Robust Visible-Infrared Person Re-Identification Based on Polymorphic Mask and Wavelet Graph Convolutional Network

Rui Sun, *Member, IEEE*, Long Chen, Lei Zhang, Ruirui Xie, and Jun Gao

***Abstract*— When deploying re-identification (ReID) models in the field of public safety, understanding the robustness of models to various types of corrupted images is crucial. Unfortunately, in the real world, images are always contaminated (e.g., noise, blur, and weather changes), which is ignored by existing visible-infrared person re-identification (VI-ReID) models. The performance of existing models tested in corrupted scenes is severely degraded. Therefore, learning corruption-invariant representations for corrupted images in VI-ReID is valuable and deserves further investigation. We design a polymorphic masked wavelet graph convolutional network for VI-ReID under corrupted scenes. Firstly, a cross-modality data augmentation algorithm is designed to construct a mixed image set that merges multi-modality attributes to improve robustness against interference. Secondly, a dual-branch network consisting of a global branch and a graph structure branch is designed. The global branch extracts overall information. While the graph structure branch is a wavelet-based graph convolutional module that utilizes the robustness of human structural information to corruptions and modalities, it can filter noise and extract discriminative features specifically targeted for cross-modality scenes. Finally, the global branch and the graph structure branch are integrated, and modality consistency loss is designed to match the branches with hetero-center triplet loss. Experiments show that our method can effectively alleviate degradation problems under corrupted environments such as noise, blur, digitization, and weather changes, and achieve state-of-the-art on corrupted datasets. Besides, it still maintains good performance on clean datasets, facilitating the reliable deployment of VI-ReID in real-world scenarios.**

***Index Terms*— Visible-infrared person re-identification, corruption robustness, cross-modality data augmentation, wavelet graph convolutional network.**

I. Introduction

PERSION re-identification (ReID) is a cross-camera image retrieval task that aims to search given personnel from an image database collected by non-overlapping cameras [1], but its nighttime surveillance capability is limited. Therefore, visible-infrared person re-identification (VI-ReID) appeared, which retrieves a given query person through cross-checking between visible and infrared cameras. To alleviate the huge modality gap and complex intra-class variation, many methods have been proposed. These methods can be divided into feature representation learning [2], [3], [4], metric learning [5], [6] and modality translation learning [7], [8], [9], [10], [11].

Although VI-ReID has achieved promising performance, its success largely depends on clean and complete image data. Actually, the real world is full of various types of corruption, such as noise, blur, distortion, weather changes, etc., as shown in Fig. 1, so obtaining ideal and undamaged samples is extremely difficult, even impossible. Existing models are easily confused by corrupted images and difficult to extract fine-grained instance information of individuals [19].

In the corrupted scenes, the generalization ability of the model to the input is greatly challenged, and the fuzzy identity information of corrupted images leads to worse modality gap and intra-class discrepancy. Models trained on traditional closed-world datasets fail to generalize to these unseen corrupted inputs. How to maintain stable recognition performance in clean and corrupted scenes is a meaningful problem. Huang et al. [39] designed a degradation feature decoupling framework based on generative adversarial network from the perspective of image recovery, which achieved good results in single-modality corrupted ReID, but the quality of generated images determined its final performance. In VI-ReID task, Chen et al. [19] proposed a corruption invariant learning (CIL) baseline, which utilized soft random erasing and self-patch mixing to address the corruption robustness of model. Josi et al. [52] improved CIL by introducing a multimodal data augmentetion (ML-MDA) to simulate real-world data possibly, thereby enhancing the generalization performance of VI-ReID model. Besides, they proposed a new evaluation strategy [53] in corruption scenes, taking into account the impact of different camera placement situations on model performance. However, corruption robustness has not been deeply studied in the VI-ReID task, these works only

Manuscript received 18 July 2023; revised 22 November 2023 and 8 January 2024; accepted 10 January 2024. Date of publication 15 January 2024; date of current version 26 January 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 62302142 and Grant 61876057, in part by the China Postdoctoral Science Foundation under Grant 2022M720981, in part by the Anhui Province Natural Science Foundation under Grant 2208085MF158, and in part by the Key Research Plan of Anhui Province-Strengthening Police with Science and Technology under Grant 202004d07020012. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Jianjiang Feng. *(Corresponding author: Rui Sun.)*

Rui Sun is with the Key Laboratory of Knowledge Engineering with Big Data (Ministry of Education), School of Computer and Information, Anhui Province Key Laboratory of Industry Safety and Emergency Technology, Hefei University of Technology, Hefei 230009, China (e-mail: sunrui@hfut.edu.cn).

Long Chen, Lei Zhang, Ruirui Xie, and Jun Gao are with the School of Computer and Information, Hefei University of Technology, Hefei 230009, China.

Digital Object Identifier 10.1109/TIFS.2024.3354377

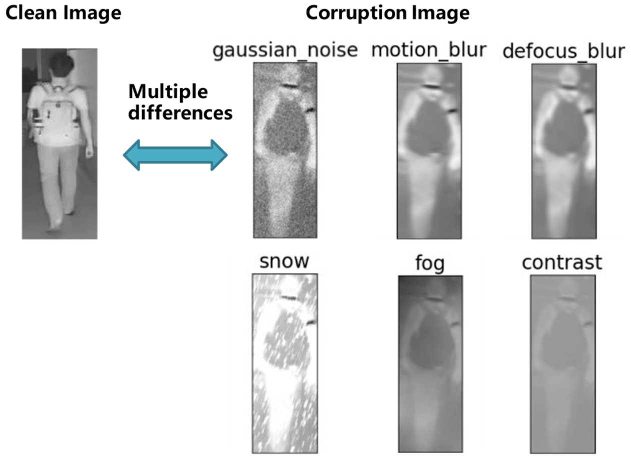


Fig. 1. Illustration of the multiple challenges caused by the problem of corruption in the VI-ReID. Apart from the modality discrepancy caused by different imaging mechanisms, there are also attacks from various corruptions in the open world, which make modality features more difficult to identify.

operate on data augmentation, without considering the physical mechanism behind noise generation. Moreover, the VI-ReID task is special in that it can take advantage of the robustness of human structural information to corruptions, whereas previous methods have not considered this perspective. To improve the corruption robustness in visible-infrared scenes, we propose a corruption-robust VI-ReID framework called Polymorphic Masked Wavelet Graph Convolutional Network (PMWGCN).

To enhance the generalization of the model to corrupted input and mitigate modality discrepancy, we introduce a Polymorphic Masked Data Augmentation algorithm (PMDA). The PMDA consists of a new data processing pipeline that uses a structurally complex fractal image set and a randomly selected cross-modality set. The convoluted fractals enable augmented images to have rich natural structural complexity and simulate real-world scenes. By introducing random perturbations with the additional image set, the PMDA algorithm allows for the utilization of complementary correlation knowledge between modalities, dynamically balancing the importance of each modality in the final prediction. Moreover, a modality-consistency loss is designed to reduce the perturbation sensitivity on cross-modality data.

Then we propose a Graph Branch module which is composed of graph construction and Wavelet Graph Convolutional Network (WGCN). We utilize human body components to construct the graph structure from interpretable frequency domain and remove the non-stationary high-frequency noise in the features as much as possible, thereby alleviating the performance degradation problem in corrupted scenes. Specifically, we take a semantic parsing model to extract patches from different regions of the person's body and generate the graph structure using a normalized graph construction method. Then, we employ WGCN to model, filter and aggregate high-level semantic relationships among local patches of individuals in cross-modality tasks. This reduces information redundancy from the source, enhancing the robustness of cross-modality recognition. Through these specially designed multimodal data augmentation algorithm and network structures, the features extracted by PMWGCN have strong modality reasoning ability and are robust to various forms of corruption.

Our main contributions are summarized below:

1. We discuss a new problem for VI-ReID, called corruption robustness. It is very challenging to overcome the huge modality discrepancy and ensure network has robust corruption tolerance in cross-modality scenes, which is different from only considering corruption problem in a single modality. To address this issue, Polymorphic Masked Wavelet Graph Convolutional Network (PMWGCN) method is proposed to effectively disentangle content and degradation features in cross-modality images.

2. A polymorphic masked data augmentation (PMDA) algorithm is proposed to enhance the corruption robustness of VI-ReID model. Fractal images and paired modality images are embedded into the original images. A new data pipeline is used to process cross-modality samples, and a modality consistency loss is introduced.

3. To consider corruptions from the frequency domain, we propose a new wavelet graph convolutional network (WGCN). WGCN models the semantic relationships between local patches of individuals and promotes filtered high-order semantic correlation. It fundamentally weakens information masking at the pixel level, providing the possibility for our model to generalize to unseen corruptions and enhancing the robustness of VI-ReID.

II. Related Work

In this section, we briefly review existing approaches in visible-infrared person re-identification, corruption robustness and graph convolutional network.

