Kaunas University of Technology

Faculty of Informatics

Investigation of Cross-Modality Person Re-Identification Techniques Between Infrared and Visible Images

Master's Final Degree Project

Gurban Shukurov

Project author

Prof. Dr. Andrius Kriščiūnas

Supervisor

Kaunas, 2026

Kaunas University of Technology

Faculty of Informatics

Investigation of Cross-Modality Person Re-Identification Techniques Between Infrared and Visible Images

Master's Final Degree Project

Artificial Intelligence in Computer Science (6211BX007)

Gurban Shukurov

Project author

Prof. Dr. Andrius Kriščiūnas

Supervisor

Prof. Dr. Vytenis Punys

Reviewer

Kaunas, 2026

Kaunas University of Technology

Faculty of Informatics

Gurban Shukurov

Investigation of Cross-Modality Person Re-Identification Techniques Between Infrared and Visible Images

Declaration of Academic Integrity

I confirm that the final project of mine, Gurban Shukurov, on the topic "Investigation of Cross-Modality Person Re-Identification Techniques Between Infrared and Visible Images" is written completely by myself; all the provided data and research results are correct and have been obtained honestly. None of the parts of this thesis have been plagiarised from any printed, Internet-based or otherwise recorded sources. All direct and indirect quotations from external resources are indicated in the list of references. No monetary funds (unless required by Law) have been paid to anyone for any contribution to this project.

I fully and completely understand that any discovery of any manifestations/case/facts of dishonesty inevitably results in me incurring a penalty according to the procedure(s) effective at Kaunas University of Technology.

Gurban Shukurov	
(name and surname filled in by hand)	(signature)

Gurban Shukurov. Investigation of Cross-Modality Person Re-Identification Techniques Between Infrared and Visible Images. Master's Final Degree Project supervisor prof. dr. Andrius Kriščiūnas; Faculty of Informatics, Kaunas University of Technology.

Study field and area (study field group): Computer science, Informatics (B01).

Keywords: cross-modality, person re-identification, infrared-visible matching, ensemble learning, knowledge distillation, occlusion, attention mechanism, feature fusion, SYSU-MM01.

Kaunas, 2026. 60 pages.

Summary

This thesis presents a systematic empirical investigation of the MACE (Modality-Aware Collaborative Ensemble Learning) framework for cross-modality infrared-to-visible person re-identification on the SYSU-MM01 benchmark. The study is structured in two experimental phases. The first phase examines whether standard architectural modifications, including alternative backbone networks, explicit attention mechanisms, and advanced feature fusion strategies, produce performance improvements when applied to an already well-optimized ensemble model. Results indicate that the tested modifications offered no consistent retrieval improvement over the ResNet-50 baseline, and in several cases reduced accuracy, suggesting that MACE's collaborative ensemble and knowledge distillation design is already well-calibrated for the modifications considered. The second phase investigates operationally relevant conditions absent from the original evaluation, specifically multi-query inference aggregation, training data efficiency, and robustness to partial occlusion. Multi-query average fusion substantially improves Rank-1 accuracy without any model retraining, revealing that single-query benchmarks underestimate practical retrieval capability. Training data experiments show that performance degrades gradually until approximately 50% of training identities are available, below which the ensemble's core learning mechanisms begin to fail due to insufficient identity diversity. Occlusion experiments demonstrate that the baseline model is sensitive to partial query occlusion, with the degree of degradation depending on which body region is obscured, and that occlusion-augmented training reduces this vulnerability while simultaneously improving clean performance. Together, the findings provide a comprehensive characterization of MACE's strengths, limitations, and practical operating boundaries.

Gurban Shukurov. Asmenų pakartotinio atpažinimo tarp infraraudonųjų ir regimųjų vaizdų metodų tyrimas. Magistro baigiamasis projektas vadovas prof. dr. Andrius Kriščiūnas; Kauno technologijos universitetas, Informatikos fakultetas.

Studijų kryptis ir sritis (studijų krypčių grupė): Informatikos mokslai, Informatika (B01).

Reikšminiai žodžiai: kryžminio modalumo asmenų pakartotinis atpažinimas, infraraudonojo ir regimojo spektro atitikimas, ansamblio mokymasis, žinių distiliavimas, atsparumas užstojimui, dėmesio mechanizmas, požymių suliejimas, SYSU-MM01.

Kaunas, 2026. 60 p.

Santrauka

Šiame magistro darbe atliekamas sisteminis MACE (modalumą suvokiančio kolaboratyvaus ansamblio mokymosi) metodo empirinis tyrimas kryžminio modalumo asmenų pakartotinio atpažinimo srityje, naudojant infraraudonojo ir regimojo spektro vaizdus iš SYSU-MM01 testų rinkinio. Tyrimas suskirstytas į dvi eksperimentines fazes. Pirmoje fazėje nagrinėjama, ar standartiniai architektūriniai pakeitimai, įskaitant alternatyvias pagrindinio tinklo architektūras, dėmesio mechanizmus ir požymių sujungimo strategijas, pagerina jau optimizuoto ansamblio modelio veikimą. Rezultatai rodo, kad išbandyti pakeitimai nesuteikė nuoseklaus paieškos pagerėjimo, lyginant su baziniu ResNet-50 modeliu, o keliais atvejais sumažino tikslumą, kas leidžia daryti išvadą, kad MACE kolaboratyvaus ansamblio ir žinių distiliavimo dizainas jau yra gerai sukalibruotas tiriamų modifikacijų atžvilgiu. Antroje fazėje tiriamos operatyvinės sąlygos, nenagrinėtos originaliame darbe: kelių užklausų agregacija, mokymo duomenų efektyvumas ir atsparumas daliniams vaizdų užstojimams. Kelių užklausų vidurkinė agregacija žymiai pagerino Rank-1 tikslumą be papildomo modelio apmokymo, parodant, kad vienos užklausos vertinimo kriterijai nepakankamai atspindi praktines paieškos galimybes. Duomenų efektyvumo eksperimentai rodo, kad veikimas nuosekliai blogėja iki maždaug 50 % mokymo tapatybių, o žemiau šios ribos ansamblio mokymosi mechanizmai praranda efektyvumą dėl nepakankamos tapatybių įvairovės. Užstojimo eksperimentai atskleidžia, kad bazinis modelis yra jautrus daliniam užklausų kūno užstojimui, degradacijos laipsnis priklauso nuo užstojamos kūno srities, o mokymas su sintetiniu užstojimu sumažina šį pažeidžiamumą kartu pagerinant veikimą su švariais vaizdais. Bendrai šie rezultatai pateikia išsamią MACE stiprybių, apribojimų ir praktinių eksploatacijos ribų charakteristiką.

Table of contents

List of figures

List of tables

List of abbreviations and terms

Abbreviations:

BNNeck – Batch Normalization Neck;

CAM – Channel Attention Module;

CBAM – Convolutional Block Attention Module;

CCA – Canonical Correlation Analysis;

CDFE – Common Discriminant Feature Extraction;

CRAFT – Camera coRrelation Aware Feature augmenTation;

CMC – Cumulative Matching Characteristic;

CNN – Convolutional Neural Network;

CPU – Central Processing Unit;

D2RL – Dual-Level Discrepancy Reduction Learning;

DAM – Dual Attention Module;

GAN – Generative Adversarial Network;

GeM – Generalized Mean Pooling;

GPU – Graphics Processing Unit;

HAT – Homogeneous Augmented Tri-modal learning;

HOG – Histogram of Oriented Gradients;

IR – Infrared;

KISSME – Keep It Simple and Straightforward Metric;

KL – Kullback-Leibler divergence;

LFDA – Local Fisher Discriminant Analysis;

LOMO – Local Maximal Occurrence representation;

MACE – Modality-Aware Collaborative Ensemble Learning;

mAP – Mean Average Precision;

MBConv – Mobile Inverted Bottleneck Convolution;

MLP – Multi-Layer Perceptron;

NIR – Near-Infrared;

Re-ID – Person Re-Identification;

RGB – Red, Green, Blue;

SAM – Spatial Attention Module;

SE – Squeeze-and-Excitation;

SGD – Stochastic Gradient Descent;

SIFT – Scale-Invariant Feature Transform;

SYSU-MM01 – Sun Yat-sen University Multimodal dataset 01;

t-SNE – t-distributed Stochastic Neighbor Embedding;

VIS – Visible light spectrum.

Terms:

All-Search mode – the SYSU-MM01 evaluation protocol in which all visible-camera images from all locations are included in the retrieval gallery, representing the most challenging evaluation setting.

Ensemble learning – a machine learning strategy that combines the predictions of multiple models or classifier branches to produce a single, more reliable output than any individual component.

Feature embedding – a fixed-length vector representation of an input image in a high-dimensional space, where distances between vectors correspond to semantic similarity between identities.

Gallery – the set of stored reference images against which a probe image is compared during retrieval.

Knowledge distillation – a training technique in which one model learns from the soft probability outputs of another, encouraging consistent identity representations across different branches or modalities.

Modality gap – the distributional difference between feature representations extracted from images captured by sensors of different types, such as infrared and visible-spectrum cameras.

Probe (or query) – a single test image submitted to a Re-ID system to retrieve matching identities from the gallery.

Rank-1 accuracy – the proportion of probe queries for which the correct identity appears as the top-ranked gallery result.

Triplet loss – a metric learning objective that trains a model so that a given anchor sample is embedded closer to a positive sample (same identity) than to a negative sample (different identity) by a fixed margin.

Introduction

Relevance of the problem

Person re-identification (Re-ID) has become a central task in computer vision, with applications in surveillance, public safety, and forensics. Conventional Re-ID frameworks rely on visible-spectrum cameras, but they exhibit limited effectiveness at nighttime or in poorly lit conditions. Infrared (IR) cameras address this limitation, yet they capture a fundamentally different spectral representation from daytime colour cameras, creating a large domain gap that complicates cross-modal feature alignment. Bridging this gap reliably is a prerequisite for 24-hour surveillance systems and motivates the growing field of cross-modality IR-VIS person Re-ID.

Motivation

Despite meaningful progress in ensemble-based and tri-modal cross-modality frameworks, the published literature evaluates these systems predominantly under idealized conditions such as clean query images, complete training datasets, a single query observation per identity and similar. Real surveillance deployments face additional constraints: queries are often partially occluded, training data collection is expensive and limited, and multiple observations of the same person are typically available at inference time. These operational realities have not been systematically characterized for ensemble-based frameworks, leaving a gap between reported benchmark performance and practical deployment readiness and this thesis addresses this gap.

Aim and objectives

The aim of this thesis is to conducts a systematic empirical investigation of the MACE cross-modality IR-VIS Re-ID framework, to examine how it responds to standard architectural modifications and how it behaves under operationally relevant conditions absent from the original evaluation. Specifically, the objectives are:

1. to evaluate whether standard architectural modifications such as backbone substitution, attention mechanisms, and feature fusion strategies improve an already well-optimized ensemble model;
2. to quantify the retrieval benefit of aggregating multiple query observations at inference time;
3. to characterize the relationship between the volume of training data and model performance;
4. to measure MACE's sensitivity to partial query occlusion and determine whether occlusion-aware training mitigates that vulnerability.

Structure of the work

The thesis is organized into three main chapters. Chapter 1 reviews the literature on cross-modality Re-ID, surveys existing solutions and their trade-offs, and formulates the research questions. Chapter 2 establishes the methodology, covering the MACE architecture in depth, the experimental dataset, and evaluation protocols. Chapter 3 presents six experiments organized in two phases where Phase 1 (Experiments 1-3) investigates architectural modifications to the baseline, while Phase 2 (Experiments 4-6) characterizes the system under multi-query aggregation, limited training data, and partial occlusion conditions. The thesis concludes with a synthesis of findings across all experiments relative to the research questions.

Use of Artificial Intelligence tools

During the preparation of this master's thesis, several digital tools were used in a limited and supplementary capacity. ChatGPT and Gemini were consulted to aid understanding of complex concepts and mathematical formulations encountered in the literature, and to review the overall structural coherence of certain sections. Grammarly and LanguageTool were used to check grammar, punctuation, and sentence readability throughout the document. None of these tools were used to generate research ideas, experimental designs, results, or conclusions; all such content originates from independent study, analysis of the scientific literature, and ongoing consultation with the thesis supervisor. The author takes full responsibility for the content and the final form of this thesis.

1. Analysis of cross-modality IR-VIS person Re-ID methods

1.1. Literature review

This section provides an in-depth overview of person re-identification (Re-ID) and, specifically, the cross-modality scenario between infrared (IR) and visible (VIS) images. Firstly, the general scope of pedestrian Re-ID is defined and its importance in diverse surveillance scenarios is underlined. Then follows an introduction to the concept of cross-modality matching, focusing on the IR-VIS setting and its unique difficulties. Subsequently, the state-of-the-art methods are surveyed, including both classical designs (e.g., handcrafted features, metric learning) and deep learning-based approaches tailored for cross-modality Re-ID. In addition, the most relevant public datasets are reviewed, benchmarks for IR-VIS Re-ID (e.g., SYSU-MM01, RegDB) are established, and a discussion on common performance metrics (CMC curves, Rank-k accuracy, mean Average Precision) is conducted. Finally, the key findings, gaps, and motivations in prior literature that pave the way for the subsequent chapters are highlighted.

1.1.1. Concept of pedestrian re-identification

Pedestrian re-identification (Re-ID) is a subtask of person recognition in computer vision that aims to match images of the same individual captured across multiple, non-overlapping camera views. Formally, given a probe image of a person of interest, a Re-ID system searches for images (or video frames) of the same person within a large gallery set of candidates [1]. By focusing on appearance-based cues such as colour, texture, or shape, Re-ID systems form a discriminative feature embedding. A retrieved Rank-k list is typically generated - if the true matching identity appears among the top k results, the retrieval is deemed successful [2].

Pedestrian Re-ID has drawn increasing attention due to its substantial utility in smart surveillance, public safety, and criminal investigation tasks [3, 4]. Firstly, with the ubiquity of camera networks in urban environments, law enforcement agencies can leverage Re-ID techniques to track suspects or locate missing persons. Moreover, Re-ID extends to applications in large public venues - airports, stations, or shopping malls - where multiple cameras must cooperatively monitor crowd movement. In settings such as forensic video analysis, Re-ID can expedite labour-intensive manual searches across extensive camera footage [5, 2]. Overall, an effective Re-ID pipeline reduces investigative time, enhances real-time situational awareness, and addresses both day-to-day safety and major security events.

Despite its significance, person Re-ID faces considerable challenges. One major issue lies in viewpoint variation: the same person may appear in drastically different poses or angles across cameras [4, 6]. Lighting differences also arise because real-world camera networks are installed in varying illumination conditions - daytime vs. nighttime, indoor vs. outdoor, and sunny vs. rainy. Furthermore, occlusions often occur in cluttered scenes, partially obscuring key body parts. Additional difficulties come from background clutter, scale changes (a person occupying different pixel sizes in separate views), and appearance similarity among different individuals, which can confuse identification when individuals wear similar clothing [2, 5]. Traditional single-modality Re-ID solutions typically exploit rich colour or texture cues under well-lit conditions but struggle when illumination is severely reduced (e.g., nighttime). These shared difficulties lay the groundwork for understanding the even greater complexity of cross-modality Re-ID problems.

1.1.2. Introduction to cross-modality (infrared vs. visible) Re-ID

While Re-ID research often assumes visible-spectrum RGB cameras, many modern surveillance systems also integrate infrared sensors for nighttime operation [1, 2, 7, 9]. Cross-modality Re-ID thus refers to matching person images captured by conventional daytime RGB cameras with person images captured by IR (or thermal) cameras [1, 2]. IR cameras receive electromagnetic waves at longer wavelengths than visible light, resulting in grayscale-like images. This yields a modality discrepancy: IR images often lack colour but preserve silhouettes under poor lighting conditions, whereas visible-spectrum cameras capture vivid colour details but may fail to see in the dark [2, 8]. Hence, "cross-modality" in IR-VIS Re-ID effectively unifies two domains-one with colour cues, and one with distinct reflectance in near- or thermal infrared.

Cross-modality IR-VIS Re-ID is indispensable in 24-hour surveillance. Visible cameras alone become unreliable in poorly lit or nighttime settings, whereas IR cameras excel in minimal illumination environments [19]. By fusing both sensor types, surveillance systems can track a person seamlessly from daytime to nighttime. In practice, many security deployments, such as perimeter surveillance or night-time monitoring, demand robust IR-VIS matching for consistent identity recognition [1, 10]. Overcoming the day/night barrier thus has far-reaching implications for policing, border control, and any setting requiring round-the-clock intelligence.

Main difficulties in IR-VIS matching. IR-VIS matching presents several compounding challenges. The most fundamental is the large domain shift: IR images contain a single intensity channel, causing severe colour and texture mismatch with RGB images, so a network trained purely on visible data cannot generalize across modalities without specialised adaptation [2]. IR imagery also discards the colour information that is pivotal to standard Re-ID, rendering classic colour-based descriptors ineffective [1, 5]. Additional complexity arises from sensor-dependent noise patterns, exposure variation under extreme temperatures, and minor calibration differences across camera types [2, 11]. These spectral difficulties interact with the intra-class variation already inherent in standard Re-ID (viewpoint change, partial occlusion, background clutter) compounding the alignment problem further [6, 7].

1.1.3. State-of-the-art pedestrian Re-ID techniques

To highlight both classical approaches and advanced deep-learning-based methods [21] that unify or adapt features across domains, the summary of three key developments in (a) visible Re-ID, (b) infrared Re-ID (or IR-IR matching), and (c) cross-modality IR-VIS Re-ID is provided below.

1.1.3.1. Visible-spectrum person Re-ID

Early Re-ID algorithms used hand-engineered descriptors, such as colour histograms, SIFT-like local features, and texture-based descriptors (e.g., LOMO or HOG) [4, 12]. They often combined these features with metric learning (e.g., KISSME [13], LFDA [14]) to ensure that same-person images are close in feature space. However, these are highly sensitive to illumination changes and lack robust invariance to viewpoint or pose variation [2].

Modern visible-spectrum Re-ID has achieved remarkable success via convolutional neural networks (CNNs) or Transformers. Pioneering CNN methods like Ahmed et al. [15] introduced two-branch siamese networks trained with contrastive or triplet loss functions, while subsequent works used deeper backbone architectures (ResNet, Inception, DenseNet) plus classification and metric losses. Strong

baselines such as “bag of tricks” or BNNeck integrated cross-entropy and triplet objectives, surpassing 90-95% rank-1 on major RGB Re-ID datasets (e.g., Market-1501, DukeMTMC-ReID) [16]. However, these colour-intensive embeddings are not directly applicable to IR images.

1.1.3.2. Infrared-Infrared (IR-IR) Re-ID

Compared to abundant RGB-based efforts, fewer studies address IR-IR. One example is Jungling et al. [17], which matched IR images under similar nighttime conditions or static cameras. The primary purpose was to handle short-range IR recognition or mild domain shifts within IR sensors. However, IR-IR matching alone is insufficient to handle day-night transitions in a broader camera network

Some works in face recognition investigate near-infrared to visible matching (NIR-VIS) [18]. These methods rely on consistent IR imaging across cameras, but the domain discrepancy remains lesser than IR-VIS. Overall, pure IR-IR solutions do not typically account for bridging colour vs. grayscale representation gaps crucial in cross-modality Re-ID [1, 10].

