2025 International Conference on Intelligent Computing
July 26-29, Ningbo, China
https://www.ic-icc.cn/2025/index.php

TMN: Bridging Modality Gap via Transition Modality Network for Visible-Infrared Person Re-Identification

Mengzhe Wang[1] and Yuhao Wang[1, 2, ✉]

[1]School of Information Science and Technology, Fudan University, Shanghai, China
[2]Faculty of Business and Law, The University of Queensland, St Lucia Campus, Brisbane, QLD 4072, Australia
yuhao.wang5@student.uq.edu.au

Abstract. Visible-infrared person re-identification (VI-ReID) task aims to match visible and infrared pedestrian images. However, due to the modality gap, VI-ReID task faces serious technical challenges. Existing methods have made significant progress but still suffer from two limitations: Unsupervised image generation methods are computational intensive and may introduce additional noise; feature-level alignment struggles with designing effective loss functions for complex and abstract features output, resulting in insufficient learning and constraint of the model. To address these issues, we propose a Transition Modality Network (TMN), which aims to construct a transitional modality between the two modalities, enabling early-stage cross-modality interaction at shallow network layers, thereby avoiding large computations and complex loss function design. First, the processed visible and infrared features are input into the Visible-infrared Transition Modality Fusion module (VI-TMF) to construct the transition modality. Secondly, we embed the Grouped Spatial-Channel Excitation block (GSCE) into the Resnet-50 for deep feature processing and extraction. Finally, we design a cross-modality bridging loss function to align the features of the three modalities. Through experiments on two benchmark datasets, TMN achieves Rank-1/mAP accuracy of 71.42%/65.91% on the SYSU-MM01 dataset, and 92.14%/83.25% on the RegDB dataset, demonstrating that transition modality construction effectively bridges cross-modality discrepancies and establishes a novel paradigm for addressing the fundamental challenges in VI-ReID tasks.

Keywords: Visible-infrared person re-identification, Transition Modality, Feature interaction and fusion.

1 Introduction

In the fundamental computer vision task of person re-identification (ReID), pedestrian images from non-overlapping cameras are matched in visible light. [1,2]. Many approaches have demonstrated impressive performance in feature identification capabilities from RGB pictures due to the rapid growth of deep learning [3, 4, 5]. Additionally, Vision Transformer (ViT) and attention mechanisms has significantly increased robustness to intra-modality fluctuations brought on by changes in posture and viewpoint [6,

7, 8]. Some research has even outperformed humans in ReID task on several datasets with innovative methodologies [9, 10]. Nonetheless, it performs poorly in low-light conditions [11], for traditional visible-visible ReID (VV-ReID) only relies on RGB images. Visible-infrared ReID (VI-ReID) can address this problem because its methodology is to match images across visible and infrared modalities. Thus, it innovatively leverages infrared cameras' ability to capture details in the darkness [12, 13, 14]. However, as shown in **Fig. 1**(a), VI-ReID still faces the challenge of the modality gap [15, 16].

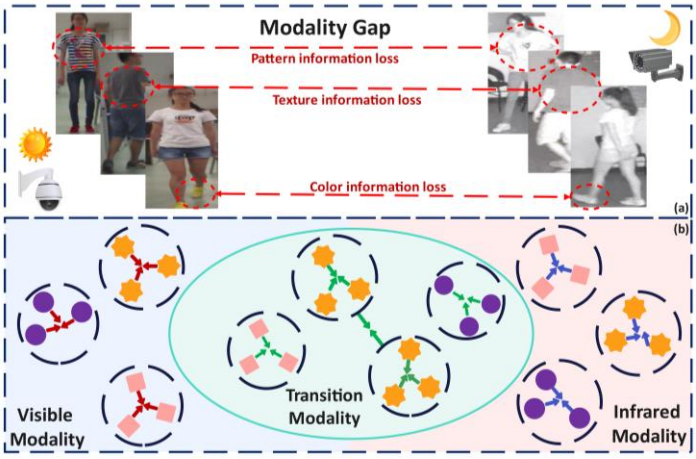

Fig. 1. Subfigure(a) illustrates the modality gap in VI-ReID task, showing the notable differences between visible and infrared images. Specifically, in visible modality, RGB images can capture the texture, color and pattern details, but in infrared modality, there is only temperature information. Subfigure(b) shows our idea to solve the reduce the modality gap by constructing a transition modality and designing loss functions to align features from three modalities.

In order to solve the modality gap problem, one approach in existing researches aims to extract cross-modality features by using the CNN-based structure [17, 18]. Notably, a dual-stream network is one of the commonly used architectures, which processes visible and infrared images through two separate networks, and then performs higher-level feature fusion to mitigate modality differences [19, 20]. Another approach is to alleviate the differences between modalities by using image generation mechanisms, such as GAN [21, 22]. Among them, Si et al. [21] proposed a three-modality consistency optimization model (TCOM). However, feature-level methods are very difficult to design a suitable loss function because the features in the deep layers are complex and extensive, and may also cause underfitting due to insufficient constraints. At the same time, most of image generation methods are based on unsupervised training, causing computationally intensive and may bring additional noise. To address these problems, our work is primarily inspired by the work of Dai et al. [23, 24], who proposed IDM and IDM++ for VV-ReID task, introducing intermediate domains between source and target domains for progressive alignment. As shown in **Fig. 1**(b), we aim to construct a transition modality between visible and infrared modalities and design loss functions to align features from these three modalities to improve the performance in completing

the VI-ReID task.

Based on AGW [16] by Ye et al., we propose a novel method named Transition Modality Network (TMN). As shown in **Fig. 2**, We firstly use separate modules to extract shallow RGB and IR features. Then, we construct a Visible-infrared Transition Modality Fusion module (VI-TMF) to dynamically blend dual-stream features, enabling features smoothly transit along geodesic paths, avoiding abrupt alignment. We design a Grouped Spatial-Channel Excitation block (GSCE) and embed it into the Resnet-50 backbone to improve feature extraction performance. Lastly, we use identity classification loss and a designed cross-modality bridging loss function to balance intra-class compactness and inter-class separation. Experiments on SYSU-MM01 and RegDB show the advanced performance our method, proving transition modality effectively harmonize modality gaps and intra-modality variations.

