation FLOPs respectively of 274 and 783 exaFLOPs according to Dosovitskiy *et al.*.

1.3 Embedding Networks

1.3.1 Metric learning

Metric learning is a subfield of ML that focuses on mapping input data points into a space where distances between these points correspond to their similarity or dissimilarity (Fig. 1.7). In other words, it seeks to learn a function to measure how similar or different two data points are. In practice, this is a very useful and powerful concept that allows answering tasks such as signature verification (Viana *et al.*, 2022), face or speaker identification (Sun *et al.*, 2014; Chen & Salman, 2011), or person ReID (Yi *et al.*, 2014; Chen *et al.*, 2018b; Yang *et al.*, 2018). Indeed, taking person ReID as an example, the idea is not to assert the identity of a given individual but to assert that two different images contain the same individual or not, rendered possible through metric learning approaches.

Let us define two data points or feature vectors $\mathbf{f_1} = (f_{11}, f_{12}, ..., f_{1d})$ and $\mathbf{f_2} = (f_{21}, f_{22}, ..., f_{2d})$, $d \in \mathbb{N}$ being the vectors dimension. To measure the distance between $\mathbf{f_1}$ and $\mathbf{f_2}$, and evaluate their similarity, the Euclidean distance (Eq. 1.9) or the cosine distance (Eq. 1.10) are commonly used distance measures that can easily be computed. However, as highlighted by Lu *et al.* (2017), the similarity measure selection must generally be task specific as each task and dataset has a specific data distribution, differently affecting the distance measures.

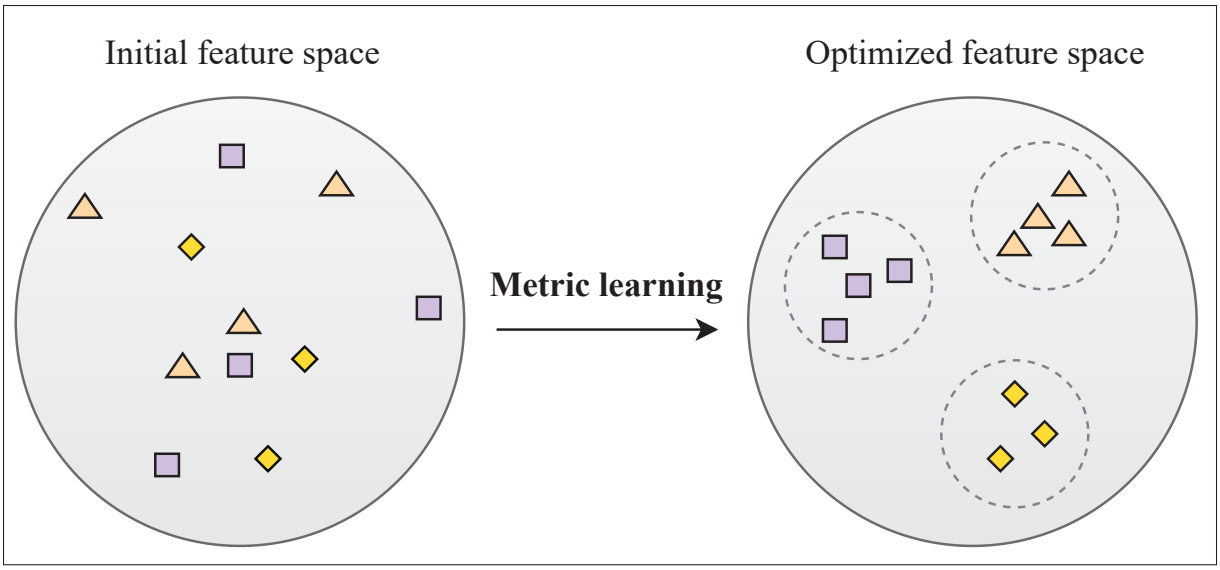

Figure 1.7 Illustration of a 2D feature space with three distinct identities (classes) before and after optimization through the usage of metric learning.

$$d_{euclidean}(\mathbf{f_1}, \mathbf{f_2}) = \|\mathbf{f_1} - \mathbf{f_2}\| = \sqrt{\sum_{i=1}^{d}(f_{1i} - f_{2i})^2} \tag{1.9}$$

$$d_{cosine}(\mathbf{f_1}, \mathbf{f_2}) = \frac{\mathbf{f_1} \cdot \mathbf{f_2}}{\|\mathbf{f_1}\|\|\mathbf{f_2}\|} = \frac{\sum_{i=1}^{d} f_{1i} f_{2i}}{\sqrt{\sum_{i=1}^{d} {f_{1i}}^2}\sqrt{\sum_{i=1}^{d} f_{2i}^2}} \tag{1.10}$$

1.3.2 Deep Siamese networks

Siamese Neural Networks (SNN) (Bromley *et al.*, 1993) is a classic ML model (Chicco, 2021) for similarity matching. An SNN automatically learns a function that maps data samples, like images or texts, into a common feature space, where their similarity or distance can be computed (Fig. 1.8.a). To build this feature space, a single backbone is replicated into two identical backbones, sharing the exact same weights, and is trained simultaneously on pairs of inputs. Data triplets can also be used with the same principle and three backbones replications (triplet

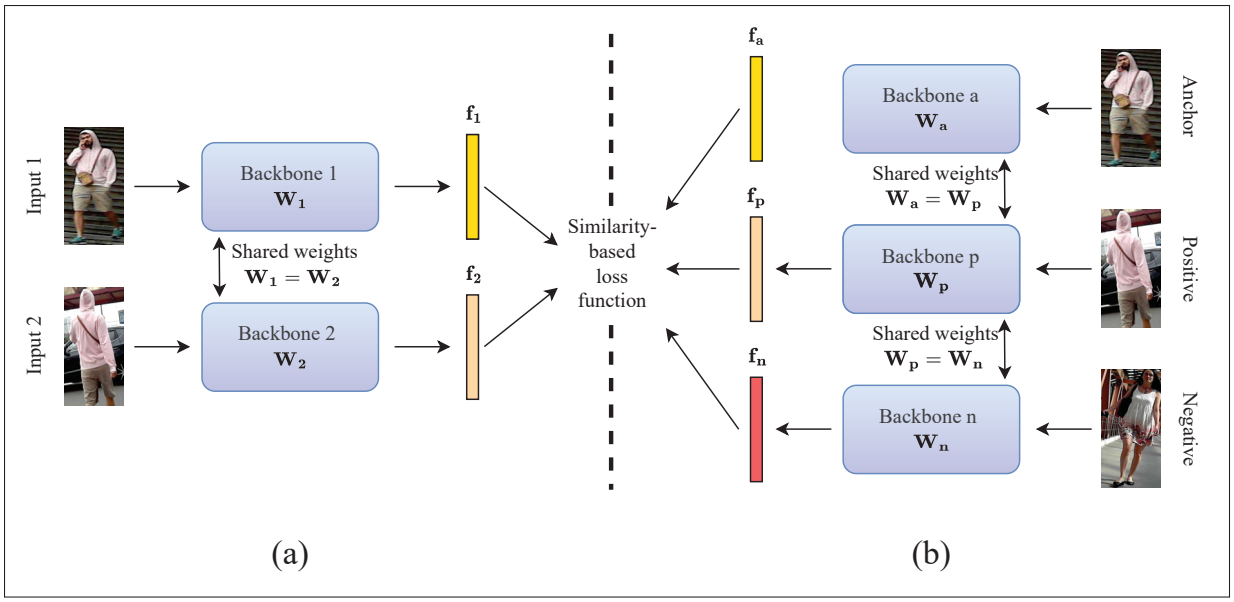

Figure 1.8 Training model using (a) pair or (b) triplet of individuals as input to a (a) Siamese or (b) Triplet network. For the Siamese network, each backbone is the same in practice, the weights being consequently the same ($\mathbf{W_1} = \mathbf{W_2}$). This is also true for the Triplet network with $\mathbf{W_a} = \mathbf{W_p} = \mathbf{W_n}$. While learning, the weights are updated using a similarity-based loss function.

network Fig. 1.8.b), this aspect being, in practice, a consequence of the loss selection. For this reason, let us detail existing metric learning losses.

Efficient and tailored for embedding networks, multiple losses have been designed and used for the person ReID task. The contrastive loss (Hadsell *et al.*, 2006) is a well-known cost function that works from pair of data and works at making similar pair closer and dissimilar pair further apart, by at least a fixed margin, in the embedding space. The triplet loss (Ding *et al.*, 2015) requires triplets instead of data pairs, using embeddings from an anchor, a positive and a negative inputs. The positive belongs to the same class as the anchor, whereas the negative belongs to another class, and the triplet loss learns the model to reduce the distance between the anchor and the positive sample while increasing the one between the anchor and the negative sample. Later, different versions were proposed to improve the original triplet loss. For example, Cheng *et al.* (2016) adds a constraint on the positive pair, or (Hermans *et al.*, 2017) proposes a semi-hard sample mining by selecting the hardest negative and the hardest positive pairs in a mini-batch.

Other approaches, such as those proposed by Yu *et al.* (2018) or Wojke & Bewley (2018), also present different versions of the triplet loss. However, in practice, one can observe that the semi-hard mining approach proposed by Hermans *et al.* (2017), usually referred to as batch-hard triplet loss, is widely used in the state-of-the-art person ReID literature (Ye *et al.*, 2021) for its efficiency and simplicity.

Because of their wide usage in literature, let us formally define the contrastive and batch-hard triplet loss. Let us consider two feature vectors $\mathbf{f_1} \in \mathbb{R}^d$ and $\mathbf{f_2} \in \mathbb{R}^d$, d being the features dimension. Let y be a binary label associated to $\mathbf{f_1}$ and $\mathbf{f_2}$. If the two vectors belong to the same class, then $y = 0$. Otherwise, $y = 1$. On the basis of this, the contrastive loss $\mathcal{L}_\text{c}$ is computed as follows:

$$\mathcal{L}_\text{c}(\mathbf{f_1}, \mathbf{f_2}) = (1 - y) \cdot \max(0, M_1 + ||\mathbf{f}_1 - \mathbf{f_2}||^2) + y \cdot ||\mathbf{f}_1 - \mathbf{f_2}||^2. \tag{1.11}$$

where M_1 is a margin hyperparameter that controls the minimum distance between samples from different classes.

For the batch-hard triplet loss, let $\mathbf{f}_a$, $\mathbf{f}_p$, and $\mathbf{f}_n$ be, respectively, the feature vectors for the anchor, the positive, and the negative samples. For a training mini-batch with $N \in \mathbb{N}$ samples and their associated labels $\{y_i\}_{i=1}^N$, the batch-hard triplet loss $\mathcal{L}_\text{BH\_tri}$ is defined as follows:

$$\mathcal{L}_\text{BH\_tri} = \frac{1}{N} \sum_{y_a = y_p \neq y_n} \max(0, m + \max_{y_a = y_p} d(\mathbf{f}_a, \mathbf{f}_p) - \min_{y_a \neq y_n} d(\mathbf{f}_a, \mathbf{f}_n)). \tag{1.12}$$

where d represents the Euclidean distance (Eq. 1.9) and M_2 is a margin hyperparameter that controls the minimum difference between the distances.

Comparing those two losses, one can notice that the contrastive loss can be less computationally expensive than the triplet loss since it considers pairs instead of triplets. However, the triplet loss explicitly considers relative relationships between samples, which makes it particularly useful when dealing with complex data distributions, where pairwise relationships may not be sufficient to capture the underlying structure effectively.

1.4 Person Re-Identification

1.4.1 Problem definition

A person ReID system aims to identify the same person across different camera views (Zheng *et al.*, 2016; Ye *et al.*, 2021). This task is helpful in various situations, primarily video surveillance, since it automatically and efficiently processes a large quantity of video data. In fact, the input to such a system can either be images of individuals (image-based ReID (Zhu *et al.*, 2020; Li *et al.*, 2021; He *et al.*, 2021)) or tracklets (video-based ReID (Yan *et al.*, 2020; Eom *et al.*, 2021)), depending on the objective and the used framework. Let us define the set of images or tracklets $\mathcal{X} = \{x_i | i \in \{1, 2, ..., N_s\}\}$, x_i being a data sample and N_s the number of data samples. The corresponding identities are stored in $\mathcal{Y} = \{y_i | i \in \{1, 2, ..., N_s\}\}$, forming the dataset $\mathcal{D} = \{\mathcal{X}, \mathcal{Y}\}$. Let $\varphi_\theta : \mathbb{R}^S \rightarrow \mathbb{R}^E$ be a function mapping from the data space $\mathbb{R}^S$ to the embedding space $\mathbb{R}^E$, parameterized by θ. In practice, for a given query individual $x_q \in \mathcal{X}$, its embedding $\mathbf{f_q} = \varphi_\Theta(x_q)$ is produced and matched with the embeddings from the gallery of references. The matching is done by comparing the query and gallery samples through a similarity (or distance) measure $S : \mathbb{R}^E \times \mathbb{R}^E \rightarrow \mathbb{R}$, the similarity values bringing insights on the similarity of the actual samples.

1.4.2 Performance measures

For evaluation, open-set and closed-set are two scenarios in which the person ReID task can be applied. For the closed-set scenario, which is more common, the considered queries always have one or more corresponding matches in the gallery. In this situation, as a match always exists, a ranking list is usually built, with the first ranks being more likely to be the same and the last different individuals. From this list, classic performance measures like the mean Average Precision (mAP) or the Cumulative Matching Characteristic (CMC) curve can be computed.

In the case of the open-set scenario, there is no insurance that the query person also appears in the gallery. Hence, the approach is different. A threshold is usually set on the similarity measure

to identify whether the individual exists in the gallery. Consequently, classification-related performance measures like the Receiver Operating Characteristic (ROC) or Precision-Recall (P-R) curves can be computed (Fawcett, 2006; Davis & Goadrich, 2006). This thesis focuses on the closed-set scenario since this approach is followed by most papers in this domain, as highlighted in the survey on person ReID provided by Ye *et al.* (2021). Still, the open-set setting could be considered for future experiments as it offers better suitability for deployment purposes by not requiring matches in the reference set.

As mentioned, the closed-set scenario is considered in this thesis to align with the current literature. Hence, a ranking list is first built for a given query thanks to a similarity measure between itself and the gallery of references. The query should be retrieved as accurately as possible, i.e., all the correct matches should have low-rank values in the list. Based on this ranking list, the performance measures can be computed. The mAP and the mean Inverse Negative Penalty (mINP) are two widely used performance measures in person ReID (Ye *et al.*, 2021), defined as follows :

- **Mean Average Precision:** To compute the mAP, the Average Precision (AP) must be defined first. For a query image q, the number of persons to retrieve in the gallery is noted R_q. Let m be a function equal to 1 if the image is a match and equal to 0 otherwise. Let P be the precision function, P being equal to the number of retrieved persons up to rank k over the rank k value. AP is computed as follows for a given query q:

$$AP_q = \frac{1}{R_q} \sum_{k=1}^{R_q} m(k).P(k) \tag{1.13}$$

  Then, the mAP can be computed through a mean over the number of query image $Q \in \mathbb{N}$:

$$\text{mAP} = \frac{1}{Q} \sum_{q=1}^{Q} AP_q \tag{1.14}$$

- **Mean Inverse Negative Penalty:** The mINP measure requires the Negative Penalty (NP) to be defined first. Again, for q a given query image, Q is the number of query images, and R_q

is the number of matches in the gallery. Also, let R_q^H be the position of the hardest match in the ranking list. Then, the NP measure is as follows:

$$NP_q = \frac{R_q^H - R_q}{R_q^H} \tag{1.15}$$

Then, the mINP is a mean over the inverse of the negative penalty ($INP = 1 - NP$), being computed as follows:

$$\text{mINP} = \frac{1}{Q}\sum_{q=1}^{Q}(1 - NP_q) \tag{1.16}$$

In practice, one can see the complementarity of the mAP and mINP performance measures. Indeed, the mAP gets computed based on all match positions in the ranking list. Hence, it reflects well an overall idea of the model precision. However, with varying positions in the ranking list, the AP measure can sometimes produce similar performance evaluations even if one of two ranking lists may have way more difficulty dealing with few samples (Fig. 1.9). Usually, a model with a similar mAP but dealing better with hard cases would likely be favored because it rubs off on a better generalization capacity. In fact, the INP measure hints at this aspect by focusing only on the latest match position in the list. Indeed, one can observe in Fig. 1.9 that the INP measure is higher for rank list 1, which deals better with hard cases.

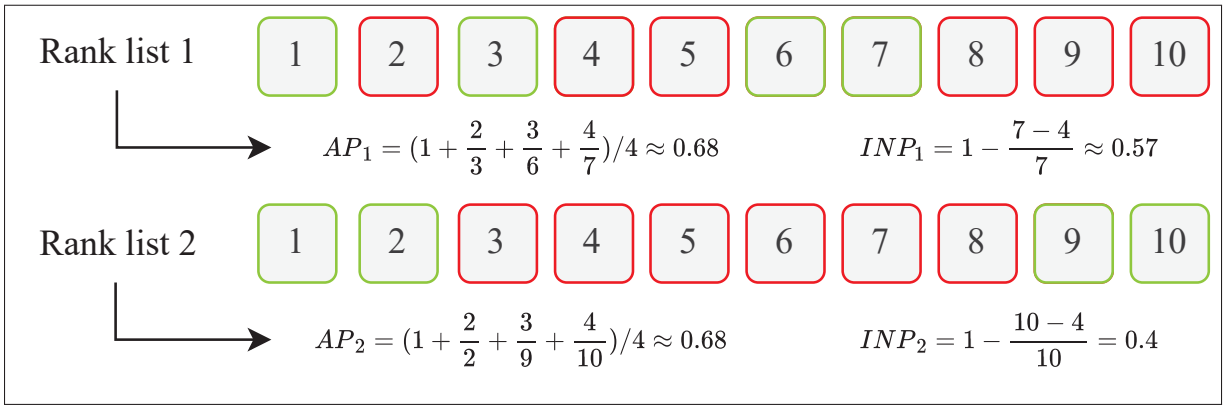

Figure 1.9 Example of the AP and INP measures for two ranking lists of ten samples. Green and red cubes represent matches and false matches, respectively.

1.4.3 Overview of the state-of-art

Over the last decade, significant progress has been made in the field of person ReID (Ye *et al.*, 2021; Ming *et al.*, 2022). According to the taxonomy proposed by Chen *et al.* (2021), the state-of-the-art approaches for person ReID can be categorized into part-based, augmentation, multi-scale, and attention methods. This section defines these categories and provides an overview of the related approaches.

Part-based approaches start from the observation that different body parts, such as the head, torso, or legs, can all contain discriminant knowledge. However, models often focus on global features and recurrent information, leading them to miss knowledge from small informative regions (Chen *et al.*, 2020b). Consequently, the objective is usually to extract features directly from these body parts or regions of interest and finally combine them to obtain a richer representation of an individual. The Clue Alignment and Conditional Embedding (CACE-Net) model (Yu *et al.*, 2020) is an example of a part-based approach, which uses a correspondence attention model to find and rely on key points between pair of images for discrimination. The Adaptive Part Division (APD) model (Lai *et al.*, 2021) is another part-based example that relies on self-attention to generate masks for part division and extract robust local features.

Augmentation techniques involve applying transformations (i.e., changes in scale, rotation, translation, or noise) on the learning data to make the models more robust to scene variations. These transformations can be more specific, as for the two DA proposed by Chen *et al.* (2021). One of the DAs applies a patch of random pixel values on the image, while the other superposes a randomly selected patch from the image on this same image. This way, the model benefits from improved robustness to local and global corruptions and a better generalization capacity. In the context of cloth-changing person ReID, Jia *et al.* (2022) proposed a positive DA augmentation that exchanges patches between positive pairs and a negative DA which translates the appearance and pose information from negative to negative samples in order to create more diversity within the negative images.

Multi-scale approaches process the input images at multiple scales or resolutions to capture both global and local information. For instance, Omni-Scale Network (OSNet) (Zhou *et al.*, 2021) is a lightweight, multiscale model architecture with different streams focusing on specific feature scales, designed for person ReID. Using the OSNet model as a basis, the Lightweight Multi-Branch Network (LightMBN) (Herzog *et al.*, 2021) uses the three first blocks of an OSNet model followed by three branches to extract global, partial, and channel-wise representations. (Huang *et al.*, 2021) presents a Multi-Scale Focusing Attention (MSFA) block, using focusing attention blocks within it with different receptive fields to capture discriminative features at different scales. For the combination of the features from the different branches, the aggregation gate from OSNet is used.

Attention approaches focus on using attention mechanisms to assign weights and selectively attend to specific features and informative regions. For example, Chen *et al.* (2019b) proposed two attention modules that are used on the last feature map obtained from a ResNet-50, one focusing and channel attention and the other on pose attention. Relying instead on self-attention mechanisms through a ViT architecture, the TransReID (He *et al.*, 2021) model is the first transformer-based approach attaining state-of-art results for person ReID.

Despite the design of strong models and strategies to improve efficiency or ReID accuracy, models must be carefully designed to meet real-time requirements, handle real-world conditions, and work with limited resources. Chen *et al.* (2021) compare 21 state-of-art person ReID models and lead us to various observations by evaluating them using challenging and realistic corrupted evaluation datasets. First, the best-performing models are inconsistent from the evaluation using normal and corrupted data, which is not handy for a model selection. For example, the LightMBN approach (Herzog *et al.*, 2021) comes first when evaluated on clean data, and the TransReID model (He *et al.*, 2021) when evaluated on corrupted data. One can also observe that the TransReID model is complex due to its transformer-based architecture. This aspect is regretful for real-world applications since such a model is unrealistic for real-time surveillance. Instead, LightMBN is lightweight but significantly under-perform on the corrupted datasets compared to TransReID. Hence, one may wonder whether there is a way to design a model that

would be consistent from clean to corrupted scenarios while being adapted to the ReID and its efficiency requirements.

1.5 Visual Interpretation

1.5.1 Methods for image classification

For years, NNs and especially CNNs were criticized for their lack of interpretability (Lipton, 2018), often characterized as "black boxes". In fact, the inability to see the cues on which a model relies to provide its prediction was seen as a major bottleneck, making their optimization more complex. Several attempts to produce CNN visualizations were first designed (Simonyan *et al.*, 2013; Zeiler & Fergus, 2014; Gan *et al.*, 2015), but did not allow for class-discriminative and input-specific visualizations, regrettable for a deep model understanding.

This has been leveraged in 2017 when Selvaraju *et al.* developed the CNN visualization approach named GradCAM, allowing for building heatmaps (Fig. 1.10) on which one can visually observe the input-wise features on which the model relies to produce each individual predictions.

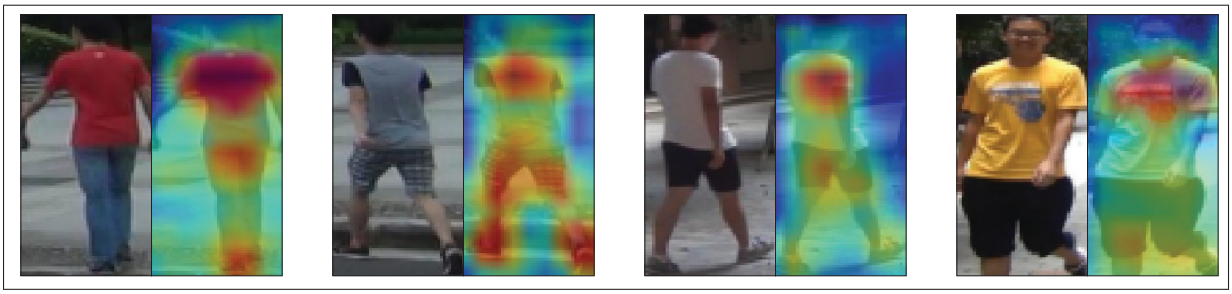

Figure 1.10 GradCAM visualizations obtained from a learned backbone on the SYSU-MM01 dataset (Wu *et al.*, 2017) and 4 images of individuals.

Let $\mathbf{F^k}$ be the k-th channel of the last feature maps, where the last feature maps have C channels, a width W, and a height H. To build a heatmap highlighting the essential regions of a given input image, GradCAM computes the importance weights of each channel w_k by performing the average pooling operation over the back-propagated gradients of the output class score noted y_c (for class c) with respect to the feature map $\mathbf{F^k}$:

$$w_c = \frac{1}{Z} \sum_{x,y} \frac{\partial y_c}{\partial \mathbf{F}^{\mathbf{k}}_{(\mathbf{x},\mathbf{y})}} \tag{1.17}$$

where $\mathbf{F}^{k}_{(x,y)}$ is the feature map $\mathbf{F}^{\mathbf{k}}$ at spacial location (x, y).

Then, the feature maps are weighted and passed through a ReLU activation function to produce the low-resolution heatmap $\mathcal{L}^{c}_{\text{GradCAM}}$:

$$\mathcal{L}^{c}_{\text{GradCAM}} = \sum_{k=1}^{C} w_c \cdot \mathbf{F}^{\mathbf{k}} \tag{1.18}$$

Finally, $\mathcal{L}^{c}_{\text{GradCAM}}$ is up-sampled to match the size of the input image using bilinear interpolation and then overlaid on the input image to create a Class Activation Map (CAM).

From GradCAM, other visualization strategies were designed, like GradCAM++ (Chattopadhay *et al.*, 2018), XGradCAM (Fu *et al.*, 2020), or LayerCAM (Jiang *et al.*, 2021). However, the previously mentioned algorithms rely on class-related gradients and consequently require class-specific activation scores, which are unavailable with embedding models.

1.5.2 Methods for similarity matching

Metric learning tasks, like person ReID, seek to match embeddings and do not rely on class predictions during inference. As a consequence, the model does not produce and use class logit values but relies on embedding similarities for predictions. As mentioned previously, this aspect discards CAM approaches like GradCAM to visualize deep metric learning models and ask for similarity-matching tailored visualization strategies.

Tackling this challenge, Stylianou *et al.* (2019) proposed a simple but effective solution based on the pooling operation and the similarity measure decomposition applied to a pair of images. For

a network using average pooling as a flattening operation, the output vector $\mathbf{f}$ is:

$$\mathbf{f} = \frac{1}{K^2} \sum_{x,y} \mathbf{F}_{(x,y)} \tag{1.19}$$

Where $\mathbf{F}_{(x,y)}$ is the C-dimensional slice of the feature maps $\mathbf{F}$ at spacial location (x, y), for $\mathbf{F}$ being the last feature map, C its number of channels, and K its width and height (considered here of same dimension). Hence, for two output vectors $\mathbf{f_1} \in \mathbb{R}$ and $\mathbf{f_2} \in \mathbb{R}$, the similarity between the two inputs can be decomposed spatially by using Eq. 1.19 and the similarity distance measure Eq. 1.10:

$$d_{cosine}(\mathbf{f_1}, \mathbf{f_2}) = \frac{\mathbf{f_1} \cdot \mathbf{f_2}}{\|\mathbf{f_1}\|\|\mathbf{f_2}\|} = \frac{\mathbf{F}_{(1,1)} \cdot \mathbf{f_2} + \ldots + \mathbf{F}_{(K,K)} \cdot \mathbf{f_2}}{Z} \tag{1.20}$$

with $Z = K^2\|\mathbf{f_1}\|\|\mathbf{f_2}\|$ being a normalizing factor.

Through this decomposition and the relative contribution of each part of the activation maps, the pixels impacting the similarity measure can get highlighted on the initial input image. Working from this same decomposition, Black *et al.* (2022) extended the approach to embedded transformer networks, applying a rollout operation to approximate each image patch contribution to the similarity measure, despite the patch mixing operation applied at each transformer layer.

For CNNs, Chen *et al.* (2020a) proposed a strategy based on the gradient of the triplet loss instead of the gradient of the classification scores for GradCAM. For each testing image, the closest training image is first chosen using the KNN algorithm. Then, the triplet loss values (from a certain quantity of triplets) are averaged and used as a gradient to produce CAMs for the test image. This approach has the advantage of requiring a single evaluation image instead of two in order to produce the CAMs. However, it requires storing the gradient values of the learning triplets and applying a KNN, which is not needed in the approach of Stylianou *et al.* (2019). Overall, the similarity-based CAM proposed by Stylianou *et al.* appears more handily while being hardly differentiable in terms of quality and is consequently used in this thesis.

CHAPTER 2

LITERATURE REVIEW

2.1 Survey of Multimodal Fusion

The principle of data fusion is ubiquitous, both around and within us. Our various senses work in harmony most of the time to provide a better understanding and description of our surroundings. For instance, we rely on sight and touch to apprehend an object, while both hearing and vision help us navigate when crossing a street. Also, we use all of these senses together when engaging in activities such as running and talking and effortlessly process numerous information combinations in our daily lives. From this observation, combining information from multiple sources appeared as an intuitive way to solve various problems. In PR applications, leveraging complementary information is one of the main motives for using multimodal approaches, this information is expected to make recognition accurate and robust.

Data fusion can be performed through various global strategies, well delimited and defined by Jain *et al.* (2004). For example, using multiple sensors is probably the most intuitive, by analogy with the use of our senses. Indeed, a scene can, for example, be encoded from distinct camera sensors (modalities), allowing for an independent but complementary encoding of the information. Instead, using plural snapshots of the same concept and from the same sensor may as well improve the system, allowing access to supplementary cues with a scene that has evolved. Similarly, multiple consecutive frames can be fused, allowing exploiting the temporal information. Finally, instead of using different input data, different algorithms can also focus on the same sample but encode it in different ways to complement the final representation and improve the accuracy. In a nutshell, it is essential to have diverse and informative sources of knowledge to benefit from fusion.

This thesis focus on the fusion of information from images captured using RGB and IR cameras for person ReID. These sensors encode the scene differently, leveraging separate segments of the electromagnetic spectrum (Sec. 1.1). To illustrate, the RGB modality encodes color

information and is ineffective for capturing nighttime scenes, whereas the infrared modality does not encode the color information while not being affected by luminosity. Additionally, as aforementioned, the cameras from the RGB to the IR modality may be co-located (CL) or not co-located (NCL). Indeed, a real-world application could use V-I cameras side-by-side (CL cameras), easily constructing V-I data pairs and benefiting from the two cameras encoding most of the time. However, visible and infrared cameras could be paired while pointing to distinct scenes (NCL cameras). In this case, a realistic setting would likely consider a short distance from the visible to the infrared cameras while pairing those but also using infrared cameras for darker scenes. Under the NCL setting, a large domain gap may be observed in the representations, as it breaks the spatial alignment existing under a CL cameras setting. This aspect is expected to make the fusion more beneficial for NCL cameras since weakly dependent or independent information usually results in a more significant fusion improvement (Jain *et al.*, 2004). However, the previous expectation might not be verified in the context of model-level fusion, as such fusion approach sometimes benefits more from spatial alignment (Wang *et al.*, 2021b; Xuan *et al.*, 2022).

2.1.1 Conventional fusion methods

When considering multimodal data fusion, common techniques usually rely on information at the sensor (Lohweg & Mönks, 2010), feature (Zhu *et al.*, 2015; Shahroudy *et al.*, 2014), or score levels (Snoek *et al.*, 2005). For sensor-level fusion (Fig. 2.1.a), the multiple modalities are usually stacked together prior to feature extraction. This can result in dense and complex inputs, but the model should encode the inter-modality correlations from the start, thanks to the shared sensor representation. However, this can also result in a loss of modality-specific knowledge as the two modalities are processed as a single representation. Feature fusion (Fig. 2.1.b) considers modality-specific backbones to extract feature representations and combines them through, e.g., aggregation or concatenation before answering the objective task. Fusion at this level allows individual modality encoding prior to the fusion, which may be a good strategy to allow for modality-specific knowledge mining. Feature fusion provides more flexibility by allowing for

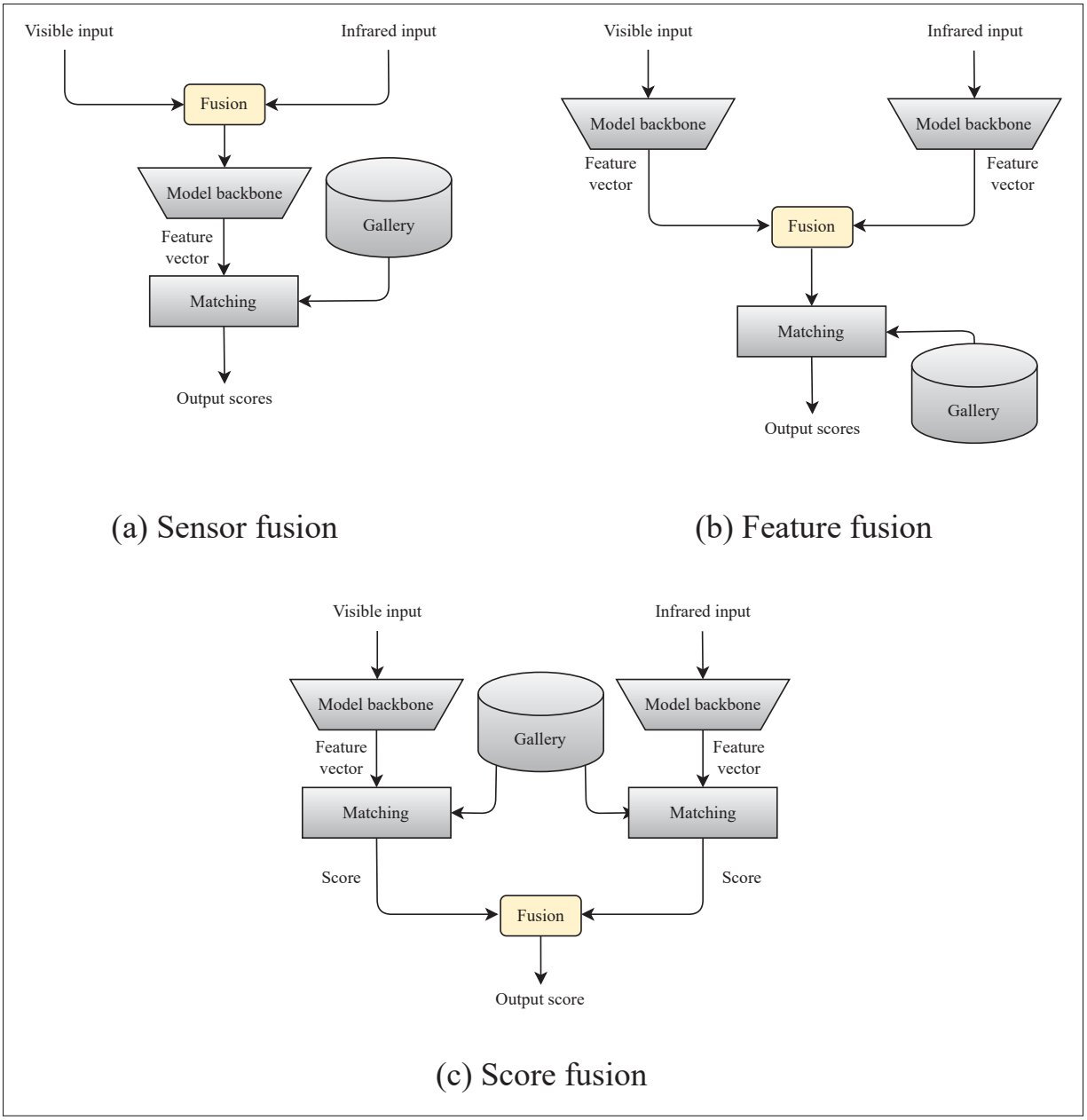

Figure 2.1 Block diagram of similarity matching systems that perform (a) early, (b) feature, and (c) score-level fusion of V and I modalities.

modality-specific encoding strategies. However, the fusion strategy must be wisely chosen not to lose information. Often, for practical reasons, people proceed with late fusion, using, for instance, weighted averaging (Nojavanasghari *et al.*, 2016) or majority voting (Wörtwein & Scherer, 2017) strategies. This level of fusion has the advantage of being easy to implement because

trained unimodal models can be integrated with other models. However, score values must be normalized or calibrated across modalities, and, assuming the score-level fusion is commonly deterministic (Zhang *et al.*, 2017) there is no learning of the eventual modality correlations. In fact, the absence of modality collaboration is one of the major limitations of this approach (Wang, 2021), modality collaboration likely being of true importance while facing real-world data.

