HAResformer: A Hybrid ResNet-Transformer Hierarchical Aggregation Architecture for Visible-Infrared Person Reidentification

Yongheng Qian and Su-Kit Tang, *Member, IEEE*

***Abstract*—Modality differences and intramodality variations make the visible-infrared person reidentification (VI-ReID) task highly challenging. Most existing methods focus on building network frameworks based on convolutional neural networks (CNN) or pure vision transformers (ViT) to extract discriminative features and address these challenges. However, these methods neglect several key issues: deeply fusing local features with global spatial information enhances comprehensive discriminative representation, patch tokens contain rich semantic information, and different feature extraction stages within the network emphasize various semantic elements. To address these issues, we propose a novel hybrid ResNet-transformer hierarchical aggregation architecture named HAResformer. HAResformer comprises three key components: 1) hierarchical feature extraction (HFE) framework; 2) deeply supervised aggregation (DSA); and 3) hierarchical global aggregate encoder (HGAE). Specifically, HFE introduces a lightweight cross-encoder feature fusion module (CFFM) to deeply integrate the local features and global spatial information of a person extracted by the ResNet encoder (RE) and transformer encoder (TE). Subsequently, the fused features are fed as global priors into the next-stage TE for deep interaction, aiming to extract specific local features and global contextual clues. Additionally, DSA and HGAE provide auxiliary supervision and aggregation on multiscale features to enhance multigranularity feature representation. HAResformer effectively alleviates modality differences and reduces intramodality variations. Extensive experiments on three benchmarks demonstrate the effectiveness and generalization of our architecture and outperform most state-of-the-art methods. HAResformer has the potential to become a new VI-ReID baseline, promoting high-quality research in the future.**

***Index Terms*—Convolutional neural networks (CNN), cross-modality, feature fusion, multiscale supervision, person reidentification (ReID), vision transformer.**

I. Introduction

PERSON reidentification (ReID) [1] plays a key role in social security, smart cities, and locating missing persons, which aims to match the person of interest across multiple nonoverlapping camera views. Some existing works [2], [3], [4], [5] have achieved unprecedented success in visible-visible person ReID (VV-ReID), with retrieval accuracy even surpassing human capabilities. However, in real scenarios, criminals often choose to act at night. Consequently, we must work with pedestrian images collected from poorly illuminated environments. Under these conditions, existing VV-ReID methods cannot adequately capture the appearance information of the person. With the development and widespread deployment of intelligent surveillance devices, cameras can automatically switch between visible and infrared modes based on illumination conditions. To address the shortcomings of VV-ReID, visible-infrared person ReID (VI-ReID) [6] was proposed and attracted extensive attention from researchers.

The VI-ReID task aims to match person images across disjoint visible and infrared camera views. Specifically, given a visible (infrared) query image, the goal is to retrieve a person of the same identity from the infrared (visible) gallery [8], [9]. VI-ReID operates continuously, making it more suitable for practical applications. However, VI-ReID faces two key challenges: 1) modality differences caused by the imaging principles of visible and infrared images and 2) intramodality variations, such as viewpoint, occlusion, pose, and illumination. To overcome these challenges, existing works focus on building VI-ReID network frameworks using convolutional neural networks (CNN) [10] or vision transformers (ViT) [11], as shown in Fig. 1(a) and (b). CNN-based VI-ReID methods [12], [13], [14] typically decouple input features into modality-specific and modality-shared features, then implement feature-level alignment in the modality-share embedding space to alleviate modality differences. However, the limited local receptive field of CNNs cannot capture global spatial context information. As shown in the second row of Fig. 1(d), CNN-based VI-ReID methods focus on the attention map of local regions of the human body (e.g., feet, hands, and legs), making it challenging to capture discriminative features of multiple different parts. Furthermore, stride convolution and downsample operations lead to the loss of fine-grained cues, restricting the discriminative feature representation of CNN-based VI-ReID methods. Due to the dynamic self-attention mechanism of ViT and its ability to model long-range dependencies, pure ViT-based VI-ReID methods [15], [16] have become popular. Compared with CNN-based methods, pure ViT-based VI-ReID methods focus more on different body parts to enhance global spatial feature representation,

Received 18 November 2024; revised 14 December 2024 and 24 January 2025; accepted 1 March 2025. Date of publication 4 March 2025; date of current version 9 June 2025. This work was supported in part by the research grant under Grant RP/FCA-09/2023 offered by Macao Polytechnic University. *(Corresponding author: Su-Kit Tang.)*

Yongheng Qian is with the Faculty of Applied Sciences, Macao Polytechnic University, Macau, SAR, China, and also with the Department of Mechatronics and Information Engineering, Zunyi Vocational and Technical College, Zunyi 563000, China (e-mail: yongheng.qian@mpu.edu.mo).

Su-Kit Tang is with the Faculty of Applied Sciences, Macao Polytechnic University, Macau, SAR, China (e-mail: sktang@mpu.edu.mo).

Digital Object Identifier 10.1109/JIOT.2025.3547920

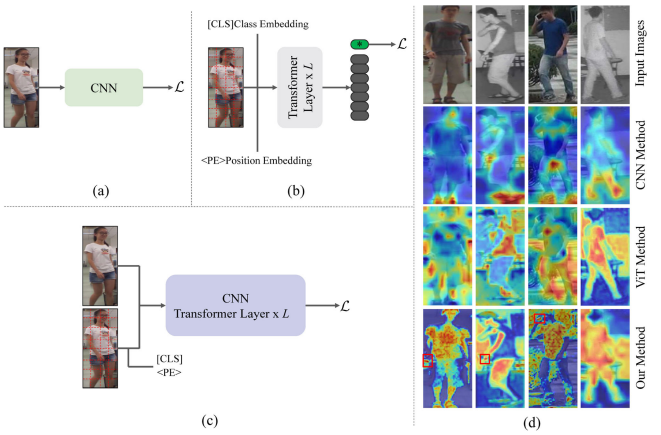

Fig. 1. Different architecture and matching methods are adopted in VI-ReID. (a) CNN method for extracting local features. (b) Pure ViT for modeling long-range dependencies. (c) Our method integrates local and global information. (d) Grad-CAM [7] visualization of the attention regions of different matching methods in VI-ReID. Deeper red indicates a higher weight.

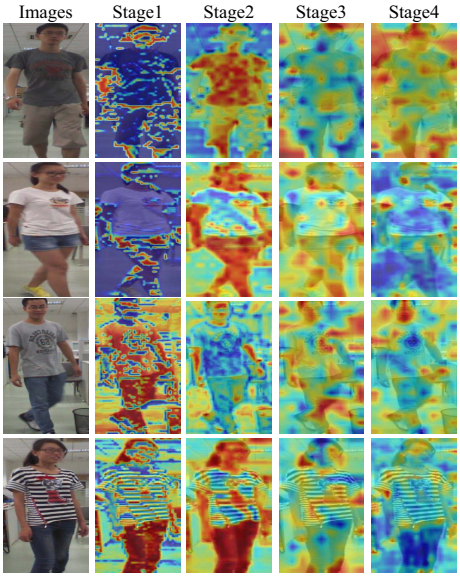

Fig. 2. Grad-CAM visualization of attention maps at different feature extraction stages in HAResformer. Deeper red indicates a higher weight.

as shown in the third row of Fig. 1(d). Pure ViT-based VI-ReID methods also fail to capture specific local features [e.g., the bracelet and phone in the second and third column input images of Fig. 1(d)]. We believe that these specific local features are salient identity-specific recognition cues. Notably, pure ViT-based VI-ReID methods adopt only class tokens as the final feature representation, ignoring the semantic information in the patch tokens.

Therefore, constructing the hybrid CNN and ViT network framework shown in Fig. 1(c) can simultaneously extract local features and global information, complementing each other and enriching the representation of features. To this end, Zhao et al. [17] used CNN and ViT alternately to extract features. The local features extracted by CNN were fed to ViT as prior knowledge for the interaction of local features and long-range dependencies. The features generated by ViT at each level were integrated as the final feature representation. However, this method cannot deeply fuse local features and global context information, potentially leading to poor performance. Intuitively, the shallow details are modality-invariant information, as shown in Fig. 2. Stages 1 and 2 networks focus on the contours and textures of person images, enriching the modality-invariant feature representation for the cross-modality ReID task. In addition, the network emphasizes different semantics in various feature extraction stages. For example, in the last row of Fig. 2, stage 3 focuses on the left arm, hair, and chest pattern, while stage 4 focuses on the right arm and shoulder. Therefore, better choices may exist than using only the last-layer features or simply fusing multiscale features. Based on the above analysis, we conclude that existing methods usually neglect several key issues: 1) Deep fusion of local features and global information enhances comprehensive discriminative feature representation. 2) Patch tokens contain rich semantic information. 3) Shallow details enrich high-level semantics. 4) The network emphasizes different semantics in different stages of feature extraction.

To address the above issues, we combine ResNet-50 [10] and ViT [11] to propose a novel hybrid ResNet-transformer hierarchical aggregation architecture for effective VI-ReID, named HAResformer, as shown in Fig. 3. HAResformer consists of three key components: 1) hierarchical feature extraction (HFE); 2) deeply supervised aggregation (DSA); and 3) hierarchical global aggregate encoder (HGAE). Specifically, HFE constructs the ResNet encoder (RE) and transformer encoder (TE) to extract local features and generate global spatial context information from person images. To deeply fuse local features and global information, and fully extract the rich semantics in patch tokens, we introduce a lightweight cross-encoder feature fusion module (CFFM). The fused features of the current stage are then fed as global priors to the TE of the next stage. Our method effectively mines specific local spatial information and models long-range dependencies. As shown in the last row of Fig. 1(d), our method captures both global cues and specific local features (e.g., red box-selected regions). In addition, we introduce the DSA strategy [18] to perform multigranularity auxiliary supervision on the class tokens generated by TE at each stage, enhancing multiscale feature representation. Different feature extraction stages focus on diverse attention regions of a person, indicating that relying only on the final high-level semantic features may be suboptimal. Therefore, we introduce an HGAE to aggregate multiscale features, enhancing comprehensive discriminative feature representation. Extensive experiments on three benchmark datasets demonstrate the effectiveness and robustness of our method.

