A Multi-Scale Semantic Feature based Relaxed Alignment Network for Visible-Infrared Person Re-identification

Zelin Deng[1], Congyu Liu[1], Ke Nai[1,*], Jiaxin Chen[1], Pei He[2]
[1]School of Computer Science and Technology, Changsha University of Science and Technology, Changsha, China
[2]School of Computer Science and Cyber Engineering, Guangzhou University, Changsha, China
zelin.deng@csust.edu.cn, cyliuwy@163.com, naike_hnu@hnu.edu.cn, jxchen@csust.edu.cn, bk_he@126.com

***Abstract*—Visible-infrared person re-identification (VI-ReID) is a challenging image retrieval task that aims to match the same identity pedestrian images across visible and infrared modalities. Feature alignment is an effective method for identifying identical entities across modalities, yet mainstream models tend to be subject to rigid constraints during alignment, making them highly susceptible to alignment failure and degrading model performance. To address this issue, we propose a Multi-Scale Semantic Feature based Relaxed Alignment Network (MSRANet). Firstly, we propose a Multi-scale Semantic Correlation Enhancement module to fuse high similarity features in adjacent layers to obtain multi-scale semantic features, which can achieve better discriminative ability by utilizing low-level fine-grained features to enhance high-level semantic features. Secondly, by fusing semantic features of pedestrians with the same identity across different scenarios within a modality to enhance feature diversity, we develop a relaxed alignment strategy that achieves coarse-grained cross-modality feature alignment, thereby improving the model's adaptability. Finally, extensive experiments on two public datasets (SYSU-MM01 and RegDB) demonstrate that our MSRANet achieves state-of-the-art performance, significantly outperforming multiple popular methods. Code is available at https://github.com/poptopcan/MSRANet.**

***Index Terms*—VI-ReID, Multi-scale, Relaxed Alignment**

I. Introduction

Re-Identification (Re-ID) is capable of identifying pedestrians of interest from images captured by multiple cameras [1], [2], [40]. Initial studies primarily concentrated on visible single-modality person re-identification, which is unable to adapt to scenarios with substantial variations in lighting conditions. To facilitate all-weather video surveillance, an increasing number of studies have begun to employ both visible and infrared cameras to capture pedestrian images. Due to the significant disparities in visible and infrared modalities, visible-infrared person re-identification(VI-ReID) [3]–[5] has emerged as a highly challenging task.

To identify pedestrians across modalities, aligning features from visible and infrared modalities is a widely adopted and effective research approach. Many studies utilize the complete features of pedestrians to perform feature alignment. Dai et al. [6] use Generative Adversarial Networks (GAN) to perform pixel alignment and alleviate modality information asymmetry. Pan et al. [7] adapt the conditional diffusion model to generate cross-modal and middle-modal images to explicitly reduce the modality discrepancy. Zhuang et al. [9] propose a Camera-based Batch Normalization (BN) operator to ensure that the input distribution is camera independent to realize feature distribution alignment.

However, the methods of aligning complete pedestrian features require matching an excessive number of features, leading to overly rigid constraints and making the model highly susceptible to interference, such as illumination changes and viewpoint variations. To reduce the constraints during the alignment, a partial alignment strategy can be developed by using local features of pedestrians. Huang et al. [10] generate mixed modality images by blending horizontal partial features to reduce pixel-level modality differences. Wei et al. [11] propose a flexible body partition (FBP) model to distinguish part representations automatically. However, partial alignment strategies usually select significant features, which may lead to overly similar features of different pedestrians being chosen and result in incorrect alignment decisions. Therefore, there is growing emphasis on mining semantic information to enhance the discriminative ability of pedestrian features. Zhang et al. [13] explore the embedding subspace to generate embeddings with multi-scale features to learn diverse information. Fang et al. [14] utilize the similarity between pixel-wise features and learnable prototypes to aggregate latent semantic part features to realize more refined semantic alignment.

Although prior methods have advanced cross-modality person re-identification, they heavily rely on local features, which exhibit limited expressive capabilities and diversity. This reliance imposes rigid constraints on models in complex scenarios, thereby hampering adaptability. To address these issues, we propose the Multi-scale Semantic Features based Relaxed Alignment Network (MSRANet). MSRANet computes feature similarity across adjacent layers and fuses highly similar features, enriching high-level semantics with fine-grained details to form multi-scale features that enhance pedestrian representation. Additionally, it constructs cross-modality relaxed patterns by fusing multi-scale features of the

*Corresponding author.

same identity under different scenarios, improving feature diversity and relaxing alignment constraints for better flexibility and adaptability.

Our contributions include:

- We propose a Multi-scale Semantic Features based Relaxed Alignment Network (MSRANet) for VI-ReID, which compensates low-level fine-grained features for high-level semantic features and realizes relaxed cross-modality feature alignment, improving the recognition ability and adaptability of the model.
- We introduce a Multi-scale Semantic Correlation Enhancement (MSCE) module that fuses high-similarity features across adjacent layers to extract multi-scale semantic features, which enables better discriminative ability by leveraging fine-grained low-level features to boost high-level semantic representations.
- We propose a Relaxed Alignment (RA) strategy, which enhances the representation diversity of semantic features by fusing instance features of the same identity and improves the adaptability of the model by relaxing the strict constraints for cross-modality semantic feature alignment.