1.1.3.3. Cross-modality IR-VIS Re-ID

The cross-modality IR-VIS domain has seen increasing research only in recent years. Thus, those approaches can be broadly categorized into image/pixel-level alignment vs. feature-level alignment, with some methods combining both, which is noted below.

a) Classical handcrafted and metric. Early attempts used hand-engineered features (HOG, LOMO) combined with cross-domain metrics (e.g., CCA, CDFE, CRAFT) [2, 13, 6]. Wu et al. [2] provided a baseline on the new SYSU-MM01 dataset using such handcrafted descriptors. The performance was limited (rank-1 < 15%) because colour-based descriptors do not directly transfer to IR images, and local gradient or texture features alone are insufficient to handle the large domain gap.

b) Pixel-level or image-to-image translation. Motivated by generative modelling, some methods aim to convert visible images into IR style or vice versa, thus training a standard single-modality Re-ID backbone. Dual-level discrepancy reduction learning (D2RL) [10] and AlignGAN [7] adopt GANs to generate cross-spectrum images that unify the pixel domain. Although these solutions can reduce domain discrepancies at a pixel level, they risk introducing generation noise and structural artefacts and often require elaborate adversarial training with large GPU overhead. Some works propose partial or intermediate alignment via grayscale transformations [6], which is simpler but can discard valuable colour cues.

c) Feature-level alignment and tri-modal solutions. Another line of research directly learns a shared embedding space where IR and VIS images for the same identity map closely. Wu et al. [2] explored single-stream vs. two-stream networks and “zero-padding” to automatically adapt domain-specific channels. Ye et al. [5] introduced homogeneous augmented tri-modal learning (HAT), augmenting each visible sample with a grayscale version to form an intermediate domain. The final network is trained with a multi-modal classification objective plus ranking losses, significantly improving IR-VIS performance. Li et al. [24] similarly propose an “X modality” bridging IR and VIS domains via a lightweight generator, obtaining robust cross-modality retrieval. Meanwhile, memory-based contrastive embedding (Cheng et al. [3]) aggregates modality-aware and modality-agnostic proxies in a large memory bank, enforcing better IR-VIS alignment.

d) Ensemble or multi-branch classifiers. Some advanced solutions incorporate ensemble learning or multi-branch classifiers to handle domain-specific cues. Ye et al. [1] design both modality-specific classifiers and a shared classifier, then unify their outputs via ensemble knowledge distillation. This approach handles the distinct IR/VIS classifier discrepancy while preserving a robust shared feature representation. Later work by Ye et al. [9] extends this direction with dual-attentive aggregation, further improving alignment across modalities.

e) Practical robustness considerations. Beyond architecture design, several operational aspects of cross-modality Re-ID have received comparatively little systematic attention in the literature. Partial occlusion is a recognized challenge in standard Re-ID due to cluttered surveillance environments [2, 6, 29] that compounds the modality gap because an occluded IR query provides even fewer discriminative cues for cross-domain matching; yet robustness to occlusion has rarely been reported for ensemble-based methods. Similarly, standard benchmarks assume a single query image per probe identity, whereas real surveillance systems often capture multiple observations of the same person; the potential benefit of aggregating these multi-query observations at inference time without retraining remains unstudied in the IR-VIS context [2]. Finally, the sensitivity of cross-modality models to the volume of labelled training identities directly affects deployment feasibility in resource-limited environments, yet this relationship has not been characterized for ensemble-based frameworks [1, 2]. These gaps provide direct motivation for the operational experiments in Phase 2 of this project.

1.1.4. Relevant public datasets and benchmarks

Due to the recency of IR-VIS Re-ID, relatively few large-scale datasets exist compared to classical visible-only Re-ID sets like Market-1501 or CUHK03. Two of the most widely utilised IR-VIS datasets are SYSU-MM01 (Wu et al. [2]), and RegDB (Nguyen et al. [19]).

The SYSU-MM01 dataset has a scope of 287,628 RGB images and 15,792 IR images, covering 491 identities which were taken by 6 cameras (4 visible, 2 infrared), capturing both indoor and outdoor. It is one of the most challenging datasets because of its huge variability in lighting conditions, day/night transitions, and multiple viewpoint changes. There is a so-called “all-search” mode that includes all visible cameras in the gallery, whereas an “indoor-search” mode restricts to indoor cameras. For both modes, rank-1 accuracy and mean Average Precision (mAP) are set as standard evaluation protocols. This dataset is considered the largest benchmark for IR-VIS Re-ID, extensively used in recent methods [1, 5, 7]. SYSU-MM01 typically has medium-resolution images (around 144 x 288) after standard resizing, but occlusions and large viewpoint changes are frequent. Outdoor cameras capture nighttime IR images, often with noise from environmental reflections [2].

With regards to the RegDB dataset, it has nearly the same number of identities, 412 to be exact, but relatively fewer images, since each identity contains 10 VIS and 10 IR captures. This dataset was collected using a dual-camera system capturing both VIS and thermal IR images. For evaluation with RegDB typically split by 206 identities for training and 206 for testing is used and repeated over random 10 splits. Consequently, both “visible-to-infrared” and “infrared-to-visible” query settings are evaluated. RegDB is smaller, with each identity having just 10 IR and 10 VIS images, leading to potential training data scarcity. However, the IR images exhibit relatively stable backgrounds, focusing more on domain shift than wide viewpoint variation [1, 19].

Other older or smaller sets sometimes appear, e.g., “NIR-VIS face” datasets for cross-modality face recognition [18, 20], but these do not contain full-body person images or multi-camera vantage points.

For example, Bi et al. [4] highlights that face-based near-infrared datasets do not generalise to full-body Re-ID. Hence, SYSU-MM01 and RegDB remain the two primary benchmarks in IR-VIS person Re-ID literature.

1.1.5. Performance metrics in Re-ID

IR-VIS person Re-ID typically inherits standard Re-ID metrics but adds nuance due to cross-domain queries. Three commonly used metrics are cumulative matching characteristic (CMC), rank-k accuracy and mean average precision (mAP).

The CMC curve at rank k indicates the proportion of probes whose true match is within the top k retrieved gallery results. In IR-VIS tasks, typically, report contains rank-1, rank-5/10/20 results [1, 2]. On the other hand, a simplified form of CMC, rank-k accuracy is a single numeric summary for $k \in \{1, 5, 10, 20\}$. mAP. This measures how well a model retrieves all correct matches across an entire gallery, averaging the Average Precision for each query. This is especially important when each identity has multiple gallery images [4, 2].

In cross-modality scenarios, some works also consider domain-specific metrics or constraints (e.g., "visible-to-infrared vs. infrared-to-visible"), analysing performance in both directions [1, 19]. However, the standard metric definitions remain largely unchanged, focusing on whether the model can effectively rank cross-domain positives at the top.

1.1.6. Summary of literature

From these extensive investigations, it can be seen that IR-VIS Re-ID has arisen to address the failure of conventional RGB-based Re-ID systems at night. The literature shows a progression from handcrafted descriptors plus classical metric learning to advanced deep networks that incorporate domain-adaptation modules, generative pixel alignment, or specialised tri-modal bridging approaches [5]. Several key points can be derived from the literature review, such as modality gap and data scarcity, and pixel versus feature-level solutions. The colour-grayscale discrepancy creates large domain shifts, compounded by limited IR-VIS datasets (SYSU-MM01, RegDB). This gap has necessitated creative architectural designs that unify features at multiple levels. Talking about approaches using generative adversarial networks (GANs), some aim to unify pixel distribution, while others prefer direct feature alignment through additional modalities (grayscale or "X") and specialised cross-modality objectives. Combining identity classification, contrastive/triplet ranking, or collaborative ensemble methods has shown promising results in bridging IR and VIS features [1, 5], which led to multi-objective learning. Despite the fact that many solutions rely on heavy computational training (adversarial generation or memory-based matching), others risk discarding colour information or introducing generation noise. The performance still lags behind single-modality Re-ID [21]. Meanwhile, the best-performing systems (e.g., HAT by Ye et al. [5], MACE by Ye et al. [1]) remain constrained by small-scale IR-VIS data. Beyond architectural design, the literature leaves open several operationally important questions: how ensemble-based cross-modality systems respond to partial query occlusion, whether aggregating multiple query observations at inference improves retrieval without retraining, and how model performance scales with the number of available training identities. These dimensions are largely absent from published evaluations and motivate the second phase of experiments in this thesis.

These findings motivate the experimental investigation pursued in the subsequent chapters, focusing on whether standard architectural modifications improve an already well-optimized ensemble system, and how practical constraints such as partial occlusion, limited training data, and multi-query inference can affect the operational readiness of cross-modality Re-ID models.

1.2. Similar existing solutions

This section examines representative IR-VIS Re-ID systems and their underlying design choices, focusing on architectural strategies, practical constraints, and performance trade-offs. The analysis concludes with the motivation for selecting a specific baseline framework for this investigation.

1.2.1. Analysis of existing cross-modality systems

Cross-modality IR-VIS Re-ID systems have emerged to address the challenge of recognizing the same individuals in daytime colour images and nighttime grayscale-like infrared imagery. This objective is usually accomplished through a combination of feature-level alignment, pixel-level style transfer, or an auxiliary modality that bridges the domain gap between IR and visible images [2].

1.2.1.1. Real-world and laboratory systems

Early large-scale IR-VIS solutions, examined in [2], showcased a scenario involving separate daytime cameras and specialized IR sensors deployed at nighttime checkpoints or in low-light areas. Such applications demand robust alignment techniques that handle camera-specific variations as well as significant shifts in illumination and texture. The SYSU-MM01 dataset [2], for instance, contains both indoor and outdoor camera views (four colour cameras and two infrared cameras), forming one of the largest IR-VIS Re-ID testbeds. In some commercial contexts, IR cameras have been integrated to observe regions during late hours, although commercial systems frequently lack systematic IR-VIS matching algorithms and rely on human operators comparing nighttime IR frames with stored daytime references.

1.2.1.2. Representative architectures and key ideas

Published systems cluster around four design families identified in the literature review. Single-stream CNNs with domain-specific normalization or zero-padding learn domain-invariant representations through largely shared weights, sometimes extended with artificial grayscale samples to improve training stability [2, 23]. GAN-based two-stream systems add a pixel-alignment generator that converts IR images into pseudo-RGB space (or vice versa) before a shared Re-ID backbone, reducing the spectral gap at the cost of added parameters and adversarial training instability [7, 10]. Memory-augmented contrastive approaches store learned cluster centroids for each modality in a dynamic memory bank, matching new embeddings against these prototypes to enforce alignment across the domain shift [3]. Tri-modal and auxiliary-modality systems introduce a bridging domain, typically grayscale or a lightweight-generated “X modality” trained jointly with the IR and VIS streams to impose shared structural representations without colour sensitivity [5, 24]. Each family reflects a distinct hypothesis about where in the pipeline the modality gap is most effectively addressed.

1.2.1.3. Reported performance

On the large SYSU-MM01 dataset [2], tri-modal and memory-driven solutions frequently surpass 50% Rank-1 accuracy under single-shot, all-search conditions, a substantial improvement over early handcrafted-descriptor baselines that barely reached 15-20% Rank-1. These figures confirm the viability of IR-VIS Re-ID in real or near-real scenarios, provided the models incorporate specialized cross-domain alignment modules. A persistent gap nevertheless remains compared to single-modality visible Re-ID, where state-of-the-art Rank-1 accuracy exceeds 90% on standard benchmarks [16].

1.2.2. Implementation environment

The experimental systems surveyed in this analysis uniformly rely on GPU-accelerated deep learning frameworks. Among published cross-modality Re-ID implementations, PyTorch [26] is the predominant choice, favoured for its dynamic computation graph and modular design, which simplify the construction of multi-branch architectures, custom ranking losses, and memory-based embedding modules [3, 5]. Open-source code releases accompanying major published methods, including those for tri-modal and ensemble-based approaches, are typically provided as PyTorch repositories, enabling direct reproducibility and extension.

1.2.3. Comparisons of existing approaches

Cross-modality Re-ID algorithms can be compared along several important dimensions: accuracy (often measured by Rank-1 or mean Average Precision), complexity (model size and training time), resource usage, and availability (open-source or proprietary). Table 1 outlines some representative approaches, drawing on references [3, 5, 7, 25].

Table 1. Comparison table of existing approaches

Approach	Rank-1 (SYSU)	mAP (SYSU)	Complexity	Resource Use	Availability
Baseline single-stream [25]	15-20%	13-16%	basic CNN (ResNet)	~ 1 GPU, moderate memory	often open-source
Tri-modal (Aux. Modality) [5]	50-60%	50-55%	standard CNN + grayscale	~ 1 GPU, standard memory usage	partially open-source
GAN-based pixel align [7]	40-50%	35-45%	larger (Gen + Disc)	2 GPUs recommended, higher load	some code, not fully
Memory-based contrastive [3]	45-55%	45-53%	CNN + memory dictionary	~ 1–2 GPUs, bigger memory overhead	code demo available

Tri-modal designs achieve the strongest SYSU-MM01 results, with rank-1 exceeding 50% [5], because grayscale bridging reduces colour sensitivity without expanding model size. Memory-based approaches reach similar accuracy levels through cluster-level alignment [3], while GAN-based systems offer pixel-level translation at the cost of additional GPU demand and artifact risk [7, 10]. Single-stream baselines are the simplest to train but yield the lowest retrieval accuracy [2]. Most

academic implementations are partially open-sourced; tri-modal and memory-based codebases offer the most direct reproducibility [3, 5], while AlignGAN [7] requires manual environment configuration.

1.2.4. Motivation for chosen methods

The four architectural families surveyed above each address the IR-VIS domain gap from a different angle, but they differ substantially in complexity, training stability, and interpretability. Among these, the ensemble-based MACE framework [1] presents the most attractive combination of properties for this investigation. Unlike GAN-based approaches such as AlignGAN [7], which target the modality gap at the pixel level but introduce generator instability and artifact risk, MACE operates directly in the feature space with a stable classification and distillation objective. Unlike single-stream architectures [2] that impose identical feature paths on both modalities, MACE's dual-branch design preserves modality-distinctive representations while shared classifiers enforce cross-domain alignment. Compared to tri-modal frameworks such as HAT [5], MACE achieves competitive performance without an auxiliary data generation step, reducing pipeline complexity and training overhead.

Three additional considerations reinforce this selection. First, MACE's modular architecture (backbone, modality-specific extractors, shared classifiers, and distillation heads) isolates each component for targeted experimental analysis, matching the investigative framing of this thesis. Second, reproducibility is straightforward: the open-source codebase [1] and standardized training procedure allow direct replication and controlled modification. Third, MACE's feature-level focus avoids the resource overhead of adversarial generation [7, 10] and the extra pipeline complexity of memory-bank updates at scale [3], making it well-suited to the computational scope of this work. These properties together make MACE the most appropriate framework for the systematic empirical investigation pursued in sections 2 and 3.

1.3. Common analysis methods and approach

This section establishes the research scope and evaluation criteria for the investigation, describes the functional modules of a cross-modality Re-ID pipeline, and identifies the key risks and limitations that shape the experimental design.

1.3.1. Research scope and evaluation criteria

This investigation is scoped to the analysis and experimental evaluation of the MACE cross-modality Re-ID framework [1] on SYSU-MM01 [2], the primary and most challenging benchmark in the cross-modality IR-VIS Re-ID literature. The focus is on characterizing how architectural modifications and practical operational conditions affect matching performance, rather than on developing a deployment-ready system. Accordingly, the analysis does not address real-time inference pipelines, camera calibration, or distributed system integration; these aspects fall outside the domain of this research-oriented investigation.

Three criteria guide all experimental work. First, retrieval accuracy measured by Rank-1, Rank-5, Rank-10, and mAP under standard evaluation protocols [2] serves as the primary comparison basis, with SYSU-MM01 all-search mode as the main reference point. Second, computational feasibility on the available hardware (NVIDIA H100 GPU, Section 3.1) constrains the scope of architectural experiments, excluding approaches with prohibitively large memory or training time requirements.

Third, reproducibility is maintained throughout by using fixed random seeds, the original MACE training procedure, and established dataset splits [1, 2], ensuring all results are directly comparable with prior published work.

1.3.2. Functional components

A cross-modality Re-ID solution can be systematically divided into major modules that reflect the data flow from image capture to final identity retrieval. Each component must be clearly defined to ensure modular development and ease of future modifications. The possible functional breakdown (workflow) is shown below:

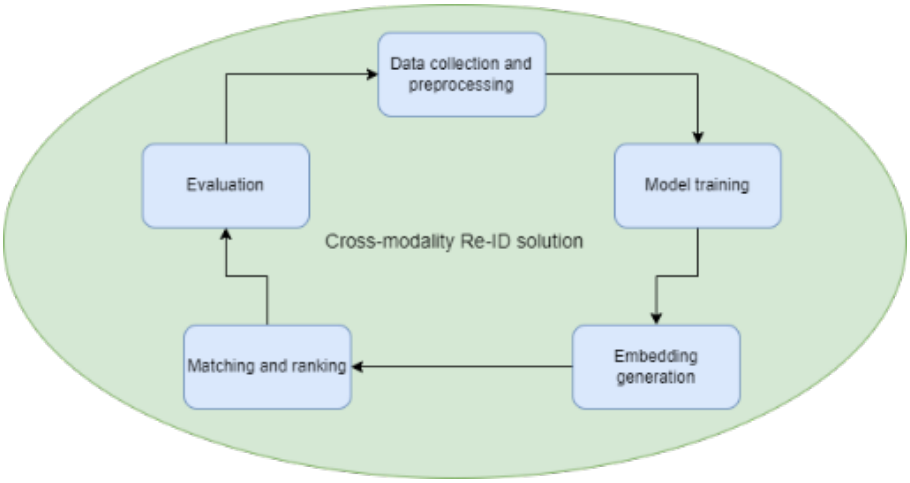

Figure 1. Workflow of the Re-ID solution

Firstly, it gathers raw images from both IR and visible cameras, then standardizes their format, resolution, and labelling conventions [2]. IR images may require intensity normalization or denoising if sensor noise is present. Visible images might be converted to grayscale or some "X" modality for bridging [5]. Then there is a model training when the system trains a deep network to align the distributions of IR and visible data, producing robust identity embeddings [3]. Some activities this part include are balancing IR vs. RGB examples, multi-modal classification or tri-modal training pipeline with separate or shared classifiers [5, 24], hyperparameter tuning. Once the network has been trained, images from the IR or visible domain are passed through the model (forward pass of the CNN or ensemble model) to obtain feature vectors (embeddings) [22] and store them in a local database or memory bank for fast retrieval. It ensures that it is suitable for the live use. If needed, normalization and other post-processing of feature vectors is conducted afterwards.

When it comes to matching and ranking, that part handles the retrieval step by computing distances between a query embedding (IR or RGB) and stored gallery embeddings. The results are sorted to generate rank lists [2, 12], distance metric is applied, ensemble is weighted if multiple classifiers are used and ranking procedure conducted for single-shot and multi-shot settings. This final part evaluates the model's performance on separate test sets (all-search or indoor-search from SYSU-MM01) or live camera data, generating standard metrics (rank-1, rank-5, mAP) for IR-VIS Re-ID [2, 5]. Additionally, confusion matrix or other charts could be plotted for domain-specific errors.

Dividing the system this way provides clear experimental boundaries for the training, evaluation, and analysis stages described in section 3.

1.3.3. Implementation framework

The implementation of this thesis builds directly on the publicly available MACE codebase [1], which is implemented in PyTorch [26]. This choice ensures direct compatibility with the original architecture and eliminates framework-related sources of experimental variance. All further details of the software environment, including library versions and hardware configuration, are documented in Section 3.2.

1.3.4. Framework selection rationale

Cross-modality Re-ID research has consolidated around PyTorch as the primary implementation framework, reflecting both its prevalence in published code releases and its suitability for the multi-branch, multi-loss training pipelines that characterize methods in this domain [1, 5, 7]. This thesis adopts the same framework, ensuring that experimental results are directly comparable with the MACE baseline [1] and that the codebase can be reproduced and extended without requiring translation between frameworks.