The main contributions of this article can be summarized as follows.

1) We propose a novel hybrid ResNet-transformer hierarchical aggregation (HAResformer) architecture for VI-ReID. HAResformer effectively alleviates modality differences and reduces intramodality variations.
2) We propose a HFE framework comprising a RE and a TE to capture local spatial information and long-range dependencies of person images. Additionally, we introduce a lightweight CFFM for deep feature fusion

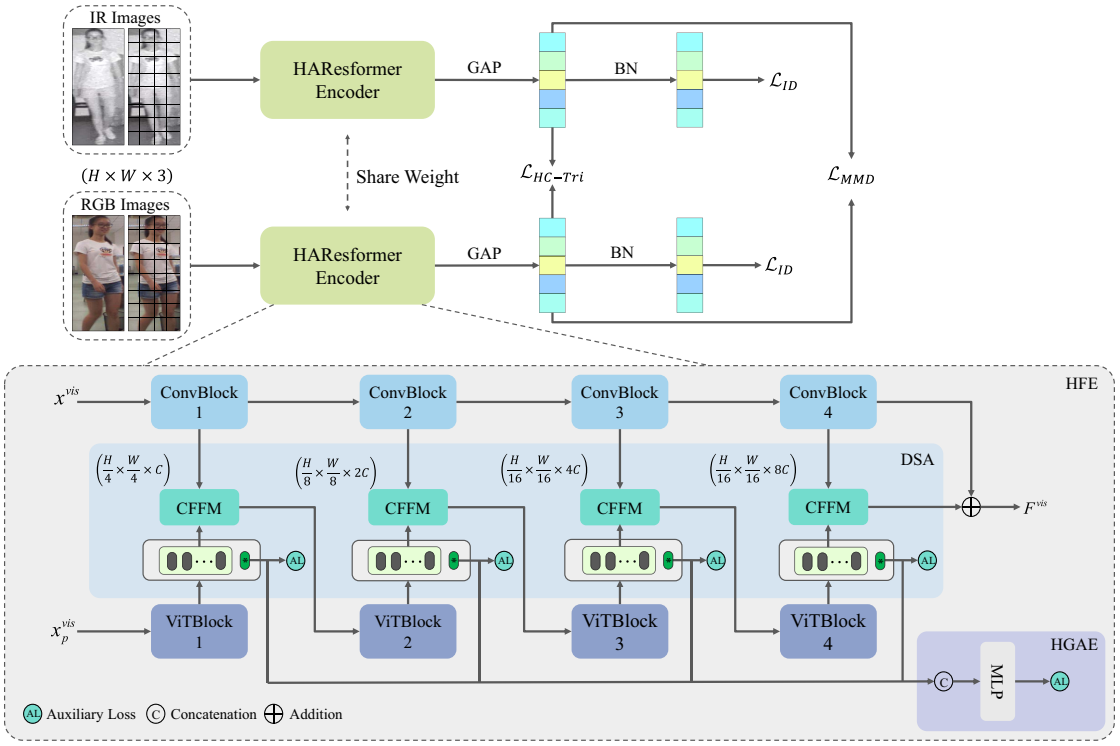

Fig. 3. Our proposed HAResformer architecture for VI-ReID contains three key components: HFE, deep supervision aggregation (DSA), and HGAE. HFE extracts local features and global representations of the input image and fusion. DSA provides auxiliary supervision for multiscale details and semantic information. HGAE aggregates multilevel class tokens to capture comprehensive global cues. It is best viewed in color.

to enhance comprehensive discriminative feature representation.

3) To harness the benefits of multigranularity features, we introduce the DSA strategy for auxiliary supervision of multiscale features. Additionally, we propose a HGAE to aggregate multigranularity features, significantly improving retrieval performance.
4) Extensive experiments on three publicly available datasets demonstrate the superiority of HAResformer over state-of-the-art methods.

The remainder of this article is organized as follows. In Section II, we analyze and summarize the related work of VI-ReID in recent years and the latest progress of ViT in VI-ReID. In Section III, we introduce the proposed HAResformer architecture, including the HFE, DSA, HGAE, and the objective function. In Section IV, we conduct extensive experiments on three publicly available benchmarks to demonstrate the effectiveness and superiority of our method. Finally, we conclude this work in Section V.

II. Related Work

A. VI-ReID

Modality differences and intramodality variations (e.g., viewpoint, illumination, pose, and occlusion) between visible and infrared images are the primary challenges in the VI-ReID task [19], [20]. To address these challenges, most existing studies focus on two approaches: 1) nongeneration methods [6], [12], [21], [22], [23], [24], [25], [26], [27], [28]; and 2) generation-based methods [29], [30], [31], [32], [33], [34]. Wu et al. [6] first defined the VI-ReID problem and proposed a deep zero-padding method to convert visible and infrared images into two-channel images, mitigating modality differences. Ye et al. [27], [28] developed a two-stream network to project modality-specific features into a shared feature space, learning a robust and discriminative shared representation to bridge the modality gap. Their work laid the groundwork for subsequent research on VI-ReID. Liu et al. [12] investigated the optimal number of shared parameters in the modality-shared section of the two-stream network and proposed a hetero-center triplet loss to relax the strict constraints of traditional triplet loss, significantly improving retrieval performance. These studies decouple person image features into modality-specific and shared features, mitigating modality differences through feature-level alignment in a shared embedding space. However, these methods overlook the contribution of modality-specific features to improving final retrieval performance. Dai et al. [31] was the first to propose a GAN-based approach to address the issue of insufficient discriminative features. Wang et al. [29] and Qian and Tang [30] bridged the modality gap by generating cross-modality paired images to achieve instance-level feature alignment. Zhong et al. [22] converted infrared images to visible images using colorization to address the VI-ReID challenge by unifying the modalities. However, GAN-based methods are computationally intensive and may introduce additional noise [12], [35]. Currently, these methods focus on constructing CNN-based frameworks to extract local spatial information from images and use the last layer's features as discriminative representations. Unlike these

methods, our HAResformer architecture simultaneously captures local features and global spatial context information, utilizing a multigranularity auxiliary supervision strategy to achieve multiscale feature learning. This approach significantly enhances modality-invariant feature representation and network robustness.

B. ViT in VI-ReID

Motivated by the success of transformers [36] in natural language processing (NLP) tasks. Dosovitskiy et al. [11] extended the transformer to computer vision tasks and proposed a vision transformer (ViT) framework that significantly enhanced the performance of various vision tasks, including medical image classification [37], object detection [38], and image segmentation [39]. Given ViT's ability to model long-range dependencies, He et al. [2] introduced a pure ViT with a jigsaw patch module for the VV-ReID task, achieving performance comparable to that of CNN-based frameworks. He et al. [2] introduced the concept of combining CNN and ViT and designed a transformer-based feature calibration module to progressively aggregate multiscale features. This approach fully exploits both local and global feature representations, generating more discriminative global features to enhance retrieval performance. However, the significant modality gap prevents these methods from being directly applied to the VI-ReID task.

Currently, many works [15], [16], [17], [35], [40], [41], [42], and [43] have explored the application of ViT in VI-ReID and achieved promising retrieval performance. Liang et al. [41] introduced a learnable modality embedding to the ViT input to bridge the gap between heterogeneous images. They also designed a modality-aware enhancement loss to explicitly capture information from each modality, representing the first pure ViT-based approach for VI-ReID. Lu et al. [35] developed a progressive modality-shared transformer framework to improve the reliability and commonality of visual features across different modalities, resulting in enhanced performance and robustness. Noting that CNN-extracted features may include identity-irrelevant information, Yang et al. [15] applied a Top-k vision tokens selection module to precisely choose Top-k discriminative vision patches, reducing the interference of irrelevant information and improving feature discrimination. While pure ViT-based frameworks effectively capture global feature information, they overlook the benefits of local spatial features critical for salient identity-specific recognition. To address this issue, Chen et al. [40] combined CNN and ViT to propose a structure-aware positional transformer network. This network integrates local interaction learning with global appearance learning to enhance semantically shareable modality representation and effectively mitigate modality differences. Jiang et al. [42] and Feng et al. [43] developed a similar network framework, where ResNet builds the local feature extraction backbone and ViT models long-range dependencies in the final layer of these features. While these methods improve global spatial feature representation, they require deeper integration of local and global information. Our method further improves discriminative feature representation by integrating local spatial information and global context through a comprehensive hybrid CNN and ViT approach. Additionally, our method fully utilizes the fine-grained cues from patch tokens to extract more semantic information. Experimental results validate the effectiveness and robustness of our architecture.

III. Proposed Method

In this section, we introduce the proposed HAResformer architecture for VI-ReID, as shown in Fig. 3. First, we are given a problem formulation and an overview of HAResformer. Second, the proposed HFE and CFFM are introduced. Third, we outline the DSA strategy. Fourth, the proposed HGAE is introduced. Finally, we introduce the objective function with auxiliary loss to optimize our architecture.

A. Problem Formulation and Overview

We take $\mathcal{D} = \{\mathcal{V}, \mathcal{R}\}$ to denote the set of cross-modality person images. $\mathcal{V} = \{x_i^{\text{vis}}, y_i^{\text{vis}} | i = 1, 2, \ldots, N^{\text{vis}}\}$ and $\mathcal{R} = \{x_i^{ir}, y_i^{ir} | i = 1, 2, \ldots, N^{ir}\}$ indicate N^{vis} visible images and N^{ir} infrared images with corresponding ground-truth labels y_i^{vis} and y_i^{ir}, respectively. VI-ReID aims to retrieve a given infrared (visible) query image with the same identity in the visible (infrared) gallery set. Existing works mainly focus on learning task-related feature representations from the final layer in CNN-based methods or the last layer class tokens in pure ViT-based methods to overcome cross-modality differences and intramodality variations. However, CNN-based methods cannot model long-range dependencies, while pure ViT-based methods have difficulty mining specific local features. Our method aims to fuse local and global features deeply to enhance discriminative feature representation.

