# Structure-Aware Positional Transformer for Visible-Infrared Person Re-Identification

Cuiqun Chen, Mang Ye, Meibin Qi, Jingjing Wu, Jianguo Jiang, and Chia-Wen Lin, *Fellow, IEEE*

***Abstract*— Visible-infrared person re-identification (VI-ReID) is a cross-modality retrieval problem, which aims at matching the same pedestrian between the visible and infrared cameras. Due to the existence of pose variation, occlusion, and huge visual differences between the two modalities, previous studies mainly focus on learning image-level shared features. Since they usually learn a global representation or extract uniformly divided part features, these methods are sensitive to misalignments. In this paper, we propose a structure-aware positional transformer (SPOT) network to learn semantic-aware sharable modality features by utilizing the structural and positional information. It consists of two main components: attended structure representation (ASR) and transformer-based part interaction (TPI). Specifically, ASR models the modality-invariant structure feature for each modality and dynamically selects the discriminative appearance regions under the guidance of the structure information. TPI mines the part-level appearance and position relations with a transformer to learn discriminative part-level modality features. With a weighted combination of ASR and TPI, the proposed SPOT explores the rich contextual and structural information, effectively reducing cross-modality difference and enhancing the robustness against misalignments. Extensive experiments indicate that SPOT is superior to the state-of-the-art methods on two cross-modal datasets. Notably, the Rank-1/mAP value on the SYSU-MM01 dataset has improved by 8.43%/6.80%.**

***Index Terms*— Visible-infrared person re-identification, transformer, structure information, interaction learning.**

## I. Introduction

PERSON re-identification (ReID) [1]–[10] task accomplishes matching the same pedestrian images photographed in different positions or spectral types of cameras. This technology provides assistance to many surveillance systems like streets and shopping malls. Most existing efforts concentrate on single-modality person re-identification using RGB images captured by visible spectrum cameras, which have limited capability for nighttime surveillance. The visible-infrared person re-identification (VI-ReID) [11]–[14] achieves cross-modality matching between visible and infrared images and promotes the application of ReID in a practical system.

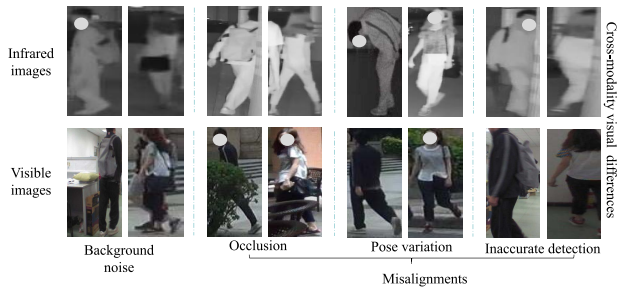



Fig. 1. VI-ReID challenges include cross-modality visual differences, misalignments, and background noise. Examples are selected from the SYSU-MM01 dataset [11].

Given a visible/infrared image of the pedestrian to be queried, VI-ReID needs to find all images of the same person from an infrared/visible gallery set. Compared with the single-modality ReID, VI-ReID is a more challenging pedestrian retrieval task due to the large intra-class differences. As shown in Fig. 1, the intra-class differences may be caused by three factors, *i.e.,* cross-modality visual differences, misalignments, and background noise. Unlike visible images, infrared images lack discriminant information caused by the different imaging principles of visible and infrared images. Similar to the single-modality ReID task, the occlusion, pose variation, and inaccurate person detection in cross-modality monitoring scenarios bring extensive misalignments for VI-ReID. Moreover, due to the particularity of data acquisition environments, VI-ReID usually suffers from severe background noise.

Most previous VI-ReID studies [11], [13]–[16] propose to learn sharable cross-modality appearance features to address the above challenges. In general, these methods first employ two identical and non-shared weight networks to extract modality-specific features of visible and infrared images separately to handle cross-modality visual differences. Then, the shared networks are adopted to learn modality-shared features and conduct cross-modal similarity optimization. Although these methods can extract certain modality-shared features and improve the performance of the models, their robustness against misalignments and background noise is bounded because they only learn coarse shared features from

Manuscript received June 11, 2021; revised November 7, 2021; accepted December 19, 2021. Date of publication March 2, 2022; date of current version March 14, 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 61771180, Grant 61876056, and Grant 62176188; in part by the Key Research and Development Program of Hubei Province under Grant 2021BAA187; and in part by the CAAI-Huawei MindSpore Open Fund. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Heng Tao Shen. *(Corresponding author: Mang Ye.)*

Cuiqun Chen, Meibin Qi, Jingjing Wu, and Jianguo Jiang are with the School of Computer Science and Information Engineering, Hefei University of Technology, Hefei, Anhui 230601, China (e-mail: chencuiqun_hfut@163.com; qimeibin@163.com; hfutwujingjing@mail.hfut.edu.cn; jgjiang@hfut.edu.cn).

Mang Ye is with the National Engineering Research Center for Multimedia Software, Hubei Key Laboratory of Multimedia and Network Communication Engineering, School of Computer Science, Wuhan University, Wuhan 430072, China (e-mail: mangye16@gmail.com).

Chia-Wen Lin is with the Department of Electrical Engineering, National Tsing Hua University, Hsinchu 30013, Taiwan (e-mail: cwlin@ee.nthu.edu.tw).

Digital Object Identifier 10.1109/TIP.2022.3141868

1941-0042 © 2022 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: Northeastern University. Downloaded on October 04,2022 at 03:44:41 UTC from IEEE Xplore. Restrictions apply.

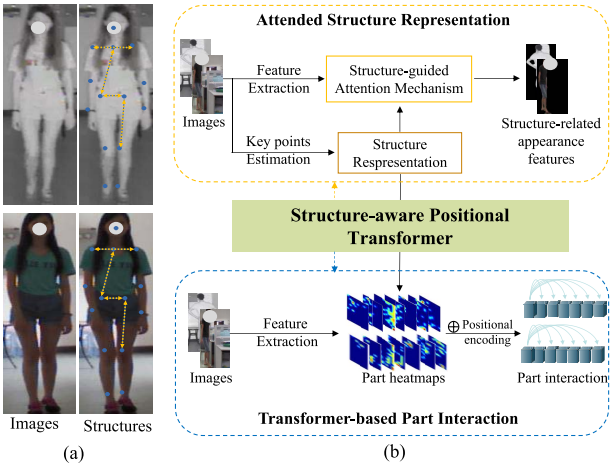



Fig. 2. (a) Illustration of structure information. It describes the connections between key points of the human body. Being sensor-invariant and identity-dependent, structure information plays a crucial role in cross-modal feature representations, which is neglected by most existing VI-ReID methods. (b) The key components in structure-aware positional transformer (SPOT): attended structure representation (ASR) learns the structure-related appearance features under the guidance of structure information to alleviate the impact of the complicated background noise in two modalities; transformer-based part interaction (TPI) explores the rich part-level contextual relations with a transformer encoder, guaranteeing the robustness against misalignment.

image-level information. In order to reduce the influence of sample noise, some works [16], [17] perform part-level feature learning for VI-ReID, capturing fine-grained local information for each modality. Nevertheless, these methods uniformly divide the person images into several regions for region-level local feature concatenation, which is sensitive to occlusions and pose variations. Worse still, they learn part-level representation independently without considering their potential interaction between different local parts, which cannot adaptively mine crucial regions. To enhance the robustness against background noise and misalignments for cross-modality learning, we firstly present our solution ideas from the following two aspects:

(1) *Semantic structure information modeling*: we propose to explore fine-grained modality-shared features rather than coarse sensor-sensitive appearance features. As a kind of soft-biometric feature, the structure information of pedestrians is identity-related and sensor-invariant, which is ignored by most existing VI-ReID methods. In other words, different pedestrians usually have different body structures, while the person images of the same pedestrian under different modalities have the same structural information. As shown in Fig. 2(a), the structure information generally refers to the relations between different key points of the human body [18]. These relations represent various biological characteristics that are discriminatory information for pedestrians. The structure information could facilitate discriminative modality-shared feature learning to achieve fine-grained cross-modality matching and ease the background noise problem. For infrared images without color cues, the structure information is extremely important in distinguishing different pedestrians.

(2) *Positional part interaction mining*: to deal with spatial misalignments in cross-modality learning, we perform interaction learning between different positions. Considering the particularity of the positions of the pedestrian body parts, the positional interaction could discover the underlying structure relationship between regions. This significantly enhances the robustness of the model when some local regions are occluded or distracted by noise. Also, it yields a more stable invariance for pose changes. Therefore, it is desirable to explore a positional part interaction to mine the positional structure relation and improve the discriminability of modality features.

Based on the above discussions, this paper proposes a structure-aware positional transformer (SPOT) network to extract semantic-level sharable cross-modality representations, which contains two main components, i.e., attended structure representation (ASR) and transformer-based part interaction (TPI), as shown in Fig. 2(b). To handle the complex background noise in each modality, ASR learns the appearance features associated with the structure through a modified attention mechanism. Firstly, the structure features are characterized utilizing the relations between key points heatmaps of the human body. These relations collect various junctions between body parts, independent of environmental noise and modality. Then, structure-guided attention updates modality appearance features by determining the importance of each node under the guidance of the structure features. With the supervision of structure information, robust sharable cross-modality appearance features are achieved.

TPI solves the spatial misalignments problem by obtaining rich contextual and structural information between different body parts to aggregate differentiated region-level features. Inspired by the success of transformer in natural language processing (NLP) [19], TPI generates a part sequence for each modality image through a node-level partitioning strategy, and a positional transformer learns part-level enhanced representations from all body parts. In this manner, TPI dynamically captures the abundant contextual and structural relations between the body regions of each modality, leading to higher discriminability of modality features and greater robustness against misalignment.

We summarize our major contributions as follows:

- A structure-aware positional transformer (SPOT) network unifies structure-related appearance learning and part-level interaction learning to reinforce the semantically sharable modality representations for VI-ReID.
- An attended structure representation (ASR) module explicitly learns the structure-related appearance feature for each modality to address the complex background noise.
- A transformer-based part interaction (TPI) module enhances the discriminability of part-level modality features by modeling contextual and positional relations, adaptively combining partially recognizable cues to improve robustness against pose changes and occlusions.
- The new baseline for VI-ReID, namely a structure-aware dual-path cross-modality network, is proposed to dynamically propagate the appearance and structure

Authorized licensed use limited to: Northeastern University. Downloaded on October 04,2022 at 03:44:41 UTC from IEEE Xplore. Restrictions apply.

information to the output for modality representations. Extensive experiment results on two representative VI-ReID datasets show the applicability of the proposed method.