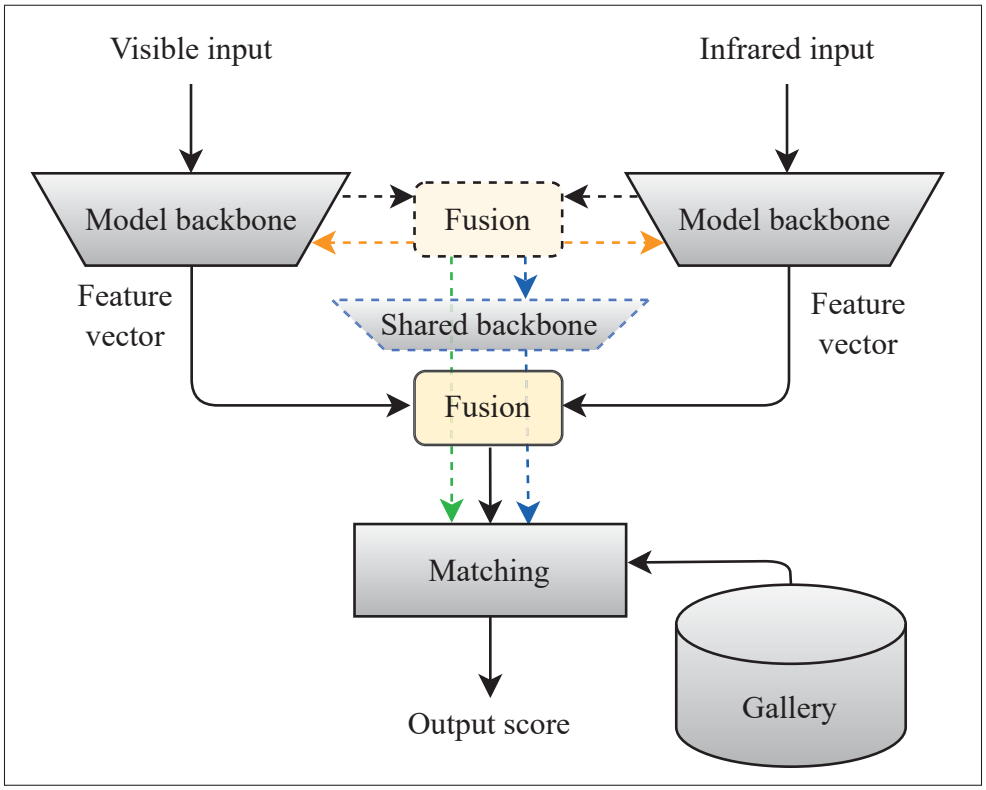

Figure 2.2 Model-level fusion representation. The fusion can be done with the feature vector produced from independent backbones (learned feature fusion). It can also be progressively done through a fusion between the backbones. In this case, the fused representation can be re-introduced into modality-specific backbones (orange arrows), into a shared backbone before matching (blue arrows), or directly matched (green arrow).

Distinct from the previous approaches, intermediate (or model-level) fusion considers combining the knowledge during feature extraction through multimodal learning (Ramachandram & Taylor, 2017; Gaw *et al.*, 2022). This fusion allows for a more progressive fusion process in the sense

that knowledge can be gradually exchanged among modality backbones at multiple stages of the data processing (Fig. 2.2). This is the case of the approach proposed by Joze *et al.* (2020), for example, with attention-based fusion modules used in between convolution blocks of two distinct modality backbones several times in the model. Unlike sensor fusion, model-level fusion usually starts by down-sampling the data through modality-specific layers before the fusion. This allows the model to intelligently encode the modality, preparing the modality collaboration for the upcoming fusion, which seems of great importance in the case of data corruption. Also, since feature fusion comes earlier in the process compared to score fusion, more knowledge is available, which should allow for more correlation findings.

The model-level fusion naturally stands out from the different approaches in the previous discussions. Indeed, model-level fusion should allow for a good balance in exploiting modality-specific and modality-shared knowledge while allowing for flexibility in the fusion design and leveraging modality collaboration. Within those models, attention-based approaches (Niu *et al.*, 2021; Guo *et al.*, 2022) could be a great strategy for making models robust to corruption. Indeed, its dynamic mechanisms may allow for better feature selection, motivating a survey of these approaches in the next section.

2.1.2 Attention-based fusion methods

This section presents multimodal attention-based methods, starting with CNN-based models, then transformer-based models, and finally hybrid models.

2.1.2.1 Attention-based CNN models.

In 2020, Ismail *et al.* proposed modality attention, producing attention weights to balance modality feature vectors' importance in the upcoming fusion. Instead of applying single soft weights to each modality, Arevalo *et al.* (2020) presented the gated multimodal unit, generating and applying attention weights for every feature in the concatenated feature vectors, discarding the less relevant information while setting an emphasis on discriminant features. Working from

richer representations by focusing on feature maps instead of feature vectors, the Multimodal Transfer Module (MMTM) (Joze *et al.*, 2020) and Multimodal Split Attention Fusion (MSAF) (Su *et al.*, 2020) approaches update the feature maps through channel-wise attention weights produced and applied at different positions of the encoding process. The weights are obtained from a modality-shared representation built using each modality feature map. Fu *et al.* (2021b) introduced a cross-modal attention block that refactors the feature maps of one modality based on the other. However, one modality representation only is refined through this attention process, whereas a bilateral refactoring would probably be a better strategy when facing corruption.

Recently, multiple approaches outperformed the MMTM and MSAF strategies performance-wise. However, the strategies are different by not solely focusing on the fusion function. For example, Luna-Jiménez *et al.* (2021) focused mainly on the modality-specific backbones to use and proceed with late fusion to avoid temporal alignment issues, which is not a concern in image-based person ReID. Middya *et al.* (2022) also improves the emotion recognition accuracy thanks to a careful selection of modality-specific backbones but trains them together with a final feature vector concatenation instead. Again, the approach is task-specific and is not directly relevant for image-based person ReID using the RGB and IR modalities. Finally, Mocanu & Tapu (2022) introduced a cross-modal attention block, which relies on cross-modal correlation exploitation. This attention block is interesting, but the model also exploits unimodal spatial and temporal attention at first, making unclear the real benefits of the fusion block.

2.1.2.2 Attention-based transformer models

Alongside improvements in CNN architectures, various multimodal transformer models emerged. Some of the considered approaches are not vision-based transformers but are easily transferable to ViT thanks to the work of Dosovitskiy *et al.* (2020).

To reinforce a target modality representation from a source modality, Tsai *et al.* (2019) proposed using a cross-modal transformer module. However, a module is used per modality combination to reinforce each modality representation. To reduce the computational cost, Sahay *et al.* (2020)

relied on a fused modality representation as the source modality to reinforce each modality (target), allowing for a linear quantity of cross-transformer modules. However, a corrupted modality would probably degrade the ReID system by corrupting the fused representation.

When a modality or a representation is weaker on average, its representation can be used only as a source and not as a target in the cross-modal transformer encoders (Yu *et al.*, 2019; Huang *et al.*, 2020; Khare *et al.*, 2021). This way, the modality supposedly positively impacts the final embedding despite its lower informative representation. In practice, if the most informative modality varies from one input to another, due to the capture conditions, for example, such a system seems not adapted. Instead, encoding both the modality-specific knowledge by the use of self-attention transformers and the modality-shared information through cross-modal transformer seems to be an interesting approach (Wei *et al.*, 2020b; Sun *et al.*, 2021; Lian *et al.*, 2021), conserving and exploiting the available knowledge as much as possible. Still, such architecture may be complex as it requires a transformer module for each modality (self-attention) and for each modality combination (cross-attention).

2.1.2.3 Hybrid CNNs and transformers architectures

Hybrid models use a combination of CNN and transformer architectures. Such models have been explored in different fields as it allows for lighter models and better performances when the data quantity is limited (Dosovitskiy *et al.*, 2020). For example, for expression recognition, Ma *et al.* (2021) first extract the features from each modality through modality-specific CNNs, fuse them, and finally feed them to a single ViT. In the medical field, the TransMed model (Dai *et al.*, 2021) feeds the multimodal data in a single stream CNN model followed by a transformer architecture instead of using one CNN stream per modality. Unlike the previous approaches, the Transfuser model (Prakash *et al.*, 2021), proposed for autonomous driving, uses self-attention transformer blocks between two CNN streams to refactor the representations during the feature extraction and finally fuses the two CNN outputs by concatenation. This approach offers new perspectives on how transformer self-attention benefits may be exploited in a hybrid architecture, limiting as well the model complexity.

2.2 Multimodal Fusion for Person ReID

2.2.1 RGB only

Various strategies were developed to improve the person ReID, using supplementary knowledge from an additional modality (Uddin *et al.*, 2023). When there is no will to bring additional sensors to the system, a supplementary modality can be directly extracted from the main modality. For example, Chen *et al.* (2019a) first extracts the contour from a given individual visible input and then encode the modalities with two backbones. With a similar approach, Bhuiyan *et al.* (2020) extract the pose information, and use this knowledge to guide a CNN backbone at different layers for the extraction of relevant discriminative appearance information. In fact, extracting contours or skeletons for multimodal guidance allows for information that is invariant to lighting and invariant to clothing changes concerning the skeletons. However, one can infer that a corrupted visible modality would affect the contour or pose extraction most of the time since those are extracted from this modality. For example, water splashes would likely corrupt the contour information or the pose extraction on the same parts of a given visible input. Hence, using one or plural supplementary sensors might be a more adapted strategy, as different sensors are independent per definition and encode the knowledge differently, consequently avoiding similar behaviors even under the same capture conditions. Using the same example, water spatter would not appear similarly on each sensor, so a model can compensate for the missing knowledge from one modality by the available knowledge from the other.

2.2.2 RGB-D

According to the survey proposed by Uddin *et al.* (2023) on multimodal person ReID, one can observe that most multimodal approaches focus on using either the skeleton or depth information. Exploiting the depth information requires another sensor, which seems to be a better strategy regarding real-world conditions. Recent RGB-D approaches consider, for example, attention mechanisms to extract the RGB foreground person thanks to a depth-based segmentation mask (Uddin *et al.*, 2020). Performing better, Uddin *et al.* (2021) proposed later to fuse the dissimilarity

representation from a trained model on the RGB modality only and a multimodal model trained on the concatenation of the visible and depth modalities. However, this approach requires both a unimodal and multimodal backbone, making it unsuitable for real-world scenarios. Also, even if the ReID accuracy is much improved through multimodal RGB-D sensors, the depth sensors cannot be used outdoors and have short-distance capture requirements. Infrared cameras are not affected by these strong limitations of the depth modality while allowing for dark environment capture as well. Still, the infrared does not allow foreground segmentation, making each sensor interesting depending on the setting and actual objective.

2.2.3 RGB-IR

With a focus on the use of the visible and the infrared modalities, a few approaches exist for person ReID. In fact, as a hint on the current research in this field, one can notice that the survey provided by Uddin *et al.* (2023) does only mention the V-I setting through the ancient approach proposed by Mogelmose *et al.* (2013), exploiting visible, depth, and infrared sensors. Nevertheless, fusion techniques have evolved much from 2013 and a person ReID literature is emerging on the V and I sensor exploitation.

First, using the V and I modalities in a multimodal way, Nguyen *et al.* (2017) is, to our best knowledge and apart from our work, the only existing approach focusing solely on fusing these two modalities. A backbone for each modality is trained independently and later used for inference by fusing the modality representations through a feature concatenation. In practice, the model is only evaluated on their proposed RegDB dataset and has no direct concurrence apart from their own model fusion experimental study. Recently, Zheng *et al.* (2021) introduced a dataset including the visible, near-infrared, and thermal-infrared (mid to far infrared) modalities and developed a new multimodal architecture based on three CNN branches, fusing the multimodal knowledge with a part-based strategy. Later, Wang *et al.* (2022) proposed a model architecture relying as well on a three-branch network but introduced a Cross-modal Interacting Module (CIM) instead. The CIM mainly relies on channel attention computed among the three modalities, allowing for refactoring each modality feature map based on how they correlate. However,

despite the interesting mechanisms engaged, using three modalities comes at the cost of larger models and increases the system's complexity.

Finally, one can say the I modality has excellent potential for a multimodal fusion with the V data, allowing for outdoor and nigh-time ReID while being promising in tacking corruptions through the added sensor representation. However, despite a vast literature for multimodal fusion, V and I fusion for person ReID remains weakly explored and must be further investigated.

2.3 Real-World Data

2.3.1 Constructing realistic evaluation datasets

Most datasets are captured in a controlled environment or over a short time period. These aspects make the collected data unrealistic in the way that a major part of the eventual capture conditions is not considered. For example, outdoor recognition is required to perform on sunny days but also under low light conditions and foggy or snowy days. Since every specific case scenario will not appear in the learning and testing data, the models likely do not handle them, leading to a substantial decrease in performance from the laboratories to the final application.

For realistic evaluation of models, recent approaches focused on building challenging evaluation datasets that cover a wider range of scenarios thanks to generated corruptions. As an example, Hendrycks & Dietterich (2019) proposed an artificially corrupted dataset, applying 15 types of common corruptions to the well-known object recognition dataset ImageNet (Deng *et al.*, 2009). The corruptions belong to different types of noise, weather, blur, and digital alterations, each corruption having 5 levels of severity. Soon after, those same corruptions were re-used to provide corrupted object detection datasets (Michaelis *et al.*, 2019), semantic segmentation datasets (Kamann & Rother, 2020), and pose estimation datasets (Wang *et al.*, 2021a). Recently, Chen *et al.* (2021) added 5 corruptions on top of the 15 provided by Hendrycks & Dietterich and especially evaluated the impact of the 20 corruptions on 21 person ReID models.

Despite the strong focus on providing corrupted datasets over the last few years, a lack of multimodal real-world datasets remains (Rahate *et al.*, 2022). In fact, to our best knowledge, with the exception of the recent work of Hong *et al.* (2023) on audiovisual data in emotion recognition, there are no existing corrupted multimodal datasets in computer vision. Plus, this work is actually focused on the speech recognition task, and consequently mainly focuses on corruptions that are not directly meaningful to other vision tasks like person ReID. For example, audio corruption has no place in a vision approach, and adding food-related occlusion patches specifically on the individual's mouth for person ReID is probably not the most accurate way to evaluate corruption robustness. Still, this work presents interesting conclusions, showing how a well-designed multimodal model can handle corruption that other models not designed for it cannot.

As the wide variety of eventual corruption "in the wild" cannot be covered, the objective is to evaluate models' generalization capacity better through these datasets and to find ways to train models that can improve it. Hence, the aforementioned datasets are only used for evaluation. As a common observation from their usage, one can say that models not trained with specific strategies to handle such data are usually poorly performing.

2.3.2 Processing real-world data

To handle real-world data and its unpredictability, models must have a great generalization capacity. Indeed, this allows for models to handle new situations better, leading to higher overall accuracy. For example, a foggy day may not appear in the learning data or may not sufficiently appear for a model to learn to handle it. However, a model must be operational under such weather conditions. Making a model better deal with unknown and new scenarios may appear complex since, by definition, no clues are available about these specific scenarios. However, in practice, solutions are numerous and must be analyzed to delimit the ones fitting the best with person ReID and its real-time requirements.

As an intuitive strategy, one can collect more data (Xie *et al.*, 2020) and cover more specific cases while learning a model. However, this comes at the cost of extra data collection and labeling work and does not ensure to improve the model's generalization power meaningfully. Working on the model's architecture can also be helpful in this manner. For example, the transformer architecture and its self-attention have been highlighted as a good answer to corruptions (Hendrycks *et al.*, 2020; Chen *et al.*, 2021). Also, a multimodal architecture can be an interesting solution as well (Hong *et al.*, 2023). Using a multimodal or transformer architecture is an interesting strategy and can be used as a basis, but improving a model architecture might be tedious.

Allowing for improved robustness to data corruption on any architecture without requiring more data collection and labeling, DA works at increasing the learning data and its distribution through data generation. For example, the Augmix strategy (Hendrycks *et al.*, 2019) mixes plural variations of the same image together, the variations being obtained through a random selection of image transformations. For person ReID, Chen *et al.* (2021) provided a data augmentation, adding random noise in a patch of the input or randomly extracting and adding a patch of the input on this same input. Plural concepts are covered through these data augmentations. For example, adding a patch on the image artificially forces the model to deal with occlusions. Also, adding noise to a patch on the image teaches the model to "read between the lines," seeking discriminant features partially appearing under a corrupted zone of the input image.

Focusing on multimodal fusion, one may wonder if multimodal data augmentation can be designed. In fact, there are only a few examples of such an approach. For image-text emotion recognition, Xu *et al.* (2020) proposed to learn a model which wisely generates image-text pairs based on unimodal image and text datasets. This process allows for building multimodal image-text pairs to be used as a pre-training dataset. Instead of using data augmentation to build pre-training data from different unimodal datasets, Hao *et al.* (2023) directly combines two image-text pairs by interpolating the same modality together. This way, a third image-text pair can be built to reinforce the model's robustness for every two image-text pairs. As an MDA closer to our task since the focus is here on V and I images for object detection, Nakamura *et al.* (2022) proposed the CutMix algorithm. CutMix generates cross-domain data for cross-domain

object detection. In practice, patches from the source domain are extracted and pasted on the target domain, forcing the model to project the same objects from each domain closer in the embedding space.

According to the previous analysis, the potential of DA for increasing corruption robustness is well-known and explored (Hendrycks *et al.*, 2019; Rusak *et al.*, 2020; Zhong *et al.*, 2020; Chen *et al.*, 2021). However, the multimodal setting is less investigated, despite its potential to allow for multiplying the DA options. Indeed, instead of learning ways to better select the knowledge intra-modality only as for unimodal DA, MDA should allow for strategies that work at improving both the intra and inter-modality knowledge selection. Consequently, MDA strategies must be further explored, especially their impact on the robustness of the multimodal models.

2.4 A Critical Analysis

To summarise the previous sections through a critical analysis of the covered approaches, Tab. 2.1 compares the different approaches. Several points are highlighted in this table and are interpreted as follows:

- The multimodal V and I setting has a strong potential for improving the person's ReID capacity, especially under real-world conditions. However, existing approaches are limited to four approaches, which use the V and the I modalities together for person ReID. Among them, only the work provided by Nguyen *et al.* (2017) is actually focused on using the V and I modalities exclusively.
- A multimodal corrupted dataset would allow for drawing a complete vision of the multimodal models' ability to adapt and ReID. However, a Corruption Robustness Analysis (CRA) through the design of a specific corruption evaluation dataset has only been proposed on unimodal data. Also, apart from Chen *et al.* (2021) work on person ReID, the others (Michaelis *et al.*, 2019; Hendrycks & Dietterich, 2019) are focused on different tasks.
- Number of existing MDA strategies is limited and MDA was never explored for person ReID, despite their strong potential in making models better adaptive and robust. Indeed, the work provided by Nakamura *et al.* (2022) on object detection shows the MDA potential

while working with the V and I modalities. However, existing techniques cannot be directly transferred to the person ReID task, which must be designed and explored.

In this thesis, the previous issues will be addressed through the design of innovative DL models specialized for V-I person ReID models, the creation of new corrupted datasets for multimodal models' CRA on this task, and the design of a MDA tailored to person ReID and its related challenges.

Table 2.1 Comparison of approaches related to this thesis. CRA stands for corruption robustness analysis. U, M, and C stands for unimodal, multimodal, and cross-modal, respectively. Aspects directly related to this thesis approach are highlighted in blue.

Approach	**Task**	**Configuration**	**CRA**	**DA**
Hendrycks *et al.* (2019)	Classification	U (RGB)	U	No
Michaelis *et al.* (2019)	Object Detection	U (RGB)	U	U DA
Xu *et al.* (2020)	Classification	M (RGB-language)	No	MDA
Nakamura *et al.* (2022)	Object Detection	C (RGB-IR)	No	MDA
Hao *et al.* (2023)	Representation learning	M (RGB-language)	No	MDA
Nguyen *et al.* (2017)	Person ReID	M (RGB-IR)	No	No
Chen *et al.* (2021)	Person ReID	U and C (RGB-IR)	U	U DA
Zheng *et al.* (2021)	Person ReID	M (RGB-IR-NIR)	No	No
Wang *et al.* (2022)	Person ReID	M (RGB-IR-NIR)	No	No
Focus of this Thesis	Person ReID	M (RGB-IR)	M	MDA

CHAPTER 3

FUSION FOR VISUAL-INFRARED PERSON REID IN REAL-WORLD SURVEILLANCE USING CORRUPTED MULTIMODAL DATA

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

[1] Laboratoire d'Imagerie, de Vision et d'Intelligence Artificielle (LIVIA), École de Technologie Supérieure,
1100 Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3

Article submitted to the "International Journal of Computer Vision (IJCV) Special Issue on Multimodal Learning" in April 2023.

Abstract

Visible-infrared person re-identification (V-I ReID) seeks to match images of individuals captured over a distributed network of RGB and IR cameras. The task is challenging due to the significant differences between V and I modalities, especially under real-world conditions, where images are corrupted by, e.g., blur, noise, and weather. Despite their practical relevance, Deep Learning (DL) models for multimodal V-I ReID remain far less investigated than for single and cross-modal V to I settings. Moreover, state-of-art V-I ReID models cannot leverage corrupted modality information to sustain a high level of accuracy. In this paper, we propose an efficient model for multimodal V-I ReID – named Multimodal Middle Stream Fusion (MMSF) – that preserves modality-specific knowledge for improved robustness to corrupted multimodal images. In addition, three state-of-art attention-based multimodal fusion models are adapted to address corrupted multimodal data in V-I ReID, allowing for dynamic balancing of the importance of each modality. The literature typically reports ReID performance using clean datasets, but more recently, evaluation protocols have been proposed to assess the robustness of ReID models under challenging real-world scenarios, using data with realistic corruptions. However, these protocols are limited to unimodal V settings. For realistic evaluation of multimodal (and cross-modal) V-I person ReID models, we propose new challenging corrupted datasets for scenarios where V and I cameras are Co-Located (CL) and Not Co-Located (NCL). Finally, the benefits of our Masking and Local Multimodal Data Augmentation (ML-MDA) strategy are explored to improve the

robustness of ReID models to multimodal corruption. Our experiments on clean and corrupted versions of the SYSU-MM01, RegDB, and ThermalWORLD datasets indicate the multimodal V-I ReID models that are more likely to perform well in real-world operational conditions. In particular, our ML-MDA is an important strategy for a V-I person ReID system to sustain high accuracy and robustness when processing corrupted multimodal images. The multimodal ReID models provide the best accuracy and complexity trade-off under both CL and NCL settings and compared to state-of-art unimodal ReID systems, except for the ThermalWORLD dataset due to its low-quality I. Our MMSF model outperforms every method under CL and NCL camera scenarios. GitHub code: https://github.com/art2611/MREiD-UCD-CCD.git.

3.1 Introduction

Real-world video monitoring and surveillance applications (e.g., recognizing individuals in airports and vehicles in traffic) are challenging problems that rely on object detection (Zou *et al.*, 2019; Zaidi *et al.*, 2022), tracking (Luo *et al.*, 2021), classification (Sen *et al.*, 2020), and re-identification (ReID) (Khan & Ullah, 2019; Ye *et al.*, 2021). Person ReID aims to recognize individuals over a set of distributed non-overlapping cameras. State-of-art ReID systems based on, e.g., deep Siamese networks (Fu *et al.*, 2021a; Sharma *et al.*, 2021; Somers *et al.*, 2023), typically learn an embedding through various metric learning losses, which seeks to make image pairs with the same identity closer, and image pairs with different identities more distant in the embedding space. Despite the recent advances with DL models, person ReID remains a challenging task due to the non-rigid structure of the human body, the different viewpoints/poses with which a person can be observed, image corruption, and the variability of capture conditions (e.g., illumination, scale, contrast) (Bhuiyan *et al.*, 2020; Mekhazni *et al.*, 2020).

Visible-infrared (V-I) person ReID aims to recognize individuals of interest across a network of V and I cameras. Unlike visible cameras, infrared ones allow night-time recognition. This has motivated research on cross-modal recognition to provide methods for V-I person ReID from night-time to day-time, or vice-versa (Ye *et al.*, 2021). In addition, a V-I person ReID approach has been proposed for a multimodal recognition (Nguyen *et al.*, 2017), where the I modality

is used in conjunction with the V, improving accuracy due to its different data encoding and perception under low light conditions. In fact, a V-I ReID can allow training a single model remains accurate over diverse capture conditions. A V-I ReID model should, however, conserve modality specific-features instead of focusing mostly on modality-shared ones (Baltrušaitis *et al.*, 2018), which is often absent, or not explicitly addressed by state-of-art approaches. Furthermore, V and I cameras may be co-located (CL) or not co-located (NCL), and variation in camera configuration affects the spacial alignment of V-I images, which is likely influencing ReID (as it is known to impact other tasks) (Wang *et al.*, 2021b; Xuan *et al.*, 2022).

Artificially corrupted datasets (Hendrycks & Dietterich, 2019; Chen *et al.*, 2021; Michaelis *et al.*, 2019) are important for evaluating V-I person ReID models, yet public datasets are often collected in controlled environments that cannot cover the range of real-world scenarios (Poria *et al.*, 2017). As highlighted by Rahate *et al.* (2022), there is a need to create multimodal real-world datasets that contain corrupted modalities. Apart from the recent approach using corrupted audio-visual data in emotion recognition Hong *et al.* (2023), the V-I ReID evaluation set proposed in our preliminary work (Josi *et al.*, 2023) is, to our best knowledge, the only existing dataset for corrupted evaluation for visual multimodal learning. However, the dataset in (Josi *et al.*, 2023) is only evaluated for a simple architecture and does not consider the correlation in the corruption from one camera to another. For example, corruption due to weather conditions should similarly occur on a V-I pair from co-located V-I cameras.

Neglecting to evaluate ReID models on corrupted data can result in large and unexpected performance gaps at deployment. To reduce this gap, one can attempt to restore corrupted input images during test time (Chang *et al.*, 2020), at the expense of pipeline complexity, by restoring the data before proceeding to the main ReID task. Using more complex DL models has been shown to improve performance on corrupted image data in object detection (Michaelis *et al.*, 2019) and image classification (Xie *et al.*, 2020). For instance, vision transformer models (Han *et al.*, 2020) have been shown some robustness to image corruption (Hendrycks *et al.*, 2020). In particular, the TransReID model He *et al.* (2021) provides state-of-art person ReID performance when facing corrupted data (Chen *et al.*, 2021). However, such complex models

limit the potential for real-time ReID applications. Using more diverse training data can improve the robustness of deep ReID models to corrupted data (Xie *et al.*, 2020), and does not increase the model's complexity at test time. Data augmentation (Shorten & Khoshgoftaar, 2019) also avoids the costs of data collection and annotation.

This paper focuses on the following research questions. How can efficient V-I ReID models be developed considering CL or NCL scenarios? How can these V-I models be trained, thanks to augmented multimodal data, to provide better robustness to real-world image corruptions than state-of-art models like TransReID?

In this paper, a cost-effective V-I ReID model named Multimodal Middle Stream Fusion (MMSF) is proposed to explicitly preserve and exploit both modality-specific and modality-shared knowledge, thereby improving robustness to corrupted images. In addition, three state-of-art attention-based models are adapted from the areas of sentiment analysis, emotion recognition, and action recognition for similarity matching, as needed for person ReID. Attention approaches are expected to address image corruptions through a dynamic feature selection, dealing with the varying availability of modality information. However, these models mainly focus on modality-shared features, eventually losing some capacity to discriminate.

Essential for the evaluation of both multimodal and cross-modal V-I person ReID models, corrupted V-I datasets are proposed for uncorrelated and correlated cases, named respectively Uncorrelated Corrupted Dataset (UCD) and Correlated Corrupted Dataset (CCD). These two sets allow for a robust evaluation of models based on 20 V and 19 I different corrupted conditions. Improving from our preliminary work, corruptions are correlated or not to suit NCL and CL camera configurations. In our experiments, we validate ReID models using clean and corrupted versions of the SYSU-MM01 (Wu *et al.*, 2017) (NCL), RegDB (Nguyen *et al.*, 2017) (CL), and ThermalWORLD (Kniaz *et al.*, 2018) (CL) datasets. Our preliminary work in (Josi *et al.*, 2023) introduced the Masking and Local Multimodal Data Augmentation (ML-MDA) strategy that improves the accuracy and robustness to strong image corruptions using simple fusion architecture. The strategy is further assessed in this paper and expected to train models leveraging

the complementary knowledge among modalities while dynamically balancing the importance of individual modalities in final predictions.

Main contributions:

(1) A novel MMSF architecture is proposed for V-I ReID that allows preserving both modality-specific and -shared features. This aspect is shown to be essential for both CL and NCL settings but is not addressed most of the time. Additionally, three state-of-art attention-based models are adapted to similarity matching, and evaluated for V-I person ReID. These models are detailed in Section 3.3.

(2) For realistic evaluation of V-I person ReID models, challenging UCD and CCD datasets are designed (see Section 3.4).

(3) The ML-MDA strategy presented Section 3.5 is introduced for training DL models for V-I ReID multimodal that are robust to corruption.

(4) Our empirical results (see Section 3.6) on clean and corrupted versions of the challenging SYSU-MM01, RegDB, and ThermalWORLD datasets provides insight about cost-effective DL models to adopt for V-I ReID, and their dependency on dataset properties and CL/NCL scenarios. Results also indicate that our V-I ReID models can outperform TransReID and related state-of-art models on clean and corrupted data in terms of accuracy and complexity.

3.2 Related Work

3.2.1 Multimodal fusion

3.2.1.1 Fusion approach and spatial alignment

To better handle or analyze a given problem, not being restricted to a single source of information is usually a powerful strategy (Baltrušaitis *et al.*, 2018; Wang, 2021). As well-known approaches, one can think of late (Snoek *et al.*, 2005) or sensor (Lohweg & Mönks, 2010) fusions. The

former considers independent learning and feature extraction for each modality before making a decision. Such fusions are easy to implement, as models can be trained independently and added to a system through minor adjustments. However, a model cannot learn the correlation between the modalities (Zhang *et al.*, 2017), like spatially related information. The latter (i.e., sensor fusion) stacks modalities together before any feature extraction, allowing inter-modality correlations to be mined and used by the model but considerably increasing the input dimension. Also, no spatial alignment may make modality correlations harder to find by the model (Wang *et al.*, 2021b).

Intermediate or model-level fusion techniques consider fusing modalities during the feature extraction and before the decision layer (Baltrušaitis *et al.*, 2018), increasing the semantic information contained in features before fusion and eventually making correlations easier to find. However, where spatial information continuously disappears through the network (Chen *et al.*, 2018a), it is unclear how much remains at each step and how it may impact a model. From experiments provided by Wang *et al.* (2021b) on fusion location and data alignment, it is important to differentiate spatially aligned and unaligned data as models may have really distinct behaviors.

3.2.1.2 Model level fusion

Model-level fusion considers fusing modality representations of a deep learning model somewhere in between the sensor representations and the feature vectors. Coordinated modality representation is seen by Baltrušaitis *et al.* (2018) as a challenging but promising fusion direction for model-level fusion approaches. Exchanging modality knowledge allows it and seems very practical as correlations may be mined by a model and as one modality may be more or less informative. However, they raise the models' lack of ability to conserve supplementary information and not only exploit complementary information.

