Video-based Visible-Infrared Person Re-Identification via Style Disturbance Defense and Dual Interaction

Chuhao Zhou
Harbin Institute of Technology (Shenzhen)
zhouchuhao99@gmail.com

Jinxing Li*
Harbin Institute of Technology (Shenzhen)
lijinxing158@gmail.com

Huafeng Li
Kunming University of Science and Technology
hfchina99@163.com

Guangming Lu
Harbin Institute of Technology (Shenzhen)
luguangm@hit.edu.cn

Yong Xu
Shenzhen Key Laboratory of Visual Object Detection and Recognition
yongxu@ymail.com

Min Zhang
Harbin Institute of Technology (Shenzhen)
zhangmin2021@hit.edu.cn

ABSTRACT

Video-based visible-infrared person re-identification (VVI-ReID) aims to retrieve video sequences of the same pedestrian from different modalities. The key of VVI-ReID is to learn discriminative sequence-level representations that are invariant to both intra- and inter-modal discrepancies. However, most works only focus on the elimination of modality-gap while ignore the distractors within the modality. Moreover, existing sequence-level representation learning approaches are limited to a single video, failing to mine the correlations among multiple videos of the same pedestrian. In this paper, we propose a **S**tyle **A**ugmentation, **A**ttack and **D**efense network with **G**raph-based dual interaction (SAADG) to guarantee the semantic consistency against both intra-modal discrepancies and inter-modal gap. Specifically, we first generate diverse styles for video frames by random style variation in image spaces. Followed by the style attack and defense, the intra- and inter-modal discrepancies are modeled as different types of style disturbance (attack), and our model achieves to keep the id-related content invariant under such attack. Besides, a graph-based dual interaction module is further introduced to fully explore the cross-view and cross-modal correlations among various videos of the same identity, which are then transferred to the sequence-level representations. Extensive experiments on the public SYSU-MM01 and HITSZ-VCM datasets show that our approach achieves the remarkable performance compared with state-of-the-arts. The code is available at https://github.com/ChuhaoZhou99/SAADG_VVIReID.

CCS CONCEPTS

• **Computing methodologies** → **Visual content-based indexing and retrieval**.

*Corresponding author.

MM '23, October 29-November 3, 2023, Ottawa, ON, Canada

ACM ISBN 979-8-4007-0108-5/23/10…$15.00
https://doi.org/10.1145/3581783.3612479

KEYWORDS

VI-ReID, Cross-Modality, Graph, Data Augmentation

ACM Reference Format:
Chuhao Zhou, Jinxing Li, Huafeng Li, Guangming Lu, Yong Xu, and Min Zhang. 2023. Video-based Visible-Infrared Person Re-Identification via Style Disturbance Defense and Dual Interaction. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29-November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3581783.3612479

1 INTRODUCTION

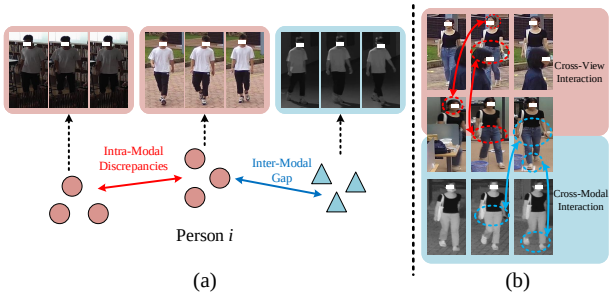

Figure 1: The insights of our methods. (a) The intra-modal discrepancies are as large as the inter-modal gap that cannot be neglected. (b) The cross-view and cross-modal correlations provide more comprehensive and discriminative features from multiple views.

Person re-identification (Re-ID) [45] is a retrieval task that aims at matching images or videos of a certain pedestrian from multiple nonoverlapping cameras. It originally focuses on visible (RGB) images and has been well solved by existing methods [4, 9, 13, 23, 40, 48–50]. However, the visible images cannot capture discriminative information when sufficient illumination is absent, making the visible Re-ID methods ineffective. Fortunately, existing cameras are capable of switching the visible mode to the infrared mode if the lighting is lower than a threshold. Thanks to this attribute, the visible-infrared person Re-ID (VI-ReID) is a feasible solution to address the aforementioned problem. Afterwards, numerous works

[11, 20, 26, 33, 37, 42–44] are proposed. Compared with still images, video sequences contain more abundant appearance features and unique motion patterns. Such additional information is significant to improve the discriminability of person representations, especially for the infrared modality [20]. Therefore, video-based VI-ReID (VVI-ReID) has been widely concerned in recent time, and some works [15, 20] have made preliminary attempts for this task.