1.3.5. Risk analysis and limitations

Cross-modality IR-VIS Re-ID projects entail certain risks that, if left unmanaged, can affect system reliability or hamper feasibility. For instance, most publicly available IR-VIS datasets, including SYSU-MM01, contain fewer images than large RGB-only datasets, which may limit deep model generalization. A solution to that could be to explore domain adaptation or synthetic IR data generation to expand the training pool [10]. Moreover, more memory-efficient sampling strategies should be used as solutions like GAN-based systems require high-end GPUs to handle large batch sizes or store memory dictionaries [3].

Additional commonly known issue is overfitting which may occur if the IR domain is captured from a limited range of scenes or times, causing poor generalization for unseen environments. Similarly, some individuals or clothing styles might dominate the dataset. Thus, training and testing sets should be carefully partitioned, ensuring varied coverage of person IDs, camera angles, and lighting. As well, cross-dataset testing from other IR-based datasets can be beneficial as it might expose bias [4]. Other risks include extreme temperatures or reflectivity differences at night could cause IR images clarity to reduce, or even sensor differences create an additional domain gap. Fine-tuning the model with domain-specific images from the actual environment [22] and using strategies such as tri-modal [5] can possibly mitigate such risks. Nevertheless, constant performance monitoring is required to identify sensor-related or other drops as soon as possible.

Limitations of current solutions. Even though top-performing approaches can exceed 50-60% rank-1 on challenging benchmarks, this level remains lower than that of standard RGB-only Re-ID, where rank-1 may exceed 90%. The tri-modal or memory-based approaches improve cross-domain matching but can introduce additional complexity in data management [5]. Generative solutions have strong alignment potential but add cost in terms of training time and potential artifacts [7, 10]. Full-scale commercial deployment must balance these trade-offs and possibly consider domain adaptation or incremental learning if the environment evolves.

Mitigation strategies, as some examples mentioned earlier, involve a combination of careful dataset expansion, hyperparameter tuning, domain adaptation, and incremental improvements based on user feedback or real-time test performance. Maintaining a thorough risk register throughout development helps ensure quick progress and supports final system validation.

1.4. Summary of analysis findings

The discussions presented in the preceding sections have underscored both the challenges and the opportunities inherent to cross-modality person re-identification (Re-ID) between infrared (IR) and visible (RGB) images. The literature review demonstrated how single-stream or two-stream architectures, domain-adaptation modules, and data augmentation strategies have been formulated to reduce the large modality gap. It was found that grayscale or "X modality" bridging techniques often yield improved alignment without incurring prohibitive complexity, whereas memory-based contrastive embedding has been shown to further enhance feature robustness by relying on dynamically updated centroid references.

Additionally, examination of real-world and research-oriented systems indicated that GPU-accelerated deep learning frameworks enable efficient training of the complex multi-branch architectures required for cross-modality Re-ID. The comparative analysis of approaches revealed how tri-modal pipelines may achieve higher accuracy with moderate computational overhead, whereas generative adversarial solutions offer pixel-level alignment but introduce extra training burden. Risk assessments were then performed to highlight potential pitfalls - limited dataset size, domain shifts, high GPU requirements - and suggest mitigation via data augmentation or domain adaptation.

This foundation leads to several initial requirements for a subsequent solution. First, the system must effectively identify persons across IR and visible cameras, targeting a baseline rank-1 accuracy benchmark (e.g., around 50% on SYSU-MM01). Second, it must handle training and inference efficiently, mitigating heavy memory overhead or instability. Third, design should integrate maintainable modules for data preprocessing, model training, inference, and evaluation, promoting future extensions. Finally, risk management strategies - including data augmentation and incremental adaptation - are recommended to address domain and hardware constraints.

With these insights in place, focus may now shift to a targeted methodological framework. The second chapter will detail the chosen network architecture, training pipeline, and specific implementation steps that align with the requirements identified above.

1.5. Research questions and aim

As established through the preceding analysis (particularly Sections 1.1.2 and 1.1.6), the primary impediments to effective cross-modality IR-VIS Re-ID stem from the significant domain shift between infrared and visible spectra, the loss of critical colour information in IR imagery, and variations inherent in sensor technology. This investigation is designed to address these persistent challenges, identified as the current bottlenecks in the literature which continue to limit the practical deployment of cross-modality person re-identification systems in real-world scenarios. The modality discrepancy represents the most fundamental obstacle, as visual features that are discriminative in one modality often fail to maintain their utility in the other, leading to degraded matching performance across domains [1, 2].

Building upon the gaps and unanswered questions identified in the analysis section, this thesis explores the following high-level research questions:

1. **Do commonly adopted architectural modifications such as alternative backbone networks, explicit attention mechanisms, and advanced feature fusion strategies produce consistent performance improvements when applied to an already well-optimized**

ensemble cross-modality Re-ID framework? Although such modifications have individually shown benefit in related recognition tasks, their interaction with MACE's collaborative ensemble learning and bidirectional knowledge distillation has not been systematically examined [1, 5, 11].

2. **To what extent can multi-query aggregation at inference time improve cross-modality retrieval performance, and which aggregation strategy is most effective?** Standard evaluation protocols in cross-modality Re-ID assume a single query image per identity, yet practical surveillance contexts frequently provide multiple observations of the same person. Whether aggregating these images compensates for modality gap limitations without any retraining has not been reported for MACE [2].

3. **How does the volume of training identity data affect the performance of an ensemble-based cross-modality Re-ID model?** The data sensitivity of MACE, specifically how performance degrades as the fraction of training identities decreases, has not been characterized, yet this directly informs the practical deployment of such systems where labelled training data may be limited [1, 2].

4. **How vulnerable is the MACE model to partial occlusion in the query modality, and can occlusion-aware training mitigate that vulnerability?** Occlusion is a well-recognized challenge in person Re-ID [2, 6], yet the original MACE evaluation does not report robustness under occluded conditions.

The aim of this thesis is to conduct a systematic empirical investigation of the MACE cross-modality IR-VIS Re-ID framework, examining the response of its architecture to standard enhancement attempts alongside its behaviour under operationally relevant conditions absent from the original evaluation. The contribution is not a new architecture, but a rigorous characterization of MACE's strengths, limitations, and practical operating boundaries providing insights relevant to researchers developing ensemble-based Re-ID methods and to practitioners assessing the deployment readiness of cross-modality systems in real surveillance environments.

2. MACE framework architecture and evaluation methodology

2.1. Baseline methodology selection and justification

After conducting a comprehensive literature review of cross-modality person Re-ID approaches, this research adopts a depth-over-breadth strategy by selecting a single, strong baseline model for systematic investigation rather than conducting a broad but shallow comparison across multiple methods. A focused study of a single advanced framework enables clearer isolation of the effects of individual modifications and produces findings that are more directly transferable to subsequent research [16].

The Modality-Aware Collaborative Ensemble Learning (MACE) framework, introduced by Ye et al. [1] in 2020 in IEEE Transactions on Image Processing, is selected as the baseline. MACE employs a dual-branch architecture with modality-specific feature extractors coupled with modality-shared classifiers, enabling the model to jointly learn representations that are both identity-discriminative and modality-invariant. A bidirectional knowledge distillation mechanism transfers soft-label distributions between the modality-specific and shared classifier branches, promoting cross-domain consistency throughout training. This design directly addresses the core IR-VIS challenge aligning representations across heterogeneous spectral domains without recourse to generative models or auxiliary data pipelines and aligns naturally with the research questions formulated in Section 1.5.

MACE achieved Rank-1 accuracy of 51.64% and mAP of 50.11% on SYSU-MM01 all-search mode [1], representing a significant advance over prior methods and establishing a stable quantitative reference for measuring the effect of modifications. Its modular structure (backbone, modality-specific extractors, shared classifiers, and distillation heads) makes each component independently accessible for experimental intervention. Compared to GAN-based approaches [7, 8], MACE's feature-level design avoids adversarial instability; compared to tri-modal frameworks [5], it achieves competitive accuracy without auxiliary generation steps. The combination of strong performance, architectural transparency, open-source availability, and reproducible training procedure makes MACE the most appropriate framework for the investigative program pursued in this thesis [25].

2.2. Research objectives and experimental program

Building upon the MACE framework established in Section 2.1, the experimental program is organized in two phases that reflect the research questions formulated in Section 1.5.

Phase 1: Architectural modification investigation.

The first phase examines whether standard modifications to MACE's internal components alter cross-modality retrieval performance. The aim is not to assume improvement but to evaluate empirically whether each modification class interacts constructively or destructively with MACE's ensemble design.

1. Backbone evaluation. Assess the effect of substituting ResNet-50 with SE-ResNet-50, EfficientNet-B3, and TResNet-M [27] on Rank-1, Rank-5, Rank-10, and mAP, keeping all other components of the MACE training procedure fixed.

2. Attention mechanism analysis. Evaluate the integration of Channel Attention (CAM), Spatial Attention (SAM), and Dual Attention (DAM) at key positions in the MACE pipeline. Attention map visualizations will be used to interpret learned behaviour alongside performance metrics.

3. Feature fusion strategy examination. Compare adaptive weighted fusion and feature transform fusion against the baseline concatenation-based knowledge distillation, determining whether more expressive fusion improves or interferes with modality alignment.

Phase 2: Operational analysis.

The second phase characterizes MACE's performance under conditions relevant to real-world deployment that were not examined in the original paper.

4. Multi-query aggregation. Quantify the retrieval gains achievable by aggregating multiple query images per identity at inference time using average, maximum, and distance-weighted fusion strategies, without any model retraining.

5. Training data efficiency. Characterize the relationship between the fraction of available training identities and retrieval performance, revealing MACE's data sensitivity and informing minimum training data requirements.

6. Occlusion robustness. Measure performance degradation under four types of synthetic partial occlusion applied to the query modality and evaluate whether occlusion-augmented training reduces this vulnerability.

2.3. Dataset for projecting and experimenting

All experiments in this thesis are conducted on SYSU-MM01 [2], the first and most widely used large-scale benchmark for RGB-infrared cross-modality person Re-ID [1, 5, 7]. SYSU-MM01 was selected over other available benchmarks because it presents the most challenging and realistic evaluation conditions: 491 identities captured by 6 cameras (4 visible, 2 infrared) across both indoor and outdoor environments, yielding 29,033 RGB and 15,792 IR images with substantial variation in illumination, viewpoint, and background. Its multi-camera topology, day-to-night capture range, and large identity count make it the most demanding and informative testbed for the architectural and operational investigations described in Section 2.2. The dataset is also the standard benchmark against which the original MACE results were reported [1], ensuring that all experimental comparisons in this thesis are directly interpretable relative to the baseline.

The evaluation protocol defined by Wu et al. [2] partitions the dataset into 395 training identities (22,258 RGB and 11,909 IR images) and 96 test identities (6,775 RGB and 3,883 IR images). All experiments use the all-search evaluation mode, in which IR images serve as queries and all visible-camera images form the gallery. Performance is reported using Rank-1, Rank-5, Rank-10, and Rank-20 accuracy and mean Average Precision (mAP), consistent with the evaluation conventions established in the literature [1, 5, 7] and described in Section 2.5.

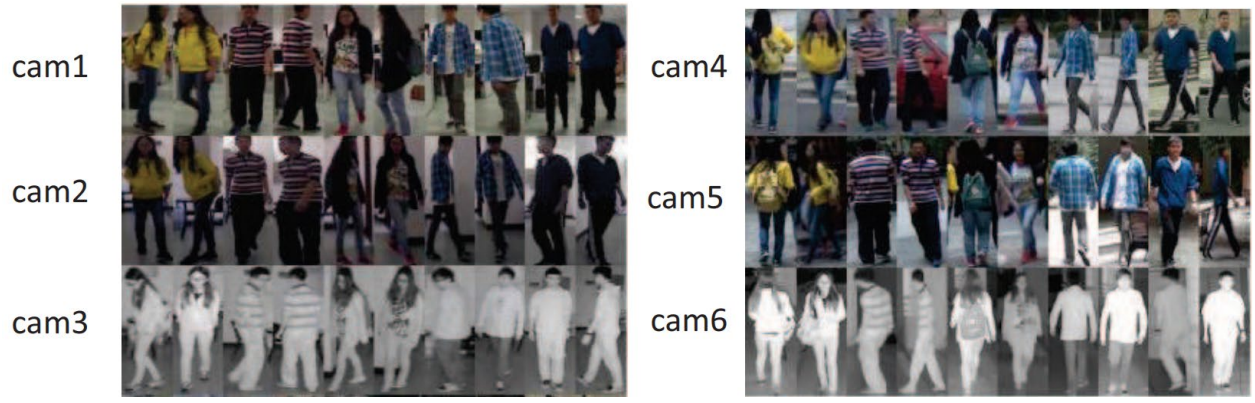

Figure 2. Sample images from SYSU-MM01 showing the same identity captured in RGB and IR modalities across different locations (indoor/outdoor), Wu et al. [2]

2.4. MACE architecture in-depth

The MACE architecture consists of four main components:

1. Modality-specific feature extractors
2. Modality-shared feature classifiers
3. Ensemble learning strategy
4. Bidirectional knowledge distillation mechanism

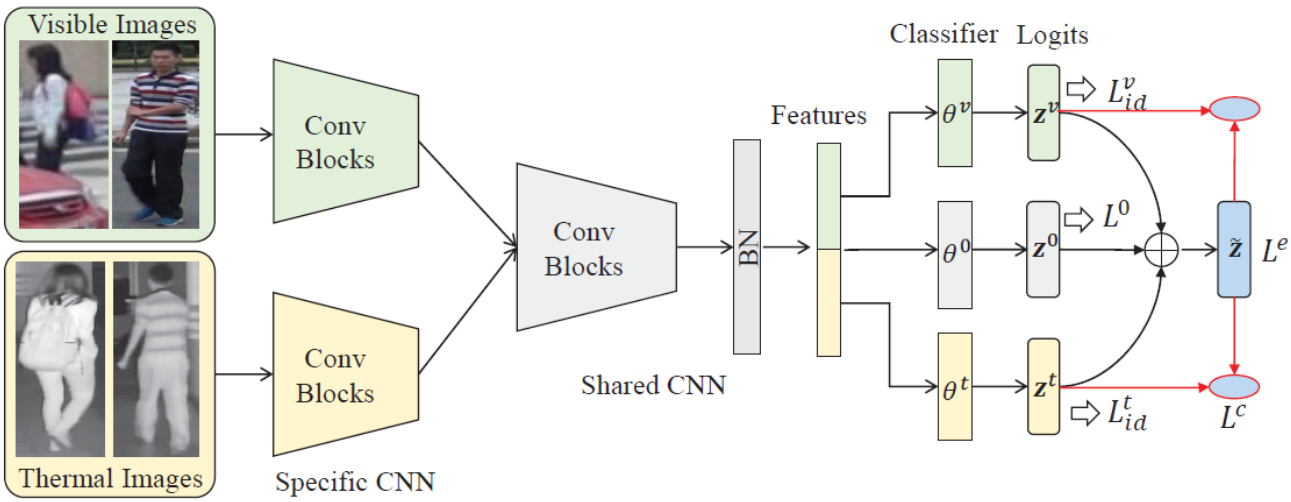

Figure 3. Framework of MACE method. [1]

The framework processes visible and infrared images through separate but parallel paths, extracting both modality-specific and modality-invariant features. These features are then used for identity classification through an ensemble of classifiers, with knowledge shared between modalities through a bidirectional distillation process.

2.4.1. Backbone architecture

MACE employs ResNet-50 as the backbone for both visible and infrared feature extractors, with several key modifications:

- Removal of the final fully connected layer
- Global average pooling after the final convolutional layer
- Pre-training on ImageNet for better generalization

In the original MACE paper [1], input images of 384×192 pixels yield feature maps of size 2048×24×12 after the fourth residual block. In the experiments conducted below, a standard 288×144 input size was used, consistent with common SYSU-MM01 practice, producing feature maps of 2048×18×9, which are then pooled to a 2048-dimensional embedding vector. These feature maps serve as the foundation for both modality-specific and modality-shared feature learning.

2.4.2. Modality-Specific and Modality-Shared feature extractors

Each modality branch passes its backbone output through a dedicated 1×1 convolutional layer (with batch normalization and ReLU) to produce modality-specific features:

$$\boldsymbol{f_V} = \boldsymbol{\varphi_{V(F_V)}}, \boldsymbol{f_I} = \boldsymbol{\varphi_{I(F_I)}}; \tag{1}$$

here F_V and F_I are the visible and infrared backbone feature maps respectively, and φ_V, φ_I are the modality-specific extraction functions. In parallel, a shared extractor φ_S that is identical in architecture but with tied weights across both branches produces modality-invariant representations:

$$\boldsymbol{f_{VS}} = \boldsymbol{\varphi_{S(F_V)}}, \boldsymbol{f_{IS}} = \boldsymbol{\varphi_{S(F_I)}}. \tag{2}$$

This parallel design ensures that each image is represented by both a modality-specific embedding (capturing domain-distinctive characteristics) and a modality-shared embedding (capturing identity-relevant features common to both spectra).

2.4.3. Ensemble classification strategy

MACE maintains separate classifiers for each branch output. Modality-specific classifiers W_V and W_I operate on the specific embeddings, while a shared classifier W_S operates on the modality-invariant embeddings:

$$\boldsymbol{p_V} = \boldsymbol{W_V} \cdot \boldsymbol{f_V}, \boldsymbol{p_I} = \boldsymbol{W_I} \cdot \boldsymbol{f_I}, \tag{3}$$

$$\boldsymbol{p_{VS}} = \boldsymbol{W_S} \cdot \boldsymbol{f_{VS}}, \boldsymbol{p_{IS}} = \boldsymbol{W_S} \cdot \boldsymbol{f_{IS}}. \tag{4}$$

At inference, the final prediction for each image is the sum of its modality-specific and modality-shared classifier outputs:

$$\boldsymbol{p_{V_{final}}} = \boldsymbol{p_V} + \boldsymbol{p_{VS}}, \tag{5}$$

$$\boldsymbol{p_{I_{final}}} = \boldsymbol{p_I} + \boldsymbol{p_{IS}}. \tag{6}$$

This ensemble combination ensures that both the domain-distinctive and the domain-invariant feature paths contribute to retrieval, improving robustness over either path alone.

2.4.4. Knowledge distillation technique

A key innovation in MACE is its bidirectional knowledge distillation mechanism, which facilitates information exchange between modalities.

Visible to Infrared distillation:

$$\boldsymbol{L_{VtoI} = KL\left(\sigma\left(\frac{p_V}{T}\right), \sigma\left(\frac{p_{IS}}{T}\right)\right)}, \tag{7}$$

Infrared to Visible distillation:

$$\boldsymbol{L_{ItoV} = KL\left(\sigma\left(\frac{p_I}{T}\right), \sigma\left(\frac{p_{VS}}{T}\right)\right)}; \tag{8}$$

here KL represents the Kullback-Leibler divergence, σ is the softmax function, and T is a temperature parameter that controls the softness of the probability distribution.

This bidirectional knowledge distillation serves two crucial purposes: it enables each modality to benefit from the discriminative capabilities of the other, and it promotes consistency between modality-specific and shared feature spaces.

2.4.5. Loss functions

MACE training is governed by multiple loss functions that work together to optimize different aspects of the model.

Identity classification losses:

$$\boldsymbol{L_{id_V} = CE(p_V, y) + CE(p_{VS}, y)}, \tag{9}$$

$$\boldsymbol{L_{id_I} = CE(p_I, y) + CE(p_{IS}, y)}; \tag{10}$$

here CE denotes the cross-entropy loss and y represents the ground-truth identity labels.