In practice, attention-based multimodal approaches allow modality knowledge exchange, as it is the case for the MMTM proposed by Joze *et al.* (2020). The module refactors the channels

of each modality regarding how the intra- and inter-modality channels correlate. Based on the MMTM concept and inspired by Zhang *et al.* (2020) that used the split operation to improve the dynamic channel selection, Su *et al.* (2020) presented the MSAF approach. The dynamic channel refactoring in such multimodal models may allow for fine-grained feature selection and limit corruption impact. Unlike previous approaches, modality attention (Gu *et al.*, 2018), later updated by Ismail *et al.* (2020), provides soft attention weights for each modality to balance modality importance in the final embedding based on their discriminating capabilities. Again, such attention sounds to be a great approach to tackling punctually corrupted data. However, those attention models do not explicitly work at conserving the modality-specific knowledge, missing the point raised by Baltrušaitis *et al.* (2018).

Some transformers architectures tackle this aspect, conserving modality-specific knowledge through modality-specific streams and self-attention, and modality-shared knowledge thanks to modality-shared streams and cross-attention (Sun *et al.*, 2021; Lian *et al.*, 2021; Wei *et al.*, 2020a). However, transformer architectures are known to be complex and heavy Han *et al.* (2020), which do not align with video-surveillance challenges, requiring close to real-time algorithms.

3.2.1.3 Multimodal person ReID

Most approaches for person ReID (Ye *et al.*, 2021) focus on the unimodal (RGB) (Ristani & Tomasi, 2018; Luo *et al.*, 2019a) and cross-modal (Ye *et al.*, 2021; Alehdaghi *et al.*, 2023; Zhang *et al.*, 2022) settings. Few only focused on combining multimodal information. For example, Chen *et al.* (2019a) used the contour information. Bhuiyan *et al.* (2020) used pose information. However, for those approaches, the additional modality is built from the exploitation of the main modality, which would be similarly affected by image corruption and consequently not so helpful in this regard.

Using another sensor to extract a supplementary modality allows to have a distinct encoding, likely differently affected by corruptions. For example, the infrared and near-infrared are shown

to be beneficial for person ReID (Zheng *et al.*, 2021; Wang *et al.*, 2022), but leveraging the knowledge from three modalities might not be realistic for a real-world surveillance setting, asking for large models architectures.

Nguyen *et al.* (2017) represents the only approach where visible and infrared modalities only are integrated into a joint representation space. Infrared and visual features are concatenated, produced from independently trained CNNs, and used for pairwise matching at test time. This simple model attained an impressive performance on the RegDB dataset. However, RegDB data is captured with only one camera per modality, V-I cameras are co-located with only a single tracklet of ten images per modality and individual, and except for their low resolutions, captured images present no specific corruptions. For these reasons, the RegDB dataset is less consistent with a real-world scenario. In fact, the development of person ReID models that are effective in uncontrolled real-world scenarios remains an open problem (Hendrycks *et al.*, 2021).

3.2.2 Image corruption and augmentation strategies

Data augmentation (DA) consists in multiplying the available training dataset by punctually applying transformations on training images, like flips, rotations, and scaling (Ciregan *et al.*, 2012). This way, a model usually benefits from increased robustness to image variations and improved generalization performance. According to Geirhos *et al.* (2018), training a model on a given corruption is only sometimes helpful over other types of degradation. Yet, Rusak *et al.* (2020) showed that a well-tuned DA can help the model to perform well over multiple types of image corruption through Gaussian and Speckle noise augmentation. Hendrycks *et al.* (2019) proposed the Augmix strategy, for which multiple variations of an image are obtained through randomly applied transformations, variations that get mixed together. Random Erasing occludes parts of the images punctually by replacing pixels with random values (Zhong *et al.*, 2020). Previous strategies allow a large variety of augmented images, simulating eventually real-world data and hence inducing higher generalization performance.

Focusing on person ReID, Chen *et al.* (2021) proposed the CIL learning strategy to improve systems performance under corrupted data. Their strategy is partly based on two local DA methods – self-patch mixing and soft random erasing. The former replaces some of the pixels in a patch with random values, while the latter superposes a randomly selected patch from an image at a random position on this same image. Gong *et al.* (2021) show interesting improvements through local and global grayscale patch DA on RGB images. However, the previous strategies are limited to single modality stream models, even though the latter shows how grayscale data may reinforce the visible modality features using DA.

Multimodal data augmentation strategies have presented encouraging results for image-text emotion recognition (Xu *et al.*, 2020) or vision-language representation learning (Hao *et al.*, 2023). Also, Nakamura *et al.* (2022) proposed a visible-thermal cross-domain DA for few shots object thermal detection, working at closing the domain gap by augmenting data through hetero modality objects added on the main modality images. However, to our best knowledge, our preliminary work (Josi *et al.*, 2023) is the first to propose MDA with V-I person ReID applications through ML-MDA. Still, this MDA has only been investigated on a simple fusion model, which does not assure its generalization to more developed fusion architectures. Also, the evaluation is limited to corruptions set that do not consider eventual correlations between corruptions for NCL or CL cameras, which is tackled in this work.

3.3 Multimodal Fusion for V-I ReID

The main objective of our study is to find how modalities should be fused to be robust to data corruption while conserving great performances on clean data. Hence, plural multimodal models are studied, all trained and evaluated following a pairwise matching scheme (Fig. 3.1).

From our preliminary work (Josi *et al.*, 2023), the learned concatenation model is now used as a baseline, referred to as Baseline C. Baseline S stands as our second baseline with the same architecture but an element-wise sum fusion of the feature vectors instead of a concatenation.

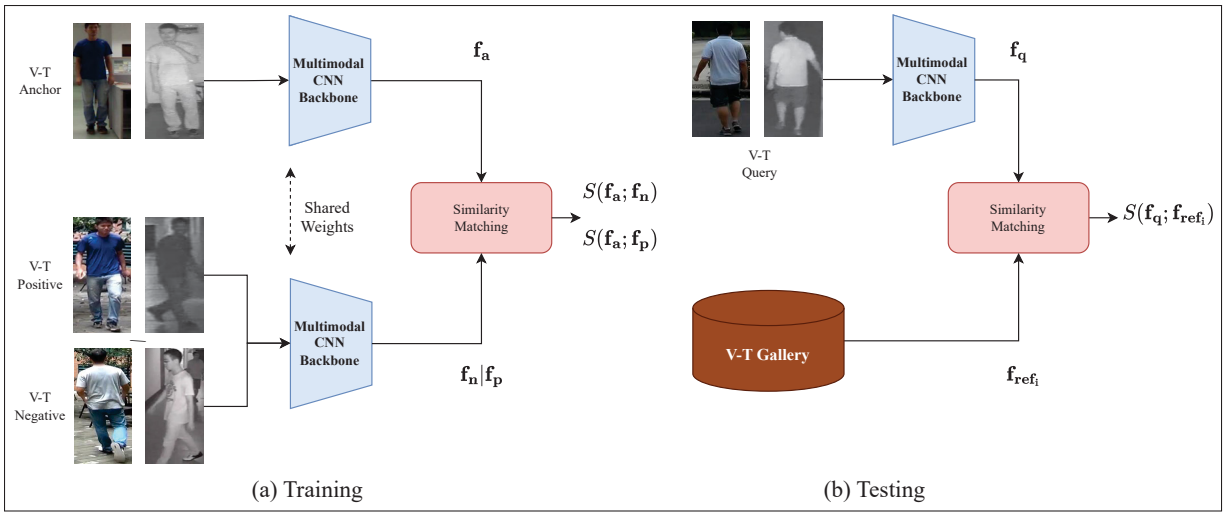

Figure 3.1 Representation of the multimodal person ReID (a) learning while using the triplet loss and (b) inference.

The selection of modality-shared and modality-specific features remains unclear in most models, whereas the importance of the conservation of both feature types has been highlighted by Baltrušaitis *et al.* (2018). Hence, the MMSF is proposed and first presented. Three attention-based models follow as attention should handle corruption well through a dynamic feature selection regarding each input. Still, the attention could also allow a modality corruption to degrade the hetero modality and require investigation. The three models are extracted from the literature and specially adapted to pairwise matching and, more precisely, to the person ReID task.

Fusion approaches are not restricted to a specific backbone, but ResNet-18 (He *et al.*, 2016) backbones are used for illustration purposes. Each model is optimized using the batch hard triplet loss (Hermans *et al.*, 2017) $\mathcal{L}_{\text{BH\_tri}}$, and cross-entropy with regularization via label smoothing (Szegedy *et al.*, 2016) $\mathcal{L}_{\text{CE\_ls}}$. We follow the usual optimization process (Ye *et al.*, 2021), except for the cross-entropy. Indeed, regularization via label smoothing is used by Chen *et al.* (2021) is is better at addressing corruption.

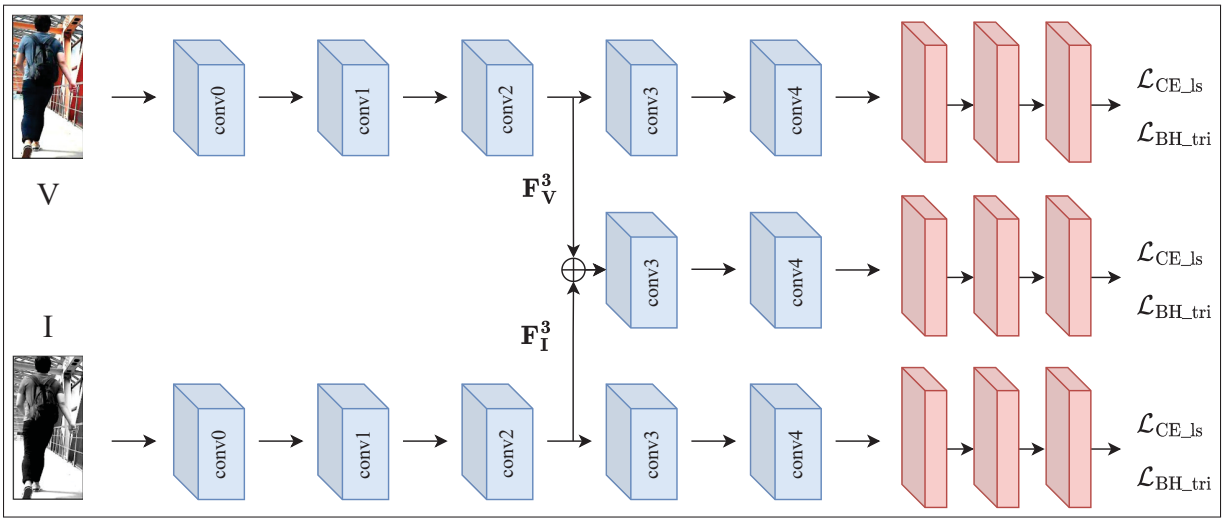

Figure 3.2 Training architecture of the MMSF model while fusing the features in the middle stream for $\ell = 3$.

3.3.1 Multimodal middle stream fusion

Assuring the conservation of the modality supplementary information, while taking advantage of the modality-shared information, we propose the MMSF model.

The model comprises two independent modality-specific CNN streams focused on the modality-specific information and a middle CNN stream that exploits the modality-shared information (Fig. 3.2). Each stream is independent and optimized through its specific loss functions, allowing it not to influence a stream representation from direct knowledge exchanges among streams. $\mathbf{F}_{\mathbf{V}}^{\ell} \in \mathbb{R}^{H \times W \times C}$ and $\mathbf{F}_{\mathbf{I}}^{\ell} \in \mathbb{R}^{H \times W \times C}$ are the visible and infrared feature maps before convolution blocks $\ell \in \mathbb{N}$. For a fusion before layer ℓ, the middle stream takes $\mathbf{F}_{\mathbf{m}} = \mathbf{F}_{\mathbf{V}}^{\ell} + \mathbf{F}_{\mathbf{I}}^{\ell}$ as input and pursues the feature extraction from this fused representation. Its middle stream size varies regarding ℓ value, being a partial backbone starting at layer block ℓ.

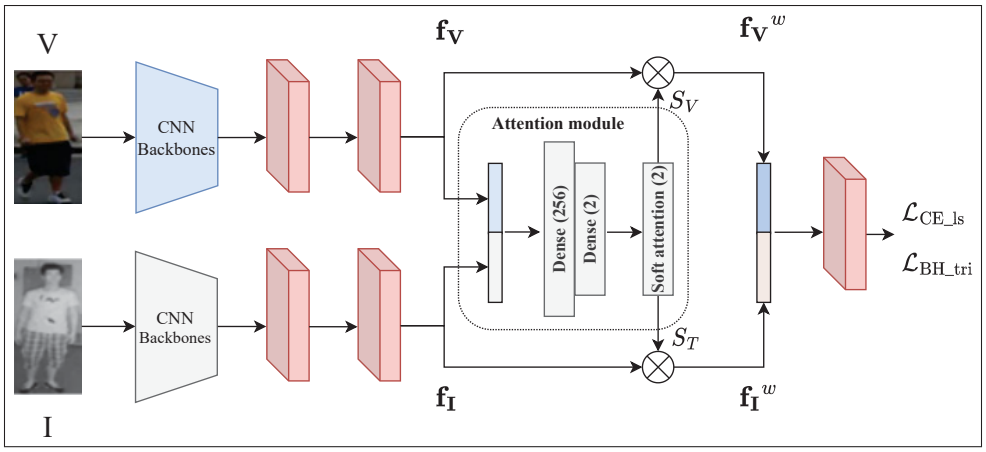

Figure 3.3 Training architecture of the MAN model.

3.3.2 Attention-based models

3.3.2.1 Modality attention network

Modality Attention Network (MAN) (Gu *et al.*, 2018) is an attention-based multimodal approach that dynamically weights feature vectors from each modality before fusing them. This model seems meaningful to explore as the dynamic weighting of each modality feature vector should help handle corrupted data. Since the model architecture has been adapted for our person ReID study, its architecture is presented in Fig. 3.3.

Two backbones first extract each visible $\mathbf{f_V} \in \mathbb{R}^d$ and infrared $\mathbf{f_I} \in \mathbb{R}^d$ modality features, with $d \in \mathbb{N}$. The obtained vectors are concatenated and passed through a modality attention module, which learns to generate soft attention weights. The soft weights allow the model to give more importance to the discriminant modality features in the final embedding. To do so, the concatenation of the two embeddings goes through two dense layers and a final softmax σ regression, which produces the soft weights $\mathrm{S_V} \in \mathbb{R}$ for the visible and $\mathrm{S_I} \in \mathbb{R}$ for the infrared modalities. Soft weights are produced as follows:

$$[\mathrm{S_V}, \mathrm{S_I}] = \sigma(\mathbf{W_2}\tanh(\mathbf{W_1}[\mathbf{f_V}, \mathbf{f_I}]^T + \mathbf{b_1}) + \mathbf{b_2}) \tag{3.1}$$

where $\mathbf{W_1} \in \mathbb{R}^{k \times d}$ and $\mathbf{W_2} \in \mathbb{R}^{1 \times k}$ are weight matrix, $k \in \mathbb{N}$ being an hyper-parameter, $\mathbf{b_1} \in \mathbb{R}^{k \times 2}$, and $\mathbf{b_2} \in \mathbb{R}^{1 \times 2}$ are biases.

Thanks to soft attention weights, visible and infrared original features are then weighted, respectively noted $w\mathbf{f_V}$ and $w\mathbf{f_I}$. For the visible modality, $w\mathbf{f_V} = \mathrm{S_V} \times \mathbf{f_V}$, and for the infrared modality, $w\mathbf{f_I} = \mathrm{S_I} \times \mathbf{f_I}$. Then, the predicted output vector $\hat{\mathbf{y}}$ is computed by passing the concatenation or the element-wise sum of the $w\mathbf{f_V}$ and $w\mathbf{f_I}$ vectors through a final softmax layer for classification.

As a consequence of the CL and NCL camera scenarios and the induced spatial alignment, which might influence the feature vector's composition, we also consider the element-wise sum fusion of the feature vectors in this work. Concatenation conserves each feature definition while fusing, but doubles the feature vector dimension. Summation makes the fused vector of the original feature vector size but may erase knowledge if the embedded concepts are not aligned.

3.3.2.2 Multimodal transfer module

The MMTM (Joze *et al.*, 2020) is an approach that focuses on channel attention to refactor the feature maps from two or more modality CNN streams regarding the spatial statistics of each. As the refactoring is done dynamically and based on the statistics of each given input, such attention should also be helpful while facing corrupted data. Two similar backbones are used to extract the features from each V and I representation. Two modules are used for our architecture (Fig. 3.4), after the third and the fourth convolution blocks, allowing for intermediate and high-level feature refactoring. For a given layer $l \in \mathbb{N}$, the visible and the infrared modality feature maps are respectively noted $\mathbf{F}_\mathbf{V}^\ell \in \mathbb{R}^{H \times W \times C}$ and $\mathbf{F}_\mathbf{I}^\ell \in \mathbb{R}^{H \times W \times C}$, with $H \in \mathbb{N}$, $W \in \mathbb{N}$ and $C \in \mathbb{N}$ being respectively the feature maps height, width and channel size. The feature map from each stream is first squeezed with a global average pooling layer over the spatial dimension, leading to two linear vectors of channel descriptors. Those vectors are concatenated and passed through a dense layer, following equation (3.2), to obtain the joint representation $\mathbf{J}^\ell \in \mathbb{R}^{C_J}$.

$$\mathbf{J}^\ell = \mathbf{W}([\mathrm{AvgPool}(\mathbf{F}_\mathbf{V}^\ell); \mathrm{AvgPool}(\mathbf{F}_\mathbf{I}^\ell)]) + \mathbf{b} \tag{3.2}$$

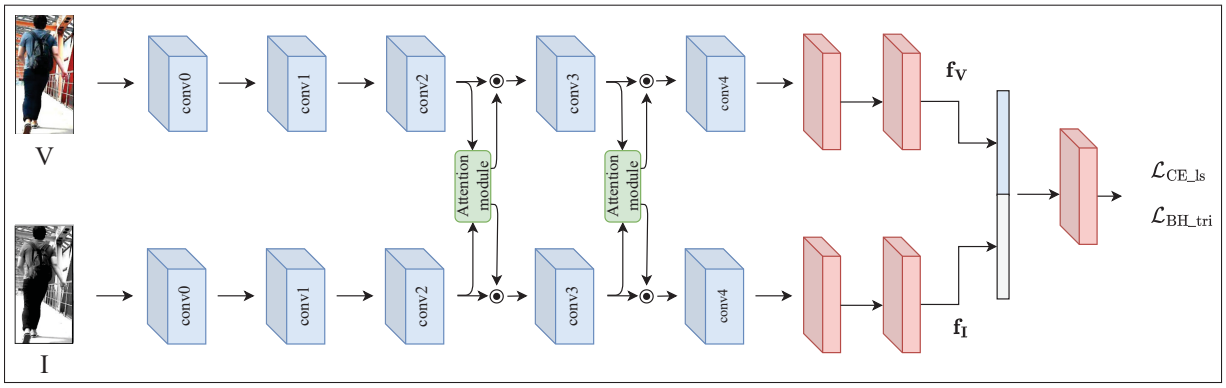

Figure 3.4 Learning model architecture for the MMTM and the MSAF approaches while concatenating the feature vectors for fusion. The attention module may be either the MMTM or the MSAF modules.

where $\mathbf{W} \in \mathbb{R}^{C_J \times C^2}$ is a weight matrix, $\mathbf{b} \in \mathbb{R}^{C_J}$ the bias of the dense layer, and $C_J = C^2/4$ to limit the model capacity and increase the generalization power (Joze *et al.*, 2020). Then, an excitation signal is produced with a distinct dense and softmax activation layer applied for each modality to the shared channel descriptor $\mathbf{J}^\ell$. Finally, this excitation signal is broadcasted through the spatial dimension for each modality with an element-wise product, following equations (3.3), forming the final weighted feature maps $w\mathbf{F}_\mathbf{V}^\ell \in \mathbb{R}^{H \times W \times C}$ and $w\mathbf{F}_\mathbf{I}^\ell \in \mathbb{R}^{H \times W \times C}$.

$$\begin{aligned} w\mathbf{F}_\mathbf{V}^\ell &= 2 \times \sigma(\mathbf{W}_\mathbf{V}\mathbf{J}^\ell + \mathbf{b}_\mathbf{V}) \odot \mathbf{F}_\mathbf{V}^\ell \\ w\mathbf{F}_\mathbf{I}^\ell &= 2 \times \sigma(\mathbf{W}_\mathbf{I}\mathbf{J}^\ell + \mathbf{b}_\mathbf{I}) \odot \mathbf{F}_\mathbf{I}^\ell \end{aligned} \tag{3.3}$$

where $\mathbf{W}_\mathbf{V} \in \mathbb{R}^{C \times C_J}$ and $\mathbf{W}_\mathbf{I} \in \mathbb{R}^{C' \times C_J}$ are weight matrix and $\mathbf{b}_\mathbf{V} \in \mathbb{R}^C$, $\mathbf{b}_\mathbf{I} \in \mathbb{R}^C$ the bias of the dense layers. σ stands for the sigmoid function. The element-wise product is represented by $\odot$.

3.3.2.3 Multimodal split attention fusion

The Multimodal Split Attention Fusion module (MSAF) proposed by Su *et al.* (2020) also works from the channel attention principle. Modules are applied at the same locations for this model (Fig. 3.4). Let us describe the MSAF module. First, the visible and infrared feature maps $\mathbf{F}_\mathbf{V}^\ell \in \mathbb{R}^{H \times W \times C}$ and $\mathbf{F}_\mathbf{I}^\ell \in \mathbb{R}^{H \times W \times C}$ are split into $n \in \mathbb{N}$ visible and infrared sub feature

maps, respectively noted $\mathbf{S}_{\mathbf{V}}^{\ell} \in \mathbb{R}^{H \times W \times \frac{C}{n}}$ and $\mathbf{S}_{\mathbf{I}}^{\ell} \in \mathbb{R}^{H \times W \times \frac{C}{n}}$. The n splits from each modality are element-wise summed and fed to a global average pooling layer to get a global channel descriptor per modality noted $\mathbf{J}_{\mathbf{V}}^{\ell} \in \mathbb{R}^{\frac{C}{n}}$ and $\mathbf{J}_{\mathbf{I}}^{\ell} \in \mathbb{R}^{\frac{C}{n}}$. Then, the global channel descriptor from each modality is element-wise summed and passed through a dense layer, followed by a batch normalization and a ReLU activation to catch the inter-channel correlations, forming the common channel descriptor $\mathbf{J}^{\ell} \in \mathbb{R}^{\frac{C}{n}}$. From $\mathbf{J}^{\ell}$, n excitation signals are produced per modality, using a dense layer and a softmax activation on $\mathbf{J}^{\ell}$ for each original feature map split. These excitation signals are then broadcasted through the spatial dimension for each split with an element-wise product, following equations 3.4, forming the final weighted splits $w\mathbf{S}_{\mathbf{V}}^{\ell} \in \mathbb{R}^{H \times W \times \frac{C}{n}}$ and $w\mathbf{S}_{\mathbf{I}}^{\ell} \in \mathbb{R}^{H \times W \times \frac{C}{n}}$.

$$
\begin{aligned}
w\mathbf{S}_{\mathbf{V}}^{\ell} &= \sigma(\mathbf{W}_{\mathbf{V}}\mathbf{J}_{\mathbf{V}}^{\ell} + \mathbf{b}_{\mathbf{V}}) \odot \mathbf{S}_{\mathbf{V}}^{\ell} \\
w\mathbf{S}_{\mathbf{I}}^{\ell} &= \sigma(\mathbf{W}_{\mathbf{T}}\mathbf{J}_{\mathbf{I}}^{\ell} + \mathbf{b}_{\mathbf{T}}) \odot \mathbf{S}_{\mathbf{T}}^{\ell}
\end{aligned}
\tag{3.4}
$$

The n excited splits are concatenated together for each modality to get the final weighted feature maps $w\mathbf{F}_{\mathbf{V}}^{\ell} \in \mathbb{R}^{H \times W \times C}$ and $w\mathbf{F}_{\mathbf{I}}^{\ell} \in \mathbb{R}^{H \times W \times C}$. One can notice that the model needs fewer parameters than the MMTM approach, thanks to the feature map splits.

3.4 Corrupted Datasets

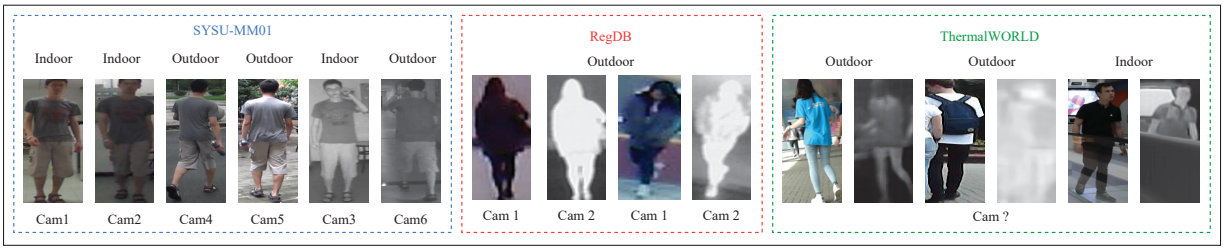

Figure 3.5 Examples from SYSU-MM01, RegDB and ThermalWORLD. ThermalWorld does not provide camera information.

To better simulate real-world conditions while evaluating a model, the focus has been on corrupted test sets over the last few years (Hendrycks & Dietterich, 2019; Chen *et al.*, 2021; Michaelis *et al.*, 2019). However, those benchmark test sets were proposed for single modality settings, whereas our objective is to evaluate the value of V-I multimodal models. As both the V

and the I modalities encode from visual cues, corruptions that affect the visual modality may also affect the infrared modality, such as occlusions or weather-related corruptions. From this observation, the 20 visible corruptions from Chen *et al.* (2021) are extended to the infrared domain in this work, allowing us to provide two corrupted datasets. Those two datasets are suited to the CL and the NCL settings.

In the following sections, the three clean datasets are first detailed. A presentation of the used modality corruptions follows. Finally, our two corrupted datasets are detailed.

Table 3.1 Datasets statistics. **V** = **V**isible and **I** = **I**nfrared. Image size and number of samples per identity are presented as: Min;Max;Avg. BRISQUE (Mittal *et al.*, 2011) measure is shown as: avg±std.

Statistic	**SYSU**	**RegDB**	**TWORLD**
V-images	29 033	4120	8125
I-images	15 712	4120	8125
V-Camera	4	1	16
I-Camera	2	1	16
Cameras setting	NCL	CL	CL
Identities	491	412	409
V-images/id	10;144;59.1	10;10;10	1;155;19.9
I-images/id	10;144;32.0	10;10;10	1;155;19.9
Image width	26;1198;111	64;64;64	10;810;141
Image height	65;879;291	128;128;128	25;897;353
V-BRISQUE	30.50±12.26	38.84±9.86	27.79±13.28
I-BRISQUE	40.52±8.42	38.81±9.56	60.25±8.67

3.4.1 Clean datasets

The three used datasets present distinct statistics (Tab. 3.1) suited to build and draw a strong study.

SYSU-MM01 (Wu *et al.*, 2017) gather 4 visible and 2 infrared cameras, with 491 distinct individuals, 29033 RGB, and 15712 I images. The V and I cameras are not co-located, so the scene's spatial description varies from one modality to another for a given V-I image pair.

RegDB (Nguyen *et al.*, 2017) is a much smaller dataset, with one camera only per modality, the V and I cameras being co-located. A single 10 images tracklet is available per identity and camera. Hence, RegDB 412 identities lead to 4120 images per modality.

[3]**ThermalWorld** (Kniaz *et al.*, 2018) has only its training part available, leading us to 409 distinct identities. 16 co-located cameras per modality captured each 8125 image. However, the infrared images are of terrible quality, with a BRISQUE (Mittal *et al.*, 2011) value of 60.25, much higher than RegDB and SYSU-MM01 ones, being at 38.81 and 40.52 respectively.

3.4.2 Modality corruptions

Hendrycks & Dietterich (2019); Chen *et al.* (2021) used 20 corruptions of the visual modality, which were regrouped into four distinct types - noise, weather, blur, and digital. In this work, the used corruptions are the same for the visual modality. However, the I modality can also be affected by multiple corruptions, which is considered. In fact, 19 of the corruptions affecting the visual modality can also apply to the infrared with a few slight adjustments (Corruptions taxonomy figure and corruptions adjustments table in the Appendix II.1).

First, the current luminosity does not impact the I modality, so brightness corruption is not used for this modality. Then, different noises, like Gaussian, Shot, Impulse, and Speckle, are applied similarly, except each noise is turned into grayscale values to respect the infrared modality single color channel encoding. Spatter and frost are two other corruptions that needed to be grayscaled before being applied to the infrared images. Indeed, blue-colored water or brown-colored dirt was applied for spatter, and frozen blue masks for frost. As a last adjustment, the saturation is expressed differently for the I modality, visually brightening the object of interest eventually if this one is too close to the camera, instead of modeling color intensity for the visual modality. Finally, all other corruptions were applied similarly for the V and I modalities.

---

[3] Download link (https://drive.google.com/file/d/1XIc_i3mp4xFlDJ_S5WJYMJAHq107irPI/view) obtained from github ThermalGAN issues (https://github.com/vlkniaz/ThermalGAN/issues/12).

3.4.3 Uncorrelated corruption dataset

The Uncorrelated Corruption Dataset (UCD) is proposed as a first way to evaluate the models' corruption robustness. To build UCD, the corruptions are randomly and independently selected and applied on each modality for a given V-I test pair, making it highly challenging. The camera corruption independence from V to I modality is suited for a NCL camera setting, as it is the case for SYSU-MM01. Indeed, for example, a visible indoor and an outdoor infrared camera would lead to weather appearing only on the infrared camera or to blur, impacting one camera only while the other is impacted independently. As applied corruptions are most of the time distinct from one modality to another under UCD, it should allow each modality to compensate for the corrupted features from the other. Hence, this setting should be a great way to evaluate the models' ability to select the information of interest from one or another modality.

3.4.4 Correlated corruption dataset

One can expect some corruption to be correlated from one camera to another, corruption type-wise as intensity-wise. As a brief example, the rain is expected to appear on both visible and infrared cameras simultaneously, especially if those are co-located. However, some other types of corruption, such as image saturation, are camera dependent and would happen punctually on one camera with no correlation with the other. The CCD dataset is proposed from these observations, suited for CL cameras and gathering the following characteristics (Tab. 3.2).

At first, weather-related corruptions such as fog, rain, frost, and snow appear much correlated, so the weather from one camera is assumed to appear with the same level of corruption on the other. Spatter expresses the water or dirt splashes on the cameras, which has a great chance to happen on both cameras considering co-located cameras, but with a level that might differ; the level is selected randomly and independently. Similar behaviors for blur-related corruptions would also make sense in real-world conditions if cameras are co-located since those corruptions are a consequence of camera settings, like exposure time or focus, for example, but which also mostly depends on the current scene. Because each modality camera might be more or less

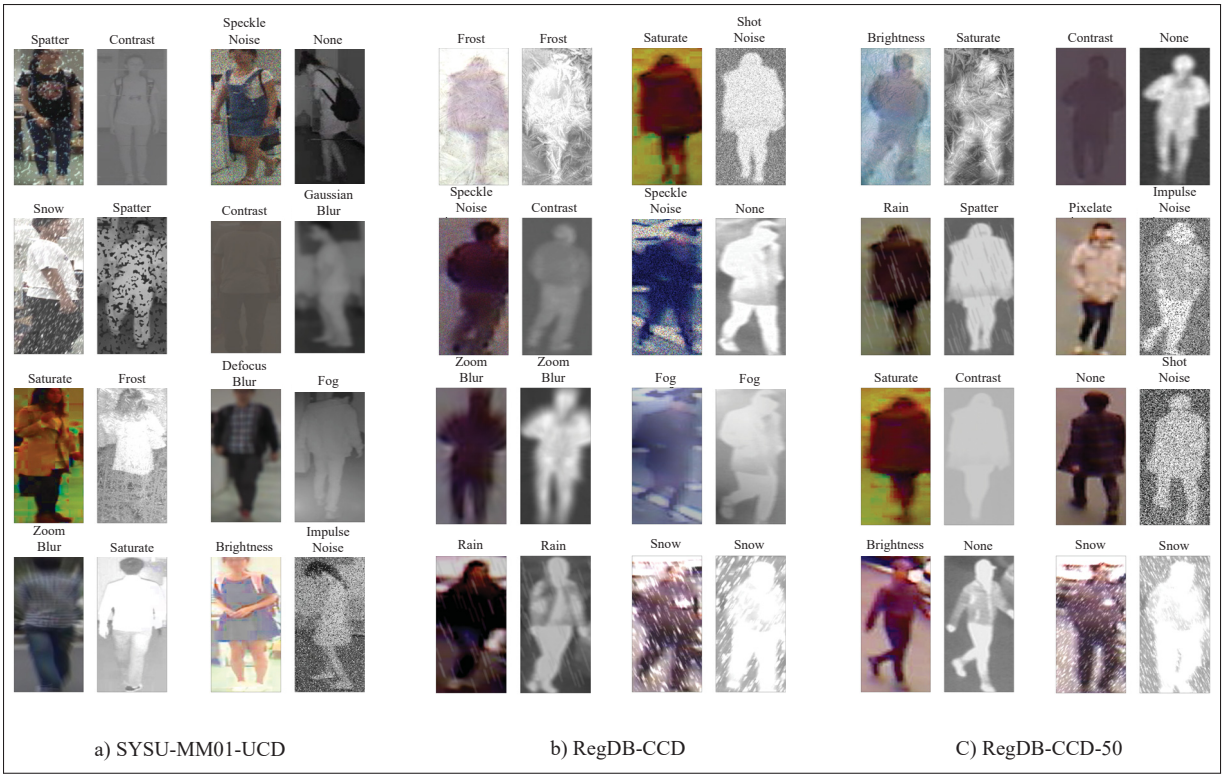

Figure 3.6 Samples from our three corrupted datasets. Visuals do not represent all available dataset versions, as each dataset has its own UCD, CCD, and CCD-50 version.

reactive regarding the situation, we consider that blur-related corruptions (i.e., defocus, gaussian, glass, zoom, motion blurs) affect the two modalities simultaneously with an intensity level that can differ. The intensity level is randomly and independently selected except for motion blur corruption. Indeed, infrared cameras usually have a higher exposure time than visible cameras, making those more affected by motion blur. Consequently, the level is always selected as equal or superior for the infrared modality compared to the visible one.