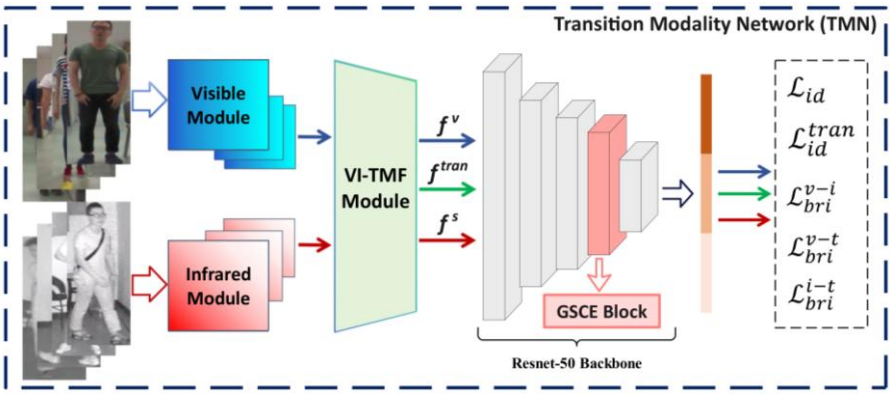

Fig. 2. The framework of TMN. Pedestrian images are first extracted by visible and infrared modules, and are fed into VI-TMF module to construct a transition modality. The GSCE module is embedded into Resnet-50 backbone for deeper feature processing and interaction. Finally, the feature alignment of the three modalities is carried out through the designed loss functions.

The key contributions of our work can be summarized as follows:

- We propose a method for constructing transition modality (TMN) between the visible and infrared modality to achieve progressive cross-modality alignment of the model, notably reducing the modality gap in cross-modality VI-ReID task.
- We design Visible-infrared Transition Modality Fusion module (VI-TMF) for transition modality construction, Grouped Spatial-Channel Excitation block (GSCE) for global feature extraction and cross-modality bridging loss function to constrain the feature alignment process. These components work together to enhance feature interaction and fusion, ensuring effective cross-modality alignment.
- We have conducted experiments on the two benchmark datasets. The experimental results demonstrate TMN achieves Rank-1/mAP accuracy of 71.42%/65.91% on the SYSU-MM01 dataset, and 92.14%/83.25% on the RegDB dataset, which reaches the advanced level of accuracy in prediction, providing a new theoretical insight for cross-modality VI-ReID task.

2 Related Works

2.1 Visible-visible ReID (VV-ReID)

In the early stages of VV-ReID research, global feature extraction evolved into local feature extraction. A Part-based Convolutional Baseline (PCB) was a symbolic method for aligning local features using Refined Part Pooling (RPP) without pose priors as suggested by Sun et al. [25], and Bai et al. [26] proposed using LSTM for capturing local feature sequences. In addition, Zhu et al. [27] introduced AAformer, which demonstrated the transition from rigid partitioning to adaptive learning using learnable "Part tokens" and optimal transport for unsupervised semantic clustering, demonstrating the shift from rigid partitioning to adaptive learning. Researchers have also explored new learning strategies as well as designed loss functions to improve feature discriminability. For example, Yan et al. [28] designed FIDI Loss with exponential penalties for fine-grained differences to address sensitivity to subtle variations; semi-supervised strategies like LSRO (Zheng et al. [3]) leveraged GAN-generated data and label smoothing. Zhang et al. [10] attempted to combine the CNN-Transformer architecture with depth-supervised aggregation (DSA) to enable hierarchical feature fusion at multiple levels. To address occlusion and lighting issues, Zeng et al. [29] suggested the IID network using adversarial training to separate identity and illumination variables. Further, Zhao et al. [30] developed IGO, which progressively generates occlusion and suppressed interference. Ge et al. [31] developed a FD-GAN framework which distilled pose-invariant features by using Siamese GANs. However, these methods are still limited to single-modal information, reducing their ability to overcome inherent data source limitations.

2.2 Visible-infrared person ReID (VI-ReID)

Visible-infrared person re-identification (VI-ReID) has advanced through dataset construction, generative alignment and feature fusion. The foundation was laid by Wu et al. [12], who took the first step as a pioneer and introduced the SYSU-MM01 dataset and demonstrated the potential of single-stream networks with deep zero-padding to implicitly bridge modality gaps. Early generative methods, like cmGAN [32] and AlignGAN [33], used adversarial training for pixel-to-feature alignment. Qi et al. [21] later improved these methods with contrastive learning and channel attention (GC-IFS), achieving 85.63% Rank-1 on SYSU-MM01. Feature fusion methods evolved with Cheng et al.'s TFFN [20] and Liu et al. [34]'s parameter-sharing dual-stream networks which optimized via Hetero-Center Triplet Loss (HCT), reaching 91.05% Rank-1 on the RegDB dataset. The View-decoupled Transformer (VDT) by Zhang et al. [35] and EDITOR [36] were two examples of dynamic architectural innovations. The former employed hierarchical tokens to distinguish between view and identification information, while the latter employed spatial-frequency selection to mitigate background noise. Additionally, as for the design of loss functions also made significant progress. Ye et al. [19, 37] combined cross-modality constraints with hierarchical metrics and BDTR, while Ren and Zhang's IDKL [38] achieved the then-current SOTA on RegDB dataset by distilling modality-specific discriminative knowledge into shared features

via triplet graph alignment.

3 Methodology

3.1 GCSE Module

As shown in **Fig. 3**, Ye et al. incorporate the Non-local block from Wang et al. [39] which aims to capture global feature dependencies, thus addressing the limitations of traditional local operations. It computes a weighted sum of features across all positions:

$$z_i = W_z * f(x_i) + x_i \tag{1}$$

where W_z is a learnable weight matrix and $f(\cdot)$ denotes the non-local operation.