II. RELATED WORK

A. Visible-Infrared Person Re-ID

Visible-infrared person re-identification (VI-ReID) aims to achieve cross-modality matching between visible images and infrared images, having received considerable interest. To bridge the modality gap, most VI-ReID methods project features of both modalities into a common feature space to mitigate cross-modality discrepancies. Wu et al. [15] first propose a deep zero padding method for training single-stream networks. Subsequently, image-level generation and transformation methods [16], [34] are proposed. These methods mainly adopt generative adversarial networks to generate heterogeneous modality images or features. In order to learn both modality-shared features and modality-specific features simultaneously, Yu et al. [8] dynamically model modality-specific features and modality-shared features to alleviate both cross-modality and intra-modality variations. Ren et al. [3] exploit implicit discriminant information by purifying the specific modality features and refined them into modality-shared features to enhance the feature discriminant ability.

B. Feature Alignment

Feature alignment is one of the key techniques for cross-modality person re-identification and has received continuous and extensive research. Zhao et al. [18] jointly use color-independent consistency learning and identity-aware modality adaptive modules to align identity-level feature distributions. Park et al. [19] enhance pixel-level cross-modality alignment by establishing dense correspondences between cross-modality images in a unified framework. Wang et al. [20] propose a multi-patch modality alignment loss to balance modality differences by mining difficult subspaces and discarding easy subspaces. Wu et al. [21] devise a pattern alignment module, splitting the feature map into multiple patterns to discover subtle differences.

Most existing models use rigid alignment, limiting adaptability to complex cases. To address this, we propose a relaxed alignment strategy that fuses semantic features of the same identity across diverse scenarios within a single modality, thereby effectively enhancing feature diversity and improving the adaptive capacity of the model.

C. Multi-scale Feature Learning

Multi-scale feature learning has been demonstrated to benefit semantic segmentation [22], detection [23] and depth estimation tasks [24]. Xu et al. [25] adopt a lightweight multi-scale feature fusion module to discover the key parts of the camouflaged object through the details, thereby enhancing the detection capability. Han et al. [26] propose a Multi-Scale Multi-Grained sequential network to generate ReID representations robust to scale changes. Li et al. [27] introduce a Multi-scale 3D convolution layer to capture complementary temporal cues, demonstrating superior performance in video-based person re-identification tasks compared to existing methods.

In VI-ReID, pedestrian features exhibit substantial cross-modality discrepancies, while features extracted by existing methods show limited discriminability, resulting in suboptimal feature alignment. To mitigate this, we propose extracting and fusing multi-scale semantic features from adjacent network layers to enhance pedestrian feature representation and improve feature discriminability.

III. METHODOLOGY

A. Overall Architecture

The overall architecture of our proposed MSRANet is shown in Fig. 1, which contains a Multi-scale Semantic Correlation Enhancement (MSCE) module and a Relaxed Alignment (RA) module. MSRANet employs ResNet-50 as backbone, embeds Semantic Correlation Enhancement (SCE) module to extract multi-scale semantic features, and leverages RA to realize cross-modality feature alignment.

In VI-ReID, the visible image set and the infrared image set are usually denoted as $V = \{x_i^{vis} | i = 1, 2, \ldots, N_v\}$ and $R = \{x_j^{inf} | j = 1, 2, \ldots, N_r\}$ respectively, where N_v and N_r represent the quantity of visible images and infrared images respectively.

In each mini-batch, the total number of images is denoted as $N = N_l \times K$, containing $N/2$ images for each modality, where N_l represents the number of identities, K indicates the number of images sampled from a single identity.

B. Multi-scale Semantic Correlation Enhancement Module

In VI-ReID tasks, traditional multi-layer convolutional neural networks (CNNs) are prone to losing fine-grained information, resulting in limited discriminability of pedestrian features and hindering cross-modality feature alignment. To handle this issue, MSCE aims to progressively incorporates semantically relevant fine-grained information from adjacent lower-level