The critical problem of VVI-ReID lies in two folds: (a) Learning the sequence-level representation that is invariant to both intra- and inter-modal discrepancies. (b) The learned sequence-level representations should be as discriminative as possible. However, most existing methods have limitations on these two aspects. For the first issue, a majority of works only concentrate on the elimination of modality-gap, without paying attention to the intra-modal distractors including background clutter, extreme illumination changes, etc. As shown in Fig.1(a), such intra-modal distractors will cause variations as large as the modality-gap and subsequently hinder the matching accuracy. As for the sequence-level representations learning, many approaches [2, 5, 21, 25, 41] tend to achieve it from a single video sequence, while the reciprocal correlations among multiple videos of the same identity have been ignored. Generally, these correlations would allow us to extract more comprehensive and modality-consistent features, as shown in Fig.1(b). On the one hand, the common issues for video-based Re-ID, including misalignment, occlusion, etc., are thorny when only one video is available. By contrary, the exploitation of cross-view correlations can refer complementary information from multiple views (red circles) to effectively alleviate such problems. On the other hand, the cross-modal correlations establish connections between discriminative features (blue circles) from different modalities, which allow the modal-specific information to refer to each other and further reduce the modality-gap.

In this paper, we propose a Style Attack and Defense (SAD) module to learn the invariant sequence-level representations against both intra- and inter-modal discrepancies. Inspired by the fact that the image style can be arbitrarily transferred by AdaIN [12] operation, the core idea of Style Attack is simulating those discrepancies by disturbing the feature extraction from a novel style disturbance perspective. At the same time, the defense module is involved to guide the model to keep the frame-level features consistent and id-discriminative before and after style attacks. By this adversarial strategy, the model becomes increasingly robust to different types of style disturbance, and eventually learns the invariant features towards intra-modal distractors and inter-modal modality-gap. Obviously, the more diverse the frame styles are, the more challenging attacks the model needs to defense, subsequently empowering it with better generalization capability. Hence, before the SAD, we further introduce the Style Augmentations (SA) to vary the image spaces for both RGB and IR frames, providing a richer range of frame styles. Once the frame-level features are produced, the Graph-based Dual Interaction (GDI) module takes over to explore the correlations among multiple videos of the same identity. To this end, the GDI regards the frame-level features as nodes and constructs the cross-view and cross-modal correlated graphs, respectively. Then, the graph convolution network is applied for correlations exploration. To overcome the absence of cross-view and cross-modal data in the testing stage, a mutual learning manner is also adopted to transfer the correlations to the final sequence-level representations.

In summary, the main contributions of our paper are:

- We propose the Style Attack and Defense (SAD) module with Style Augmentation (SA) to guide the model to extract intra- and inter-modality invariant sequence-level representations by a novel adversarial strategy.
- A Graph-based Dual Interaction (GDI) module is proposed to extend the sequence-level representations learning to multiple videos of the same identity, which is instrumental for providing more comprehensive and modality-consistent pedestrian representations.
- Extensive experiments on image-based and video-based VI-ReID datasets have been conducted, and the results show that our method is superior to the SOTAs.

2 RELATED WORK

2.1 Visible-Infrared Person Re-ID

Visible-infrared person Re-ID (VI-ReID) focuses on matching a certain pedestrian from non-overlapping cameras among different modalities. Wu *et al.* [37] and Nguyen *et al.* [24] jointly pioneered this field and contributed the most widely used benchmarks: SYSU-MM01 and RegDB datasets. The key for VI-ReID is learning the discriminative person representations that are invariant against modality-gap. To solve this issue, many works have been proposed from different perspectives including feature and metric learning [8, 18, 20, 30, 32, 33, 36, 38, 44, 46], modality generation [3, 35, 47], auxiliary information utilization [17, 19, 29] and data augmentation [6, 19, 27, 31, 43]. Our work models the intra- and inter-modal discrepancies through style augmentation and disturbance, which is more related to the modality generation and data augmentation methods.