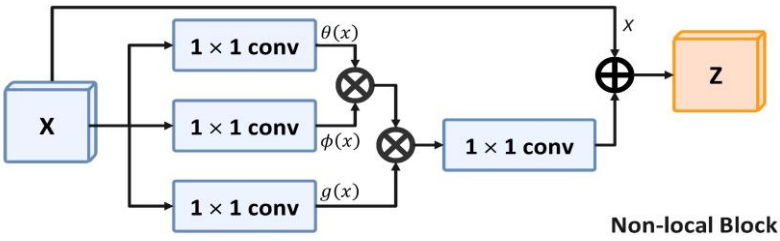

Fig. 3. The structure of Non-local Block in AGW (baseline) [16], which applies a single 1×1 convolution to generate $\theta(x)$, $\phi(x)$, and $g(x)$, the block also adds the input x to form the residual connection.

We enhance the Non-local block with the a Grouped Spatial-Channel Excitation block (GSCE), improving former method in terms of grouping and channels. The structure of GSCE block is shown in **Fig. 4** GSCE employs grouped convolutions to partition the input features into parallel interaction branches. Each branch independently computes spatial correlations through:

$$A^{(g)} = \text{Softmax}\left(\left(\theta^{(g)}(x)\right)^T \phi^g(x)\right) \tag{2}$$

where $g = 1,2,\ldots,N$ denotes the group index. Each group is then formulated as:

$$y^{(g)} = A^{(g)}\, g^{(g)}(x) \tag{3}$$

The outputs from all groups are concatenated along the channel dimension and a final 1×1 convolution maps the output back to the original channel dimension. We keep the residual connection to the input, forming the intermediate output $z = W(y) + x$.

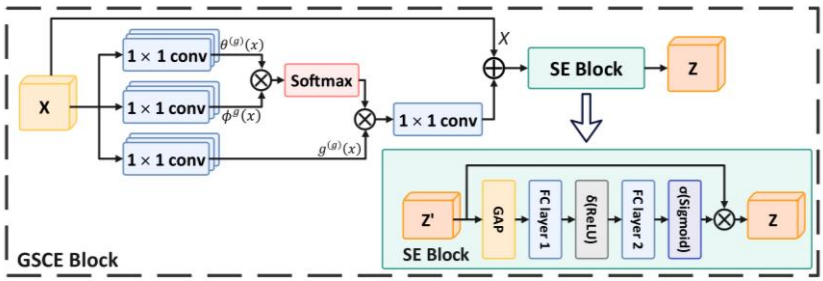

Fig. 4. The structure of our Grouped Spatial-channel Excitation block (GSCE). By leveraging grouped convolution, GSCE captures global spatial dependencies while flexibly learning fine-grained information across different feature subspaces. Softmax activation function can avoid the extreme values of the weight matrix, while the introduction of SE block can enhance the expressive ability of the block.

SE Block. Refer to [50], to further enhance the model's sensitivity to feature channels, we also integrate an SE Block (Squeeze-and-Excitation). As shown in **Fig. 4**, the SE block recalibrates the feature through two phases: Squeeze and Excitation. In the Squeeze stage, SE Block uses global average pooling (GAP) to aggregate spatial information into a vector s_c :

$$s_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} z_c(i,j), \quad \forall c \in \{1,2,\dots,C\} \tag{4}$$

In the Excitation stage, it generates channel-wise weight calibration by flowing through two fully connected layers with ReLU and Sigmoid activation functions:

$$w = \sigma\big(W_2 \delta(W_1 s)\big) \tag{5}$$

Here $W_1 \in R^{\frac{C}{r} \times C}$ and $W_2 \in R^{C \times \frac{C}{r}}$ represent two FC layers, the final weight w is expanded to match the input feature dimensions and applied to reweight the feature map:

$$z' = w \cdot z \tag{6}$$

By integrating the GSCE and SE blocks, the model captures global dependencies across different orientations and channels, accentuating essential regions and suppresses less significant features, ultimately enhancing its representational capacity.

3.2 VI-TMF Module

To address modality gap in VI-ReID task, we design a Visible-infrared Transition Modality Fusion module (VI-TMF). As shown in **Fig. 5**, dual-stream backbone extracts visible and infrared features as inputs. To capture both global and local details, the module applies adaptive average pooling and max pooling and concatenate the output along the channel dimension, forming aggregated representations for both modalities:

$$F_{vis} = concat\left(P_{avg}(f_{vis}), P_{max}(f_{vis})\right), \quad F_{ir} = concat\left(P_{avg}(f_{ir}), P_{max}(f_{ir})\right) \quad (7)$$

Next, we define the operation of generating the transition modality by adding the F_{vis} and F_{ir} and feed it into MLP and Softmax activation functions, ensuring dynamic weighting of each modality:

$$a_{vis}, a_{ir} = \text{Softmax}\big(\text{MLP}(F_{vir} + F_{ir})\big) \quad (8)$$

$$f_{tran} = a_{vis} f_{vis} + a_{ir} f_{ir} \quad (9)$$

In this way, the entire VI-TMF module can not only adaptively learn the importance of features in different modalities but also achieve feature fusion through weight adjustment, making the cross-modality features more consistent and thereby enhancing the matching capability.

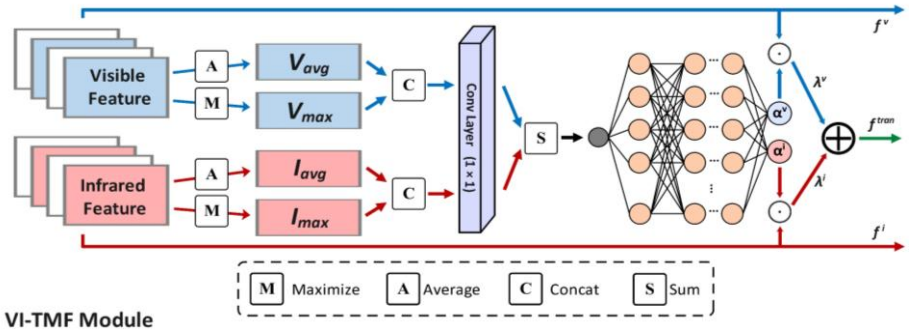

Fig. 5. The structure of VI-TMF module. The module fuses visible and infrared features by generating two combination factors, allowing the model to obtain modality-specific representations while dynamically constructing a transition modality.