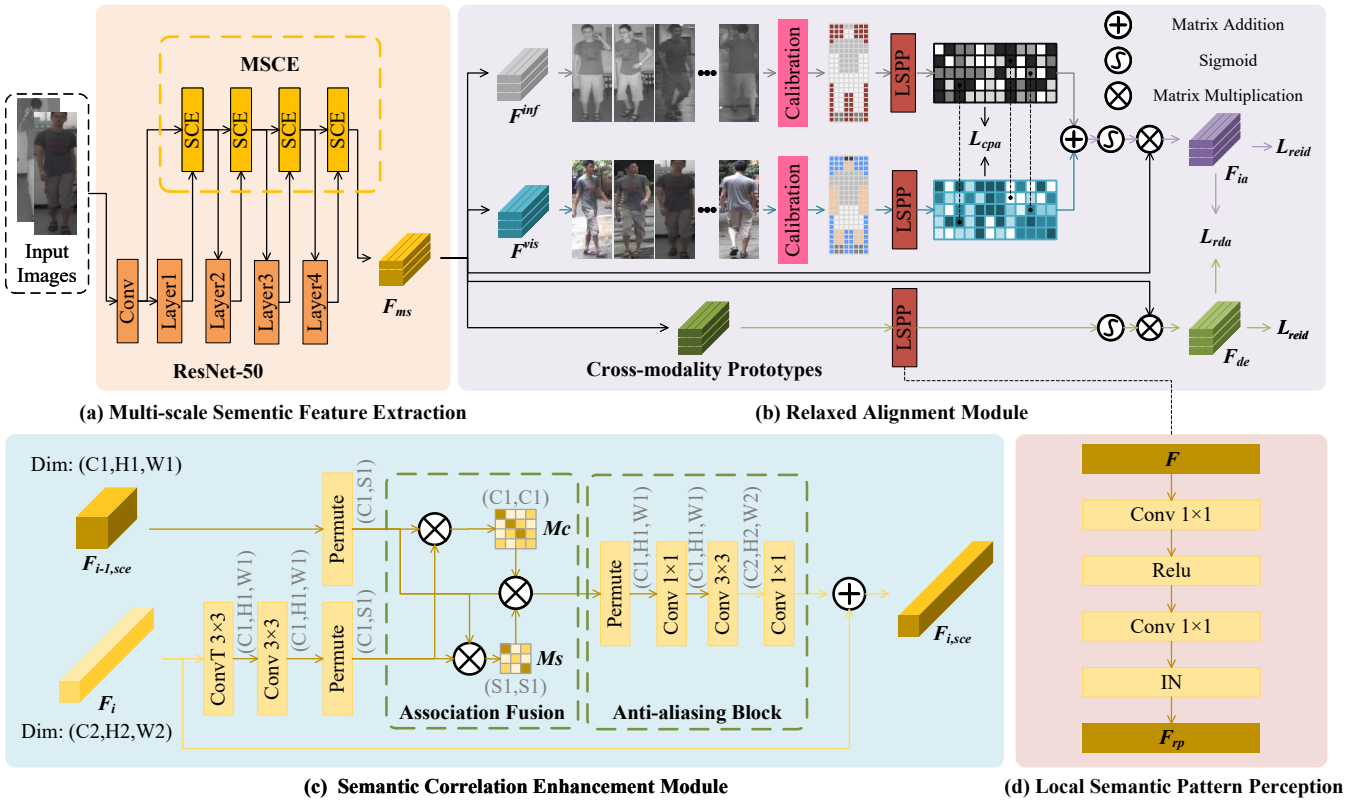

Fig. 1. The overall architecture of the proposed Multi-scale Semantic Feature based Relaxed Alignment (MSRANet) model.

features, ultimately generating more discriminative multi-scale semantic features.

As shown in Fig. 1(a), MSCE includes four SCE layers, in which the low-level fine-grained features are contrapuntally supplemented for high-level semantic features by computing element-wise similarity of adjacent layer features.

For the i-th SCE, as shown in Fig. 1(c), input features consist of the output F_i of the i-th layer (i = 1,2,3,4) and that of the $i-1$-th SCE module, $F_{0,sce} = F_0$.

Then, we project high-level semantic feature F_i into the low-level feature space to obtain $F_{i,l}$:

$$F_{i,l} = \delta_{S,3\times3}(\delta_{T,3\times3}(F_i)), \tag{1}$$

where $\delta_{T,3\times3}$ represents the transposed convolutional layer with 3×3 kernel for upsampling, and $\delta_{S,3\times3}$ represents the convolutional layer for smoothing the transition. Then, we compute the feature similarity between $F_{i,l}$ and $F_{i-1,sce}$ from both channel and spatial dimensions to generate the channel complementary features M_C and spatial complementary features M_S. We propose a Association Fusion (AF) strategy to fuse M_C, $F_{i-1,sce}$ and M_S to obtain refined feature representation $F_{i-1,af}$:

$$F_{i-1,af} = M_C \times F_{i-1,sce} \times M_S, \tag{2}$$

$$M_C = \frac{\exp(F_{i-1,sce} \times (F_{i,l})^T)}{\| \exp(F_{i-1,sce} \times (F_{i,l})^T) \|_1}, \tag{3}$$

$$M_S = \frac{\exp((F_{i-1,sce})^T \times F_{i,l})}{\| \exp((F_{i-1,sce})^T \times F_{i,l}) \|_1}, \tag{4}$$

where $\exp(\cdot)$ represents the exponential function, $\times$ represents matrix multiplication, $\| \cdot \|_1$ represents L1 norm and X^T represents the transpose of X. As the AF operation fuse semantic feature according to element-wise similarity, SCE can dynamically enhance high-level semantic features by contrapuntally supplementing low-level fine-grained features.

Finally, $F_{i-1,af}$ is restored by an Anti-aliasing Block (AB) which can mitigate the aliasing phenomenon, and fused with the high-level semantic feature to perceive feature variations and compensate for semantic information:

$$F_{i,sce} = F_i + \delta_{1\times1}(\delta_{3\times3}(\delta_{1\times1}(F_{i-1,af}))). \tag{5}$$

$F_{4,sce}$ is the final multi-scale semantic feature denoted as F_{ms}.