The modality generation aims to translate a given modality to the another one, so the modality-consistence is achieved. For example, Wang *et al.* [35] introduced AlignGAN to transform the RGB images to the IR version by confusing a discriminator, and the retrieval was then conducted in the IR feature space. Besides, Choi *et al.* [3] disentangled the features into id-discriminative and id-excluded parts. The modality-gap was then alleviated by generating cross-modal id-consistent images based on modality-mixed features. Since directly generating cross-modal images may introduce lots of noise for the sake of the significant modality-gap, Zhang *et al.* [47] conducted GAN-based generation in the feature-level and utilized the generated cross-modal features for modality compensation.

Data augmentation is another widely used strategy to boost the performance for VI-ReID. Fan *et al.* [6] split the RGB images into R, G, B channels and added an additional gray image channel to obtain more IR-similar data for training. Similarly, Ye *et al.* [43] augmented the data by extending a randomly selected color channel (R,G,B) to a three-channel image. Except for only considering the RGB modality, Qian *et al.* [27] combined the patches from both modalities and generated an intermediate modality based on patches mixing, which was further utilized to reduce the modality-gap. Furthermore, Liang *et al.* [19] directly disturbed the color information of the human body by introducing the additional human key-point heatmap and ColorJitter operation.

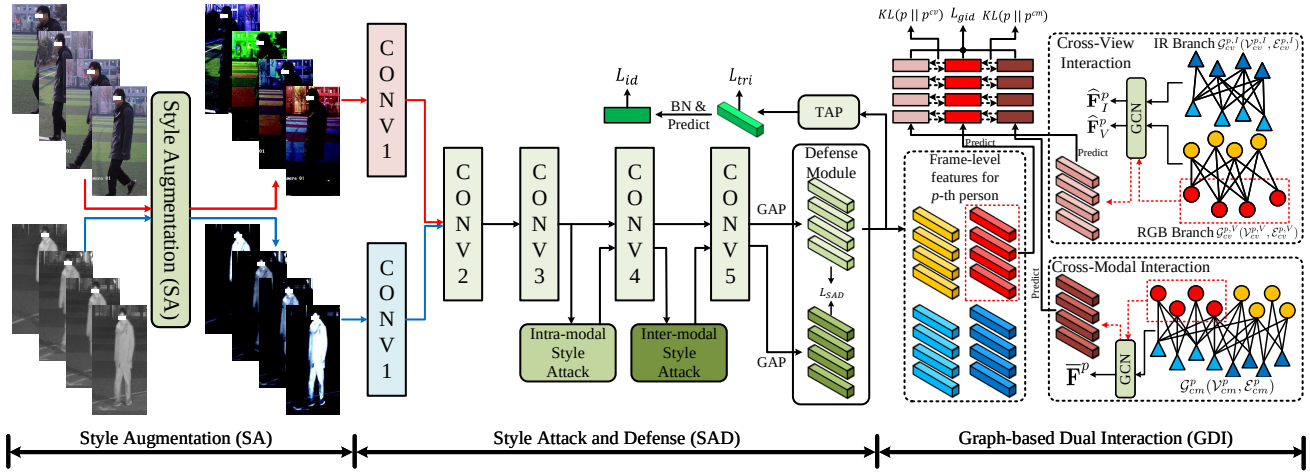

Figure 2: The pipeline of our SAADG. Given a video, the SA module is first applied to change the style of each frame. We then utilize the ResNet50 to extract frame-level features where the intra- and inter-modal style attacks are respectively embedded between the (3^{rd},4^{th}) and (4^{th}, 5^{th}) CONV blocks to disturb the feature extraction, and the defense module will force the model to keep feature consistent. Followed by the GDI, the cross-view and cross-modal correlations are explored and transferred to the sequence-level representations.

2.2 Video-based Person Re-ID

Since videos contain much richer information than still images, video-based person Re-ID has attracted extensive attention in recent time. The critical issue for it is to learn discriminative sequence-level representations. To achieve that, many methods based on Recurrent Neural Network (RNN) [21, 22, 39], 3D-convolution [1, 10, 16], and Graph Neural Network (GNN) [2, 25, 41] have been proposed.