An overview of our architecture is shown in Fig. 3, which employs a HAResformer two-stream network with weight sharing to extract cross-modality person features. HAResformer contains three key components: 1) HFE; 2) DSA; and 3) HGAE. The visible image is similar to the feature extraction process of the infrared image, and we describe it in terms of the visible image. First, CFFM receives the output features f_1^{vis} from the RE of stage 1 and the output patch tokens f_{1-p}^{vis} from the TE, and the size is $(H/4) \times (W/4) \times C$ after deep fusion, where H, W, and C indicate the height, width, and channel dimensions of the input image, respectively. The fused feature map $\hat{f}_1^{\text{vis}}$ is then fed to the next stage TE. Second, a deep supervision strategy is exploited to provide auxiliary supervision for the global representation class tokens $f_{1-\text{cls}}^{\text{vis}}$ output by TE to capture multiscale detail information. Third, the class tokens of each stage are fed to HGAE by skipping connections to capture the multigranularity feature representation. Finally, the fusion features were combined with local features to obtain F^{vis} end-to-end training. Similarly, we can get the final feature F^{ir} of the infrared image. In addition, considering the limitation of GPU memory, we set the number of heads H_i in each stage and the number of transformer layers L_i to 1, and the comprehensive performance is good. The detailed settings are listed in Table I.

TABLE I
DETAILED SETTINGS OF OUR HARESFORMER INCLUDE PATCH SIZE (P_i), STRIDE SIZE (S_i), HEAD NUMBER (H_i), AND NUMBER OF TRANSFORMERS (L_i) IN STAGES

Stage	Output Size	ResNet Encoder	Transformer Encoder
1	$\frac{H}{4} \times \frac{W}{4} \times 256$	3 × Bottleneck	$P_1 = 4; S_1 = 4; H_1 = 1; L_1 = 1$
2	$\frac{H}{8} \times \frac{W}{8} \times 512$	4 × Bottleneck	$P_2 = 2; S_2 = 2; H_2 = 1; L_2 = 1$
3	$\frac{H}{16} \times \frac{W}{16} \times 1024$	6 × Bottleneck	$P_3 = 2; S_3 = 2; H_3 = 1; L_3 = 1$
4	$\frac{H}{16} \times \frac{W}{16} \times 2048$	3 × Bottleneck	$P_4 = 1; S_4 = 1; H_4 = 1; L_4 = 1$

B. Hierarchical Feature Extraction

HFE contains four stages, denoted as $\{S1, S2, S3, S4\}$. Each stage consists of three components: ResNet-50 [10] pretrained by ImageNet [44] builds a RE block at each stage, which is applied to extract local features of the image. TE block built by ViT-based [11] to capture global representations. A lightweight CFFM is introduced to fuse local and global features deeply.

RE: While a transformer can model long-range dependencies, it lacks local inductive bias. The fixed local receptive field characteristics of CNN enable it to generate specific local spatial information. Following some previous works [19], [20], we use ResNet-50 [10] to build an RE block to extract the local features f_i^{vis}, and $i(i = 1, 2, 3, 4)$ denote the feature extraction stage. To simplify the description, the feature representation function of the RE block is denoted as $\mathcal{F}(\cdot)$. Given an input image $x^{\text{vis}} \in \mathbb{R}^{C\times H\times W}$, its corresponding feature representation is expressed as

$$\begin{aligned} f_i^{\text{vis}} &= \mathcal{F}_i\left(x^{\text{vis}}\right), i = 1 \\ f_i^{\text{vis}} &= \mathcal{F}_i\left(f_{i-1}^{\text{vis}}\right), i = 2, 3, 4 \end{aligned} \tag{1}$$

where settings and output size of $\mathcal{F}_i(\cdot)$ are shown in Table I.

TE: Attributed to the success of ViT in computer vision tasks. We introduce TE to model the long-range dependence of the input image, as shown in Fig. 4. Given an input image $x^{\text{vis}} \in \mathbb{R}^{C\times H\times W}$, we apply a convolutional layer to obtain nonoverlapped image patches of size $P \times P$. The stride is set to $S(S = P)$, a total of $N = \lfloor([H + S - P]/S)\rfloor \times \lfloor([W + S - P]/S)\rfloor$ patches are obtained, and each patch is flattened into an embedding of size $\mathbb{R}^{P^2\cdot C}$. Then, each patch embedding x_p^{vis} is linearly projected into a D-dimensional vector. Like BERT [45], we add learnable class embedding $x_{\text{cls}}^{\text{vis}}$ to capture the global feature representation. Since VI-ReID requires position information between patches, a standard learnable 1-D position embedding $x_{\text{pos}}^{\text{vis}}$ is added to the projected patch embedding. The patch embedding fed into transformer layers can be expressed as follows:

$$z_0^{\text{vis}} = \left[x_{\text{cls}}^{\text{vis}}, x_p^{\text{vis}}\right] + x_{\text{pos}}^{\text{vis}} \tag{2}$$

where $x_{\text{cls}}^{\text{vis}} \in \mathbb{R}^{1\times D}$, $x_p^{\text{vis}} \in \mathbb{R}^{N\times D}$, and $x_{\text{pos}}^{\text{vis}} \in \mathbb{R}^{(N+1)\times D}$. After obtaining the patch embedding z_0^{vis}, it is fed into the transformer layer. Each transformer layer contains a multihead self-attention (MHSA) block, a multilayer perception (MLP) block, and layer normalization (LN), and residual connections

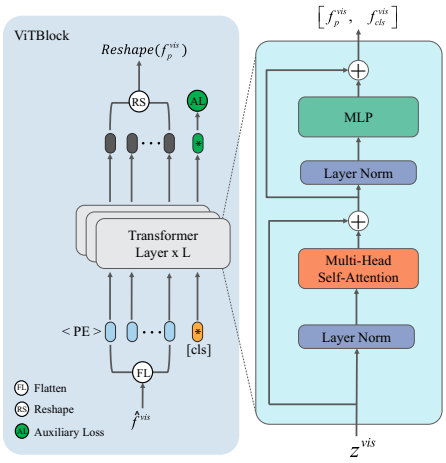

Fig. 4. Our proposed TE block with auxiliary loss.

are applied before and after each block, which models long-range dependencies and drives TE to focus on diverse human body regions. We can obtain the output of each transformer layer as follows:

$$\begin{aligned} z_{L,\text{MHSA}}^{\text{vis}} &= z_{L-1}^{\text{vis}} + \text{MHSA}\left(LN(z_{L-1}^{\text{vis}})\right) \\ z_L^{\text{vis}} &= z_{L,\text{MHSA}}^{\text{vis}} + \text{MLP}\left(LN(z_{L,\text{MHSA}}^{\text{vis}})\right) \end{aligned} \tag{3}$$

where L indicates the number of transformer layers. After the transformer layer calculates, the output class tokens $f_{\text{cls}}^{\text{vis}}$ are used as global feature representations. Meanwhile, to mine the rich semantic information contained in patch tokens f_p^{vis}, we fuse them with the local features extracted by RE as the input of TE in the next stage. We denote by f_{i-cls}^{vis} and f_{i-p}^{vis} the class tokens and patch tokens generated by TE at the stage $i(i = 1, 2, 3, 4)$, respectively.

CFFM: Patch tokens also exhibit strong global representations thanks to the excellent ability of transformers to model long-range dependencies. To bridge the shortcomings of RE and TE and mine the rich semantic information of patch tokens, we propose a lightweight CFFM, as shown in Fig. 5. CFFM fuses the local features f^{vis} extracted by RE and the global feature sequence f_p^{vis} generated by TE. We reshape the global feature sequence f_p^{vis} to have the same shape as f^{vis}, then deeply fuse the concatenation features through a point-wise convolutional layer. The fusion process can be expressed as follows:

$$\hat{f}_i^{\text{vis}} = PW\left(f_i^{\text{vis}} \| \text{Reshape}\left(f_{i-p}^{\text{vis}}\right)\right), i = 1, 2, 3, 4 \tag{4}$$

where $\hat{f}_i^{\text{vis}}$ indicates the fused features of the ith stage, PW indicates point-wise convolution, and $\|$ indicates channel-wise concatenation. The fused features of the current stage are fed to the TE of the next stage, which complements the local spatial information of the TE.

C. Deeply Supervised Aggregation

As shown in Fig. 2, we consider shallow features (e.g., texture or contour) as modality-invariant information, which greatly aids the cross-modality ReID task. Inspired

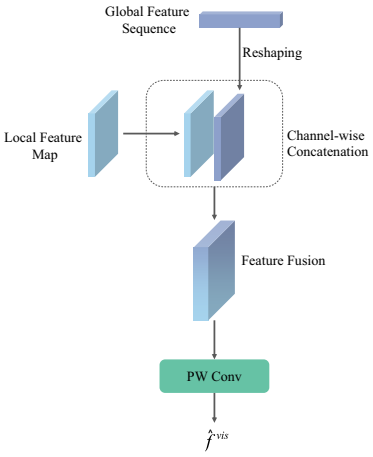

Fig. 5. Illustration of the proposed CFFM.

by the deep supervised learning strategy [18], previous work [46], [47], and [21] in the VI-ReID task demonstrated the effectiveness of low-level detail information in alleviating the modality difference between visible and infrared images. However, these works achieved poor performance primarily because the multiscale features in CNN-based methods lack sufficient global context information. Additionally, we observed that different feature extraction stages emphasize different semantic information. Considering the comprehensive global spatial representation of class tokens, we perform multiscale supervision on the class tokens generated by the TE at each stage to enhance multigranularity feature representation.