C. Relaxed Alignment Module

Note that VI-ReID images are collected in complex conditions, instance features have significant local feature differences. Most existing models have limited adaptability and still employ rigid alignment strategies, which inevitably lead to alignment failures and constrain model performance. To solve these issues, we propose dual-branch RA to realize cross-modality semantic feature alignment, as illustrated in Fig. 1(b),

Firstly, we aggregate visible instances F^{vis} and infrared instances F^{inf} of same identity by calibration module:

$$F^{mod} = \delta_{cali}(f_{concat}(F^{mod}_{ms})), \tag{6}$$

where mod denotes the modality ($mod \in \{vis,\ inf\}$), $f_{concat}(\cdot)$ represents concatenating features along channel dimension, δ_{cali} represents extracting internal features by a convolutional layer with 1×1 kernel and reducing the number of channels.

Secondly, as shown in Fig. 1(d), we generate relaxed patterns by local semantic pattern perception module (LSPP):

$$F^{mod}_{rp} = \text{IN}(\delta_{P,1\times1}(\text{RELU}(\delta^{mod}_{1\times1}(F^{mod})))), \tag{7}$$

where RELU denotes the Relu function. $\delta^{mod}_{1\times1}$ and $\delta_{P,1\times1}$ represent the $\text{Conv}_{1\times1}$ layer with C channels, IN denotes Instance Normalization.

Then, the cross-modality pattern alignment is performed under the constraint of cross-modality pattern alignment loss L_{cpa}:

$$\begin{aligned} L_{cpa} = &\text{avg}(1 - \frac{F^{vis}_{rp} \cdot F^{inf}_{rp}}{\|F^{vis}_{rp}\|\|F^{inf}_{rp}\|}) \\ &+\text{avg}(1 - \frac{F^{vis,T}_{rp} \cdot F^{inf,T}_{rp}}{\|F^{vis,T}_{rp}\|\|F^{inf,T}_{rp}\|}), \end{aligned} \tag{8}$$

where $\text{avg}(\cdot)$ represents the mean of all elements. By maximizing the cosine similarity across both channel and spatial dimensions of cross-modality features, the semantic consistence can be enhanced in a more fine granularity manner. Subsequently, F^{vis}_{rp} and F^{inf}_{rp} are fused as F^{m}_{rp}.

Thirdly, we maintain a cross-modality prototype bank to assist generating relaxed patterns that can perceive feature diversity. The prototype bank $P = \{p_j \in \mathbb{R}^{C\times H\times W} | j = 1, 2, \ldots, M\}$ is initialized by [29], where M represents the number of prototypes, it is set to 30. For each p in P, we randomly select an instance F_{ins} to update p and select the top-K samples closest to p (denoted as F_p) for aggregating local semantic features with diversity:

$$p = r2 \cdot F_p + (1 - r2) \cdot (r1 \cdot F_{ins} + (1 - r1) \cdot p), \tag{9}$$

where $r1$ and $r2$ are the update rate, which are set to 0.1 and 0.2 respectively. It is worth noting that the features F_{ms} vary in each mini-batch. Equation (9) guarantees the diversity and representational capabilities of prototype patterns by introducing subtle random perturbations and aggregating stable, similar local semantic features across multiple samples.

To form the identity-irrelevant semantic embedding F_{np}, for each sample, we selecte K nearest prototype patterns (denoted as F_k) from the prototype based on pairwise distances:

$$F_{np} = \delta_{1\times1}(f_{concat}(F_k)), \tag{10}$$

F_{np} is then fed into LSPP to generate cross-modality relaxed patterns F^{cm}_{rp}. Then, we fuse F_{ms} with relaxed patterns through attention residual connection, generating identity-aware feature F_{ia} and diversity enhanced feature F_{de}:

$$F_{ia} = F_{ms} + F_{ms} \cdot \text{Sigmoid}(F^{m}_{rp}), \tag{11}$$

$$F_{de} = F_{ms} + F_{ms} \cdot \text{Sigmoid}(F^{cm}_{rp}), \tag{12}$$

F_{ia} and F_{de} are then converted into $logits_{ia}$ and $logits_{de}$ via batch normalization (BN) and separate classifiers.

Finally, we perform relaxed distillation alignment L_{rda} across branches in RA to achieve feature distribution alignment, simultaneously enhancing feature diversity and feature discrimination:

$$L_{rda} = \sum_{i=1}^{N} \Big(\text{KL}\left(\text{A}(logits_{ia}), \text{A}(logits_{de})\right) \Big). \tag{13}$$

where KL represents the Kullback-Leibler (KL) divergence, A represents the attention module which contains spatial attention and channel attention [41]. L_{rda} loss realizes mutual knowledge distillation between branches and ensure the extracted features are both discriminative and robust.

D. Overall Objective Function

In this paper, we adopt the widely-used loss L_{reid} as our baseline loss functions. The Loss L_{reid} contains the identity loss L_{id} and the triplet loss $L_{triplet}$.