Concerning the ten remaining corruptions, those are much related to data encoding and can affect visible or infrared cameras independently. The hetero-modality is consequently corrupted if the selected corruption lies in the correlated corruptions. Otherwise, we randomly apply another corruption among the uncorrelated ones to the hetero-modality. Considering modalities as always corrupted is an extreme scenario, which is attractive to frame models' behaviors but not entirely realistic. Hence, the UCD-X dataset is proposed. In this configuration, $X\%$ of the

Table 3.2 Correlated (center) and uncorrelated (right) corruptions are presented, along with the relation between levels of corruption (left) from the V to the I modality for correlated corruptions.

Level	Correlated	Uncorrelated
V = I	Fog	Gaussian noise
V = I	Frost	Shot noise
V = I	Snow	Impulse noise
V = I	Rain	Speckle noise
V ≠ I	Spatter	Elastic transform
V ≠ I	Defocus blur	Saturation
V ≠ I	Gaussian Blur	JPEG compression
V ≠ I	Glass Blur	Pixelate
V ≠ I	Zoom Blur	Contrast
V ≤ I	Motion Blur	Brightness

(a) S-REA (b) MS-REA (c) Masking

Figure 3.7 Soft random erasing (S-REA) (Chen *et al.*, 2021) and our MDA based on multimodal soft random erasing (MS-REA) and modality masking.

corrupted pairs affected by uncorrelated corruptions are formed with one of the two modalities remaining clean. In practice, we fixed it at 50%, but this value can be tweaked to make the datasets more or less challenging for further experiments.

3.5 Multimodal Data Augmentation

The explored models are based on co-learning, allowing each modality stream to adapt to the other one (Baltrušaitis *et al.*, 2018; Rahate *et al.*, 2022). We propose a new MDA approach,

the Masking and Local Multimodal Data Augmentation (ML-MDA), for better learning of the models. In practice, ML-MDA is based on two components: Multimodal Soft Random Erasing (MS-REA) and modality masking (Fig. 3.7). Those two data augmentations are used together during the learning process to make the learned co-learning model robust and accurate in a challenging inference environment.

3.5.1 Multimodal soft random erasing

Making a multimodal model focus on modality-specific features is challenging, as the model usually mainly focuses on shared features (Baltrušaitis *et al.*, 2018). Augmenting the multimodal data with local occlusions may help the model to emphasize modality-specific feature importance, as some features will be available only from one or another modality. The soft random erasing (Chen *et al.*, 2021) (S-REA) (Fig. 3.7.a.) uses local occlusions to learn the model not to rely only on the most important features, but consider unimodal learning and consequently not exploit this aspect.

The MS-REA data augmentation is proposed to close this gap. Instead of replacing a proportion of the pixels in a given image patch with random pixel values for the visible modality only as S-REA, MS-REA applies a patch on both the visible and infrared modalities. Grayscaled random pixel values are used for patches on the thermal modality to respect the infrared thermal image definition, encoded on one channel, and potentially aligning better with real-world corruptions. The spatial patch location is randomly and independently selected from the visible and infrared images for a given V-I pair. To close the occlusion gap brought by the applied patches through MS-REA, the model must learn how to select each modality feature when partial information is available from each modality. Such behavior is expected to extend well to real-world corruption.

3.5.2 Modality masking

A modality might be punctually unavailable or primarily uninformative. Thus, the model shall learn how to cancel a modality to reduce its impact on the final prediction. The modality masking approach is expected to allow it by punctually replacing one or another modality with an entirely blank image. Instead of masking the multimodal representation as it has been done by Gabeur *et al.* (2022), a representation is extracted from the masked input, so the model has to learn how to cancel its influence on the final results. This also forces the model to focus more on modality-specific features since one modality only contains all the meaningful knowledge for ReID. This DA should supposedly complement the previously presented MS-REA approach by balancing each modality's importance in the final embedding regarding each modality level of corruption, whereas MS-REA should learn the model to select the features within each modality better. MS-REA should also make models' put more emphasis on the modality-specific features, this time thanks to the independent occlusions locations on each image.

3.6 Results and Discussion

3.6.1 Experimental methodology

Data division. SYSU-MM01 and RegDB datasets have well-established V-I cross-modal protocols (Wang *et al.*, 2019a,b; Ye *et al.*, 2019), but multimodal protocols were not existing prior to our preliminary work (Josi *et al.*, 2023). Following them again, 395 and 96 identities from SYSU-MM01 are respectively used for the training and the testing set. For RegDB, 412 identities are divided into two identical sets of 206 individuals for learning and testing. The SYSU-MM01 train/test ratio is kept for ThermalWORLD, leading to 325 training identities and 84 for testing. A 5-fold validation (Raschka, 2018) is performed over the data used for training, using folds of respectively 79, 41, and 65 distinct identities for SYSU-MM01, RegDB, and ThermalWORLD.

Data augmentation. Our proposed multimodal extensions MS-REA is used with the same appearance augmentation probability as S-REA (Chen *et al.*, 2021). Modality Masking is applied randomly on one or another modality, with equiprobability, and occurs with a default probability of $1/8$. When used on unimodal models, the CIL (Chen *et al.*, 2021) DA is used the same way as the original authors. For the RegDB dataset only, the validation set is given the same DA as the training set as the maximum performances were reached in the early epochs otherwise. This way, better convergence was observed, allowing learning complex cues by the model.

Pre-processing. A data normalization is done at first by re-scaling V and I images to 144×288. Random cropping with zero padding and horizontal flips are adopted for base DA. Those parameters were proposed by Ye *et al.* (2021) on RegDB and SYSU-MM01 datasets. The same normalization is kept under ThermalWORLD for consistency among protocols.

Performance measures. The mean Average Precision (mAP), and the mean Inverse Penalty (mINP) are used as performance measures, commonly used for person ReID (Ye *et al.*, 2021). The mAP is the mean computed over all query image ratio of retrieved matches over total matches. However, mAP does not reflect the worst-case scenario, unlike the mINP measure, which applies a penalty on the hardest matches, making it a great complementary measure.

Hyperparameters. The hyperparameters values in our models were set based on the default AGW (Ye *et al.*, 2021) baseline. The SGD is used for training optimization, combined with a Nesterov momentum of 0.9 and a weight decay of $5e-4$. Our models are trained through 100 epochs. Early stopping is applied based on validation mAP performances. The learning rate is initialized at 0.1 and follows a warming-up strategy (Luo *et al.*, 2019b). The batch size is 32, with 8 distinct individuals and 4 images per individual. The paired image is selected by default for RegDB and ThermalWORLD. For the SYSU-MM01 dataset, the images from the hetero modality are randomly selected through the available ones to form a pair for a given identity.

Losses. The Batch Hard triplet loss (Hermans *et al.*, 2017) $\mathcal{L}_{\text{BH\_tri}}$ and the cross-entropy with regularization via Label smoothing (Szegedy *et al.*, 2016) $\mathcal{L}_{\text{CE\_ls}}$ are used as loss functions for

our models. Indeed, the former is widely used in person ReID approaches (Wang *et al.*, 2019a; Choi *et al.*, 2020; Ye *et al.*, 2021), so the same margin value is fixed at 0.3, and the latter is part of the CIL implementation (Chen *et al.*, 2021). The total loss corresponds to the sum of both losses. The batch hard triplet loss aims at reducing the distance in the embedding space for the hardest positives while increasing the distance for the hardest negatives. The regularization with label smoothing reduces the gap between logits, making the model less confident in predictions and hence improving generalization (Müller *et al.*, 2019).

Models details. MMSF is used with $\ell = 4$ for NCL and $\ell = 0$ for CL cameras (Appendix II.2). The influence of concatenation or sum of the feature vectors is explored in the Appendix II.3 and allowed to converge to use MMTM S (Sum) and MSAF C (Concatenation) for RegDB, and MMTM C and MSAF S for ThermalWORLD and SYSU-MM01.

Leave-one-out query strategy. The Leave-One-Out Query (LOOQ) strategy, proposed in our preliminary work (Josi *et al.*, 2023), is used the same way in this study. The LOOQ treats the extreme but meaningful case in which one would have only a unique image of the person to ReID and multiple footages containing images of this same person in the gallery. Every pair of images is alternatively used as a probe set while all the other pairs join the gallery. While an interesting evaluation strategy, this also allows us to respect the original dataset statistics (Tab. 3.1) by authorizing the number of used gallery images per individual to vary.

3.6.2 Scenario with not co-located cameras

NCL V-I cameras imply that a pair of images for a given individual is built from two distinct viewpoints. Consequently, images in a given V-I pair will not be spatially aligned from one modality to another. Having two viewpoints for a given V-I pair should allow more cues and be more discriminant to ReID than a CL setting. Indeed, if the person is occluded or partially visible from one camera modality, for example, the hetero-modality camera might have a better view and compensate for the missing features. However, correlations from one modality to another may be harder to find for NCL cameras as the scene appears much different between

modalities (Wang *et al.*, 2021b). For example, the spatial information remaining in the features when the fusion is done may act as noise for the model due to the absence of alignment.

Since various corruptions can impact either modality, a multimodal model might be disturbed by the supplementary modality and could consequently be less able to ReID than a well-trained single-modal model. The upcoming study is proposed to determine whether or not the multimodal framework is worthwhile given the above statements and to seek the best approach to follow.

3.6.2.1 Robustness to corruption

Table 3.3 Unimodal and multimodal models performances while evaluated on clean and corrupted SYSU-MM01 datasets. Unimodal V and I stands respectively for unimodal visible and thermal models. In bold and blue are the first and second best approaches respectively.

	Model	Clean		UCD		CCD		CCD-50	
		mAP	mINP	mAP	mINP	mAP	mINP	mAP	mINP
No DA	Unimodal V	86.72	41.16	**32.16**	1.86	**32.11**	1.89	**37.70**	2.15
	Unimodal I	77.06	30.44	13.97	1.25	13.51	1.25	18.26	1.31
	Baseline S	95.96	71.14	22.55	1.82	19.26	1.71	31.90	2.37
	Baseline C	96,47	73,69	25.01	1.90	24.35	1.86	31.24	2.26
	MAN	91.05	55.00	27.76	1.84	27.72	1.84	33.99	2.17
	MMTM	95.71	71.35	20.00	1.59	18.31	1.70	30.59	2.25
	MSAF	96.77	77.27	25.64	**2.03**	21.77	**1.93**	34.58	**2.54**
	MMSF	**97.80**	**80.93**	22.23	1.65	17.70	1.60	31.15	2.12
ML-MDA / CIL	Unimodal V	86.72	42.70	52.37	3.89	52.58	3.93	55.48	4.67
	Unimodal I	78.33	35.41	33.38	2.32	32.78	2.32	36.26	2.39
	Baseline S	96.54	74.49	64.00	9.72	61.81	7.53	64.69	8.26
	Baseline C	96.77	76.01	63.40	9.51	61.94	7.72	65.79	8.71
	MAN	97.13	77.91	63.50	8.24	61.87	6.39	64.75	7.10
	MMTM	95.81	74.23	64.41	**11.49**	62.30	**8.55**	64.91	**9.22**
	MSAF	96.36	73.70	**67.78**	10.09	**65.49**	8.00	**68.91**	9.14
	MMSF	**97.66**	**79.52**	65.24	10.41	63.16	7.44	65.58	8.36

Multimodal models are compared while evaluated on each clean, UCD, CCD, and CCD-50 version of the SYSU-MM01 evaluation data. Clean data is important as a reference, observing performances under the best-case scenario. UCD and CCD should complete each other. The

former will allow observing how the models can adapt and select information from differently corrupted V-I inputs. The latter will present how the models can deal with similarly corrupted inputs, which should make the task harder as it should happen more often that the same features for a given pair get corrupted from V to I. Finally, the CCD-50 should be the easiest evaluation set, with 50% of the pairs having one over two modalities remaining clean. This last set should allow observing if some models better deal with punctual unilateral corruption.

3.6.2.1.1 Natural models corruption robustness

To begin with, the models are trained without any data augmentation technique and evaluated on the original and corrupted versions of SYSU-MM01 (Upper half Tab. 3.3). The considered unimodal models are fine-tuned ResNet-18 models, trained from visible (unimodal V) or infrared (unimodal I) modality only.

Before seeking the models' robustness to corruption, observing good performance on clean evaluation data is essential. In practice, each multimodal approach improves over the unimodal models. From the unimodal to the multimodal setting, the greatest improvement comes between the unimodal visible and the proposed MMSF approach, improving the mAP and mINP percentile point (PP) by $11,08$ and $39,77$, respectively. The impressive performance improvement shows how the infrared modality and the NCL cameras through clean SYSU-MM01 strongly benefit the multimodal ReID.

For each corrupted test set (i.e., UCD, CCD, and CCD-50), as both modalities can impact the ReID either way, one can observe here that each multimodal model (learned without a specific strategy) is less efficient than the unimodal V specialist. Indeed, the unimodal V model reaches 32.16% mAP, followed by MAN at 27.76% mAP for SYSU-MM01-UCD, for example.

Focusing on the corrupted datasets, the multimodal models globally reach lower performances from UCD to CCD as expected, with, for example, the MMSF model being respectively at 22.23% or 17.70% mAP. From CCD to CCD-50, one can see that some models seem to react

better to unilateral corruptions, as the mAP improvement in PP for the proposed MMSF is about 13.45, for MSAF about 12.81, and for Baseline C about 6.89.

Among multimodal models, the ranking is inconsistent, from the clean to the corrupted setting. On clean data, the proposed MMSF model presents the highest performances, with mAP about 97.80% and mINP about 80.93%, closely followed by the attention-based MSAF approach, reaching 96.77% mAP and 77.27% mINP. On corrupted data, MAN and MSAF appear as better at handling corruption than MMSF.

From there, it would be hard to advise one or another model with the aim of ReID under real-world conditions. Indeed, the evaluated multimodal models are shown not to learn how to select the right modality information in the face of corruption without using a corruption-dedicated learning strategy.

3.6.2.1.2 DA impact on models robustness

Performances for the unimodal specialists learned using CIL and the multimodal models learned using ML-MDA are presented in the bottom half of Tab. 3.3.

The CIL use for unimodal models increases the models' performances on clean data. Indeed, especially for the unimodal infrared model, mAP and mINP are respectively improved by 1.27 PP and 4.97 PP for example. ML-MDA has an impact that is model dependant. Still, most models conserve similar mAPs, except for the baseline sum and MAN that see it considerably improve.

Considering corrupted evaluation sets, using DA brings an impressive corruption robustness improvement to every model. The unimodal V model under UCD improves, for example, from 32.16% to 52.37% mAP, or the baseline C model from 24.35% to 61.94% mAP under the CCD set. Similar improvements are happening under each corrupted setting.

Using ML-MDA, the multimodal models' performances are now ahead of the unimodal ones by a strong margin on corrupted datasets. Indeed, the models learn to select information from each corrupted modality way better. Also, its usage brings more consistency from the clean to the corrupted setting, making it essential to handle real-world conditions. For example, the proposed MMSF model was, and remains, the most discriminant approach under clean data but is now the second-best approach on corrupted datasets. In contrast, this one came in the fifth position at best without DA.

One may wonder if the multimodal models benefit more from clean data pairs than the unimodal specialists. Results from CCD to CCD-50 (that has 50% of its pairs containing one clean modality) should help for this analysis. In fact, the mAP gap from unimodal V to MSAF increases by 0.62 percentage points from CCD to CCD-50, and decreases from unimodal V to MMSF by 0.48 points. Hence, the multimodal setting seems to benefit globally as much from the clean data pairs as the unimodal V model. However, as 50% of V-I pair have a clean image, it means 25% of the data is clean for the unimodal V, which shows, in a way, that the multimodal models are benefiting less from a clean modality but keep up with the unimodal V model thanks to the doubled amount of clean pairs. For deeper analysis, each corruption impact and unilateral corruption are further explored in the next section, so as a qualitative analysis through class activation maps (CAMs) generation (Appendix II.5).

3.6.2.2 Specific corruption impact

Corruption can sometimes be one-sided, as with NCL cameras or digital corruption. Hence, we may wonder whether some corruptions of the infrared modality will make the unimodal V model advantageous against multimodal models. This question is also raised for visible corruptions and the unimodal I specialist against multimodal models. To answer those, performances of the unimodal visible and thermal models and those of MSAF and MMSF are observed regarding each corruption (Tab. 3.4) while corrupting only either modality. The MSAF and the proposed MMSF models are selected as those that performed the best over the evaluated multimodal models.

Weather-related corruptions are the most challenging over the infrared modality, reflected in the lower MSAF and MMSF performances under those data alterations. Compared to the unimodal V model, MSAF is under for 4 corruptions and MMSF for 5 (in red), and both are, on average, much higher for the other. When the RGB modality only is corrupted, MMTM and MSAF models globally conserve a great performance margin over the unimodal I model without corruption. Indeed, it only happens twice among the 20 V corruptions, with contrast and saturation, that those two multimodal models get under unimodal I (in blue). This leads us to affirm that unimodal corruptions are globally very well handled by the multimodal models that can extract some interesting cues from the corrupted modality while not getting regrettably impacted on the clean modality input most of the time.

Comparing the proposed MMSF to the proposed MSAF model, we may observe that MMSF deals better with most corruptions, except for the very challenging ones. Indeed, the I weather alterations are very challenging, and one can see the snow corruption leading, for example, the MSAF model to 81.77% mAP, against 46.99% mAP for MMSF. In fact, strong corruptions may completely alter $2/3$ of the MMSF fused embedding (Corrupted modality stream feature and the modality shared one), whereas the MSAF attention may simply refactor features in the corrupted modality so that they do not influence too much the final embedding. For weaker corruptions, having a specific stream to mine the right cues while having a specific stream that exploits the correlations among modalities is better.

Table 3.4 Corruption-wise performance comparison between unimodal, MSAF and MMSF models and while corrupting one or the other modality only. Models were trained using DA. In Red are visible model performance without corruption and multimodal models performances that gets lower those due to thermal corruptions. In blue are thermal model performances without corruption and models that get lower those due to an RGB corruption.

Corruption	**V Corrupted**						**I Corrupted**					
	Unimodal V		**MSAF**		**MMSF**		**Unimodal I**		**MSAF**		**MMSF**	
	mAP	**mINP**	**mAP**	**mINP**	**mAP**	**mINP**	**mAP**	**mINP**	**mAP**	**mINP**	**mAP**	**mINP**
No corruption	86.72	42.70	96.36	73.70	97.66	79.52	78.33	35.41	96.36	73.70	97.66	79.52
Gaussian noise	73.88	23.58	93.49	62.30	95.97	70.89	43.82	6.07	90.90	50.72	92.20	55.58
Shot noise	79.53	30.95	94.75	67.63	96.85	75.42	43.97	6.48	91.07	51.43	92.09	55.23
Impulse noise	73.46	23.44	93.22	61.46	95.91	70.54	36.78	4.26	89.87	47.89	90.44	49.53
Speckle noise	82.76	35.37	95.53	70.57	97.31	77.67	53.22	10.31	92.84	57.31	93.96	62.02
Defocus blur	75.03	23.68	94.75	67.08	96.58	73.83	59.68	13.56	94.21	63.25	95.58	68.22
Glass blur	80.12	30.75	95.42	69.74	97.08	76.53	62.38	15.83	94.58	65.26	96.08	70.89
Motion blur	79.52	29.55	95.21	68.63	97.03	76.16	65.51	17.03	94.87	65.55	96.56	72.79
Zoom blur	76.50	26.27	94.73	67.55	96.48	73.91	54.17	11.24	93.15	59.67	94.73	64.56
Gaussian blur	74.51	22.79	94.66	66.55	96.50	73.21	59.43	13.33	94.09	62.87	95.40	67.80
Snow	36.18	4.88	85.46	43.74	85.02	41.23	5.87	1.15	81.77	33.67	46.99	5.16
Frost	28.08	2.36	83.09	38.02	77.10	28.61	14.82	1.46	85.30	38.31	74.31	21.06
Fog	34.22	3.82	84.48	40.91	87.71	42.69	16.23	1.43	86.36	40.06	79.71	27.57
Brightness	66.96	16.83	92.46	59.21	94.69	64.89	/	/	/	/	/	/
Rain	50.66	7.99	87.86	47.65	89.85	51.28	31.79	2.85	89.56	47.24	86.52	41.69
Spatter	70.28	21.06	93.08	61.45	95.09	67.49	29.18	3.27	88.34	45.37	78.86	32.04
Contrast	23.00	1.45	77.64	29.44	78.08	26.19	33.04	2.90	85.32	37.07	88.18	39.89
Elastic trsf	75.67	26.93	94.50	66.67	96.39	73.27	42.88	6.43	92.16	55.63	92.46	55.91
Pixelate	84.62	37.65	96.24	73.24	97.70	79.99	73.35	25.02	96.01	71.51	97.47	78.46
JPEG compr	69.43	18.43	93.69	62.20	95.32	67.99	70.08	20.83	95.47	68.53	96.92	75.24
Saturation	66.39	16.59	72.06	23.99	57.25	10.71	45.61	5.83	90.85	51.01	93.58	59.07

3.6.2.3 Comparison with state-of-art

The multimodal baseline, MSAF, and the proposed MMSF models get compared in terms of complexity and accuracy against the unimodal V and the state-of-art unimodal models LightMBN and TransREID (Fig. 3.8). The Appendix II.4 provides detailed performance and additional comparison. Unimodal models are learned using CIL DA, and the multimodal models using our ML-MDA. The accuracy is obtained over both SYSU-MM01 clean and CCD evaluation sets and gathered Fig. 3.8 along with the models' number of parameters (params) and FLOPs.

Performance-wise, each multimodal model is more interesting on clean data than the best unimodal approach, LightMBN. For the best ReID overall, MMSF is the best model, although it performs slightly under MSAF regarding mAP on corrupted data. In fact, MSAF would be favored for a highly challenging environment, especially when facing strong unilateral corruption.

Complexity-wise, LightMBN is the best model to adopt but comes with a considerable performance decrease from our MMSF, the gap being about 3.32 mAP PP and 15.59 mINP PP on clean data.

3.6.2.4 Discussion

Experiments over the SYSU-MM01 dataset give us an excellent overview of the multimodal power under the NCL configuration. The main conclusions are as follows:

- The proposed ML-MDA is essential for the multimodal models to handle corruption. This way, models learn how to select the right information from each modality and not get disturbed by noisy features.

- For the best ReID, the proposed MMSF should be used in priority, followed by MSAF, and finally by the unimodal LightMBN models if the memory resources do not allow it.

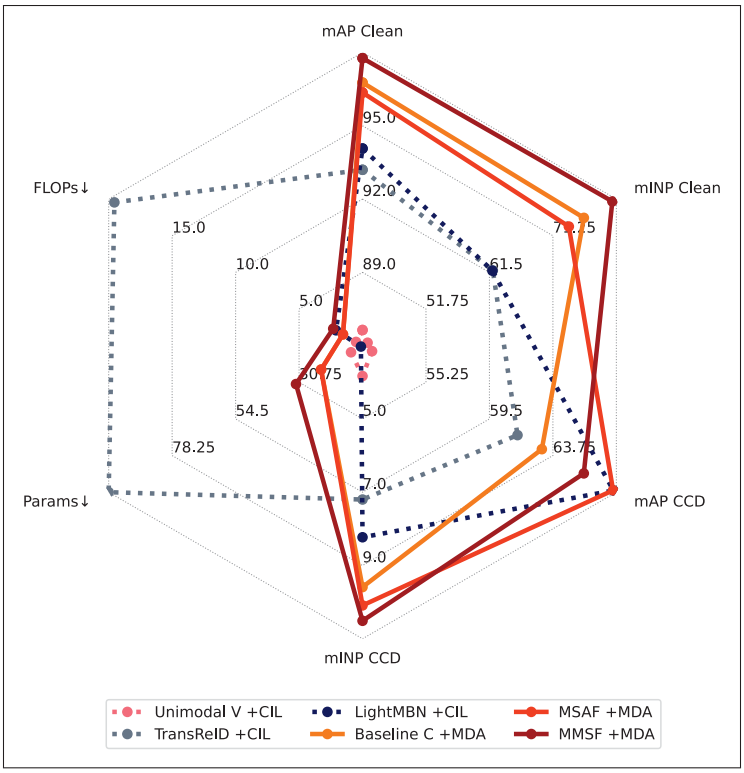

Figure 3.8 Complexity and accuracy trade-off on the SYSU-MM01 clean and CCD sets. Dashed lines and plain lines are, respectively, unimodal and multimodal approaches. Measures marked with '↓' should be minimized for an optimized model.

- The high multimodal accuracy on corrupted data is to highlight as both modalities always get corrupted through the UCD evaluation set, making the task highly challenging.

- The multimodal setting appears as a much better answer to data corruption than the transformer-based approach TransReID, both regarding complexity and accuracy. In fact, TransReID performs less than expected from its performances without DA (Chen *et al.*, 2021), the CIL strategy making, for example, the LightMBN more interesting.

3.6.3 Scenario with co-located cameras

The spatial alignment brought by co-located V-I cameras should make the correlations from one modality to another easier to find for a model.However, this might not make much difference for fusions that come late in the model, as the spatial information will be much diminished and supposedly replaced by semantic information. Also, a corrupted V-I input brings some disequilibrium in how each modality contains relevant information, which should perturb the multimodal models and eventually influence the correlation benefits of spatial alignments. Previous assumptions are explored in the next sections.

3.6.3.1 Robustness to corruption

Table 3.5 Unimodal and multimodal models performances while evaluated on clean and corrupted RegDB datasets.

	Model	Clean		UCD		CCD		CCD-50	
		mAP	**mINP**	**mAP**	**mINP**	**mAP**	**mINP**	**mAP**	**mINP**
No DA	Unimodal V	99.19	96.71	**40.54**	**5.13**	**40.70**	**5.01**	**45.43**	**6.19**
	Unimodal I	98.92	96.03	21.94	1.33	21.71	1.31	27.89	1.72
	Baseline S	99.39	97.60	18.66	1.75	20.73	1.57	26.86	2.17
	Baseline C	99.64	98.46	21.73	2.39	23.45	2.10	29.64	2.83
	MAN	99.36	97.51	29.02	3.47	29.07	3.15	35.33	4.06
	MMTM	99.53	98.01	18.78	2.06	19.67	1.76	26.15	2.48
	MSAF	99.86	99.26	23.42	2.82	24.23	2.37	31.05	3.32
	MMSF	**99.88**	**99.36**	32.63	5.07	31.54	3.79	38.99	5.41
ML-MDA / CIL	Unimodal V	99.51	98.21	54.61	12.58	54.61	12.51	58.43	14.56
	Unimodal T	98.92	96.12	44.62	6.46	44.27	6.41	50.13	8.49
	Baseline S	99.87	99.37	62.48	20.34	59.33	14.60	63.89	17.34
	Baseline C	99.90	99.45	61.92	20.14	59.06	14.64	64.15	18.08
	MAN	99.90	99.43	62.24	23.38	60.64	18.49	65.15	21.62
	MMTM	99.84	99.24	69.06	25.32	63.34	17.81	67.27	20.17
	MSAF	99.88	99.33	61.70	19.99	58.82	15.12	63.87	18.28
	MMSF	**99.95**	**99.69**	**76.47**	**39.51**	**71.52**	**30.43**	**74.25**	**33.24**

Table 3.6 Unimodal and multimodal models performances while evaluated on clean and corrupted ThermalWORLD datasets.

	Model	Clean		UCD		CCD		CCD-50	
		mAP	**mINP**	**mAP**	**mINP**	**mAP**	**mINP**	**mAP**	**mINP**
No DA	Unimodal V	87.38	51.71	28.74	4.50	28.97	4.47	35.28	5.18
	Unimodal I	56.17	10.65	24.45	3.78	24.33	3.78	27.65	3.99
	Baseline S	86.44	46.55	30.34	4.84	29.99	4.76	36.32	5.55
	Baseline C	87.92	50.41	30.43	4.77	30.51	4.80	36.96	5.65
	MAN	87.50	51.98	29.10	4.56	29.15	4.54	35.62	5.26
	MMTM	88.01	49.97	30.15	4.73	29.95	4.71	36.58	5.52
	MSAF	88.13	51.28	29.68	4.64	29.36	4.63	35.94	5.40
	MMSF	**89.43**	**52.83**	**30.91**	**5.20**	**30.86**	**5.07**	**37.44**	**6.23**
ML-MDA / CIL	Unimodal V	86.37	47.42	52.77	9.51	52.83	9.43	56.28	10.79
	Unimodal I	55.29	9.81	32.21	4.61	32.26	4.60	34.01	4.68
	Baseline S	82.18	36.89	54.49	10.59	52.97	9.72	55.68	10.55
	Baseline C	86.34	43.24	56.10	11.04	55.20	9.93	58.01	11.02
	MAN	87.11	45.47	59.22	11.19	57.54	10.56	60.23	11.64
	MMTM	**87.82**	47.95	59.98	12.55	58.12	11.53	60.51	12.36
	MSAF	87.62	**50.02**	60.38	11.30	58.10	10.03	60.78	10.93
	MMSF	86.10	44.50	**62.58**	**14.45**	**60.75**	**13.33**	**62.77**	**14.24**

3.6.3.1.1 Natural models corruption robustness

To begin with, the RegDB and ThermalWORLD models are learned without the use of data augmentation, and their performances are respectively gathered in the upper half Tab. 3.5 and 3.6.

The models must be robust to corrupted data but must also be accurate on clean data at first. Indeed, an optimal model would perform well under the two scenarios. On clean data, the multimodal models are improving over the unimodal visible and Thermal specialists, except for the ThermalWORLD sum model. Precisely, our MMSF model comes first for the two datasets, both regarding mAP and mINP.

On corrupted evaluation sets, RegDB presents a unimodal visible accuracy considerably ahead of every multimodal model, showing the multimodal model's lack of adaptation while facing

corrupted data. Indeed, the unimodal V model is, for example, at 45.43% mAP, when the following approach is our MMSF model reaching only 38.99% mAP under the CCD-50 set. In reverse, ThermalWORLD observes a considerable improvement with the multimodal setting. Indeed, the unimodal model is behind every multimodal approach for each corrupted dataset version. The most significant improvement comes from the proposed MMSF model again, reaching 40.01% mAP, whereas the unimodal V reaches 35.28% mAP.

The lousy thermal modality quality makes the ThermalWORLD dataset distinct from RegDB, which can explain why the multimodal models naturally better handle corruptions. Indeed, this might seem counter-intuitive as a lower quality modality should help less for the ReID, but this more likely indicates that the challenging ThermalWORLD learning environment helps the multimodal models to handle corruption better. This learning configuration forces the models to learn how to adapt regarding each input quality. Under this assumption, higher corruption robustness can be expected from MDA strategies since it works on related concepts by synthetically bringing noisy samples into the learning process.

As a supplementary observation, the gap from unimodal to multimodal models performance is much lower under the CL datasets than under SYSU-MM01 and its NCL cameras. Here, the highest performance improvement in PP from the visible to the best multimodal model is about $0,69$ mAP and $2,65$ mINP for RegDB, and about $2,1$ mAP and $1,77$ mINP for ThermalWORLD. In comparison, the gap in PP was about 11.8 mAP and 39.77 mINP for SYSU-MM01. This performance gap change might result from the CL cameras concerning RegDB and ThermalWORLD, the additional modality bringing fewer supplementary cues than the NCL setting, as an expected consequence of the spatial alignment. For ThermalWORLD, the gap change is likely also due to the terrible thermal modality quality (BRISQUE value Tab. 3.1), reflected in the mAP gap from the unimodal V to the unimodal T model, being of 31.21 PP. For RegDB, the unique camera per modality probably influence this aspect as well, making the problem easier, leading to almost maxed-out performances that do not allow similar improvement through the multimodal setting.

3.6.3.1.2 DA impact on models robustness

Models performances while considering data augmentation strategies are presented lower half Tab. 3.5 and 3.6 respectively for RegDB and ThermalWORLD.

Moving from no use of DA to its usage leads to impressive performance improvements on corrupted data. Where the RegDB multimodal models performed lower than the unimodal visible model using no DA, all multimodal models learned with our ML-MDA become way ahead of the visible model. The greatest improvement comes from unimodal V to our proposed MMSF model, which increases the mAP by 16.91 PP for CCD and 15.82 PP for CCD-50.

For corrupted versions of ThermalWORLD, for which multimodal models already had better performances than unimodal specialists before DA, the performance gap significantly increases with ML-MDA usage. Considering CCD evaluation, for example, the gap from unimodal V to the best approach being MMSF is about 7.92 mAP, where it was about 1.89 mAP percentage points without DA.

The massive multimodal corruption robustness improvement from the proposed multimodal data augmentation on the two datasets makes it a crucial approach. With it, the MMSF model becomes the best working approach for RegDB, followed by MMTM. In fact, modalities are both corrupted most of the time, so the attention through MMTM and MSAF probably becomes tough to adjust for the models. MMSF does not allow another modality to bring additional noise in its modality-specific streams and consequently better benefits from each input. Also, its central stream can focus only on the encoding of the modality correlations and eventually improve the ReID even more.