D. Hierarchical Global Aggregation Encoder

We observe from Fig. 2 that different feature extraction stages focus on different attention regions of a person, which indicates that using only high-level semantic features of the network's last layer may be a suboptimal problem. To address this issue, we propose a HGAE to aggregate multiscale features to enhance discriminative feature representation. Specifically, we integrate the class tokens generated by TE at different stages to construct a global aggregate embedding $\hat{F}_{GA}^{\text{vis}}$, which can be expressed as follows:

$$\hat{F}_{GA}^{\text{vis}} = \text{Concat}\left(f_{1-\text{cls}}^{\text{vis}}, f_{2-cls}^{\text{vis}}, f_{3-cls}^{\text{vis}}, f_{4-cls}^{\text{vis}}\right) \tag{5}$$

where Concat indicates the channel-wise concatenation, previous work [5] confirmed that simply concatenating multiscale features will lead to poor performance. The main reason is that integrating multilevel features may increase task-irrelevant error information. Therefore, to reduce task-irrelevant cues and enhance useful global context information, similar to the MLP feature extraction process in ViT, we feed $\hat{F}_{GA}^{\text{vis}}$ to the fully connected layer and GeLU activation function [48] to extract a comprehensive and effective global aggregation representation, which can be expressed as follows:

$$F_{GA}^{\text{vis}} = \text{GeLU}\left(FC\left(\hat{F}_{GA}^{\text{vis}}\right)\right) \tag{6}$$

a similar process can obtain the global aggregation representation F_{GA}^{ir} of the infrared image.

E. Objective Function

In this section, we introduce the objective function for optimizing HAResformer, including ID loss $\mathcal{L}_{ID}$, HC-Tri loss $\mathcal{L}_{HC-Tri}$ [12], margin-based Maximum mean discrepancy (MMD) loss $\mathcal{L}_{\text{MMD}}$ [13], and the designed auxiliary loss $\mathcal{L}_{AL}$ to optimize the model in an end-to-end learning framework jointly.

ID Loss: Following previous work [20], [49], we employ cross-entropy-based ID loss to treat images of different modalities of the same identity as the same class, aiming to extract feature representations of stable identities for classification. ID loss can be defined as

$$\mathcal{L}_{ID} = -\frac{1}{n}\sum_{i=1}^{n}\log\left(p\left(y_i^{\text{vis}}|x_i^{\text{vis}}\right)\right) - \frac{1}{n}\sum_{i=1}^{n}\log\left(p\left(y_i^{ir}|x_i^{ir}\right)\right) \tag{7}$$

where n indicates the number of training images in a mini-batch and $p(\cdot)$ indicates the predicted probability after the softmax function.

Hetero-Center Triplet Loss: In the ReID task, triplet loss [50] pulls the same class and pushes away different classes, effectively bridging the gap between modalities. However, triplet loss is a strict constraint metric that may harm other well-learned pairwise distances when there are some terrible samples. Hetero-center triplet loss [12] is a variant of triplet loss, which calculates the distance between the anchor center and all other sample centers, thereby relaxing the constraint and reducing the computational complexity. In a mini-batch, we randomly sample P identities, and each identity randomly selects K visible and K infrared images for a total of $2 \times PK$ images. The modality-specific feature centers for each identity can be expressed as follows:

$$c_{\text{vis}}^{i} = \frac{1}{K}\sum_{j=1}^{K} v_j^i,\ c_{ir}^{i} = \frac{1}{K}\sum_{j=1}^{K} t_j^i \tag{8}$$

where v_j^i indicates the jth visible image feature of the ith identity, and t_j^i is the infrared image feature. Then, the HC-Tri loss can be expressed as

$$\mathcal{L}_{HC-\text{Tri}} = \sum_{i=1}^{P}\left[\rho + \left\|c_{\text{vis}}^{i} - c_{ir}^{i}\right\|_2 - \min_{\substack{n\in\{\text{vis},ir\}\\ j\neq i}}\left\|c_{\text{vis}}^{i} - c_{n}^{j}\right\|_2\right]_+ + \sum_{i=1}^{P}\left[\rho + \left\|c_{ir}^{i} - c_{\text{vis}}^{i}\right\|_2 - \min_{\substack{n\in\{\text{vis},ir\}\\ j\neq i}}\left\|c_{ir}^{i} - c_{n}^{j}\right\|_2\right]_+ \tag{9}$$

where ρ is a margin, following [12], $\rho = 0.3$. $[\cdot]_+$ indicates the hinge function $\max(0, x)$.

Margin-Based MMD Loss: MMD [51] measures the closeness between two feature distributions. However, in the VI-ReID task, directly applying the MMD distance can easily lead to overfitting and feature degradation. Margin-based MMD is a variant of MMD that has the following advantages: 1) Only uses global features and does not rely on part-level features. 2) Align two modalities and class distributions at the same time. 3) Alleviates the overfitting problem. Margin-based MMD loss can be expressed as

$$\text{MMD}^2(V_c, R_c) = E_V\left[k\left(x_c^{\text{vis}}, x_{c'}^{\text{vis}}\right)\right] + E_R\left[k\left(x_c^{ir}, x_{c'}^{ir}\right)\right]$$

$$- 2E_{VR}\left[k\left(x_c^{\text{vis}}, x_c^{ir}\right)\right] \tag{10}$$

$$\text{MMD}'^2(V_c, R_c) = \begin{cases} \text{MMD}^2(V_c, R_c), & \text{if dist} - \rho > 0 \\ 0, & \text{otherwise} \end{cases} \tag{11}$$

$$\mathcal{L}_{\text{MMD}} = \frac{1}{C}\sum_{c=1}^{C} \text{MMD}'^2(V_c, R_c) \tag{12}$$

where V_c and R_c indicate the c-th identify-specific visible and infrared sample distributions. $k(\cdot)$ is a Gaussian kernel. dist $=$ $\text{MMD}^2(V_c, R_c)$, ρ is a margin, following [13], $\rho = 1.4$.

Auxiliary Loss: HAResformer focuses on different semantic elements at various feature extraction stages, with shallow details serving as an excellent complement to high-level semantic information. To achieve this, we provide auxiliary supervision on the global feature representation class tokens generated by the TE at each stage and the global aggregate embedding fused by the HGAE, enhancing the network's discriminative feature representation. In this work, the auxiliary loss $\mathcal{L}_{AL}$ consists of hetero-center triplet loss for jointly supervised training of the HAResformer architecture.

Therefore, combining (7), (9), and (12), the overall loss function of our HAResformer architecture can be expressed as

$$\mathcal{L} = \mathcal{L}_{ID} + \lambda_1\mathcal{L}_{HC-tri} + \lambda_2\mathcal{L}_{\text{MMD}} + \lambda_3\sum_{i=1}^{N_{AL}}\mathcal{L}_{AL} \tag{13}$$

where λ_1, λ_2, and λ_3 are tradeoff parameters. We follow [12] and [13], setting $\lambda_1 = 2.0$, $\lambda_2 = 0.25$. λ_3 is discussed in detail in section C of the experiment. N_{AL} is the number of stage blocks and HGAE blocks.

IV. Experiments

A. Datasets and Settings

Datasets: The effectiveness of our method is evaluated on three publicly available real-world benchmarks: 1) SYSU-MM01 [6]; 2) RegDB [52]; and 3) LLCM [53], respectively. The SYSU-MM01 dataset was captured by four visible and two near-infrared cameras in indoor and outdoor environments. The training set contains 22 258 visible images and 11 909 near-infrared images of 395 identities. The test set contains 3 803 near-infrared images of 96 identities for the query. The gallery has All-search or Indoor-search and single-shot according to the evaluation mode versions, details of each mode can be found in [19] and [20]. In this work, we adopt the All-search and Indoor-search evaluation modes in the most challenging and popularly utilized single-shot setting. The RegDB dataset contains 8 240 images of 412 identities captured by a visible and a far-infrared camera, 206 identities for training, and the rest for testing. Each identity contains ten visible and ten far-infrared images. We alternately utilize visible (far-infrared) images for query (gallery) to evaluate Visible-to-Infrared and Infrared-to-Visible modes. LLCM is a cross-modality dataset of 46 767 images from 1 064 identities, captured using nine cameras under challenging low-light conditions. It includes person images captured in various real-world scenarios, covering different climate conditions, low-resolution, and clothing styles. The training set includes 16 946 visible images and 13,975 near-infrared images from 713 identities, with the remaining 351 identities designated for testing. We alternately utilize visible (near-infrared) images for query (gallery) to evaluate Visible-to-Infrared and Infrared-to-Visible search modes.

Evaluation Protocols: Following standard protocols [6], query and gallery images come from different modalities. Standard cumulative matching characteristics (CMC) curves and mean average precision (mAP) are used for performance evaluation. We perform ten gallery set selection experiments and report the average performance.

Implementation Details: Our proposed method is implemented using version 1.12.1 + cu11.3 of the PyTorch framework, and we use an NVIDIA GPU for training. The ResNet-50 [10] model pretrained on ImageNet [44] is adopted as the backbone of RE, and the stride of the last convolutional block is set to 1 to obtain fine-grained feature maps [20]. We adopt multilayer ViT [11] without pretraining parameters as the backbone of TE, and the detailed settings of each layer are shown in Table I. All images were resized to 288×144, and random cropping, horizontal flipping, and random erasing (erasing probability $\rho = 0.5$) data augmentation were performed on the training set. In addition, we randomly convert visible images to grayscale images with a probability of 0.5 [54], which helps to alleviate the cross-modality differences at the image-level. We use the stochastic gradient descent (SGD) algorithm to optimize our architecture with a momentum parameter of 0.9 and a decay rate of 5e-4. During the training phase, the learning rate is initialized to 0.01 and decayed by 0.1 after the 30th epoch, and the entire network is trained for 80 epochs. For sampling, we randomly sample four identities, each with four visible and four infrared images, for 32 images per training batch. In addition, the tradeoff parameters λ_1, λ_2, and λ_3 of the SYSU-MM01 dataset are set to 2.0, 0.25, and 0.5, respectively. The tradeoff parameters λ_1, λ_2, and λ_3 of the RegDB dataset are set to 2.0, 0.25, and 0.25, respectively. As the LLCM dataset is collected similarly to SYSU-MM01, we adopt the same hyperparameter settings used for the SYSU-MM01 dataset during LLCM training.