Knowledge distillation losses:

$$\boldsymbol{L_{KD} = L_{VtoI} + L_{ItoV}}, \tag{11}$$

as defined in formulas (7) and (8).

Triplet loss for feature embedding:

$$\boldsymbol{L_{tri} = \Sigma_{\{(a,p,n)\in T\}}\left[\left||f_a - f_p|\right|^2 - \left||f_a - f_n|\right|^2 + \alpha\right]_+}; \tag{12}$$

here T is the set of valid triplets mined from the combined batch of visible and infrared features, $||\cdot||^2$ denotes Euclidean distance, f_a, f_p, and f_n represent the feature vectors of the anchor, positive, and negative samples respectively (each drawn from either modality), $\alpha = 0.3$ is the margin hyperparameter, and $[\cdot]_+ = \max(\cdot, 0)$. Each training batch contains $B = batch_{size} \times num_{pos} = 16 \times 4 = 64$ combined samples per modality; hard negative mining is performed over this joint pool, making cross-modal pairs (e.g., visible anchor with infrared negative of the same identity) valid triplet candidates.

Total loss:

$$\boldsymbol{L_{total} = \lambda_{1(L_{id_V} + L_{id_I})} + \lambda_{2(L_{VtoI} + L_{ItoV})} + L_{tri};} \quad (13)$$

here $\lambda_1 = 1.0$ is the identity classification weight. The distillation weight λ_2 is not a fixed constant but follows a sigmoid ramp-up schedule, increasing smoothly from 0 toward 1.0 over $\tau = 60$ training epochs, allowing the model to stabilize its identity representations before cross-modal consistency is enforced. The triplet loss L_tri enters the total loss without a separate multiplier (effective weight 1.0); its magnitude is governed internally by the margin parameter $\alpha = 0.3$.

2.5. Evaluation metrics

The evaluation of cross-modality Re-ID methods employs standard metrics designed to assess ranking accuracy:

The Cumulative Matching Characteristic (CMC) curve represents the probability that the correct match appears within the top-k ranked gallery results. Results are reported at $k \in 1,5,10,20$, with Rank-1 accuracy serving as the primary performance indicator [1, 2].

Mean Average Precision (mAP) metric provides a more comprehensive evaluation by considering the retrieval order of all correct matches in the gallery:

$$\boldsymbol{AP = (1/M) * \Sigma\big(Prec(k) * rel(k)\big)}, \quad (14)$$

$$\boldsymbol{mAP = (1/Q) * \Sigma\big(AP_q\big)}; \quad (15)$$

here M is the number of relevant items, $Prec(k)$ is the precision at cut-off k, $rel(k)$ is an indicator function equalling 1 if the item at rank k is relevant, Q is the number of queries, and AP_q is the average precision for query q.

mAP is particularly important for cross-modality Re-ID evaluation as it accounts for scenarios where multiple correct matches exist in the gallery.

3. Experimental analysis of MACE performance and robustness

3.1. Experimental configuration

All experiments were conducted using the university's shared AI Jupyter Notebook environment equipped with an NVIDIA H100 GPU (100 GB HBM3 memory), an Intel Xeon Gold 6438Y+ CPU, 1 TiB system RAM, and 446 GB of storage. The H100's large GPU memory was the determining factor for batch size selection and for accommodating the multi-branch MACE architecture with multiple simultaneous loss computations; training duration varied between approximately 2 and 8 hours per run depending on the backbone architecture, as reported in each experiment section.

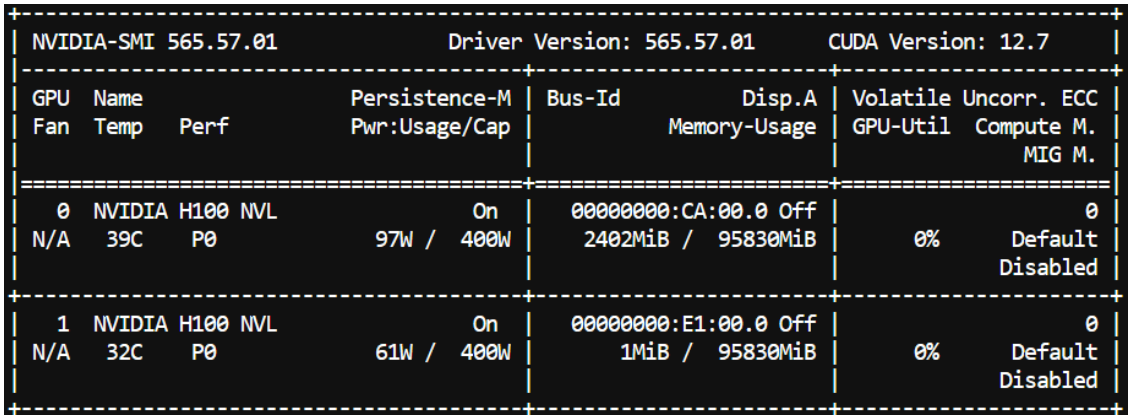

Figure 4. Screenshot from terminal with information about all GPUs available

3.2. Experimentation plan

This section provides an overview of the complete experimental program, with each experiment corresponding to a research question or operational scenario defined in Section 2.2. The program is organized in two phases: Phase 1 examines whether standard architectural modifications improve MACE's performance, while Phase 2 characterizes its behaviour under operationally relevant conditions absent from the original evaluation. Table 2 summarizes the experiments and detailed results for all the experiments are reported in Sections 3.4-3.10. The dataset SYSU-MM01 was used for model training across all the experiments.

Table 2. Overview of experimental program

Nr.	Experiment summary	Phase
0	Validate MACE baseline implementation	Setup
1	Evaluate alternative backbone architectures	1
2	Assess integrated attention modules	1
3	Optimize feature fusion strategies	1
4	Aggregate multi-query inference features	2
5	Analyse training data scaling	2
6	Test occlusion-aware training robustness	2

3.2.1. Experiment 0: Baseline replication

The initial experiment replicates the MACE framework to validate the training pipeline and establish a reliable performance baseline for all subsequent comparisons. The model was trained using the original hyperparameters and evaluated under the all-search protocol (Rank-1, Rank-5, Rank-10, Rank-20, mAP). Success requires Rank-1 ≥ 45% and mAP ≥ 43% on SYSU-MM01 all-search mode, within 6% of the figures reported in [1].

3.2.2. Experiment 1: Backbone architecture comparison

This experiment assesses whether substituting the ResNet-50 backbone with architectures incorporating modern design improvements - SE-ResNet-50 (squeeze-and-excitation channel recalibration), EfficientNet-B3 (compound scaling), and TResNet-M [27] (GPU-optimized stem and anti-alias downsampling) - produces measurable improvements within MACE's ensemble framework. All other training settings are kept identical to the baseline. Reported metrics include Rank-1, Rank-5, Rank-10, Rank-20, mAP, training time, and total parameter count.

3.2.3. Experiment 2: Attention mechanism enhancement

Three attention modules are integrated at the modality-specific extractor stage: Channel Attention Module (CAM, recalibrating feature responses across channels), Spatial Attention Module (SAM, weighting spatial locations), and Dual Attention Module (DAM, combining both in sequence). Performance metrics (Rank-1, mAP) are reported alongside attention map visualizations for both modalities to support qualitative interpretation of the learned behaviour.

3.2.4. Experiment 3: Feature fusion strategy refinement

This experiment replaces the default concatenation-based knowledge distillation with two alternative fusion strategies: adaptive weighted fusion (learned dynamic weights balancing modality-specific and shared feature contributions), feature transform fusion (non-linear projection of each modality to a common representation space before combination). Each strategy is evaluated against the baseline, reporting standard retrieval metrics and qualitative examples of how fusion affects cross-modal matching behaviour.

3.2.5. Experiment 4: Multi-query aggregation

Rather than retraining the model, this experiment tests whether aggregating multiple query images per identity at inference time improves retrieval. Three aggregation strategies are evaluated: average feature fusion, maximum feature fusion, and distance-weighted fusion. Experiments are conducted using the trained baseline MACE model, reporting Rank-1 and mAP gains relative to the standard single-query evaluation.

3.2.6. Experiment 5: Training data efficiency

This experiment characterizes how MACE's performance degrades as the fraction of available training identities decreases. The model is retrained at multiple data fractions (e.g., 25%, 50%, 75%, 100% of the 395 training identities) and evaluated on the full test set, revealing minimum data requirements for practical deployment.

3.2.7. Experiment 6: Occlusion robustness

Four types of synthetic partial occlusion are applied to the IR query images at test time (top, bottom, left, and right half-occlusion). Performance degradation under each condition is measured. A second model variant is then trained with occlusion-augmented data to evaluate whether occlusion-aware training reduces vulnerability. Results are compared against the unoccluded baseline.

Together, the seven experiments provide a comprehensive empirical characterization of MACE's architectural flexibility (Phase 1) and operational boundaries (Phase 2), addressing all four research questions formulated in Section 1.5.

3.3. Experiment 0: Initial MACE implementation and environment setup validation

The primary objective of this initial experiment was to implement the MACE model in the university's AI environment, validate the computational toolchain (PyTorch, CUDA, H100 GPU), establish a functional training pipeline, and obtain preliminary performance figures on available version of the SYSU-MM01 dataset. This experiment served as a critical preparatory step for the more comprehensive investigations planned for the remainder of the research.

3.3.1. Implementation details

The implementation of the MACE framework was based on code adapted from an author's GitHub repository that implemented the architecture described in Ye et al. [1]. The original code required several modifications to function properly in the university's AI environment and accommodate the available hardware resources.

Training parameters:

- Optimizer: Stochastic Gradient Descent (SGD) with momentum 0.9
- Initial learning rate: 0.1 with decay by factor of 10 after 20 and 30 epochs
- Batch size: 8 images per modality (modified from original due to memory constraints)
- Training duration: 60 epochs (with early stopping if no improvements in metrics)
- Weight decay: 5×10^{-4}
- Loss function weights (λ): 1.0 for identity loss, 0.1 for knowledge distillation, 1.0 for triplet loss

For this initial experiment, a version of the SYSU-MM01 dataset was obtained from Kaggle and Google Drive repositories, as the official dataset required formal approval from the original authors. This version maintained the same identities and image structure as the official release but may have had minor differences in preprocessing. The official dataset was later obtained for all subsequent experiments.

3.3.2. Implementation decisions and training configuration

Several decisions were made during implementation to align the training procedure with available hardware and to ensure compatibility between the original MACE codebase and the current software environment.

Batch size and gradient accumulation. The MACE architecture simultaneously maintains four classifier branches and computes a triplet loss over combined visible and infrared feature batches,

resulting in substantially higher GPU memory demand than a single-stream network. GPU memory profiling indicated that the original batch size of 32 images per modality exceeded available memory during the multi-branch forward pass. A batch size of 8 images per modality was therefore selected, with gradient accumulation over 4 steps to maintain an effective batch size of 32 consistent with the original training procedure [1]. This decision directly affects triplet mining quality, as larger effective batches provide more hard-negative candidates; gradient accumulation is a standard technique to recover this property under memory constraints.

Software compatibility. The original MACE codebase was written for an earlier PyTorch version. Deprecated API calls were updated to match PyTorch 2.6.0, and tensor dimension conventions in the data augmentation pipeline were adjusted to match current torchvision behaviour. These modifications are functional adaptations with no impact on the mathematical computation performed by the model.

Dataset preprocessing. The SYSU-MM01 version obtained initially from a public repository used a different directory structure from what the original data loader expected. Preprocessing scripts were written to reorganize the files into the standard layout, ensuring that identity labels, camera indices, and train/test splits matched the official protocol defined by Wu et al. [2].

3.3.3. Preliminary results

After resolving the technical challenges, a complete 40-epoch training run was successfully completed. The trained model was then evaluated on the SYSU-MM01 test set according to the standard evaluation protocol described in Section 2.5. Figure 5 shows MACE code snippet and how the training was logged in AI Jupyter Notebook environment, this is just visualization of the process just to have an overview of how the code was running and used.

Table 3. SYSU-MM01 results (All-Search mode), experiment 0

Metric	Rank-1	Rank-5	Rank-10	Rank-20	mAP
Initial implementation	45.59%	75.19%	85.84%	93.44%	43.25%

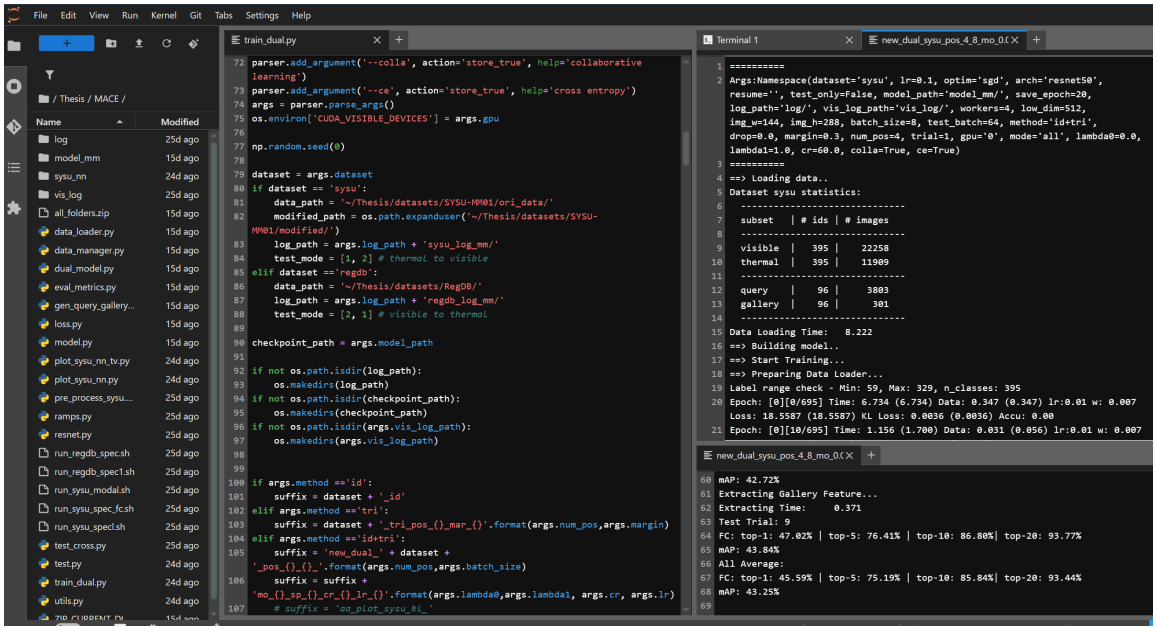

Figure 5. Visualization of MACE code snippet and logging of each epoch, as well, metrics results

These results constitute the experimental baseline for all subsequent comparisons in this thesis. As documented in Section 3.3.2, the implementation required several non-trivial adaptations relative to the original codebase: batch size was reduced due to GPU memory constraints, API calls were updated for PyTorch 2.6.0 compatibility, and the dataset directory structure was restructured to match the data loader. Because the original MACE code could not be executed without modification, a direct numerical comparison with the figures reported in [1] would conflate reproduction differences with the effects of the architectural modifications under investigation. All improvements and regressions reported in Experiments 1-3 are therefore measured relative to this reproduced implementation, ensuring a consistent and internally valid comparison throughout the experimental program.

Nevertheless, the primary goal of Experiment 0 was achieved: the university's AI environment was successfully configured for cross-modality Re-ID research, the MACE implementation was made functional, and preliminary results demonstrated that the approach could achieve reasonable performance even under non-optimal conditions.

3.3.4. Conclusions of experiment 0

The initial experiment provided several important outcomes that will inform subsequent research. First, the university's AI environment with the H100 GPU is capable of training complex cross-modality Re-ID models, though careful memory management is required. Furthermore, the adapted implementation of MACE is functional and produces reasonable results, providing a solid foundation for further experiments. Also, the preliminary results (45.59% Rank-1, 43.25% mAP) establish a baseline upon which improvements can be measured. The experiment also identified several aspects requiring attention, including the need for official datasets to ensure valid comparisons, optimization of hyperparameters for the specific hardware setup, potential for longer training to achieve convergence, and opportunities for architectural modifications as outlined in the experimental plan.

This initial experiment has successfully laid the groundwork for the comprehensive investigation planned in Section 2.2, demonstrating the feasibility of the approach and identifying concrete paths forward for improving cross-modality person Re-ID performance.

3.4. Experiment 1: Testing different backbones

The first experiment investigates the impact of replacing the baseline ResNet-50 backbone with alternative architectures that incorporate modern design principles. The hypothesis underlying this experiment is that architectural innovations such as squeeze-and-excitation (SE) blocks, efficient compound scaling, or GPU-optimized operations could improve feature extraction for cross-modality person re-identification. Three alternative backbones were evaluated: SE-ResNet-50, EfficientNet-B3, and TResNet-M.

3.4.1. SE-Resnet50 implementation

SE-ResNet-50 extends the standard ResNet-50 architecture by inserting Squeeze-and-Excitation blocks after each residual layer. The SE block performs channel-wise recalibration by first applying global average pooling to squeeze spatial information into a channel descriptor, then passing this descriptor through two fully connected layers with a reduction ratio of 16 to learn channel interdependencies and finally applying sigmoid activation to produce channel attention weights. These weights are used to rescale the original feature maps, allowing the network to emphasize informative channels while suppressing less useful ones.

In addition to SE blocks, the implementation replaced the standard average pooling with Generalized Mean (HAT) pooling, which learns an optimal pooling strategy between average and max pooling through a learnable parameter. This combination was expected to improve feature discriminability by adaptively weighting channel importance.

Training process and results. The SE-ResNet-50 model was trained under identical conditions to the baseline. Despite the channel recalibration introduced by SE blocks and the more flexible pooling strategy of GeM, the resulting performance was essentially unchanged from the baseline, as reported in Table 4. The addition of SE attention and GeM pooling offered no measurable benefit, suggesting that the global channel weighting these components provide does not address the source of error in cross-modal feature matching.

3.4.2. EfficientNet-B3 implementation

EfficientNet-B3 represents a fundamentally different architectural philosophy based on compound scaling, which uniformly scales network depth, width, and resolution using fixed ratios derived through neural architecture search. The architecture employs Mobile Inverted Bottleneck Convolution (MBConv) blocks with depthwise separable convolutions and integrated SE attention, offering a favourable accuracy-to-computation trade-off.

The implementation required adapting the EfficientNet-B3 stem and feature extraction blocks to the dual-stream MACE framework. The visible and thermal modules each utilized the EfficientNet stem (a single fused convolution layer), while the shared backbone employed the remaining EfficientNet blocks. GeM pooling was applied before the final embedding layer. The architecture produces 1536-dimensional features compared to 2048 dimensions in ResNet-50.

Training process and results. EfficientNet-B3 converged more slowly than ResNet-based models and plateaued noticeably earlier in training, with evaluation performance stabilizing around epoch 20 and showing no meaningful subsequent improvement despite continued loss reduction. The final retrieval results represent a substantial drop relative to the baseline (Table 4), making EfficientNet-B3 the weakest-performing architecture in this experiment.

Several factors may explain this underperformance. First, EfficientNet was designed and pretrained primarily for ImageNet classification with natural images, and its learned features may not transfer effectively to the cross-modality Re-ID domain where matching between visible and infrared images requires modality-invariant representations. Second, the smaller feature dimension (1536 vs. 2048) may limit the model's capacity to encode the complex relationships needed for cross-modality matching. Third, the compound scaling approach optimized for classification may not align with the metric learning objectives central to person re-identification.