## II. Related Work

*Visible-Infrared Person ReID:* With the rapid development of deep learning, various computer vision tasks [20]–[26] have been widely studied. Visible-infrared person re-identification (VI-ReID) [11], [27], [28] is designed to enable video surveillance at night. Given a visible/infrared image of a person, the VI-ReID system finds the corresponding images from the infrared/visible gallery set. Furthermore, existing VI-ReID methods can broadly be classified into two main categories.

The first category [29]–[33] attempts to reduce modality-differences by compensating for specific cues missing from one modality to another. They apply generative adversarial networks [34] to align the image style of the two modalities. Wang *et al.* [31] design an alignment generative adversarial network to realize the pixel-level and feature-level alignments between the two modalities for the VI-ReID task. To mitigate both intra-modality and cross-modality differences, a hierarchical cross-modality disentanglement [30] method decomposes the image information into the ID-discriminative (*e.g.*, body shape, clothes style) factors and ID-excluded (*e.g.*, illumination, pose) factors by an image generation network. Since the color of an image converted from infrared modality to visible modality is arbitrary, it is difficult for these methods to select the target for VI-ReID.

The second category focuses on mining sharable cross-modality features representations through network design [11], [12], [27], [35], [36] or loss function optimization [13], [15], [17]. Wu *et al.* [11] propose a one-path network with a zero-padding strategy to solve the cross-modality ReID problem. Chen *et al.* [36] automate the cross-modal feature selection process based on neural architecture search. Huang *et al.* [27] design a multi-level modality-shared feature extraction network to learn modality-shared appearance representations and modality-invariant relation representations. Ye *et al.* [15] employ the augmented grayscale modality to form a tri-modal metric learning framework to eliminate modal variations. Due to global-level feature learning, the methods mentioned above have limited robustness against the background noise and modality differences. Recently, similar to single-modality ReID [6], [37], [38], some VI-ReID methods [16], [17], [35] propose to explore part-level feature representation to solve the misalignments problem in VI-ReID. Wei *et al.* [17] design a flexible body partition model-based adversarial learning framework to explore detailed information between different modalities. In [35], an intra-modality part-level feature aggregation and cross-modality graph learning [6], [39], [40] are proposed for VI-ReID, effectively selecting discriminative modality representations for cross-modality matching.

However, these methods mainly focus on exploiting the appearance information for cross-modality learning while ignoring the modality-invariant structural information, resulting in limited model performance. In this paper, we explore the semantic-level fine-grained modality features guided by structure information from two aspects. For one thing, we propose an attended structure representation (ASR) that extracts the structure-related appearance features to mitigate the impact of background noises in two modalities. For another thing, to improve the discriminability of part-level modality features, we design a transformer-based part interaction (TPI) module to mine complex positional structure relations with stronger robustness to occlusion and pose variation.

*Pose Information for Person ReID:* Human pose estimation [20] is a popular task in computer vision. In recent researches, many human pose estimation methods based on deep learning have been introduced, and their performance far exceeds that of traditional methods. The pose information obtained by the pose estimation assists in the visual feature extraction for person re-identification, which has been applied to some ReID sub-tasks, such as image/video-based ReID [7], [41]–[44], occluded ReID [25], [45], and text-based person search [5]. Literatures [7], [41], [42] divide the images into several parts according to key points for alignment. Some works [43], [46] combine appearance and posture features to distinguish areas through bilinear pooling, gate mechanism fusion, and other aggregation methods. Due to the direct fusion of the two types of information, these methods are vulnerable to the gap between appearance information and key point information. Different from these methods, Qian *et al.* [18] encode the key point coordinates to obtain the shape features to handle the cloth-changing problem. Gao *et al.* [45] employ key point heatmaps for part feature pooling and then predict the visible regions in a self-supervised manner. Following Gao *et al.* [45], we encode key point heatmaps through a relation network to characterize the structural features of the human body in each modality. Based on the structure information, we further explore the structure-related appearance feature and part-level feature to reduce the influence of background noise and misalignments for VI-ReID.

*Transformer for Vision Tasks:* Transformer [19] based on self-attention has been widely applied in various computer vision tasks, such as image classification [47]–[49], object detection [50], action recognition [51], and segmentation [52]. For person re-identification, He *et al.* [49] first present a pure transformer-based architecture to learn robust feature representation for single-modality ReID. Zhang *et al.* [48] propose a spatial-temporal transformer to handle information extraction on spatial and temporal dimensions for video-based ReID. In this paper, to alleviate the influence of misalignments caused by occlusion and pose variation, we perform the part-level interaction through a transformer encoder to mine the underlying positional relations and improve the discriminant ability of part-level modality representations. To our best knowledge, this is the first study that evaluates the validity of the transformer in visible-infrared person re-identification.

## III. Proposed Method

In this section, we first elaborate on the major components of the structure-aware positional transformer (SPOT) network, namely the attended structure representation (ASR) in

Authorized licensed use limited to: Northeastern University. Downloaded on October 04,2022 at 03:44:41 UTC from IEEE Xplore. Restrictions apply.

Section III-A and the transformer-based part interaction (TPI) in Section III-B. Then, in Section III-C, we introduce the overall network architecture as well as a new baseline for transferring appearance and structure information to the output.

### *A. Attended Structure Representation*

For VI-ReID, the complex background clutter in each modality makes it challenging to learn discriminative appearance features. Pedestrian structural information describes the proportion of body parts, which is identity-related and modality-invariant. If we precisely represent it, the structure information could mitigate the impact of background clutter and promote sharable modality feature learning. Therefore, in this paper, we devise an attended structure representation (ASR) module to explicitly characterize the structure feature and extract the structure-related appearance feature for each modality, as shown in Fig. 3.

*Structure Representation:* Since the structure information is not labeled in ReID task, the structure feature cannot be obtained directly. Motivated by [18], [45], as shown in Fig. 3, we adopt a relation network with four convolutional layers and two pooling layers to deal with the relation reasoning between different key point heatmaps to represent the structure feature, as formulated below:

$$F_S^o = N_r^o(S) \tag{1}$$

where $N_r^o(\cdot)$ represents the $o$-th layer of the relation network. We adopt OpenPose [20] to obtain the key point heatmaps $S \in \mathbb{R}^{K_S \times H_S \times W_S}$ for each modality, of which $K_S$ denotes the number of key points and $H_S \times W_S$ represents the size of each key point heatmap. $F_S^o$ represents the $o$-th layer structure feature, which collects the rich contextual information and shows a stronger feature representation capability.

*Structure-Guided Attention Mechanism:* Due to the intricate background noise in each modality, it is difficult to extract the discriminative and modality-shared features. In this paper, the structure-related appearance features are learned through structure-guided attention which dynamically selects regions of interest. Given a visible/infrared image $I$, the appearance feature map $F_I \in \mathbb{R}^{C_I \times H \times W}$ is computed through a modality-specific network $\mathrm{N}_{\mathrm{specific}\,(\cdot)}$ followed by a modality-shared network $\mathrm{N}_{\mathrm{shared}\,(\cdot)}$, formulated as:

$$F_I = N_{\mathrm{shared}}\left(N_{\mathrm{spectfic}}(I)\right) \tag{2}$$

Before computing the attention matrix, we fuse the structure feature and appearance feature to reduce the impact of the differences in semantic space between these two features. The updated structure feature $\bar{F}_S$ is defined as:

$$\bar{F}_S = W_{\phi 2}\left(ReLU\left(W_{\phi 1}\left([F_I, F_S]\right)\right)\right) \tag{3}$$

where $F_S \in \mathbb{R}^{C_S \times H \times W}$ represents the final structure feature that is the output of the last layer of relation network $N_r$ and has the same size as the appearance feature. $W_{\phi 1} \in \mathbb{R}^{(C_I+C_S)\times(C_I+C_S)}$ and $W_{\phi 2} \in \mathbb{R}^{C_I \times (C_I+C_S)}$ are both parameters of the embedding network $\phi(\cdot)$, which are implemented through a $1 \times 1$ convolution-BN-ReLU layer. With the updated

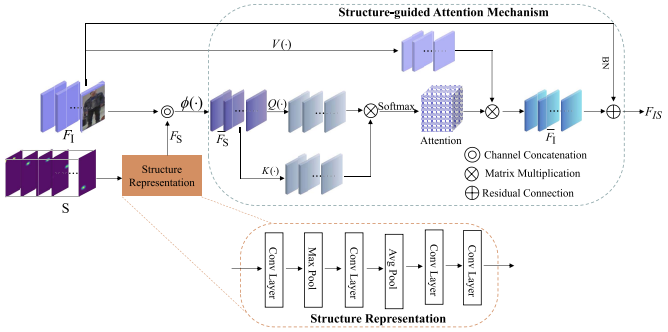



Fig. 3. Illustration of attended structure representation (ASR) module. It contains the structure representation module and structure-guided attention mechanism. For each image, we first extract the structure information by encoding key point heatmaps of the human body, and then the critical areas are highlighted under the guidance of structure information.

structure feature, the spatial relation attention $m_{t,j}$ between the structure feature nodes is then:

$$m_{t,j} = \frac{e^{D(\bar{F}_{S,t}, \bar{F}_{S,j})}}{\sum_{\forall j} e^{D(\bar{F}_{S,t}, \bar{F}_{S,j})}} \tag{4}$$

where $D\left(\bar{F}_{S,t}, \bar{F}_{S,j}\right) = K\left(\bar{F}_{S,t}\right)^T \times Q\left(\bar{F}_{S,j}\right)$ denotes the similarity between feature node $t$ and feature node $j$. $K(\cdot)$ and $Q(\cdot)$ are two embedding functions which are implemented by a $1 \times 1$ convolution layer, mapping structure feature into different semantic spaces. The attention locates discriminatory regions and can be used to guide appearance feature learning:

$$\bar{F}_I = \sum_{t=1}^{n} m_{j,t} \times V\left(F_{I,t}\right) \tag{5}$$

where $V(F_{I,t})$ represents the embedded appearance feature with convolution operation $V(\cdot)$. $\bar{F}_I$ defines the structure-related appearance feature, which highlights the appearance regions associated with the structure information and allows more discriminative appearance information to participate in similarity matching. Specifically, we also adopt the residual batchnorm (BN) connection [35] to keep the training process steady. Therefore, the final structure-related feature is formulated by:

$$F_{IS} = \bar{F}_I + BN\left(F_I\right) \tag{6}$$

With the structure-guided attention mechanism, the identity-related and modality-shared appearance information can be selected, which assists in avoiding the influence of the background noise and minimizing the modality difference.