3.3 Loss function

Identity Classification Loss. A cross-entropy loss is applied on both the visible-infrared and transition modality features. For an input batch of RGB–IR image pairs $\{x_i^{RGB}, x_i^{IR}\}_{i=1}^{N}$ with identity labels y_i, the identity loss is defined as:

$$\mathcal{L}_{id} = -\frac{1}{N}\sum_{i=1}^{N} y_i \log p\,(y_i|x_i) \quad (10)$$

Additionally, a transition modality identity loss $\mathcal{L}_{id}^{tran}$ is also used to enforce hierarchical feature learning, its calculation method is basically the same.

$$\mathcal{L}_{id}^{tran} = -\frac{1}{N}\sum_{i=1}^{N} y_i \log p\,(y_i|x_i^{tran}) \quad (11)$$

Cross-modality Bridging Loss. Additionally, to bridge the gap between different modalities, we introduced a bridging loss function based on maximum mean discrepancy

(MMD). It minimizes the statistical distance between feature distributions in a reproducing kernel Hilbert space (RKHS) from different modalities. As an illustration, we first discuss the bridging loss for the visible and infrared modality. We first map the visible feature f_{vis} and the infrared feature f_{ir} into the RKHS using a mapping function $\phi(\cdot)$, thereby obtaining their means respectively:

$$\mu_{vis} = \frac{1}{N}\sum_{i=1}^{N}\phi\left(f_v(x_i)\right), \quad \mu_{ir} = \frac{1}{N}\sum_{i=1}^{N}\phi\left(f_i(x_i)\right) \tag{12}$$

We defined the cross-modality bridging loss as the Euclidean distance between them:

$$\mathcal{L}_{bri}^{v-i} = \|\mu_{vis} - \mu_{ir}\|^2 \tag{13}$$

Practically, we describe the similarity between every pair of samples in a Gaussian kernel as $K\left(z_i, z_j\right)$ .Therefore, the calculation of the bridging loss is partitioned into three components according of the modality: within the visible modality (XX), within the infrared modality (YY), and the cross-modality (XY). Their average values are calculated to get the empirical assessment of the cross-modality bridging loss:

$$\mathcal{L}_{bri}^{v-i} = \frac{1}{N^2}\sum_{i,j}^{N}\left[K\left(f_v(x_i), f_v(x_j)\right) + K\left(f_i(x_i), f_i(x_j)\right) - 2K\left(f_v(x_i), f_i(x_j)\right)\right] \tag{14}$$

As shown in **Fig. 6**, in order to ensure the consistency between the three modes, the entire cross-modality bridging loss covers the bridging loss between every two modalities, and finally consists of three parts:

$$\mathcal{L}_{bri} = \mathcal{L}_{bri}^{v-i} + \mathcal{L}_{bri}^{v-t} + \mathcal{L}_{bri}^{i-t} = \mathcal{L}_{bri}(f_v, f_i) + \mathcal{L}_{bri}(f_v, f_{tran}) + \mathcal{L}_{bri}(f_i, f_{tran}) \tag{15}$$

The final loss function is formulated as:

$$\mathcal{L} = \mathcal{L}_{id} + \lambda_1\mathcal{L}_{id}^{tran} + \lambda_2\mathcal{L}_{bri} \tag{16}$$

where λ_1, λ_2 are the weighs to balance losses.

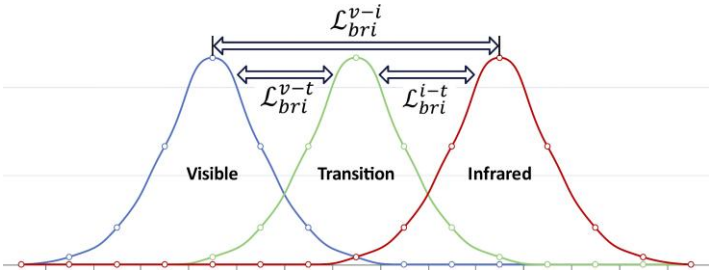

Fig. 6. Cross-modality bridging loss function diagram. We use three bridging losses to represent the Euclidean distance between the features of different modalities in RKHS, and realize the cross-modality features alignment.

4 Experiments

4.1 Datasets and Evaluation Metrics

We conduct comprehensive validation of our model on two cross-modality (visible-infrared) benchmarks: SYSU-MM01 (Wu et al. [12], 2017) and RegDB (Nguyen et al. [40],2017). The SYSU-MM01 dataset consists of 34,166 multimodal images (22,257 visible and 11,909 infrared samples) and includes 395 training identities and 96 test identities. The dataset offers two evaluation modes: all-search mode (hybrid indoor-outdoor scenarios) and indoor-search mode, capturing images using four visible and two infrared cameras under different illumination and conditions. The RegDB dataset has 8,240 paired images across 412 identities, with 206 identities for training and the rest for testing. We perform bidirectional cross-modality retrieval (visible↔infrared) for evaluation. In our experimental setting, referring to the evaluation approach proposed by [16], we used two key metrics to evaluate the performance of two datasets: mean Average Precision (mAP) and cumulative matching features (CMC), which centers on ranking precision, assessing a model's capability to correctly position true matches within ordered candidate lists. In particular, we select the Rank-1 accuracy as the one of evaluation metrics of our method.

