Learning Progressive Modality-Shared Transformers for Effective Visible-Infrared Person Re-identification

Hu Lu[1], **Xuezhang Zou**[1], **Pingping Zhang**[2*]

[1]School of Computer Science and Communication Engineering, Jiangsu University
[2]School of Artificial Intelligence, Dalian University of Technology
luhu@ujs.edu.cn;2222108016@stmail.ujs.edu.cn;zhpp@dlut.edu.cn

Abstract

Visible-Infrared Person Re-Identification (VI-ReID) is a challenging retrieval task under complex modality changes. Existing methods usually focus on extracting discriminative visual features while ignoring the reliability and commonality of visual features between different modalities. In this paper, we propose a novel deep learning framework named Progressive Modality-shared Transformer (PMT) for effective VI-ReID. To reduce the negative effect of modality gaps, we first take the gray-scale images as an auxiliary modality and propose a progressive learning strategy. Then, we propose a Modality-Shared Enhancement Loss (MSEL) to guide the model to explore more reliable identity information from modality-shared features. Finally, to cope with the problem of large intra-class differences and small inter-class differences, we propose a Discriminative Center Loss (DCL) combined with the MSEL to further improve the discrimination of reliable features. Extensive experiments on SYSU-MM01 and RegDB datasets show that our proposed framework performs better than most state-of-the-art methods. For model reproduction, we release the source code at https://github.com/hulu88/PMT.

Introduction

Person Re-Identification (ReID) aims to retrieve the same person under different cameras and times. It can be utilized in many real-world applications, such as video surveillance, smart security, etc. Recently, with the advances of deep learning, person ReID has witnessed great success in performance and deployment. However, most of the existing ReID methods target on the visible environment. Thus, they can be regarded as visible-visible ReID. In fact, most of visible-visible ReID methods can not work well at nighttime. To address this problem, images captured by infrared cameras are considered in practical scenarios, which greatly help the ReID under different modalities and result in Visible-Infrared Person Re-Identification (VI-ReID).

Compared with single-modality ReID, VI-ReID has three main challenges: 1) The large modality gap will make it difficult to align the identify-related features of the two modalities. 2) Infrared images are more sensitive to light conditions than visible images, resulting in less discriminative features for cross-modality matching. 3) Modality-based clothing changes can occur due to the large time span, which further increases the difficulty of robust feature extraction.

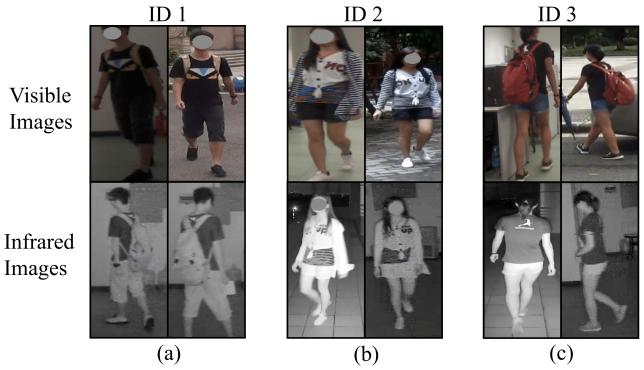

Figure 1: Several typical cases in visible-infrared person re-identification. (a) Discriminative information is not always visible due to posture or viewpoint changes. (b) Partial information may disappear due to the modality shift and different lighting conditions. (c) Modality-based clothing change due to the large time span.

To reduce the heterogeneous differences between two modalities, existing approaches (Ye et al. 2021; Gao et al. 2021; Chen et al. 2021b) mainly use a dual-stream network structure. The non-shared weight components are first used to extract modality-specific features separately before learning modality-shared features. Although these methods can effectively benefit from modality-specific features and deal with inter-modality differences, they can hardly extract effective modality-shared features. Meanwhile, there are also some Generative Adversarial Network (GAN)-based methods (Li et al. 2020; Dai et al. 2018; Wang et al. 2019) that generate cross-modality images by learning modality transformed patterns. However, these methods usually introduce additional image noises and huge computational costs. Thus, they are difficult to deploy in practical scenarios.

In addition, some outstanding methods aim to extract more discriminative information. For example, (Zhu et al. 2020; Zhang et al. 2021b) horizontally divide the portrait into multiple regions to align independent local features

*The corresponding author.

and focus on extracting fine-grained discriminative features. However, due to the great challenges, overreliance on these features may lead to wrong matches, as shown in Fig. 1. Therefore, recent VI-ReID works mainly focus on reducing the feature differences between modalities.

In this work, we propose a novel deep learning framework named Progressive Modality-shared Transformer (PMT) to extract reliable modality-invariant features for effective VI-ReID. To this end, we first propose a progressive learning strategy with Transformers (Dosovitskiy et al. 2020) to reduce the gap between visible and infrared modalities. More specifically, we improve the hard triplet loss and introduce gray-scale images as an auxiliary modality to learn modality-independent patterns. Besides, we propose a Modality-Shared Enhancement Loss (MSEL) to reduce the negative effects of modality differences and enhance the features with modality-shared information. Finally, we propose a Discriminative Center Loss (DCL) to deal with the large intra-class variance, further enhancing the discrimination of reliable modality-shared features. Extensive experiments on SYSU-MM01 and RegDB datasets show that our framework performs better than most state-of-the-art methods.

Our main contributions are summarized as follows:

- We propose a novel deep learning framework (*i.e., **PMT***) for effective VI-ReID, focusing on extracting more robust modality-shared features.
- We propose a new Modality-Shared Enhancement Loss (MSEL) to enhance the modality-shared features, thus effectively addressing the problem of feature unreliability.
- We propose a new Discriminative Center Loss (DCL) to deal with large intra-class differences and further enhance the discrimination of modality-invariant features.
- Extensive experimental results on the SYSU-MM01 and RegDB datasets show that our proposed method achieves a new state-of-the-art performance.