A. Visible-Infrared Person Re-Identification

To alleviate cross-modality discrepancy, many VI-ReID methods have been proposed in recent years. According to the differences in mitigating discrepancy, existing methods can be divided into three categories. In terms of feature representation learning methods, Wu et al. [2] first constructed the VI-ReID benchmark SYSU-MM01 and proposed a deep zero-padding method that can be used for cross-modality training. Wu et al. [46] designed a joint modality and pattern alignment network to mine the cross-modality nuances from modality-independent feature maps. Ye et al. [4] discovered the noise problem in VI-ReID and designed a dynamic tri-level relation mining framework to learn partial discriminative features. In terms of metric learning methods, Ye et al. [5] introduced bi-directional dual-constrained top-ranking loss to learn discriminative feature representations. Liu et al. [6] explored the problem of parameter sharing and designed the hetero-center triplet loss to constrain different category centers from the same and cross modalities. Regarding modality conversion learning methods, some work [7], [8], [9] applied generative adversarial networks to learn modal-shared representations to alleviate modality differences. Others [10] introduced an auxiliary modality to bridge the cross-modality gap through three-modality learning. Lu et al. [11] utilized a channel-interactive generator to generate confused

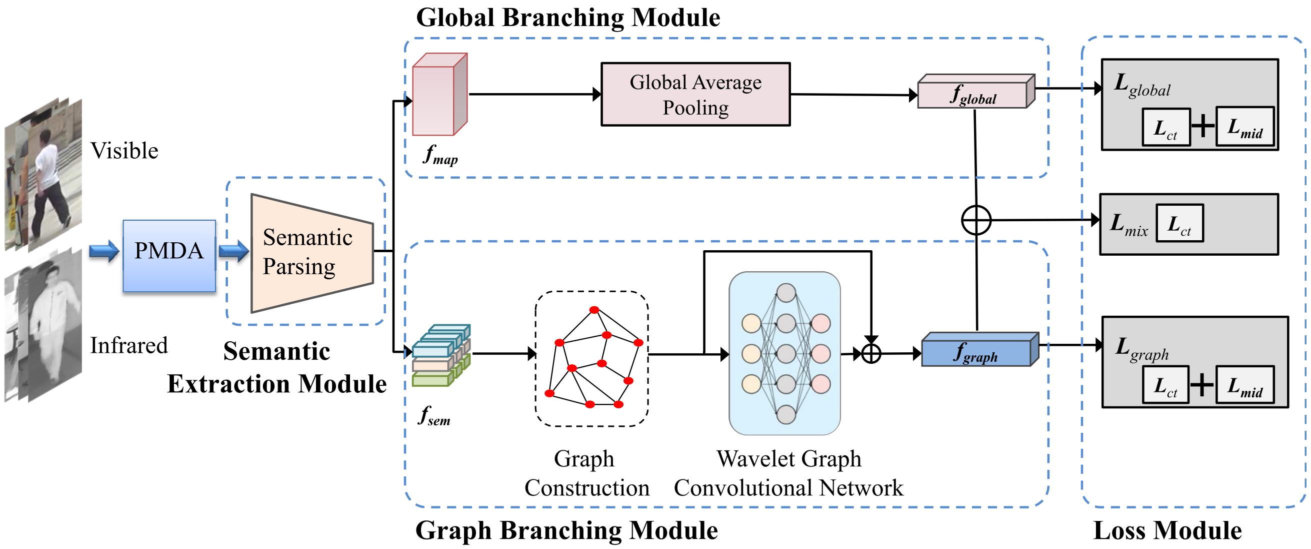


Fig. 2. The overall architecture of our proposed Polymorphic Masked Wavelet Graph Convolution Network (PMWGCN). It consists of four components: the Polymorphic Masked Data Augmentation (PMDA) algorithm, the semantic extraction module, the graph branching module, and the loss module. Given an image, we first preprocess it using PMDA and use the semantic extraction module to obtain the part features. Then, the graph branching module is used to further aggregate higher-order relational features. Both hetero-center loss and Jensen-Shannon divergence based cross-modality consistency loss are used for robust feature representation learning.

modalities for extracting modality-invariant representations. Wan et al. [15] tried to tackle the task from a pre-training perspective, and proposed a modality-aware multiple granularity learning model.

B. Corruption Robustness

The study of corruption robustness has been developing in computer vision for several years. Related research shows that current deep neural networks are susceptible to a wide variety of image corruptions, such as noise, blur, distortion, or a combination of them. To improve the robustness of corruption, various data augmentation [12], [13], [14] strategies have been proposed. AugMix [13] mixed multiple augmented images by designing algorithms and obtains significant improvements on RGB data. Rusak et al. [20] successfully improved the generalization ability and robustness of the neural network by introducing data augmentation techniques such as random distortion and color perturbation in the training process. Unlike adversarial training, Hendrycks et al. [14] argued for Pareto improvement of existing security measures and proposed the PixMix data augmentation algorithm. Some other researchers argued that pre-training [16], [17], [18] can improve robustness. Both Hendrycks et al. [16] and Si et al. [21] discussed the robustness of pre-training to out-of-distribution samples, and they find that a suitable training strategy can improve security. Focusing on ReID, Chen et al. [19] constructed corrupted ReID datasets and proposed a data augmentation strategy consisting of a mixture of self-patch mixing and soft random erasing to improve the system performance under corrupted data. However, these methods are extremely general and badly adapted in VI-ReID.

C. Graph Convolutional Network

The graph convolutional network (GCN) [22] model usually refers to a spectral-based graph neural network, which defines the neighborhood aggregation of nodes based on the graph Laplace decomposition. Recently, many studies have successfully applied GCN to ReID. AAGCN [23] exploited the low-pass property of GCN to reduce intra-class variance. Zhang et al. [24] proposed a local graph attention network to learn intra and inter-local relations. CTL [25] extracted multi-granularity semantic local features from human body as graph nodes and designs cross-scale graph convolution to capture hierarchical spatio-temporal dependence and structural information. Because CTL captured complex spatio-temporal information based on GCN, SOTA performance was achieved in single-modality ReID. However, only a small number of researchers have explored GCN applications to the visible-infrared domain. GLGCN [26] proposed global-local graph convolutional networks to model the underlying relationships of each modality body part. This is the initial attempt to apply GCN to VI-ReID. GCN facilitates learning higher-order visual attributes and has better descriptive, explanatory, and robustness. But, it has received little attention in VI-ReID.

III. METHOD

In this section, we introduce the proposed Polymorphic Masked Wavelet Graph Convolutional Network (PMWGCN) in detail, as illustrated in Fig 2. We first briefly introduce the polymorphic masked data augmentation algorithm (PMDA) in Section III-A. In Section III-B, we describe the semantic extraction process in detail. Then, in Section III-C, we describe the various components of the graph branching module and the process of embedding the wavelet transform in the graph convolutional network in detail. Finally, we introduce the proposed multi-branch loss learning strategy in Section III-D.

A. Polymorphic Masked Data Augmentation

In this subsection, we will introduce how PMDA enhances the robustness of the model to unseen corruptions and

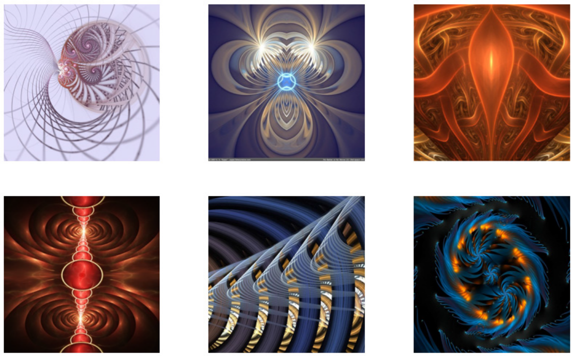


Fig. 3. Example images from the fractal image set. The fractal image set is usually generated by iterative function systems, producing images of various shapes.

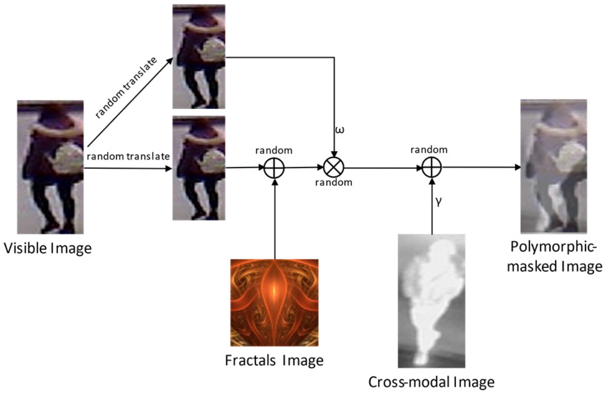


Fig. 4. The operation pipeline of our proposed PMDA algorithm, which aims to mix natural and modality attributes.

mitigates modality gap in VI-ReID. The PMDA algorithm consists of two components: the compensated image set and the algorithm pipeline for expanding images of the clean training set.

1) The Compensated Image Set: The compensated image set includes fractals image set and cross-modality image set. According to [27], the fractal image set is usually generated by iterative function systems, producing images of various shapes. As illustrated in Fig. 3, these images have "non-random" attributes that humans may use, namely the structural attributes of contours (direction, length, curvature) and contour connections (type and angle) drawn from natural scenes. Existing study [14] has shown that combining these fractals images with training samples can greatly expand the natural elements contained in images, thereby improving the security of neural networks. Therefore, to improve the corruption robustness of model, we selected and downloaded 14,230 fractal images from DeviantArt, and used them in the algorithm to mix natural attributes.

However, it is insufficient to consider natural attributes to improve the corruption robustness. Corruptions result in worse modality gap in VI-ReID, the network also needs to capture modality-shared attributes across different modalities. We introduce a cross-modality image set during data augmentation process. Unlike the fixed fractals image set, it consists of images from another modality randomly selected from the same training batch, with same identity as the image being processed. By mixing cross-modality images, model exchanges cross-modality information within the same identity. Since the Fractals image set does not belong to any specific pedestrian identity and the cross-modality image set does not affect identity discrimination, the augmented images are still classified into their original categories during network training, just like standard data augmentation.