3.6.3.2 Comparison with state-of-the-art

For CL cameras, multimodal MMSF and MMTM models get compared to the state-of-art unimodal models under both RegDB and ThermalWORLD Clean and CCD evaluation sets. The accuracy is put in perspective of the models' complexity through their number of parameters

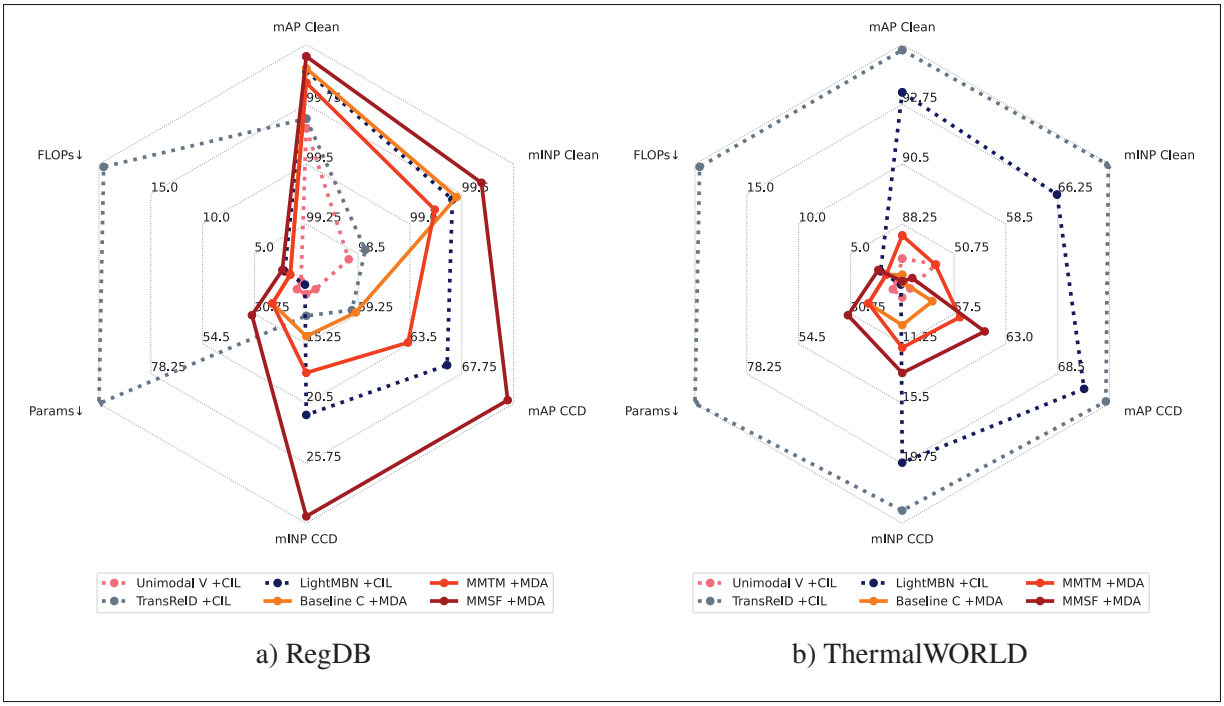

Figure 3.9 Complexity and accuracy trade-off using clean and CCD evaluation sets. Dashed lines and plain lines are, respectively, unimodal and multimodal approaches. Measures marked with '↓' should be minimized for an optimized model.

(params) and FLOPs (Fig. 3.9). Only the CCD evaluation set is considered as this configuration is the most adapted to CL cameras (Section 3.4) and should allow drawing the main conclusions.

For RegDB (Fig. 3.9a), the best-performing model is our proposed MMSF model in terms of accuracy, both on clean and corrupted data. The model is followed by LightMBN and then by MMTM. Hence, the complexity and accuracy trade-off comes between LightMBN and the MMSF model, MMSF being the best way for a strong ReID, and LightMBN for a lighter but lesser efficient approach.

Focusing on ThermalWORLD (Fig. 3.9b), the story is different. Despite the same CL camera configuration as RegDB, the two compared multimodal models are much less accurate than the unimodal LightMBN and TransReID models while having more parameters and needing more FLOPs than LightMBN. This large gap in behavior from RegDB to ThermalWORLD comes from the latter dataset's infrared quality again. Still, for a similar ResNet-18 backbone

architecture through Unimodal V, we observe that the multimodal models are more accurate. This shows how the multimodal models can benefit from the additional modality even if this one is of low quality, but that it is not enough to compare with LightMBN and TransReID discriminant power. Finally, among TransReID and LightMBN models, it is again an accuracy and complexity trade-off. Heavier but more discriminant is TransReID for ThermalWORLD, and much lighter but also less discriminant is LightMBN.

3.6.3.3 Discussion

The previous analysis under the CL setting from RegDB and ThermalWORLD datasets allowed us to reinforce some observations from the NCL setting and draw additional conclusions that are as follows:

- The proposed ML-MDA data augmentation is crucial for a multimodal model to handle challenging data in NCL and CL settings well. Also, models still benefit much from the MDA when the original dataset is challenging, as observed through ThermalWORLD.
- Our MMSF model deals substantially better with clean and corrupted data than every other approach, including TransReID, despite its highlighted corruption dealing (Chen *et al.*, 2021). The early fusion likely allows the model to better apprehend and disentangle the corrupted features from the clean ones between modalities. Considered attention approaches exchange stream information later in the process and consequently have already lost an essential part of the modality correlations. Plus, they do not have modality-specific streams as MMSF, whereas it assures that the final embedding conserves features from a good modality definition and also ensure the model does not only focus on modality-shared features.
- The deficient infrared data quality of the ThermalWORLD dataset does not allow the multimodal setting to compare with unimodal state-of-art.

3.7 Conclusion

Real-world surveillance and especially person ReID is a complex task that requires models to handle complex and abstract concepts, handle data corruption and remain lightweight. To

address these challenges, the multimodal setting can be a powerful tool, as an additional modality brings supplementary information that can help to reach higher accuracy while it allows reaching competitive complexity thanks to lightened backbones. However, real-world conditions and the subsequent data corruptions (e.g., weather, blur, illumination) have to be considered. To this aim, our study proposes a strong V-I multimodal evaluation through the first V-I corrupted evaluation sets (UCD and CCD) for multimodal (and cross-modal) V-I person ReID, tackling the lack of multimodal real-world datasets (Rahate *et al.*, 2022). Precisely, 20 visible and 19 infrared corruptions are considered, 3 datasets, 2 camera settings (NCL and CL), 2 state-of-art person ReID models, a MDA, 6 multimodal models, comprising 3 attention-based, 2 baselines, and our proposed MMSF architecture.

Experiments on the clean and proposed corrupted datasets converge to present the proposed ML-MDA as a must-use to make any multimodal model way more robust to real-world events. The multimodal models observe a larger margin of improvement from the NCL rather than the CL scenario as a consequence of the additional information provided by the NCL complementary view. Still, the benefits of plural modalities are unequivocal for both scenarios, the TransReID model being way more complex and less accurate than plural multimodal approaches (except under really low-quality infrared through the ThermalWORLD dataset). Especially, among multimodal approaches, our MMSF model comes ahead of every considered model for the two scenarios, highlighting the importance of considering modality-specific features not tackled in attention state-of-art models.

To extend this work, vision-based MDA could be further explored as it showed great benefits but remains not much investigated in the literature. Also, the proposed MMSF has shown weakness while facing strongly and unilaterally corrupted data, which has less impact on attention-based models. Hence, adding the right attention modules may allow getting the best of both worlds. Finally, different backbones could be explored for a better accuracy/complexity ratio.

CONCLUSION AND RECOMMENDATIONS

This research presents a solution for addressing the fundamental challenges in ReID, including the diverse range of individual representations, real-world conditions, and near real-time requirements. A three-component framework is proposed to tackle these challenges. The first component focuses on multimodal fusion, specifically exploring the fusion of visible and infrared modalities to enhance the performance of person ReID. The second component involves building evaluation data that reflects real-world uncertainties by designing V-I corrupted datasets. The proposed datasets unlock the V-I multimodal person ReID evaluation under realistic and challenging conditions for the community, making a huge difference in the current multimodal computer vision literature. The last component revolves around finding effective learning strategies to handle different modalities and challenging data, with a particular emphasis on multimodal data augmentation. The main findings for each component are summarised in the following paragraphs.

This thesis explores distinct fusion processes through CNN and transformer architectures, allowing or not knowledge exchanges during the feature extraction and using attention or not to select features better. Attention mechanisms have been shown to be interesting and especially valuable under strong unilateral corruption with channel attention and the MMTM model. In this case, the attention allows for a strong modality refactoring, resulting in a less corrupted final representation. Transformer-based approaches with hybrid architectures have been shown as considerably heavy but especially not so beneficial (Appendix II). Indeed, transformers bring complexity through a large number of parameters while having a disputable potential, eventually bringing a better ReID on the hardest corrupted samples but globally lowering performances compared to the proposed MMSF or the attention-based CNN models. However, valuable in a broader spectrum of cases, the proposed MMSF model has been able to overpass every other approach under most scenarios, relying on a stable encoding process through its three branches architecture. This allows us to explicitly highlight the necessity to conserve modality-specific knowledge and show that attention is not the only solution to real-world conditions, despite its highlighted potential.

The proposed V-I corrupted datasets are the first multimodal computer vision datasets that involve corruptions. Their design enables the evaluation of models under broader and more realistic conditions and is shown as essential for a multimodal model evaluation. To do so, this required the construction of infrared corruptions, inspired in practice by existing corruptions of the visible modality. A first V-I corrupted dataset named "-C*" is proposed in Appendix I and then renamed UCD in Chapter 3. The UCD dataset does not consider the correlation of corruption between cameras, which is meaningful for the NCL setting but not tailored for CL cameras, which would likely be similarly corrupted depending on the data modification considered. The subsequent version of UCD, named CCD, is introduced in Chapter 3, tackling this aspect by considering the corruption correlation between cameras.

Developing multi-modal models robust to corrupted data through the adapted learning strategy is shown as a necessity. Precisely, the ML-MDA multimodal data augmentation is proposed in this regard and demonstrated to be highly effective in each multimodal model studied, without bringing complexity to the final pipeline. Based on two distinct approaches, the masking and the multimodal soft random erasing, ML-MDA learns the models to discard a modality influence in the final fused embedding when this one is highly corrupted and poorly informative compared to the other. Plus, the models learn to rely on a wider scale of features within a modality through local occlusions, but also on the hetero-modality features as the occlusions are independently positioned on a pair of images.

In conclusion, this thesis offers new perspectives in the field of person ReID and multimodal fusion. It explores distinct yet interconnected aspects, including models, learning data, and learning strategies, to develop a comprehensive framework that addresses multimodal fusion tasks and that is suited to real-world ReID. Furthermore, despite the challenges posed by multimodal corrupted data, multimodal fusion demonstrates its benefits over unimodal state-of-the-art person ReID algorithms with moderate complexity increase. Looking ahead, it is crucial to deepen research on multimodal fusion due to its high potential in addressing current tasks and challenges and regarding the advent of accurate and accessible sensors.

Limitations.

Despite valuable contributions to the research community, a few limitations can be identified for this work.

- Implementing a multimodal approach involving two distinct modalities comes with supplementary challenges from the perspective of practical deployment. First, more channels (camera streams) must be processed, requiring more computational resources but also likely impacting the system's processing efficiency. Second, using the V and I modality in our case typically results in an embedding size twice that of a unimodal approach. In the context of a practical person ReID application, reference images must be processed as they come, and their embedding stored in a database, allowing for faster later matching. It is important to note that the adoption of a multimodal framework would consequently require doubling the storage capacity due to the larger embedding size, which might compromise this choice. However, one must notice that the multimodal setting can eventually allow for using lower complexity backbones while achieving equivalent or better performances than the unimodal approaches. As a consequence, doubling the embedding size may be avoidable, tackling this limitation.
- The proposed corrupted datasets allow for evaluating the models under highly challenging conditions, but the frequency of the used corruption does not accurately mirror real-world conditions. Indeed, the UCD and CCD datasets corrupt each and every visible-infrared data pair. Hence, while these datasets allow for a highly challenging evaluation and must be used for better model comprehension, real-world performance cannot be determined through them and would likely fall somewhere in between the evaluation of those models when using clean and corrupted data.

Future work.

This work and its conclusions suggest the following future work:

- As highlighted, attention mechanisms could be included in the framework, and especially model-level feature map refactoring. Indeed, this could allow for better handling of hard

image corruption. However, this must be done carefully to avoid losing the modality-specific streams' benefits.

- The proposed multimodal V-I corrupted datasets could be reused to evaluate state-of-art cross-modal person ReID models. Indeed, V and I modalities are likely similarly impacted by corruption under the multimodal and cross-modal settings. Such an approach would allow for reevaluating cross-modal models under more realistic scenarios, assessing their generalization power, and eventually uncovering robustness weaknesses.
- Considering each modality under the Learning Using Privileged Information (LUPI) paradigm might improve the overall process without additional complexity. LUPI approaches take advantage of privileged information (which can be a modality) available at training only to reinforce the main modality representation. Although the multimodal setting conserves each modality during inference, we believe that each modality stream could be reinforced due to the knowledge of the other modality during training following LUPI approaches. Then, even if the two modalities are, in our case, available at inference, we believe that such an approach can help the whole framework. For example, it can allow for more corruption robustness as each modality stream should be reinforced and able to mine the hetero-modality-related features from the main modality.
- Adapting the proposed model for a real-world deployment could require supplementary steps. First, the open-set scenario (Ye *et al.*, 2021) could be explored. Indeed, a practical scenario could not ensure that the query appears in the gallery, which is expected for the considered closed-set scenario. Also, lower complexity backbones could be investigated, like OsNet (Zhou *et al.*, 2021), for a more efficient approach. Furthermore, this could allow for adding new modalities to the approach and, consequently, more knowledge diversity, more model flexibility, and better ReID.

APPENDIX I

MULTIMODAL DATA AUGMENTATION FOR VISUAL-INFRARED PERSON REID WITH CORRUPTED DATA

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

[1] Laboratoire d'Imagerie, de Vision et d'Intelligence Artificielle (LIVIA), École de Technologie Supérieure,
1100 Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3

Article Published in « Winter Conference on Applications of Computer Vision (WACV) » 2023.

Abstract

The re-identification (ReID) of individuals over a complex network of cameras is a challenging task, especially under real-world surveillance conditions. Several deep learning models have been proposed for visible-infrared (V-I) person ReID to recognize individuals from images captured using V and I cameras. However, performance may decline considerably if V and I images captured at test time are corrupted (e.g., noise, blur, and weather conditions). Although various data augmentation (DA) methods have been explored to improve the generalization capacity, these are not adapted for V-I person ReID. In this paper, a specialized DA strategy is proposed to address this multimodal setting. Given both the V and I modalities, this strategy allows to diminish the impact of corruption on the accuracy of deep person ReID models. Corruption may be modality-specific, and an additional modality often provides complementary information. Our multimodal DA strategy is designed specifically to encourage modality collaboration and reinforce generalization capability. For instance, punctual masking of modalities forces the model to select the informative modality. Local DA is also explored for advanced selection of features within and among modalities. The impact of training baseline fusion models for V-I person ReID using the proposed multimodal DA strategy is assessed on corrupted versions of the SYSU-MM01, RegDB, and ThermalWORLD datasets in terms of complexity and efficiency. Results indicate that using our strategy provides V-I ReID models the ability to exploit both

shared and individual modality knowledge so they can outperform models trained with no or unimodal DA. GitHub code: https://github.com/art2611/ML-MDA.

1. Introduction

Real-world monitoring and surveillance application (e.g., individuals in airport, and vehicles in traffic) rely on challenging tasks, like object detection (Zou *et al.*, 2019; Zaidi *et al.*, 2022), tracking (Luo *et al.*, 2021), and re-identification (ReID) (Khan & Ullah, 2019; Ye *et al.*, 2021). The aim of person ReID is to recognize individuals over a set of distributed non-overlapping cameras. State-Of-Art (SOA) systems for person re-identification (e.g., deep Siamese networks) typically learn an embedding through various metric learning losses, which aim at making similar image pairs (with the same identity) closer to each other and dissimilar image pairs (with different identities) more distant from each other. Despite the recent advances with DL models, person ReID remains a challenging task due to the non-rigid structure of the human body, the different viewpoints/poses with which a person can be observed, image corruption, and the variability of capture conditions (e.g., illumination, scale, contrast) (Bhuiyan *et al.*, 2020; Mekhazni *et al.*, 2020).

Visible-infrared (V-I) person ReID aims to recognize individuals of interest across a network of V and I cameras. I cameras are often employed in conjunction with V cameras for, e.g., night time recognition in outdoor environments. Most approaches for V-I person ReID focus on the cross-modal matching problem. This paper focuses on person ReID systems that allow for fusion of visible and infrared modalities based on a joint representation space. Although several techniques have been proposed for dynamic and attention-based fusion (Ismail *et al.*, 2020; Su *et al.*, 2020), few V-I person ReID methods have been proposed for V-I fusion (Nguyen *et al.*, 2017). In this setting, it is difficult to extract discriminant modality-specific features when one modality becomes corrupted, while conserving the shared modality features (Baltrušaitis *et al.*, 2018).

In real-world surveillance applications, the accuracy of person ReID models often declines when image data is corrupted by noise, occlusions, saturation, blur, weather conditions, etc. (Chen *et al.*, 2021). Several strategies have been developed to improve the generalization performance of person ReID models in response to corrupted image data. Using more complex DL models, trained with more data have been shown to improve the performances in object detection (Michaelis *et al.*, 2019), and image classification (Xie *et al.*, 2020) tasks. For instance, using transformer-based models may be more suitable to tackle corruption (Hendrycks *et al.*, 2020; Chen *et al.*, 2021). However, using more complex models, like vision-transformers (Han *et al.*, 2020) limits real-time ReID applications. In addition, using more diverse training data can help (Xie *et al.*, 2020), and therefore data augmentation (DA) (Chen *et al.*, 2021) methods may improve performance, without increasing the models complexity, and while avoiding the costs of data collection and annotation (Shorten & Khoshgoftaar, 2019).

In this paper, we propose a MDA strategy to improve the accuracy of V-I person ReID systems. Chen *et al.* (2021) recently proposed a DA learning strategy, called the Consistent ID Loss, with Inference before BNNeck, and Local-based Augmentation (CIL). It is mainly based on local DA, and provides improvements in accuracy for unimodal (RGB) person ReID. However, the multimodal aspect has not been explored in the literature to tackle corruptions. Yet, such approach might be helpful to tackle corruption as modalities are not similarly affected by corruptions and can still benefit by DA strategies (Hao *et al.*, 2023).

To manage corrupted image data in multimodal settings, a multimodal DA (MDA) strategy is introduced, allowing to leverage the complementary knowledge among modalities, while dynamically balancing the importance of individual modality in the final predictions. Consequently, the strategy should reduce the corruption impact. Having in mind the multimodal person ReID aspect, and regarding that person ReID datasets were only used for cross-modal ReID, protocols are provided along with a comprehensive study over three V-I person ReID datasets, SYSU-MM01 (Wu *et al.*, 2017), RegDB (Nguyen *et al.*, 2017) and (less explored) ThermalWORLD (Kniaz *et al.*, 2018). Finally, as the focus is made on corruption robustness for

the multimodal setting, the corruption benchmark proposed by Chen *et al.* (2021) is extended to the infrared thermal modality.

Our main contributions are summarized as follows. (1) A MDA strategy is proposed to improve the accuracy of DL models for V-I person ReID. To optimize the collaboration among modalities, discriminant joint feature representations in the DL model, our MDA strategy relies on local occlusions and global modality masking data augmentation. (2) A comprehensive V-I multimodal experimental protocol is proposed to evaluate the impact on performance of clean and corrupted image data using the well-known SYSU-MM01, RegDB, and ThermalWORLD datasets. Corruptions from Chen *et al.* (2021) are extended to the infrared domain to analyse multimodal data corruption impact. (3) An extensive set of experiments is conducted, showing that the used V-I fusion model outperforms the related SOA models. The limitations of unimodal models are shown by comparing a basic fusion model learned with the adapted DA to the unimodal SOA person ReID models.

2. Related Work

2.1 Multimodal person ReID

Most approaches for person ReID in the last decade (Ye *et al.*, 2021) focus on the unimodal (RGB) (Ristani & Tomasi, 2018; Luo *et al.*, 2019a) and cross-modal (Hao *et al.*, 2021; Alehdaghi *et al.*, 2023; Zhang *et al.*, 2022) settings. Few focused on combining multimodal information, despite the potential to improve performance in the joint representation setting (Baltrušaitis *et al.*, 2018). For example, Chen *et al.* (2019a) extracted contours from the V modality and used a two-stream CNN architecture to combine information. Bhuiyan *et al.* (2020) proposed to use pose information to gate the flow of visual information through a CNN backbone. These approaches used the knowledge extracted from the main modality, which would be similarly affected by image corruption.

Some approaches sought to leverage the complementarity of V and depth modalities for an accurate person ReID (Paolanti *et al.*, 2018; Lejbolle *et al.*, 2018; Martini *et al.*, 2020). However,

Nguyen *et al.* (2017) represents the only approach where visible and infrared modalities are integrated into a joint representation space. Infrared and visual features are concatenated from embeddings extracted independently trained CNNs, and used for pairwise matching at test time. This simple model attained an impressive performances on the RegDB dataset. However, RegDB data is captured with only one camera per modality, and V-I cameras are co-located, with only a single tracklet of ten images per modality and individual. For these reasons, the RegDB dataset is less consistent with a real-world scenario. In fact, the development of person ReID models that are effective in uncontrolled real-world scenario remains an open problem (Hendrycks *et al.*, 2021).

2.2 Corruption and data augmentation strategies

Data augmentation (DA) consists in multiplying the available training dataset by punctually applying transformations on training images, like flips, rotations, and scaling (Ciregan *et al.*, 2012). This way, a model usually benefits from increased robustness to image variations, and improved generalization performance. According to Geirhos *et al.* (2018), training a model on a given corruption is not often helpful over other types of degradation. Yet, Rusak *et al.* (2020) showed that a well-tuned DA can help the model to perform well over multiple types of image corruption, through Gaussian and Speckle noise augmentation. Hendrycks *et al.* (2019) proposed the Augmix strategy , where various transformations are randomly applied to an image, and then mix multiple of those augmented images. Random Erasing punctually occludes parts of the images by replacing pixels with random values (Zhong *et al.*, 2020). Those strategies allow a large variety of augmented image, simulating eventually real-world data, and hence inducing higher generalization performance.

Focusing on person ReID, Chen *et al.* (2021) proposed both a corrupted V dataset (adapted from Hendrycks & Dietterich (2019)) and the CIL learning strategy to improve systems performance under corrupted data. Their strategy is partly based on two local DA methods – self-patch mixing and soft random erasing. The former replaces some of the pixels in a patch with random values, while the latter superposes a randomly selected patch from an image at a random position

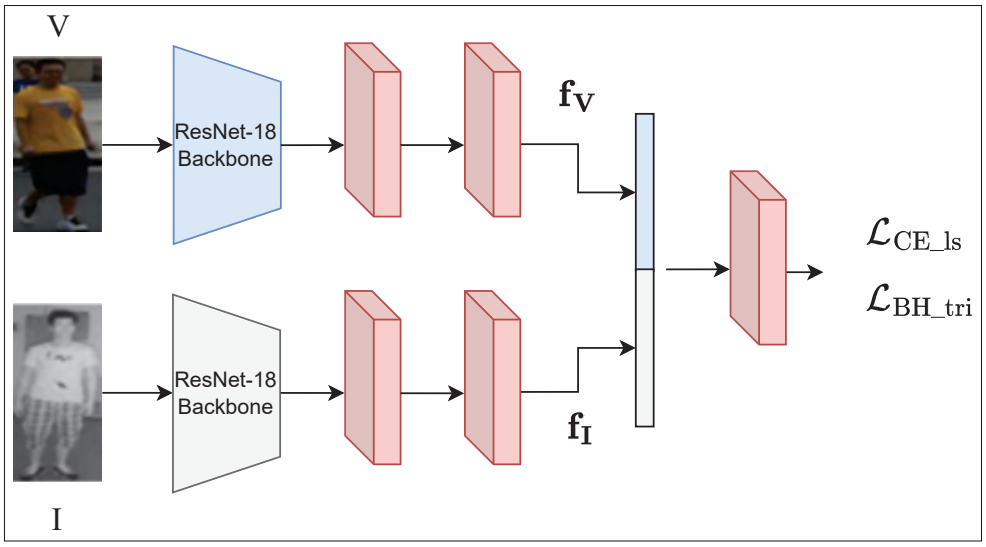

Figure-A I-1 Training architecture considered for V-I person ReID. It learns a joint multimodal representation by concatenating features produced by independent I and V ResNet-18 CNN backbones.

on this same image. Gong *et al.* (2021) show interesting improvements through local and global grayscale patch DA on V images. The previous strategies are limited to single modality stream models, even though the latter shows how greyscale data may reinforce the visible modality features using DA. MDA strategies have presented encouraging results for image-text emotion recognition (Xu *et al.*, 2020) or vision-language representation learning (Hao *et al.*, 2023). However, to our best knowledge, our work is first to propose MDA with V-I person ReID applications.

3. Proposed Strategies

Our strategy is based on co-learning, allowing each modality stream to adapt to the other (Baltrušaitis *et al.*, 2018). Using our MDA strategy, we expect to adapt DL models from one modality stream to another, and consequently provide better robustness to corrupted multimodal data. The low-cost multimodal architecture that we considered for V-I person ReID is based on two parallel ResNet-18 (He *et al.*, 2016) backbones pre-trained on ImageNet (Deng *et al.*, 2009). Rather than having a large single stream model, such architecture might allow us to

present a competitive model both in size and efficiency. After the two backbones, each stream has an average pooling and a batch normalization layer. The final prediction is obtained by concatenating features from each embedding, right before presenting it to a fully connected layer (Fig. I-1). Embeddings are concatenated during the test phase for pairwise similarity matching, from which the final ranking is obtained.

3.1 Multimodal patch mixing and soft random-erasing

Making a multimodal model focus on modality-specific features is challenging, as the model usually mainly focuses on shared features (Baltrušaitis *et al.*, 2018). Augmenting data with local occlusions may help the model to emphasize modality-specific feature importance, as some features will be available only from one or the other modality.

Multimodal soft Random Erasing (MS-REA): The soft random erasing (S-REA) (Chen *et al.*, 2021) might play this role, as it occludes parts of the V image punctually, potentially letting the opportunity for the hetero modality to close this occlusion gap. For S-REA, a proportion of the pixels in a given patch are given random values. To make the model close the occlusion gap in a bi-directional manner, the MS-REA is proposed (Fig. I-2), applying grayscaled random pixel values on a given path of the thermal modality, as well as the random values pixel values on the V modality. Grayscale values respect the infrared thermal image definition as I thermal is encoded on one channel, potentially aligning better with real-world corruptions.

Multimodal Patch Mixing (M-PATCH): Our M-PATCH DA inspired by the Self Patch (S-PATCH) DA (Chen *et al.*, 2021). Through M-PATCH, the idea is to extract a patch from each modality and superimpose it on the hetero-modality. The I modality receives the V patch from the same individual and vice versa. As the patches come from the same individual, the model has the option to rely on the patch features to discriminate. Three variants are explored which have different disturbance levels. From the less disturbing to the most disturbing, the first variation is extracting the patch from the Same part of the image, and applying it at the Same location on both modalities (-SS). The second extracts from the Same location but apply at Different locations

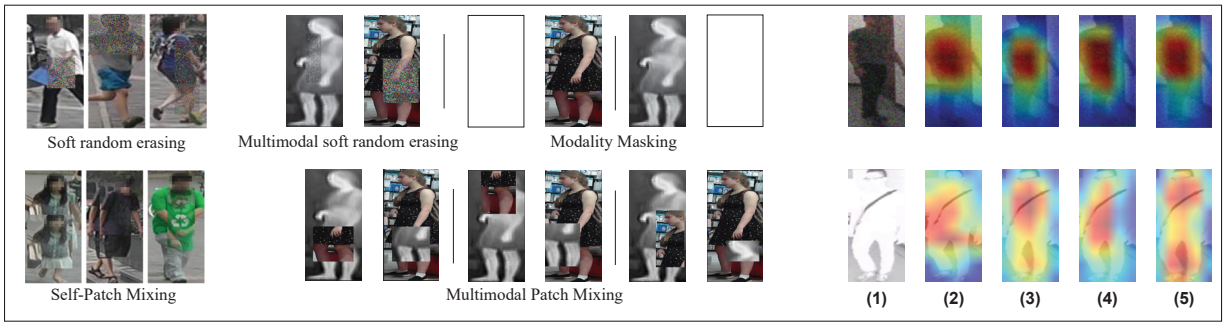

Figure-A I-2 **Left** present data augmentation methods from Chen *et al.* (2021). **Center** are our augmentation methods extensions from those, along with the proposed Modality Masking approach. **Right** shows visualizations of activation maps from corrupted a sample and from differently trained models. **(1)** Pair input data, V corrupted through Gaussian noise and I through saturation. **(2)** Augmix. **(3)** Multimodal Patch Mixing. **(4)** Modality Masking. **(5)** Multimodal soft random erasing. The discriminability increases from left to right.

(-SD), and the third extracts from Different locations and also apply at Different locations (-DD) (Fig. I-2). The M-PATCH approach might gather the best of both RandomPatch (Zhou *et al.*, 2019) and S-PATCH (Chen *et al.*, 2021) strategies. RandomPatch is strongly disturbing, and the model is forced to focus on out-of-patch features as the patch gathers information related to a different individual. S-PATCH less disturbing – it allows the model to focus on in-patch features as it contains features related to the same individual. Ours also allows in-patch feature selection by using the same individual, but provides more disturbance since the patch comes from the hetero modality. This approach may reinforce the model's shared features finding, while also pushing the model to exchange information across modalities.

3.2 Modality masking

A modality might be punctually unavailable or primarily uninformative. Though, the model has to know how to cancel a modality so that this one should not have a high impact on the final prediction. The modality masking approach is expected to make the model learns such behavior by punctually replacing one or another modality with an entirely blank image. Instead of masking the multimodal representation as it has been done by Gabeur *et al.* (2022), a representation is

extracted from the masked input, so the model has to learn how to cancel its influence on the final results. This masking DA is expected to complement the previously presented DA. The M-PATCH and MS-REA approaches supposedly focus on making the model better at selecting the right features within a modality. The idea is here to balance the importance of each modality in the final embedding regarding the level of corruption of each.

4. Results and Discussion

4.1 Datasets and performance measures

Since our study is focused on V-I multimodal person ReID, we employ the widely known SYSU-MM01 (SYSU) (Wu *et al.*, 2017) and RegDB (Nguyen *et al.*, 2017) datasets, along with the lesser-known ThermalWORLD (TWORLD) (Kniaz *et al.*, 2018) dataset. Details on these datasets are shown in Table I-1), allowing us to evaluate under diverse conditions.

Table-A I-1 Statistics of SYSU, RegDB, and TWORLD datasets. **V**: **V**isible and **I**: **I**nfrared. Image size and number per identity is presented as: Min;Max;Avg. BRISQUE (Mittal *et al.*, 2011) metric as: avg±std.

Statistic	**SYSU**	**RegDB**	**TWORLD**
V-images	29,033	4,120	8,125
I-images	15,712	4,120	8,125
V-Camera	4	1	16
I-Camera	2	1	Generated
Identities	491	412	409
Paired cameras	No	Yes	Yes
V-images/id	10;144;59.1	10;10;10	1;155;19.9
I-images/id	10;144;32.0	10;10;10	1;155;19.9
Image width	26;1198;111	64;64;64	10;810;141
Image height	65;879;291	128;128;128	25;897;353
V-BRISQUE	30.50±12.26	38.84±9.86	27.79±13.28
I-BRISQUE	40.52±8.42	38.81±9.56	60.25±8.67

SYSU-MM01. (Wu *et al.*, 2017) gather 4 V and 2 I cameras, with 491 distinct individuals, 29033 V, and 15712 I thermal images. The specificity of this dataset is that its V and I cameras are not co-located.

RegDB. (Nguyen *et al.*, 2017) is a much smaller dataset, with one camera only per modality, co-located cameras, and a single 10 images tracklet per identity and camera. RegDB 410 identities lead to 4120 images per modality.

ThermalWORLD. (Kniaz *et al.*, 2018) is only partially available, leading us to 409 distinct identities and 8125 V images from 16 cameras. I images were generated synthetically. Hence, cameras can be considered as co-located. However, the thermal images are of poor quality (see BRISQUE (Mittal *et al.*, 2011) value of 60.25 in Table I-1).

Corruptions. For comparison reasons, the corruptions used by Chen *et al.* (2021) are the same in this study. However, the V corruptions were adapted to the thermal modality (detailed in supplementary material) as the thermal modality would more likely get impacted in a real scenario. The V data corruptions proposed by Chen *et al.* (2021) are mentioned through the notation **-C**, and its extension with both modalities corrupted through the notation **-C***. Corruptions are applied independently and randomly for the V and the I modalities and on both the query and the gallery images to match real-world conditions.

Performance Measures. The mean Average Precision (mAP), and the mean Inverse Penalty (mINP) are used as performance metrics, commonly used for person ReID (Ye *et al.*, 2021).

4.2 Implementation details

Data division. SYSU-MM01 and RegDB datasets have well-established V-I cross-modal protocols (Wang *et al.*, 2019a,a,b; Ye *et al.*, 2019), but multimodal protocols remain to be built. Following the SYSU-MM01 authors' cross-modal protocol, 395 identities were used for the training set, and 96 identities were used for the testing set. For RegDB, the 412 identities are kept as well into the two identical sets of 206 individuals. The SYSU-MM01 train/test ratio is kept

for ThermalWORLD, leading to 325 training identities and 84 for testing. A 5-fold validation (Raschka, 2018) is performed over the data used for training, using folds of respectively 79, 41, and 65 distinct identities for SYSU-MM01, RegDB, and ThermalWORLD.

Data augmentation (DA). The Augmix, S-PATCH, or S-REA were evaluated following the original papers settings. Our proposed multimodal extensions M-PATCH and MS-REA were used with the same appearance augmentation probability as S-PATCH and S-REA. Modality Masking is applied randomly on one or another modality, with equiprobability, and occurs with a default probability of $1/8$. For RegDB, the validation set uses the same DA as the training set. This way, better performances were observed, since they maxed out in the early epochs, or otherwise do not learn complex cues for the model.

Pre-processing. A data normalization is done at first by re-scaling RBG and I images to 144×288. Random cropping with zero padding and horizontal flips are adopted for base DA. Those parameters were proposed by Ye *et al.* (2021) on RegDB and SYSU-MM01 datasets. The same normalization is kept under ThermalWORLD for consistency among protocols.