The total loss L of MSRANet is defined as

$$L = L_{reid} + \lambda \cdot L_{cpa} + L_{rda}. \tag{14}$$

In the formula, λ is the loss coefficient and is set to 0.8.

IV. Experiments

A. Experimental Settings

Datasets: SYSU-MM01 is a large-scale dataset for VI-ReID, which includes 491 identities collected by 4 visible cameras and 2 near-infrared cameras. For test modes, all camera images are used for the all-search mode, and the images captured by 1st, 2nd, and 3rd cameras are used for the indoor-search mode. The visible and infrared images are serve as the gallery and query sets, respectively, in both test modes. RegDB dataset is recorded by a dual-camera system, in which visible images and infrared images are captured in pairs. RegDB contains 412 identities, where 206 identities are randomly selected for training, and the remaining for testing.

Evaluation metrics: The cumulative matching characteristic (CMC), mean average precision (mAP), and mean inverse negative penalty (mINP) are adopted as evaluation metrics.

Implementation details: All the experiments are implemented on a single NVIDIA A6000 GPU. All input images are resized to 384×144. The mini-batch size is set to 120, including 12 random identities with 5 visible images and 5 infrared images. The Adam optimizer is used with an initial learning rate of 3.5×10^{-5}.

B. Comparison with State-of-the-art Methods

We compare the proposed MSRANet model with state-of-the-art VI-ReID methods on two public datasets SYSU-MM01 and RegDB. The comparison results of SYSU-MM01 and RegDB in terms of CMC, mAP, and mINP are listed in Table I and Table II, respectively.

TABLE I
COMPARISON OF CMC (%), MAP (%) AND MINP (%) PERFORMANCES WITH THE STATE-OF-THE-ART METHODS ON SYSU-MM01 DATASET. BEST RESULTS ARE HIGHLIGHTED IN BOLD, AND SECOND-BEST RESULTS ARE UNDERLINED.

Method	Publish	All-Search						Indoor-Search					
		Single-Shot			Multi-Shot			Single-Shot			Multi-Shot		
		R=1	mAP	mINP	R=1	mAP	mINP	R=1	mAP	mINP	R=1	mAP	mINP
Zero-Pad [15]	ICCV 17	14.80	15.95	-	19.13	10.89	-	20.58	26.92	-	24.43	18.64	-
BDTR [33]	IJCAI 18	17.01	19.66	-	-	-	-	-	-	-	-	-	-
D^{2}RL [34]	CVPR 19	28.90	29.20	-	-	-	-	-	-	-	-	-	-
AlignGAN [17]	ICCV 19	42.40	40.70	-	51.50	33.90	-	45.90	54.30	-	57.10	45.30	-
Hi-CMD [16]	CVPR 20	34.94	35.94	-	-	-	-	-	-	-	-	-	-
MPANet [21]	CVPR 21	70.58	68.24	-	75.58	62.91	-	76.74	80.95	-	84.22	75.11	-
CMT [35]	ECCV 22	71.88	68.57	-	80.23	63.13	-	76.9	79.91	-	84.87	74.11	-
DEEN [13]	CVPR 23	74.70	71.80	-	-	-	-	80.30	83.30	-	-	-	-
MUN [8]	ICCV 23	76.24	73.81	-	-	-	-	79.42	82.06	-	-	-	-
PartMix [12]	CVPR 23	77.78	74.62	-	80.54	69.84	-	81.52	84.38	-	87.99	79.95	-
CAJ [36]	TPAMI 24	71.48	68.15	54.57	-	-	-	78.36	81.98	78.44	-	-	-
DMA [28]	TIFS 24	74.57	70.41	-	-	-	-	80.30	83.30	-	-	-	-
HOS-Net [37]	AAAI 24	75.60	74.20	-	-	-	-	84.20	86.70	-	-	-	-
IDKL [3]	CVPR 24	81.42	79.85	-	84.34	78.22	-	87.14	89.37	-	94.30	88.75	-
MIP [38]	TCSVT 25	76.92	75.84	-	78.59	68.55	-	83.53	85.88	-	87.04	80.85	-
STAR [30]	TIFS 25	82.93	80.47	-	-	-	-	88.04	89.58	-	-	-	-
MDANet [32]	TMM 25	79.46	77.03	63.50	**89.21**	70.40	-	86.71	91.62	80.03	93.37	81.20	-
MSRANet (ours)		**84.61**	**82.48**	**73.28**	87.00	**79.43**	**38.52**	**91.72**	**92.30**	**90.05**	**96.41**	**91.52**	**67.27**

Comparisons on SYSU-MM01. The results of the comparative experiments are shown in Table I. The results show that the MSRANet is superior to most existing cutting-edge methods. Specifically, MSRANet achieves an accuracy of 84.61% in Rank-1 and 82.48% in mAP in the all-search mode. Meanwhile, for indoor-search mode, the accuracy of Rank-1 at 91.72% and mAP at 92.30% are achieved.