Related Work

Visible-infrared Person ReID

Visible-infrared person ReID aims to retrieve the same person under different image modalities. In fact, Wu *et al.* (Wu et al. 2017) first explicitly defined the VI-ReID task and built a large-scale and challenging dataset. Existing VI-ReID methods usually adopt dual-stream networks and mine modality-shared features. For example, Ye *et al.* (Ye et al. 2018) propose an effective dual-stream network to explore modality-specific and modality-shared features simultaneously. Lu *et al.* (Lu et al. 2020) propose a mechanism of feature information complementarity to exploit the potential of modality-specific features. Gao *et al.* (Gao et al. 2021) propose a multi-feature space joint optimization network to enhance modality-shared features. Zhang *et al.* (Zhang et al. 2021b) introduce a dual-stream network to achieve the global-local multiple granularity learning. Based on dual-stream networks, Zhu *et al.* (Zhu et al. 2020) propose a Heterogeneous-Center (HC) loss to reduce the modality gaps. Liu *et al.* (Liu, Tan, and Zhou 2020) further design a heterogeneous-center triplet loss and explore the parameter sharing methods to improve the feature representation ability. Ye *et al.* (Ye, Shen, and Shao 2020) introduce gray-scale images as auxiliary modalities and realize the homogeneous augmented tri-modal learning. Fu *et al.* (Fu et al. 2021) propose the cross-modality neural architecture search and improve the structural effectiveness for VI-ReID. Hao *et al.* (Hao et al. 2021) introduce a modality confusion mechanism and a center aggregation method to reduce the differences between modalities. Meanwhile, many image generation-based methods are developed to mitigate the large modality gap. For example, Dai *et al.* (Dai et al. 2018) introduce the GAN framework for cross-modality image generation, and propose the so-called cmGAN for feature learning. Furthermore, Wang *et al.* (Wang et al. 2019) propose the AlignGAN and convert visible images to infrared images with joint pixel and feature alignment. Choi *et al.* (Choi et al. 2020) attempt to disentangle cross-modality representations with hierarchical structures. Li *et al.* (Li et al. 2020) introduce an auxiliary X-modality to generate robust features and bridge the different modalities. Although the above methods are somewhat effective, they usually introduce additional image noises and huge computational costs. Thus, they are difficult to deploy in practical scenarios.

Transformer in Person ReID

Transformers (Vaswani et al. 2017) are initially proposed in Natural Language Processing (NLP). Recently, they have been utilized for some computer vision tasks, including person Re-ID. For visible-visible person ReID, He *et al.* (He et al. 2021) improve the Vision Transformer (ViT) (Dosovitskiy et al. 2020) with a side information embedding and a jigsaw patch module to learn discriminative features. Zhu *et al.* (Zhu et al. 2021) add the learnable vectors of part tokens to learn part features and integrate the part alignment into the self-attention. Lai *et al.* (Lai, Chai, and Wei 2021) utilize Transformers to generate adaptive part divisions. Zhang *et al.* (Zhang et al. 2021a) propose a hierarchical aggregation Transformer for integrating multi-level features. Chen *et al.* (Chen et al. 2021a) propose an omni-relational high-order Transformer for person Re-ID. Ma *et al.* (Ma, Zhao, and Li 2021) propose a pose-guided inter-and intra-part relational Transformer for occluded person Re-ID. As for VI-ReID, Liang *et al.* (Liang et al. 2021) improve the single-modality Transformer and try to remove modality-specific information. Chen *et al.* (Chen et al. 2022) utilize the human key-point information and introduce a structure-aware positional Transformer to learn semantic-aware modality-shared features. Although all the above Transformer-based methods have achieved superior performances, they generally lack of desirable modality-invariant properties. In this work, we explore the progressive modality-shared Transformers and learn reliable features for the VI-ReID task.

The Proposed Method

In this section, we introduce the details of the proposed Progressive Modality-shared Transformer (PMT) for VI-ReID. As shown in Fig. 2, we first take the gray-scale images as an

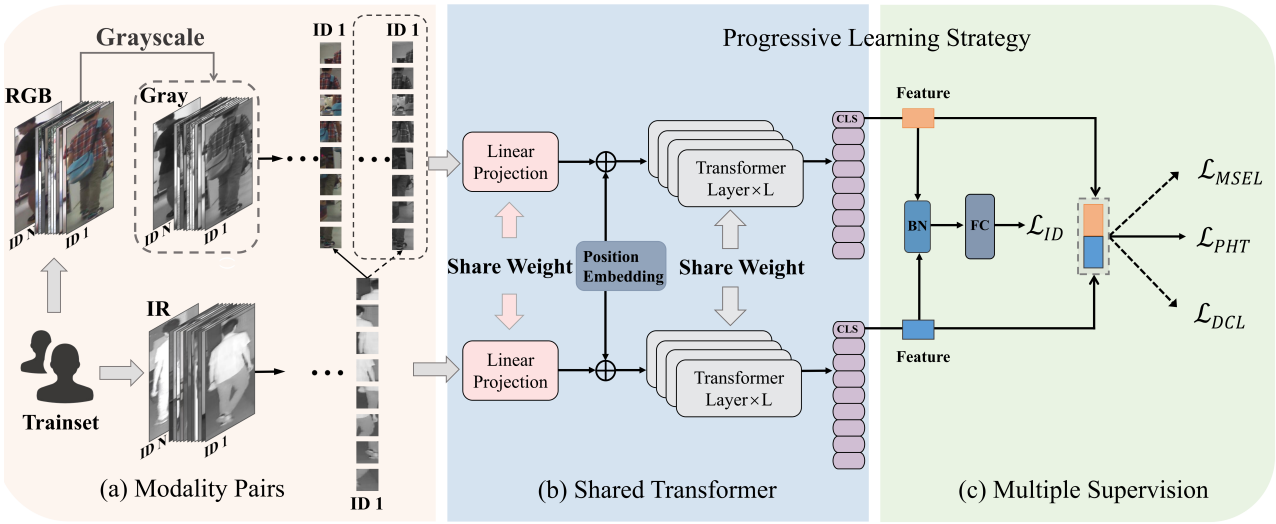

Figure 2: The framework of our proposed Progressive Modality-shared Transformer (PMT). To deal with the large modality gap, we propose a progressive learning strategy: 1) At the first stage, we feed gray-scale images and infrared images into a weight-shared Transformer supervised by L_{ID} and L_{PHT} for modality-independent feature extraction. 2) At the second stage, we utilize visual images and infrared images to improve the modality-shared features with L_{MSEL} and L_{DCL}.