3.4.3. TResNet-M implementation

TResNet-M (Transferable ResNet-Medium) incorporates several GPU-optimized modifications to the standard ResNet architecture. The most distinctive feature is the SpaceToDepth stem, which replaces the conventional 7x7 convolution and max pooling with a spatial rearrangement operation that converts spatial dimensions into channel depth. For an input of size [B, 3, 288, 144] with block size 4, the SpaceToDepth operation produces output of size [B, 48, 72, 36], effectively performing 4x downsampling through a single reshape operation rather than strided convolutions.

Additional TResNet enhancements include anti-alias downsampling layers that apply Gaussian blur before strided operations to improve shift invariance, SE blocks after each residual layer for channel attention, and GeM pooling for adaptive feature aggregation. The implementation utilized pretrained ResNet-50 weights for the backbone layers while training the SpaceToDepth stem from scratch.

Training process and results. TResNet-M training presented initial challenges due to the from-scratch training of the SpaceToDepth stem. Early experiments with a fully custom backbone (without pretrained weights) showed extremely slow learning with Rank-1 accuracy below 2% after several epochs. The solution adopted was to combine the SpaceToDepth stem with pretrained ResNet-50 backbone layers, allowing the stem to learn while leveraging established feature representations.

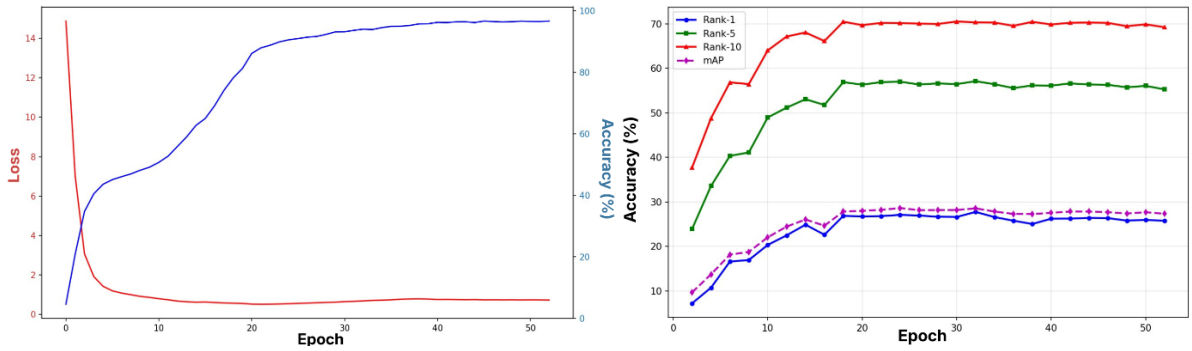

Figure 6. TResNet-M training progress and evaluation metrics change during training

After adopting the hybrid initialization approach, training loss decreased consistently and training accuracy reached a reasonable level by epoch 40. Evaluation performance, however, plateaued early and remained the lowest among all architectures tested in this experiment (Table 4), with no improvement after the first 20 epochs.

The poor performance can be attributed to the domain mismatch introduced by the SpaceToDepth stem. While this operation is GPU-efficient for general image classification, it fundamentally alters the spatial relationships in the input image by interleaving pixels from different spatial locations into the channel dimension. For person re-identification, where fine-grained spatial details (clothing patterns, body proportions, carried items) are critical for discrimination, this spatial shuffling may destroy important local features before they can be processed by subsequent layers. The pretrained backbone layers, having learned to expect features from a conventional stem, may be unable to effectively process the rearranged representations from SpaceToDepth.

3.4.4. Conclusions of experiment 1

Table 4. Comparative results of different backbones

Architecture	Rank-1	Rank-5	Rank-10	mAP	Training time	Parameters	Best epoch
SE-ResNet-50	45.54%	74.02%	85.17%	44.74%	~2h	49.3M	22
EfficientNet-B3	36.84%	66.71%	79.07%	37.70%	~7h	52.1M	46
TResNet-M	27.71%	57.09%	70.31%	28.52%	~7h	51.8M	32

The backbone experiments demonstrate that architectural innovations successful in general image classification do not necessarily transfer to cross-modality person re-identification. The baseline ResNet-50 architecture, while simpler, provides robust features that align well with the MACE framework's modality alignment objectives. SE-ResNet-50 shows that channel attention does not

provide marginal benefit when applied globally, despite the fact that attention mechanisms usually can be more effective when applied strategically at specific points in the cross-modality pipeline. The significant underperformance of EfficientNet-B3 and TResNet-M indicates that architectures optimized for efficiency or GPU throughput may sacrifice the spatial precision required for fine-grained person matching across modalities.

3.5. Experiment 2: Testing three attention mechanisms

The second experiment investigates the integration of explicit attention mechanisms into the MACE framework. Unlike the global SE blocks evaluated in Experiment 1, this experiment focuses on attention modules applied at strategic locations: within the modality-specific stems and before the final pooling layer. Three attention mechanisms were implemented and evaluated: Channel Attention Module (CAM), Spatial Attention Module (SAM), and Dual Attention Module (DAM) combining both.

3.5.1. Channel attention module (CAM)

The Channel Attention Module focuses on learning "what" features are meaningful by recalibrating channel-wise responses. Unlike the SE block which uses only average pooling, CAM employs both global average pooling and global max pooling to capture richer channel statistics. The pooled features are passed through a shared two-layer MLP, and the outputs are summed before applying sigmoid activation to produce channel attention weights.

CAM was integrated at two locations: immediately after the visible and thermal stems to allow early modality-specific channel weighting, and before the final pooling layer to refine the shared feature representation. The hypothesis was that channel attention could help the network identify which feature channels are most informative for cross-modality matching, potentially suppressing modality-specific channels while enhancing modality-invariant ones.

Training process and results. Training converged smoothly, with loss and accuracy curves following a similar trajectory to the baseline. The evaluation metrics curve shows an initial improvement phase, followed by a period of instability around the learning rate transition, after which performance settled at a stable level below the baseline. Best-epoch results are reported in Table 5.

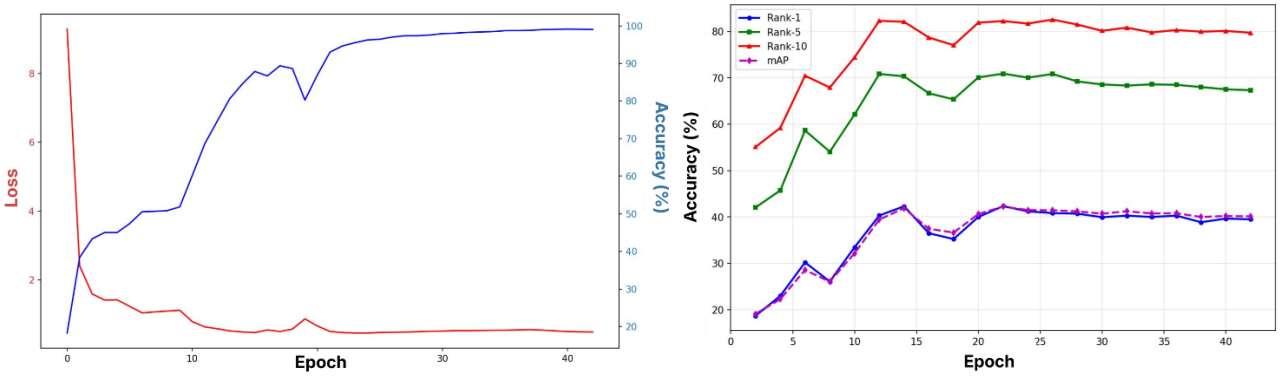

Figure 7. CAM training progress and evaluation metrics change during training

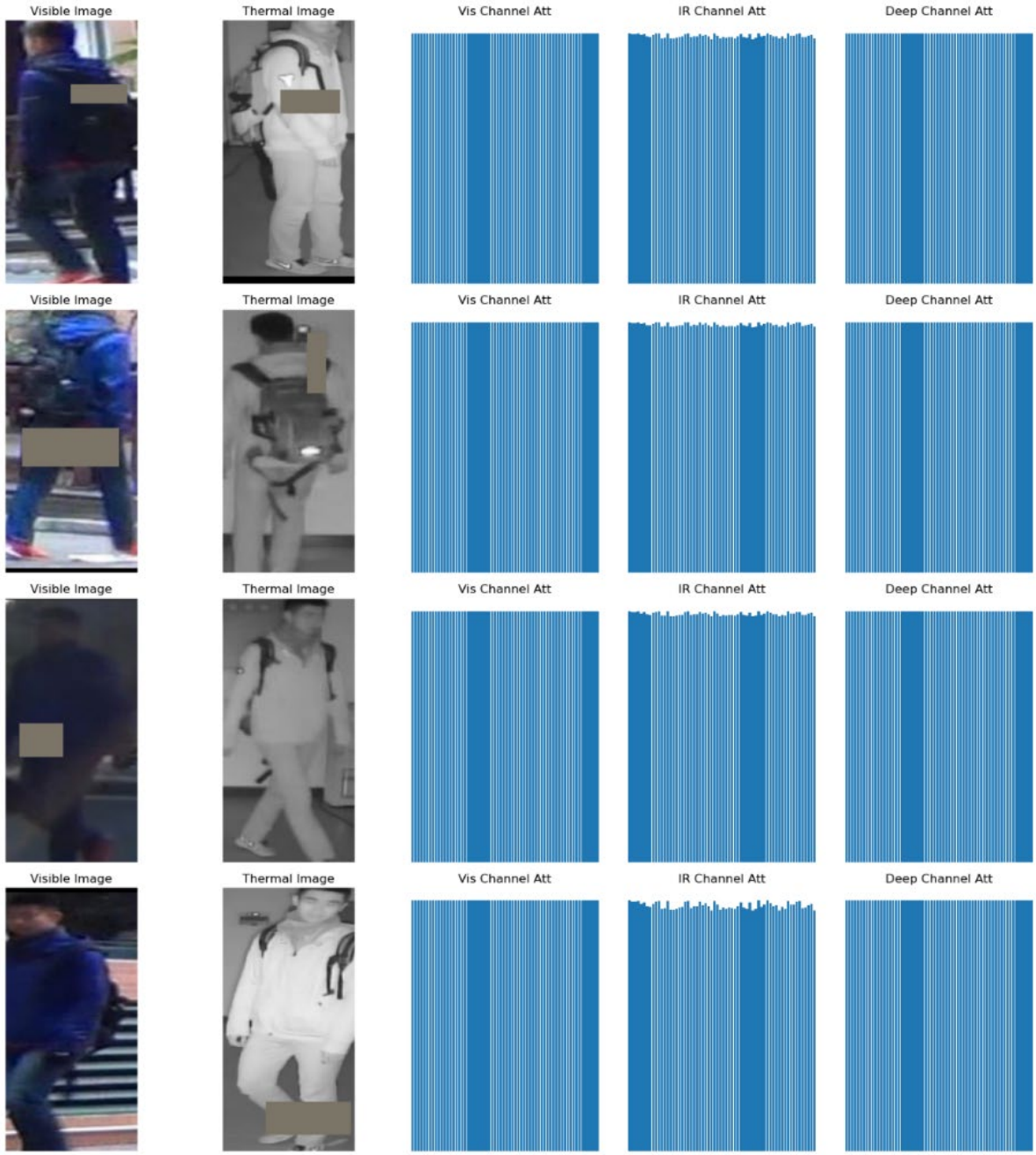

Figure 8. Attention map of CAM at epoch 40

Attention visualization analysis. To understand the behaviour of the learned attention, channel attention maps were visualized at epoch 40 when the model had reached a stable state. The visualization displays the attention weights for the first 64 channels as bar charts.

Figure 8 reveals a concerning pattern that the attention weights appear nearly uniform across all channels, with minimal variation between channels. This suggests that the CAM module failed to learn discriminative channel weighting and instead converged to an approximately uniform attention distribution. In an effective channel attention mechanism, one would expect to see significant variation, with some channels receiving high weights, indicating importance for the task, and others receiving low weights, indicating redundancy or noise.

The uniform attention pattern indicates that CAM may be facing an optimization challenge in the cross-modality setting. When features from visible and infrared modalities must be aligned, the notion of

"important channels" becomes ambiguous because different channels may be important for different modalities. The network may have learned to hedge by weighting all channels similarly rather than making strong channel selection decisions that could harm one modality's representation.

3.5.2. Spatial attention module (SAM)

The Spatial Attention Module focuses on learning where to attend by generating a spatial attention map that highlights important regions. SAM first applies channel-wise average and max pooling to produce two spatial descriptors, concatenates them, and applies a 7x7 convolution followed by sigmoid activation to produce a 2D attention map with values between 0 and 1.

SAM was integrated at the same locations as CAM: after the modality-specific stems and before final pooling. The motivation for spatial attention in cross-modality Re-ID is that certain body regions (torso, distinctive clothing areas) may be more reliable for matching across modalities than others (face, which appears very different in IR vs. visible).

Training process and results. ResNet50-SAM followed a broadly similar training trajectory to CAM. A characteristic fluctuation in the loss curve appeared around epochs 15 to 20, coinciding with a visible dip in evaluation metrics before both recovered and stabilized. The final best-epoch performance (Table 5) is marginally better than CAM but still falls below the baseline.

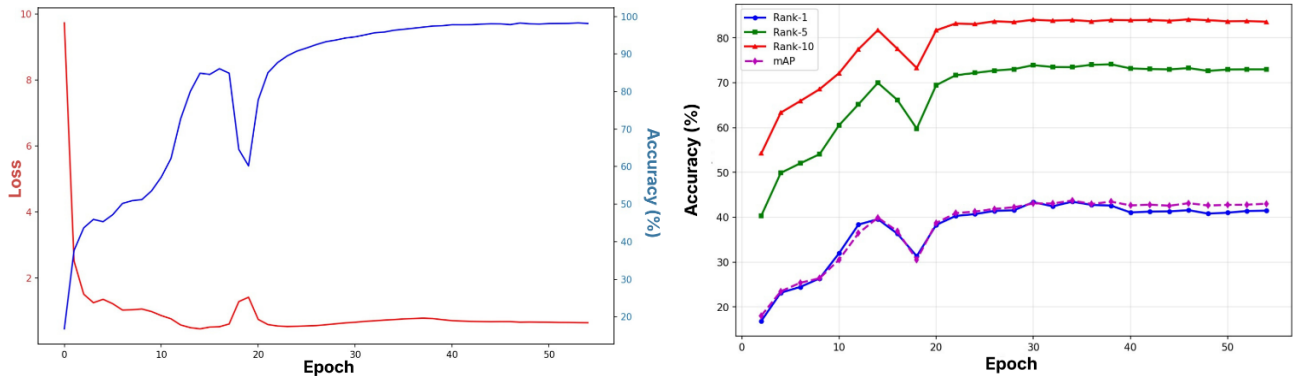

Figure 9. SAM training progress and evaluation metrics change during training

Attention visualization analysis. The spatial attention maps at epoch 50 provide insight into what regions the model learned to emphasize.

Figure 10 shows that SAM learned to attend broadly to the person region while suppressing background areas. The attention maps appear as mostly uniform high values (red/orange) across the person silhouette, with lower values only at the extreme edges of the images. This behaviour, while reasonable, is essentially learning a coarse person segmentation rather than fine-grained discriminative region selection.

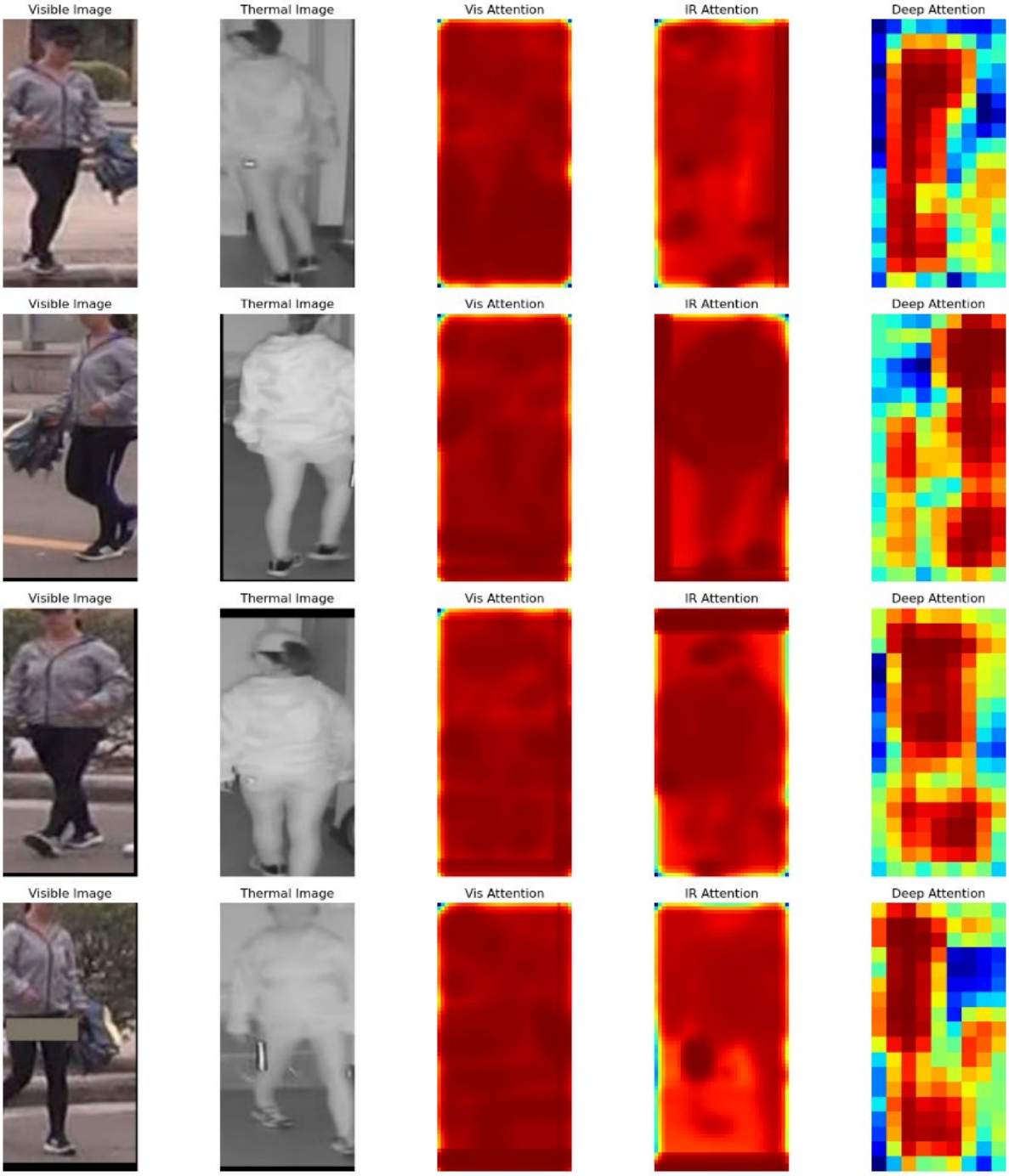

Figure 10. Attention map of SAM at epoch 50

For effective person re-identification, one would expect the attention to highlight specific discriminative regions such as distinctive clothing patterns, carried objects, or body proportions that help distinguish between individuals. The observed uniform attention over the entire person suggests that SAM failed to learn identity-discriminative spatial weighting. The attention mechanism may be primarily learning to distinguish foreground from background rather than identifying which foreground regions are most useful for person matching.

The similarity between visible and infrared attention maps is encouraging as it indicates some degree of modality-invariant spatial attention learning. However, the lack of fine-grained spatial discrimination limits the potential benefit of the attention mechanism.

3.5.3. Dual attention module (DAM)

The Dual Attention Module combines channel and spatial attention in a sequential manner, following the Convolutional Block Attention Module (CBAM) [28] architecture. Features first pass through the Channel Attention Module, which recalibrates channel responses, then through the Spatial Attention Module, which applies spatial weighting. This sequential combination allows the network to first determine "what" is meaningful, then "where" to focus.