2) Algorithm Pipeline: The operation pipeline of the PMDA algorithm is shown in Fig. 4. First, assuming that the original image is X_{orig}, we perform a standard data augmentation operation (including rotate, solarize, posterize, etc.) on it to obtain X_{hide}, and randomly select one from X_{orig} and X_{hide} as the standard image X_{stad}. Then, we perform a random number of corruption augmentation operations on the image, with a maximum value of k. Each operation is implemented by mixing with the newly augmented image X_{temp} or an image in the fractals image set. The mixing mode is divided into additive mixing and multiplicative mixing. Additive mixing directly superimposes pixel values on the channel level, and the execution of multiplicative mixing is similar to geometric mean, where alpha is the hyperparameter ω that adjusts the weight of multiplicative mixing. After random corruption augmentation, we select a heterogeneous image from the cross-modality image set for additive mixing, with a weight of γ, to obtain the PMDA data augmentation result X_{pmda}. Cause images come from the same identity, cross-modality mixing helps model discover shared features and choose distinguishing features to rely on. Standard data augmentation usually operates in the RGB domain, thus this algorithm only processes RGB images in the dataset. In the preprocessing operation, PMDA integrates natural and modality attributes into the training set, increasing generalization ability to unseen corruptions of VI-ReID model. The algorithm flowchart is shown in Algorithm 1.

Algorithm 1 Polymorphic Masked Data Augmentation

```
Input: $X_{orig}$, $X_{frac}$, $X_{modal}$;
aug_op ← random choice(rotate, solarize ...);
$X_{hide}$ ← augment(aug_op($X_{orig}$));
$X_{stad}$ ← random choice ($X_{hide}$, $X_{orig}$);
$X_{mult}$ ← $X_{stad}$;
for k ← {0...k} do
    aug_op ← random choice (rotate, solarize ...);
    $X_{hide}$ ← augment(aug_op($X_{stad}$));
    $X_{mix}$ ← random choice ($X_{hide}$, $X_{frac}$);
mix_op ← random choice(add, multiply);
$X_{mult}$ ← mix(mix_op($X_{stad}$, $X_{mix,}\omega$));
end for
$X_{pmda}$ ← add ($X_{mult}$, $X_{modal,}\gamma$);
Output: $X_{pmda}$
```

B. Semantic Extraction Module

The purpose of the semantic extraction module is to extract modality independent human body part features from the pixel level and exclude background interference through semantic analysis. The background noise and body parts are effectively

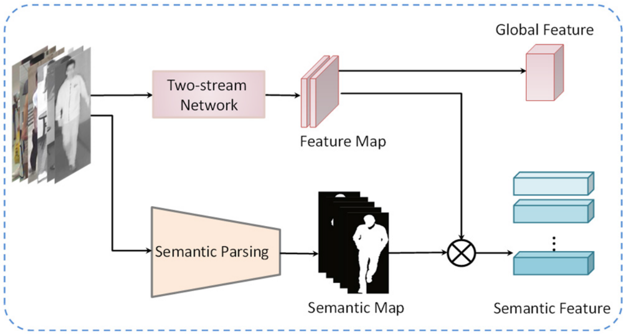


Fig. 5. The processing of the semantic extraction module. It generates semantic features of each region on the human body through semantic parsing network.

disentangled. In addition, the part features can achieve local alignment, further promote the correlation on the overall structure, and reduce the impact of corruption on the human subject. The semantic extraction module is shown in Fig. 5.

We use the self-correcting for human parsing (SCHP) [28] model pre-trained on the look into person (LIP) dataset [29] as our semantic extraction network. It predicts four body semantic parts, namely head, arms, trunk and lower body, and generates the corresponding semantic parse maps m_k (k=1,2,3,4). The semantic parse maps are pixel-level, and each pixel point represents the confidence level of the corresponding position in the original image.

In addition, the two-stream feature extraction network is a commonly used global feature extraction method in visible-infrared cross-modality person re-identification. Given a pre-processed visible or infrared image X_{pmda}, we obtain its global feature graph f_{map} using a two-stream feature extraction network. Then, by computing a matrix multiplication operation for each channel on f_{map} and m_k, we can obtain the semantic features corresponding to each body part of the person, which is used to construct the graph, named as f_{sem}^k. For the global branch, the global feature f_{global} is obtained by simply subjecting f_{map} to a global average pooling operation. the equation is formulated as:

$$f_{sem}^k = g(f_{map} \otimes m_k), \quad k = 1, 2, 3, 4 \tag{1}$$

$$f_{global} = g(f_{map}) \tag{2}$$

where g denotes the global average pooling, the token $\otimes$ denotes the matrix multiplication operation, and k denotes the number of semantic component blocks. Thus, we can obtain a semantic feature f_{sem}^k and transmit it to the graph branching module later.

C. The Graph Branching Module

Human body components are less affected by the modality gap, and the hidden semantic information is more robust to corruptions. Besides, most noises in nature are non-stationary signals. Wavelet transform is a widely used denoising method in the field of signal processing with rigorous physical background and interpretability, which can deal with non-stationary processes. But it loses the deserved attention in corruption-robust VI-ReID. To leverage these advantages, we design two components in this module, graph construction and wavelet graph convolution network (WGCN). The graph construction aims to construct normalized graphs using semantic features generated by the semantic extraction module. Subsequently, we introduce a graph convolution network embedded with wavelet transform, which leverages wavelet transform to disentangle corruption noise and effective content in the frequency domain. Then it filters out the degradation features and capture higher-order feature patterns by wavelet graph convolution operator.

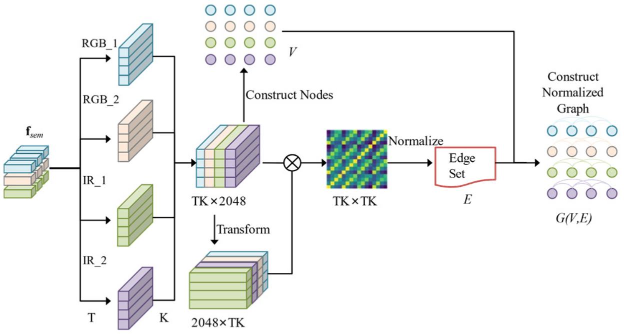


Fig. 6. Illustration of the normalized graph construction. Semantic feature patches are generated by the semantic extraction module, where different colors represent different pedestrians, and both visible and infrared features are included here. Assuming a batch contains T images, each image is segmented into K semantic feature blocks, and a total of T×K patches are obtained. The pairwise relationships are represented by computing the affinity matrix and normalizing the first n relationship values for each patch.

1) Graph Construction: Semantic component blocks have the advantage of feature detail and positional alignment, which facilitate the interaction of information between different modalities in a cross-modal scene. Moreover, semantic component blocks further narrow the influence of corruption factors and are useful in mitigating problems caused by noise, ambiguity, and distortion. Therefore, we construct the graph using $\mathbf{f}_{sem}^k$. As shown in Fig. 6, we first simplify $\mathbf{f}_{sem}^k$ to $\mathbf{f}_k = \{\mathbf{f}_1, \mathbf{f}_2 \ldots \mathbf{f}_K\}$. Let $G(V, E)$ denote a graph structure with K nodes, where $V = \{v_i\}_{i=1}^K$ is the vertex set of nodes and the edge set $(v_i, v_j) \in E$ is used to represent the correspondence of feature blocks. The k semantic vectors $\mathbf{f}_k$ are used as the vertex set of each graph structure.

The edge set E is represented by the affine adjacency matrix $\mathbf{A}$. Each element of the matrix represents the relationship between two component blocks. The two graph nodes v_i and v_j are known to represent features $\mathbf{f}_i$ and $\mathbf{f}_j$ respectively. The pairwise relationship between each two component blocks is represented by their inner product:

$$\mathbf{r}_{i,j} = \mathbf{f}_i^{\mathrm{T}} \cdot \mathbf{f}_j \tag{3}$$

The adjacency matrix is obtained by normalizing $\mathbf{A}$ and formulating it as follows:

$$\mathbf{A}_{i,j} = \frac{\mathbf{r}_{i,j}}{\sum_{j=1}^{K} \left|\mathbf{r}_{i,j}\right|} \tag{4}$$

We need to normalize $\mathbf{A}$. For each element in $\mathbf{A}$, we keep only the first n relevant elements to obtain the weighted canonical adjacency matrix $\tilde{\mathbf{A}}$. The specific formula is as follows:

$$\widetilde{\mathbf{A}}_{i,j} = \begin{cases} \widetilde{\mathbf{r}}_{i,j} & i \neq j \, and \, \widetilde{\mathbf{r}}_{i,j} \in CUT(\widetilde{\mathbf{A}}_{i,:,k}), \\ 0 & otherwise, \end{cases} \tag{5}$$

$\tilde{\mathbf{r}}_{i,j}$ denotes the normalized association score from node i to node j. $CUT(\widetilde{\mathbf{A}}_{i,:,k})$ means that for the i-th row of matrix $\mathbf{A}$, only the first n element values associated with node i are retained. In this way, the number of edges on the graph is reduced from K^2 to nK. Finally, the pairwise relationship E between nodes is represented by the adjacency matrix $\tilde{\mathbf{A}}$, and the graph is generated from V and E. In a batch, visible and infrared images are always present in pairs, so the final graph structure combines visible features and infrared features. Through the above operations, we construct the body structure graph that is modality independent and robust to corruptions.

2) Wavelet Graph Convolution Network: Wavelet transform can be applied in graph convolutional network to filter non-stationary degradation noise. Different from normal graph Fourier transform, graph wavelet matrix can be obtained via a fast algorithm without requiring Laplacian matrix eigendecomposition with high computational cost. Moreover, graph wavelets are sparse and localized in vertex domain, offering high efficiency and good interpretability for graph convolution.