auxiliary modality and adopt a weight-shared ViT (Dosovitskiy et al. 2020) as our feature extractor to capture modality-invariant features. Then, we propose a progressive learning strategy to deal with the large modality gap. Besides, a Modality-Shared Enhancement Loss (MSEL) is employed for enhancing modality-shared features. Finally, a Discriminative Center Loss (DCL) is introduced to further improve the discrimination of reliable modality-shared features.

Progressive Learning Strategy

Although previous weight-shared structures can capture more modality-shared features, they are also susceptible to modality-specific noises. Besides, the pre-trained weights on ImageNet (Deng et al. 2009) generally have a stronger reliance on low-level features, such as color or texture. Thus, directly using these pre-trained models may miss some modality-specific information. Considering above facts, we design a progressive learning strategy. The key idea is to remove color information of visible images through gray-scale images. It also helps to learn the modality-independent discriminative patterns. In this way, the negative effects from large modality gaps are effectively mitigated.

Formally, we denote the visible image and the infrared image as x^{vis} and x^{ir}, respectively. Then, the gray-scale image corresponding to x^{vis} can be denoted as x^{gray}. By feeding $\{x^{vis}, x^{gray}, x^{ir}\}$ into a weight-shared Transformer $\mathcal{F}(\cdot)$, we can obtain their corresponding embedding vectors:

$$f^v = \mathcal{F}\left(x^{vis}\right), f^g = \mathcal{F}\left(x^{gray}\right), f^{ir} = \mathcal{F}\left(x^{ir}\right). \quad (1)$$

To reduce the impact of modality-specific information, we further propose a Progressive Hard Triplet Loss (PHT). Similar to most VI-ReID methods, in each mini-batch, we randomly select P identities and then select each identity's K visible and K infrared images. Then, the proposed progressive hard triplet loss can be denoted as:

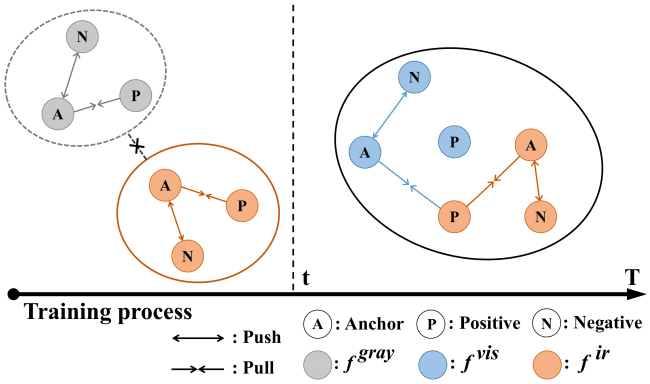

Figure 3: Illustration of the Progressive Learning Strategy.

$$L_{PHT} = \begin{cases} L_{Intra}, & X = \{x^{gray}, x^{ir}\} \\ L_{Global}, & X = \{x^{vis}, x^{ir}\} \end{cases} \quad (2)$$

$$L_{Intra} = \sum_{i=1}^{PK} \left[\max_{\forall y_i = y_j} D\left(f_i^g, f_j^g\right) - \min_{\forall y_i \neq y_k} D\left(f_i^g, f_k^g\right) + m \right]_+ + \sum_{i=1}^{PK} \left[\max_{\forall y_i = y_j} D\left(f_i^{ir}, f_j^{ir}\right) - \min_{\forall y_i \neq y_k} D\left(f_i^{ir}, f_k^{ir}\right) + m \right]_+ . \quad (3)$$

$$L_{Global} = \sum_{i=1}^{2PK} \left[\max_{\forall y_i = y_j} D\left(f_i, f_j\right) - \min_{\forall y_i \neq y_k} D\left(f_i, f_k\right) + m \right]_+ . \quad (4)$$

where $D(\cdot, \cdot)$ represents a distance metric. y_i is the identity label of the i-th image. $[z]_+ = \max(z, 0)$. m is a margin.

As shown in Fig. 3, we divide the entire training process into two stages. At the first stage, the framework

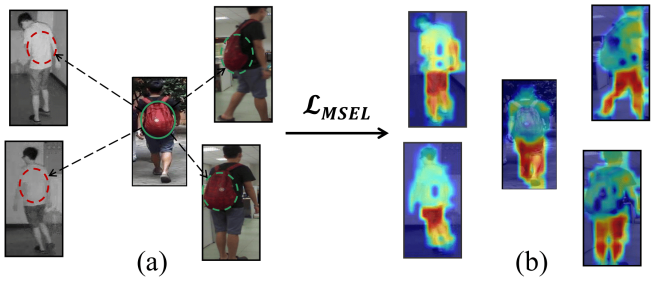

Figure 4: Motivations of the proposed Modality-Shared Enhancement Loss. It suppresses the unreliable features that only appear in one modality and enhances the modality-shared features.

takes $\{x^{gray}, x^{ir}\}$ as input, independently sampling positive and negative samples within each modality. With L_{intra}, the framework mainly focuses on learning the modality-independent discriminative patterns, thus effectively alleviating the negative effects caused by the large gap between visible and infrared modalities. At the second stage, we replace the inputs with $\{x^{vis}, x^{ir}\}$ to take full advantages of modality-specific information for more fine-grained learning. With L_{global}, the framework will no longer distinguish different modalities and select positive and negative samples only based on feature distances. This can keep the raw image information, and allow the model to be benefited from modality-specific information.