B. Ablation Experiments

Evaluation of Different Loss: In this part, we conduct experiments to evaluate the effectiveness of each loss term in HAResformer on the SYSU-MM01 and RegDB datasets. All hyperparameters are fixed during evaluation for consistency. From Table II, we draw the following conclusions.

1) With ID loss alone, performance on SYSU-MM01 and RegDB datasets is poor. This is because, in the open-set setting, the identities of persons in the test and training sets do not overlap. Therefore, ID loss is usually regarded as a regularization term in VI-ReID. By integrating other metric losses in the embedding space, performance is greatly improved. As shown in the second row of Table II, HC-Tri loss, compared with the method with only ID loss, achieves a gain of 6.12% Rank-1 and 4.18% mAP on the SYSU-MM01 dataset. A significant improvement of 23.80% Rank-1

TABLE II
EVALUATION OF EACH LOSS COMPONENT ON SYSU-MM01 AND REGDB DATASETS. ID DENOTES ID LOSS BASED ON CROSS-ENTROPY, HC-TRI IS HETERO-CENTER TRIPLET LOSS, MMD DENOTES MARGIN-BASED MMD LOSS, AND AL INDICATES MULTISCALE AUXILIARY LOSS. REID RATES AT RANK-1 (%) AND MAP (%)

Method	ID	HC-Tri	MMD	AL	SYSU-MM01		RegDB	
					Rank-l	mAP	Rank-l	mAP
1	✓				47.57	45.67	56.38	56.08
2	✓	✓			53.69	49.85	80.18	76.44
3	✓	✓	✓		60.45	56.23	91.00	86.62
4	✓	✓	✓	✓	**61.00**	**57.92**	**92.15**	**87.87**

and 20.36% mAP is achieved on the RegDB dataset, indicating that HC-Tri loss relaxes constraints, effectively pulling closer to the same class and pushing away different classes, alleviating intramodality variations. The performance of integrated margin-based MMD loss is further improved based on method 2, indicating that aligning intermodality and intramodality feature distributions can significantly alleviate modality differences.

2) By integrating auxiliary loss based on method 3, Rank-1 for SYSU-MM01 and RegDB datasets is improved by 0.55% and 1.15%, respectively, and mAP is improved by 1.69% and 1.25%, respectively. This gain indicates that low-level detail cues effectively complement high-level semantic information with multiscale supervision, enhancing discriminative feature representations.

Impact of Supervision Scope: In this part, we compare the effects of auxiliary supervision on the SYSU-MM01 and RegDB datasets, where {*Si*} denotes the *i*th feature extraction stage. As shown in Table III, Rank-1 and mAP for the SYSU-MM01 and RegDB datasets tend to increase when auxiliary supervision is applied at each feature extraction stage. Compared to applying auxiliary supervision independently at each feature extraction stage, the Rank-1 and mAP for {*S*1, *S*4} on the SYSU-MM01 and RegDB datasets drop significantly. When {*S*1, *S*4} is combined with other feature extraction stages, the Rank-1 and mAP on both datasets show a slight upward trend. These results demonstrate that low-level details can effectively complement high-level semantic information. However, the noisy information in low-level features can seriously affect final retrieval performance. Therefore, we discard the auxiliary loss imposed at {*S*1}.

Observing the last three rows of Table III, {*S*2, *S*4} achieve comparable Rank-1 and mAP performance on the SYSU-MM01 dataset, but the results are lower than {*S*4}. The best Rank-1 and mAP results on the RegDB dataset are achieved by {*S*3, *S*4}. Based on the above analysis, we conclude that the network focuses on different regions of person images at various feature extraction stages, inevitably, including identity-irrelevant information. Therefore, blindly performing joint supervision on multiscale features may significantly hinder discriminative feature representation. In this work, we apply auxiliary supervision at {*S*4} for the SYSU-MM01 dataset and at {*S*3, *S*4} for the RegDB dataset.

Evaluation of Different Components: In this part, we analyze the effectiveness of various components in HAResformer on

TABLE III
EVALUATION OF AUXILIARY SUPERVISION AT DIFFERENT FEATURE EXTRACTION STAGES ON SYSU-MM01 AND REGDB DATASETS. REID RATES AT RANK-1 (%) AND MAP (%)

Method	$S1$	$S2$	$S3$	$S4$	SYSU-MM01		RegDB	
					Rank-l	mAP	Rank-l	mAP
1	✓				60.58	56.50	88.88	85.58
2		✓			62.24	58.85	90.87	87.42
3			✓		62.27	58.83	91.46	87.24
4				✓	**64.00**	**59.96**	91.16	87.60
5	✓			✓	58.59	56.47	90.87	85.83
6	✓	✓		✓	60.03	56.96	88.76	83.45
7	✓		✓	✓	61.79	58.04	91.86	87.66
8	✓	✓	✓	✓	61.00	57.92	92.15	87.87
9		✓		✓	62.35	59.13	91.70	87.84
10			✓	✓	61.98	58.39	**92.75**	**88.97**
11		✓	✓	✓	59.19	56.00	91.96	88.46

TABLE IV
ANALYSIS OF THE EFFECTIVENESS OF DIFFERENT COMPONENTS ON SYSU-MM01 AND REGDB DATASETS. REID RATES AT RANK-1 (%) AND MAP (%)

Method	B	CFFM	DSA	ADD	HGAE	SYSU-MM01		RegDB	
						Rank-l	mAP	Rank-l	mAP
1	✓					52.54	48.34	77.59	72.74
2	✓	✓				60.45	56.23	91.00	86.62
3	✓	✓	✓			64.00	59.96	92.75	**88.97**
4	✓	✓	✓	✓		62.85	57.87	93.12	86.90
5	✓	✓	✓		✓	60.01	55.92	89.83	85.31
6	✓	✓	✓	✓	✓	**65.92**	**60.37**	**93.44**	87.02

SYSU-MM01 and RegDB datasets. As shown in Table IV, B indicates the HAResformer foundational framework, while ADD denotes the residual connection of local features in the final stage. HGAE is a HGAE. From Table IV, we can draw the following conclusions.

1) Based on B, CFFM integrates local features with global information, significantly enhancing retrieval accuracy. DSA enhances comprehensive discriminative representations by capturing multigranularity features, resulting in Rank-1 improvements of 3.55% and 1.75% and mAP gains of 3.73% and 2.35% on the SYSU-MM01 and RegDB datasets, respectively.
2) Integrating Method 3 with ADD significantly decreased Rank-1 and mAP scores on the SYSU-MM01 dataset but slightly increased Rank-1 on the RegDB dataset. This experimental result indicates that element-wise addition may have caused a loss of discriminative features from the CFFM output. Furthermore, integrating Method 3 with HGAE reduced Rank-1 and mAP scores on both datasets compared to Method 3 and Method 4, potentially due to noisy information within the multigranularity features.
3) Encouraging performance was achieved by combining ADD and HGAE with Method 3. Compared to Method 3, the Rank-1 and mAP of SYSU-MM01 increased by 1.92% and 0.41%, respectively, while RegDB achieved a Rank-1 gain of 0.69%. In conclusion, these comparison results demonstrate the effectiveness of the HAResformer architecture.

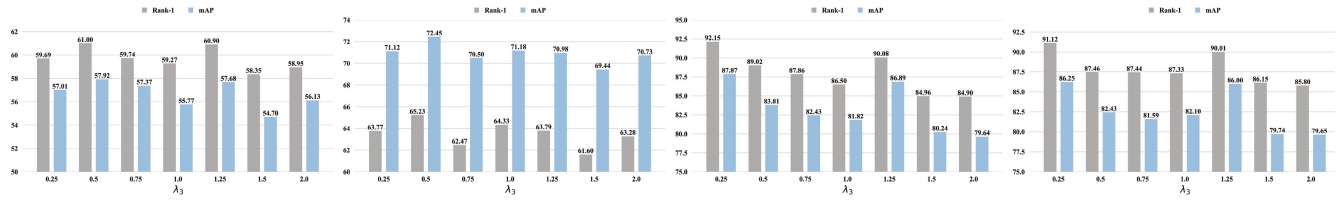

Fig. 6. Evaluation of the tradeoff parameter λ_3 on the SYSU-MM01 and RegDB datasets. ReID rates at Rank-1 (%) and mAP (%). (a) All-search. (b) Indoor-search. (c) Visible-to-Infrared. (c) Infrared-to-Visible.

Scaling Study: In this part, we evaluate the impact of the TE setting on the computational cost and retrieval performance of HAResformer on the SYSU-MM01 and RegDB datasets, as shown in Table V. Table I lists the setting details of the TE, where each stage has a uniform number of heads and depth. We fix the head embedding and MLP dimensions, and vary the number of heads H and transformer depth L. Set 1-head and 1-depth achieved the best Rank-1 scores on both datasets. However, increasing the transformer depth led to a drop in model quality. Impressively, increasing the number of heads on the SYSU-MM01 dataset achieved the best mAP performance. These results indicate that a multihead transformer can effectively capture rich semantic information. Given the GPU memory constraints, this work sets 1-head and 1-depth to achieve good comprehensive performance.