### *B. Transformer-Based Part Interaction*

To address the spatial misalignment in cross-modality learning, this subsection introduces a transformer-based part interaction (TPI) module that boosts part-level modality representations by incorporating the local appearance and position information. TPI explores the abundant contextual information as well as establishes the positional structure relations between regions, strengthening the discriminability of modality representations.

Authorized licensed use limited to: Northeastern University. Downloaded on October 04,2022 at 03:44:41 UTC from IEEE Xplore. Restrictions apply.

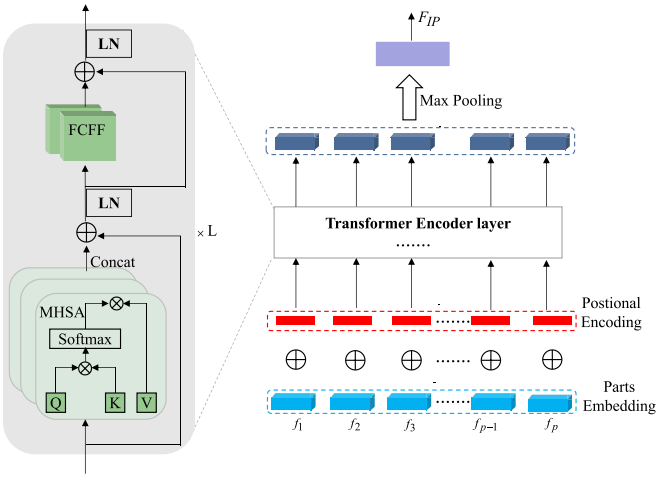



Fig. 4. Illustration of transformer-based part interaction (TPI) module. LN represents the LayerNorm [53]. Given a part-level feature sequence of modality image, TPI selects discriminative part features to gain robustness against occlusion and pose variation. It adaptively enhances part features by mining part-level contextual relations and then adopts a max-pooling to choose part-level features with the most excellent discriminatory power for modality feature representation.

*Transformer:* As shown in Fig. 4, a transformer encoder is made up of $L$ layers. Each layer $l$ consists of two sub-layers: the multi-head self-attention (MHSA) network, and the fully connected feed-forward (FCFF) network. Specifically, MHSA maps the query, key, and value to $h$ different sub-spaces to mine abundant relations between elements and enhance the performance of the model while the total number of model parameters remains unchanged. The FCFF realizes the space transformation by adding the nonlinear function, which can further enhance the differentiating power of features. Similar to literature [19], [51], we employ the sine and cosine functions to denote the positional encoding:

$$PE(pos, i) = \begin{cases} \sin\left(pos/Q^{2k/d_m}\right), & i = 2k \\ \cos\left(pos/Q^{2k/d_m}\right), & i = 2k+1 \end{cases} \tag{7}$$

where $d_m$ represents the dimension of the elements and $i$ is one dimension of the feature vector. $Q$ represents a constant parameter. The positional embedding is an essential component of the transformer and captures the location information of each element. By adding the positional embedding directly to the feature embedding, the positional information of each sequence element is fully integrated with its embedding information and transferred to the high-level features. Moreover, with the embedding of position information, the distance between the elements is distinguished, which conduces to the mining of structural information of sequence.

*Part Interaction:* Assuming the part sequence $F_P = \{f_1, f_2 \ldots f_i, .. f_p \mid f_i \in \mathbb{R}^C, i = \{1, 2, \ldots, p\}\}$ ($p$ is the number of parts.), TPI attempts to explore discriminative contextual information and structural relations between regions to enhance the part-level feature representation. This process involves two steps: part division and part interaction.

First, the simple horizontal pooling methods [16], [35] have limited accuracy due to the harsh noise in VI-ReID. Finer segmentation methods are needed for learning more accurate part-level representations. In this paper, following [45], we adopt the structure information to realize fine-grained local partitioning for cross-modality person ReID. Specifically, the node-level part maps are learned from the structure feature and describe the likelihood that each node belongs to a specific region. And the part map $M \in \mathbb{R}^{p \times H \times W}$ can be denoted as:

$$M = W_\theta (F_S) \tag{8}$$

where $\theta(\cdot)$ with the parameter $W_\theta$ is an embedding function, which encodes the structure feature as the part attention maps through a $1 \times 1$ convolution layer followed by a Sigmoid function. Then, the $i$-th part feature $f_i$ can be obtained by:

$$f_i = \frac{1}{K^i} \sum_{h=1}^{H} \sum_{w=1}^{W} \bar{M}^i_{h,w} \otimes F_{I,h,w} \tag{9}$$

where $\bar{M}^i$ represents the maximum activation along the part dimension. This main idea is to make the network locate non-overlapping parts [45] to extract more discriminant part features. $K^i$ represents the spatial average pooling of $\bar{M}^i$. Then, in order to improve the robustness of part features to misalignments, we propose to model the positional interaction between parts by means of a transformer encoder. And the enhanced part-level feature $\bar{F}_P$ can be formulated by:

$$\bar{F}_P = \text{ Transformer } (F_P + PE) \tag{10}$$

where $PE$ represents the positional encoding of part sequence and measures the order of part maps. Finally, for each modality, the most distinguishing information is selected by maximum pooling along the part dimension to obtain the final part-level feature $F_{IP} \in \mathbb{R}^C$.

Compared with the previous part-based methods [6], [35] in VI-ReID, our TPI presents the following three merits: (1) The node-level part division strategy guided by the structure feature helps to mitigate the effect of occlusion and achieve feature alignment. (2) TPI captures contextual and structural relations among human body parts with multi-head self-attention, strengthening the cross-modality feature representations. (3) The structure features obtained by modeling key point relationships have finer granularity and greater stability, yielding more accurate part-level modality relations than existing methods.

### C. Overall Architecture

The architecture of structure-aware positional transformer (SPOT) is shown in Fig. 5, which contains the structure-aware dual-path cross-modality network, ASR, TPI, and fusion block. Specifically, the structure-aware dual-path cross-modality network extracts modality-shared features by aggregating the appearance and structure information for each modality. The fusion block combines the structure-related appearance feature and part-level feature to obtain the final feature representation of each modality image.

*Structure-Aware Dual-Path Cross-Modality Network:* Structure information is a modality-invariant cue, guiding the

Authorized licensed use limited to: Northeastern University. Downloaded on October 04,2022 at 03:44:41 UTC from IEEE Xplore. Restrictions apply.

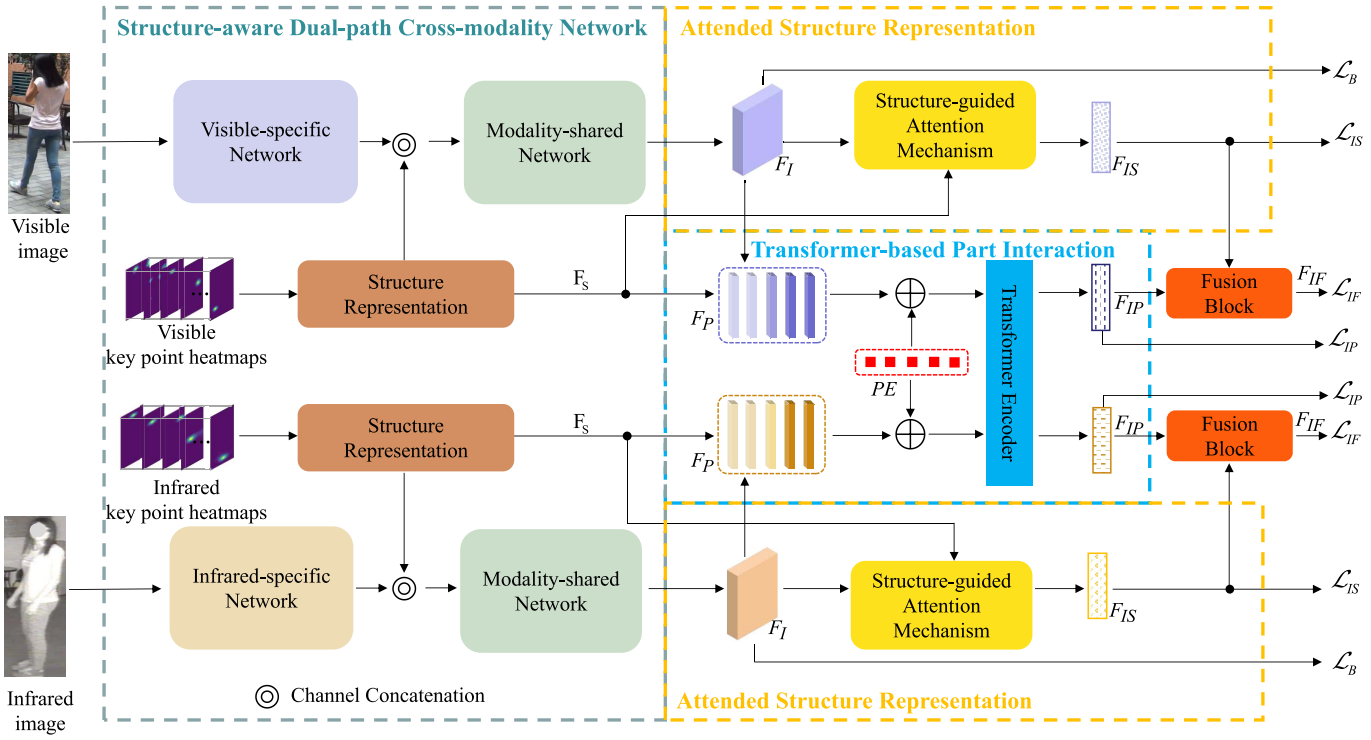



Fig. 5. Structure-aware positional transformer (SPOT) architecture consists of three main parts. The structure-aware dual-path cross-modality network combines structure information with appearance information in the modality-shared network to enhance the discriminability of modality features. Attended structure representation (ASR) alleviates the influence of background noise by learning structure-related appearance features. To improve the robustness against misalignments, transformer-based part interaction (TPI) employs a transformer encoder to enhance part-level features obtained by a node-level partition strategy. The fusion block aggregates the structure-related appearance and part-level features to output the final modality representations for cross-modal matching. Notably, in our experiments, except that the parameters of modality-specific networks do not share weights, other networks of the two modalities all share parameters.