Modality-Shared Enhancement Loss

In real scenes, there are large modality differences between visible and infrared images. Thus, it is essential to extract modality-invariant features. As shown in Fig. 4 (a), the red backpack appears only in the visible modality, so over-reliance on such features will lead to failure in cross-modality retrieval. Therefore, we introduce the MSEL to appropriately suppress the unreliable features that only appear in one modality and enhance the utilization of reliable modality-invariant features.

To the above goal, we explore potential information from all samples in a mini batch. Formally, we denote the anchor features of the infrared and visible modality as f_a^{ir} and f_a^{vis}, respectively. Without loss of generality, we take f_a^{ir} as an example. Firstly, we calculate its average distance to other positive samples under the intra modality and cross modality, denoted as:

$$\mathrm{D}^{intra} = \frac{1}{K-1}\sum_{\substack{i=1 \\ i\neq a}}^{K} D\left(f_a^{ir}, f_i^{ir}\right), \tag{5}$$

$$\mathrm{D}^{cross} = \frac{1}{K}\sum_{i=1}^{K} D\left(f_a^{ir}, f_i^{vis}\right). \tag{6}$$

Then, the L_{MSEL} is defined as:

$$L_{MSEL} = \frac{1}{2PK}\sum_{p=1}^{P}\left[\sum_{a=1}^{2K}\left(\mathrm{D}_a^{intra} - \mathrm{D}_a^{cross}\right)^2\right]. \tag{7}$$

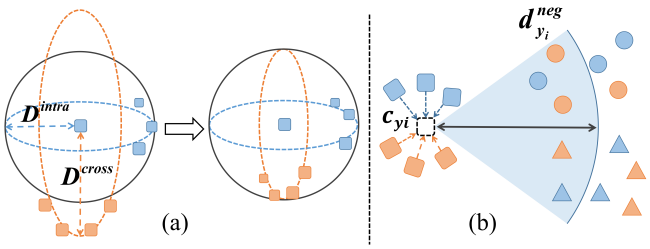

Figure 5: Geometric illustrations of the (a) Modality-Shared Enhancement Loss and (b) Discriminative Center Loss.

In Eq. 7, L_{MSEL} penalizes the difference between D^{intra} and D^{cross}. When discriminative features that appear only within one modality, then the difference between D^{intra} and D^{cross} will increase, and such anomalies will be captured by L_{MSEL}. During the bi-direction optimization process of D^{intra} and D^{cross}, the unreliable features that appear in only one modality will be suppressed, while the more reliable features that appear in both modalities will be enhanced as shown in Fig. 4 (b). Fig. 5 (a) shows the geometric illustrations of the MSEL. It encourages feature embeddings subject to a spherical distribution.

Discriminative Center Loss

Similar to visible-visible person ReID, the same person may suffer large intra-class differences due to typical variations in pose, point of view, illumination, etc. They greatly increase the difficulty of feature alignment between different modalities. To address this problem, we propose a Discriminative Center Loss (DCL) to exploit the example relationships between center instances and enhance the discriminative power of reliable modality-shared features.

Firstly, to obtain the robust representation of each identity, we compute the feature center under the two modalities by:

$$c_{y_i} = \frac{1}{2K}\left(\sum_{j=1}^{\mathrm{K}} f_j^{vis} + \sum_{k=1}^{\mathrm{K}} f_k^{ir}\right). \tag{8}$$

Here, c_{y_i} denotes the feature center of the y_i^{th} identity. Then we calculate the average distance of c_{y_i} to all other negative samples as a dynamic margin, which can be denoted as:

$$d_{y_i}^{neg} = \frac{1}{2K(P-1)}\sum_{\forall y_j \neq y_i} \|f_j - c_{y_i}\|_2. \tag{9}$$

Finally, the L_{DCL} is defined as:

$$L_{DCL} = \frac{\sum_{i=1}^{P} \underset{y_j=y_i}{mean} \|f_j - c_{y_i}\|_2}{\sum_{i=1}^{P} \underset{\substack{\|f_k - c_{y_i}\|_2 < d_{y_i}^{neg} \\ y_k \neq y_i}}{mean} \|f_k - c_{y_i}\|_2}. \tag{10}$$

By minimizing Eq. 10, the intra-class compactness and inter-class separability will be improved. Fig. 5 (b) shows the geometric illustrations of the DCL. The utilization of L_{DCL} has two main advantages: 1) It can utilize modality-specific features and capture more potential relationships than the center-center solution. 2) The dynamic sampling through $d_{y_i}^{neg}$ can effectively focus on relatively difficult examples. The effectiveness will be verified by experiments.

Overall Objective Function

For model training, we adopt a hybrid loss function for our progressive learning framework. At the first stage, we utilize the identity loss L_{ID} (Zheng, Zheng, and Yang 2017) and L_{Intra} to learn modality-independent features:

$$L_1 = L_{Intra} + L_{ID}. \tag{11}$$

At the second stage, we further extract the reliable modality-shared features with L_{MSEL} and enhance the discrimination with L_{DCL}. The loss function can be defined as:

$$L_2 = L_{Global} + L_{ID} + \lambda_1 L_{MSEL} + \lambda_2 L_{DCL}. \tag{12}$$

Here, the parameters λ_1 and λ_2 are used to balance the terms of L_{MSEL} and L_{DCL}, respectively.

Experiments

Experimental Setting

Datasets. In this work, we follow previous methods and conduct experiments on two public VI-ReID datasets.