4.2 Experimental Methodology

According to the experimental setup, we use Python 3.10.16 and PyTorch 2.6.0+cu124, and train our model on NVIDIA GeForce RTX 4090. As for our method, we adopt the ResNet50 [41] pre-trained on ImageNet [42] as the backbone. All input images are resized to a fixed resolution of 288 × 144 pixels. For training data augmentation, we apply horizontal flipping, random cropping and random erasing (with an erasing probability of ρ=0.5).The architecture is optimized using the Stochastic Gradient Descent (SGD) algorithm, configured with a momentum parameter of 0.9 and a weight decay rate of 5×10^{-4}. During training, the learning rate is initialized to 0.1, reduced by a factor of 0.1 after first 20 epochs, and further adjusted to a decay rate of 0.01 after 20 epochs. The entire model undergoes 80 training epochs. For batch sampling, we randomly select 4 identities per batch, each containing 4 visible and 4 infrared images, resulting in a total batch size of 32 images.

4.3 Visualization Analysis

Visualization of Feature Distribution. In order to visualize the training results of our model, the features before prediction phase were extracted and reduced to two-dimensional space through t-SNE as shown in **Fig. 7**. The points with the same color in the figure indicate that the samples that belong to the same class and to ensure the discriminability of the visualization results, we randomly selected 20 samples in a batch for image rendering. In Fig.7(a), we compare the visual results between the AGW baseline and our TMN. It can be seen that for the baseline model, although it has some feature differentiation ability, many samples from different categories still overlap, limiting its

performance in the VI-ReID task. When we gradually complete all components and apply the TMN method, the model can almost identify all 20 samples, indicating that our TMN method has strong feature differentiation ability and effectiveness. In Fig.7(b), we present the results for different training epochs. After several epochs of training, samples with the same label increasingly cluster together, further visualizing the feature differentiation ability and effectiveness of our method.

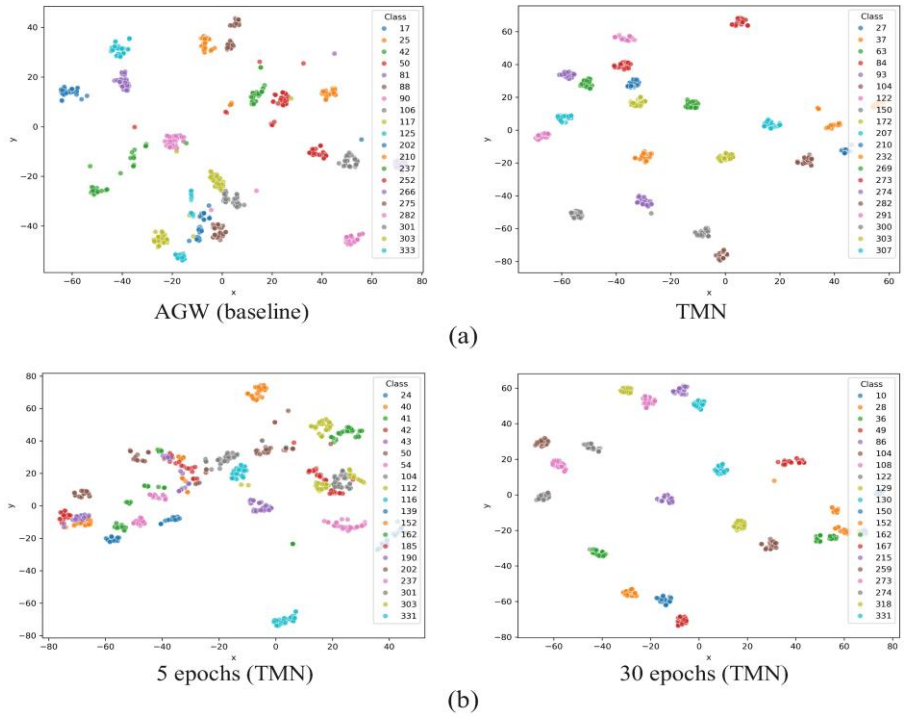

Fig. 7. The feature distribution after t-SNE dimensionality reduction. Compared to the baseline, TMN exhibits strong performance in distinguishing different classes. Furthermore, its performance improves as the model training progresses, confirming the effectiveness and robustness of our method TMN.

Influence of hyperparameters. We evaluated the effect of the tradeoff parameters λ_1 and λ_2 in Eq. (16) on TMN prediction performance. **Fig. 8** shows the effect of parameter values on Rank-1 and mAP in the all-search mode of the SYSU-MM01 and the visible to infrared mode of RegDB dataset. With the other parameter fixed at 1, we gradually increased the values of λ_1 and λ_2 from 0 to 2 for independent experiments. As shown in Fig.8(a), the R1 and mAP of the TMN in the all-search mode of the SYSU-MM01 dataset exhibited a trend of first increasing and then decreasing with the increase of the parameters. When $\lambda_1 = 0.5$ and $\lambda_2 = 0.5$, TMN achieved the best R1 and mAP on the SYSU-MM01 dataset. Using the same method, we conducted parameter analysis on the visible-to-infrared search mode on the RegDB dataset. Similar to the SYSU-MM01 dataset, R1 and mAP on RegDB also showed a trend of first increasing and then

decreasing. However, as shown in Fig.8(b), when $\lambda_1 = 1$ and $\lambda_2 = 0.75$, the best R1 and mAP were achieved on the RegDB dataset. Therefore, during the training of the TMN architecture, we set the tradeoff parameters λ_1 and λ_2 to 0.5 and 0.5, respectively, for training on the SYSU-MM01 dataset, and λ_1 and λ_2 to 1 and 0.75 for training on the RegDB dataset.