In traditional GCN, the Laplacian matrix $\mathcal{L}$ of a graph is defined as $\mathcal{L} = \mathbf{D} - \tilde{\mathbf{A}}$, where $\mathbf{D}$ is the pairwise angle matrix and $\mathbf{D}_{i,j} = \sum_j \tilde{\mathbf{A}}_{i,j}$. A normalized Laplace operator matrix is defined as $\mathbf{L} = \mathbf{I}_k - \mathbf{D}^{-1/2}\widetilde{\mathbf{A}}\mathbf{D}^{-1/2}$, where $\mathbf{I}_k \in \mathbb{R}^{K \times K}$ is the identity matrix, and since $\mathcal{L}$ is a real symmetric matrix, it has a set of standard orthogonal eigenvectors $\mathbf{U} = (\mathbf{u}_1, \mathbf{u}_1, \ldots, \mathbf{u}_K)$, called Laplace eigenvectors. The non-negative eigenvalues $\{\lambda_l\}_{l=1}^{K}$ corresponding to these eigenvectors are called the frequencies of the graph.

Similar to the graph Fourier transform, the graph wavelet transform maps the graph signal from the vertex domain to the frequency domain. The graph wavelet transform uses a set of wavelets $\psi_s = (\psi_{s1}, \psi_{s2}, \ldots, \psi_{sK})$ as bases, where each wavelet basis ψ_{si} corresponds to the signal diffused from node i on the graph, and s is a scale parameter that constrains the extent of the domain. The formulation is as follows:

$$\psi_s = \mathbf{U}\mathbf{G}_s\mathbf{U}^{\mathrm{T}} \tag{6}$$

where $\mathbf{U}$ is a Laplacian vector, $\mathbf{G}_s$ is a scale matrix, and $G_s = diag(g(s\lambda_1), \ldots, g(s\lambda_K))$, $g(s\lambda_i) = e^{\lambda_i s}$.

The graph wavelet transform of the signal $\mathbf{f}$ on the graph with the graph wavelet as the base is defined as $\widehat{\mathbf{f}} = \psi_s^{-1}\mathbf{f}$. Where ψ_s^{-1} can be obtained by replacing $g(-s\lambda_i)$ in ψ_s with $g(s\lambda_i)$. Correspondingly, the graph wavelet inverse transform is defined as $\mathbf{f} = \psi_s\widehat{\mathbf{f}}$. By graph wavelet transform, we can transform the constructed body structure graph to the spectral domain. Since the valid content and degradation noise are often distributed in different frequency bands, we can build a filter operator to disentangle them.

According to the time-domain convolution theorem, the convolution of the graph wavelet operator can be expressed as the product of the corresponding spectral domains, denoted as $*_g$. Using $\mathbf{y}$ to represent the convolution kernel, $*_g$ is defined as:

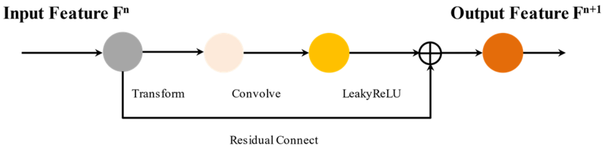


Fig. 7. The illustration of a wavelet graph convolution layer. The process is as in Eq. 9, where the transformation and convolution can be obtained from Eq. 6 and Eq. 7.

$$\mathbf{f} *_g \mathbf{y} = \psi_s((\psi_s^{-1}\mathbf{y}) \odot (\psi_s^{-1}\mathbf{f})) \tag{7}$$

where $\odot$ denotes the Hadamard product at the elemental level.

After defining the wavelet graph convolution operation, we can construct an M-layer WGCN network to disentangle and filter degradation features, thereby capturing higher-order semantic content. Each layer of the WGCN network consists of feature transformation and graph convolution operations, as shown in Fig. 7. For the n-th layer (1≤n≤M), the feature transformation is defined as:

$$\hat{\mathbf{F}}^n = \mathbf{F}^n\mathbf{W} \tag{8}$$

where $\mathbf{F}^n \in \mathbb{R}^{K \times c}$ is the hidden feature of all feature blocks in the nth layer and c refers to the feature dimension; $\mathbf{F}^0 \in \mathbb{R}^{K \times c}$ is the original feature output from the graph construction stage; and $\mathbf{W}$ is the parameter matrix to be learned. For the $(\text{n}+1)_{th}$ layer, we define graph convolution as follows:

$$\mathbf{F}^{n+1} = h(\psi_s\mathbf{X}^n\psi_s^{-1}\hat{\mathbf{F}}^n) \tag{9}$$

where ψ_s is the wavelet bases and ψ_s^{-1} is the graph wavelet transform matrix with scale s. The inverse transform projects the signal from the vertex domain into the frequency domain. The input is the nth layer feature $\mathbf{F}^n \in \mathbb{R}^{K \times c}$. $\mathbf{X}^n$ is the diagonal filter matrix learned in the spectral domain. h is a nonlinear activation function (LeakyReLU) with negative input slope $\alpha = 0.1$. Finally, residual connections are used to facilitate network optimization, as shown in Fig. 7. A final wavelet graph convolution layer is defined as follows:

$$\mathbf{F}^{n+1} := \mathbf{F}^{n+1} + \mathbf{F}^n, \quad 1 \leq n \leq M-1 \tag{10}$$

After M layers of wavelet graph convolution, the module models semantic relationship among human local feature parches and reduces the information redundancy from the source (frequency domain). The output for each patch group is $\mathbf{F}^M \in \mathbb{R}^{K \times c}$. Finally, by applying average pooling to $\mathbf{F}^M$, we obtain the feature $f_{graph} \in \mathbb{R}^c$ for the graph branching module. By concatenating the f_{global} and f_{graph} channel dimensions, we obtain the final feature f_{mix}.

D. Multi-Branch Loss Learning Strategy

Our method combines representation learning and metric learning, combining modality consistency loss L_{mid} and hetero-central triplet loss L_{ct} to train the model.

Given an image x_i with label y_i, a softmax function is used to encode the probability of x_i being recognized as class y_i, denoted as $p_{\text{ori}}(y||x_{ori})$. Since any input image is randomly augmented twice in the PMDA operation, an image triple including x_{ori}, x_{pmda1} and x_{pmda2} is generated. The predicted probability distributions are $p_{\mathrm{ori}}(y||x_{ori})$, $p_{\mathrm{pmda1}}(y||x_{pmda1})$ and $p_{\mathrm{pmda2}}(y||x_{pmda2})$, respectively. The semantic content in augmented images is retained and no other identity information is introduced, so we hope the model to learn the consistency between original and augmented images. We leverage the improved Jensen-Shannon divergence [13] to constrain the similarity between the posterior distributions of original and augmented samples to define L_{mid}:

$$\begin{aligned} L_{mid} &= \mathrm{JS}\left(p_{ori}; p_{pmda1}; p_{pmda2}\right) \\ &= \frac{1}{3}(KL[p_{ori}||M] + KL[p_{pmda1}||M] \\ &\quad + KL[p_{pmda2}||M]) \end{aligned} \tag{11}$$

where the mean distribution M is calculated as: $M = (p_{ori} + p_{pmda1} + p_{pmda2})/3$. We can understand this loss by imagining a sample from one of the above three distributions, where JS divergence can be understood as a measure of the average identity information that this sample reveals about the distribution it was sampled from. This loss motivates the model to be stable, consistent, and insensitive across a variety of input ranges.

Hetero-center triplet loss L_{ct} was first proposed in [6]. The authors introduced the concept of feature centers for guiding the network to push the feature centers of different modalities of the same pedestrian together and to move the feature centers of different people away from each other. The definition is as follows:

$$\begin{aligned} L_{ct}(\mathbf{C}) = &\sum_{i=1}^{N}\left[\rho + \left\|\mathbf{c}_v^i - \mathbf{c}_t^i\right\|_2 - \min_{\substack{n\in\{v,t\} \\ j\neq i}}\left\|\mathbf{c}_v^i - \mathbf{c}_n^j\right\|_2\right]_+ \\ &+ \sum_{i=1}^{N}\left[\rho + \left\|\mathbf{c}_t^i - \mathbf{c}_v^i\right\|_2 - \min_{\substack{n\in\{v,t\} \\ j\neq i}}\left\|\mathbf{c}_t^i - \mathbf{c}_n^j\right\|_2\right]_+ \end{aligned} \tag{12}$$

where N denotes N pedestrian identities and $\mathbf{C}$ is the batch of feature centers containing visible feature centers $\mathbf{c}_v^i|i = 1, \ldots, N$ and infrared feature centers $\mathbf{c}_t^i|\{i = 1, \ldots, N\}$, and $||\cdot||_2$ denotes the Euclidean distance.

For each branch, we combine L_{mid} and L_{ct} to denote the total branch loss function. Introducing the weight parameter λ, the loss functions L_{global} and L_{graph} of the graph branch module and the global branch module are calculated as follows:

$$L_{global} = L_{graph} = L_{mid} + \lambda L_{ct} \tag{13}$$

Specially, by concatenating f_{global} and f_{graph} in channel dimensions and then calculating the hetero-center triplet loss, we can get the loss L_{mix} of the mixed features.

$$L_{mix} = L_{ct} \tag{14}$$

Finally, the total loss L_{total} is represented by the sum of the three branch loss functions, as shown in Eq. (15).

$$L_{total} = L_{global} + L_{graph} + L_{mix} \tag{15}$$

IV. Experiments

In this section, we use the standard dataset RegDB, SYSU-MM01 [2] and the dataset RegDB, SYSU-MM01 after being randomly corrupted to evaluate the effectiveness of our proposed method. In addition, extensive ablation experiments were performed to verify the role of each component of the network and the effect of various parameters.