DAM was expected to provide complementary benefits by combining the strengths of both attention types. The channel attention could identify modality-invariant feature channels, while the spatial attention could highlight discriminative body regions.

Training process and results. DAM training dynamics closely resembled those of the individual attention modules, including the characteristic dip around epochs 18 to 20. The combined module did not produce better evaluation results than either individual attention type (Table 5), indicating that stacking channel and spatial attention sequentially does not recover the discriminability lost by either component in isolation.

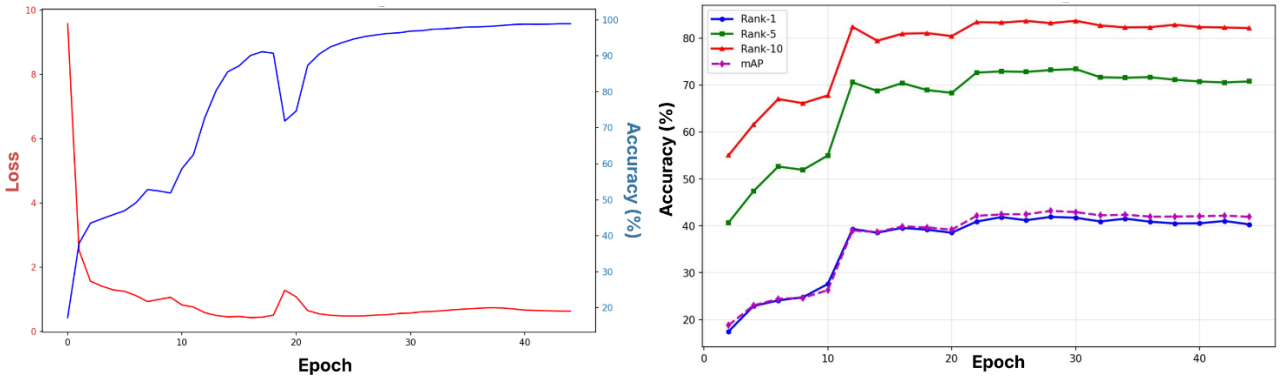

Figure 11. DAM training progress and evaluation metrics change during training

Attention visualization analysis. The DAM attention visualizations at epoch 40 display both spatial and channel attention maps.

The spatial attention maps from DAM (figure 12) show more defined patterns compared to standalone SAM, with some images showing attention concentrated on specific regions such as the upper body or torso area. This suggests that the channel attention preprocessing helps the subsequent spatial attention learn more focused patterns. The channel attention bars, however, show the same near-uniform distribution observed in standalone CAM.

An interesting observation is that the deep spatial attention, applied before final pooling on the 18x9 feature maps, shows more structured patterns than the stem-level attention, with clear vertical bands corresponding to different body parts. This indicates that spatial attention may be more effective at later stages of the network where features are more semantically meaningful.

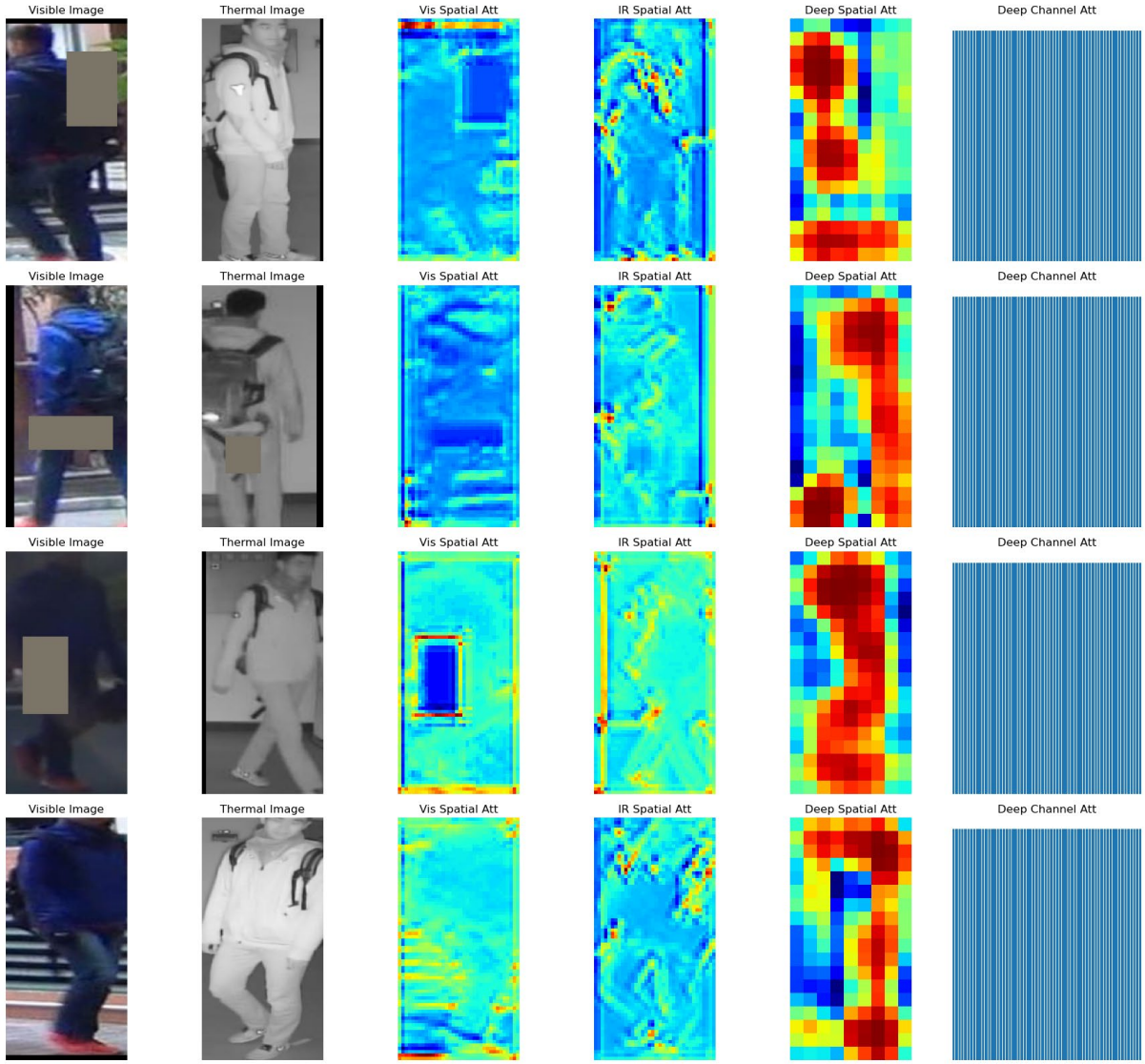

Figure 12. Attention map of DAM at epoch 40

3.5.4. Conclusions of experiment 2

Table 5. Comparative results of attention mechanisms

Architecture	Rank-1	Rank-5	Rank-10	mAP	Best epoch
ResNet50-CAM	42.33%	70.89%	82.22%	42.24%	22
ResNet50-SAM	43.47%	73.47%	83.96%	43.73%	34
ResNet50-DAM	41.89%	73.18%	83.17%	43.16%	28

The attention mechanism experiments reveal that simply adding attention modules to an already well-designed framework does not guarantee improvement. All three attention variants performed below the baseline, with Rank-1 accuracy decreasing by 2-4 pp. The attention visualizations provide insight into why the learned attention patterns are either too uniform (channel attention) or too coarse (spatial attention) to provide discriminative benefit.

Several factors may contribute to this outcome. First, the MACE framework already incorporates implicit attention through its modality-specific and modality-shared learning branches, and explicit attention modules may interfere with this learned specialization. Second, the cross-modality nature of

the task makes attention learning challenging because the important features differ between visible and infrared modalities, leading to attention patterns that compromise between modalities rather than optimize for either. Third, the additional parameters introduced by attention modules may increase overfitting risk without providing commensurate representational benefit.

3.6. Experiment 3: Feature fusion strategy refinement

The third experiment investigates alternative feature fusion strategies for combining modality-specific representations. The baseline MACE framework uses simple concatenation followed by batch normalization to combine visible and infrared features. This experiment explores whether more sophisticated fusion mechanisms could better leverage complementary information between modalities. Two fusion strategies were implemented: Adaptive Weighted Fusion and Feature Transform Fusion.

3.6.1. Adaptive weighted fusion

Adaptive weighted fusion learns to dynamically balance the contribution of each modality based on the input features. The implementation includes learnable base weights for each modality plus a gate network that predicts sample-specific fusion weights. The gate network takes the concatenated visible and infrared features as input, passes them through a two-layer MLP, and applies softmax to produce normalized weights. The final fused representation is a weighted combination of the two modality features.

The motivation for adaptive weighting is that different image pairs may benefit from different modality emphasis. For example, when the visible image has poor lighting, the infrared features might be more reliable and should receive higher weight, while well-lit visible images might provide more discriminative colour information.

Training process and results. Training with adaptive fusion converged reliably, with the fusion loss component decreasing alongside the main identity and triplet objectives. Analysis of the learned fusion weights showed that the gate network settled on nearly equal contributions from both modalities for the large majority of samples, indicating that the gating mechanism did not learn meaningful sample-dependent preferences. Evaluation results are summarized in Table 6.

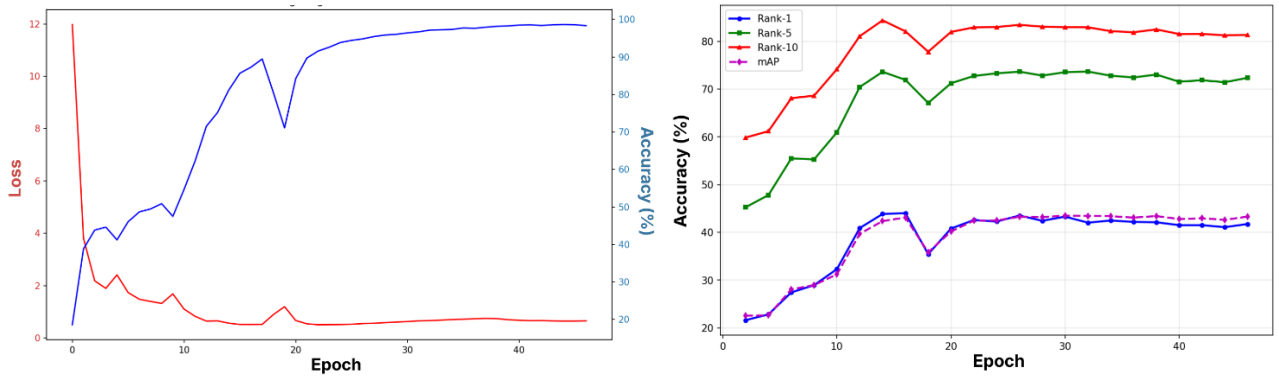

Figure 13. Adaptive weighted fusion training progress and evaluation metrics change during training

3.6.2. Feature transform fusion

Feature transform fusion applies non-linear transformations to each modality's features before combination, with the goal of projecting both modalities into a common representation space where

they can be more effectively fused. The implementation uses separate transformation networks for visible and infrared features, each consisting of a two-layer MLP with batch normalization and ReLU activation, followed by a residual connection. The transformed features are concatenated and passed through a fusion layer.

The hypothesis underlying this approach is that raw features from different modalities may not be directly comparable due to the domain gap. By learning modality-specific transformations, the network could potentially align the feature distributions before fusion, similar to domain adaptation techniques.

Training process and results. Feature transform fusion trained stably and converged within a similar number of epochs to adaptive fusion. The additional transformation layers did not improve retrieval, and t-SNE analysis of the transformed feature distributions suggested that the projection reduced inter-class distance alongside intra-class distance, compressing the embedding space in a way that lowered overall discriminability (Table 6).

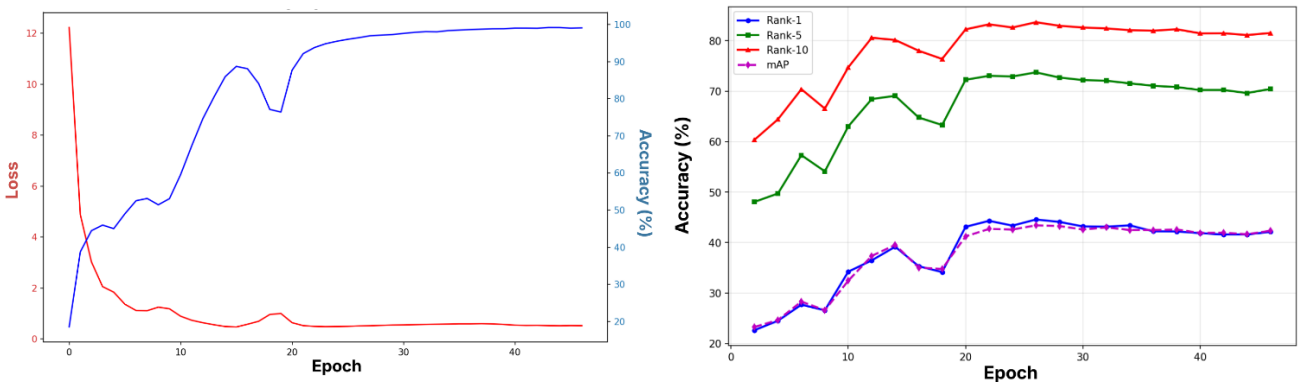

Figure 14. Feature transform fusion training progress and evaluation metrics change during training

3.6.3. Conclusions of experiment 3

Table 6. Comparative results of fusion strategies

Fusion strategy	Rank-1	Rank-5	Rank-10	mAP	Best epoch
Adaptive weighted	44.02%	71.94%	82.12%	43.10%	16
Feature transform	44.57%	73.73%	83.67%	43.42%	26

The two fusion strategies evaluated in Experiment 3 produced modest decreases in retrieval performance relative to the baseline ResNet-50. Both strategies converged reliably and without training instability. The marginal difference between the two fusion approaches suggests that the fundamental challenge in cross-modal matching lies upstream, in the feature extraction and modality alignment stages, rather than in how the final representations are combined. The MACE baseline's knowledge distillation mechanism already acts as an implicit feature alignment step, and the additional fusion modules may introduce redundant transformation without a commensurate benefit.

3.7. Experiment 4: Multi-query fusion

Standard cross-modality Re-ID evaluation protocols assume a single query image per probe identity [2], which is a practical simplification that does not reflect how surveillance systems actually operate. In a real deployment, the infrared cameras covering a scene will typically capture multiple images of the same person before a matching decision is required. Aggregating these observations before

computing distances against the visible gallery is a purely inference-side operation requiring no changes to the trained model. This experiment examines whether this approach can recover meaningful performance from MACE's already-trained representations, and which aggregation strategy best consolidates a pool of query features into a single representative vector.

The experiment uses the trained baseline checkpoint evaluated across 10 random gallery trials under the standard SYSU-MM01 all-search protocol. Three aggregation strategies are tested at query pool sizes n = {2, 5, 10, 15, 20}. The query set contains 3,803 infrared images across 96 test identities, with an average of 39.6 images per identity, making all tested pool sizes feasible for the majority of identities.

Three aggregation strategies are these: average fusion computes the element-wise mean of all n query feature vectors, producing a centroid embedding before distance computation against the gallery; maximum fusion takes the element-wise maximum across all n query feature vectors, retaining the highest activation in each feature dimension from any query in the pool; distance-weighted fusion first computes the simple average of all n query features as a reference, then rescales each individual query's contribution by its cosine similarity to this reference before computing the final weighted average.

Analysis and results.

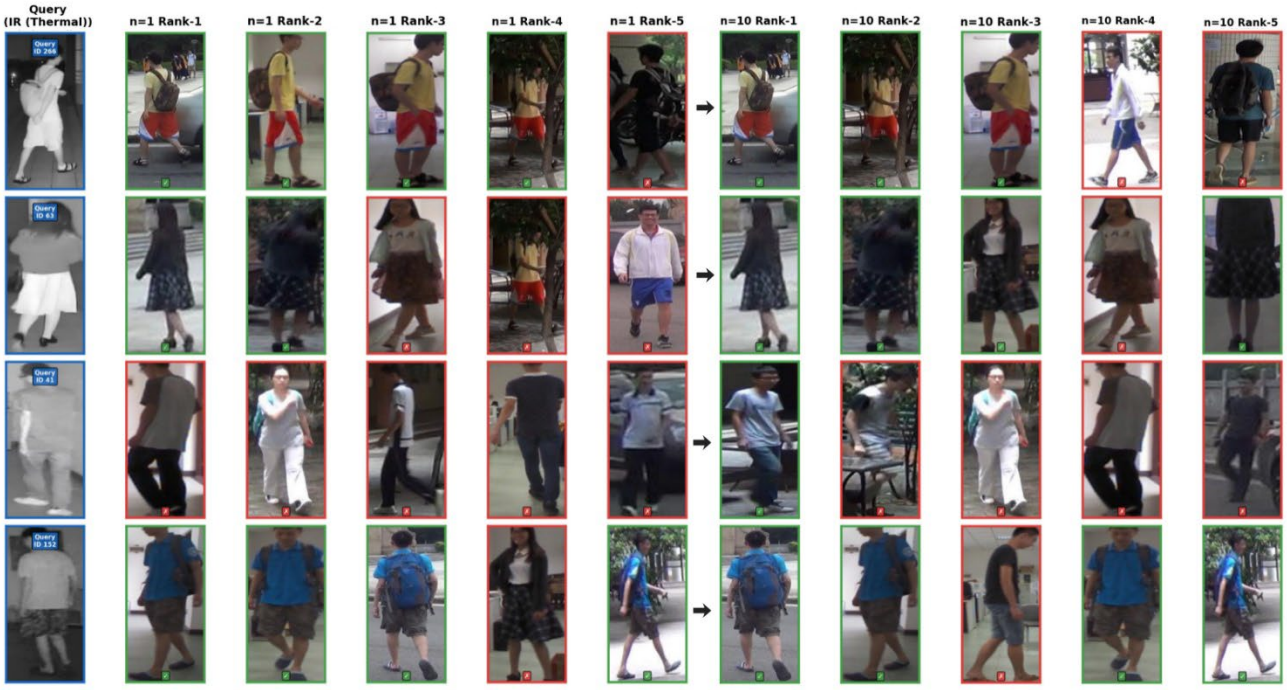

Figure 15. Qualitative retrieval comparison between single-query and multi-query average fusion for 4 randomly selected example identities

Figure 15 illustrates the core effect that a single infrared query can be captured at an ambiguous angle or with partial background interference, placing wrong gallery candidates above the true match. With 10 queries averaged, the representation stabilizes around the identity's consistent appearance characteristics, and correct visible-spectrum matches rise to the top positions.