sharable cross-modality feature learning for VI-ReID. To augment the intra-modality visual representation, the structure-aware dual-path network adds the structure information during feature extraction. Considering the visual differences between the two modalities, the backbone network is designed with a dual-path framework, including two components: modality-specific network and modality-shared network, as shown in Fig. 5. The low-level appearance features of each modality are extracted through modality-specific networks which are not parameter-shared. In contrast, the modality-shared networks share the goal of learning the high-level sharable features for two modalities. Right after the modality-specific network, the structure information is integrated to optimize the modality-shared feature representation learning. We assume that the contextual structure information represents a special modality-shared feature. Therefore, we combine the modality-specific feature with the structure feature as the input of the modality-shared network. This encourages more sharable modality information propagation to the high-level convolution layers. Besides, the experimental results in Table II demonstrate this conclusion. We utilize the cross-entropy loss after a BN layer ($\mathcal{L}_{id}$) and triplet loss ($\mathcal{L}_{tri}$) [15], [33], [35] to train the baseline. The objective function of the baseline is defined as:

$$\mathcal{L}_B = \mathcal{L}_{id} + \mathcal{L}_{tri} \tag{11}$$

*Fusion Block:* After obtaining the structure-related appearance and parts features, we propose to aggregate these two features via a fusion block and output the final modality features for similarity matching. Through the joint optimization of these two branches, our model improves the discriminability of each modality feature and obtains stronger robustness against background noise and misalignments. Concretely, to preserve as many of these two features as possible, we employ the channel concatenation for aggregation in the fusion block. This process can be represented by:

$$F_{IF} = \mathrm{W}_\gamma\left([F_{IS}, F_{IP}]\right) \tag{12}$$

where $W_\gamma \in \mathbb{R}^{C\times 2C}$ is learned parameter of the function $\gamma(\cdot)$, which realizes the adaptive selection of differentiated modality channel features.

*Loss Function:* In this paper, we disintegrate the total objective function of our network into four segments, *i.e.*, the baseline feature learning $\mathcal{L}_B$, the structure-level appearance feature learning $\mathcal{L}_{IS}$, the part-level feature learning $\mathcal{L}_{IP}$, and the fused feature learning $\mathcal{L}_{IF}$. And the total loss function is denoted as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_B + \lambda_1\mathcal{L}_{IS} + \lambda_2\mathcal{L}_{IP} + \mathcal{L}_{IF} \tag{13}$$

$$\mathcal{L}_f = -\frac{1}{N}\sum_{i=1}^{N} y_i \log\left(p\left(y_i \mid f\right)\right) \tag{14}$$

Authorized licensed use limited to: Northeastern University. Downloaded on October 04,2022 at 03:44:41 UTC from IEEE Xplore. Restrictions apply.

where $\mathcal{L}_f$ represents the cross-entropy loss and $p(y_i \mid f)$ describes the probability that feature $f$ is predicted as the ground truth $y_i$. $\mathcal{L}_{IS}$, $\mathcal{L}_{IP}$, and $\mathcal{L}_{IF}$ are calculated by eq. 14. Since the appearance and part branches are different tasks, we set the parameters $\lambda_1$ and $\lambda_2$ as the balance weights for joint training of the two branches to promote the learning of more vital discriminating modality features.

## IV. Experiments

### A. Datasets and Experimental Settings

*Datasets:* SYSU-MM01 [11] contains 491 persons captured from six different cameras, which are four visible cameras (cameras 1, 2, 4, and 5) and two infrared cameras (cameras 3 and 6). It collects 287628 RGB images and 15792 IR images. According to [11], [15], [36], we randomly select 395 persons for training and 96 persons for testing in the experiments. The testing process includes two test modes. In *all search* mode, where gallery images come from all visible cameras. In *indoor search* mode with gallery images from cameras 1 and 2. For each test mode, the infrared images are generally set as the probe, and the visible images are set as the gallery.

RegDB [54] is a small dataset, comprising 412 persons, captured by a dual-camera system (with one visible camera and one thermal camera). Each identity has ten optical images and ten thermal images. Following [15], [22], [28], two evaluation ways are used, *i.e.*, visible-to-infrared, infrared-to-visible. This means that if the visible/infrared images are used as the gallery set, the infrared/visible images will be set as the probe set. For each evaluation way, the performance is based on the average experimental results of 10 randomly divided training sets and testing sets.

*Evaluation Metrics:* Following existing works [1], [11], [23], [32], we adopt the Cumulative Matching Characteristics (CMC), mean Average Precision (mAP), and mean Inverse Negative Penalty (mINP) [1] as the evaluation protocols.

*Implementation Details:* For a fair comparison, we adopt the ResNet-50 [55] as the backbone network. The sub-network before the layer2 is used as the modality-specific networks for different modalities, and the remaining two layers are used as the modality-shared network. Specifically, for each modality image, we employ the OpenPose model, which is trained on the COCO dataset [56] to generate $K_S = 56$ heatmaps, including key points heatmaps and part affinity fields [20]. In order to decrease the calculated amount, the parameter $L$ in the transformer is set as 1. And the number of heads (h) in multi-head self-attention is set as 8. For positional encoding, the parameter $d_m$ is set as 2048, and $Q$ is set as 10000 [19].

Following [1], [35], in the training process, we resize the images to $288 \times 144$ for each modality. Besides, horizontal flipping and random erasing for data augmentation are adopted to alleviate the over-fitting issue when training the SPOT model. We train our model with the stochastic gradient descent (SGD) optimizer for 60 epochs. The initial learning rate is 0.1. We select eight identities in a batch, each with four visible images and four infrared images. In experiments, the parameters $\lambda_1$ and $\lambda_2$ are set as 0.6 and 0.5, respectively. And the number of parts ($p$) is set as 7. All experiments are trained end-to-end by Huawei MindSpore [57].

### B. Ablation Study

In this subsection, we first examine the validity of the proposed components. Then, we explore the layers of structure embedding in the structure-aware dual-path cross-modality network. Finally, different fusion ways in the fusion block are discussed. All the experiments and analyses are based on the all search mode and indoor search mode of the SYSU-MM01 dataset.

*1) Evaluation of Each Component:* We design six experiments as shown in Table I. Specifically, 'B' represents the dual-path baseline which shares the parameters after layer2 of ResNet-50 and is trained by $\mathcal{L}_B$. 'S' refers to structure information.

*Structure-Aware Dual-Path Cross-Modality Network:* The index 2 represents the structure-aware dual-path cross-modality network that aggregates the structure and appearance features after the modality-specific network to amplify visual feature representations. Comparing index 1 with index 2. In all search mode and indoor search mode, the accuracy of Rank-1/mAP/mINP is increased by 4.96%/3.73%/2.35% and 3.97%/1.67%/3.89%, respectively. These results indicate that structure information favors enhancing the feature representation and improving model performance.

*ASR:* We add the ASR module to learn the structure-related appearance feature to alleviate the influence of background noise on similarity matching. As shown in index 2 and index 3, the improvements of Rank-1/mAP/mINP are 3.14%/3.32%/3.84% on all search mode. These significant improvements demonstrate that learning structure-related appearance features captures the sharable modality information and reduces the cross-modality difference.

*TPI:* In order to increase the representation of part-level features, the TPI module is adopted to explore the rich part-level positional relations, as shown in index 4. Compared with index 2, the Rank-1/mAP/mINP is increased by 3.93%/3.50%/3.22% in all-search mode, which indicates that the contextual structure relations between parts strengthen the feature robustness against misalignments.

*SPOT:* SPOT jointly optimizes the ASR and TPI to capture the semantically fine-grained information for cross-modality ReID. And the results are shown in index 5. Comparing index 2 with index 5, on two modes, our SPOT outperforms the structure-aware dual-path cross-modality network by 9.93% /8.96% /8.84% and 9.11%/9.09% /6.49% in Rank-1/mAP/mINP, respectively.

Compared with index 5, SPOT with fusion block (index 6) achieves 2.33%/1.99%/2.32% and 2.86%/2.19%/2.59% improvements in Rank-1/mAP/mINP on two modes, respectively. These meaningful improvements indicate that combining structure-related appearance information and discriminative local information realizes the complementarity of performance and further enhances feature discrimination.

*2) Different layers for structure information embedding:* In order to explore the optimal layer for combining appearance feature and structure feature, we inject the structure

Authorized licensed use limited to: Northeastern University. Downloaded on October 04,2022 at 03:44:41 UTC from IEEE Xplore. Restrictions apply.