A. Dataset and Evaluation Protocol

1) SYSU-MM01: It is the definitive dataset for cross-modal person re-identification. The SYSU-MM01 dataset is a large-scale dataset collected by four visible and two infrared cameras, which contains a total of 287628 visible and 15792 infrared images of 491 different pedestrians. Its training set contains 22258 visible and 11909 infrared images of 395 pedestrians, and the test set contains images of other 96 pedestrians. The dataset provides two different search settings named All-Search mode and Indoor-Search mode. The All-Search mode tests using both indoor and outdoor images and the Indoor-Search mode tests using only indoor images. Following the validation protocol in [2], 10 random experiments were performed under the single-shot setting and the average value was taken as the final result.

2) RegDB: It is a small dataset captured by a visible-thermal dual-mode camera, using both visible and thermal cameras. A total of 412 different person identities are captured in the dataset and each identity has 10 pose aligned visible and thermal images. According to the evaluation protocol [5], 206 pedestrian identities (2060 images) were generally selected randomly for training and the remaining 206 pedestrian identities (2060 images) were used for testing. In the testing phase, there are two retrieval modes. The visible image is used as the retrieved image, while the image of the infrared modality is used as the retrieved image, called the visible retrieval mode (V to T). The infrared image is used as the retrieved image called the infrared retrieval mode (T to V). The final result is taken as the average result of 10 experiments.

3) SYSU-MM01-C and RegDB-C: We use the SYSU-MM01-C and RegDB-C datasets to test the corruption performance of the model, which stands for the SYSU-MM01, RegDB datasets in the corruption scenes. However, it should be noted that SYSU-MM01-C and RegDB-C are not constant datasets, but are obtained by applying random corruption to the SYSU-MM01, RegDB clean datasets. There are four categories of corruption, namely noise, blur, weather, and digitization. The noise category corruptions are: Gaussian noise, shot noise, impulse noise and speckle noise; the blur category corruptions are: defocus blur, gross glass blur, motion blur, zoom blur, and Gaussian blur. The weather category corruptions are: rain, snow, frost, fog, brightness, and spatter; and the digitization category corruptions are: contrast, elastic, pixel, JPEG compression, and saturate. Each type of corruption has 5 different levels of corruption. In total, more than 100 different types of corruption were generated. For a fair comparison, all experiments followed a standard evaluation

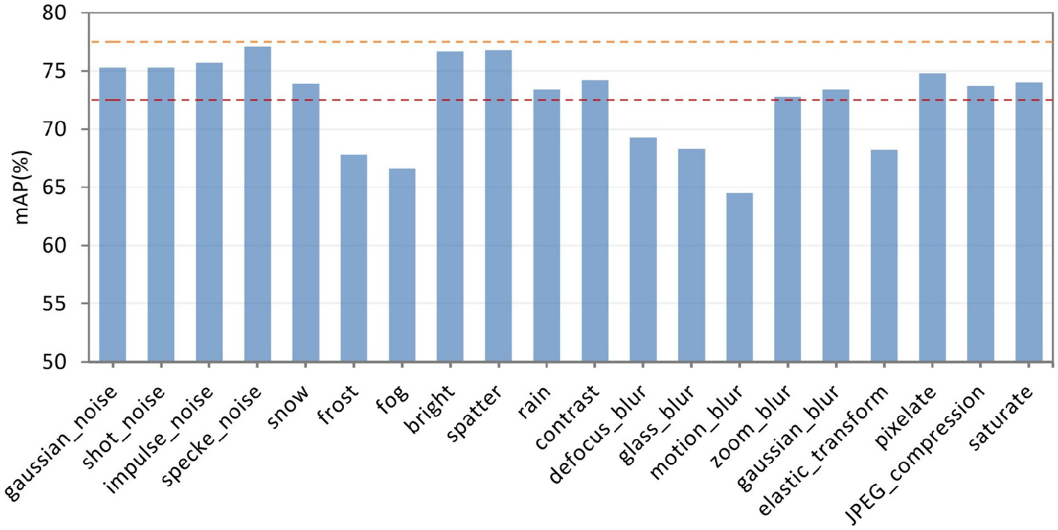


Fig. 8. The experimental results of different corruption types on SYSU-MM01-C.

protocol [19], and the average value was selected as the test result by testing 10 times on the datasets.

4) Evaluation Metrics: Following the standard evaluation protocol of VI-ReID [1], [19], we used cumulative matching characteristics (CMC), mean average precision (mAP), and mean of inverse negative penalty (mINP) for performance evaluation. CMC evaluates the probability of a correct cross-modality personal image appearing in the retrieval results. The most significant and intuitive metric is Rank-1, which reflects the probability that the top image in the search results is correct. mAP measures the retrieval performance of all correct samples when there are multiple images in the gallery. mINP which considers the hardest correct match is a supplement metric to CMC and mAP, indicating the workload of the observer.

B. Implementation Details

Our experiments were implemented on an NVIDIA RTX 3090 GPU with Pytorch version 3.10.1. For the dual-stream network, we used a modified ResNet50 [6] and set the step size of the last convolutional layer of ResNet50 to 1. In the image preprocessing stage, we first resized the input image to 256 × 128, and then used the random horizontal flipping and PMDA algorithm. Where ω is set to 4 and γ is set to 2. In the training phase, a constrained random sampling strategy is used to randomly select eight personnel categories from the training set, with eight images from each category. The parameters are updated using the Adam optimizer [30] with an initial learning rate of 3×10^{-4} and a weight decay of 5×10^{-4}. In total, the network is trained for 120 epochs and the learning rate decays to one tenth of the original rate every 40 epochs. For the semantic extraction module, we set the number of semantic feature blocks k to 4. For the graph branching module, in the normalized graph construction, we set the number of neighboring edges n for each node to 3. In the WGCN, we set the number of layers M to 2 and the scale parameter s to 1. λ is set to 0.8. Finally, the features of the global branch and graph branch are concatenated to match the final person features using cosine similarity as a distance metric.

C. Comparison With the State-of-the Arts

In this section, the proposed PMWGCN is compared with some existing methods on four datasets, RegDB-C, SYSU-MM01-C, RegDB and SYSU-MM01, to show the superiority of our method. As illustrated in Table I and Table II, Ist and 2nd best results are indicated by red and blue color, respectively.

1) RegDB-C and SYSU-MM01-C: The focus of this work. We conduct comprehensive experiments on two corrupted datasets and reproduce many methods that perform well on clean datasets.

a) Robustness to corruptions: Firstly, we evaluate the robustness of our PMWGCN in corruption scenes. We test the proposed model under 20 single corruption types, and the experiments are conducted in the indoor-search mode of SYSU-MM01-C, and the experimental results are shown in Fig. 8. The bar chart displays the specific experimental values for each type of corruptions. Among them, the yellow dashed line indicates the clean data set results and the red dashed line indicates the ten times random damage average results. As a whole, the PMWGCN shows a low degradation of mAP on the four categories of noise, blur, weather, and digital corruption, which range from 0.60% to 12.82%. Keeping the parameters constant and conducting ten random corruption tests, the mean value of mAP is stable at around 72%, which is only 4.19% lower than the clean dataset. These results fully demonstrate the relatively stable performance of our model under different corruption scenes. For the noise class, PMWGCN can play a good anti-noise role. It performs better in the face of variations such as impulse noise and speckle noise, and the mAP only degrades within 2%. The difference is that for weather and digital type corruptions, there is a huge difference between the different types. When dealing with brightness, spatter, and pixel corruptions, the mAP values all remain above the mean value, which indicates that the network is not sensitive to

TABLE I
COMPARISON WITH THE STATE-OF-THE-ART MRTHODS ON SYSU-MM01-C AND REGDB-C

Dataset		SYSU-MM01-C			RegDB-C		
Methods	Venue	Rank-1	mAP	mINP	Rank-1	mAP	mINP
ResNet-18 [50]	CVPR19	30.96	32.36	21.91	47.21	45.15	5.68
TransReID [41]	ICCV21	33.57	40.03	38.94	48.08	45.64	5.69
LightMBN [40]	ICIP21	22.61	30.10	18.73	34.10	32.40	3.25
ResNet-18+CIL	-	37.69	45.82	40.77	52.42	46.69	38.78
TransReID+CIL	-	57.84	58.46	48.90	62.78	55.85	45.18
LightMBN+CIL	-	59.37	56.70	48.58	63.34	57.00	**46.80**
ML-MDA [52]	WACV23	**61.46**	**62.21**	**52.82**	**64.38**	**59.94**	46.75
AGW [1]	TPAMI21	33.80	40.98	35.39	45.44	43.09	32.88
CAJ [37]	ICCV21	42.18	43.53	41.25	52.72	48.86	37.95
DGTL [38]	SPL21	42.22	47.90	41.15	53.50	48.35	37.02
DART [47]	CVPR22	36.46	43.82	37.26	56.84	53.60	40.96
DEEN [48]	CVPR23	44.19	45.81	37.90	36.71	37.34	28.10
PMT [49]	AAAI23	46.66	51.90	44.59	51.50	48.98	38.03
MUN [54]	ICCV23	41.17	38.63	-	52.69	50.18	-
DSCNet [51]	TIFS23	44.37	49.25	42.18	55.53	52.66	41.71
Ours	-	**65.72**	**72.55**	**66.05**	**73.35**	**65.13**	**50.41**

color information. However, the performance of the network degrades by 10.62% to 12.10% when encountering frost, fog, and elastic transform. For blur corruption, the performance is basically below the mean value, with motion blur reaching the lowest value of 64.5%. It indicates that the cross-modal model relies on information about the body structure of the person. When the personnel body elements are distorted or obscured, it is difficult for the network to determine the identity of the personnel.

b) Comparison on SYSU-MM01-C and RegDB-C: Table I shows the performance on RegDB-C and SYSU-MM01-C. We test the overall robustness of the model in corruption scenes and compare it with advanced models. There are not many researchers in the field of VI-ReID in corruption scenes. At present, CIL [19] and ML-MDA [52] are designed for corrupted VI-ReID. In view of the excellent performance of LightMBN [40] and TransReID [41] reported in [19], we reproduced BoT's [50] ResNet-18 version, LightMBN and TransReID. Some code was modified to enable them to work under VI-ReID, and CIL data augmentation was added to them for comparison (+CIL). In addition, we also reproduced some advanced VI-ReID methods such as AGW, CAJ [37], DGTL, DART [47], DEEN [48], DSCNet [51] and PMT [49], among which PMT adopts a different Transformer structure, and then test their performance in corruption scenes for comparison. Note that all methods are trained under clean datasets and only introduce corruptions in testing. For a fair comparison, all test parameters are kept consistent. the V to T mode is used uniformly on the RegDB-C dataset and the Indoor-search mode is used uniformly on the SYSU-MM01-C dataset.