Comparisons on the RegDB. As shown in the Table II, we evaluate MSRANet on RegDB dataset. Compared with the existing methods, MSRANet achieves the best performance in terms of Rank-1, mAP, and mINP in both the V2I and I2V modes. In the V2I mode, MSRANet achieves the accuracy of 96.50% Rank-1 and 92.32% mAP. In the I2V mode, MSRANet achieves the accuracy of 95.49% Rank-1 and 92.11% mAP.

TABLE II
COMPARISON OF CMC (%), MAP (%) AND MINP (%) PERFORMANCES WITH THE STATE-OF-THE-ART METHODS ON REGDB DATASET.

Method	Publish	V2I		I2V	
		R=1	mAP	R=1	mAP
Zero-Pad [15]	ICCV 17	17.75	18.90	16.63	17.82
D^{2}RL [34]	CVPR 19	43.40	44.10	-	-
AlignGAN [17]	ICCV 19	57.90	53.60	56.30	53.40
Hi-CMD [16]	CVPR 20	70.93	66.04	-	-
MPANet [21]	CVPR 21	83.70	80.90	82.80	80.70
CMT [35]	ECCV 22	95.17	87.30	91.97	84.46
SAAI [14]	ICCV 23	91.07	91.45	92.09	92.01
CAL [39]	ICCV 23	94.51	88.67	93.64	87.61
MUN [8]	ICCV 23	95.19	87.15	91.86	85.01
DMA [28]	TIFS 24	93.30	88.34	91.50	86.80
HOS-Net [37]	AAAI 24	94.70	90.40	93.30	89.20
IDKL [3]	CVPR 24	94.72	90.19	94.22	90.43
STAR [30]	TIFS 25	90.48	91.77	91.89	**93.31**
MDANet [32]	TMM 25	90.52	81.74	92.37	82.16
MIP [38]	TCSVT 25	90.32	87.59	91.98	88.11
DSAF [31]	TMM 25	92.62	86.37	93.25	87.17
ours		**96.50**	**92.32**	**95.49**	92.11

C. Ablation Study

In this section, we conduct ablation experiments to evaluate the performance of each module.

Effectiveness of each component. We evaluate the effectiveness of the proposed modules on the SYSU-MM01 dataset. As shown in Table III, when MSCE is integrated into the baseline, the Rank-1, mAP, mINP accuracy of the model are improved by 8.83%, 11.28% and 11.89%. Furthermore, when L_{cpa} and L_{rda} are incorporated alongside the baseline, the model's Rank-1, mAP, and mINP performance is improved by 17.66%, 21.35% and 25.36%.

D. Visualization Analysis

To further illustrate the validity of MSRANet, visual comparison results are presented in this section. In Fig. 2, by comparing the attention heatmaps of the baseline and MSRANet, it can be observed that our model finds more person semantic cues and can identify more localized discriminative regions. In addition, Fig. 3(a) shows that there are significant modality discrepancies between different instances of the same identity in the baseline. Fig. 3(b) shows the reduction of the modality gap and cluster area, which is realized by the proposed RA strategy.

V. CONCLUSION

This paper provides a novel relaxed alignment network based on multi-scale semantic features. Firstly, the proposed MSCE module aims to obtain powerful multi-scale semantic features for robust feature representation. By compensating

TABLE III
ABLATION STUDIES OF THE PROPOSED MSRANET ON THE SYSU-MM01 DATASET. RANK-1 (%), MAP (%), AND MINP (%) ARE REPORTED. W/O AF MEANS WITHOUT THE ASSOCIATION FUSION MODULE.

Settings				ALL-Search			Indoor-Search		
w/oAF	MSCE	L_{cpa}	L_{rda}	R=1	mAP	mINP	R=1	mAP	mINP
				66.88	61.13	47.45	69.91	73.88	67.57
√	√			74.06	69.12	56.20	80.39	82.04	78.11
	√			75.71	72.41	59.34	82.37	85.14	81.23
		√		83.73	81.99	71.85	90.36	91.00	88.52
		√	√	84.54	82.41	72.81	91.13	91.53	89.11
	√	√	√	**84.61**	**82.48**	**73.28**	**91.72**	**92.30**	**90.05**

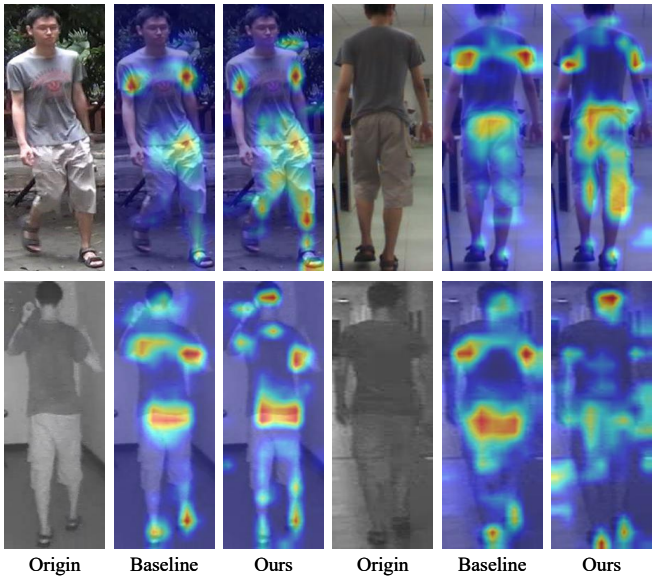

Fig. 2. Visualization of GradCAM on the SYSU-MM01 dataset.