Additionally, to evaluate the practicality of HAResformer in real-world IoT systems, we analyzed the impact of varying the number of heads and depth on the computational cost of HAResformer. The last two columns of Table V display the time and space complexity associated with the HAResformer test experiments. In this work, time complexity and space complexity are measured in terms of floating-point operations (FLOPs) and the number of parameters, respectively. Several patterns are evident: First, the internal parallel computation of the multihead attention mechanism [36] ensures that increasing the number of heads does not increase the HAResformer's complexity. However, the 1-head and 1-depth achieve a superior performance-to-computation tradeoff compared to the 2-head and 1-depth setting. Second, increasing the depth results in approximately 1.6 times higher time and space complexity. Surprisingly, increasing the depth of the transformer did not yield optimal performance. Although larger models are often assumed to capture more robust visual representations and discriminative features. However, the multilayer encoding structure of large models introduces significant matrix computations, potentially challenging their deployment in real-world IoT systems. Notably, both FLOPs and the number of parameters remain within the acceptable range for HAResformer. These analyses establish the feasibility of deploying HAResformer in industrial IoT smart cameras.

TABLE V
EVALUATE THE IMPACT OF THE TE SETTING ON THE RETRIEVAL PERFORMANCE AND COMPUTATIONAL COST OF THE HARESFORMER. REID RATES AT RANK-1 (%) AND MAP (%)

Setting		SYSU-MM01		RegDB		HAResformer Testing	
H	L	Rank-1	mAP	Rank-1	mAP	FLOPs, G	Parameters, M
1	1	**65.92**	60.37	**93.44**	**87.02**	25.00	113.75
1	2	64.05	59.60	92.84	85.90	39.34	180.64
2	1	64.69	**60.44**	91.87	83.12	25.00	113.75

C. Visualization Analysis

Influence of Hyperparameters: We evaluate the impact of the tradeoff parameter λ_3 in (13) on the retrieval performance of the HAResformer architecture. Fig. 6 shows the effects of tradeoff parameter values on Rank-1 and mAP in two retrieval modes on the SYSU-MM01 and RegDB datasets. We progressively increase the value of λ_3 from 0.25 to 2.0 for independent experiments. It can be intuitively seen from Fig. 6 that consistent improvements can be obtained when we set different λ_3. Specifically, as shown in Fig. 6(a) and (b), the Rank-1 and mAP of the two search modes on the SYSU-MM01 dataset both show a trend of first increasing and then decreasing. Set $\lambda_3 = 0.5$, HAResformer obtains the best Rank-1 and mAP in the two search modes on the SYSU-MM01 dataset. Fig. 6(c) and (d) show that the Rank-1 and mAP of the two search modes on the RegDB dataset achieve the best performance in $\lambda_3 = 0.25$. Therefore, during the training of the HAResformer architecture, we set the tradeoff parameter λ_3 to 0.5 and 0.25 for the SYSU-MM01 and RegDB datasets, respectively.

Visualization of Feature Distribution: To investigate why HAResformer is effective, we visualize the intraclass and interclass similarity distributions on SYSU-MM01 and RegDB datasets, and the similarity is obtained by cosine distance calculation, as shown in the first row of Fig. 7. AL indicates multiscale auxiliary supervision, and B indicates HAResformer w/o AL. We observe that the means of intraclass and interclass distances (i.e., vertical dashed lines) are pushed apart by B + AL and HAResformer, where $\delta_{1-1} < \delta_{1-3} < \delta_{1-2}$ and $\delta_{2-1} < \delta_{2-3} < \delta_{2-2}$, as shown in Fig. 7(a) and (b). The results demonstrate that our introduction of AL can effectively alleviate the modal differences between visible and infrared images. Meanwhile, we find that the intraclass distance of HAResformer becomes more concentrated, indicating that our method can effectively reduce intramodality variation.

Additionally, we visualized the feature distribution in the 2-D feature space using t-SNE [55], as shown in the second row of Fig. 7. Compared with B, the interclass distributions of B + AL are more spaced and have no intersecting feature families. Compared to B + AL, the feature distribution between the HAResformer modalities becomes more compact. These results demonstrate the effectiveness and generalization of our method.

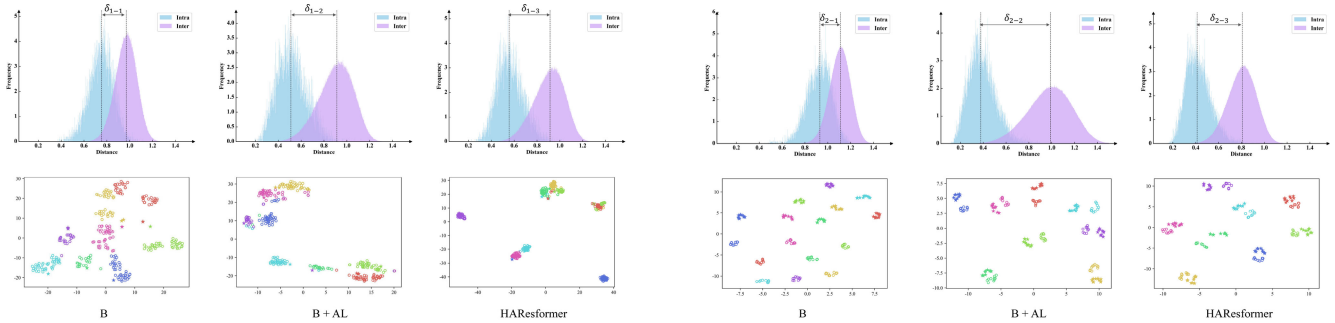

Fig. 7. Intraclass and interclass similarity distribution and t-SNE [55] visualization of feature distribution of different methods on the SYSU-MM01 and RegDB datasets, where $\delta_{1-1} = 0.21$, $\delta_{1-2} = 0.40$, $\delta_{1-3} = 0.36$, $\delta_{2-1} = 0.17$, $\delta_{2-2} = 0.61$, and $\delta_{2-3} = 0.40$. Different colors indicate the image features of a person with various identities, and the star and circle symbols indicate the visible and infrared modalities, respectively. It is best viewed in color. (a) SYSU-MM01. (b) RegDB.

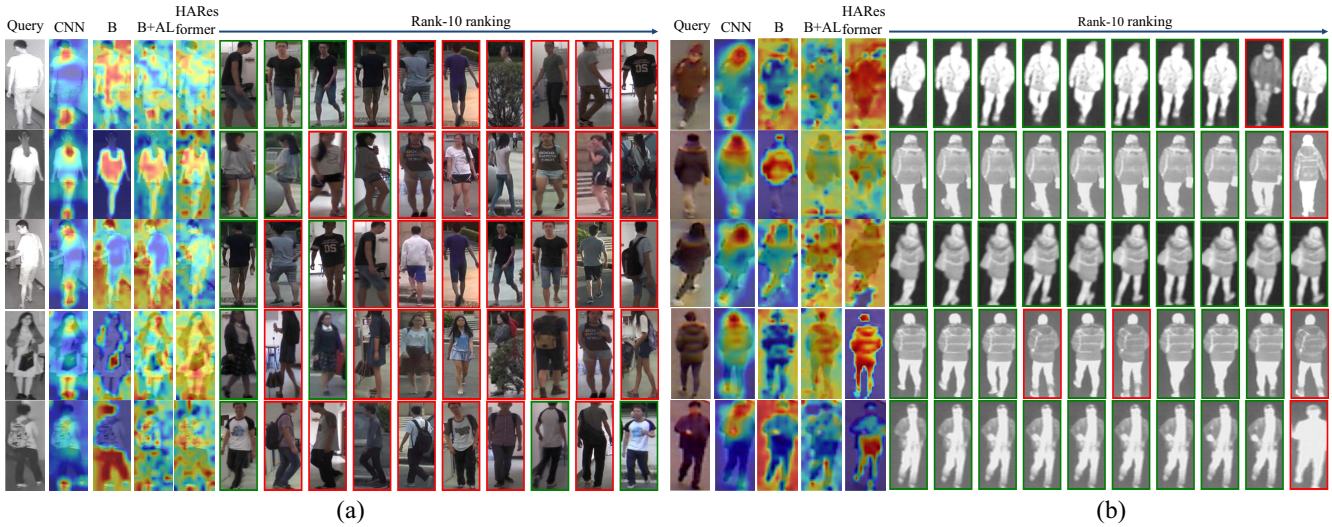

Fig. 8. Grad-CAM visualization of regions emphasized in different methods and top-10 results for some example queries retrieved by HAResformer on SYSU-MM01 and RegDB datasets. The green and red bounding boxes indicate query results matching the same and different identities from the gallery, respectively. Deeper red indicates a higher weight. It is best viewed in color. (a) SYSU-MM01. (b) RegDB.

Visualization of Feature Heatmaps and Top Retrieved Examples: To intuitively demonstrate the attention regions and effectiveness of CNN-based methods and our approach, we randomly selected five identities from the SYSU-MM01 and RegDB datasets to visualize the attention maps using Grad-CAM. Following the settings in the previous section, we used the following settings: CNN, B, B + AL, and HAResformer. The visualization results are shown in Fig. 8. We draw the following conclusions.

1) CNN-based methods focus on local regions of person parts (e.g., neck and calves), as shown in the CNN column in Fig. 8(a). CNN-based methods need to consider global spatial information to avoid local overfitting.
2) Compared to CNN-based methods, our architecture combines the advantages of CNN and ViT, capturing global contextual cues and specific local regions, as shown in column B in Fig. 8(a) and (b). Compared to B, B + AL emphasizes more fine-grained features, enhancing comprehensive global feature representation. Furthermore, based on B + AL, our HAResformer architecture shows remarkable results, focusing more on effective discrimination regions of the person. These conclusions demonstrate the effectiveness of our method.