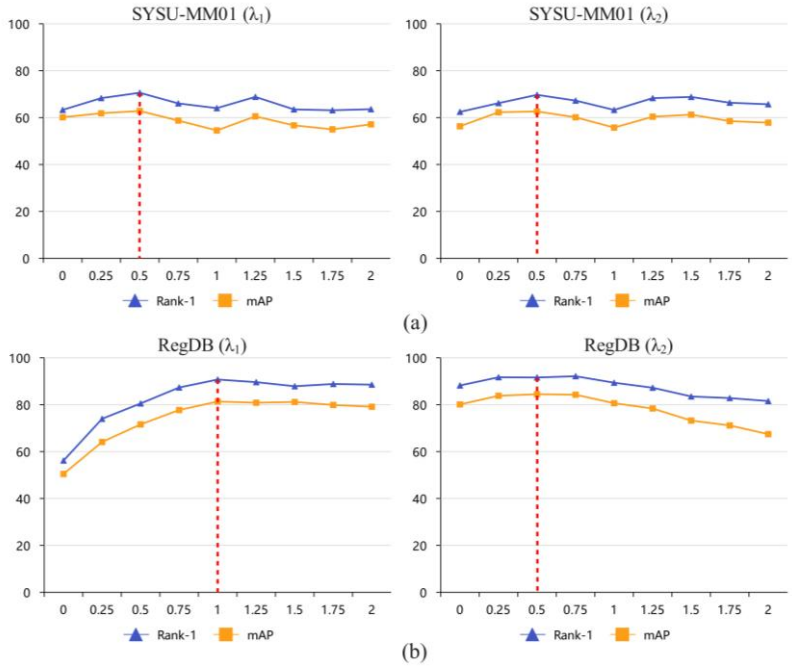

Fig. 8. The influence of tradeoff parameters λ_1 and λ_2 in the TMN loss function. The red dotted line marks the value of the hyperparameter when the evaluation metrics is the highest. For the SYSU-MM01 dataset, the best parameter combination is λ_1=0.5 and λ_2=0.5, while for the RegDB dataset, the best parameter combination is λ_1=1 and λ_2=0.75.

4.4 Comparison with Current Advanced Methods

As shown in **Table 1**, TMN showed excellent performance beyond most of the methods shown in the table, especially on the larger SYSU-MM01 dataset, where our TMN obtained optimal results for both Rank-1 and mAP in both the all-search and indoor-search modes. More specifically, TMN has achieved 71.42% R1 accuracy and 65.91% mAP in the all-search mode, and 73.74% Rank-1 and 77.05% mAP in the indoor-search mode. Compared with other best-performing models that have been listed in the table, TMN has respectively improved Rank-1 and mAP by 1.63% and 2.41% in the all-search mode, and 0.33% of the R1 as well as 0.38% of the mAP accuracy in the indoor-search mode. The remarkable performance on the larger SYSU-MM01 dataset more convincingly demonstrates its generalization and robustness. Even when confronted with a wide variety of image viewpoints and pedestrian motion postures, TMN demonstrates formidable discrimination and matching prowess.

In the RegDB dataset, TMN's R1 performance in the infrared to visible retrieval mode is 0.15% lower than the best result of the existing methods. However, it achieves a best R1 result of 92.14%, 83.25% mAP in the visible to infrared mode and 81.27% mAP in the infrared to visible mode, demonstrating superior performance while improving upon the previous best results by 4.03%, 1.59% and 0.48%. Although TMN does not achieve optimal results across all metrics under every search mode in the two benchmark datasets, its overall performance surpasses most existing models listed in the table, demonstrating its effectiveness in achieving progressive feature alignment by constructing a transition modality.

Table 1. Comparison of TMN and current advanced methods on two benchmark datasets SYSU-MM01 and RegDB datasets. Our TMN has reached the current advanced level of accuracy on most of the evaluation metrics.

Method	**SYSU-MM01**				**RegDB**			
	All-search		Indoor-Search		Vis to Ir		Ir to Vis	
	R1	mAP	R1	mAP	R1	mAP	R1	mAP
Zero-Pad ([12], 2017)	14.80	15.95	20.58	26.92	17.75	18.90	16.63	17.82
HSME ([17], 2019)	20.68	23.12	-	-	50.85	47.00	50.15	46.16
DDAG ([15], 2020)	54.75	53.02	61.02	67.98	69.34	63.46	68.06	61.80
AlignGAN ([33], 2021)	42.40	40.70	45.90	54.30	57.90	53.60	56.30	53.40
AGW (baseline) ([16], 2021)	47.50	47.65	54.17	62.97	70.05	66.37	70.49	65.90
SPOT ([43],2022)	65.34	62.25	69.42	74.63	80.35	72.46	79.37	72.26
G^2DA ([44], 2023)	63.94	60.73	71.06	68.88	73.95	65.49	69.67	61.98
CMTR ([45], 2023)	65.45	62.90	71.46	76.67	88.11	81.66	84.92	80.79
PRAISE ([46], 2024)	59.44	53.27	61.03	66.35	72.54	68.46	73.15	69.85
SDCL ([47], 2024)	64.49	63.24	71.37	73.50	86.91	78.34	80.05	75.32
NLDC ([48],2025)	57.09	51.02	58.24	65.05	84.03	78.34	80.05	75.32
CM^2GT ([49],2025)	69.79	63.50	73.41	76.63	86.72	77.99	**86.47**	77.51
TMN (Ours)	**71.42**	**65.91**	**73.74**	**77.05**	**92.14**	**83.25**	86.32	**81.27**

4.5 Ablation study

As shown in **Table 2**, by gradually introducing a combination of different modules and loss functions, we verify the contribution of each component to the performance of TMN. In the experiment, AGW is used as the baseline, and the initial performance are Rank-1 (R1) 47.50% and average precision mean (mAP) 47.65%. Subsequently, we introduced GSCE and SE modules, and R1 is further increase to 51.45% (+3.95%) and mAP to 49.51% (+1.86%), indicating that global dependency modeling enhanced the ability to capture global features such as pedestrian posture and contour. On this basis, we add VI-IMF module and introduce transition modality identity loss $\mathcal{L}_{id}^{inter}$, further push R1 to 57.85% (+10.30%), mAP to 55.39% (+7.74%). The introduction of visible-

infrared bridging loss $\mathcal{L}_{tri}^{v-i}$ further increases R1 and mAP to 62.82% (+15.32%) and 60.26% (+12.61%) respectively. Finally, when combined the entire cross-modality bridging loss $\mathcal{L}_{tri}$, TMN achieved optimal performance (R1=71.42%, mAP=65.91%), with 23.92% (R1) and 18.26% (mAP) improvements over the baseline.