TABLE I

PERFORMANCE (%) COMPARISON OF EACH COMPONENTS OF OUR PROPOSED METHOD. WE CARRY OUT THESE EXPERIMENTS ON THE SYSU-MM01 DATASET. THE RANK-$r$ (R-$r$), MAP, AND MINP ARE SHOWN

| Index | Methods | | | | | All search | | | | | Indoor search | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | B | S | $\mathcal{L}_{IS}$ | $\mathcal{L}_{IP}$ | $\mathcal{L}_{IF}$ | R-1 | R-10 | R-20 | mAP | mINP | R-1 | R-10 | R-20 | mAP | mINP |
| 1 | ✓ | | | | | 48.12 | 85.44 | 92.71 | 47.57 | 35.35 | 53.48 | 89.80 | 95.52 | 61.68 | 57.87 |
| 2 | ✓ | ✓ | | | | 53.08 | 87.84 | 94.22 | 51.30 | 37.70 | 57.45 | 92.13 | 97.47 | 63.35 | 61.76 |
| 3 | ✓ | ✓ | ✓ | | | 56.22 | 89.38 | 94.91 | 54.62 | 41.54 | 61.89 | 93.55 | 98.15 | 68.53 | 64.61 |
| 4 | ✓ | ✓ | | ✓ | | 57.01 | 90.12 | 95.61 | 54.80 | 40.92 | 61.89 | 94.92 | 98.72 | 68.98 | 65.08 |
| 5 | ✓ | ✓ | ✓ | ✓ | | 63.01 | 92.46 | 96.90 | 60.26 | 46.54 | 66.56 | 95.92 | 98.65 | 72.44 | 68.25 |
| 6 | ✓ | ✓ | ✓ | ✓ | ✓ | 65.34 | 92.73 | 96.97 | 62.25 | 48.86 | 69.42 | 96.22 | 98.99 | 74.63 | 70.84 |

TABLE II

ANALYSIS OF THE POSITION FOR STRUCTURE INFORMATION ADDING. WE CARRY OUT THESE EXPERIMENTS ON THE SYSU-MM01 DATASET. THE RANK-$r$ (R-$r$), MAP, AND MINP ARE SHOWN

| Methods | All search | | | Indoor search | | |
|---|---|---|---|---|---|---|
| | R-1 | mAP | mINP | R-1 | mAP | mINP |
| B | 48.12 | 47.57 | 35.35 | 53.48 | 61.68 | 57.87 |
| B+$S_1$ | 51.88 | 51.36 | 39.15 | 54.56 | 63.39 | 59.72 |
| B+$S_2$ | 53.08 | 51.30 | 37.70 | 57.45 | 63.35 | 61.76 |
| B+$S_3$ | 52.47 | 52.21 | 40.26 | 56.68 | 65.00 | 61.50 |
| B+$S_4$ | 48.63 | 47.99 | 35.71 | 51.36 | 59.86 | 55.78 |
| B+$K_2$ | 45.93 | 44.56 | 31.17 | 45.95 | 55.00 | 50.67 |

information into different layers of ResNet-50. As shown in Table II, 'B+$S_i$' represents that the structure feature is added to the output of layer-$i$ for each modality. By integrating the structure information into the input of modality-shared networks ('B+$S_2$'), the model achieves a high recognition rate with 53.08% and 57.45% Rank-1 on two modes, respectively, which demonstrates that the structure information promotes the learning of sharable cross-modality features. To further verify the effectiveness of structure representation for cross-modality feature learning, we directly combine the appearance information and key point heatmaps information at the layer2 of ResNet-50, as shown in 'B+$K_2$'. From the results, we observe that a straightforward integration is detrimental to cross-modal feature learning due to the large gap between appearance information and key points information.

*3) Different Fusion Methods in Fusion Block:* The fusion strategy between the structure-related appearance feature and the part feature plays an influential role in matching similarity. In order to verify the effectiveness of the proposed fusion block, we design several variants for features aggregation, such as maximum-based aggregation, mean-based aggregation, and weighting-based aggregation. The results shown in Table III demonstrate that the fusion block based on cascading aggregation can effectively improve model performance through channel interaction with convolution operation. The weighed concatenation adaptively enhances the discriminating features for each modality.

## C. Parameter Analysis

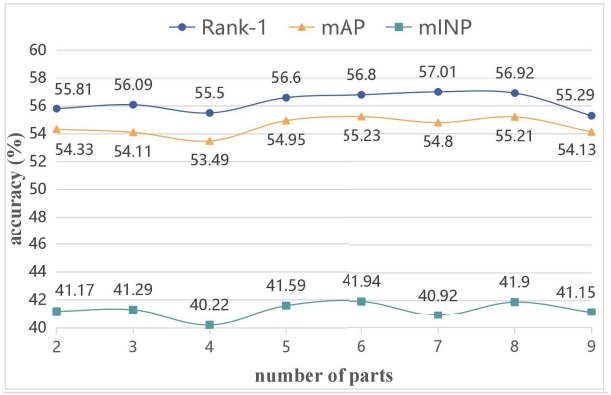
We analyze the impact of the number of parts and the different values of parameters $\lambda_1$ and $\lambda_2$ on the SYSU-MM01 dataset. The results are illustrated in Fig. 6 and Fig. 7.

*The Impact of the Number of Parts:* We extensively experiment on the TPI module by changing the number of parts $p$ from 2 to 9 to explore the most effective setting, as shown in Fig. 6. When $p = 7$, the model realizes the best Rank-1 of two modes, which are 57.01% and 61.89%, respectively.

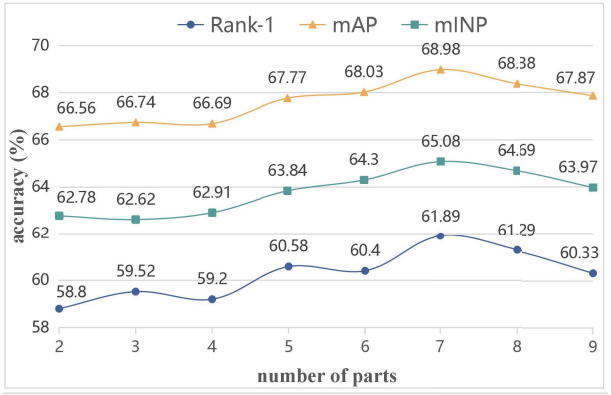
TABLE III

ANALYSIS OF DIFFERENT FUSION METHODS FOR COMBINING STRUCTURE-RELATED APPEARANCE FEATURE AND PART-LEVEL FEATURE. WE CARRY OUT THESE EXPERIMENTS ON THE SYSU-MM01 DATASET. THE RANK-$r$ (R-$r$), MAP, AND MINP ARE SHOWN

| Methods | All search | | | Indoor search | | |
|---|---|---|---|---|---|---|
| | R-1 | mAP | mINP | R-1 | mAP | mINP |
| No fusion | 63.01 | 60.26 | 46.54 | 66.56 | 72.44 | 68.25 |
| $\text{Fusion}_{max}$ | 63.44 | 60.36 | 46.68 | 65.94 | 72.23 | 67.94 |
| $\text{Fusion}_{mean}$ | 63.09 | 60.25 | 46.76 | 67.06 | 72.60 | 65.04 |
| $\text{Fusion}_{weight}$ | 62.94 | 59.81 | 45.52 | 66.04 | 71.86 | 67.48 |
| $\text{Fusion}_{concat}$ | 65.34 | 62.25 | 48.86 | 69.42 | 74.63 | 70.84 |




(a) All search mode




(b) Indoor search mode

Fig. 6. Performance (%) comparison with different number of parts on the SYSU-MM01 dataset. The Rank-1, mAP, and mINP are shown.

Authorized licensed use limited to: Northeastern University. Downloaded on October 04,2022 at 03:44:41 UTC from IEEE Xplore. Restrictions apply.

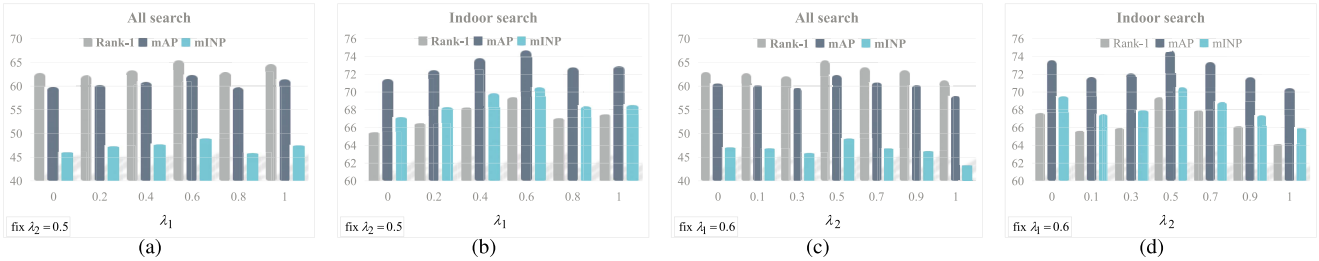



Fig. 7. Performance (%) comparison with different values for parameters $\lambda_1$ and $\lambda_2$ on the SYSU-MM01 dataset. The Rank-1, mAP, and mINP are shown. (a) $\lambda_1$ in all search mode. (b) $\lambda_1$ in indoor search mode. (c) $\lambda_2$ in all search mode. (d) $\lambda_2$ in indoor search mode.

TABLE IV

COMPARISON WITH STATE-OF-THE-ART METHODS ON THE SYSU-MM01 DATASET UNDER TWO EVALUATION SETTINGS. THE RANK-*r* (R-*r*), MAP, AND MINP ARE SHOWN