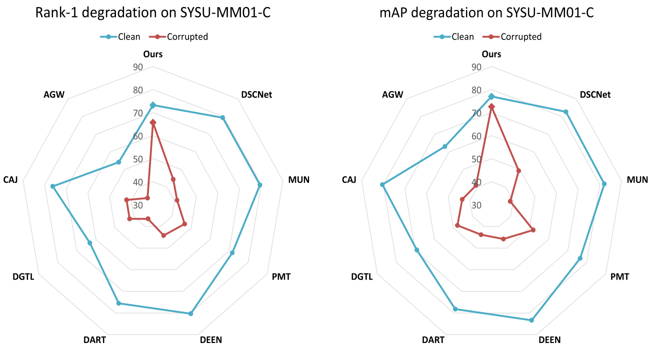


Fig. 9. Performance degradation of each model on SYSU-MM01-C.

We can see by Table I that the PMWGCN model provides significant performance improvements over the mentioned [1], [37], [38], [40], [41], [47], [48], [49], [50], [51], [52], [54], models when corruption occurs on the SYSU-MM01 and RegDB datasets, i.e., when RGB modal corruption (-C) is present. In fact, our model achieves 65.72% Rank1, 72.55% mAP, 66.05% mINP for SYSU-MM01-C and 73.35% Rank1,

TABLE II
COMPARISON WITH THE STATE-OF-THE-ART MRTHODS ON SYSU-MM01 AND REGDB

Dataset		SYSU-MMO1 (All-search)		SYSU-MMO1 (Indoor-search)		RegDB (V to T)		RegDB (T to V)	
Methods	Venue	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
X-Modal [10]	AAAI20	49.92	50.73	-	-	62.21	50.18	-	-
cm-SSFT [3]	CVPR20	61.60	63.20	70.50	72.60	71.00	71.30	72.30	72.90
MLC [36]	AISP20	62.22	59.56	69.68	71.00	81.02	78.73	-	-
HAT [32]	TIFS21	55.29	53.89	62.10	69.37	71.83	67.56	70.02	66.30
NFS [35]	CVPR21	56.91	55.45	62.79	69.79	80.54	72.10	77.95	69.79
DGTL [38]	SPL21	57.34	55.13	63.11	69.20	83.92	73.78	81.59	71.65
FMI [33]	CVPR21	60.02	58.80	66.05	72.98	73.20	71.60	71.80	70.10
CAJ [37]	ICCV21	69.88	66.89	76.26	80.37	85.03	79.14	84.75	77.82
MPANet [46]	CVPR21	**70.58**	**68.24**	**76.74**	**80.95**	83.70	80.90	82.80	80.70
DML [42]	TCSVT22	58.40	56.10	62.40	69.50	77.60	84.30	77.00	83.60
MAUM [43]	CVPR22	61.59	59.96	67.07	73.58	83.39	78.75	81.07	78.89
SPOT [44]	TIP22	65.34	62.25	69.42	74.63	80.35	72.46	79.37	72.26
FMCNet [45]	CVPR22	66.34	62.51	68.15	74.09	89.12	84.43	88.38	**83.86**
CMTR [31]	TMM23	65.45	62.90	71.46	76.67	88.11	81.66	84.92	80.79
SFANet [34]	TNNLS23	65.74	60.83	71.60	80.05	76.31	68.00	70.15	63.77
MUN [54]	CVPR23	**76.24**	**73.81**	**79.42**	**82.06**	**95.19**	**87.15**	**91.86**	**85.01**
Ours	-	**66.82**	**64.88**	**72.64**	**76.19**	**90.61**	**84.53**	**88.77**	**81.61**

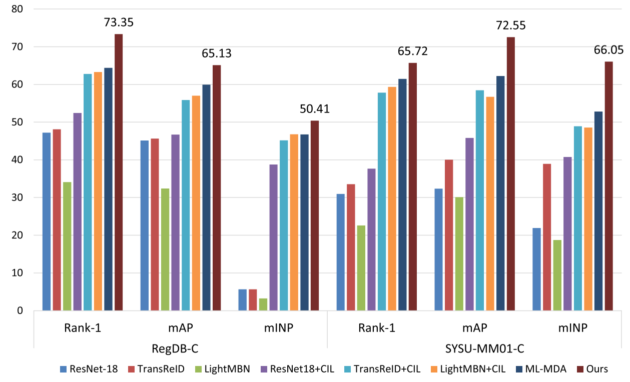
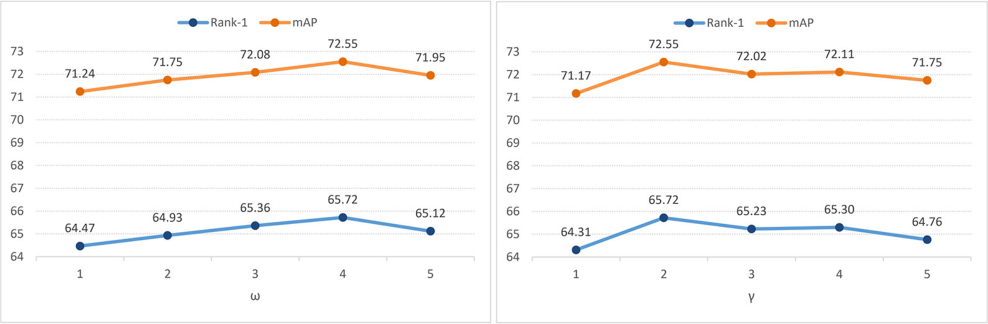


Fig. 10. Performance comparison of each model in corruption scenes.

65.13% mAP, 50.41% mINP for RegDB-C. Both CAJ and MUN [54] use channel augmentation for data augmentation, but MUN with special network design for clean scenes performs worse than original CAJ. This indicates that the current SOTA model in clean scenes is not suitable for corrupted scenes, and overfitting to clean scenes reduces the generalization performance of the model. Compared with the suboptimal method, Rank-1 scores improve by 4.26% and 8.97%, mAP scores improve by 10.34% and 5.19%, and mINP scores improve by 13.23% and 3.61%, respectively. This outperforms the current SOTAs by a remarkable margin, proving that both the polymorphic masked data augmentation algorithm and the wavelet graph convolution network can increase the model's robustness to corruption.

c) Degradation comparison: We compare the performance degradation of the SOTA methods in Fig. 9. We directly refer to the performance of clean scene reported in their papers. We see that methods perform well on clean datasets exhibit a sharp performance degradation (more than 20% Rank-1 and mAP) when encountering corruptions. On the more challenging SYSU-MM01-C, our PMWGCN only degrades 6.92% Rank-1 and 3.64% mAP compared with the clean scene. This proves the meaning of our work. The PMT adopts the Transformer structure and achieves good results compared with other methods without considering corruptions, which shows that the Transformer structure can extract robust pedestrian features in corrupted scenes. But our method still [3], [10], [31], [32], [33], [34], [35], [36], [37], [38], [42], [43], [44], [45], [46], [54] achieves more excellent results. When a modality is corrupted (visible modality), the model can still acquire valid cross-modality pedestrian features. This



(a) (b)

Fig. 11. The experimental results for different values of ω and γ on the SYSU-MM01-C dataset.

indicates that the model has the ability to adapt to real-world interference information.

d) Comparison of corrupted VI-ReID methods: In particular, we compare our method with the CIL and ML-MDA that aimed at enhancing corruption robustness of VI-ReID, as shown in Fig. 10. It can be found that adding data augmentation can effectively improve the corruption robustness of models. Especially after considering multimodal data augmentation (ML-MDA, Ours), the performance increases significantly, indicating that cross-modality information interaction is more important when image corrupt occurs. However, our PMDA introduces a fractal image set to consider natural elements and WGCN constructs a robust human body structure from a frequency domain perspective to filter out corruptions. So we achieve the best performance. All of these results demonstrate that our method can extract robust human semantic information in clean and corrupted scenes, promoting the real-world deployment of VI-ReID.