The ablation study demonstrates that TMN achieves progressive performance enhancement through a hierarchical integration of multi-scale feature extraction and channel attention mechanisms (AGW-GSCE), cross-modal interaction (VI-TMF), and joint loss functions. Each component contributes distinctively: AGW-GSCE captures discriminative patterns across scales and prioritizes critical feature channels, VI-TMF aligns cross-modality representations, and the combined loss functions refine feature consistency. With full components integration, TMN notably outperform the baseline method, highlighting our method's effectiveness and superiority in VI-ReID task.

Table 2. Results of ablation experiments on the SYSU-MM01 dataset. In order to maintain the consistency of the results, our ablation experiments were uniformly conducted in the all-search mode of the SYSU-MM01 dataset.

AGW Baseline	GSCE (without SE)	GSCE (with SE)	VI-TMF $+\mathcal{L}_{id}^{inter}$	$\mathcal{L}_{bri}^{v-i}$	$\mathcal{L}_{bri}$	R1	mAP
√						47.50	47.65
√	√					50.54	49.04
√	√	√				51.45	49.51
√	√	√	√			57.85	55.39
√	√	√	√	√		62.82	60.26
√	√	√	√	√	√	71.42	65.91

5 Conclusion

In this paper, we propose a framework TMN by constructing a transition modality for cross-modality feature interaction and fusion, which addresses the challenge of distribution discrepancies between visible and infrared modalities. We first introduce a Grouped Spatial-channel Excitation block (GSCE) to capture the global feature of pedestrian images, thereby effectively mitigating local feature ambiguities caused by occlusions and viewpoint variations. Subsequently, the Visible-infrared Transition Modality module (VI-TMF) is applied to fuse the dual-modal features, constructing a cross-modality transitional representation that bridges the modality gap and enhances feature compatibility. Based on this, a multi-level loss constraint system is designed: the visible-infrared and transition modality identity loss enforces consistency of similar features in the cross-modality space, and the cross-modality bridging loss forces the alignment of higher-order features between modalities. Extensive experiments on two datasets demonstrate that TMN achieves superior performance. This method, following a progressive optimization path of "transition modality construction – feature extraction and interaction – distribution alignment" realizes hierarchical fusion in cross-modality person, thereby providing robust technical support for all-weather security scenarios.

However, our proposed method is relatively simple. In future work, we can consider multi-level feature fusion and interaction, such as performing multiple feature fusions at different stages of the ResNet backbone. More complex feature fusion methods can also be explored, and comparisons between different fusion strategies can be conducted. These are all directions that need to be further improved in our subsequent work.

References

1. Zheng, L., Yang, Y., Hauptmann, A. G.: Person re-identification: Past, present and future. arXiv preprint arXiv:1610.02984 (2016)
2. Leng, Q., Ye, M., Tian, Q.: A survey of open-world person re-identification. IEEE Transactions on Circuits and Systems for Video Technology 30(4), 1092–1108 (2019)
3. Zheng, Z., Zheng, L., Yang, Y.: Unlabeled samples generated by GAN improve the person re-identification baseline in vitro. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3754–3762 (2017)
4. Li, W., Zhao, R., Xiao, T., Wang, X.: DeepReID: Deep filter pairing neural network for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 152–159 (2014)
5. Wei, L., Zhang, S., Gao, W., Tian, Q.: Person transfer GAN to bridge domain gap for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 79–88 (2018)
6. Yang, F., Yan, K., Lu, S., Jia, H., Xie, X., Gao, W.: Attention driven person re-identification. Pattern Recognition 86, 143–155 (2019)
7. Zhao, J., Wang, H., Zhou, Y., Yao, R., Chen, S., Saddik, A.: Spatial-Channel Enhanced Transformer for Visible-Infrared Person Re-Identification. IEEE Transactions on Multimedia 25, 3668–3680 (2023)
8. Hu, B., Wang, X., Liu, W.: PersonViT: Large-scale Self-supervised Vision Transformer for Person Re-Identification. arXiv:2408.05398 (2024)
9. He, S., Luo, H., Wang, P., Wang, F., Li, H., Jiang, W.: TransReID: Transformer-based object re-identification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 15 013–15 022 (2021)
10. Zhang, G., Zhang, P., Qi, J., Lu, H.: HAT: Hierarchical Aggregation Transformers for person re-identification. In: Proceedings of the 29th ACM International Conference on Multimedia, pp. 516–525 (2021)
11. Wanru, S. O. N. G., Qingqing, Z. H. A. O., Changhong, C., Zong-Liang, G., Feng, L.: Survey on pedestrian re-identification research. CAAI Transactions on Intelligent Systems 12(6), 770–780 (2017)
12. Wu, A., Zheng, W. S., Yu, H. X., Gong, S., Lai, J.: RGB-infrared cross-modality person re-identification. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 5380–5389 (2017)
13. Liu, J., Wang, J., Huang, N., Zhang, Q., Han, J.: Revisiting modality-specific feature compensation for visible-infrared person re-identification. IEEE Transactions on Circuits and Systems for Video Technology 32(10), 7226–7240 (2022)
14. Zhang, D., Zhang, P., Qi, J., Lu, H.: Dual mutual learning for cross-modality person re-identification. IEEE Transactions on Circuits and Systems for Video Technology 32(8), 5361–5373 (2022)