| Methods | Venue | All search | | | | | Indoor search | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | R-1 | R-10 | R-20 | mAP | mINP | R-1 | R-10 | R-20 | mAP | mINP |
| One-stream [11] | ICCV17 | 12.04 | 49.68 | 66.74 | 13.67 | - | 16.94 | 63.55 | 82.10 | 22.95 | - |
| HCML [12] | AAAI18 | 14.32 | 53.16 | 69.17 | 16.16 | - | 24.52 | 73.25 | 86.73 | 30.08 | - |
| $D^2$RL [33] | CVPR19 | 28.90 | 70.60 | 82.40 | 29.20 | - | - | - | - | - | - |
| AlignGAN [31] | ICCV19 | 42.40 | 85.00 | 93.70 | 40.70 | - | 45.90 | 87.60 | 94.40 | 54.30 | - |
| DFE [58] | MM19 | 48.71 | 88.86 | 95.27 | 48.59 | - | 52.25 | 89.86 | 95.85 | 59.68 | - |
| TCMDL [59] | TCSVT20 | 16.91 | 58.83 | 76.64 | 19.30 | - | 21.60 | 54.26 | 71.83 | 87.91 | - |
| eBDTR [13] | TIFS20 | 27.82 | 67.34 | 81.34 | 28.42 | - | 32.46 | 77.42 | 89.62 | 42.46 | - |
| SDL [60] | TCSVT20 | 28.12 | 70.23 | 83.67 | 29.01 | - | 32.56 | 80.45 | 90.67 | 39.56 | - |
| Hi-CMD [30] | CVPR20 | 34.94 | 77.58 | - | 35.94 | - | - | - | - | - | - |
| JSIA [29] | AAAI20 | 38.10 | 80.70 | 89.90 | 36.90 | - | 43.80 | 86.20 | 94.20 | 52.90 | - |
| CMSP [14] | IJCV20 | 43.56 | 86.25 | - | 44.98 | - | 48.62 | 89.50 | - | 57.50 | - |
| XIV [32] | AAAI20 | 49.92 | 89.79 | 93.96 | 50.73 | - | - | - | - | - | - |
| DDAG [35] | ECCV20 | 54.75 | 90.39 | 95.81 | 53.02 | - | 61.02 | 94.06 | 98.41 | 67.98 | - |
| expAT Loss [61] | TIP21 | 38.57 | 76.64 | 86.39 | 38.61 | - | - | - | - | - | - |
| AGW [1] | TPAMI21 | 47.50 | 84.39 | 92.14 | 47.65 | 35.30 | 54.17 | 91.14 | 95.98 | 62.97 | 59.23 |
| DLS [62] | TMM21 | 48.80 | - | - | 49.00 | - | - | - | - | - | - |
| HAT [15] | TIFS21 | 55.29 | 92.14 | **97.36** | 53.89 | - | 62.10 | 95.75 | **99.20** | 69.37 | - |
| NFS [36] | CVPR21 | 56.91 | 91.34 | 96.52 | 55.45 | - | 62.79 | **96.53** | 99.07 | 69.79 | - |
| SPOT | This paper | **65.34** | **92.73** | 97.04 | **62.25** | **48.86** | **69.42** | 96.22 | 99.12 | **74.63** | **70.48** |

On the other side, as the number of parts increases, the performance gradually degrades. The reason may be that many parts destroy some useful regional information and cause regional mismatching.

*Balance Parameters $\lambda_1$ and $\lambda_2$:* The parameters $\lambda_1$ and $\lambda_2$ control the weights of two modules in the joint training, which facilitate the learning of stronger discriminant features. By fixing the $\lambda_2 = 0.5$, we increase the parameter $\lambda_1$ from 0 to 1.0 to find the best setting of $\lambda_1$. As shown in Fig. 7(a)(b), the best results of two modes are achieved when $\lambda_1 = 0.6$. In addition, we fix the $\lambda_1 = 0.6$ to change parameter $\lambda_2$ and the results are shown in Fig. 7(c)(d). We can conclude that when $\lambda_1 = 0.6$ and $\lambda_2 = 0.5$, two modes reach optimal performance with the 65.43%/62.25%/48.86% and 69.42%/74.63%/70.48% Rank-1/mAP/mINP, respectively.

### D. Comparison With State-of-the-Art Methods

This subsection compares with the state-of-the-arts on SYSU-MM01 and RegDB datasets in Tables IV and V.

*SYSU-MM01:* As shown in Table IV, for all search mode and indoor search mode, our SPOT surpasses the second-best method, NFS [36] which achieves automated feature selection based on neural architecture search, by 8.43%/6.80% and 6.63%/5.14% in Rank-1/mAP, respectively. DDAG [35] proposes to aggregate the part-level features through weighted-part attention. Compared with DDAG, our method adopts a transformer to mine the rich part-level relations and unites the structure-related global appearance features for feature representation, which shows more outstanding performance. More importantly, to decrease the cross-modality difference in the VI-ReID, $D^2$RL [33], AlignGAN [31], and Hi-CMD [30] always use the generative adversarial network to convert cross-modality images into same-modality images. And XIV [32] design third modality to constraint the modality gap. Compared with these methods, SPOT shows superior performance without complex adversarial learning.

*RegDB:* The results on the RegDB dataset are shown in Table V. SPOT achieves comparable performance with NFS [36] on visible-to-infrared evaluation setting but performs significantly better than NFS on challenging infrared-to-visible evaluation setting, achieving the best results with 79.37%/72.26% in Rank-1/mAP. From these results, we conclude that exploring structure-related appearance information and modeling sophisticated part-level feature relations mine rich visual and structural information from infrared and visible images, resulting in discriminative modality-shared features.

### E. Visualizations

We adopt two feature visualization methods to verify the effectiveness of our method, namely feature heatmaps and

Authorized licensed use limited to: Northeastern University. Downloaded on October 04,2022 at 03:44:41 UTC from IEEE Xplore. Restrictions apply.

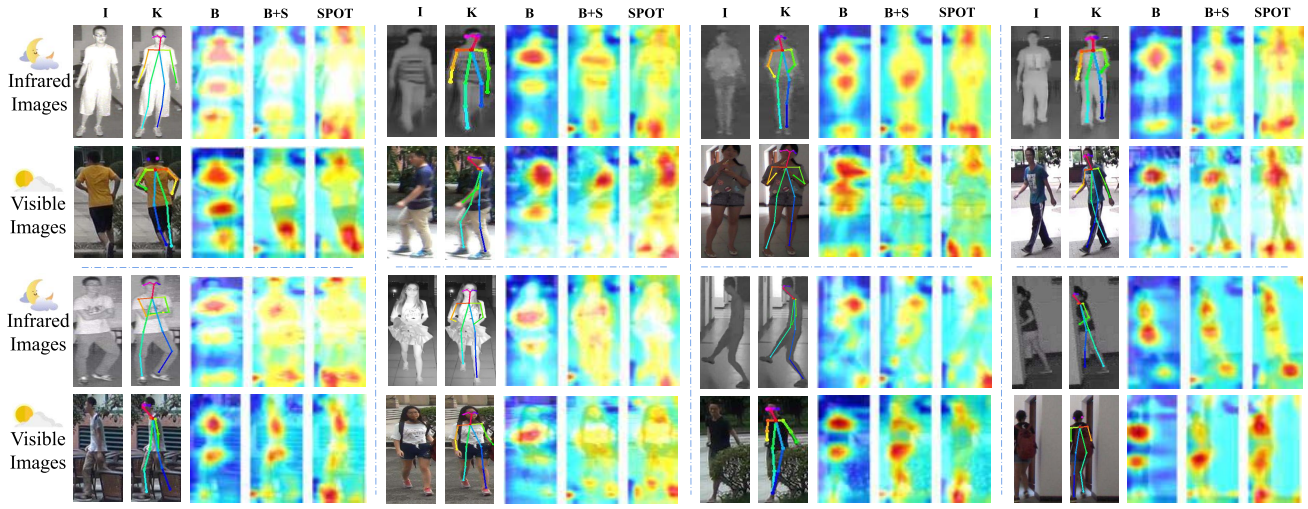
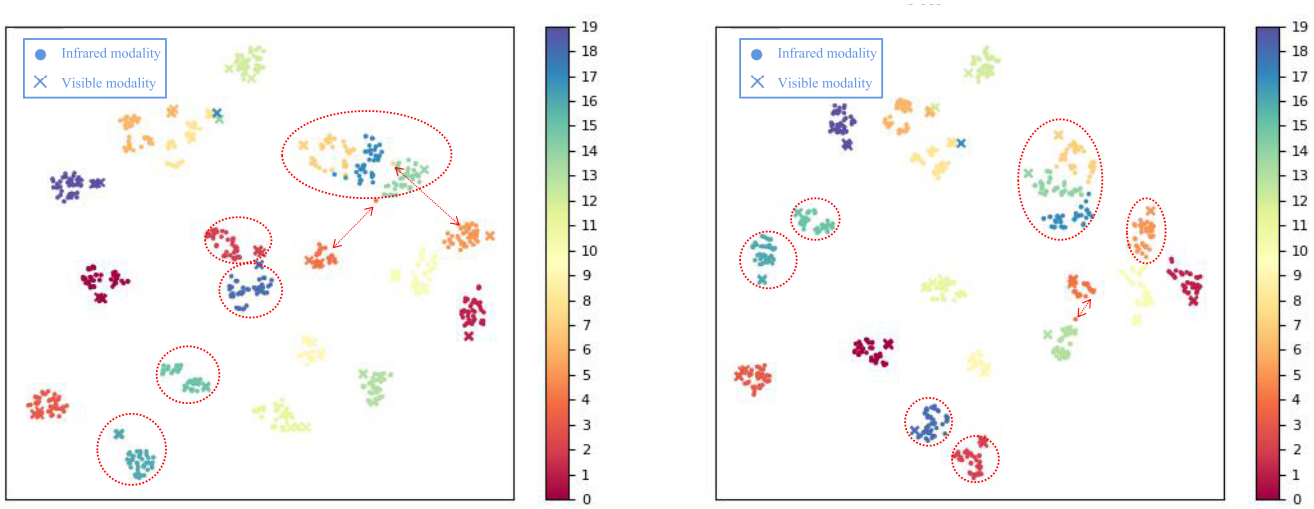



Fig. 8. The visualization results of some images from the SYSU-MM01 test set. We randomly select eight identities and visualize their class activation maps from the B, B+S, and SPOT models, respectively. 'I' refers to the original infrared/visible images, and 'K' represents the detected key points. **Best viewed in color and zoomed in**.




(a) B (b) SPOT

Fig. 9. The visualization results of feature distributions from the baseline and SPOT models on the SYSU-MM01 dataset. Different colors represent different persons. **Best viewed in color and zoomed in**.

feature distribution. The heatmaps [64] directly demonstrates whether the algorithm pays attention to the expected target regions. Furthermore, the feature distribution intuitively observes the compactness within the class and the degree of separation between classes to verify whether the algorithm learns separable features related to identity.

*Visualization of Feature Heatmaps:* We randomly select eight identities from the test set of the SYSU-MM01 dataset and visualize their feature maps learned by the baseline (B), structure-aware dual-path cross-modality network (B+S), and SPOT models, respectively. The following conclusions can be drawn from Fig. 8: (1) The baseline usually reinforces different local regions, which may cause mismatches between persons when misalignments occur. Compared with the baseline, our methods enhance the overall pedestrian features under the guidance of structure representations. For example, in Fig. 8, the leg and foot regions of infrared and visible images. This proves that modeling structure information for each modality effectively reduces the influence of modality differences. (2) Compared with B+S, our SPOT encourages more differentiating features, resulting in greater robustness against occlusion. Through the collaborative research of structure-related features and part-level features, SPOT significantly strengthens the identity-related information regions and weakens the noise information regions [65]. For example, as shown in the last two rows of Fig. 8, when some areas of the pedestrian's body are occluded, SPOT focuses well on the non-obscured and discriminative parts. These results show that the combination of ASR and TPI facilitates shareable cross-modal feature learning, mitigating the modality discrepancy for VI-ReID

*Visualization of Feature Distribution:* Besides, we visualize the feature distributions by randomly sampling twenty

Authorized licensed use limited to: Northeastern University. Downloaded on October 04,2022 at 03:44:41 UTC from IEEE Xplore. Restrictions apply.