2) RegDB and SYSU-MM01: We also evaluate our PMWGCN on traditional clean datasets RegDB and SYSU-MM01, and compare with existing methods to show the stability of performance.

a) Comparison on SYSU-MM01: From Table II, we can conclude that our proposed PMWGCN outperforms most of existing methods on SYSU-MM01 and RegDB and is little worse than some methods (e.g., MUN, MPANet, CAJ), but its performance on both scenes (corrupted and clean) is the most balanced. For the SYSU-MM01, our PMWGCN achieves 66.82% Rank1 and 64.88% mAP on the all-search mode. Compared to those models based on modality compensation (e.g., X-Modal, FMCNet) and models based on modality-sharing feature learning (e.g., HAT, FMI, DML, SFANet), our PMWGCN outperforms them on two search mode. In particular, our model even achieves better results than transformer-based models (e.g., SPOT, CMTR) on both of two datasets. The reason can be attributed that the proposed polymorphic masked wavelet graph convolutional network can capture robust human semantic features and learn complementary cross-modality knowledges.

b) Comparison on RegDB: For the RegDB dataset, our proposed model achieves 90.61% Rank1 and 84.53% mAP on the visible to infrared mode (V to T), which is better than all mentioned SOTA models except MUN. Close results are also obtained in the infrared to visible mode (T to V), proving that our method is also competitive in clean scenes. This improvement can be attributed to the fact that our PMDA facilitates inter-modality information exchange within the same identity. These results validate the effectiveness of the VI-ReID model proposed in this paper on clean datasets.

D. Ablation Experiments

In this subsection, we have done comprehensive ablation experiments to evaluate the effectiveness of each component of the network structure, which mainly includes the PMDA algorithm, the graph branching module and the loss module. All experiments in this section are carried out in indoor-search mode of SYSU-MM01-C and visible to thermal query setting of RegDB-C. The baseline model is based on the Resnet-50 dual-stream backbone network [1] with CIL data augmentation. The first two stages are set as the modality-specific modules with independent parameters to learn the modality-specific feature. The remaining three stages are set as the modality-shared module with shared parameters to learn the modality-sharable feature. In other words, the baseline model only represents the global branching module. The ablation experiments of different modules are shown in Table III and the ablation experiments of loss functions are shown in Table IV.

1) The Effectiveness of the PMDA Algorithm: In this subsection, we first discuss the effectiveness of the PMDA algorithm, which introduces fractal image sets and cross-modal image sets to generate expanded images by means of appropri ate fusion. At the algorithmic pipeline level, we improve the popular data augmentation approach in the image processing field, aiming to accommodate cross-modal scenes. Is such a change effective? According to Table III, we can conclude that our PMDA data augmentation algorithm produces a 10.59% and 11.37% Rank-1 gain in performance. Compared to the baseline CIL data augmentation, it also increases mAP and mINP accuracy significantly on RegDB-C and SYSU-MM01-C. This indicates that the PMDA data augmentation algorithm is more suitable for visible-infrared cross-modal scenes. A possible explanation is that the PMDA algorithm introduces biased data augmentation. It makes the network more inclined to focus on the most discriminative cues in the infrared modality during RGB data corruption.

TABLE III
THE RESULTS OF ABLATION EXPERIMENTS FOR EACH MODULE ON REGDB-C AND SYSU-MM01-C

Method	RegDB-C			SYSU-MM01-C		
	Rank-1	mAP	mINP	Rank-1	mAP	mINP
Baseline	55.87	52.64	40.70	44.91	50.65	46.23
+PMDA	66.46	58.14	44.56	56.58	63.80	58.06
+FTN	66.92	60.92	42.60	57.63	64.79	58.64
+WGCN	67.86	61.84	47.11	59.45	66.94	63.45
+WGCN(2 layers)	**70.68**	**64.85**	**50.18**	**62.86**	**70.58**	**65.17**
Ours	**73.35**	**65.13**	**50.41**	**65.72**	**72.55**	**66.05**

2) The Effectiveness of the Graph Branching Module: In this subsection, we demonstrate the effectiveness of the graph branching module. The research of the graph branching module focuses on wavelet graph convolution network (WGCN). To verify the effectiveness of the wavelet graph convolution network, we first replace the wavelet graph convolution layer with the feature transform layer, and then train and test it under the same experimental setup. The equation of the feature transform network (FTN) can be written as $\mathbf{Y} = h(\mathbf{XW})$, where $\mathbf{X}$ is the input and $\mathbf{W}$ is the parameter matrix of the FTN layer. In contrast to WGCN, FTN can be considered as an ordinary graph convolution layer with the graph wavelet transform removed. In particular, considering single-layer WGCN and two-layer WGCN, the computational procedure can be obtained from Eq. (8) and Eq. (9). It can be found by Table III the "+FTN" model is limited in improving the Rank-1 and mINP accuracies. The reason is that the relationship between different block features cannot be modeled separately using the feature transformation layer, and FTN cannot further explore the modal common information in the visible-infrared. Compared with "+FTN", "+WGCN" increases the Rank-1 by 14.54% and mAP by 16.29% on the harder SYSU-MM01-C. The accuracy improvement proves the effectiveness of wavelet graph convolution network. The reason is that the wavelet graph convolution layer exploits human semantic information across modalities, and filters out the degradation features to extract more discriminative and corruption-invariant individual features. In addition, our proposed "+WGCN (2 layers)" achieves better results in terms of Rank-1 and mAP accuracy. Finally, the overall network obtained 73.35% Rank-1, 65.13% mAP, 50.41% mINP on RegDB-C and 65.72% Rank-1, 72.55% mAP, 66.05% mINP on SYSU-MM01-C, respectively.

3) The Effectiveness of the Loss Module: As shown in Equation (15), the loss module consists of L_{global} acting on the global features, L_{graph} acting on the graph branch features, and L_{mix} acting on the fused features.

Table IV shows the accuracy values of Rank-1, mAP, and mINP under the action of different loss functions, where w/o indicates that they are not used. We can observe the following phenomena: when the modal consistency loss L_{mid} is not used in the training phase the performance is severely degraded and the network fails to converge if both global branch and graph branch do not use L_{mid} loss. This is because the real identity labels of the persons are not used, so person matching is not possible. This shows that L_{mid} loss is essential in corrupted scenes. At the same time, L_{ct} is also very important. If L_{ct} is not used for fused features, the features used for similarity metrics cannot be captured, making the feature representation less powerful. If both L_{mid} and L_{ct} are removed from the graph branch, the graph branch fails and the image cannot pass features through the wavelet graph convolution network freely correctly, resulting in a 9.22% decrease in Rank-1, a 7.75% decrease in mAP, and a 7.93% decrease in mINP. The results of the ablation experiments in this subsection verify that the proposed each loss function in this paper is effective for the visible-infrared person re-identification task in corrupted scenes.

TABLE IV
THE RESULTS OF THE EFFECTIVENESS OF LOSS MODULE ON SYSU-MM01-C

Model	Rank-1	mAP	mINP
w/o L_{mid}(global)	63.82	68.48	66.13
w/o L_{mid}(graph)	51.22	60.16	57.21
w/o L_{mid}(global +graph)	2.10	1.02	0.62
w/o L_{ct}(graph)	**64.39**	**71.66**	**66.80**
w/o $L_{mid} + L_{ct}$ (graph)	56.50	64.80	58.12
w/o L_{mix}	62.45	70.80	64.83
Ours	**65.72**	**72.55**	**66.05**

E. Parameter Discussion

The proposed PMWGCN involves two compromise parameters ω, γ in Algorithm 1. ω controls the additive or multiplicative mixing ratio and γ controls the mixing ratio of the cross-modal image set. We analyze the two hyperparameters on the SYSU-MM01-C dataset in all-search mode.

And when one hyperparameter changes, we keep the other hyperparameters at their optimal values. The Rank-1 and the mAP results of PMWGCN with different ω and γ are shown in Fig. 11. We can see that our method is robust to the hyperparameters. From Fig. 11(a), we can observe that Rank1 and mAP rise and then fall as the weights increase. The optimal solution is obtained at $\omega = 4$. A similar phenomenon can be observed for the hyperparameter γ. The most suitable parameter setting is $\gamma = 2$.

V. Conclusion

In this paper, we propose a polymorphic masked wavelet graph convolutional network for processing VI-ReID tasks in corrupted scenes. Firstly, a cross-modal data augmentation algorithm PMDA is designed. Robustness to corruption is improved by constructing a mixed image set fusing multi-modal attributes. Secondly, a two-branch network consisting of global branches and graph branches is designed. Among them, the global branch is used to extract the overall information; the graph branch successfully extracts discriminative features while filtering noise through two steps of graph construction and wavelet graph convolution, which can respond to cross-modal scenes in a targeted manner. Finally, the corruption problems such as noise, blur, weather and digitization are alleviated by designing the corresponding loss functions according to different branches. Extensive experiments on four public datasets, RegDB, SYSU-MM01, RegDB-C and SYSU-MM01-C, demonstrate the effectiveness and superior robustness to corruptions of our PMWGCN as well as each proposed component.

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.*, vol. 44, no. 6, pp. 2872–2893, Jun. 2022.

[2] 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)*, Venice, Italy, Oct. 2017, pp. 5390–5399.

[3] Y. Lu et al., "Cross-modality person re-identification with shared-specific feature transfer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, Jun. 2020, pp. 13376–13386.

[4] M. Ye, C. Chen, J. Shen, and L. Shao, "Dynamic tri-level relation mining with attentive graph for visible infrared re-identification," *IEEE Trans. Inf. Forensics Security*, vol. 17, pp. 386–398, 2022.

[5] M. Ye, Z. Wang, X. Lan, and P. C. Yuen, "Visible thermal person re-identification via dual-constrained top-ranking," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, vol. 1, no. 2, pp. 1092–1099.

[6] 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.

[7] P. Dai, R. Ji, H. Wang, Q. Wu, and Y. Huang, "Cross-modality person re-identification with generative adversarial training," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 677–683.

[8] 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)*, Seoul, Oct. 2019, pp. 3622–3631.

[9] G.-A. Wang et al., "Cross-modality paired-images generation for RGB-infrared person re-identification," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 12144–12151.

[10] 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.*, Apr. 2020, vol. 34, no. 4, pp. 4610–4617.