The GNN-based methods draw more attention in recent time for the sake of its light-weight and the ability to model multi-level correlations among human parts, videos frames and instances. For example, Yang *et al.* [41] introduced both spatial and temporal graph to simultaneously explore the intra-frame structural and cross-frame complementary information. Taking the auxiliary keypoint information into consideration, Chen *et al.* [2] established graph among keypoints of all frames and introduced the graph convolution network (GCN) to assist representation learning on CNN features. In addition to the graph among keypoints, Ning *et al.* [25] further built up connection networks between frame-level features and keypoints, fully exploiting the global and local information. However, almost all of the GNN-based methods limited the feature learning into a single video but ignored the correlations of multiple videos of the same identity.

Compared to these methods, our SAADG addresses the VVI-ReID by jointly considering the cross-view and cross-modal correlations among different videos, which achieves more discriminative sequence-level representations.

3 THE PROPOSED METHOD

In this paper, we propose a novel method SAADG to obtain discriminative sequence-level representations that are invariant against both intra- and inter-modal discrepancies. The pipeline of SAADG is shown in Fig.2. Here, we denote a RGB and an IR sequence of a same pedestrian as $\mathbf{V} = \{\mathbf{V}^t | \mathbf{V}^t \in \mathbb{R}^{H_{im} \times W_{im} \times 3}\}_{t=1}^{T}$ and $\mathbf{I} = \{\mathbf{I}^t | \mathbf{I}^t \in \mathbb{R}^{H_{im} \times W_{im}}\}_{t=1}^{T}$, where H_{im} and W_{im} are the height and width of each frame, t means the t-th frame and there are T frames in total.

Initially, the RGB and IR sequences are passed into the Style Augmentation, where the style information of each frame is significantly varied (Section. 3.1). Afterwards, we take ResNet50 as the backbone [20] to extract frame-level features. Meanwhile, the intra- and inter-modal Style Attack is respectively embedded between the (3^{rd}, 4^{th}) and (4^{th}, 5^{th}) CONV blocks for disturbance. The Defense module is then attached to keep the feature to be consistent and discriminative before and after the attack (Section. 3.2). Once the frame-level features generated, they are further passed into the Graph-based Dual Interaction module to mine both cross-view and cross-modal correlations. Finally, a mutual learning manner is introduced to fully transfer the correlations into the sequence-level representations (Section. 3.3).

3.1 Style Augmentation (SA)

The more diverse styles feeded to the model, the better generalization capability of the model can achieve from training. To obtain multiple frame styles, our SA randomly disturbs the image space of each frame for both RGB and IR modalities.

As shown in Fig.3, we first sample a series of style variation factors α, β, γ and δ from a uniform distribution $\mathbb{U}(0.5, 1, 5)$. They are then utilized for channel-wise style variation to change the appearance information of each frame:

$$\begin{aligned} \mathbf{V}_{aug}^t &= [\alpha \mathbf{V}_R^t, \beta \mathbf{V}_G^t, \gamma \mathbf{V}_B^t] \\ \mathbf{I}_{aug}^t &= \delta \mathbf{I}^t \end{aligned} \tag{1}$$

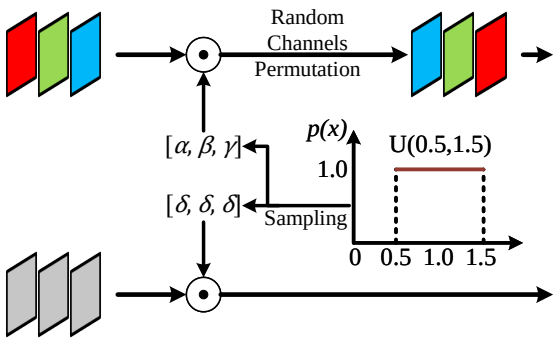

Figure 3: The insight of the SA module. The style variation factors are sampled from the uniform distribution $\mathbb{U}(0.5, 1, 5)$ and utilized to change the appearance. For visible modality, the random channel permutation is further introduced for richer styles.

where the $\mathbf{V}_R, \mathbf{V}_G, \mathbf{V}_B$ means the R,G,B channels for a RGB frame. Besides, since the IR frames only have one channel, we extend them to three-channel images by replication.