low-level fine-grained features for high-level semantic features, MSCE can fully utilize the advantages of low-level fine-grained features and high-level semantic features to construct multi-scale semantic features, achieving better discriminative capability. Secondly, the proposed RA module attempts to perform feature alignment in a more flexible, coarse-grained manner. By constructing the cross-modality relaxed pattern from multi-scenario instances of the same identity within a single modality, RA can implicitly utilize multiple pedestrian samples to relax rigid constraints to realize relax feature alignment, thereby effectively enhancing the model's adaptability. The performance improvement on two benchmark datasets validates our method's effectiveness.

ACKNOWLEDGMENT

This work is supported in part by the National Natural Science Foundation of China under Grants 62106071; the Hunan Provincial Key Research and Development Program under Grant 2024AQ2027, 2025AQ2022; the Research Foundation of Education Bureau of Hunan Province, China under Grant 25B0221; the Natural Science Foundation of Changsha under Grant No. kq2202215; the Practical Innovation and Entrepreneurship Enhancement Program for Professional Degree Postgraduates of Changsha University of Science and Technology under Grant CLSJCX25071.

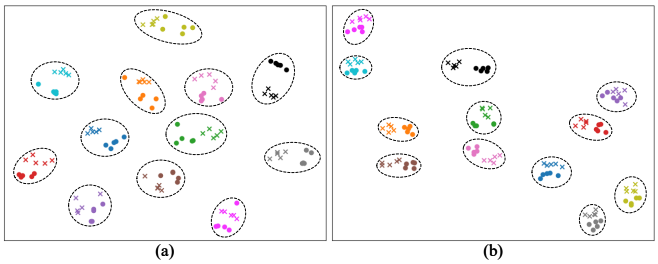

Fig. 3. Visualization of feature distributions by t-SNE. Figure a depicts the feature distribution of the baseline, while Figure b illustrates that of our method. Different colors represent different identity, and different shapes represent different modality.

REFERENCES

[1] Z. Deng, M. Tang, K. Nai, G. Li, S. Chen, and P. He, "A semantic-guided occlusion simulation based local feature semantic expansion network for person re-identification," *Pattern Recognition*, p. 112567, 2026.
[2] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, "Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 480–496.
[3] K. Ren and L. Zhang, "Implicit discriminative knowledge learning for visible-infrared person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recogn.*, 2024, pp. 393–402.
[4] M. Ye, C. Chen, J. Shen, and L. Shao, "Dynamic tri-level relation mining with attentive graph for visible infrared re-identification," *IEEE Trans. Inf. Forensics Security*, vol. 17, pp. 386–398, 2022.
[5] X. Li, Y. Lu, B. Liu, Y. Liu, G. Yin, et al., "Counterfactual intervention feature transfer for visible-infrared person re-identification," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 381–398.
[6] H. Dai, Q. Xie, Y. Ma, Y. Liu, and S. Xiong, "RGB-infrared person re-identification via image modality conversion," in *Proc. Int. Conf. Pattern Recognit.*, 2021, pp. 592–598.
[7] H. Pan, W. Pei, X. Li, and Z. He, "Unified conditional image generation for visible-infrared person re-identification," *IEEE Trans. Inf. Forensics Security*, vol. 19, pp. 9026–9038, 2024.
[8] H. Yu, X. Cheng, W. Peng, W. Liu, and G. Zhao, "Modality unifying network for visible-infrared person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 11185–11195.
[9] Z. Zhuang, L. Wei, L. Xie, H. Ai, and Q. Tian, "Camera-based batch normalization: An effective distribution alignment method for person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 1, pp. 374–387, 2022.
[10] Z. Huang, J. Liu, L. Li, K. Zheng, and Z.-J. Zha, "Modality-adaptive mixup and invariant decomposition for RGB-infrared person re-identification," in *Proc. AAAI Conf. Artif. Intell.*, vol. 36, 2022, pp. 1034–1042.
[11] Z. Wei, X. Yang, N. Wang, and X. Gao, "Flexible body partition-based adversarial learning for visible infrared person re-identification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 9, pp. 4676–4687, 2022.
[12] M. Kim, S. Kim, J. Park, S. Park, and K. Sohn, "Partmix: Regularization strategy to learn part discovery for visible-infrared person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recogn.*, 2023, pp. 18621–18632.
[13] Y. Zhang and H. Wang, "Diverse embedding expansion network and low-light cross-modality benchmark for visible-infrared person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recogn.*, 2023, pp. 2153–2162.
[14] X. Fang, Y. Yang, and Y. Fu, "Visible-infrared person re-identification via semantic alignment and affinity inference," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 11236–11245.
[15] A. Wu, W.-S. Zheng, H.-X. Yu, S. Gong, and J. Lai, "RGB-infrared cross-modality person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 5380–5389.