[11] Z. Lu, R. Lin, and H. Hu, "Modality and camera factors bi-disentanglement for NIR-VIS object re-identification," *IEEE Trans. Inf. Forensics Security*, vol. 18, pp. 1989–2004, 2023.

[12] J. Wu, S. Zhang, Y. Zhang, Q. Huang, and J. Tian, "Improving robustness without sacrificing accuracy with patch Gaussian augmentation," *IEEE Trans. Image Process.*, vol. 30, pp. 2674–2685, 2021.

[13] D. Hendrycks et al., "AugMix: A simple data processing method to improve robustness and uncertainty," in *Proc. Int. Conf. Learning Represent.*, Apr. 2020, pp. 1–15.

[14] D. Hendrycks et al., "PixMix: Dreamlike pictures comprehensively improve safety measures," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, Jun. 2022, pp. 16762–16771.

[15] L. Wan, Q. Jing, Z. Sun, C. Zhang, Z. Li, and Y. Chen, "Self-supervised modality-aware multiple granularity pre-training for RGB-infrared person re-identification," *IEEE Trans. Inf. Forensics Security*, vol. 18, pp. 3044–3057, 2023.

[16] D. Hendrycks, X. Liu, E. Wallace, A. Dziedzic, R. Krishnan, and D. Song, "Pretrained transformers improve out-of-distribution robustness," 2020, *arXiv:2004.06100*.

[17] R. Shao, Z. Shi, J. Yi, P.-Y. Chen, and C.-J. Hsieh, "On the adversarial robustness of vision transformers," 2021, *arXiv:2103.15670*.

[18] G. K. Nayak, R. Rawal, and A. Chakraborty, "DE-CROP: Data-efficient certified robustness for pretrained classifiers," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2023, pp. 4622–4631.

[19] M. Chen, Z. Wang, and F. Zheng, "Benchmarks for corruption invariant person re-identification," 2021, *arXiv:2111.00880*.

[20] E. Rusak et al., "A simple way to make neural networks robust against diverse image corruptions," in *Proc. 16th Eur. Conf. Comput. Vis.*, vol. 12348, Glasgow, U.K., 2020, pp. 53–69.

[21] C. Si et al., "Better robustness by more coverage: Adversarial and mixup data augmentation for robust finetuning," in *Proc. Findings Assoc. Comput. Linguistics (ACL-IJCNLP)*, 2021, pp. 1569–1576.

[22] T. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *Proc. Int. Conf. Learn. Represent.*, 2017, pp. 1–14.

[23] H. Pan, Y. Bai, Z. He, and C. Zhang, "AAGCN: Adjacency-aware graph convolutional network for person re-identification," *Knowl.-Based Syst.*, vol. 236, Jan. 2022, Art. no. 107300.

[24] Z. Zhang, H. Zhang, and S. Liu, "Person re-identification using heterogeneous local graph attention networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 12131–12140.

[25] J. Liu, Z.-J. Zha, W. Wu, K. Zheng, and Q. Sun, "Spatial-temporal correlation and topology learning for person re-identification in videos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Nashville, TN, USA, Jun. 2021, pp. 4368–4377.

[26] J. Zhang, X. Li, C. Chen, M. Qi, J. Wu, and J. Jiang, "Global-local graph convolutional network for cross-modality person re-identification," *Neurocomputing*, vol. 452, pp. 137–146, Sep. 2021.

[27] D. B. Walther and D. Shen, "Nonaccidental properties underlie human categorization of complex natural scenes," *Psychol. Sci.*, vol. 25, no. 4, pp. 851–860, Apr. 2014.

[28] P. Li, Y. Xu, Y. Wei, and Y. Yang, "Self-correction for human parsing," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 6, pp. 3260–3271, Jun. 2022.

[29] X. Liang, K. Gong, X. Shen, and L. Lin, "Look into person: Joint body parsing & pose estimation network and a new benchmark," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 4, pp. 871–885, Apr. 2019.

[30] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[31] 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, 2023, doi: 10.1109/TMM.2023.3237155.

[32] 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, 2021.

[33] X. Tian, Z. Zhang, S. Lin, Y. Qu, Y. Xie, and L. Ma, "Farewell to mutual information: Variational distillation for cross-modal person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 1522–1531.

[34] H. Liu, S. Ma, D. Xia, and S. Li, "SFANet: A spectrum-aware feature augmentation network for visible-infrared person reidentification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 4, pp. 1958–1971, Apr. 2023.

[35] 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)*, Nashville, TN, USA, Jun. 2021, pp. 587–597.

[36] Z. Sun, Y. Zhu, S. Song, J. Hou, S. Du, and Y. Song, "The multi-layer constrained loss for cross-modality person re-identification," in *Proc. Int. Conf. Artif. Intell. Signal Process. (AISP)*, Amaravati, India, Jan. 2020, pp. 1–6.

[37] 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. (ICCV)*, Montreal, QC, Canada, Oct. 2021, pp. 13547–13556.

[38] H. Liu, Y. Chai, X. Tan, D. Li, and X. Zhou, "Strong but simple baseline with dual-granularity triplet loss for visible-thermal person re-identification," *IEEE Signal Process. Lett.*, vol. 28, pp. 653–657, 2021.

[39] Y. Huang, X. Fu, L. Li, and Z.-J. Zha, "Learning degradation-invariant representation for robust real-world person re-identification," *Int. J. Comput. Vis.*, vol. 130, no. 11, pp. 2770–2796, Nov. 2022.

[40] F. Herzog, X. Ji, T. Teepe, S. Hörmann, J. Gilg, and G. Rigoll, "Lightweight multi-branch network for person re-identification," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Anchorage, AK, USA, Sep. 2021, pp. 1129–1133.

[41] 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. (ICCV)*, Montreal, QC, Canada, Oct. 2021, pp. 14993–15002.

[42] D. Zhang, Z. Zhang, Y. Ju, C. Wang, Y. Xie, and Y. Qu, "Dual mutual learning for cross-modality person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 8, pp. 5361–5373, Aug. 2022.

[43] J. Liu, Y. Sun, F. Zhu, H. Pei, Y. Yang, and W. Li, "Learning memory-augmented unidirectional metrics for cross-modality person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, Jun. 2022, pp. 19344–19353.

[44] 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.

[45] Q. Zhang, C. Lai, J. Liu, N. Huang, and J. Han, "FMCNet: Feature-level modality compensation for visible-infrared person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, Jun. 2022, pp. 7339–7348.

[46] Q. Wu et al., "Discover cross-modality nuances for visible-infrared person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Nashville, TN, USA, Jun. 2021, pp. 4328–4337.

[47] M. Yang, Z. Huang, P. Hu, T. Li, J. Lv, and X. Peng, "Learning with twin noisy labels for visible-infrared person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, Jun. 2022, pp. 14288–14297.

[48] 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. (CVPR)*, Vancouver, BC, Canada, Jun. 2023, pp. 2153–2162.

[49] 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.*, Washington, DC, USA, 2023, pp. 1835–1843.

[50] H. Luo, Y. Gu, X. Liao, S. Lai, and W. Jiang, "Bag of tricks and a strong baseline for deep person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Long Beach, CA, USA, Jun. 2019, pp. 1487–1495.

[51] Y. Zhang, Y. Kang, S. Zhao, and J. Shen, "Dual-semantic consistency learning for visible-infrared person re-identification," *IEEE Trans. Inf. Forensics Security*, vol. 18, pp. 1554–1565, 2023.

[52] A. Josi, M. Alehdaghi, R. M. O. Cruz, and E. Granger, "Multimodal data augmentation for visual-infrared person ReID with corrupted data," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. Workshops (WACVW)*, Waikoloa, HI, USA, Jan. 2023, pp. 1–10.

[53] A. Josi, M. Alehdaghi, R. M. O. Cruz, and E. Granger, "Fusion for visual-infrared person ReID in real-world surveillance using corrupted multimodal data," 2023, *arXiv:2305.00320*.

[54] H. Yu, X. Cheng, W. Peng, W. Liu, and G. Zhao, "Modality unifying network for visible-infrared person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Paris, France, Oct. 2023, pp. 11185–11195.

Rui Sun (Member, IEEE) received the B.S. degree from Central South University, China, in 1998, the M.S. degree from Harbin Engineering University, China, in 2000, and the Ph.D. degree from the Huazhong University of Science and Technology, China, in 2003. He was a Visiting Scholar with the Computer Science Department, University of Missouri, Columbia, USA, from 2010 to 2011. He is currently a Professor with the School of Computer and Information, Hefei University of Technology, China. His research interests include object recognition and tracking, computer vision, and machine learning.

Long Chen received the B.S. degree from the Hefei University of Technology, China, in 2022, where he is currently pursuing the M.S. degree. His research interests include machine learning and computer vision, especially cross modal person re-identification and corruption robustness.

Lei Zhang received the B.S. degree from Huainan Normal University, China, in 2020. He is currently pursuing the M.S. degree with the Hefei University of Technology. His research interests include machine learning and computer vision, cross modal person re-identification, and graph learning.

Ruirui Xie received the B.S. degree from Anhui Jianzhu University, China, in 2021. He is currently pursuing the M.S. degree with the Hefei University of Technology. His research interests include machine learning, computer vision, and vision pretraining.

Jun Gao received the bachelor's degree in electronic engineering and the master's degree in signal and information processing from the Hefei University of Technology (HFUT), Hefei, China, in 1985 and 1991, respectively, and the Ph.D. degree in information and communication engineering from the University of Science and Technology, Hefei, in 1999. From March 1995 to October 1996, he was invited to work with the University of Stuttgart, Stuttgart, Germany. He is currently a Professor with the Laboratory of Image Information Processing, HFUT. His current research interests include image processing and intelligent information processing.