TABLE V

COMPARISON WITH STATE-OF-THE-ART METHODS ON THE REGDB DATASET UNDER TWO EVALUATION SETTINGS. THE RANK-$r$ (R-$r$), MAP, AND MINP ARE SHOWN

| Methods | Venue | R-1 | R-10 | R-20 | mAP | mINP |
|---|---|---|---|---|---|---|
| Visible to Infrared | | | | | | |
| HCML [12] | AAAI18 | 24.44 | 47.53 | 56.78 | 20.08 | - |
| $D^2$RL [33] | CVPR19 | 43.40 | 66.10 | 76.30 | 44.10 | - |
| AlignGAN [31] | ICCV19 | 57.90 | - | - | 53.60 | - |
| DFE [58] | MM19 | 70.13 | 86.32 | 91.96 | 69.14 | - |
| SDL [60] | TCSVT20 | 26.47 | 51.34 | 61.22 | 23.58 | - |
| eBDTR [13] | TIFS20 | 34.62 | 58.96 | 68.72 | 33.46 | - |
| JSIA [29] | AAAI20 | 48.10 | - | - | 48.90 | - |
| XIV [32] | AAAI20 | 62.21 | 83.13 | 91.72 | 60.18 | - |
| CMSP [14] | IJCV20 | 65.07 | 83.71 | - | 64.50 | - |
| DDAG [35] | ECCV20 | 69.34 | 86.19 | 91.49 | 63.46 | - |
| Hi-CMD [30] | CVPR20 | 70.93 | 86.39 | - | 66.04 | - |
| MACE [63] | TIP20 | 72.37 | 88.40 | 93.59 | 69.09 | - |
| expAT Loss [61] | TIP21 | 66.48 | - | - | 67.31 | - |
| AGW [1] | TPAMI21 | 70.05 | - | - | 66.37 | 50.19 |
| DLS [62] | TMM21 | 71.10 | - | - | 68.10 | - |
| HAT [15] | TIFS21 | 71.83 | 87.16 | 92.16 | 67.56 | - |
| NFS [36] | CVPR21 | **80.54** | 91.96 | 95.07 | 72.10 | - |
| SPOT | This paper | 80.35 | **93.48** | **96.44** | **72.46** | **56.19** |
| Infrared to Visible | | | | | | |
| HCML [12] | AAAI18 | 21.70 | 45.02 | 55.58 | 22.24 | - |
| AlignGAN [31] | ICCV19 | 56.30 | - | - | 53.40 | - |
| DFE [58] | MM19 | 67.99 | 85.56 | 91.41 | 66.70 | - |
| SDL [60] | TCSVT20 | 25.74 | 50.23 | 59.66 | 22.89 | - |
| eBDTR [13] | TIFS20 | 34.21 | 58.74 | 68.64 | 32.49 | - |
| JSIA [29] | AAAI20 | 48.50 | - | - | 49.30 | - |
| DDAG [35] | ECCV20 | 68.06 | 85.15 | 90.31 | 61.80 | - |
| MACE [63] | TIP20 | 72.12 | 88.07 | 93.07 | 68.57 | - |
| expAT Loss [61] | TIP21 | 67.45 | - | - | 66.51 | - |
| HAT [15] | TIFS21 | 70.02 | 86.45 | 91.61 | 66.30 | - |
| NFS [36] | CVPR21 | 77.95 | 90.45 | 93.62 | 69.79 | - |
| SPOT | This paper | **79.37** | **92.79** | **96.01** | **72.26** | **56.06** |

identities from the test set of the SYSU-MM01 dataset, as shown in Fig. 9. For each identity, we select several infrared and visible images from different cameras and extract features from the baseline and SPOT models, respectively. Compared with the baseline, our SPOT exhibits the following two advantages. On the one hand, SPOT decreases the distance within a class, including both intra-modality and cross-modality. In Fig. 9(b), most identities have clusters that are more compact than the baseline model, such as dashed circles. On the other hand, SPOT increases the distance between classes, making some persons with similar appearances separable. For example, the yellow identity and blue identity in the ellipse dotted box are indistinguishable for baseline, while our SPOT does a good job separating these samples. To sum up, our method by jointly estimating the structure-related appearance information and part-level positional relations adaptively reduces the impact of cross-modality variations and noises on visible-infrared ReID.

## V. CONCLUSION

In this paper, we propose a structure-aware positional transformer (SPOT) network to learn semantic-level sharable cross-modality representations for visible-infrared person re-identification. For one thing, to reduce the influence of background noise, we present an attended structure representation (ASR) module to extract structure-related appearance information. ASR models the structure features for each modality and enhances identity-related feature nodes with structure-guided attention to facilitate the learning of shared modality information. For another, a transformer-based part interaction (TPI) module explores plentiful contextual information and structural relations between body parts to improve part-level modality feature representations. For each part feature, it selectively aggregates information from other parts to improve the robustness against misalignments. With the combination of the above two modules, SPOT effectively extracts fine-grained modality-shared features and decreases the impact of cross-modality discrepancy. Extensive ablation experiments demonstrate the validity of proposed methods, and our SPOT is superior to the state-of-the-art methods on two cross-modality datasets. In the future, for cross-modality ReID, the exploration of finer-grained features, such as attribute information, is a research direction worthy of consideration.

## REFERENCES


[1] 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.*, early access, Jan. 26, 2021, doi: 10.1109/TPAMI.2021.3054775.

[2] Z. Zhong, L. Zheng, Z. Luo, S. Li, and Y. Yang, "Learning to adapt invariance in memory for person re-identification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 8, pp. 2723–2738, 2021.

[3] Y. Lin, L. Xie, Y. Wu, C. Yan, and Q. Tian, "Unsupervised person re-identification via softened similarity learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3390–3399.

[4] Z. Zheng, L. Zheng, and Y. Yang, "Pedestrian alignment network for large-scale person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 10, pp. 3037–3045, Oct. 2018.

[5] Y. Jing, C. Si, J. Wang, W. Wang, L. Wang, and T. Tan, "Pose-guided multi-granularity attention network for text-based person search," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, Apr. 2020, pp. 11189–11196.

[6] J. Yang, W.-S. Zheng, Q. Yang, Y.-C. Chen, and Q. Tian, "Spatial-temporal graph convolutional network for video-based person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3289–3299.

[7] G. Chen, J. Lu, M. Yang, and J. Zhou, "Spatial-temporal attention-aware learning for video-based person re-identification," *IEEE Trans. Image Process.*, vol. 28, no. 9, pp. 4192–4205, Sep. 2019.

[8] Y. Liu, W. Zhou, J. Liu, G.-J. Qi, Q. Tian, and H. Li, "An end-to-end foreground-aware network for person re-identification," *IEEE Trans. Image Process.*, vol. 30, pp. 2060–2071, 2021.

[9] H. Zhang, H. Cao, X. Yang, C. Deng, and D. Tao, "Self-training with progressive representation enhancement for unsupervised cross-domain person re-identification," *IEEE Trans. Image Process.*, vol. 30, pp. 5287–5298, 2021.

[10] K. Wang, P. Wang, C. Ding, and D. Tao, "Batch coherence-driven network for part-aware person re-identification," *IEEE Trans. Image Process.*, vol. 30, pp. 3405–3418, 2021.

[11] 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. (ICCV)*, Oct. 2017, pp. 5380–5389.

[12] 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, no. 1, 2018, pp. 7501–7508.

[13] 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.

[14] A. Wu, W.-S. Zheng, S. Gong, and J. Lai, "RGB-IR person re-identification by cross-modality similarity preservation," *Int. J. Comput. Vis.*, vol. 128, no. 6, pp. 1765–1785, Jun. 2020.

[15] 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.


Authorized licensed use limited to: Northeastern University. Downloaded on October 04,2022 at 03:44:41 UTC from IEEE Xplore. Restrictions apply.

[16] 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, 2021.

[17] 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.*, early access, Mar. 2, 2021, doi: 10.1109/TNNLS.2021.3059713.

[18] X. Qian *et al.*, "Long-term cloth-changing person re-identification," in *Proc. Asian Conf. Comput. Vis.*, Nov. 2020, pp. 1–24.

[19] A. Vaswani *et al.*, "Attention is all you need," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 6000–6010.

[20] Z. Cao, G. Hidalgo, T. Simon, S. E. Wei, and Y. Sheikh, "OpenPose: Realtime multi-person 2D pose estimation using part affinity fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 1, pp. 172–186, Jan. 2021.

[21] X. Zhu *et al.*, "Joint prediction and time estimation of COVID-19 developing severe symptoms using chest CT scan," *Med. Image Anal.*, vol. 67, Jan. 2021, Art. no. 101824.

[22] Q. Wu *et al.*, "Discover cross-modality nuances for visible-infrared person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 4330–4339.

[23] N. Pu, W. Chen, Y. Liu, E. M. Bakker, and M. S. Lew, "Dual Gaussian-based variational subspace disentanglement for visible-infrared person re-identification," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 2149–2158.

[24] R. Hu *et al.*, "Multi-band brain network analysis for functional neuroimaging biomarker identification," *IEEE Trans. Med. Imag.*, vol. 40, no. 12, pp. 3843–3855, Dec. 2021.

[25] J. Miao, Y. Wu, P. Liu, Y. Ding, and Y. Yang, "Pose-guided feature alignment for occluded person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 542–551.

[26] H. T. Shen, Y. Zhu, W. Zheng, and X. Zhu, "Half-quadratic minimization for unsupervised feature selection on incomplete data," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 7, pp. 3122–3135, Jul. 2021, doi: 10.1109/TNNLS.2020.3009632.