15. Ye, M., Shen, J., Crandall, D. J., Shao, L., Luo, J.: Dynamic dual-attentive aggregation learning for visible-infrared person re-identification. In: Springer European Conference on Computer Vision, pp. 229–247 (2020)
16. Ye, M., Shen, J., Lin, G., Xiang, T., Shao, L., Hoi, S. C.: Deep learning for person re-identification: A survey and outlook. IEEE Transactions on Pattern Analysis and Machine Intelligence 44(6), 2872–2893 (2021)
17. Hao, Y., Wang, N., Li, J., Gao, X.: HSME: Hypersphere manifold embedding for visible thermal person re-identification. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33(01), pp. 8385–8392 (2019)
18. Liu, H., Cheng, J., Wang, W., Su, Y., Bai, H.: Enhancing the discriminative feature learning for visible-thermal cross-modality person re-identification. Neurocomputing 398, 11–19 (2020)
19. Ye, M., Lan, X., Wang, Z., Yuen, P. C.: Bi-directional center-constrained top-ranking for visible thermal person re-identification. IEEE Transactions on Information Forensics and Security 15, 407–419 (2019)
20. Cheng, Y., Xiao, G., Tang, X., Ma, W., Gou, X.: Two-phase feature fusion network for visible-infrared person re-identification. In: 2021 IEEE International Conference on Image Processing (ICIP), pp. 1149–1153 (2021)
21. Qi, J., Zhang, G., et al.: A generative-based image fusion strategy for visible-infrared person re-identification. IEEE Transactions on Circuits and Systems for Video Technology 34(1), 518–533 (2023)
22. Si, T., He, F., Li, P., Gao, X.: Tri-modality consistency optimization with heterogeneous augmented images for visible-infrared person re-identification. Neurocomputing 523, 170–181 (2023)
23. Dai, Y., Liu, J., Sun, Y., Tong, Z., Zhang, C., Duan, L. Y.: IDM: An intermediate domain module for domain adaptive person re-ID. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 11 864–11 874 (2021)
24. Dai, Y., Sun, Y., Liu, J., Tong, Z., Duan, L. Y.: Bridging the source-to-target gap for cross-domain person re-identification with intermediate domains. International Journal of Computer Vision 133(1), 410–434 (2025)
25. Sun, Y., Zheng, L., Yang, Y., Tian, Q., Wang, S.: Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 480–496 (2018)
26. Bai, X., Yang, M., Huang, T., Dou, Z., Yu, R., Xu, Y.: Deep-Person: Learning discriminative deep features for person re-identification. Pattern Recognition 98, 107036 (2020)
27. Zhu, K., Guo, H., Zhang, S., Wang, Y., Liu, J., Wang, J., Tang, M.: AAFormer: Auto-aligned transformer for person re-identification. IEEE Transactions on Neural Networks and Learning Systems (2023)
28. Yan, C., Pang, G., Bai, X., Liu, C., Ning, X., Gu, L., Zhou, J.: Beyond triplet loss: Person re-identification with fine-grained difference-aware pairwise loss. IEEE Transactions on Multimedia 24, 1665–1677 (2021)
29. Zeng, Z., Wang, Z., Wang, Z., Zheng, Y., Chuang, Y. Y., Satoh, S. I.: Illumination-adaptive person re-identification. IEEE Transactions on Multimedia 22(12), 3064–3074 (2020)
30. Zhao, C., Lv, X., Dou, S., Zhang, S., Wu, J., Wang, L.: Incremental generative occlusion adversarial suppression network for person ReID. IEEE Transactions on Image Processing 30, 4212–4224 (2021)
31. Ge, Y., Li, Z., Zhao, H., Yin, G., Yi, S., Wang, X.: FD-GAN: Pose-guided feature distilling GAN for robust person re-identification. Advances in Neural Information Processing Systems 31 (2018)

32. Dai, P., Ji, R., Wang, H., Wu, Q., Huang, Y.: Cross-modality person re-identification with generative adversarial training. In: IJCAI, vol. 1(03), p. 6 (2018)
33. Wang, G. A., Zhang, T., Cheng, J., Liu, S., Yang, Y., Hou, Z.: RGB-infrared cross-modality person re-identification via joint pixel and feature alignment. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3623–3632 (2019)
34. Liu, H., Tan, X., Zhou, X.: Parameter sharing exploration and hetero-center triplet loss for visible-thermal person re-identification. IEEE Transactions on Multimedia 23, 4414–4425 (2020)
35. Zhang, Q., Wang, L., Patel, V. M., Xie, X., Lai, J.: View-decoupled transformer for person re-identification under aerial-ground camera network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 22 000–22 009 (2024)
36. Zhang, P., Wang, Y., Liu, Y., Tu, Z., Lu, H.: Magic Tokens: Select diverse tokens for multi-modal object re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 17 117–17 126 (2024)
37. Ye, M., Lan, X., Li, J., Yuen, P. C.: Hierarchical discriminative learning for visible thermal person re-identification. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32(01) (2018)
38. Ren, K., Zhang, L.: Implicit discriminative knowledge learning for visible-infrared person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 393–402 (2024)
39. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7794–7803 (2018)
40. Nguyen, D. T., Hong, H. G., Kim, K. W., Park, K. R.: Person recognition system based on a combination of body images from visible light and thermal cameras. Sensors 17(3), 605 (2017)
41. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
42. Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255 (2009)
43. Chen, C., Ye, M., Qi, M., Wu, J., Jiang, J., Lin, C.-W.: Structure-aware positional transformer for visible-infrared person re-identification. IEEE Transactions on Image Processing 31, 2352–2364 (2022)
44. Wan, L., Sun, Z., Jing, Q., Chen, Y., Lu, L., Li, Z.: G2DA: Geometry-guided dual-alignment learning for RGB-infrared person re-identification. Pattern Recognition 135, 109150 (2023)
45. Liang, T., Jin, Y., Liu, W., Li, Y.: CMTR: Cross-modality transformer with modality mining for visible-infrared person re-identification. IEEE Transactions on Multimedia 25, 8432–8444 (2023)
46. Liu, Y., Zhang, W., Vasilakos, A. V., Wang, L.: Unsupervised visible-infrared ReID via pseudo-label correction and modality-level alignment. arXiv:2404.06683 (2024)
47. Yang, B., Chen, J., Ye, M.: Shallow-deep collaborative learning for unsupervised visible-infrared person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16870–16879 (2024)
48. Xu, J., Xu, X., Cai, W.: Negative learning and dual contrastive for unsupervised visible-infrared person re-identification. In: ICASSP 2025 – IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 1–5 (2025)

49. Feng, Y., Chen, F., Sun, G., Wu, F., Ji, Y., Liu, T., … Luo, J.: Learning multi-granularity representation with transformer for visible-infrared person re-identification. Pattern Recognition, 111510 (2025)
50. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7132–7141 (2018)