Additionally, we present the top-10 results of example queries retrieved by our method on the SYSU-MM01 and RegDB datasets, as shown in Fig. 8. It can be observed that even humans struggle to judge the correct match based on body pose and clothing style, e.g., Rank-4, Rank-6, and Rank-10 in the fourth row of Fig. 8(b), making the VI-ReID task extremely challenging. Notably, our HAResformer architecture shows promising query results. Five randomly selected query examples correctly match individuals of the same and different identities from the gallery. The visualization results further verify the effectiveness and superiority of HAResformer.

D. Comparison With State-of-the-art Methods

In this section, we compare the proposed HAResformer architecture with some state-of-the-art VI-ReID methods published in recent years, including CNN-based methods (JSIA-ReID [29], DDAG [19], HC-Tri [12], GECNet [22], AGW [20], NFS [21], DTRM [23], DMiR [24], G^{2}DA [25],

TABLE VI
COMPARISON WITH THE STATE-OF-THE-ART METHODS ON THE SYSU-MM01 DATASET UNDER TWO SEARCH MODES. REID RATES AT RANK-1, 10, 20 (%), AND MAP (%)

Method	Classification of Methods	All-search				Indoor-search			
		Rank-1	Rank-10	Rank-20	mAP	Rank-1	Rank-10	Rank-20	mAP
JSIA-ReID [29]	CNN-based	38.10	80.70	89.90	36.90	43.80	86.20	94.20	52.90
DDAG [19]		54.75	90.39	95.81	53.02	61.02	94.06	98.41	67.98
LbA [56]		55.40	-	-	54.10	58.50	-	-	66.30
HC-Tri [12]		61.68	93.10	97.17	57.51	63.41	91.69	95.28	68.17
GECNet [22]		53.37	89.86	95.66	51.83	60.60	94.29	98.10	62.89
AGW [20]		47.50	84.39	92.14	47.65	54.17	91.14	95.98	62.97
NFS [21]		56.91	91.34	96.52	55.45	62.79	96.53	99.07	69.79
DTRM [23]		63.03	93.82	97.56	58.63	66.35	95.58	98.80	71.76
DMiR [24]		50.54	88.12	94.86	49.29	53.92	92.50	97.09	62.49
G^2DA [25]		57.07	90.99	96.28	55.05	63.70	94.06	98.35	69.83
FAM [26]		55.75	87.51	93.27	51.52	58.24	91.08	96.42	65.65
DFLN-ViT [17]	ViT-based	59.84	92.49	97.20	57.70	62.13	94.83	98.24	69.03
SPOT [40]		65.34	92.73	97.04	62.25	69.42	96.22	99.12	74.63
TVTR [15]		65.30	-	-	**64.15**	**72.21**	-	-	**77.94**
CMTR [41]		65.45	**94.47**	**98.16**	62.90	71.46	97.16	99.22	76.67
HAResformer (Ours)	ViT-based	**65.92**	92.51	97.27	60.37	67.69	**97.66**	**99.57**	73.82

FAM [26], LbA [56]) and ViT-based methods (DFLN-ViT [17], SPOT [40], TVTR [15], CMTR [41], PMT [35]). Due to these methods all following the standard evaluation protocols on the SYSU-MM01, RegDB, and LLCM datasets, we directly use the experimental results of published papers for comparison.

Table VI shows the comparison of our HAResformer architecture with other methods on the SYSU-MM01 dataset. HAResformer obtains competitive results on all performance metrics. We find that ViT-based methods outperform CNN-based methods overall in both evaluation modes, which indicates that ViT can effectively model the long-range dependencies of a person to enhance comprehensive discriminative feature representation in the VI-ReID task. Specifically, compared with the best DTRM with CNN-based backbone construction, our method obtained 2.86% and 1.74% gains in key indicators Rank-1 and mAP on All-search mode, respectively. Comparison results demonstrate that our method can effectively enhance visual representation by hybrid CNN and ViT. Especially, compared to the recent pure ViT-based method TVTR, our method performs poorly in terms of Rank-1 and mAP in the Indoor-search mode. TVTR takes advantage of Top-K vision tokens to reduce identity-irrelevant information, which may be why TVTR performs well in indoor scenes with simple backgrounds. However, in the most challenging All-search mode, our method improves by 0.62% on Rank-1. Compared with DFLN-ViT, which alternates CNN and ViT to extract features, our method significantly improves Rank-1 by 6.08% and 5.56% in both evaluation modes, respectively. This impressive comparison indicates that the HAResformer architecture can better mine local spatial information and global cues. In conclusion, these results verify the superiority of our method for VI-ReID.

Comparison results on the RegDB dataset are listed in Table VII. Due to the person images in the RegDB dataset having better instance alignment and fewer intramodality variations, the performance of all methods outperforms the SYSU-MM01 dataset. Our method achieves superior performance in both evaluation modes. Specifically, compared with the CNN-based two-stream network construction method HC-Tri, in the Visible-to-Infrared mode, the two key indicators Rank-1 and mAP increased by 2.39% and 3.73%, respectively. Similar significant gains also occurred in the Infrared-to-Visible mode. Compared with DFLN-ViT, Rank-1 decreases by 0.18% in Infrared-to-Visible mode, while mAP significantly improves by 2.56%. These comparison results demonstrate that our HAResformer has strong discriminative feature representations.

Combining Table VI and Table VII to compare the experimental results of different datasets, we find that ViT-based methods CMTR, TVTR, and SPOT have smaller gaps in comparing results with our proposed HAResformer architecture on the SYSU-MM01 dataset on various performance metrics in both evaluation modalities. Even these methods outperform HAResformer on Rank-1 and mAP in Indoor-search mode. However, a fantastic fact happens on the RegDB dataset, where HAResformer achieves a significant performance improvement compared to these methods. For instance, compared with SPOT, the Rank-1 in the two evaluation modes is improved by 13.09% and 11.66%, respectively. These comparison results on different datasets demonstrate that our HAResformer architecture has better generalization and robustness.

In addition, this work validates the scalability of HAResformer on the large-scale LLCM dataset, captured under complex low-light conditions in a real-world environment, experimental results are shown in Table VIII. Our method obtains encourages performance in both search modes. Compared to the LbA method, our method achieves significant improvements of 2.30% in Rank-1 and 1.14% in mAP on the Visible-to-Infrared. On the Infrared-to-Visible, it achieves a Rank-1 gain of 0.86%. The experimental results confirm the strong robustness and generalization capabilities of our method. However, it is undeniable that the performance of all methods is noticeably lower than their results on other datasets. This observation underscores the challenges of complex low-light conditions and motivates further research to address these issues.

TABLE VII
COMPARISON WITH THE STATE-OF-THE-ART METHODS ON THE REGDB DATASET UNDER TWO SEARCH MODES. REID RATES AT RANK-1, 10, 20 (%), AND MAP (%)

Method	Classification of Methods	Visible-to-Infrared				Infrared-to-Visible			
		Rank-l	Rank-10	Rank-20	mAP	Rank-l	Rank-10	Rank-20	mAP
JSIA-ReID [29]	CNN-based	48.50	-	-	49.30	48.10	-	-	48.90
DDAG [19]		69.34	86.19	91.49	63.46	68.06	85.15	90.31	61.80
LbA [56]		72.40	-	-	67.60	67.50	-	-	72.40
HC-Tri [12]		91.05	97.16	98.57	83.29	89.30	96.41	98.16	81.46
GECNet [22]		82.33	92.72	95.49	78.45	78.93	91.99	95.44	75.58
AGW [20]		70.05	86.21	91.55	66.37	70.49	87.12	91.84	65.90
NFS [21]		80.54	91.96	95.07	72.10	77.95	90.45	93.62	69.79
DTRM [23]		79.09	92.25	95.66	70.09	78.02	91.75	95.19	69.56
DMiR [24]		75.79	89.86	94.18	69.97	73.93	89.87	93.98	68.22
G^2DA [25]		71.72	87.13	91.92	65.90	69.50	84.87	89.85	63.88
FAM [26]		87.31	95.67	97.49	76.70	84.81	94.33	96.48	74.73
DFLN-ViT [17]	ViT-based	92.10	97.97	**99.17**	82.11	**91.21**	**98.20**	**99.08**	81.62
SPOT [40]		80.35	93.48	96.44	72.46	79.37	92.79	96.01	72.26
TVTR [15]		84.10	-	-	79.50	83.70	-	-	78.00
PMT [35]		84.83	-	-	76.55	84.16	-	-	75.13
CMTR [41]		88.11	-	-	81.66	84.92	-	-	80.79
HAResformer (Ours)	ViT-based	**93.44**	**98.21**	99.01	**87.02**	91.03	97.05	98.26	**84.18**

TABLE VIII
COMPARISON WITH THE STATE-OF-THE-ART METHODS ON THE LLCM DATASET UNDER TWO SEARCH MODES. REID RATES AT RANK-1, 10, 20 (%), AND MAP (%)

Method	Classification of Methods	Visible-to-Infrared				Infrared-to-Visible			
		Rank-l	Rank-10	Rank-20	mAP	Rank-l	Rank-10	Rank-20	mAP
DDAG [19]	CNN-based	48.50	81.00	87.80	53.00	41.00	73.40	81.90	49.60
LbA [56]		50.80	84.60	91.10	55.90	44.60	**78.20**	**86.80**	**53.80**
HAResformer (Ours)	ViT-based	**53.10**	**85.42**	**91.11**	**57.04**	**45.46**	76.86	84.21	52.52

V. CONCLUSION

In this article, we proposed a hybrid ResNet-transformer hierarchical aggregation architecture (HAResformer) for VI-ReID. HAResformer aims to enhance comprehensive discriminative feature representations, effectively alleviating modality differences and reducing intramodality variations. To capture specific local spatial information and global contextual cues, we propose a HFE framework. This framework consists of a RE for local features and a TE for global information. Additionally, we introduce a lightweight CFFM to merge these features, which are then fed into the next stage TE for deep interaction. Furthermore, recognizing that different feature extraction stages emphasize different semantic information, we introduce a DSA strategy for auxiliary supervision of multiscale features. We also propose a HGAE to aggregate these features and enhance multigranularity feature representation. Extensive experiments on three datasets demonstrate the effectiveness and generalization of our HAResformer architecture, outperforming most state-of-the-art methods.