[27] N. Huang, J. Liu, Q. Zhang, and J. Han, "Exploring modality-shared appearance features and modality-invariant relation features for cross-modality person re-identification," 2021, *arXiv:2104.11539*.

[28] Y. Lu *et al.*, "Cross-modality person re-identification with shared-specific feature transfer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13379–13389.

[29] G.-A. Wang, T. Zhang, Y. Yang, J. Cheng, J. Chang, X. Liang, and Z.-G. Hou, "Cross-modality paired-images generation for RGB-infrared person re-identification," in *Proc. AAAI Conf. Artificial Intell.*, vol. 34, Apr. 2020, pp. 12144–12151.

[30] 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 Recognit. (CVPR)*, Jun. 2020, pp. 10257–10266.

[31] 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. (ICCV)*, Oct. 2019, pp. 3623–3632.

[32] D. Li, X. Wei, X. Hong, and Y. Gong, "Infrared-visible cross-modal person re-identification with an X modality," in *Proc. AAAI*, vol. 34, no. 4, 2020, pp. 4610–4617.

[33] 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 Recognit. (CVPR)*, Jun. 2019, pp. 618–626.

[34] I. Goodfellow *et al.*, "Generative adversarial networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 3, Jun. 2014, pp. 469–477.

[35] J. Luo, "Dynamic dual-attentive aggregation learning for visible-infrared person re-identification," in *Proc. 16th Eur. Conf. Comput. Vis. (ECCV)*, in Lecture Notes in Computer Science, vol. 12362. Glasgow, U.K.: Springer, Aug. 2020, pp. 229–247.

[36] 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. (CVPR)*, Jun. 2021, pp. 587–597.

[37] 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. ECCV*, Sep. 2018, pp. 480–496.

[38] G. Wang, Y. Yuan, X. Chen, J. Li, and X. Zhou, "Learning discriminative features with multiple granularities for person re-identification," in *Proc. 26th ACM Int. Conf. Multimedia*, Oct. 2018, pp. 274–282.

[39] Y. Zhu, J. Ma, C. Yuan, and X. Zhu, "Interpretable learning based dynamic graph convolutional networks for Alzheimer's disease analysis," *Inf. Fusion*, vol. 77, pp. 53–61, Jan. 2022.

[40] X. Zhu, X. Li, S. Zhang, C. Ju, and X. Wu, "Robust joint graph sparse coding for unsupervised spectral feature selection," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 6, pp. 1263–1275, Jun. 2017.

[41] X. Luo, H. K. Luan Duong, and W. Liu, "Person re-identification via pose-aware multi-semantic learning," in *Proc. IEEE Int. Conf. Multimedia Expo. (ICME)*, Jul. 2020, pp. 1–6.

[42] J. Li, S. Zhang, Q. Tian, M. Wang, and W. Gao, "Pose-guided representation learning for person re-identification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 2, pp. 622–635, Feb. 2022.

[43] Y. Suh, J. Wang, S. Tang, T. Mei, and K. M. Lee, "Part-aligned bilinear representations for person re-identification," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 418–437.

[44] L. Zheng, Y. Huang, H. Lu, and Y. Yang, "Pose-invariant embedding for deep person re-identification," *IEEE Trans. Image Process.*, vol. 28, no. 9, pp. 4500–4509, Sep. 2019.

[45] S. Gao, J. Wang, H. Lu, and Z. Liu, "Pose-guided visible part matching for occluded person Reid," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11744–11752.

[46] A. Bhuiyan, Y. Liu, P. Siva, M. Javan, I. B. Ayed, and E. Granger, "Pose guided gated fusion for person re-identification," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 2675–2684.

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

[48] T. Zhang *et al.*, "Spatiotemporal transformer for video-based person re-identification," 2021, *arXiv:2103.16469*.

[49] S. He, H. Luo, P. Wang, F. Wang, H. Li, and W. Jiang, "TransReID: Transformer-based object re-identification," 2021, *arXiv:2102.04378*.

[50] Z. Sun, S. Cao, Y. Yang, and K. M. Kitani, "Rethinking transformer-based set prediction for object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2021, pp. 3611–3620.

[51] X. Li, Y. Hou, P. Wang, Z. Gao, M. Xu, and W. Li, "Trear: Transformer-based RGB-D egocentric action recognition," *IEEE Trans. Cognit. Develop. Syst.*, early access, Jan. 1, 2021, doi: 10.1109/TCDS.2020.3048883.

[52] Y. Wang *et al.*, "End-to-end video instance segmentation with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 8741–8750.

[53] J. Lei Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," 2016, *arXiv:1607.06450*.

[54] 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.

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

[56] T.-Y. Lin *et al.*, "Microsoft coco: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.

[57] (2020). *Mindspore*. [Online]. Available: https://www.mindspore.cn/

[58] Y. Hao, N. Wang, X. Gao, J. Li, and X. Wang, "Dual-alignment feature embedding for cross-modality person re-identification," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 57–65.

[59] P. Zhang, J. Xu, Q. Wu, Y. Huang, and J. Zhang, "Top-push constrained modality-adaptive dictionary learning for cross-modality person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 12, pp. 4554–4566, Dec. 2020.

[60] K. Kansal, A. V. Subramanyam, Z. Wang, and S. Satoh, "SDL: Spectrum-disentangled representation learning for visible-infrared person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 10, pp. 3422–3432, Oct. 2020.

[61] H. Ye, H. Liu, F. Meng, and X. Li, "Bi-directional exponential angular triplet loss for RGB-infrared person re-identification," *IEEE Trans. Image Process.*, vol. 30, pp. 1583–1595, 2021.

[62] Y. Huang, Q. Wu, J. Xu, Y. Zhong, P. Zhang, and Z. Zhang, "Alleviating modality bias training for infrared-visible person re-identification," *IEEE Trans. Multimedia*, early access, Mar. 23, 2021, doi: 10.1109/TMM.2021.3067760.

[63] M. Ye, X. Lan, Q. Leng, and J. Shen, "Cross-modality person re-identification via modality-aware collaborative ensemble learning," *IEEE Trans. Image Process.*, vol. 29, pp. 9387–9399, 2020.

[64] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2921–2929.


Authorized licensed use limited to: Northeastern University. Downloaded on October 04,2022 at 03:44:41 UTC from IEEE Xplore. Restrictions apply.

[65] M. Ye, W. Ruan, B. Du, and M. Z. Shou, "Channel augmented joint learning for visible-infrared recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2021, pp. 13567–13576.

**Cuiqun Chen** received the B.E. degree from Fuyang Normal University, China, in 2016. She is currently pursuing the master's and Ph.D. degrees with the Hefei University of Technology. Her research interests include digital image analysis and processing, computer vision, and pattern recognition.

**Mang Ye** received the B.S. and M.S. degrees from Wuhan University, Wuhan, China, in 2013 and 2016, respectively, and the Ph.D. degree in computer science from Hong Kong Baptist University, Hong Kong, in 2019. He was a Research Scientist with the Inception Institute of Artificial Intelligence, Abu Dhabi, United Arab Emirates. He is currently a Full Professor with Wuhan University. His research interests focus on multimedia retrieval, computer vision, and pattern recognition.

**Meibin Qi** received the B.E. degree in radio technology from Chongqing University in 1991, and the M.E. and Ph.D. degrees in signal and information processing from the Hefei University of Technology in 2001 and 2007, respectively. He is currently a Professor with the School of Computer and Information, Hefei University of Technology. His research interests include pattern recognition, video coding, video surveillance, and the application of DSP technology.

**Jingjing Wu** received the B.E. degree in communication engineering from the Hefei University of Technology, where she is currently pursuing the master's and Ph.D. degrees. Her research interests include digital image analysis and processing, computer vision, and pattern recognition.

**Jianguo Jiang** received the B.E. degree in radio technology and the M.E. degree in signal and information processing from the Hefei University of Technology in 1982 and 1989, respectively. He is currently a Professor with the School of Computer and Information, Hefei University of Technology. His research interests include digital image analysis and processing, distributed intelligent systems, and the application of DSP technology.

**Chia-Wen Lin** (Fellow, IEEE) received the Ph.D. degree in electrical engineering from National Tsing Hua University (NTHU), Hsinchu, Taiwan, in 2000.

He was with the Department of Computer Science and Information Engineering, National Chung Cheng University, Taiwan, from 2000 to 2007. Prior to joining academia, he worked with Information and Communications Research Laboratories, Industrial Technology Research Institute, Hsinchu, from 1992 to 2000. He is currently a Professor with the Department of Electrical Engineering and the Institute of Communications Engineering, NTHU. He is also the Deputy Director of the AI Research Center, NTHU. His research interests include image and video processing, computer vision, and video networking. His articles received the Best Paper Award of IEEE VCIP in 2015, the Top 10% Paper Awards of IEEE MMSP in 2013, and the Young Investigator Award of VCIP in 2005. He received the Outstanding Electrical Professor Award presented by the Chinese Institute of Electrical Engineering in 2019 and the Young Investigator Award presented by the Ministry of Science and Technology, Taiwan, in 2006. He served as a Steering Committee Member for IEEE TRANSACTIONS ON MULTIMEDIA from 2014 to 2015 and the Chair for the Multimedia Systems and Applications Technical Committee of IEEE CASS from 2013 to 2015. He is also the Chair of the Steering Committee of IEEE ICME. He served as the President for the Chinese Image Processing and Pattern Recognition Association, Taiwan, from 2019 to 2020; the Technical Program Co-Chair for IEEE ICIP in 2019 and IEEE ICME in 2010; and the General Co-Chair for IEEE VCIP in 2018. He served as an Associate Editor for IEEE TRANSACTIONS ON IMAGE PROCESSING (IEEE TIP), IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY (IEEE TCSVT), IEEE TRANSACTIONS ON MULTIMEDIA (IEEE TMM), and IEEE MULTIMEDIA. He has been serving on the IEEE Circuits and Systems Society (CASS) Fellow Evaluating Committee since 2021. He served as an IEEE CASS Distinguished Lecturer from 2018 to 2019.

Authorized licensed use limited to: Northeastern University. Downloaded on October 04,2022 at 03:44:41 UTC from IEEE Xplore. Restrictions apply.