To further expand the variety of styles, we randomly permute the color channels for each RGB frame:

$$\mathbf{V}^t_{aug} = Randperm(\mathbf{V}^t_{aug}) \tag{2}$$

where the $Randperm(\cdot)$ means the random permutation of color channels for current RGB frame with equal probability.

After the style variation, the appearance of each frame has been significantly changed, as shown in Fig.4(a). These diverse styles contribute to the more challenging style attacks, which helps the model be more robust to different scenarios.

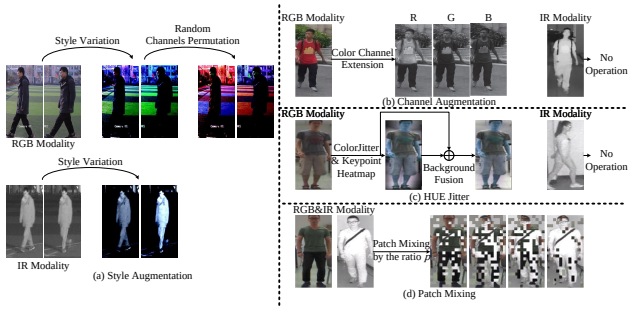

Figure 4: Comparison between SA and other data augmentation methods. (a) Our SA module; (b) Channel Augmentation [43]; (c) HUE Jitter [19]; (d) Patch Mixing [27]

Discussion. Here, we illustrate the differences between our SA and other data augmentation methods. We list three typical methods in Fig.4(b)-(d). The Channel Augmentation [43] and the Hue Jitter [19] only take the RGB modality into consideration. The former generates more IR-similar data by extending the single R/G/B channel to a three-channel image, while the latter changes the color information of the human body by using ColorJitter and keypoints heatmap. The Patch Mixing [27] fuses the patches from RGB and IR modalities to generate an intermediate modality. Different from the above methods, our SA creates frames with a richer variety of styles for both RGB and IR modalities, separately. Furthermore, except for the generated data that can be regarded as data augmentation, our method also highly concentrates on the style information (i.e., the mean and variance of the frame-level features). We will elaborate how to utilize it for the style attack in the following subsection.

3.2 Style Attack and Defense (SAD)

In the VI-ReID task, the intra-modal distractors including pedestrian postures, viewpoints, and camera styles, etc., usually increase the difficulty for Re-ID (as shown Fig.1(a)). Furthermore, compared with still images, the video sequences contain multiple frames where the frame-level variances may cause even more contamination on the sequence-level representations. To simultaneously defense against all these distractors and the modality-gap, a robust model on the intra- and inter-modal disturbance is significant. To this end, we propose the Style Attack and Defense module (SAD) by introducing an adversarial strategy from a novel style disturbance perspective. Particularly, the SAD models the intra-modal distractors and the inter-modal modality-gap as the style disturbance on a pedestrian. By embedding different types of disturbance into the learned features, our model is enforced to extract id-discriminative features that are invariant to both intra- and inter-modal discrepancies.

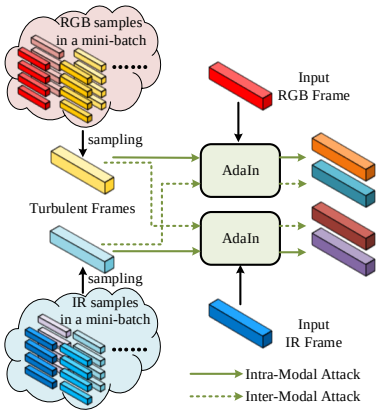

Figure 5: The process of style attack. The turbulent frame utilize its style information to disturb that of the input frame.

Mathematically, let E_{conv3}, E_{conv4} and E_{conv5} be the 3-rd, 4-th and 5-th CONV blocks in the backbone, respectively. In a mini-batch, there are N video-sequences and totally $N \times T$ frames for each modality. Here, we denote the frame-level features corresponding to RGB and IR from the X-th CONV block as $\{\mathbf{f}^{convX}_{V,i}\}^{NT}_{i=1}$ and $\{\mathbf{f}^{convX}_{I,j}\}^{NT}_{j=1}$. Without the loss of generality, we take the RGB modality as the example to show the mechanism of the SAD.