REFERENCES

[1] H. Luo et al., "A strong baseline and batch normalization neck for deep person re-identification," *IEEE Trans. Multimedia*, vol. 22, no. 10, pp. 2597–2609, Oct. 2020.

[2] S. He, H. Luo, P. Wang, F. Wang, H. Li, and W. Jiang, "Transreid: Transformer-based object re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 15013–15022.

[3] G. Zhang, P. Zhang, J. Qi, and H. Lu, "HAT: Hierarchical aggregation transformers for person re-identification," in *Proc. 29th ACM Int. Conf. Multimedia*, 2021, pp. 516–525.

[4] C. Song, Y. Huang, W. Ouyang, and L. Wang, "Mask-guided contrastive attention model for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1179–1188.

[5] X. Chen et al., "Salience-guided cascaded suppression network for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 3300–3310.

[6] A. Wu, W.-S. Zheng, H.-X. Yu, S. Gong, and J. Lai, "RGB-infrared cross-modality person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2017, pp. 5380–5389.

[7] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2017, pp. 618–626.

[8] N. Huang, J. Liu, Y. Luo, Q. Zhang, and J. Han, "Exploring modality-shared appearance features and modality-invariant relation features for cross-modality person re-identification," *Pattern Recognit.*, vol. 135, Mar. 2023, Art. no. 109145.

[9] N. Huang, B. Xing, Q. Zhang, J. Han, and J. Huang, "Co-segmentation assisted cross-modality person re-identification," *Inf. Fusion*, vol. 104, Apr. 2024, Art. no. 102194.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[11] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[12] H. Liu, X. Tan, and X. Zhou, "Parameter sharing exploration and hetero-center triplet loss for visible-thermal person re-identification," *IEEE Trans. Multimedia*, vol. 23, pp. 4414–4425, Dec. 2020.

[13] C. Jambigi, R. Rawal, and A. Chakraborty, "MMD-ReID: A simple but effective solution for visible-thermal person ReID," 2021, *arXiv:2111.05059*.

[14] J. Gong, S. Zhao, and K.-M. Lam, "Interaction and alignment for visible-infrared person re-identification," in *Proc. IEEE Int. Conf. Pattern Recognit.*, 2022, pp. 2253–2259.

[15] B. Yang, J. Chen, and M. Ye, "Top-k visual tokens transformer: Selecting tokens for visible-infrared person re-identification," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2023, pp. 1–5.

[16] Z. Chai, Y. Ling, Z. Luo, D. Lin, M. Jiang, and S. Li, "Dual-stream transformer with distribution alignment for visible-infrared person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 11, pp. 6764–6776, Nov. 2023.
[17] J. Zhao, H. Wang, Y. Zhou, R. Yao, S. Chen, and A. El Saddik, "Spatial-channel enhanced transformer for visible-infrared person re-identification," *IEEE Trans. Multimedia*, vol. 25, pp. 3668–3680, Mar. 2022.
[18] L. Zhang, X. Chen, J. Zhang, R. Dong, and K. Ma, "Contrastive deep supervision," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 1–19.
[19] M. Ye, J. Shen, D. J. Crandall, L. Shao, and J. Luo, "Dynamic dual-attentive aggregation learning for visible-infrared person re-identification," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 229–247.
[20] 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, Jun. 2022.
[21] Y. Chen, L. Wan, Z. Li, Q. Jing, and Z. Sun, "Neural feature search for RGB-infrared person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 587–597.
[22] X. Zhong, T. Lu, W. Huang, M. Ye, X. Jia, and C.-W. Lin, "Grayscale enhancement colorization network for visible-infrared person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 3, pp. 1418–1430, Mar. 2022.
[23] 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, 2021.
[24] W. Hu, B. Liu, H. Zeng, Y. Hou, and H. Hu, "Adversarial decoupling and modality-invariant representation learning for visible-infrared person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 8, pp. 5095–5109, Aug. 2022.
[25] L. Wan, Z. Sun, Q. Jing, Y. Chen, L. Lu, and Z. Li, "G2DA: Geometry-guided dual-alignment learning for RGB-infrared person re-identification," *Pattern Recognit.*, vol. 135, Mar. 2023, Art. no. 109150.
[26] B. Wu, Y. Feng, Y. Sun, and Y. Ji, "Feature aggregation via attention mechanism for visible-thermal person re-identification," *IEEE Signal Process. Lett.*, vol. 30, pp. 140–144, Feb. 2023.
[27] M. Ye, X. Lan, Z. Wang, and P. C. Yuen, "Bi-directional center-constrained top-ranking for visible thermal person re-identification," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 407–419, 2020.
[28] M. Ye, X. Lan, J. Li, and P. Yuen, "Hierarchical discriminative learning for visible thermal person re-identification," in *Proc. AAAI Conf. Artif. Intell.*, vol. 32, 2018, pp. 1–8.
[29] G.-A. Wang et al., "Cross-modality paired-images generation for RGB-infrared person re-identification," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, 2020, pp. 12144–12151.
[30] Y. Qian and S.-K. Tang, "Pose attention-guided paired-images generation for visible-infrared person re-identification," *IEEE Signal Process. Lett.*, vol. 31, pp. 346–350, Jan. 2024.
[31] P. Dai, R. Ji, H. Wang, Q. Wu, and Y. Huang, "Cross-modality person re-identification with generative adversarial training," in *Proc. Int. Joint Conf. Artif. Intell.*, vol. 1, 2018, p. 6.
[32] D. Li, X. Wei, X. Hong, and Y. Gong, "Infrared-visible cross-modal person re-identification with an x modality," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, 2020, pp. 4610–4617.
[33] M. Ye, J. Shen, and L. Shao, "Visible-infrared person re-identification via homogeneous augmented tri-modal learning," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 728–739, 2020.
[34] Y. Zhang, Y. Yan, Y. Lu, and H. Wang, "Towards a unified middle modality learning for visible-infrared person re-identification," in *Proc. 29th ACM Int. Conf. Multimedia*, 2021, pp. 788–796.
[35] H. Lu, X. Zou, and P. Zhang, "Learning progressive modality-shared transformers for effective visible-infrared person re-identification," in *Proc. AAAI Conf. Artif. Intell.*, vol. 37, 2023, pp. 1835–1843.
[36] A. Vaswani et al., "Attention is all you need," in *Proc. 31st Conf. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.
[37] X. Huo et al., "HiFuse: Hierarchical multi-scale feature fusion network for medical image classification," *Biomed. Signal Process. Control*, vol. 87, Jan. 2024, Art. no. 105534.
[38] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 213–229.
[39] S. Zheng et al., "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 6881–6890.
[40] C. Chen, M. Ye, M. Qi, J. Wu, J. Jiang, and C.-W. Lin, "Structure-aware positional transformer for visible-infrared person re-identification," *IEEE Trans. Image Process.*, vol. 31, pp. 2352–2364, 2022.
[41] T. Liang, Y. Jin, W. Liu, and Y. Li, "Cross-modality transformer with modality mining for visible-infrared person re-identification," *IEEE Trans. Multimedia*, vol. 25, pp. 8432–8444, May 2023.
[42] 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.
[43] Y. Feng et al., "Visible-infrared person re-identification via cross-modality interaction transformer," *IEEE Trans. Multimedia*, vol. 25, pp. 7647–7659, Nov. 2022.
[44] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.
[45] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.
[46] Y. Cheng, X. Li, G. Xiao, W. Ma, and X. Gou, "Dual-path deep supervision network with self-attention for visible-infrared person re-identification," in *Proc. IEEE Int. Symp. Circuits Syst.*, 2021, pp. 1–5.
[47] Y. Cheng, G. Xiao, X. Tang, W. Ma, and X. Gou, "Two-phase feature fusion network for visible-infrared person re-identification," in *Proc. IEEE Int. Conf. Image Process.*, 2021, pp. 1149–1153.
[48] D. Hendrycks and K. Gimpel, "Gaussian error linear units (GELUs)," 2016, *arXiv:1606.08415*.
[49] Y. Qian, X. Yang, and S.-K. Tang, "Dual-space aggregation learning and random erasure for visible infrared person re-identification," *IEEE Access*, vol. 11, pp. 75440–75450, 2023.
[50] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 815–823.
[51] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *J. Mach. Learn. Res.*, vol. 13, no. 25, pp. 723–773, 2012.
[52] D. T. Nguyen, H. G. Hong, K. W. Kim, and K. R. Park, "Person recognition system based on a combination of body images from visible light and thermal cameras," *Sensors*, vol. 17, no. 3, p. 605, 2017.
[53] 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 Recognit.*, 2023, pp. 2153–2162.
[54] W. Li, K. Qi, W. Chen, and Y. Zhou, "Unified batch all triplet loss for visible-infrared person re-identification," in *Proc. IEEE Int. Joint. Conf. Neural Netw.*, 2021, pp. 1–8.
[55] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 86, pp. 2579–2605, 2008.
[56] 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.

Yongheng Qian received the M.S. degree from Qingdao University of Science and Technology, Qingdao, China, in 2017. He is currently pursuing the Ph.D. degree with Macao Polytechnic University, Macau, SAR, China.

His current research interests include computer vision and pattern recognition.

Su-Kit Tang (Member, IEEE) received the Doctor degree in computer science from Sun Yat-sen University, Guangzhou, China.

He is an Associate Professor with Macao Polytechnic University, Macau, China, where he is serving as the Coordinator of the Computing Program. His research interests encompass a wide range of topics, including machine learning, blockchain technology, smart city development, Internet of Things, networking protocols, network security, and ad hoc wireless networks.