Table-A I-2 The performance of various multimodal DA strategies using a standard model (V-I ReID model trained without DA) as the baseline. Augmix DA is applied with and without other proposed DA approaches.

	SYSU		**SYSU-C***		**RegDB**		**RegDB-C***		**TWORLD**		**TWORLD-C***	
DA Strategy	mAP	mINP	mAP	mINP	mAP	mINP	mAP	mINP	mAP	mINP	mAP	mINP
Standard	96.47	73.69	25.01	1.90	99.64	98.46	21.80	2.40	87.90	49.05	29.30	3.93
Augmix	95.37	68.60	35.23	2.56	99.88	99.40	40.75	9.10	87.12	46.33	42.26	5.69
+ S-REA	96.21	74.36	43.24	4.06	99.90	99.51	43.84	10.25	89.24	50.10	54.14	8.92
+ MS-REA	**96.81**	**77.02**	**61.44**	**8.34**	99.86	99.35	**57.84**	**19.38**	88.95	49.92	**58.10**	**9.89**
+ S-PATCH	96.40	74.89	31.39	2.14	99.90	99.53	41.83	9.39	89.12	50.53	40.73	5.63
+ MS-PATCH	94.70	69.10	33.69	2.17	99.89	99.41	40.97	9.34	**89.26**	**51.26**	41.75	5.57
+ M-PATCH-SS	96.10	73.40	35.49	2.44	99.86	99.34	43.28	10.68	88.35	50.16	44.41	5.61
+ M-PATCH-SD	95.94	72.93	35.10	2.40	99.87	99.35	42.95	10.31	88.58	51.59	43.49	5.53
+ M-PATCH-DD	94.98	68.95	33.90	2.42	99.89	99.48	41.98	9.71	88.49	51.35	43.90	5.51
+ Masking	95.61	73.49	40.92	2.90	**99.90**	**99.52**	49.27	12.10	86.01	42.76	39.91	6.16

Hyperparameters. The hyperparameters values in our models were set based on the default AGW (Ye *et al.*, 2021) baseline. The SGD is used for training optimization, combined with a Nesterov momentum of 0.9 and a weight decay of $5e-4$. Our models are trained through 100 epochs. Early stopping is applied based on validation mAP performances. The learning rate is initialized at 0.1 and follows a warming-up strategy (Luo *et al.*, 2019b). The batch size is set to 32, with 8 distinct individuals and 4 images per individual. The paired image is selected by default for RegDB and ThermalWORLD. For the SYSU-MM01 dataset, the images from the hetero modality are randomly selected through the available ones for a given identity.

Losses. The Batch Hard triplet loss (Hermans *et al.*, 2017) $\mathcal{L}_{\text{BH\_tri}}$ and the cross-entropy with regularization via Label smoothing (Szegedy *et al.*, 2016) $\mathcal{L}_{\text{CE\_ls}}$ are used as loss functions for our models. Indeed, the former is widely used in person ReID approaches (Wang *et al.*, 2019a; Choi *et al.*, 2020; Ye *et al.*, 2021), so the same margin value is fixed at 0.3, and the latter is part of the CIL implementation (Chen *et al.*, 2021). The total loss corresponds to the sum of both losses. The batch hard triplet loss aims at reducing the distance in the embedding space for the hardest positives while increasing the distance for the hardest negatives. The regularization with label smoothing works at reducing the gap between logits, which makes the model less confident in predictions and hence improves generalization (Müller *et al.*, 2019).

Leave-one-out query strategy. The single-shot and the multi-shot settings (Wang *et al.*, 2019a) are widely used in cross-modal papers to form the query and gallery sets. For these settings, one or ten images from the hetero modality are selected per identity and camera to join the gallery, while the other modality forms the probe set. However, such an approach is not so realistic in a surveillance context, as the video makes the gallery number of frames per person vary much. These variations cannot be controlled as individuals are unknown in the final environment. Hence, a new strategy is developed, inspired by the leave-one-out cross-validation strategy (Raschka, 2018), named Leave-One-Out Query (LOOQ). The LOOQ strategy treats the extreme but meaningful case in which one would have only a unique image of the person to ReID and multiple footages containing images of this same person in the gallery. Every pair of images is alternatively used as a probe set while all the other pairs join the gallery. This allows us

to respect the original dataset statistics (see Table I-1) by authorising the gallery images per individual to vary. Also, the mINP metric relates to the hardest test sample from the same individual. Hence, computing this metric over multiple gallery images makes it more consistent, appearing even more important in a corrupted context.

Concerning the implementation, the images are paired for both RegDB and ThermalWORLD datasets, so the paired image from the hetero modality joins the query and gallery set directly during the formation of those sets. However, SYSU-MM01 needs personal treatment since its images are not paired. Plus, the image number per modality for a given individual varies (Table I-1). To solve this issue, as many pairs of images as possible are randomly selected with the constraint that one image from one modality or another must not appear in two distinct pairs. Because random image pairs are formed for SYSU-MM01, a mean of 30 trials is performed to present robust and reliable results according to the Central Limit Theorem.

4.3 Benchmarking data augmentation strategies

Table I-2 shows the impact on person ReID performance of each DA strategy is investigated over the three datasets under clean and corrupted (-C*) settings. First, we compare the model learned without DA (Standard) with the model learned with Augmix, and the models learned with Augmix plus other augmentation. The other DA strategies can be S-REA, S-PATCH, or one of our proposed augmentation.

4.3.1 Multimodal soft random erasing

The S-REA strategy applies random values to a certain proportion of the pixels in a given patch of the V image. A good improvement can be seen from the Augmix to the S-REA strategy for each dataset and the clean and corrupted settings. Still, a more significant improvement happened for ThermalWORLD-C* compared to SYSU-MM01-C* and RegDB-C*, respectively, with a 11.88% improvement against 8.01% and 3.09%. While extending the DA to the multimodal setting through MS-REA, we observe a remarkable improvement for each corrupted setting, and especially that the improvement is much higher on both SYSU-MM01 and RegDB compared to

ThermalWORLD. Indeed, mAP increases by 18.20% and 14.00% for SYSU-MM01-C* and RegDB-C* respectively against 3.96% for ThermalWORLD-C*. ThermalWORLD has a much weaker I modality, so the model probably focuses much on the visible modality. Consequently, the model probably almost fully benefits from S-REA as if it were a unimodal architecture. The other datasets do not allow to benefit as much from this DA, as the model has presumably learned to focus more on I due to the unbalanced augmentation (applied only on V). In contrast, the equilibrium brought by MS-REA probably allows the full exploitation of the approach and explains the impressive improvement from S-REA to MS-REA. Also, MS-REA comes first among approaches under the clean setting for SYSU-MM01 and RegDB datasets, except for ThermalWORLD. With a 95% confidence, results using MS-REA compared to the best approach are not statistically significant for RegDB, whereas it is for ThermalWORLD according to the Cochran p-values (Raschka, 2018) of respectively 0.29 and $4.89e-5$. Thanks to MS-REA and partial occlusions, the model might have learned not to only focus on the most discriminant cues, as confirmed by the I activation map comparison from Augmix to MS-REA (see Fig. I-2). Also, this approach presents important improvement over biased data augmentation, denoting a great generalization power (detailed in supplementary material).

Table-A I-3 Data augmentation combination. Each is used along with Augmix and MS-REA. C1 stands for Masking, C2 for M-PATCH-SS and C3 for Masking and M-PATCH-SS.

Strategy	**SYSU**		**SYSU-C***		**RegDB**		**RegDB-C***		**TWORLD**		**TWORLD-C***	
	mAP	mINP	mAP	mINP	mAP	mINP	mAP	mINP	mAP	mINP	mAP	mINP
MS-REA	**96,81**	**77,02**	61,44	8,34	99.86	99.35	57.84	19.3	88,95	49,92	58,10	9,89
MS-REA + C1	96.77	76.01	63.01	9.59	99.90	99.45	**61.92**	**20.14**	86.34	43.24	56.10	11.04
MS-REA + C2	96.85	75.87	61.19	9.13	99.85	99.26	56.23	17.98	**89.16**	**50.68**	57.45	9.64
MS-REA + C3	96.78	75.87	**63.83**	**9.77**	**99.89**	**99.48**	61.53	20.17	86.65	43.75	**57.95**	**11.53**

4.3.2 Multimodal patch mixing

Observing the results obtained for SYSU-MM01 and ThermalWORLD, the performances globally improved from the Augmix strategy to the S-PATCH approach for the clean datasets,

while those are reduced under the corrupted setting. While applying the self patch mixing on both modalities through MS-PATCH, performances are questionable, as performances remains lower or equivalent to Augmix on corrupted data, while conserving or decreasing from clean S-PATCH results. In practice, it is only while considering the modality patch exchange in our M-PATCH strategy, especially the less disturbing version M-PATCH-SS, that the best improvement is obtained on the corrupted setting, while conserving great performances on the clean one. Indeed, mAP is respectively improved by 2.15% and 2.53% over the Augmix strategy for ThermalWORLD-C* and RegDB-C*. The cameras might need to be co-located for the approach to perform, as SYSU-MM01-C* pretty much conserve similar performances as Augmix on corrupted data, and as the standard model on clean data. Spatial alignment is probably helping much the model to find correlations between the hetero modality patch and the current modality image. Still, there is a performance improvement on two datasets even if this one remains much lower than the previous MS-REA approach.

4.3.3 Masking

The modality masking approach presents interesting improvements under the SYSU-MM01 and RegDB datasets. Indeed, performances on corrupted datasets are increased by 5.69% mAP and 8.52% mAP over the Augmix approach, while those are pretty much matching Augmix performances on the clean datasets. The modality masking DA consists of punctually feeding a modality stream with a fully uninformative modality. Hence, those results show that the approach can make the model better able to give more importance to the discriminant modality for a given pair of images. The ability to balance modality influence on not co-located cameras through SYSU-MM01 dataset is important to highlight. Concerning the ThermalWORLD dataset, the Masking model's performances decrease for the clean and the corrupted setting compared to Augmix. Indeed, the mAP is respectively lower by 1.11% and 2.35%. Such a decrease is not surprising, as this dataset's thermal modality is very uninformative. Hence, while learning, a masked visible modality probably acts as noise by creating not discriminant V-I pairs.

4.3.4 Combination

As the DA approaches have distinct expected roles in the way they help the model to get more robust against corruption, combining them might allow to benefit from each of their specificities (Table I-3). It is interesting to see that the real combining improvement comes from the Masking approach used with MS-REA (C1) on both SYSU-MM01-C* and RegDB-C*, with respectively 1.57% and 4.08% improvement over MS-REA used by itself. ThermalWORLD did not benefit from the masking DA, which could be expected as the Masking was already decreasing its performance when used alone. Adding M-PATCH to MS-REA (C2) or to MS-REA and Masking (C3) seems not to bring meaningful additional improvements. Indeed, MS-REA + (C3) matches the performances of MS-REA + (C1) under the clean and corrupted settings on both RegDB and SYSU-MM01. Similar observations can be done from MS-REA alone and MS-REA + M-PATCH. Hence, even if M-PATCH has shown improvements on RegDB and ThermalWORLD when used alone, those improvements are probably mainly due to the benefits of occlusions, which are already part of the MS-REA approach. Visual results observing especially I activation maps seem to confirm this aspect (Fig. I-2). Though, using MS-REA with M-PATCH appear as not being meaningful. From the previous conclusions, we propose the Masking and Local Multimodal Data Augmentation (ML-MDA) strategy, which combines both the local approach MS-REA with the modality masking DA.

4.4 Comparison with the state-of-art

4.4.1 Performance

As there is no true competitor in the area of V-I multimodal person ReID, the ML-MDA strategy is compared with SOTA unimodal person ReID models, with or without the CIL strategy used. According to results obtained in Chen *et al.* (2021), the LightMBN (Herzog *et al.*, 2021) and TransReID (He *et al.*, 2021) models are respectively the most performing unimodal models under the clean and corrupted scenarios. For fair comparison, Table I-4 shows two scenarios for the multimodal test data. First, both V and I are corrupted (-C*), and second, to observe how

Table-A I-4 Performance of our multimodal model using ML-MDA compared against SOTA unimodal person ReID models, and a ResNet-18 unimodal model while using CIL or not. The two last rows show the performance of the same model when V is corrupted (-C), and when V and I are corrupted (-C*). Note that performance on clean datasets are the same and presented in fused cells.

Model	**SYSU**		**SYSU-C**		**RegDB**		**RegDB-C**		**TWORLD**		**TWORLD-C**	
	mAP	mINP	mAP	mINP	mAP	mINP	mAP	mINP	mAP	mINP	mAP	mINP
ResNet-18	86,25	39,97	32,36	1,91	99,26	96,64	45,15	5,68	86.44	49.44	28.06	3.86
TransReID	94,33	64,79	52,03	3,60	99,34	97,35	45,64	5,69	**95.86**	**77.98**	65.47	17.20
LightMBN	94,45	64,06	40,90	2,13	99,90	99,41	32,40	3,25	93.02	65.94	37.34	5.60
ResNet-18 + CIL	86,64	42,78	51,64	3,83	99,65	98,41	55,76	10,98	86.95	48.07	52.85	7.97
TransReID + CIL	93,20	62,02	61,38	7,20	99,69	98,57	58,74	12,89	94.79	73.82	**73.61**	**23.16**
LightMBN+CIL	94,07	61,95	67,80	8,23	99,89	99,41	66,55	21,53	93.20	66.14	71.30	19.73
Ours+ML-MDA (-C)	**96.77**	**76.01**	**87.89**	**42.5**	**99.90**	**99.45**	**92.37**	**75.71**	86.34	43.24	69.20	18.47
Ours+ML-MDA (-C*)			63.01	9.59			61.92	20.14			56.10	11.04

a clean I modality can help when the V modality is corrupted, performance is also compared when only V is corrupted (-C).

Considering a clean data setting, the ML-MDA model outperforms the second-best approach by 2.32% mAP and especially by 11.22% mINP on SYSU-MM01. This significant mINP improvement shows that the multimodal setting helps considerably on the more challenging images. Indeed, the multimodal model can compensate for challenging V samples with the I modality. On the RegDB dataset, our approach outperforms the others, with a statistically significant improvement. Indeed, with a 95% confidence interval, the Cochran (Raschka, 2018) p-value between LightMBN, LightMBN + CIL and our approach is of 0.02. On ThermalWORLD, the performance of the multimodal model cannot compare with TransReID and LightMBN models, not even improving over the ResNet-18 model. Again, the poor quality I mostly acts as a source of perturbation for the model.

When the V modality only is corrupted (-C) on both SYSU-MM01 and RegDB datasets, the ML-MDA model provides a considerable performance improvement over TransReID and

LightMBN models. Indeed, our model reaches 87.98% mAP and 92.37% mAP for SYSU-MM01 and RegDB, respectively improving by 20.09% and 25.82% over the second-best approach. These improvements highlight the benefits of a well-trained multimodal model, relying mainly on the clean modality (I) when the other is corrupted (V). A performance gap between -C to -C* settings can be observed, as -C* is much more challenging with two corrupted modalities. The multimodal model appears as the second-best approach for both SYSU-MM01-C* and RegDB-C* datasets. Indeed, LightMBN reaches respectively 67.80% and 66.55% mAP against 63.01% and 61.92% mAP for our multimodal model. Still, the multimodal setting improves the mINP for SYSU-MM01, from 8.23% to 9.59%, and is only below RegDB by 1.39% mINP, showing that the multimodal setting can help on the hardest cases. For the -C* setting, our approach outperforms other models except for LightMBN + CIL, or on ThermalWORLD data, apparently unable to encode discriminant cues from the corrupted I to counterbalance the well designed unimodal models. However, our architecture remains very simple, and obtaining such performance improvement on our light architecture is already promising. More complex fusion strategies, with more knowledge exchange between modality streams, and a more robust backbones like ResNet-50, may allow exceeding the performance of LightMBN.

Note that both the -C and the -C* settings might not be the most accurate for a great multimodal evaluation. Indeed, considering the I always clean (-C) is not so accurate as weather would, for example, probably happen on both modalities for a given co-located pair. On the other hand, considering both modalities always corrupted (-C*) hardly allows the hetero modality to help the primary modality, but is not so realistic either. Indeed, digital corruption or noise would probably not affect V-I modalities simultaneously. In fact, the real-world setting would allow [Clean V, Corrupted I] pairs, and would especially be a mixture of -C and -C* settings. In practice, there should be more pairs in which one of the two modalities remains clean, so the true potential of the multimodal setting probably lies somewhere in between the -C and the -C* settings.

Table-A I-5 Memory (number of parameters) and time (FLOPs) complexity of proposed and baseline ReID models, FLOPs computed from a single or multi-modal input.

Model	No. Params (M)	FLOPs (G)
ResNet-18	11.3	0.51
TransReID	102.0	19.55
LightMBN	7.6	2.09
Ours	22.5	1.54

4.4.2 Complexity

Multimodal person ReID with I and V is more complex than regular ReID with V, so models are compared in terms of number of parameters and FLOPs (Table I-5). The TransReID (He *et al.*, 2021) model is known for being computationally expensive as its architecture is transformer based, with a total of 102M parameters and 19.55 G-FLOPs. In contrast, LightMBN (Herzog *et al.*, 2021) is based on the Os-Net architecture, which makes it very light, requires 7.6M parameters and 2.09G FLOPs. Even if our multimodal model has more parameters (22.5M) to adjust than LightMBN, it requires less memory compared to the SOTA unimodal person ReID models, with its 1.54G FLOPs. Although our model seems equivalent to LightMBN in terms of complexity, it provides a significant performance improvement. Its robustness to corrupted data makes it an excellent trade-off in the face of uncontrollable scenarios.

5. Conclusion

Real-world surveillance often requires light models that perform well on corrupted data. In this paper, image corruptions were extended to the infrared modality, and MDA strategy was proposed to improve the performance of the V-I person ReID. Experiments on the SYSU-MM01, RegDB and ThermalWORLD datasets showed the benefits of the multimodal setting over SOTA unimodal ReID models, especially when combined with the specialized MDA strategy. Indeed, our ML-MDA strategy has allowed for significant improvements in terms of robustness to

corruption using the proposed modality masking and MS-REA MDA. The former learns the model to dynamically balance the importance of each modality in the final embedding. The latter works on the occlusion concept and teaches the model to better select features among modalities and not to focus only on the most discriminant features. ML-MDA improves performances, yet does not incur additional model complexity, and allows for a light ReID architecture.

Given multiple modalities, MDA allows addressing image data corruption, as these corruptions impact V and I modalities in a different ways, allowing the hetero-modality to compensate. MDA could be studied more independently from person ReID, and our methods can be applied to more general datasets (e.g., RGB-D data). Moreover, increasing the number modalities could further reduce the impact of corruption.

Note that potential improvements are possible using more advanced fusion methods (Su *et al.*, 2020; Ismail *et al.*, 2020). Finally, we believe that our multimodal corrupted test set might not entirely reflect the true potential of the multimodal setting, as discussed section 4.4. To better fit real-world conditions, corruption correlations among modalities should be considered in the test set design. This would probably allow the multimodal setting to perform even better.

APPENDIX II

HYBRID MODELS

1. Introduction

Computer vision approaches, when targeting tasks like classification or person ReID under the supervised setting, are mainly governed by two model types of architectures, the transformers (Dosovitskiy *et al.*, 2020) and the CNNs (LeCun *et al.*, 1998). Facing real-world data, transformer architectures have been highlighted as being more robust than CNN state-of-art models on several occasions (Hendrycks *et al.*, 2020; Chen *et al.*, 2021). However, to build a strong multimodal framework that remains lightweight for a person ReID approach, transformers' backbones may first be seen as too complex (Han *et al.*, 2020). Allowing benefiting from each model architecture while being less complex than transformers only model, the use of hybrid architecture (Dosovitskiy *et al.*, 2020) is an interesting exploration path that has to be investigated.

As mentioned in the literature review Section 2.1, different multimodal hybrid approaches were designed through models like the transfuser (Prakash *et al.*, 2021) or approaches like the one proposed by Dai *et al.* (2021). The transfuser model is promising and will be explored in our task, as its architecture has been proposed for autonomous vehicle driving, requiring efficient data processing. In this model, two specific CNN streams encode the knowledge along with in-between streams transformer self-attention modules. The transformer modules allow for selecting and refactoring the relevant features with the knowledge of the two modality representations and without the local limitations of the convolution operations. Unlike the transfuser model, the model from Dai *et al.* (2021) first processes the data through a CNN architecture, extracting low-level features passed through a transformer self-attention module to exploit long-range spatial dependencies among features. However, the different modalities are processed through the same layers in the whole model, which might lead to a lack of modality-specific feature exploitation, shown as essential through our MMSF model. Still, using

a CNN followed by transformer self-attention modules might be an interesting way to process data.

Based on previous observations, one can follow a similar strategy while adapting it to our task. In practice, we have seen the importance of considering modality-specific backbones, allowing for encoding the modality-specific features. Hence, instead of having a single backbone like Dai *et al.* (2021), modality-specific backbones can be used and followed by modality-specific transformer self-attention layers and a fusion of the final scores. Plus, CNN models such as MMTM or MSAF can be used as well as backbones before using the transformer self-attention layers, which would likely make the self-attention more beneficial as they could allow a better feature selection among the refactored features. Selecting the relevant features with the knowledge of the hetero-modality, the cross-attention transformer (Tsai *et al.*, 2019; Wei *et al.*, 2020b; Sun *et al.*, 2021) could be used instead of the MSA operations in those same architectures, the cross attention simply replacing queries from the main modality by queries from the hetero modality in the MSA equation (Eq. 1.7). This aspect leads us to three distinct models to explore on our task: The transfuser model, and models with distinct modality backbones followed by self or cross attention transformer layers. These models will be detailed and experimentally explored in the next sections.

2. Models

2.1 Transfuser

The transfuser model has been developed for autonomous driving tasks by Prakash *et al.* (2021). In the original paper, the model extracts the features from two modalities using a ResNet-34 for the visible data and a ResNet-18 for the lidar data. In between the two backbones and right after each convolutional block, a transformer self-attention module is used to refine each feature representation based on their global shared representation. The model architecture is slightly adapted in our work, by using two ResNet-18 backbones and adapting the model to the specific

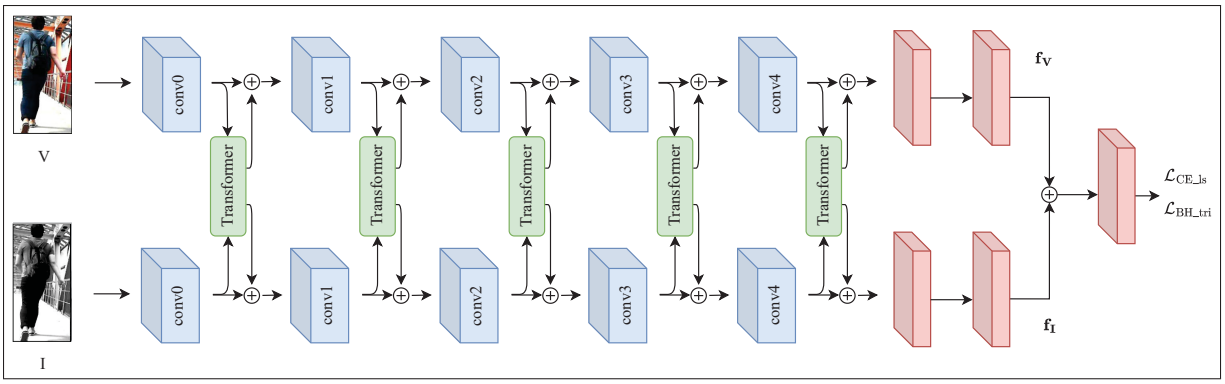

Figure-A II-1 Block diagram for the Transfuser model adapted to the person ReID task.

person ReID task (Fig. II-1). Indeed, the model requires class scores for learning and feature embeddings for evaluation instead of vehicle waypoints.

Before feeding the feature maps $\mathbf{F}_{\mathbf{V}}^{\ell}$ and $\mathbf{F}_{\mathbf{I}}^{\ell}$ to a transformer block, their size is down-sampled through global average pooling at a fixed resolution $H = W = 8$. The original paper proposes this resolution to reduce the model's complexity. Then, the downsampled features from each modality are stacked together, leading to a sequence of dimension $(2 \times H \times W) \times C$, $C \in \mathbb{N}$ being the original feature maps channel size. To this sequence is added a positional embedding, therefore given to the transformer. One can notice that there is no need for a classification token since the transformer module is used to refactor the feature maps and not to provide a discriminant feature vector for later classification (Dosovitskiy *et al.*, 2020). The output sequence is reshaped into 2 feature maps of dimension $H \times W \times C$, which get upsampled to the original feature resolution. Finally, each refactored feature map gets summed to the original feature map before being fed to the subsequent CNN layers.

2.2 Multimodal self-attention and cross-attention transformers

The model architectures proposed in this section are based on modality-specific ResNet-18, later followed by $L_s \in \mathbb{N}$ transformer self-attention or $L_c \in \mathbb{N}$ cross-attention modules, this value being referred as the depth of the model. In practice, the used CNN backbones can eventually consider interactions between each modality stream, through the use of architectures

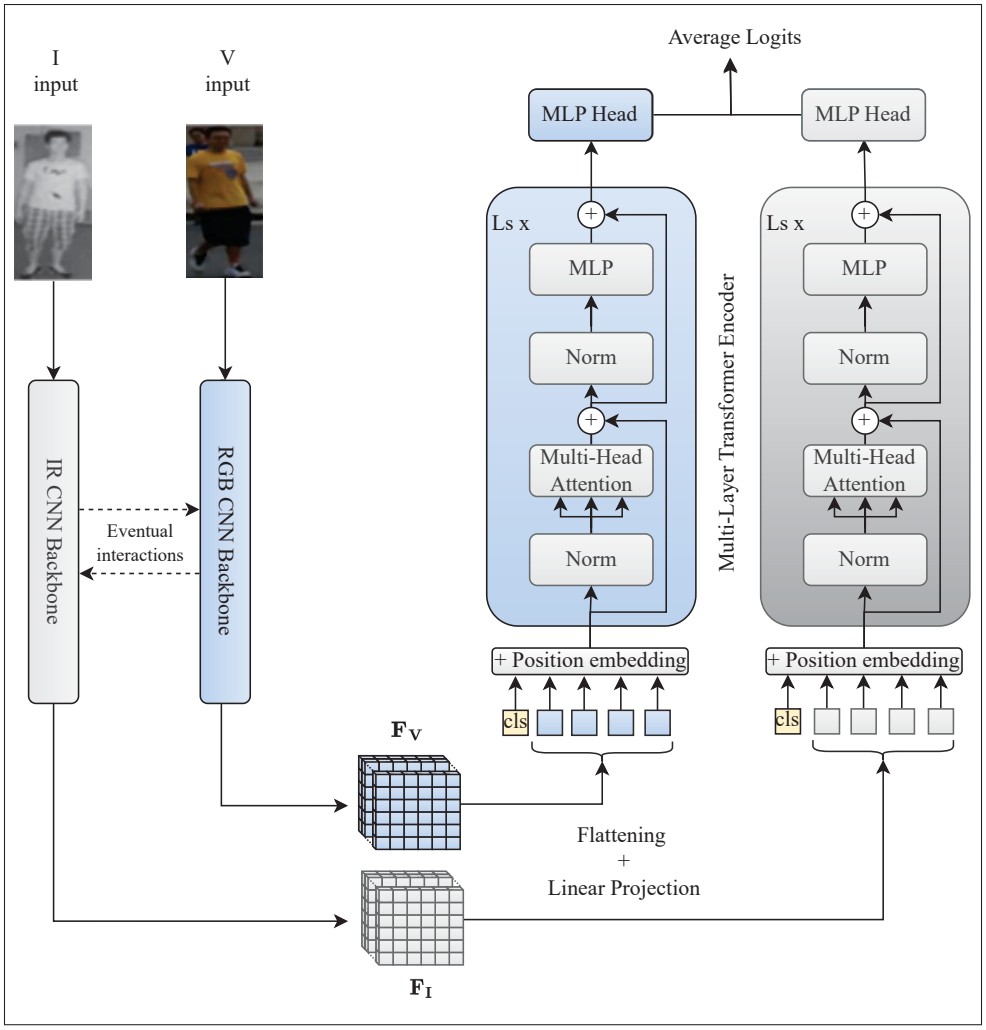

Figure-A II-2 Multimodal Self Attention Transformer

like MMTM or MSAF. Based on the self-attention transformer, the multimodal self-attention transformer (MSAT) model is represented Fig. II-2. Using cross-attention transformer instead, the multimodal cross-attention transformer (MCAT) is presented Fig. II-3.

Before getting through the transformer modules, the last feature map representations get split into tokens, to which a position embedding is added and a classification token is concatenated. Then, the two transformers' output representations (classification tokens) are individually passed through modality-specific MLPs for final classification, from which the scores get averaged while learning. For inference matching, the classification tokens are concatenated.

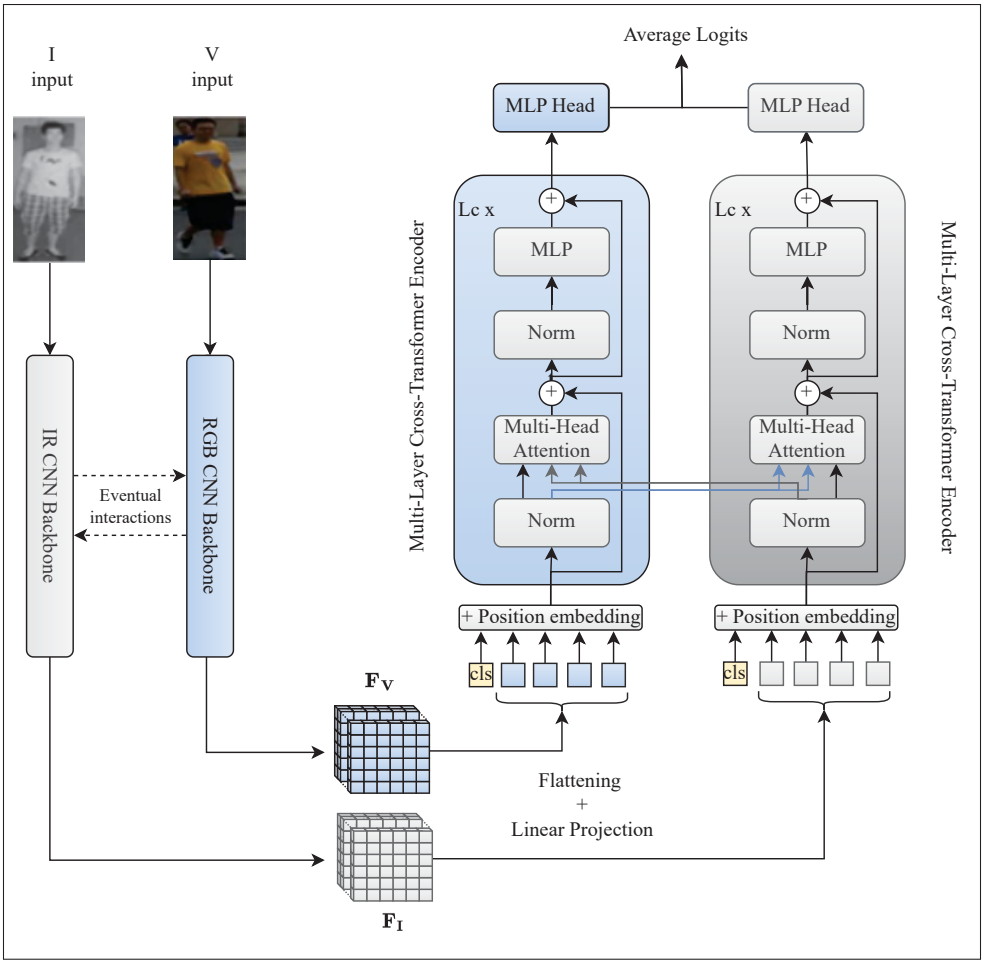

Figure-A II-3 Multimodal Self Attention Transformer

3. Results

The experimental methodology followed in this section is the same as the one followed in Chapter 3. First, except for the transfuser model, which already has a fixed architecture from the version proposed by Prakash *et al.* (2021), the two other hybrid models get optimized through distinct parameters. The best versions of these two models get compared to the transfuser and to our best multimodal models from Chapter 3.

3.1 Model selection

Table-A II-1 Hybrid models accuracy on the RegDB clean and CCD dataset regarding the used CNN backbone, pretraining (P), freezing (F), and the transformer's depth L. Separated CNN stands for modality-specific ResNet-18 backbones with no interactions between the two streams.