SYSU-MM01 (Wu et al. 2017) has a total of 286,628 visible images and 15,792 infrared images with 491 different person identities. The training set contains 22,258 visible images and 11,909 infrared images of 395 persons, and the test set contains images of another 96 different identities. 3803 infrared images are used as the query set, and from the other visible images, 301 images are randomly selected as the gallery set. Besides, there are two search modes. The *all-search mode* uses all images for testing, while the *indoor-search mode* only uses the indoor images.

RegDB (Nguyen et al. 2017) contains a total of 412 different person identities. For each person, 10 visible images and 10 infrared images are captured. We follow the evaluation protocol in (Ye et al. 2018) and randomly select all images of 206 identities for training and the remaining 206 identities for testing. To obtain the stable results, we randomly divide this dataset ten times for independent training and testing.

Evaluation metrics. We use Cumulative Matching Characteristics (CMC), Mean Average Precision (mAP), and Mean Inverse Negative Penalty (mINP) (Ye et al. 2021) as our main evaluation metrics.

Implementation details. Our proposed method is implemented with the Huawei-Mindspore toolbox and one NVIDIA RTX3090 GPU. We adopt the ViT-B/16 (Dosovitskiy et al. 2020) pre-trained on ImageNet (Deng et al. 2009) as our backbone and set the overlap stride to 12 to balance speed and performance. All person images are resized to 256×128 with horizontal flipping and random erasing for data augmentation. For infrared images, color jitter and gaussian blur are additionally applied. The batch size is set to 64, containing a total of 8 different identities. For each identity, 4 visible images and 4 infrared images are sampled. We adopt AdamW optimizer with a cosine annealing learning rate scheduler for training. The basic learning rate is set to $3e^{-4}$ and weight decay is set to $1e^{-4}$. We train 24 epochs for the SYSU-MM01 and 36 epochs for the RegDB. For both datasets, the epoch t of the first stage is set to 6, the trade-off parameters λ_1 and λ_2 are set to 0.5, and the margin parameter m is set to 0.1. The 768-dimensional features after the BN layer are used for testing.

Ablation Studies

In this subsection, we conduct experiments to verify the effects of different modules on the SYSU-MM01 dataset under the all-search mode.

Effectiveness of the progressive learning strategy. We evaluate the effectiveness in terms of both image modality and loss function. For the image modality, the comparison results are shown in Tab. 1. "Baseline (RGB)" and "Baseline (Gray)" indicates the baseline model trained with the RGB-IR modality and Grayscale-IR modality, respectively. "Baseline (Gray-RGB)" means that the first t epochs of training uses the Grayscale-IR modality and the rest uses the RGB-IR modality. From the results, one can see that directly using the RGB-IR modality and Grayscale-IR modality shows inferior performances. With a progressive learning strategy, the model can show better results, indicating the effectiveness of reducing the modality differences. Our proposed strategy significantly improves the "Baseline (RGB)" model by 5.16% Rank-1, 5.04% mAP, and 5.74% mINP.

Besides, based on the "Baseline (Gray-RGB)" model, we replace the hard triplet loss with WRT (Ye et al. 2021), HCT (Liu, Tan, and Zhou 2020) and our proposed PHT. The comparison results are shown in Tab. 2. These results further prove the effectiveness of our progressive learning strategy.

Effects with MSEL and DCL. Tab. 3 shows the comparison results of different settings with MSEL and DCL. "Base(PL)" only adopts the progressive learning strategy. "MSEL (Cosine)" and "MSEL (Euclid)" mean that the model uses the cosine distance and Euclidean distance, respectively. The experimental results show that the model with MSEL consistently improves the performance. The "MSEL (Euclid)" model brings the best results with 3.44% Rank-1, 2.21% mAP, and 1.91% mINP improvement compared to "Base (PL)".

As for the DCL, "DCL (Hard)" means only selecting the closest negative sample for each identity center. "DCL (All)" means selecting all negative samples for each identity center, and "DCL (Dyn)" means dynamically selecting negative samples based on Eq. 9. As shown in Tab. 3, the L_{DCL} can bring a consistent improvement. Our dynamic selection shows much better relative results with 2.38% Rank-1, 3.50% mAP, and 4.89% mINP when compared with "Base

Transition Schemes	Rank-1	Rank-10	mAP	mINP
Baseline (RGB)	52.51	88.21	51.30	38.51
Baseline (Gray)	56.51	91.76	54.22	39.75
Baseline (RGB-Gray)	57.24	91.87	54.95	40.25
Baseline (Gray-RGB)	59.07	92.53	56.86	42.81

Table 1: Effects of different image modalities.

Methods	Rank-1	Rank-10	mAP	mINP
HardTri ($m = 0.1$)	59.07	92.53	56.86	42.81
WRT	54.90	92.33	54.74	42.30
HCT ($m = 0.3$)	59.51	92.38	56.68	42.05
PHT ($m = 0.1$)	61.67	93.02	59.26	45.49

Table 2: Effects of progressive learning losses.

Methods	Rank-1	Rank-10	mAP	mINP
Base (PL)	61.67	93.02	59.26	45.49
+MSEL (Cosine)	64.19	93.45	60.67	46.15
+MSEL (Euclid)	65.11	93.81	61.47	47.40
+DCL (Hard)	63.22	94.09	62.13	49.97
+DCL (All)	62.86	94.15	61.32	48.73
+DCL (Dyn)	64.05	94.71	62.76	50.38
+MSEL (Euclid)+DCL (Dyn)	67.53	95.36	64.98	51.86

Table 3: Comparison results with MSEL and DCL.

Methods	Rank-1	mAP	mINP
AGW	58.19	56.50	43.52
AGW + MSEL	62.16	59.66	46.38
AGW + MSEL + PL	65.97	62.15	47.30
AGW + MSEL + PL + DCL	67.09	64.25	50.89
Ours	67.53	64.98	51.86

Table 4: Comparison results with CNN-based backbones.