Table 7. Multi-query fusion results on SYSU-MM01 (All-Search mode)

n queries	Rank-1			mAP		
	Average	Maximum	Weighted	Average	Maximum	Weighted
2	53.54%	51.15%	53.54%	49.11%	47.90%	49.11%
5	60.10%	49.69%	59.58%	54.79%	45.66%	54.64%
10	60.62%	43.96%	60.52%	55.81%	40.72%	55.70%
15	61.35%	40.10%	61.77%	56.76%	37.78%	56.83%
20	62.19%	36.98%	62.08%	57.28%	35.14%	57.24%

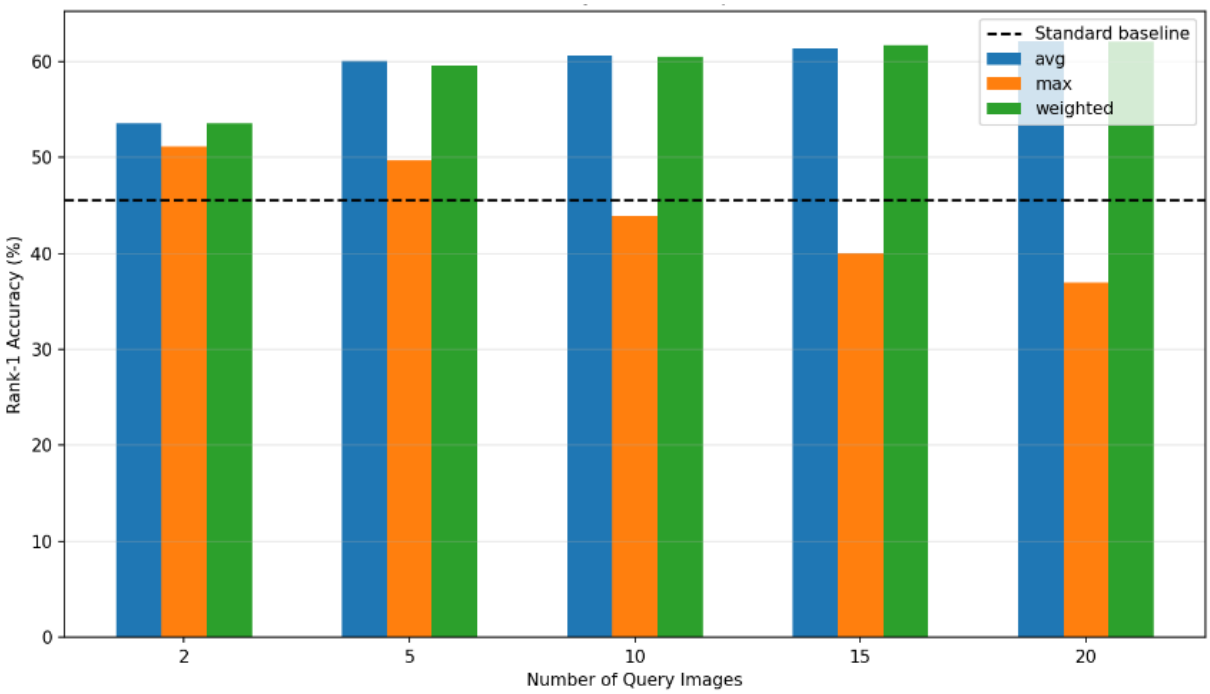

Figure 16. Rank-1 accuracy of average, maximum, and distance-weighted fusion strategies at different query pool sizes

The diverging trajectories of average and maximum fusion in Figure 16 reveal something fundamental about how MACE represents identity. Average fusion improves monotonically because MACE's joint identity and triplet training objective specifically shapes the embedding space so that feature vectors belonging to the same identity form a compact cluster. Averaging multiple samples moves the aggregated representation toward the centre of that cluster, suppressing the noise inherent in any individual frame. The more samples are averaged, the closer the result gets to the true identity centroid.

Maximum fusion behaves oppositely for the same reason. In a metric-learned embedding space, the most extreme activation in any dimension is not the most discriminative one it is the most unusual one. A frame where the subject is backlit, partially cut off, or captured mid-step will produce atypical activations in certain dimensions, and maximum pooling over a growing pool guarantees that such outlier activations accumulate and dominate the final representation. This strategy might benefit a shallow or noisy feature extractor, but it actively degrades a well-trained embedding.

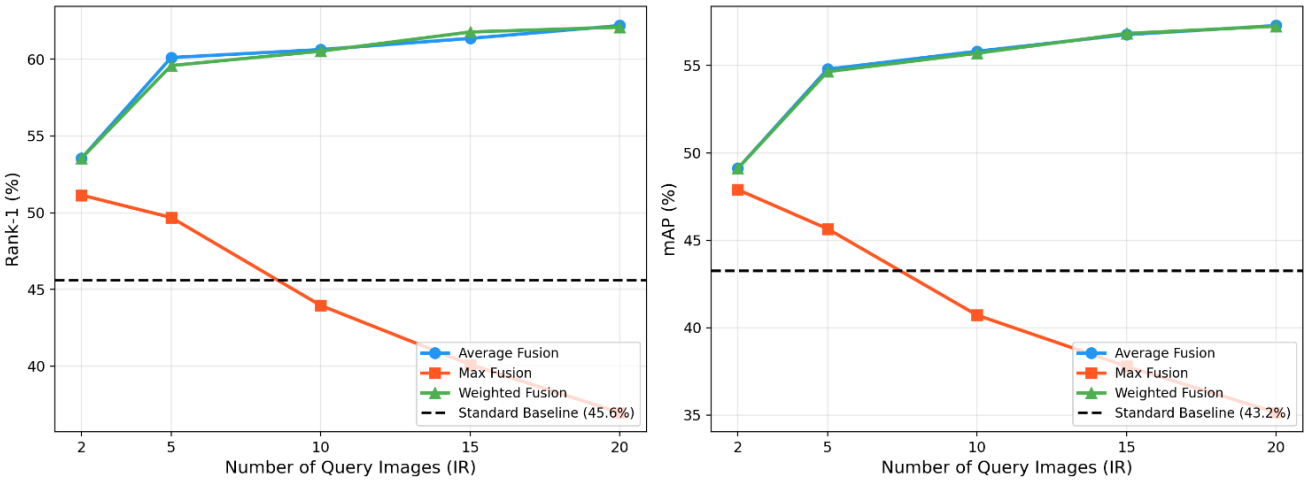

Figure 17. Rank-1 accuracy and mAP as a function of query pool size for each fusion strategy

Figure 17 shows that the largest gains from average fusion occur between n=2 and n=5, with diminishing returns beyond that point. This saturation indicates that five samples already capture the effective diversity of an identity's infrared appearance under MACE's learned representation. Adding further queries does not introduce new discriminative information, it only adds redundant copies of what the model has already seen. The practical implication is that five queries is the efficient operating point: beyond it, the cost of collecting additional observations is not matched by proportional retrieval gain.

Distance-weighted fusion produces results nearly identical to simple averaging throughout all tested pool sizes. This is not surprising given how MACE is trained because identity loss and triplet loss together minimize intra-class feature distances, embeddings for the same person are already highly consistent. When the embeddings within a pool are already tightly clustered, re-weighting by similarity to the centroid has almost no effect on the final vector.

A broader observation from these results concerns how single-query evaluation characterizes MACE's performance. The standard protocol, which underpins all comparisons in Experiments 0 through 3, evaluates the model under conditions where one potentially unrepresentative frame must carry the entire retrieval burden. The gap between the single-query baseline and five-query average fusion suggests that single-query metrics meaningfully underestimate the practical retrieval capability of the model when it is used as it would be in a real deployment.

3.7.1. Conclusions of experiment 4

Multi-query average fusion substantially improves MACE's cross-modality retrieval without any model modification or retraining. The improvement is not simply additive noise cancellation - it reflects the geometric structure of MACE's learned embedding space, where identity centroids are stable and accessible once enough samples are averaged. Five queries per identity captures most of the available gain and represents the practical operating point for deployment. Maximum fusion degrades with additional queries and should not be used when more than two observations are available. Distance-weighted fusion offers no advantage over plain averaging, making simple mean aggregation the recommended approach. The finding also has an implication for how single-query benchmark results should be interpreted: they represent a lower bound on what the trained model can achieve, not its operational ceiling.

3.8. Experiment 5: Data efficiency analysis

Cross-modality Re-ID systems require labelled paired data from both modalities, which is expensive to collect in practice. Knowing how performance scales with the number of available training identities directly informs deployment decisions: how much labelled data is actually needed to reach a useful operating level, and at what point collecting additional data stops being worthwhile. This experiment addresses that question by training separate MACE models from scratch at five data fractions (10%, 25%, 50%, 75%, and 100%) of the 395 available SYSU-MM01 training identities and evaluating each on the full test set under the standard all-search protocol.

Training process and results.

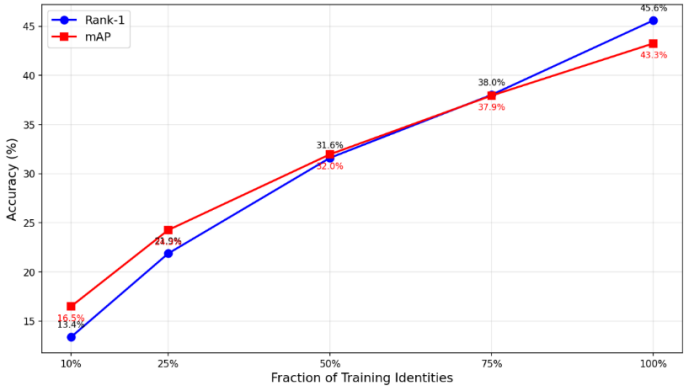

Figure 18. Best Rank-1 and mAP achieved at each training data fraction

Figure 18 shows that performance scales roughly linearly with data fraction across most of the range, but the curve is not simply proportional and the two extremes behave differently. Table 8 provides the full numerical summary.

Table 8. Best results per training data fraction

Data fraction	10%	25%	50%	75%	100%
Best Rank-1	13.4%	21.9%	31.6%	38.0%	45.6%
Best mAP	16.5%	24.3%	32.0%	37.9%	43.3%

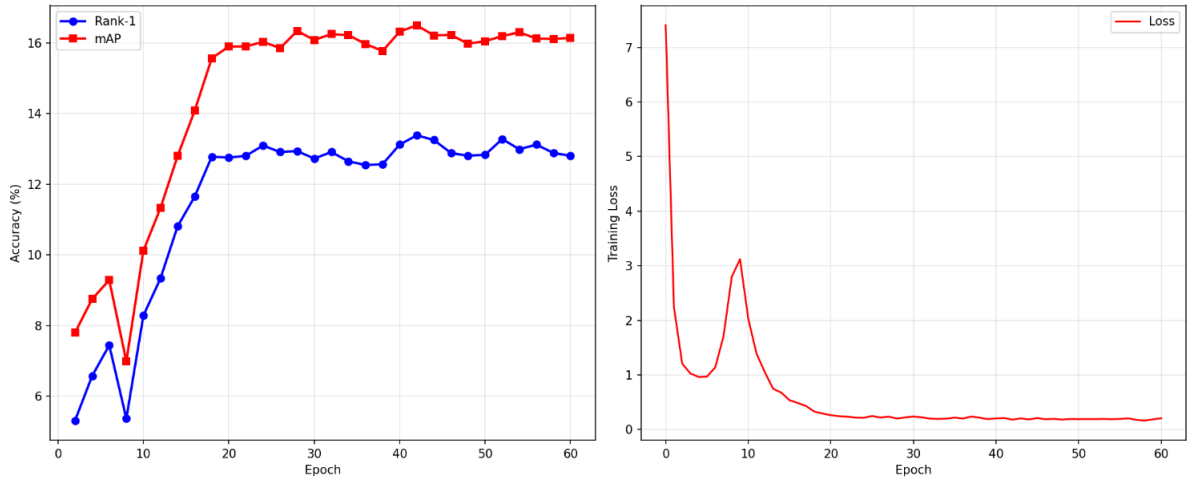

Figure 19. Evaluation metrics and training loss for the 10% data condition with 39 training identities

The 10% condition exposes a qualitative failure mode rather than just reduced performance. The evaluation curve in Figure 19 saturates around epoch 18 and shows no meaningful improvement across the remaining 40 epochs of training, despite the loss continuing to decrease. A spike in the training loss around epoch 8 10 coincides with the scheduled learning rate decay: with only 39 training identities, the identity classifier is severely underdetermined, and the sharper gradient updates at the decay transition cause temporary instability. More fundamentally, MACE's two core learning mechanisms both break down under such a narrow identity space. The triplet loss requires diverse within-batch negative pairs to learn a discriminative metric. With 39 identities, the hard negative mining pool is too small to drive meaningful embedding separation. The bidirectional knowledge distillation between modalities equally depends on the classifier logits having enough identity-class structure to be worth transferring across modalities. When neither mechanism has sufficient material to work with, the model memorizes the training identities without learning transferable representations.

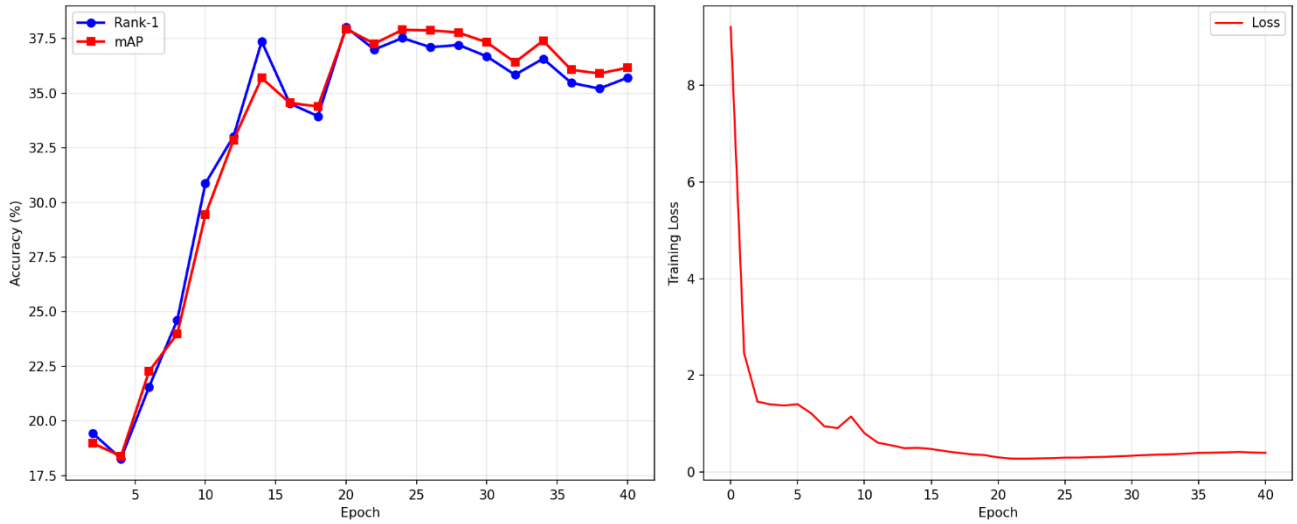

Figure 20. Evaluation metrics and training loss for the 75% data condition with 296 training identities

The 75% condition in Figure 20 shows the opposite pattern: the loss decreases smoothly with no instability, and the evaluation curve rises consistently until around epoch 14 before stabilizing. The absence of the loss spike confirms that the learning rate transition no longer destabilizes training once a sufficient number of identities are present. A mild evaluation decline after the peak (from epoch 20 onward) is visible and reflects the same overfitting tendency seen in experiments 1 and 2 the learning rate decay that drives the best epoch performance also makes the model more susceptible to overfitting on whatever training identities are available.

Analysis. Two structural observations follow from these results. First, the performance gap between 75% and 100% data (about 7.6 Rank-1 percentage points for adding only 99 identities) is disproportionately large relative to the gains in the lower fractions. This suggests that the final quarter of training identities in SYSU-MM01 contains appearance and scene diversity that is particularly relevant to the test set. The training and test identities are not uniformly matched in difficulty, and the identities present only at higher data fractions cover challenging cases that the test queries require.

Second, the convergence epoch decreases with data size: the 10% model peaks at epoch 42 (still essentially flat since epoch 18), the 25% model at epoch 16, the 75% model at epoch 20, and the full data model at epoch 32. At first this appears contradictory more data should take longer to learn from. The explanation lies in what the model is optimizing against: with few identities, training loss reaches

near-zero quickly because the classifier overfits, but the evaluation performance has already stalled. With more identities, the model continues to find informative training signal at later epochs and therefore peaks later.

From a deployment perspective, the results indicate that approximately 200 identities (50% of the training set) is a practical minimum below which the model's cross-modal alignment mechanisms become unreliable. Above that threshold, performance scales reasonably with additional data up to the full set.

3.8.1. Conclusions of experiment 5

MACE's retrieval performance degrades gradually but without a single catastrophic cutoff as training data is reduced. However, the 10% condition reveals a qualitative change in learning behaviour rather than just a quantitative reduction: with too few training identities, the mechanisms that make MACE work hard negative triplet mining and cross-modal knowledge distillation lose the identity diversity they depend on, and the model fails to learn transferable representations regardless of how long training continues. Approximately 200 training identities represents the threshold below which this qualitative failure mode begins to emerge. Above 50% of the training set, the model operates in a regime where adding more data produces predictable and consistent improvement, making it feasible to estimate the cost-benefit of additional data collection for a target deployment accuracy.

3.9. Experiment 6: Synthetic occlusions

Person re-identification in real surveillance environments rarely produces clean, unobstructed query images. People walk behind barriers, partially past frame edges, or through crowds, leaving the IR camera with only a portion of the body visible. This experiment asks two related questions: how much does partial occlusion degrade MACE's cross-modal retrieval performance, and can occlusion-aware training reduce that vulnerability?

Four synthetic occlusion types are applied to IR query images at evaluation time, each blocking 20 to 40 percent of the image area: a randomly placed rectangle, a horizontal band spanning the full image width, a fixed upper body block covering the torso and head region, and a fixed lower body block covering the legs. Figure 21 shows representative examples of each type on IR images from the SYSU-MM01 test set.

The experiment proceeds in two stages. First, the trained baseline MACE model is evaluated directly on occluded queries without any retraining, to establish how sensitive the unaugmented model is to each occlusion type. Second, the model is retrained from scratch with random rectangle and horizontal band occlusion applied stochastically during training at 50% probability, and the results on both clean and occluded test sets are measured against the baseline.

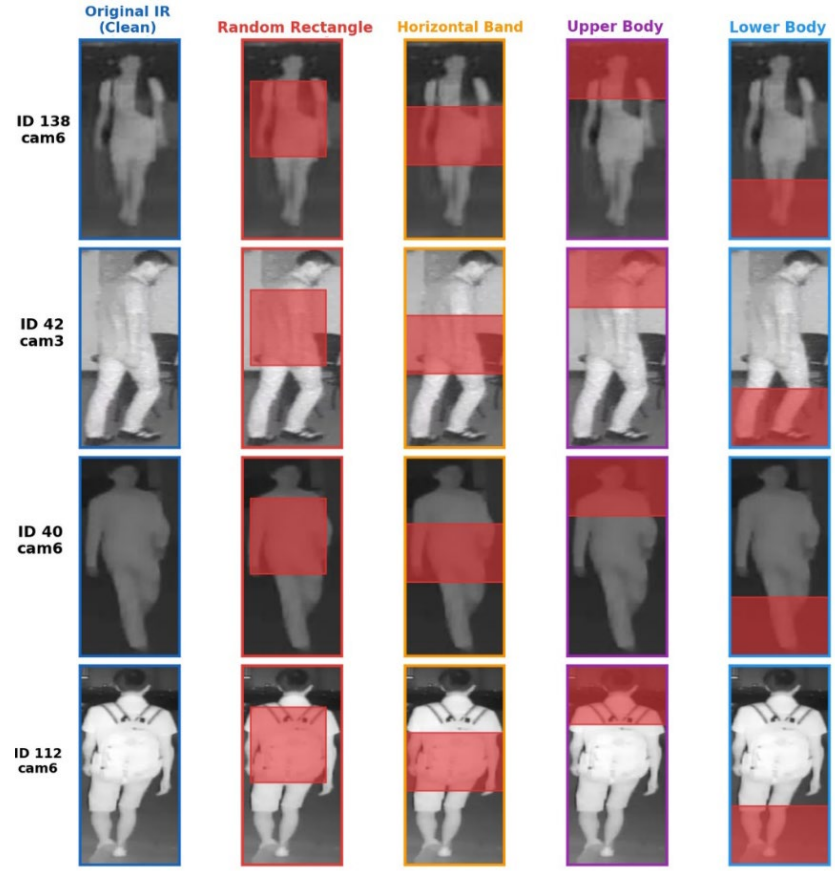

Figure 21. Four synthetic occlusion types applied to IR (thermal) query images

Baseline vulnerability. Applying occlusion at test time without any model adaptation reveals consistent and substantial performance drops across all four conditions. The severity follows a clear ordering, shown in Figure 22, that reflects which body regions MACE relies on most for cross-modal identity matching.