Model	CNN	P	F	L	RegDB		RegDB-CCD	
					mAP	mINP	mAP	mINP
MSAT	Separated	Yes	No	1	**99.77±0.04**	**99.00±0.25**	59.83±1.19	17.43±1.11
	Separated	Yes	No	2	99.75±0.06	99.02±0.22	59.90±1.12	17.41±0.93
	Separated	Yes	No	3	99.71±0.06	98.82±0.21	59.33±0.80	17.09±1.03
	MMTM	Yes	No	1	99.58±0.15	98.38±0.33	63.10±1.85	24.59±2.40
	MMTM	Yes	No	2	99.48±0.21	98.19±0.37	62.16±1.00	23.40±1.21
	MMTM	Yes	No	3	99.66±0.11	98.56±0.37	**63.41±1.08**	**25.22±1.27**
	MMTM	Yes	Yes	1	99.59±0.06	98.33±0.21	62.68±2.67	24.66±2.76
	MMTM	Yes	Yes	2	99.58±0.17	98.33±0.49	62.90±1.06	24.40±1.41
	MMTM	Yes	Yes	3	99.65±0.10	98.48±0.38	62.38±1.24	23.87±1.46
	MMTM	No	No	1	99.59±0.08	98.26±0.38	57.01±1.63	18.04±1.40
	MMTM	No	No	2	99.49±0.09	98.07±0.25	56.61±1.01	17.54±1.00
	MMTM	No	No	3	99.61±0.13	98.37±0.44	56.15±0.33	17.10±0.99
MCAT	Separated	Yes	No	1	**99.74±0.05**	**98.96±0.19**	60.68±1.02	17.56±0.71
	Separated	Yes	No	2	99.64±0.08	98.49±0.26	59.80±1.70	19.40±1.70
	Separated	Yes	No	3	99.68±0.07	98.57±0.15	58.72±2.42	18.83±1.70
	MMTM	Yes	No	1	99.56±0.07	98.26±0.38	**64.04±1.46**	**25.61±1.60**
	MMTM	Yes	No	2	99.58±0.14	98.21±0.55	63.03±0.74	25.21±0.36
	MMTM	Yes	No	3	99.60±0.11	98.22±0.38	62.87±0.75	25.06±0.75
	MMTM	Yes	Yes	1	99.54±0.14	98.22±0.35	63.63±1.99	25.49±2.54
	MMTM	Yes	Yes	2	99.60±0.11	98.33±0.31	62.71±1.81	24.18±0.90
	MMTM	Yes	Yes	3	99.50±0.09	98.02±0.36	62.69±1.09	23.96±1.51
	MMTM	No	No	1	99.69±0.05	98.57±0.17	57.42±2.69	18.28±2.57
	MMTM	No	No	2	99.42±0.16	97.89±0.44	57.61±1.29	18.11±1.31
	MMTM	No	No	3	99.43±0.13	97.95±0.36	57.67±1.28	19.03±1.17

This section investigates the parameters to set for the optimization of the two proposed hybrid models. Let us present those parameters. First, the backbones used for low-level feature map representation before the transformer architecture must be selected. In practice, modality-specific

ResNet-18 backbones are explored, along with backbone including modality collaborations like MMTM for RegDB or MSAF for SYSU-MM01 since those two were performing very well on those respective datasets. Then, the used backbones are eventually pre-trained (P) and frozen (F) if pre-trained. Finally, the depth of the transformer architecture is explored from $L = 1$ to $L = 3$. These parameters are selected for each MSAT and MCAT model and for RegDB Tab. II-1 or SYSU-MM01 Tab. II-2.

For the RegDB dataset, results on the clean data are much closer from one configuration to another than the CCD version results. For this reason, and for selecting a model robust to corruption, the performances are mainly observed from the CCD data. First, one can observe that the best performances come with a pre-trained MMTM backbone for the two MSAT and MCAT models, with no freeze of the CNN while learning the transformer. Pretraining is necessary, probably making learning the added transformer easier. Freezing, on the other hand, leads to lower performances, probably restricting too much the model weights. Then, the model's depth to choose depends on the model and its parameters. Indeed, for MSAT and a pre-trained MMTM backbone, three layers lead to the highest results, while those are reached with $L = 1$ for MCAT with the same backbone.

On the SYSU-MM01 dataset, the best model configuration for both MSAT and MCAT comes with no pretraining or freezing of the CNN backbone while using MSAF as CNN backbone. The difference in pretraining needs from RegDB to SYSU-MM01 dataset might result from the data quantity, which is much lower for RegDB. Then, under this configuration, a depth of 2 for the MSAT model and 3 for the MCAT optimize their performances. For the SYSU-MM01 dataset, one can see more consistency from the clean to the UCD performances than for the RegDB dataset, the clean data being more challenging for the SYSU dataset, probably allowing for a better model evaluation.

Table-A II-2 Hybrid models accuracy on the RegDB the SYSU-MM01 clean and UCD dataset regarding the used CNN backbone, pretraining (P), freezing (F), and the transformer's depth L. Separated CNN stands for modality-specific ResNet-18 backbones with no interactions between the two streams.

Model	CNN	P	F	L	SYSU-MM01		SYSU-MM01-UCD	
					mAP	mINP	mAP	mINP
MSAT	Separated	Yes	No	1	94.12±0.08	66.80±0.57	60.33±0.73	8.02±0.30
	Separated	Yes	No	2	94.22±0.36	67.02±1.10	60.62±1.57	8.06±0.74
	Separated	Yes	No	3	94.16±0.30	67.19±0.86	60.35±1.02	8.06±0.38
	MSAF	Yes	No	1	93.41±0.27	64.16±1.13	62.11±0.67	9.62±0.28
	MSAF	Yes	No	2	93.34±0.34	64.79±0.27	62.30±0.77	9.66±0.25
	MSAF	Yes	No	3	93.07±0.41	63.86±1.14	61.68±0.99	9.60±0.59
	MSAF	Yes	Yes	1	93.20±0.56	63.53±1.54	62.32±1.52	9.82±0.74
	MSAF	Yes	Yes	2	93.19±0.64	64.09±1.16	61.49±0.82	9.47±0.33
	MSAF	Yes	Yes	3	93.49±0.24	64.85±0.43	61.69±1.95	9.55±0.82
	MSAF	No	No	1	94.90±0.17	69.64±0.89	62.33±0.81	9.95±0.47
	MSAF	No	No	2	**95.19±0.45**	**70.84±1.46**	**63.21±0.86**	**10.15±0.72**
	MSAF	No	No	3	95.17±0.13	70.83±0.50	61.77±0.88	9.45±0.73
MSAT	Separated	Yes	No	1	94.27±0.32	67.28±0.74	60.24±1.12	7.89±0.48
	Separated	Yes	No	2	94.01±0.80	66.78±3.07	61.58±2.94	9.81±1.93
	Separated	Yes	No	3	94.41±0.20	68.56±0.52	63.60±1.10	11.04±0.44
	MSAF	Yes	No	1	93.10±0.89	63.31±2.24	61.06±2.08	8.78±0.94
	MSAF	Yes	No	2	94.11±0.14	66.64±0.93	64.87±0.54	11.20±0.37
	MSAF	Yes	No	3	93.85±0.46	65.80±0.84	64.25±0.84	11.26±0.56
	MSAF	Yes	Yes	1	93.14±0.61	64.19±1.17	62.35±0.70	10.00±0.22
	MSAF	Yes	Yes	2	93.75±0.17	65.27±1.00	64.77±0.99	10.94±0.60
	MSAF	Yes	Yes	3	93.78±0.27	65.72±1.08	64.72±0.97	11.60±0.67
	MSAF	No	No	1	95.29±0.38	71.04±1.14	62.57±1.27	10.05±0.63
	MSAF	No	No	2	**95.64±0.18**	**72.87±0.65**	65.10±2.14	11.68±1.13
	MSAF	No	No	3	95.60±0.21	72.63±1.05	**65.85±0.83**	**11.84±0.50**

3.2 Hybrid and CNN-based comparison

On the best hybrid model configurations, the performance on the RegDB (Tab. II-3) and SYSU-MM01 (Tab. II-4) datasets are compared to the best multimodal approaches, both in terms of accuracy and complexity.

Table-A II-3 Performance comparison of the three hybrid models with our best approaches on the clean and CCD RegDB datasets.

Model	No params (M)	FLOPs (G)	Clean		CCD	
			mAP	mINP	mAP	mINP
MMTM	**23.8**	**1.54**	99.84	99.24	63.34	17.81
MMSF	34.6	3.09	**99.95**	**99.69**	**71.52**	**30.43**
Transfuser	57.1	3.35	99.89	99.51	57.71	15.24
MSAT	40.75	1.71	99.66	98.56	63.41	25.22
MCAT	30.24	1.60	99.56	98.26	64.04	25.61

Table-A II-4 Performance comparison of the three hybrid models with our best approaches on the clean and UCD SYSU-MM01 datasets.

Model	No params (M)	FLOPs (G)	Clean		UCD	
			mAP	mINP	mAP	mINP
MSAF	**22.5**	**1.54**	96.36	73.70	**67.78**	10.09
MMSF	31.9	2.31	**97.77**	**80.38**	65.82	10.51
Transfuser	57.1	3.35	95.58	69.24	65.88	9.43
MSAT	34.44	1.66	95.19	70.84	63.21	10.15
MCAT	39.69	1.71	95.60	72.63	65.85	**11.84**

On both RegDB and SYSU-MM01 datasets, the transfuser model performs considerably under the other approaches while being the most complex. For example, despite similar performances with our MMSF model under SYSU-MM01 and its UCD version, its mAP and mINP are lower by 2.19 and 11.14 percentile points, respectively. This might be a consequence of the data amount, as transformers used in the transfuser model bring much parameters and as transformers are known to require much data for optimization. Instead, MSAF and MSAT uses transformer later in the process, making it lighter and consequently easier to learn.

Comparing MSAT and MCAT together, the MCAT better deals with corrupted data, whereas the performances are pretty close under the clean setting on both datasets. This aspect probably comes from the knowledge exchange the cross-attention allows, adapting better each modality representation thanks to the knowledge of how each correlate. In terms of complexity, the number

of parameters the FLOPs depend on the dataset as the depth varies from one model to another, but where the parameters globally increases much, the required FLOPs is almost equivalent, varying from 1.60G to 1.71G FLOPs. In fact, the transformer does not increase FLOPs so much when used as late in the process, despite bringing many parameters. Indeed, MSAT and MCAT use the MMTM model as backbone for RegDB, and their FLOPs requirements increase only by 0.17G and 0.06G, respectively, from MMTM. Also, knowing the MMTM backbone is used, we can see how the transformer helps improve the model robustness, increasing its mINP by 7.80 percentile points through MCAT. On the SYSU-MM01 dataset, the mINP is also improved, but this comes at the cost of a lower mAP.

Despite some robustness improvements from the used CNN backbone on the RegDB dataset, and especially mINP increase on each dataset for the MCAT model, the hybrid models remain all under the proposed MMSF architecture in terms of performance. Indeed, for RegDB, the MMSF is strongly ahead of other models on clean and corrupted data. For SYSU-MM01, the MMSF performances on the clean version of the dataset with more than 2 mAP and 7 mINP percentile points higher than the best hybrid model, for example. This improvement makes the loss of 0.03 mAP and 1.33 mINP percentile points from MMSF to MCAT less relevant, and allows to present MMSF as the most effective approach.

4. Conclusion

This appendix explores hybrid models based on CNN and transformers architectures through three distinct models: the transfuser, the MSAT, and the MCAT. These three models have distinct functioning, transfuser including self-attention transformers between the modality-specific backbones, and MSAT and MCAT including self-attention and cross-attention transformers, respectively, after the CNN backbones.

Among the hybrid architectures, the proposed models, especially the MCAT model, outperform the transfuser in terms of complexity and performance. The cross-attention transformer is shown as more beneficial than the self-attention transformer, thanks to the knowledge exchange

it permits. More knowledge exchanges are also favored; those models benefit more from using a CNN backbone that allows for modality knowledge exchanges.

Comparing the performances between a given CNN backbone and this same model while adding a transformer to build the feature vectors from the feature map representation, we observed better handling of the hardest scenarios through the mINP under the corrupted datasets. Hence, even if such a strategy might be interesting in specific cases, it also weighs up the model with no performance benefits in most cases, which makes it not worth it. Instead, our MMSF model remains the best strategy, with a competitive complexity but especially great performances overall.

APPENDIX III

SUPPLEMENTARY MATERIAL FOR THE IJCV SUBMISSION

1. Details regarding infrared corruptions

Further details are provided Tab. III-1 concerning the way infrared corruptions were obtained from the existing visible ones. Also, a figure gathering an example of each 19 infrared corruptions is presented Fig. III-1.

Table-A III-1 Applied corruption adjustments to extend Visible (V) corruptions to the Infrared (I) modality. V corruptions that get grayscaled to perform I corruptions appear in red.

Type	V corruption	I corruption
Noise	Gaussian noise Shot noise Impulse noise Speckle noise	Each noise is used similarly but is first grayscaled.
Blur	Defocus blur Glass blur Motion blur Zoom blur Gaussian blur	No change in the way blurs are extended to infrared.
Weather	Snow Frost Fog Rain Brightness Spatter	Brightness is not used for Infrared. Spatter (water or dirt splash) and frost get grayscaled. Others are similarly applied.
Digital	Contrast Elastic trsf Pixelate JPEG compr Saturation	Digital corruptions are the same except for saturation. Saturation for infrared make close objects brighter.

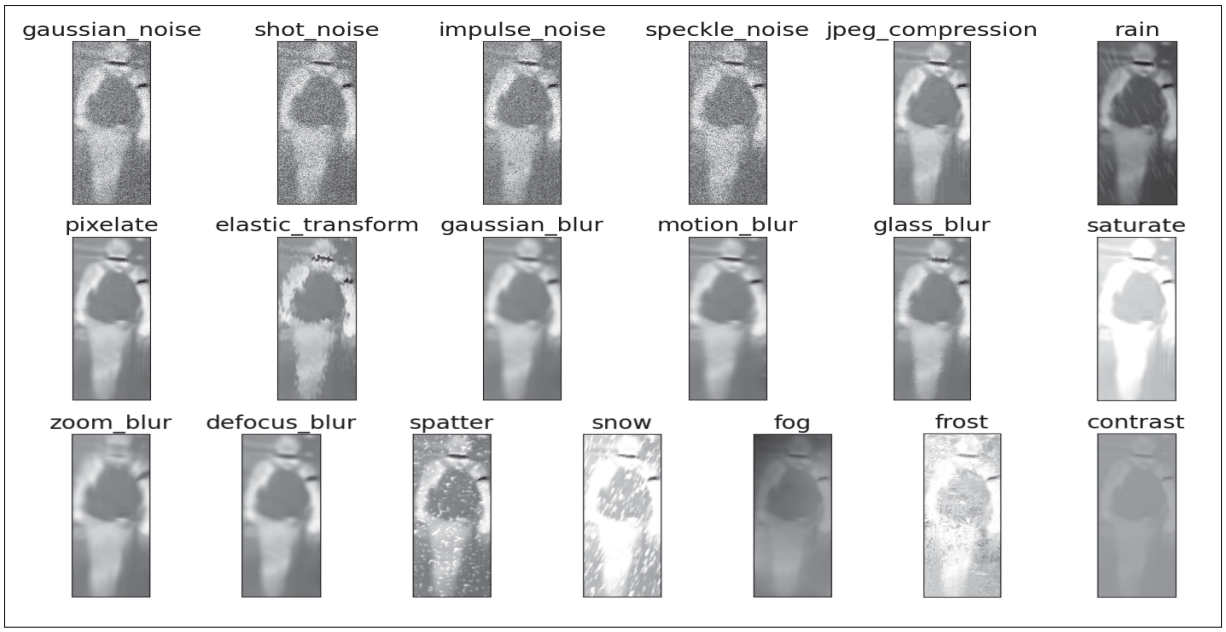

Figure-A III-1 Taxonomy of the 19 thermal corruptions, all applied with an intensity level 3.

2. MMSF optimization

2.1 MMSF and not co-located cameras

The MMSF model fuses the features from each visible and infrared backbone in its middle stream (Sec. 3.2 in main document). The proper fusion location has to be determined. The fusion can either be early in the process, fusing directly original images by element-wise sum, or later by fusing feature maps from each modality stream the same way for a given layer. Intuitively, as the cameras are not co-located and, as a consequence, the images not spatially aligned for the SYSU dataset, an early fusion in the middle stream might result in a noised fused representation. Indeed, the model in early stages might not be able to extract meaningful representation and adapt them according to the used fusion. In reverse, later-stage feature maps have a superior degree of abstraction and should suit better such fusion. Also, considering a corrupted evaluation setting, corruptions may increase the representation gap from one modality to another and thus eventually make the model further benefit from a later fusion. Still, as earlier

representation gather more information and hence more potential correlations from one modality to another, one can only be assured of where to fuse data in the middle stream with an empirical study.

Obtained results are gathered in Tab. III-2. As expected, fusing at later stages for the middle stream leads to a more discriminant final representation. Indeed, performances in both mAP and mINP gradually improve from fusing at $\ell = 0$ to fusing at $\ell = 4$ for clean data. For example, the mAP improves by $1,18\%$ and the mINP by $7,27\%$ from $\ell = 0$ to $\ell = 4$ respectively. Also, for the UCD corrupted setting, mAP improves by $2,09\%$ and $0,24\%$ mINP for the same ℓ values. Later fusion is more beneficial to the model, confirming the drawn hypothesis on NCL data. Most complex cases are similarly handled by all configurations on corrupted data according to the mINP witch evolves from 10.27% to 10.51% mINP for $\ell = 0$ to $\ell = 4$ respectively. .

Table-A III-2 MMSF performances regarding the fusion location in the middle stream, and for the clean and UCD SYSU datasets.

Model	**SYSU**		**SYSU-UCD**	
	mAP	**mINP**	**mAP**	**mINP**
MMSF 0	96.59	73.11	63.73	10.27
MMSF 1	97.27	77.01	64.28	9.57
MMSF 2	97.28	77.37	63.43	9.60
MMSF 3	97.76	79.91	64.81	10.38
MMSF 4	**97.77**	**80.38**	**65.82**	**10.51**

2.2 MMSF and co-located cameras

The MMSF model under CL cameras may behave differently than under NCL cameras due to the alignment of the visible and thermal images in a given pair. In fact, earlier fusion should allow more correlation findings as the feature representation is less compressed than later in the process. Plus, the spatial alignment should make the sum of the feature maps relevant even in the early process. Still, an earlier MMSF fusion comes with a more complex architecture since it requires more layers in the central stream, which needs to be kept in mind.

Performances regarding the fusion location l for RegDB and ThermalWORLD datasets are presented Tab. III-3. As expected, it is interesting to observe the performance decrease from $l = 0$ to $l = 4$ on RegDB for clean and CCD-50 data. In practice, the mAP decreases by 2.46%, and the mINP by 3.74% on CCD-50 dataset. ThermalWORLD results are not following this same scheme, as the results on clean data are the highest for $l = 1$, followed by $l = 2$, $l = 4$, and $l = 0$. The model act as an in-between the NCL and CL settings, which probably comes from the thermal modality being of terrible quality, messing with the expected impact of spatial alignment. Still, the RegBD model acts similarly as the ThermalWORLD one on corrupted data, performing the best through earlier fusions. In fact, earlier fusion may allow the model to get less impacted by corrupted features as the model can directly find and discard them while benefiting from the most correlations.

Table-A III-3 MMSF performances regarding the fusion location in the middle stream, and for the clean and CCD-50 versions of RegDB and ThermalWORLD datasets.

	Model	Clean		CCD-50	
		mAP	mINP	mAP	mINP
RegDB	MMSF 0	**99.95**	**99.69**	**74.25**	**33.24**
	MMSF 1	99.93	99.60	73.16	30.36
	MMSF 2	99.93	99.66	73.17	30.68
	MMSF 3	99.93	99.64	72.55	29.74
	MMSF 4	99.94	99.64	71.79	29.50
TWorld	MMSF 0	86.10	44.50	**62.77**	**14.24**
	MMSF 1	**86.27**	**45.96**	62.27	13.59
	MMSF 2	86.28	45.26	62.06	13.44
	MMSF 3	86.14	44.24	61.21	12.98
	MMSF 4	86.58	44.89	61.30	12.95

3. Element-wise sum or concatenation

3.1 Fusion with co-located cameras

The absence of spatial alignment may favor the concatenation over the element-wise sum of the feature vectors or vice-versa. Indeed, a summation might require the vector information to be aligned from one modality to another, not to erase it. Unlike element-wise sum, concatenation conserves features from each modality the same way and could consequently better fit with NCL configuration. Also, in the case of corrupted data, corruption should also tend to make the produced modality-specific feature vector representation different and hence favored concatenation as well. Still, this is only hypothetical as the information may be only semantic and aligned at this point of the data encoding.

Table-A III-4 Baseline, MMTM, and MSAF performances on SYSU-MM01 dataset while summing (S) or concatenating (C) their feature vectors. Clean and UCD evaluation only are considered since UCD respects the most NCL corruptions (Section 4.3 main document).

Model	**Clean**		**UCD**	
	mAP	**mINP**	**mAP**	**mINP**
Baseline S	96.54	74.49	**64.00**	**9.72**
Baseline C	**96.77**	**76.01**	63.40	9.51
MMTM S	94.97	68.33	63.29	9.45
MMTM C	**95.81**	**74.23**	**64.41**	**11.49**
MSAF S	**96.36**	**73.70**	**67.78**	**10.09**
MSAF C	96.04	71.13	66.20	9.68

To confirm or invalidate the previous assumptions, the baseline, MMTM, and MSAF models are compared in terms of mAP and mINP regarding a sum or a concatenation of the feature vectors, and on clean and UCD SYSU-MM01 datasets (Table III-4). The corrupted UCD set is only considered here as uncorrelated corruptions are the most suited for the NCL configuration and as it should allow answering the previous hypothesis. Models were trained using ML-MDA, but

the ML-MDA is beyond this section's scope, only used as a tool (for now) to bring consistency from clean to corrupted evaluation. Observing the baseline results, it seems beneficial to concatenate the features as expected while looking at clean data results. Indeed, concatenation improves mINP by 1.52% while conserving similar mAPs. However, performances under the UCD dataset show that summing is more beneficial, slightly improving the mAP and mINP respectively by 0.60 and $0,21\%$. These results on corrupted data are going against our hypothesis, as concatenation was expected to overpass summation under UCD. Observing attention MMTM and MSAF models results on clean and UCD data; the former considerably improves from summation to concatenation, whereas the second considerably decreases. Hence, the concatenation or summation of the features at such a level of abstraction probably allows the model to deal with the absence of spacial alignment and to align features according to the fusion used. Consequently, the best feature vector fusion strategy is model dependent and needs to be assessed experimentally. For MMTM and MSAF, the upcoming NCL analysis will consider only their best fusion version MMTM C and MSAF S.

3.2 Fusion with co-located cameras

The best strategy between element-wise sum and concatenation of the feature maps was shown to be model-dependent for NCL cameras (Section 3.1). Unlike NCL cameras, CL ones bring spacial alignment that might impact the preferred fusion differently. An empirical analysis is provided Tab. III-5 to determine which fusion to follow and if it remains model dependent by applying it on the baseline, MMTM and MSAF models. In practice, where it behaves similarly for each dataset by favoring the fusion by concatenation for the baseline models, it becomes more complex for MMTM and MSAF models. Indeed, the MMTM and the MSAF models, which exchange information between the visible and thermal CNN streams, seem not to follow a specific rule again. More than being model-dependent, performances appear as being data-dependent. For example, MMTM S performs better under both clean and corrupted RegDB settings, whereas it is MMTM C for ThermalWORLD. It is important to notice that the performance gap can be important from sum to concatenation, making such analysis important while seeking the right way to fuse feature vectors in a model. Models performing best for

Table-A III-5 Baseline, MMTM and MSAF performances on RegDB and ThermalWORLD datasets while summing (S) or concatenating (C) the feature vectors as fusion.

	Model	Clean		CCD-50	
		mAP	**mINP**	**mAP**	**mINP**
RegDB	Baseline S	99.87	99.37	63.89	17.34
	Baseline C	**99.90**	**99.45**	**64.15**	**18.08**
	MMTM S	**99.84**	**99.24**	**67.27**	**20.17**
	MMTM C	99.80	99.12	63.92	17.97
	MSAF S	99.84	99.19	59.22	13.26
	MSAF C	**99.88**	**99.33**	**63.87**	**18.28**
TWorld	Baseline S	82.18	36.89	55.68	10.55
	Baseline C	**86.34**	**43.24**	**58.01**	**11.02**
	MMTM S	86.17	45.50	59.60	10.91
	MMTM C	**87.82**	**47.95**	**60.51**	**12.36**
	MSAF S	**87.62**	**50.02**	60.78	10.93
	MSAF C	87.73	48.00	**60.57**	**12.01**

MMTM and MSAF are kept for the rest of CL cameras study.

4. Detailed complexity and accuracy trade-off

The accuracy and complexity analysis is provided in the main document Section 6.2.3 for NCL and 6.3.2 for CL cameras. However, detailed performances and complexity was not provided. Hence, this section focus on the detailed models performances for NCL and CL cameras at first and finally present the complexity in terms of parameters and FLOPs for each and every considered model.

4.1 Accuracy with not co-located cameras

Table-A III-6 Multimodal comparison with the state-of-art unimodal models. MDA refer to our ML-MDA approach.

	Model	SYSU			
		Clean		UCD	
		mAP	mINP	mAP	mINP
No DA	Unimodal V	86.25	39.97	32.36	1.91
	TransReID	94.33	64.79	52.03	3.60
	LightMBN	94.45	64.06	40.90	2.13
CIL	Unimodal V	86.64	42.78	51.64	3.83
	TransReID	93.20	62.02	61.38	7.20
	LightMBN	94.07	61.95	67.80	8.23
MDA	Baseline C	96.77	76.01	63.01	9.59
	MSAF	96.36	73.70	**67.78**	10.09
	MMSF	**97.77**	**80.38**	65.82	**10.51**

State-of-art unimodal models, along with the unimodal V model, are compared to the baseline C, MMSF, and MSAF multimodal approaches learned using our ML-MDA (Table III-6). Unimodal models are evaluated while being learned with and without the CIL strategy. As a first observation, multimodal models are all considerably over the unimodal models in terms of both mAP and mINP on clean data. Indeed, the highest improvement in mAP and mINP from the best unimodal model performances is respectively about 3.32% and 15.59%. Then, if we compare the multimodal models among themselves, MMSF comes first by improving mAP of the baseline by 1.00% while it improves its mINP by 4.37%. Surprisingly, MSAF is below the baseline's mINP by 2.31% while conserving its mAP.

Looking now at the UCD performances, the best working model is MSAF with 67.78% mAP and 10.09% mINP. Then, LightMBN and MMSF are pretty equivalent, with respectively mAPs about 67.80% and 65.82% but mINPs about 8.23% and 10.51%. From the previous observations, both the MSAF and MMSF models can be used to improve over the state-of-art unimodal models, considering both clean and corrupted data. However, the benefits from the proposed MMSF are

higher than the ones from the MSAF approach. Still, if the real-world conditions were expected as tough, MSAF would eventually be favored. However, if conditions were varying or tending to be clean, MMSF should be used.

4.2 Accuracy with co-located cameras

Table-A III-7 RegDB and ThermalWORLD - Comparison with SOTA

	Model	Clean		CCD		CCD-50	
		mAP	mINP	mAP	mINP	mAP	mINP
RegDB	Unimodal V	99.26	96.64	45.15	7.01	45.42	6.20
	TransReID	99.34	97.35	45.64	5.69	48.60	7.01
	LightMBN	99.90	99.41	32.40	7.01	33.63	3.85
	Unimodal V + CIL	99.65	98.41	55.76	10.9	58.53	14.8
	TransReID + CIL	99.69	98.57	58.74	12.8	60.48	16.2
	LightMBN + CIL	99.89	99.41	66.55	21.5	69.40	26.2
	Baseline C + ML-MDA	99.90	99.45	59.06	14.6	64.15	18.0
	MMTM + ML-MDA	99.84	99.2!4	63.34	17.8	67.27	20.1
	MMSF + ML-MDA	**99.95**	**99.69**	**71.52**	**30.4**	**74.25**	**33.2**
ThermalWORLD	Unimodal V	86.44	49.44	28.06	3.86	35.27	5.18
	TransReID	**95.86**	**77.98**	65.47	17.2	68.66	20.0
	LightMBN	93.02	65.94	37.34	5.60	44.01	6.70
	Unimodal V + CIL	86.95	48.07	52.85	7.97	56.33	10.6
	TransReID + CIL	94.79	73.82	**73.61**	**23.1**	**76.21**	**25.8**
	LightMBN + CIL	93.20	66.14	71.30	19.7	73.62	21.4
	Baseline C + ML-MDA	86.34	43.24	56.10	9.93	58.01	11.02
	MMTM + ML-MDA	87.82	47.95	58.12	11.53	60.51	12.36
	MMSF + ML-MDA	86.10	44.50	60.75	13.33	62.77	14.24

The multimodal models trained using our ML-MDA are compared with state-of-art unimodal frameworks learned using CIL DA under the CL setting Table III-7. First, observing performances on RegDB clean data, the multimodal baseline C and the proposed MMSF are ahead of the unimodal models. MMSF improving LightMBN mAP and mINP respectively from 99.89 to 99.92 and 99.45 to 99.57. If we observe corrupted performances, only the proposed MMSF can improve over the best unimodal model LightMBN + CIL, increasing the mAP by 4.97% on

CCD and by 4.85 on CCD-50. Hence, our MMSF model is the way to go for both clean and corrupted data under the CL configuration performance-wise.

About ThermalWORLD, models behave really differently. The TransReID and LightMBN models perform much better than the best multimodal approach MSAF on clean data. Indeed, for example, TransReID reaches 95.86% mAP when MSAF reaches 87.82% mAP. In fact, the slight 0.87% mAP improvement from the Unimodal V to the MSAF model shows how hard the multimodal setting benefits from the bad thermal modality. This is confirmed by the results under corrupted settings, as the best multimodal approach MMSF is 12.86% and 13.44% mAP below the TransReID approach for CCD and CCD-50 respectively. Consequently, favoring stronger unimodal models is a better strategy when the supplementary modality is far behind in terms of quality.

4.3 Models complexity

Table-A III-8 Size (Number of parameters) and computation complexity regarding FLOPs.

Model	No params (M)	FLOPs (G)
Unimodal V or I	11.3	0.51
TransReID	102.0	19.55
LightMBN	7.6	2.09
Baseline	22.5	1.54
MAN	22.5	1.54
MMTM	23.8	1.54
MSAF	22.5	1.54
MMSF l=0	34.6	3.09
MMSF l=4	31.9	2.31

Thanks to the additional modality and knowledge, a multimodal setting might allow the use of lighter backbones than a given unimodal pipeline while matching or even improving accuracy. From the previous experiments, the multimodal accuracy comes ahead unimodal approaches, but a complexity analysis remains needed, and is provided Tab. III-8. The analysis is presented regarding the models' number of parameters and the FLOPs needed to compute a single input.

First, one can observe that the TransReID complexity appeals at first sight, much heavier through 102.0M parameters than any other models, followed by MMSF with $\ell = 0$ and its 34.6M parameters. The lighter model is LightMBN, being more than ten times lighter than TransReID with 7.6M parameters. Based on the obtained results for NCL, the multimodal setting improves much the ReID accuracy. Especially, LightMBN comes first among unimodal approaches but is way less performing than MSAF and MMSF. In practice, the proposed MMSF works the best ($\ell = 4$ for NCL) under NCL cameras and should be used if resources allow it, requiring 2.31 GFLOPs and 31.9M parameters. Otherwise, MSAF would be the next model to go with 1.54 GFLOPs and 22.5M parameters, finally followed by the unimodal LightMBN approach with 2.09GFLOPs and 7.6M parameters.

Considering the CL setting, the proposed MMSF ($\ell = 0$) model is ahead, followed directly by the LightMBN model performance-wise. Similarly LightMBN comes with less complexity than MMSF, thus making a compromise between precision and complexity.

5. Qualitative analysis

Models learned through ML-MDA were compared over clean and corrupted data in terms of performances Section 6.2.1 in the main document. However, observing what the models are focusing on to discriminate and ReID would be a great way to draw additional conclusions, or at least to better understand why a model is better than another. To this end, adapted for pairwise matching algorithms, similarity based Class Activation Maps (CAMs) from Stylianou *et al.* (2019) is used. It is important to notice that MMSF CAMs are produced from its two modality specific streams only, and that the shared modality stream cannot be analysed from this CAM technique for the NCL cameras. Indeed, CAMs could be determined for the middle stream but there would be no way to dissociate from which spatial part of the V or I modality comes the shared activation.

To put visualizations in perspective, models ranking performance-wise on clean data start from MMSF, followed by MAN, Baseline C, MSAF, and MMTM. Observing Fig. III-2a., one can

see that activation on the V modality is more or less similar from one model to another, focusing mainly on the torso. Actually, MMSF might appear a bit less accurate but tend to focus on the same region. However, it seems that the most discriminant models consider both the torso and the legs of the person concerning the I modality. Indeed, from MMSF (6) to MAN (3), Baseline C (2), MSAF (5) and finally to MMTM (4), legs activation just decreases.

Switching to corrupted data, models ranking was the following under CCD: MSAF, MMSF, MMTM, Baseline C, MAN. Looking at Fig. III-2b. one may observe that the best working models focus both on the short and on the t-shirt of the individual concerning the V modality. About the I modality, it is harder to interpret, as the added snow made the focus of the models much less accurate, which in fact correlate well with the snow corruption impact on the thermal modality (Tab. 3.4). In fact, for I, both MMSF and MSAF focus on waist, but MMSF adding feet where MSAF adds shoulders to it. Also, MMTM seems much perturbed, as its attention is not so much on the person, and baseline C with MAN are both looking pretty fuzzy, mainly looking at the whole back of the individual.

If we look at corruptions that seem to less affect each modalities, with Fig. III-2c., one can see that the thermal modality (gaussian noise) is much better apprehended by each model. For the most discriminant ones, MMSF (6), MSAF (5) and MMTM (4), it is interesting to again observe the importance of feet in the ReID process.