(PL)". Combined with MSEL, the model can bring a further 2.42% Rank-1, 3.51% mAP and 4.46% mINP improvement. These results fully demonstrate the effectiveness of our MSEL and DCL.

Effects of CNN backbones. To further study the effectiveness and generalization of our proposed methods, we also carry out experiments with CNN-based frameworks. As a typical example, we take the outstanding AGW method with random erasing (Ye et al. 2021). The experimental results are listed in Tab. 4. The results show that by adding MSEL, the model delivers a performance gain of 3.97% Rank-1, 3.16% mAP, and 2.86% mINP. The results indicate that L_{MSEL} can also be compatible with variants of different triples. With the full modules ("MSEL+PL+DCL"), the model can achieve a performance gain of 8.90% Rank-1, 7.75% mAP, and 8.34% mINP. These facts clearly demonstrate the generalization of our proposed methods on CNN-based frameworks. However, our Transformer-based framework shows better results than CNN-based ones, as shown in the last row.

Trade-off parameters. We conduct additional experiments to evaluate the effect of trade-off parameters λ_1 and λ_2. As shown in Fig. 6, L_{MSEL} is not sensitive to the parameter settings, while L_{DCL} is stable within a certain range.

Visualization analysis. To analyse the visual effect of our proposed model, we present some typical visual examples. As shown in Fig. 7, we use the Grad-CAM (Selvaraju et al. 2017) to generate attention maps of query images with our models. Besides, the top 10 retrieval results are also provided in Fig. 7 (d). One can observe that with our MSEL, the model focuses on more discriminative regions and extracts decent modality-shared features. Thus, the model can effectively deal with complex scenarios in Fig. 1. In addition, we randomly sample 10,000 positive and negative matching pairs from the test set and visualize their cosine similarity distribution, as shown in Fig. 8. One can observe that the similarity of positive cross-modality matching pairs increases by introducing MSEL, which indicates that MSEL enhances the utilization of reliable modality-invariant features. By further introducing DCL, the range of cosine similarity between negative and positive pairs is significantly expanded. The visualization shows the ability of L_{DCL} to explore potential information and effectively improve the discrimination of feature embeddings.

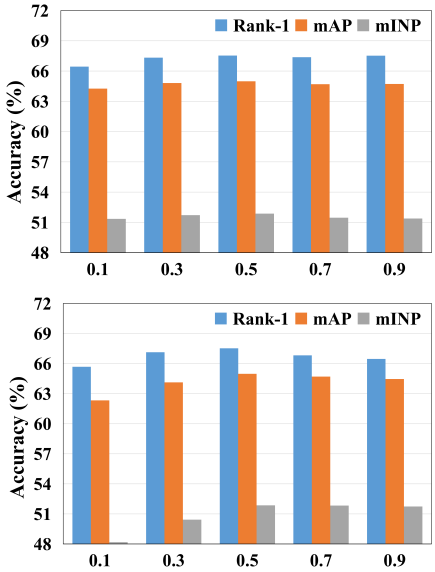

Figure 6: Performance effects of trade-off parameters λ_1 and λ_2. In the top sub-figure, $\lambda_2 = 0.5$, $\lambda_1 \in [0.1, 0.9]$ and in the bottom sub-figure, $\lambda_1 = 0.5$, $\lambda_2 \in [0.1, 0.9]$.

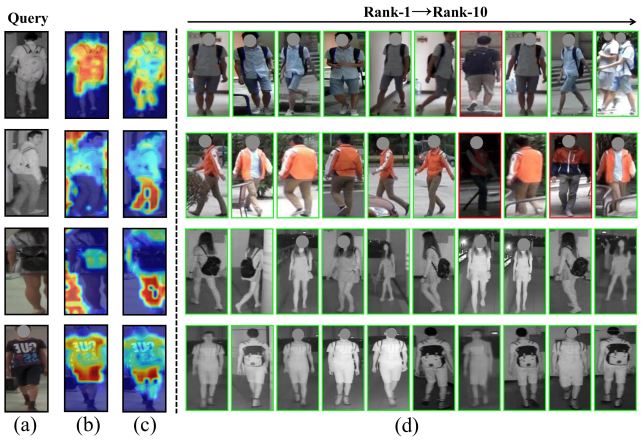

Figure 7: Attention maps and retrieval results. (a) Query images. (b) Attention maps w/o MSEL. (c) Attention maps w/ MSEL. (d) Top 10 retrieval results.

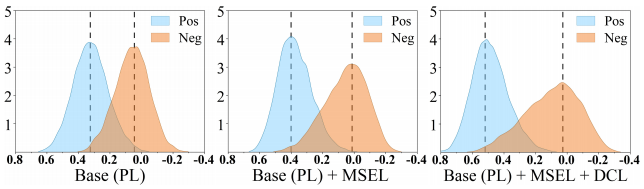

Figure 8: The cosine similarity distribution of positive and negative matching pairs from the test set.