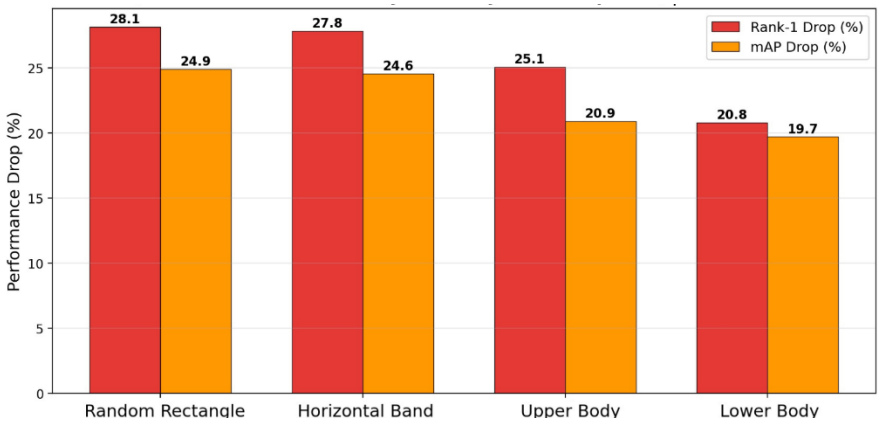

Figure 22. Rank-1 accuracy and mAP performance drop of the baseline MACE model when tested on synthetically occluded queries for all four occlusion types

Lower body occlusion is the least damaging of the four types. When legs are hidden, the model retains access to the torso and upper clothing area, which carries the most discriminative cross-modal signal in near-IR imagery. Thermal imaging produces a strong response over the body core and upper garment region, making those zones the primary anchor for identity matching. The legs, by contrast, present more homogeneous thermal profiles across identities and contribute less to the retrieval decision.

Upper body occlusion is correspondingly more harmful, and this asymmetry between the two structured types is consistent across both Rank-1 and mAP. Random rectangle and horizontal band produce the largest drops because neither is confined to a low-information region. A horizontal band cutting across the mid-torso disrupts the spatial feature maps that ResNet50 builds continuously across the full image height, and a randomly placed rectangle can land anywhere, with meaningful probability of covering the most discriminative zone.

Figure 23 illustrates these effects on a single probe identity. Under clean conditions the model correctly places the true match at Rank-1. With horizontal band occlusion, which strips away the chest and waist area, none of the top five retrieved candidates correspond to the correct person. Under lower body occlusion, the correct identity remains in the top-five result set. This example reflects the consistent quantitative pattern across the test set.

Figure 23. Top-5 retrieval results for Identity 27 under clean and four occluded query conditions

Occlusion-augmented training. Retraining with occlusion augmentation substantially changes the picture. Figure 24 compares the baseline and occlusion-trained models on both clean and occluded test conditions for the two augmentation types.

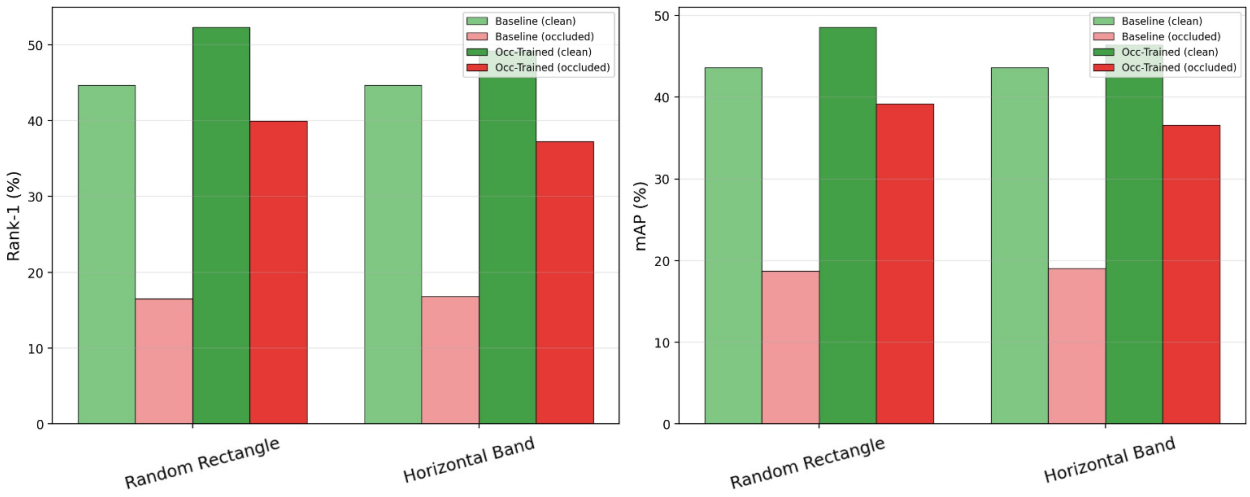

Figure 24. Rank-1 and mAP performance before (baseline) and after occlusion-augmented training, for random rectangle and horizontal band types

Two effects stand out. First, augmented training does not reduce clean performance, it improves it. The model trained with random rectangle augmentation achieves a notably higher clean Rank-1 than the original baseline, and the horizontal band model shows a similar clean improvement. This is a regularization benefit: when the model is required to identify people from only partial observations, it develops more spatially distributed representations that generalize better even on complete images.

Second, the gap between clean and occluded performance narrows substantially after augmented training, dropping to roughly half of the baseline gap for both occlusion types. The model is not fully immune to occlusion after augmented training, and a residual clean-to-occluded gap remains. This is expected because partial information inherently limits matching precision when the missing region would have been discriminative. What augmented training achieves is that the model no longer collapses under occlusion; it continues to extract meaningful partial-body features and retrieve plausible matches.

Training dynamics for both occlusion-augmented runs follow the same general pattern as the baseline: rapid metric improvement through the first 20 epochs, followed by stabilization around epoch 28 to 34. No additional training instability is introduced by the augmentation, and no extra epochs are required to converge. Table 9 summarizes the best epoch results across all evaluated conditions.

Table 9. Occlusion robustness results

Model	Occlusion Type	Clean Rank-1 (%)	Occluded Rank-1 (%)	Rank-1 Drop (pp)	mAP Drop (pp)
Baseline	Random Rectangle	44.65	16.55	28.1	24.9
Baseline	Horizontal Band	44.65	16.85	27.8	24.6
Baseline	Upper Body	44.65	19.59	25.1	20.9
Baseline	Lower Body	44.65	23.85	20.8	19.7
Occ-Trained	Random Rectangle	52.35	40.36	12.0	9.3
Occ-Trained	Horizontal Band	49.25	37.18	12.1	9.6

3.9.1. Conclusions of experiment 6

MACE is sensitive to partial occlusion in the query image, with performance degrading substantially across all four tested types when the model is evaluated without any adaptation. The pattern of sensitivity is informative: upper body features carry more identity-relevant information in cross-modal IR-VIS matching than lower body features, and occlusion types that disrupt the torso or mid-body region are consistently more damaging. This confirms that MACE's feature representation is not spatially uniform across the image and concentrates its discriminative weight on the upper portion of the body.

Occlusion-augmented training is an effective and low-cost mitigation. Training with random rectangle or horizontal band masking applied stochastically at 50% probability reduces the clean-to-occluded performance gap to roughly half of what the baseline shows, without any penalty on clean retrieval. The same augmentation additionally improves clean performance through a regularization effect, indicating that exposure to partial observations encourages the model to build less spatially concentrated representations. For deployment in environments where partial occlusions are expected, including indoor surveillance with barriers or outdoor scenes with crowd overlap, incorporating occlusion augmentation during training is a straightforward and effective way to improve operational robustness.

Conclusions

1. The MACE framework proves resilient against standard architectural modifications. Alternative backbones, attention modules, and fusion strategies each produced neutral to negative retrieval outcomes when applied within the MACE pipeline, suggesting that the collaborative ensemble and bidirectional knowledge distillation already provide effective implicit feature selection and cross-modal alignment. For researchers working with ensemble-based frameworks, this finding points toward the alignment and distillation components as more productive areas for further development than the surrounding architectural elements.
2. Multi-query average fusion at inference time produces a substantial retrieval improvement over the single-query baseline, with most of the gain captured by a pool of five queries. Maximum fusion degrades as pool size grows, consistent with MACE's metric-learned embedding space where identity representations favor centroid-based aggregation. This finding also has an implication for benchmark interpretation that a single-query Rank-1 accuracy represents a lower bound on what the trained model can achieve, not its practical operating level.
3. Performance scales predictably with training data down to approximately 200 identities, below which a qualitative shift in learning behaviour occurs. At very low data fractions, the triplet loss and bidirectional distillation lose the identity diversity they depend on, and evaluation metrics plateau regardless of continued training. Above this threshold, the data-to-performance relationship is smooth and consistent.
4. The baseline MACE model is sensitive to partial query occlusion, with upper body and horizontal band conditions producing greater retrieval degradation than lower body occlusion, reflecting a concentration of discriminative weight in the torso region. Occlusion-augmented training substantially closes this gap and simultaneously improves clean performance through a regularization effect, making it a practical and low-cost addition for any deployment where partial observations are expected.

List of references

1. YE, Mang, LAN, Xiangyuan, LENG, Qingming and SHEN, Jianbing. Cross-Modality Person Re-Identification via Modality-Aware Collaborative Ensemble Learning. *IEEE Transactions on Image Processing*. 2020. Vol. 29, p. 9387–9399. DOI 10.1109/TIP.2020.2998275.
2. WU, Ancong, ZHENG, Wei Shi, YU, Hong Xing, GONG, Shaogang and LAI, Jianhuang. RGB-Infrared Cross-Modality Person Re-identification. *Proceedings of the IEEE International Conference on Computer Vision*. 22 December 2017. Vol. 2017-October, p. 5390–5399. DOI 10.1109/ICCV.2017.575.
3. CHENG, De, WANG, Xiaolong, WANG, Nannan, WANG, Zhen, WANG, Xiaoyu and GAO, Xinbo. Cross-Modality Person Re-identification with Memory-Based Contrastive Embedding. *Proceedings of the AAAI Conference on Artificial Intelligence*. 26 June 2023. Vol. 37, no. 1, p. 425–432. DOI 10.1609/aaai.v37i1.25116.
4. BI, Yihan, WANG, Rong, ZHOU, Qianli, ZENG, Zhaolong, LIN, Ronghui and WANG, Mingjie. Cross-Modality Person Re-Identification Method with Joint-Modality Generation and Feature Enhancement. *Entropy*. 13 August 2024. Vol. 26, no. 8, p. 681. DOI 10.3390/e26080681.
5. YE, Mang, SHEN, Jianbing and SHAO, Ling. Visible-Infrared Person Re-Identification via Homogeneous Augmented Tri-Modal Learning. *IEEE Transactions on Information Forensics and Security*. 2021. Vol. 16, p. 728–739. DOI 10.1109/TIFS.2020.3001665.
6. LIAO, Shengcai, HU, Yang, ZHU, Xiangyu and LI, Stan Z. Person re-identification by Local Maximal Occurrence representation and metric learning. In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2015. p. 2197–2206. ISBN 978-1-4673-6964-0. DOI 10.1109/CVPR.2015.7298832.
7. WANG, Guan'an, ZHANG, Tianzhu, CHENG, Jian, LIU, Si, YANG, Yang and HOU, Zengguang. RGB-Infrared Cross-Modality Person Re-Identification via Joint Pixel and Feature Alignment. In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, October 2019. p. 3622–3631. ISBN 978-1-7281-4803-8. DOI 10.1109/ICCV.2019.00372.
8. DAI, Pingyang, JI, Rongrong, WANG, Haibin, WU, Qiong and HUANG, Yuyu. Cross-Modality Person Re-Identification with Generative Adversarial Training. In: *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*. California : International Joint Conferences on Artificial Intelligence Organization, July 2018. p. 677–683. ISBN 9780999241127. DOI 10.24963/ijcai.2018/94.
9. YE, Mang, SHEN, Jianbing, CRANDALL, David J., SHAO, Ling and LUO, Jiebo, 2020. Dynamic Dual-Attentive Aggregation Learning for Visible-Infrared Person Re-Identification. *arXiv (Cornell University)*. Online. 18 July 2020. DOI 10.48550/arxiv.2007.09314.
10. WANG, Zhixiang, WANG, Zheng, ZHENG, Yinqiang, CHUANG, Yung-Yu and SATOH, Shin'ich. Learning to Reduce Dual-Level Discrepancy for Infrared-Visible Person Re-Identification. In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2019. p. 618–626. ISBN 978-1-7281-3293-8. DOI 10.1109/CVPR.2019.00071.
11. FENG, Zhanxiang, LAI, Jianhuang and XIE, Xiaohua. Learning Modality-Specific Representations for Visible-Infrared Person Re-Identification. *IEEE Transactions on Image Processing*. 2020. Vol. 29, p. 579–590. DOI 10.1109/TIP.2019.2928126.
12. MA, Bingpeng, SU, Yu and JURIE, Frédéric. Covariance descriptor based on bio-inspired features for person re-identification and face verification. *Image and Vision Computing*. June 2014. Vol. 32, no. 6–7, p. 379–390. DOI 10.1016/j.imavis.2014.04.002.

13. KOSTINGER, M., HIRZER, M., WOHLHART, P., ROTH, P. M. and BISCHOF, H. Large scale metric learning from equivalence constraints. In: *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, June 2012. p. 2288–2295. ISBN 978-1-4673-1228-8. DOI 10.1109/CVPR.2012.6247939.
14. PEDAGADI, Sateesh, ORWELL, James, VELASTIN, Sergio and BOGHOSSIAN, Boghos. Local Fisher Discriminant Analysis for Pedestrian Re-identification. In: *2013 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, June 2013. p. 3318–3325. ISBN 978-0-7695-4989-7. DOI 10.1109/CVPR.2013.426.
15. AHMED, Ejaz, JONES, Michael and MARKS, Tim K. An improved deep learning architecture for person re-identification. In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2015. p. 3908–3916. ISBN 978-1-4673-6964-0. DOI 10.1109/CVPR.2015.7299016.
16. LUO, Hao, JIANG, Wei, GU, Youzhi, LIU, Fuxu, LIAO, Xingyu, LAI, Shenqi and GU, Jianyang. A Strong Baseline and Batch Normalization Neck for Deep Person Re-Identification. *IEEE Transactions on Multimedia*. October 2020. Vol. 22, no. 10, p. 2597–2609. DOI 10.1109/TMM.2019.2958756.
17. JUNGLING, Kai and ARENS, Michael. Local Feature Based Person Reidentification in Infrared Image Sequences. In: *2010 7th IEEE International Conference on Advanced Video and Signal Based Surveillance*. IEEE, August 2010. p. 448–455. ISBN 978-1-4244-8310-5. DOI 10.1109/AVSS.2010.75.
18. WU, Xiang, SONG, Lingxiao, HE, Ran and TAN, Tieniu. Coupled Deep Learning for Heterogeneous Face Recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*. 25 April 2018. Vol. 32, no. 1. DOI 10.1609/aaai.v32i1.11500.
19. NGUYEN, Dat, HONG, Hyung, KIM, Ki and PARK, Kang. Person Recognition System Based on a Combination of Body Images from Visible Light and Thermal Cameras. *Sensors*. 16 March 2017. Vol. 17, no. 3, p. 605. DOI 10.3390/s17030605.
20. LEI, Zhen and LI, Stan Z. Coupled Spectral Regression for matching heterogeneous faces. In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, June 2009. p. 1123–1128. ISBN 978-1-4244-3992-8. DOI 10.1109/CVPRW.2009.5206860.
21. YE, Mang, SHEN, Jianbing, LIN, Gaojie, XIANG, Tao, SHAO, Ling and HOI, Steven C. H., 2020. Deep Learning for Person Re-identification: A survey and outlook. *arXiv (Cornell University)*. Online. 13 January 2020. DOI 10.48550/arxiv.2001.04193.
22. SUN, Yifan, XU, Qin, LI, Yali, ZHANG, Chi, LI, Yikang, WANG, Shengjin and SUN, Jian. Perceive Where to Focus: Learning Visibility-Aware Part-Level Features for Partial Person Re-Identification. In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2019. p. 393–402. ISBN 978-1-7281-3293-8. DOI 10.1109/CVPR.2019.00048.
23. ZHONG, Xian, LU, Tianyou, HUANG, Wenxin, YE, Mang, JIA, Xuemei and LIN, Chia-Wen. Grayscale Enhancement Colorization Network for Visible-Infrared Person Re-Identification. *IEEE Transactions on Circuits and Systems for Video Technology*. March 2022. Vol. 32, no. 3, p. 1418–1430. DOI 10.1109/TCSVT.2021.3072171.
24. LI, Diangang, WEI, Xing, HONG, Xiaopeng and GONG, Yihong. Infrared-Visible Cross-Modal Person Re-Identification with an X Modality. *Proceedings of the AAAI Conference on Artificial Intelligence*. 3 April 2020. Vol. 34, no. 04, p. 4610–4617. DOI 10.1609/aaai.v34i04.5891.
25. XIONG, Fu, XIAO, Yang, CAO, Zhiguo, GONG, Kaicheng, FANG, Zhiwen and ZHOU, Joey Tianyi. Good practices on building effective CNN baseline model for person re-identification. In: YU, Hui, PU, Yifei, LI, Chunming and PAN, Zhigeng (eds.), *Tenth International Conference on Graphics*

and Image Processing (ICGIP 2018). SPIE, 6 May 2019. p. 145. ISBN 9781510628281. DOI 10.1117/12.2524386.
26. PASZKE, Adam, GROSS, Sam, MASSA, Francisco, LERER, Adam, BRADBURY, James, CHANAN, Gregory, KILLEEN, Trevor, LIN, Zeming, GIMELSHEIN, Natalia, ANTIGA, Luca, DESMAISON, Alban, KÖPF, Andreas, YANG, Edward, DEVITO, Zach, RAISON, Martin, TEJANI, Alykhan, CHILAMKURTHY, Sasank, STEINER, Benoit, FANG, Lu, BAI, Junjie and CHINTALA, Soumith, 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. *arXiv (Cornell University)*. Online. 3 December 2019. Vol. 32, p. 8026–8037. DOI 10.48550/arxiv.1912.01703.
27. RIDNIK, Tal, LAWEN, Hussam, NOY, Asaf, BEN, Emanuel, SHARIR, Baruch Gilad and FRIEDMAN, Itamar. TResNet: High Performance GPU-Dedicated Architecture. In: *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, January 2021. p. 1399–1408. ISBN 978-1-6654-0477-8. DOI 10.1109/WACV48630.2021.00144.
28. LU, Wufu. Research on Pulmonary Tuberculosis Diagnosis Method Based on Dual Attention Module. In: *2023 4th International Conference on Information Processing and Computer Applications (ICIPCA)*. IEEE, 2023. p. 580–585. DOI 10.1109/icipca59209.2023.10257711.
29. ZHONG, Zhun, ZHENG, Liang, KANG, Guoliang, LI, Shaozi and YANG, Yi, 2020. Random erasing data augmentation. *Proceedings of the AAAI Conference on Artificial Intelligence*. Online. 3 April 2020. Vol. 34, no. 07, p. 13001–13008. DOI 10.1609/aaai.v34i07.7000.