[16] S. Choi, S. Lee, Y. Kim, T. Kim, and C. Kim, "Hi-cmd: Hierarchical cross-modality disentanglement for visible-infrared person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recogn.*, 2020, pp. 10257–10266.

[17] G. Wang, T. Zhang, J. Cheng, S. Liu, Y. Yang, and Z. Hou, "RGB-infrared cross-modality person re-identification via joint pixel and feature alignment," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 3622–3631.

[18] Z. Zhao, B. Liu, Q. Chu, Y. Lu, and N. Yu, "Joint color-irrelevant consistency learning and identity-aware modality adaptation for visible-infrared cross modality person re-identification," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, 2021, pp. 3520–3528.

[19] H. Park, S. Lee, J. Lee, and B. Ham, "Learning by aligning: Visible-infrared person re-identification using cross-modal correspondences," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 12046–12055.

[20] P. Wang, Z. Zhao, F. Su, et al., "Deep multi-patch matching network for visible thermal person re-identification," *IEEE Trans. Multimedia*, vol. 23, pp. 1474–1488, 2021.

[21] Q. Wu, P. Dai, J. Chen, et al., "Discover cross-modality nuances for visible-infrared person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recogn.*, 2021, pp. 4330–4339.

[22] Q. Tong, Z. Yuan, X. Liao, et al., "A joint multi-scale convolutional network for fully automatic segmentation of the left ventricle," in *Proc. IEEE Int. Conf. Image Process.*, 2017, pp. 3110–3114.

[23] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, 2017, pp. 936–944.

[24] X. Xu, Z. Chen, and F. Yin, "Monocular depth estimation with multi-scale feature fusion," *IEEE Signal Process. Lett.*, vol. 28, pp. 678–682, 2021.

[25] Z. Xu, Z. Wang, H. Wang, C. Liu, Y. Zhou, and M. Sun, "Boosting lightweight camouflaged object detection with multi-scale context and boundary awareness," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2025, pp. 1–5.

[26] Z. Han and B. Ma, "Enhancing identification for person search with multi-scale multi-grained representation learning," *Pattern Recognition*, vol. 150, p. 110361, 2024.

[27] J. Li, S. Zhang, and T. Huang, "Multi-scale temporal cues learning for video person re-identification," *IEEE Trans. Image Process.*, vol. 29, pp. 4461–4473, 2020.

[28] Z. Cui, J. Zhou, and Y. Peng, "DMA: Dual modality-aware alignment for visible-infrared person re-identification," *IEEE Trans. Inf. Forensics Security*, vol. 19, pp. 2696–2708, 2024.

[29] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1026–1034.

[30] Y. Qiu, L. Wang, W. Song, J. Liu, Z. Shi, and N. Jiang, "Advancing visible-infrared person re-identification: Synergizing visual-textual reasoning and cross-modal feature alignment," *IEEE Trans. Inf. Forensics Security*, vol. 20, pp. 2184–2196, 2025.

[31] Y. Jiang, X. Cheng, H. Yu, X. Liu, H. Chen, and G. Zhao, "DSAF: Dual space alignment framework for visible-infrared person re-identification," *IEEE Trans. Multimedia*, pp. 1–13, 2025.

[32] X. Cheng, H. Yu, K. H. M. Cheng, Z. Yu, and G. Zhao, "MDANet: Modality-aware domain alignment network for visible-infrared person re-identification," *IEEE Trans. Multimedia*, vol. 27, pp. 2015–2027, 2025.

[33] M. Ye, Z. Wang, X. Lan, and P. C. Yuen, "Visible thermal person re-identification via dual-constrained top-ranking," in *Proc. Int. Joint Conf. Artif. Intell.*, 2018, pp. 1092–1099.

[34] Z. Wang, Z. Wang, Y. Zheng, Y.-Y. Chuang, and S. Satoh, "Learning to reduce dual-level discrepancy for infrared-visible person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recogn.*, 2019, pp. 618–626.

[35] K. Jiang, T. Zhang, X. Liu, B. Qian, Y. Zhang, and F. Wu, "Cross-modality transformer for visible-infrared person re-identification," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 480–496.

[36] M. Ye, Z. Wu, C. Chen, and B. Du, "Channel augmentation for visible-infrared re-identification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 4, pp. 2299–2315, 2024.

[37] L. Qiu, S. Chen, Y. Yan, J.-H. Xue, D.-H. Wang, and S. Zhu, "High-order structure based middle-feature learning for visible-infrared person re-identification," in *Proc. AAAI Conf. Artif. Intell.*, vol. 38, 2024, pp. 4596–4604.

[38] R. Wu, B. Jiao, M. Liu, S. Wang, W. Wang, and P. Wang, "Enhancing visible-infrared person re-identification with modality- and instance-aware adaptation learning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 35, pp. 8086–8103, 2025.

[39] J. Wu, H. Liu, Y. Su, W. Shi, and H. Tang, "Learning concordant attention via target-aware alignment for visible-infrared person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 11088–11097.

[40] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. H. Hoi, "Deep learning for person re-identification: A survey and outlook," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 6, pp. 2872–2893, 2022.

[41] S. Woo, J. Park, J.-Y. Lee, and I.-S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.