Methods	*All search*					*Indoor search*				
	$r=1$	$r=10$	$r=20$	mAP	mINP	$r=1$	$r=10$	$r=20$	mAP	mINP
Zero-Pad (Wu et al. 2017)	14.80	54.12	71.33	15.95	-	20.58	68.38	85.79	26.92	-
Hi-CMD (Choi et al. 2020)	34.94	77.58	-	35.94	-	-	-	-	-	-
CMSP (Wu et al. 2020)	43.56	86.25	-	44.98	-	48.62	89.50	-	57.50	-
expAT Loss (Ye et al. 2020a)	38.57	76.64	86.39	38.61	-	-	-	-	-	-
AGW (Ye et al. 2021)	47.50	84.39	92.14	47.65	35.30	54.17	91.14	95.98	62.97	59.23
HAT (Ye, Shen, and Shao 2020)	55.29	92.14	97.36	53.89	-	62.10	95.75	99.20	69.37	-
LBA (Park et al. 2021)	55.41	91.12	-	54.14	-	58.46	94.13	-	66.33	-
NFS (Chen et al. 2021b)	56.91	91.34	96.52	55.45	-	62.79	96.53	99.07	69.79	-
MSO (Gao et al. 2021)	58.70	92.06	97.20	56.42	42.04	63.09	96.61	-	70.31	-
CM-NAS (Fu et al. 2021)	61.99	92.87	97.25	60.02	-	67.01	**97.02**	**99.32**	72.95	-
MID (Huang et al. 2022)	60.27	92.90	-	59.40	-	64.86	96.12	-	70.12	-
SPOT (Chen et al. 2022)	65.34	92.73	97.04	62.25	48.86	69.42	96.22	99.12	74.63	70.48
FMCNet (Zhang et al. 2022)	66.34	-	-	62.51	-	68.15	-	-	74.09	-
PMT (Ours)	**67.53**	**95.36**	**98.64**	**64.98**	**51.86**	**71.66**	96.73	99.25	**76.52**	**72.74**

Table 5: Comparisons with state-of-the-art methods under all-search and indoor-search modes on the SYSU-MM01 dataset.

Methods	*V to T*		*T to V*	
	$r=1$	mAP	$r=1$	mAP
DDAG (Ye et al. 2020b)	69.34	63.46	68.06	61.80
expAT Loss (Ye et al. 2020a)	66.48	67.31	67.45	66.51
AGW (Ye et al. 2021)	70.05	66.37	70.49	65.90
HAT (Ye, Shen, and Shao 2020)	71.83	67.56	70.02	66.30
MSO (Gao et al. 2021)	73.6	66.9	74.6	67.5
LBA (Park et al. 2021)	74.17	67.64	72.43	65.46
NFS (Chen et al. 2021b)	80.54	72.10	77.95	69.79
MCLNet (Hao et al. 2021)	80.31	73.07	75.93	69.49
SPOT (Chen et al. 2022)	80.35	72.46	79.37	72.26
PMT (Ours)	**84.83**	**76.55**	**84.16**	**75.13**

Table 6: Comparison results under Visible-Thermal and Thermal-Visible modes on the RegDB dataset.

Comparison with State-of-the-Arts

In this subsection, we compare our proposed PMT with other state-of-the-art methods on SYSU-MM01 and RegDB.

SYSU-MM01: Tab. 5 shows the comparison results on the SYSU-MM01 dataset. One can observe that our proposed method outperforms other weight-shared methods [expAT (Ye et al. 2020a), HAT (Ye, Shen, and Shao 2020)] by at least 12.24% in Rank-1 and 11.09% in mAP under the all-search mode. Moreover, compared with dual-stream-based methods [LBA (Park et al. 2021), NFS (Chen et al. 2021b), SPOT (Chen et al. 2022)], our proposed method also has substantial performance advantages. Under the indoor-search mode, our proposed method shows much better results in terms of Rank-1, mAP and mINP. In terms of Rank-10 and Rank-20, CM-NAS (Fu et al. 2021) shows best results. The main reason may be that the searched network by CM-NAS is more helpful for global discrimination. However, our proposed method delivers very comparable results. All the above results fully demonstrate that our proposed PMT can effectively reduce the large modality gap and utilize more reliable modality-shared features.

RegDB: In Tab. 6, we report the comparison results on the RegDB dataset. The results show that our proposed method achieves excellent performances in both Visible to Thermal ($VtoT$) and Thermal to Visible ($TtoV$) modes. More specifically, our proposed method achieves an expressive performance of 84.83% Rank-1 and 76.55% mAP under the $VtoT$ mode, showing a 4% performance gain than other best methods. For the more challenging $TtoV$ mode, our method also shows great performance advantages. These results also indicate that our proposed method is more robust against different datasets and query patterns.

Conclusion

In this paper, we propose a novel deep learning-based framework named PMT, which effectively improves the performance of VI-ReID by fully exploring reliable modality-invariant features. With gray-scale images as an auxiliary modality, our framework mitigates the large gap between RGB-IR modalities through a progressive learning strategy. Meanwhile, our proposed MSEL and DCL can effectively extract more reliable and discriminative features, bringing stronger performance and robustness. Moreover, the proposed methods have a good generalization. By applying our methods to CNN-based backbones, they can also bring significant performance improvements. Experimental results on two public VI-ReID benchmarks verify the effectiveness of our proposed framework. In the future, we will explore more effective Transformer structures to further improve the feature representation ability.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (NSFC) (No. 91538201, 62101092), the CAAI-Huawei MindSpore Open Fund under Grant CAAIXSJLJJ-2021-067A, and the Fundamental Research Funds for the Central Universities (No. DUT20RC(3)083).

References

Chen, C.; Ye, M.; Qi, M.; Wu, J.; Jiang, J.; and Lin, C.-W. 2022. Structure-Aware Positional Transformer for Visible-Infrared Person Re-Identification. *IEEE Transactions on Image Processing*, 31: 2352–2364.

Chen, X.; Xu, J.; Xu, J.; and Gao, S. 2021a. OH-Former: Omni-Relational High-Order Transformer for Person Re-Identification. *arXiv:2109.11159*.

Chen, Y.; Wan, L.; Li, Z.; Jing, Q.; and Sun, Z. 2021b. Neural feature search for rgb-infrared person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 587–597.

Choi, S.; Lee, S.; Kim, Y.; Kim, T.; and Kim, C. 2020. Hi-CMD: Hierarchical cross-modality disentanglement for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10257–10266.

Dai, P.; Ji, R.; Wang, H.; Wu, Q.; and Huang, Y. 2018. Cross-modality person re-identification with generative adversarial training. In *International Joint Conference on Artificial Intelligence*, volume 1, 6.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, 248–255. IEEE.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.