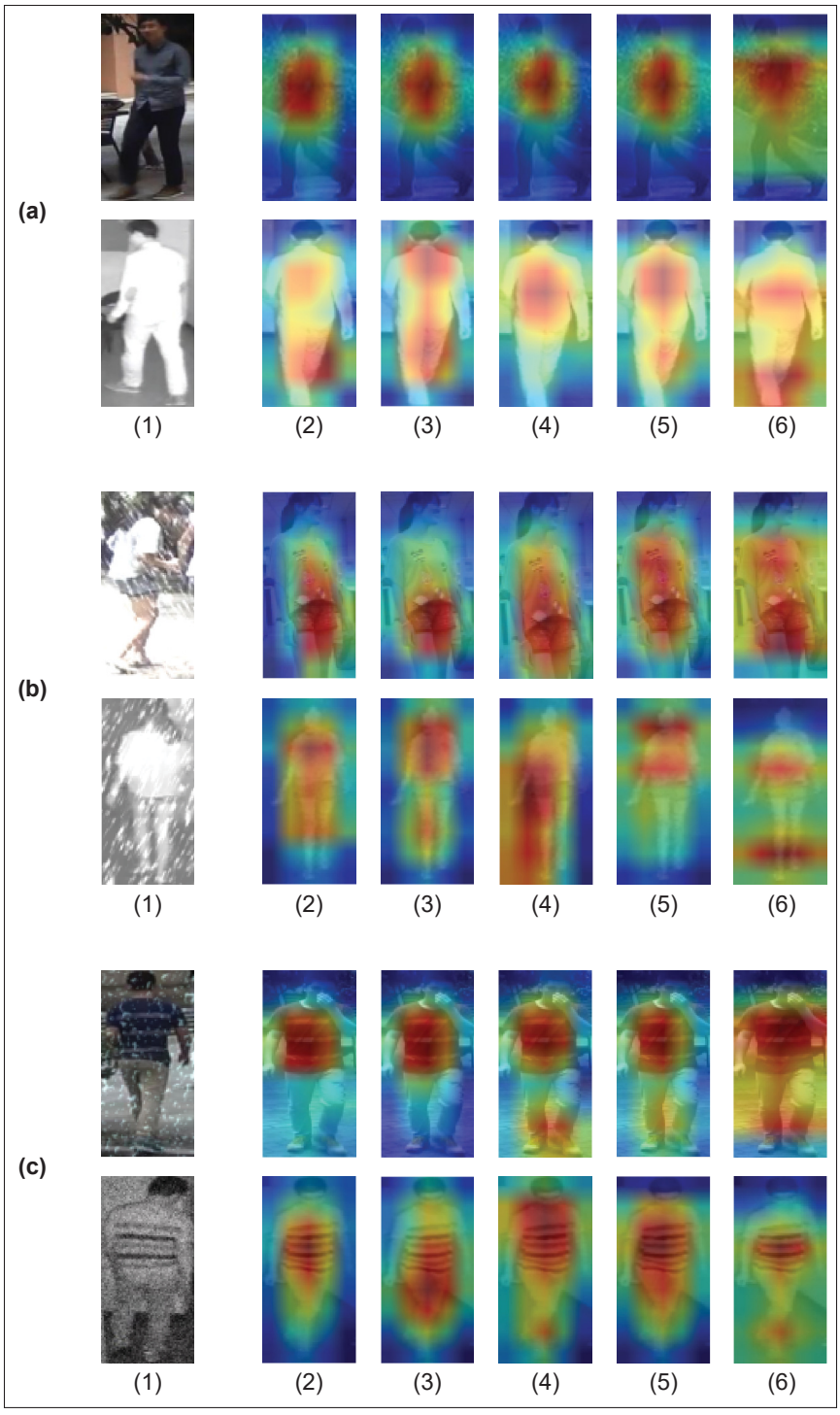

Figure-A III-2 Three examples of similarity-based CAMs using SYSU-MM01 (a) clean V-I pairs, (b) snow corrupted V-I pairs, (c) differently corrupted V (spatter) and I (gaussian noise) pairs. CAMs are computed from (2) baseline C, (3) MAN, (4) MMTM, (5) MSAF, and (6) MMSF. (1) is the reference V-I pair.

BIBLIOGRAPHY

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

Arevalo, J., Solorio, T., Montes-y Gomez, M. & González, F. A. (2020). Gated multimodal networks. *Neural Computing and Applications*, 32, 10209–10228.

Atrey, P. K., Hossain, M. A., El Saddik, A. & Kankanhalli, M. S. (2010). Multimodal fusion for multimedia analysis: a survey. *Multimedia systems*, 16(6), 345–379.

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

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

Birjali, M., Kasri, M. & Beni-Hssane, A. (2021). A comprehensive survey on sentiment analysis: Approaches, challenges and trends. *Knowledge-Based Systems*, 226, 107134.

Black, S., Stylianou, A., Pless, R. & Souvenir, R. (2022). Visualizing Paired Image Similarity in Transformer Networks. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3164–3173.

Bromley, J., Guyon, I., LeCun, Y., Säckinger, E. & Shah, R. (1993). Signature verification using a" siamese" time delay neural network. *Advances in neural information processing systems*, 6.

Chang, Y., Jung, C., Sun, J. & Wang, F. (2020). Siamese dense network for reflection removal with flash and no-flash image pairs. *International Journal of Computer Vision*, 128, 1673–1698.

Chattopadhay, A., Sarkar, A., Howlader, P. & Balasubramanian, V. N. (2018). Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. *2018 IEEE winter conference on applications of computer vision (WACV)*, pp. 839–847.

Chen, J., Yang, Q., Meng, J., Zheng, W.-S. & Lai, J.-H. (2019a). Contour-Guided Person Re-identification. *PRCV*.

Chen, K. & Salman, A. (2011). Extracting speaker-specific information with a regularized siamese deep network. *Advances in Neural Information Processing Systems*, 24.

Chen, L., Chen, J., Hajimirsadeghi, H. & Mori, G. (2020a). Adapting Grad-CAM for embedding networks. *WACV*.

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F. & Adam, H. (2018a). Encoder-decoder with atrous separable convolution for semantic image segmentation. *ECCV*.

Chen, M., Ge, Y., Feng, X., Xu, C. & Yang, D. (2018b). Person re-identification by pose invariant deep metric learning with improved triplet loss. *IEEE Access*, 6, 68089–68095.

Chen, M., Wang, Z. & Zheng, F. (2021). Benchmarks for corruption invariant person re-identification. *arXiv preprint arXiv:2111.00880*.

Chen, T., Ding, S., Xie, J., Yuan, Y., Chen, W., Yang, Y., Ren, Z. & Wang, Z. (2019b). Abd-net: Attentive but diverse person re-identification. *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 8351–8361.

Chen, X., Fu, C., Zhao, Y., Zheng, F., Song, J., Ji, R. & Yang, Y. (2020b). Salience-guided cascaded suppression network for person re-identification. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3300–3310.

Cheng, D., Gong, Y., Zhou, S., Wang, J. & Zheng, N. (2016). Person re-identification by multi-channel parts-based cnn with improved triplet loss function. *Proceedings of the iEEE conference on computer vision and pattern recognition*, pp. 1335–1344.

Chicco, D. (2021). Siamese neural networks: An overview. *Artificial neural networks*, 73–94.

Choi, S., Lee, S., Kim, Y., Kim, T. & Kim, C. (2020). HI-CMD: hierarchical cross-modality disentanglement for visible-infrared person re-identification. *CVPR*.

Ciregan, D., Meier, U. & Schmidhuber, J. (2012). Multi-column deep neural networks for image classification. *CVPR*.

Čížek, P. & Sadıkoğlu, S. (2020). Robust nonparametric regression: A review. *Wiley Interdisciplinary Reviews: Computational Statistics*, 12(3), e1492.

Coppola, D., Laiolo, M., Cigolini, C., Donne, D. D. & Ripepe, M. (2016). Enhanced volcanic hot-spot detection using MODIS IR data: results from the MIROVA system. *Geological Society, London, Special Publications*, 426(1), 181–205.

Cortes, C. & Vapnik, V. (1995). Support vector machine. *Machine learning*, 20(3), 273–297.

Dabre, R., Chu, C. & Kunchukuttan, A. (2020). A survey of multilingual neural machine translation. *ACM Computing Surveys (CSUR)*, 53(5), 1–38.

Dai, Y., Gao, Y. & Liu, F. (2021). Transmed: Transformers advance multi-modal medical image classification. *Diagnostics*, 11(8), 1384.

Davis, J. & Goadrich, M. (2006). The relationship between Precision-Recall and ROC curves. *Proceedings of the 23rd international conference on Machine learning*, pp. 233–240.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K. & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. *CVPR*.

Ding, S., Lin, L., Wang, G. & Chao, H. (2015). Deep feature learning with relative distance comparison for person re-identification. *Pattern Recognition*, 48(10), 2993–3003.

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

Elert, G. (1998). The electromagnetic spectrum, the physics hypertextbook. *Hypertextbook. com*.

Eom, C., Lee, G., Lee, J. & Ham, B. (2021). Video-based person re-identification with spatial and temporal memory networks. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12036–12045.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern recognition letters*, 27(8), 861–874.

Fu, D., Chen, D., Bao, J., Yang, H., Yuan, L., Zhang, L., Li, H. & Chen, D. (2021a). Unsupervised pre-training for person re-identification. *CVPR*.

Fu, R., Hu, Q., Dong, X., Guo, Y., Gao, Y. & Li, B. (2020). Axiom-based grad-cam: Towards accurate visualization and explanation of cnns. *BMVC*.

Fu, Z., Liu, F., Wang, H., Qi, J., Fu, X., Zhou, A. & Li, Z. (2021b). A cross-modal fusion network based on self-attention and residual structure for multimodal emotion recognition. *arXiv preprint arXiv:2111.02172*.

Gabeur, V., Nagrani, A., Sun, C., Alahari, K. & Schmid, C. (2022). Masking modalities for cross-modal video retrieval. *WACV*.

Gan, C., Wang, N., Yang, Y., Yeung, D.-Y. & Hauptmann, A. G. (2015). Devnet: A deep event network for multimedia event detection and evidence recounting. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2568–2577.

Gandhi, A., Adhvaryu, K., Poria, S., Cambria, E. & Hussain, A. (2022). Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. *Information Fusion*.

Gardner, M. W. & Dorling, S. (1998). Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. *Atmospheric environment*, 32(14-15), 2627–2636.

Gaw, N., Yousefi, S. & Gahrooei, M. R. (2022). Multimodal data fusion for systems improvement: A review. *IISE Transactions*, 54(11), 1098–1116.

Geirhos, R., Temme, C. R., Rauber, J., Schütt, H. H., Bethge, M. & Wichmann, F. A. (2018). Generalisation in humans and deep neural networks. *NIPS*.

Giancola, S., Cioppa, A., Deliège, A., Magera, F., Somers, V., Kang, L., Zhou, X., Barnich, O., De Vleeschouwer, C., Alahi, A. et al. (2022). SoccerNet 2022 Challenges Results. *Proceedings of the 5th International ACM Workshop on Multimedia Content Analysis in Sports*, pp. 75–86.

Gong, Y., Zeng, Z., Chen, L., Luo, Y., Weng, B. & Ye, F. (2021). A Person Re-identification Data Augmentation Method with Adversarial Defense Effect. *arXiv:2101.08783*.

Greenspan, H., Van Ginneken, B. & Summers, R. M. (2016). Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique. *IEEE transactions on medical imaging*, 35(5), 1153–1159.

Gu, Y., Yang, K., Fu, S., Chen, S., Li, X. & Marsic, I. (2018). Hybrid attention based multimodal network for spoken language classification. *Proceedings of the conference. Association for Computational Linguistics. Meeting*.

Guo, M.-H., Xu, T.-X., Liu, J.-J., Liu, Z.-N., Jiang, P.-T., Mu, T.-J., Zhang, S.-H., Martin, R. R., Cheng, M.-M. & Hu, S.-M. (2022). Attention mechanisms in computer vision: A survey. *Computational visual media*, 8(3), 331–368.

Hadsell, R., Chopra, S. & LeCun, Y. (2006). Dimensionality reduction by learning an invariant mapping. *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, 2, 1735–1742.

Han, K., Wang, Y., Chen, H., Chen, X., Guo, J., Liu, Z., Tang, Y., Xiao, A., Xu, C., Xu, Y. et al. (2020). A survey on visual transformer. *arXiv:2012.12556*.

Hao, X., Zhu, Y., Appalaraju, S., Zhang, A., Zhang, W., Li, B. & Li, M. (2023). Mixgen: A new multi-modal data augmentation. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 379–389.

Hao, X., Zhao, S., Ye, M. & Shen, J. (2021). Cross-modality person re-identification via modality confusion and center aggregation. *ICCV*.

He, K., Zhang, X., Ren, S. & Sun, J. (2016). Deep residual learning for image recognition. *CVPR*.

He, S., Luo, H., Wang, P., Wang, F., Li, H. & Jiang, W. (2021). Transreid: Transformer-based object re-identification. *arXiv:2102.04378*.

Hendrycks, D. & Dietterich, T. (2019). Benchmarking neural network robustness to common corruptions and perturbations. *arXiv:1903.12261*.

Hendrycks, D., Mu, N., Cubuk, E. D., Zoph, B., Gilmer, J. & Lakshminarayanan, B. (2019). Augmix: A simple data processing method to improve robustness and uncertainty. *arXiv:1912.02781*.

Hendrycks, D., Liu, X., Wallace, E., Dziedzic, A., Krishnan, R. & Song, D. (2020). Pretrained transformers improve out-of-distribution robustness. *arXiv:2004.06100*.

Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M. et al. (2021). The many faces of robustness: A critical analysis of out-of-distribution generalization. *ICCV*.

Hermans, A., Beyer, L. & Leibe, B. (2017). In defense of the triplet loss for person re-identification. *arXiv:1703.07737*.

Herzog, F., Ji, X., Teepe, T., Hörmann, S., Gilg, J. & Rigoll, G. (2021). Lightweight multi-branch network for person re-identification. *ICIP*.

Hong, J., Kim, M., Choi, J. & Ro, Y. M. (2023). Watch or Listen: Robust Audio-Visual Speech Recognition with Visual Corruption Modeling and Reliability Scoring.

Hu, J., Shen, L. & Sun, G. (2018). Squeeze-and-excitation networks. *CVPR*.

Huang, J., Tao, J., Liu, B., Lian, Z. & Niu, M. (2020). Multimodal transformer fusion for continuous emotion recognition. *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3507–3511.

Huang, W., Li, Y., Zhang, K., Hou, X., Xu, J., Su, R. & Xu, H. (2021). An Efficient Multi-Scale Focusing Attention Network for Person Re-Identification. *Applied Sciences*, 11(5), 2010.

Ismail, A. A., Hasan, M. & Ishtiaq, F. (2020). Improving Multimodal Accuracy Through Modality Pre-training and Attention. *arXiv:2011.06102*.

Jaderberg, M., Simonyan, K., Zisserman, A. & Kavukcuoglu, K. (2015). Spatial transformer networks. *arXiv:1506.02025*.

Jain, A. K., Ross, A. & Prabhakar, S. (2004). An introduction to biometric recognition. *IEEE Transactions on circuits and systems for video technology*, 14(1), 4–20.

Jia, X., Zhong, X., Ye, M., Liu, W. & Huang, W. (2022). Complementary data augmentation for cloth-changing person re-identification. *IEEE Transactions on Image Processing*, 31, 4227–4239.

Jiang, P.-T., Zhang, C.-B., Hou, Q., Cheng, M.-M. & Wei, Y. (2021). Layercam: Exploring hierarchical class activation maps for localization. *IEEE Transactions on Image Processing*, 30, 5875–5888.

Josi, A., Alehdaghi, M., Cruz, R. M. & Granger, E. (2023). Multimodal Data Augmentation for Visual-Infrared Person ReID with Corrupted Data. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 32–41.

Joze, H. R. V., Shaban, A., Iuzzolino, M. L. & Koishida, K. (2020). MMTM: Multimodal transfer module for CNN fusion. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13289–13299.

Kamann, C. & Rother, C. (2020). Benchmarking the robustness of semantic segmentation models. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8828–8838.

Khan, S. D. & Ullah, H. (2019). A survey of advances in vision-based vehicle re-identification. *CVIU*.

Khare, A., Parthasarathy, S. & Sundaram, S. (2021). Self-Supervised learning with cross-modal transformers for emotion recognition. *2021 IEEE Spoken Language Technology Workshop (SLT)*, pp. 381–388.

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. *ECCV Workshops*.

Krišto, M., Ivasic-Kos, M. & Pobar, M. (2020). Thermal object detection in difficult weather conditions using YOLO. *IEEE access*, 8, 125459–125476.

Krizhevsky, A., Sutskever, I. & Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6), 84–90.

Lai, S., Chai, Z. & Wei, X. (2021). Transformer meets part model: Adaptive part division for person re-identification. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4150–4157.

LeCun, Y., Bottou, L., Bengio, Y. & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), 2278–2324.

Lejbolle, A. R., Krogh, B., Nasrollahi, K. & Moeslund, T. B. (2018). Attention in multimodal neural networks for person re-identification. *CVPR Workshops*.

Li, X., Wang, W., Hu, X. & Yang, J. (2019). Selective kernel networks. *CVPR*.

Li, Y., He, J., Zhang, T., Liu, X., Zhang, Y. & Wu, F. (2021). 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.

Lian, Z., Liu, B. & Tao, J. (2021). CTNet: Conversational transformer network for emotion recognition. *TASLP*.

Lipton, Z. C. (2018). The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3), 31–57.

Lohweg, V. & Mönks, U. (2010). Fuzzy-pattern-classifier based sensor fusion for machine conditioning. *Sensor Fusion and its Applications*.

Long, J., Shelhamer, E. & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431–3440.

Lu, J., Hu, J. & Zhou, J. (2017). Deep metric learning for visual understanding: An overview of recent advances. *IEEE Signal Processing Magazine*, 34(6), 76–84.

Luna-Jiménez, C., Griol, D., Callejas, Z., Kleinlein, R., Montero, J. M. & Fernández-Martínez, F. (2021). Multimodal emotion recognition on ravdess dataset using transfer learning. *Sensors*, 21(22), 7665.

Luo, H., Gu, Y., Liao, X., Lai, S. & Jiang, W. (2019a). Bag of tricks and a strong baseline for deep person re-identification. *CVPR workshops*.

Luo, H., Jiang, W., Gu, Y., Liu, F., Liao, X., Lai, S. & Gu, J. (2019b). A strong baseline and batch normalization neck for deep person re-identification. *IEEE Transactions on Multimedia*.

Luo, W., Xing, J., Milan, A., Zhang, X., Liu, W. & Kim, T.-K. (2021). Multiple object tracking: A literature review. *Artificial Intelligence*.

Ma, F., Sun, B. & Li, S. (2021). Robust facial expression recognition with convolutional visual transformers. *arXiv:2103.16854*.

Ma, J., Ma, Y. & Li, C. (2019). Infrared and visible image fusion methods and applications: A survey. *Information Fusion*, 45, 153–178.

Martini, M., Paolanti, M. & Frontoni, E. (2020). Open-world person re-identification with rgbd camera in top-view configuration for retail applications. *IEEE Access*.

Mekhazni, D., Bhuiyan, A., Ekladious, G. & Granger, E. (2020). Unsupervised Domain Adaptation in the Dissimilarity Space for Person ReID. *ECCV*.

Mendes-Moreira, J., Soares, C., Jorge, A. M. & Sousa, J. F. D. (2012). Ensemble approaches for regression: A survey. *Acm computing surveys (csur)*, 45(1), 1–40.

Michaelis, C., Mitzkus, B., Geirhos, R., Rusak, E., Bringmann, O., Ecker, A. S., Bethge, M. & Brendel, W. (2019). Benchmarking robustness in object detection: Autonomous driving when winter is coming. *arXiv:1907.07484*.

Middya, A. I., Nag, B. & Roy, S. (2022). Deep learning based multimodal emotion recognition using model-level fusion of audio–visual modalities. *Knowledge-Based Systems*, 244, 108580.

Minaee, S., Boykov, Y. Y., Porikli, F., Plaza, A. J., Kehtarnavaz, N. & Terzopoulos, D. (2021). Image segmentation using deep learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Ming, Z., Zhu, M., Wang, X., Zhu, J., Cheng, J., Gao, C., Yang, Y. & Wei, X. (2022). Deep learning-based person re-identification methods: A survey and outlook of recent works. *Image and Vision Computing*, 119, 104394.

Mittal, A., Moorthy, A. K. & Bovik, A. C. (2011). Blind/referenceless image spatial quality evaluator. *ASILOMAR*.

Mocanu, B. & Tapu, R. (2022). Audio-Video Fusion with Double Attention for Multimodal Emotion Recognition. *2022 IEEE 14th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP)*, pp. 1–5.

Mogelmose, A., Bahnsen, C., Moeslund, T., Clapés, A. & Escalera, S. (2013). Tri-modal person re-identification with rgb, depth and thermal features. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 301–307.

Müller, R., Kornblith, S. & Hinton, G. E. (2019). When does label smoothing help? *NeurIPS*.

Nair, V. & Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. *Proceedings of the 27th international conference on machine learning (ICML-10)*, pp. 807–814.

Nakamura, Y., Ishii, Y., Maruyama, Y. & Yamashita, T. (2022). Few-shot Adaptive Object Detection with Cross-Domain CutMix. *ACCV*.

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*.

Niu, Z., Zhong, G. & Yu, H. (2021). A review on the attention mechanism of deep learning. *Neurocomputing*.

Nojavanasghari, B., Gopinath, D., Koushik, J., Baltrušaitis, T. & Morency, L.-P. (2016). Deep multimodal fusion for persuasiveness prediction. *ICMI*.

Paolanti, M., Romeo, L., Liciotti, D., Pietrini, R., Cenci, A., Frontoni, E. & Zingaretti, P. (2018). Person re-identification with RGB-D camera in top-view configuration through multiple nearest neighbor classifiers and neighborhood component features selection. *Sensors*.

Penate-Sanchez, A., Freire-Obregon, D., Lorenzo-Melian, A., Lorenzo-Navarro, J. & Castrillon-Santana, M. (2020). TGC20ReId: A dataset for sport event re-identification in the wild. *Pattern Recognition Letters*, 138, 355–361.

Poria, S., Cambria, E., Bajpai, R. & Hussain, A. (2017). A review of affective computing: From unimodal analysis to multimodal fusion. 37, 98–125.

Prakash, A., Chitta, K. & Geiger, A. (2021). Multi-Modal Fusion Transformer for End-to-End Autonomous Driving. *CVPR*.

Rahate, A., Walambe, R., Ramanna, S. & Kotecha, K. (2022). Multimodal co-learning: challenges, applications with datasets, recent advances and future directions.

Ramachandram, D. & Taylor, G. W. (2017). Deep multimodal learning: A survey on recent advances and trends. *IEEE signal processing magazine*, 34(6), 96–108.

Raschka, S. (2018). Model evaluation, model selection, and algorithm selection in machine learning. *arXiv:1811.12808*.

Remigereau, F., Mekhazni, D., Abdoli, S., Cruz, R. M., Granger, E. et al. (2022). Knowledge distillation for multi-target domain adaptation in real-time person re-identification. *2022 IEEE International Conference on Image Processing (ICIP)*, pp. 3853–3557.

Rensink, R. A. (2000). The dynamic representation of scenes. *Visual cognition*, 7(1-3), 17–42.

Ristani, E. & Tomasi, C. (2018). Features for multi-target multi-camera tracking and re-identification. *CVPR*.

Rumelhart, D. E., Hinton, G. E. & Williams, R. J. (1986). Learning representations by back-propagating errors. *nature*, 323(6088), 533–536.

Rusak, E., Schott, L., Zimmermann, R. S., Bitterwolf, J., Bringmann, O., Bethge, M. & Brendel, W. (2020). A simple way to make neural networks robust against diverse image corruptions. *ECCV*.

Sahay, S., Okur, E., Kumar, S. H. & Nachman, L. (2020). Low Rank Fusion based Transformers for Multimodal Sequences. *arXiv:2007.02038*.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D. & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. *Proceedings of the IEEE international conference on computer vision*, pp. 618–626.

Sen, P. C., Hajra, M. & Ghosh, M. (2020). Supervised classification algorithms in machine learning: A survey and review. *Emerging Technology in Modelling and Graphics: Proceedings of IEM Graph 2018*.

Shahroudy, A., Wang, G. & Ng, T.-T. (2014). Multi-modal feature fusion for action recognition in rgb-d sequences. *2014 6th International Symposium on Communications, Control and Signal Processing (ISCCSP)*, pp. 1–4.

Sharma, C., Kapil, S. R. & Chapman, D. (2021). Person re-identification with a locally aware transformer. *arXiv:2106.03720*.

Shorten, C. & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of big data*.

Simonyan, K. & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Simonyan, K., Vedaldi, A. & Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.

Snoek, C. G., Worring, M. & Smeulders, A. W. (2005). Early versus late fusion in semantic video analysis. *ICMI*.

Somers, V., De Vleeschouwer, C. & Alahi, A. (2023). Body Part-Based Representation Learning for Occluded Person Re-Identification. *WACV*.

Stylianou, A., Souvenir, R. & Pless, R. (2019). Visualizing deep similarity networks. *WACV*.

Su, L., Hu, C., Li, G. & Cao, D. (2020). MSAF: Multimodal Split Attention Fusion. *arXiv:2012.07175*.

Sun, L., Liu, B., Tao, J. & Lian, Z. (2021). Multimodal cross-and self-attention network for speech emotion recognition. *ICASSP*.

Sun, Y., Chen, Y., Wang, X. & Tang, X. (2014). Deep learning face representation by joint identification-verification. *Advances in neural information processing systems*, 27.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. *CVPR*.

Tran, D., Bourdev, L., Fergus, R., Torresani, L. & Paluri, M. (2015). Learning spatiotemporal features with 3d convolutional networks. *Proceedings of the IEEE international conference on computer vision*, pp. 4489–4497.

Tsai, Y.-H. H., Bai, S., Liang, P. P., Kolter, J. Z., Morency, L.-P. & Salakhutdinov, R. (2019). Multimodal transformer for unaligned multimodal language sequences. *Proceedings of the conference. Association for Computational Linguistics. Meeting*, 2019, 6558.

Uddin, M. K., Lam, A., Fukuda, H., Kobayashi, Y. & Kuno, Y. (2020). Depth guided attention for person re-identification. *Intelligent Computing Methodologies: 16th International Conference, ICIC 2020, Bari, Italy, October 2–5, 2020, Proceedings, Part III 16*, pp. 110–120.

Uddin, M. K., Lam, A., Fukuda, H., Kobayashi, Y. & Kuno, Y. (2021). Fusion in dissimilarity space for RGB-D person re-identification. *Array*, 12, 100089.

Uddin, M. K., Bhuiyan, A., Bappee, F. K., Islam, M. M. & Hasan, M. (2023). Person Re-Identification with RGB–D and RGB–IR Sensors: A Comprehensive Survey. *Sensors*, 23(3), 1504.

Valade, S., Ley, A., Massimetti, F., D'Hondt, O., Laiolo, M., Coppola, D., Loibl, D., Hellwich, O. & Walter, T. R. (2019). Towards global volcano monitoring using multisensor sentinel missions and artificial intelligence: The MOUNTS monitoring system. *Remote Sensing*, 11(13), 1528.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. & Polosukhin, I. (2017). Attention is all you need. *arXiv:1706.03762*.

Viana, T. B., Souza, V. L., Oliveira, A. L., Cruz, R. M. & Sabourin, R. (2022). Contrastive learning of handwritten signature representations for writer-independent verification. *2022 International Joint Conference on Neural Networks (IJCNN)*, pp. 01–09.

Wang, B., Chen, X., Wang, Q., Liu, L., Zhang, H. & Li, B. (2010). Power line inspection with a flying robot. *2010 1st International Conference on Applied Robotics for the Power Industry*, pp. 1–6.

Wang, G., Zhang, T., Cheng, J., Liu, S., Yang, Y. & Hou, Z. (2019a). RGB-infrared cross-modality person re-identification via joint pixel and feature alignment. *ICCV*.

Wang, J., Jin, S., Liu, W., Liu, W., Qian, C. & Luo, P. (2021a). When human pose estimation meets robustness: Adversarial algorithms and benchmarks. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11855–11864.

Wang, X., Shu, K., Kuang, H., Luo, S., Jin, R. & Liu, J. (2021b). The role of spatial alignment in multimodal medical image fusion using deep learning for diagnostic problems. *ICIMH*.

Wang, Y. (2021). Survey on deep multi-modal data analytics: Collaboration, rivalry, and fusion. *TOMM*.

Wang, Z., Wang, Z., Zheng, Y., Chuang, Y.-Y. & Satoh, S. (2019b). Learning to reduce dual-level discrepancy for infrared-visible person re-identification. *CVPR*.

Wang, Z., Li, C., Zheng, A., He, R. & Tang, J. (2022). Interact, Embed, and EnlargE: Boosting Modality-Specific Representations for Multi-Modal Person Re-identification. *Proceedings of the AAAI Conference on Artificial Intelligence*.

Wei, X., Zhang, T., Li, Y., Zhang, Y. & Wu, F. (2020a). Multi-modality cross attention network for image and sentence matching. *CVPR*.

Wei, X., Zhang, T., Li, Y., Zhang, Y. & Wu, F. (2020b). Multi-modality cross attention network for image and sentence matching. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10941–10950.

Wojke, N. & Bewley, A. (2018). Deep cosine metric learning for person re-identification. *2018 IEEE winter conference on applications of computer vision (WACV)*, pp. 748–756.

Wörtwein, T. & Scherer, S. (2017). What really matters—an information gain analysis of questions and reactions in automated PTSD screenings. *ACII*.

Wu, A., Zheng, W.-S., Yu, H.-X., Gong, S. & Lai, J. (2017). RGB-infrared cross-modality person re-identification. *ICCV*.

Xie, Q., Luong, M.-T., Hovy, E. & Le, Q. V. (2020). Self-training with noisy student improves imagenet classification. *CVPR*.

Xu, N., Mao, W., Wei, P. & Zeng, D. (2020). MDA: Multimodal Data Augmentation Framework for Boosting Performance on Sentiment/Emotion Classification Tasks. *IEEE Intelligent Systems*.

Xuan, K., Xiang, L., Huang, X., Zhang, L., Liao, S., Shen, D. & Wang, Q. (2022). Multimodal MRI Reconstruction Assisted With Spatial Alignment Network. *TMI*. doi: 10.1109/TMI.2022.3164050.

Yan, Y., Qin, J., Chen, J., Liu, L., Zhu, F., Tai, Y. & Shao, L. (2020). Learning multi-granular hypergraphs for video-based person re-identification. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2899–2908.

Yang, X., Zhou, P. & Wang, M. (2018). Person reidentification via structural deep metric learning. *IEEE transactions on neural networks and learning systems*, 30(10), 2987–2998.

Ye, M., Lan, X., Wang, Z. & Yuen, P. C. (2019). Bi-directional center-constrained top-ranking for visible thermal person re-identification. *TIFS*.

Ye, M., Shen, J., Lin, G., Xiang, T., Shao, L. & Hoi, S. C. (2021). Deep learning for person re-identification: A survey and outlook. *TPAMI*.

Yi, D., Lei, Z., Liao, S. & Li, S. Z. (2014). Deep metric learning for person re-identification. *2014 22nd international conference on pattern recognition*, pp. 34–39.

Yu, F., Jiang, X., Gong, Y., Zhao, S., Guo, X., Zheng, W.-S., Zheng, F. & Sun, X. (2020). Devil's in the details: Aligning visual clues for conditional embedding in person re-identification. *arXiv preprint arXiv:2009.05250*.

Yu, J., Li, J., Yu, Z. & Huang, Q. (2019). Multimodal transformer with multi-view visual representation for image captioning. *IEEE transactions on circuits and systems for video technology*, 30(12), 4467–4480.

Yu, R., Dou, Z., Bai, S., Zhang, Z., Xu, Y. & Bai, X. (2018). Hard-aware point-to-set deep metric for person re-identification. *Proceedings of the European conference on computer vision (ECCV)*, pp. 188–204.

Zaidi, S. S. A., Ansari, M. S., Aslam, A., Kanwal, N., Asghar, M. & Lee, B. (2022). A survey of modern deep learning based object detection models. *DSP*.

Zeiler, M. D. & Fergus, R. (2014). Visualizing and understanding convolutional networks. *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*, pp. 818–833.

Zhang, H., Wu, C., Zhang, Z., Zhu, Y., Lin, H., Zhang, Z., Sun, Y., He, T., Mueller, J., Manmatha, R. et al. (2020). Resnest: Split-attention networks. *arXiv:2004.08955*.

Zhang, H., Luo, C., Wang, Q., Kitchin, M., Parmley, A., Monge-Alvarez, J. & Casaseca-De-La-Higuera, P. (2018). A novel infrared video surveillance system using deep learning based techniques. *Multimedia tools and applications*, 77, 26657–26676.

Zhang, Q., Lai, C., Liu, J., Huang, N. & Han, J. (2022). FMCNet: Feature-Level Modality Compensation for Visible-Infrared Person Re-Identification. *CVPR*.

Zhang, S., Zhang, S., Huang, T., Gao, W. & Tian, Q. (2017). Learning affective features with a hybrid deep model for audio–visual emotion recognition. *TCSVT*.

Zheng, A., Wang, Z., Chen, Z., Li, C. & Tang, J. (2021). Robust Multi-Modality Person Re-identification. *Proceedings of the AAAI Conference on Artificial Intelligence*.

Zheng, L., Yang, Y. & Hauptmann, A. G. (2016). Person re-identification: Past, present and future. *arXiv preprint arXiv:1610.02984*.

Zhong, Z., Zheng, L., Kang, G., Li, S. & Yang, Y. (2020). Random erasing data augmentation. *Proceedings of the AAAI conference on artificial intelligence*.

Zhou, K., Yang, Y., Cavallaro, A. & Xiang, T. (2019). Omni-scale feature learning for person re-identification. *ICCV*.

Zhou, K., Yang, Y., Cavallaro, A. & Xiang, T. (2021). Learning generalisable omni-scale representations for person re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zhu, Y., Chen, W. & Guo, G. (2015). Fusing multiple features for depth-based action recognition. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 6(2), 1–20.

Zhu, Z., Jiang, X., Zheng, F., Guo, X., Huang, F., Sun, X. & Zheng, W. (2020). Aware loss with angular regularization for person re-identification. *Proceedings of the AAAI conference on artificial intelligence*, 34(07), 13114–13121.

Zou, Z., Shi, Z., Guo, Y. & Ye, J. (2019). Object detection in 20 years: A survey. *arXiv:1905.05055*.

Zou, Z., Chen, K., Shi, Z., Guo, Y. & Ye, J. (2023). Object detection in 20 years: A survey. *Proceedings of the IEEE*.