Fu, C.; Hu, Y.; Wu, X.; Shi, H.; Mei, T.; and He, R. 2021. CM-NAS: Cross-modality neural architecture search for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11823–11832.

Gao, Y.; Liang, T.; Jin, Y.; Gu, X.; Liu, W.; Li, Y.; and Lang, C. 2021. MSO: Multi-feature space joint optimization network for rgb-infrared person re-identification. In *Proceedings of the 29th ACM International Conference on Multimedia*, 5257–5265.

Hao, X.; Zhao, S.; Ye, M.; and Shen, J. 2021. Cross-modality person re-identification via modality confusion and center aggregation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 16403–16412.

He, S.; Luo, H.; Wang, P.; Wang, F.; Li, H.; and Jiang, W. 2021. Transreid: Transformer-based object re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15013–15022.

Huang, Z.; Liu, J.; Li, L.; Zheng, K.; and Zha, Z.-J. 2022. Modality-Adaptive Mixup and Invariant Decomposition for RGB-Infrared Person Re-Identification. *arXiv preprint arXiv:2203.01735*.

Lai, S.; Chai, Z.; and Wei, X. 2021. Transformer Meets Part Model: Adaptive Part Division for Person Re-Identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4150–4157.

Li, D.; Wei, X.; Hong, X.; and Gong, Y. 2020. Infrared-visible cross-modal person re-identification with an x modality. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 4610–4617.

Liang, T.; Jin, Y.; Gao, Y.; Liu, W.; Feng, S.; Wang, T.; and Li, Y. 2021. CMTR: Cross-modality Transformer for Visible-infrared Person Re-identification. *arXiv preprint arXiv:2110.08994*.

Liu, H.; Tan, X.; and Zhou, X. 2020. Parameter sharing exploration and hetero-center triplet loss for visible-thermal person re-identification. *IEEE Transactions on Multimedia*, 23: 4414–4425.

Lu, Y.; Wu, Y.; Liu, B.; Zhang, T.; Li, B.; Chu, Q.; and Yu, N. 2020. Cross-modality person re-identification with shared-specific feature transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13379–13389.

Ma, Z.; Zhao, Y.; and Li, J. 2021. Pose-guided inter- and intra-part relational transformer for occluded person re-identification. In *Proceedings of the 29th ACM International Conference on Multimedia*, 1487–1496.

Nguyen, D. T.; Hong, H. G.; Kim, K. W.; and Park, K. R. 2017. Person recognition system based on a combination of body images from visible light and thermal cameras. *Sensors*, 17(3): 605.

Park, H.; Lee, S.; Lee, J.; and Ham, B. 2021. Learning by aligning: Visible-infrared person re-identification using cross-modal correspondences. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 12046–12055.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, 618–626.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

Wang, G.; Zhang, T.; Cheng, J.; Liu, S.; Yang, Y.; and Hou, Z. 2019. RGB-infrared cross-modality person re-identification via joint pixel and feature alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3623–3632.

Wu, A.; Zheng, W.-S.; Gong, S.; and Lai, J. 2020. Rgb-ir person re-identification by cross-modality similarity preservation. *International Journal of Computer Vision*, 128(6): 1765–1785.

Wu, A.; Zheng, W.-S.; Yu, H.-X.; Gong, S.; and Lai, J. 2017. RGB-infrared cross-modality person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, 5380–5389.

Ye, H.; Liu, H.; Meng, F.; and Li, X. 2020a. Bi-directional exponential angular triplet loss for RGB-infrared person re-identification. *IEEE Transactions on Image Processing*, 30: 1583–1595.

Ye, M.; Lan, X.; Li, J.; and Yuen, P. 2018. Hierarchical discriminative learning for visible thermal person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Ye, M.; Shen, J.; J Crandall, D.; Shao, L.; and Luo, J. 2020b. Dynamic dual-attentive aggregation learning for visible-infrared person re-identification. In *European Conference on Computer Vision*, 229–247. Springer.

Ye, M.; Shen, J.; Lin, G.; Xiang, T.; Shao, L.; and Hoi, S. C. 2021. Deep learning for person re-identification: A survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6): 2872–2893.

Ye, M.; Shen, J.; and Shao, L. 2020. Visible-infrared person re-identification via homogeneous augmented tri-modal learning. *IEEE Transactions on Information Forensics and Security*, 16: 728–739.

Zhang, G.; Zhang, P.; Qi, J.; and Lu, H. 2021a. Hat: Hierarchical aggregation transformers for person re-identification. In *Proceedings of the 29th ACM International Conference on Multimedia*, 516–525.

Zhang, L.; Du, G.; Liu, F.; Tu, H.; and Shu, X. 2021b. Global-local multiple granularity learning for cross-modality visible-infrared person reidentification. *IEEE Transactions on Neural Networks and Learning Systems*.

Zhang, Q.; Lai, C.; Liu, J.; Huang, N.; and Han, J. 2022. FMCNet: Feature-Level Modality Compensation for Visible-Infrared Person Re-Identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7349–7358.

Zheng, Z.; Zheng, L.; and Yang, Y. 2017. A discriminatively learned cnn embedding for person reidentification. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 14(1): 1–20.

Zhong, Z.; Zheng, L.; Kang, G.; Li, S.; and Yang, Y. 2020. Random erasing data augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 13001–13008.

Zhu, K.; Guo, H.; Zhang, S.; Wang, Y.; Huang, G.; Qiao, H.; Liu, J.; Wang, J.; and Tang, M. 2021. Aaformer: Auto-aligned transformer for person re-identification. *arXiv preprint arXiv:2104.00921*.

Zhu, Y.; Yang, Z.; Wang, L.; Zhao, S.; Hu, X.; and Tao, D. 2020. Hetero-center loss for cross-modality person re-identification. *Neurocomputing*, 386: 97–109.