The intra-modal attack is embedded between the 3-rd and 4-th CONV blocks. As shown in Fig.5, for each RGB frame $\mathbf{f}^{conv3}_{V,i}$,

we randomly select another turbulent frame $\mathbf{f}_{V,k}^{conv3}$ from the mini-batch to disturb it via a style disturbance manner. Then, a turbulent feature for $\mathbf{f}_{V,i}^{conv3}$ is obtained:

$$\widetilde{\mathbf{f}}_{V,i}^{conv3} = \delta(\mathbf{f}_{V,k}^{conv3}) \left(\frac{\mathbf{f}_{V,i}^{conv3} - \mu(\mathbf{f}_{V,i}^{conv3})}{\delta(\mathbf{f}_{V,i}^{conv3})} \right) + \mu(\mathbf{f}_{V,k}^{conv3}) \tag{3}$$

where $\delta(\cdot)$ and $\mu(\cdot)$ represent the variance and the mean of the input feature. Based on the Eq.(3), the feature extraction is under strong intra-modal attack from both frame-level and instance-level variations. Obviously, if the model is capable of defending these attacks and capturing discriminative features, its robustness would be highly boosted.

Except for the distractors within a single modality, the modality-gap is another thorny problem for VI-ReID. Therefore, we further introduce the inter-modal attack to simulate such gap and guide the model to alleviate it, as shown in the dash line in Fig.5. After the intra-modal attack, the turbulent features are passed to the 4-th CONV block to get $\widetilde{\mathbf{f}}_{V,i}^{conv4} = E_{conv4}(\widetilde{\mathbf{f}}_{V,i}^{conv3})$ and the inter-modal attack is subsequently conducted. Being similar to Eq.(3), an intra-modal-attacked IR feature $\widetilde{\mathbf{f}}_{I,j}^{conv4}$ is randomly selected as the disturbance and the $\widetilde{\mathbf{f}}_{V,i}^{conv4}$ can be further contaminated through inter-modal attack:

$$\widetilde{\mathbf{f}}_{V,i}^{conv4} = \delta(\mathbf{f}_{I,j}^{conv4}) \left(\frac{\mathbf{f}_{V,i}^{conv4} - \mu(\mathbf{f}_{V,i}^{conv4})}{\delta(\mathbf{f}_{V,i}^{conv4})} \right) + \mu(\mathbf{f}_{I,j}^{conv4}) \tag{4}$$

After the inter-modal attack, the final CONV block and a global average pooling layer are used to produce final turbulent features for the RGB modality $\widetilde{\mathbf{f}}_{V,i} = GAP(E_{conv5}(\widetilde{\mathbf{f}}_{V,i}^{conv4}))$. Referring to the IR modality, turbulent features can be obtained in the symmetrical way. As a result, the features for RGB and IR modalities after the style attack can be formulated as $\widetilde{\mathbf{F}}_V = \{\widetilde{\mathbf{f}}_{V,i}\}_{i=1}^{NT}$ and $\widetilde{\mathbf{F}}_I = \{\widetilde{\mathbf{f}}_{I,j}\}_{j=1}^{NT}$.

Based on style attack, the frame-level features are severely contaminated by intra-modal variations at both frame-level and instance-level as well as the inter-modal discrepancies. Intuitively, if the model can keep the **discriminability** and **consistency** of the learned features before and after the attack, the robustness of it against both intra- and inter-modal discrepancies in various scenarios is achieved. To this end, the defense module, which consists of the discriminative and the consistent terms, is presented to guide the model to survive from the inferior influences caused by style attacks. In detail, denoting the attack-free features for the RGB and IR modalities as $\mathbf{F}_V = \{\mathbf{f}_{V,i}\}_{i=1}^{NT}$ and $\mathbf{F}_I = \{\mathbf{f}_{I,j}\}_{j=1}^{NT}$, the defense module can be formulated as:

$$\begin{aligned} L_{dis} &= \frac{1}{2NT}\Big(\sum_{i=1}^{NT} CE(\widetilde{p}_{V,i}, y_i) + \sum_{j=1}^{NT} CE(\widetilde{p}_{I,j}, y_j)\Big) \\ L_{con} &= \left\| \mathbf{F}_V - \widetilde{\mathbf{F}}_V \right\|_2^2 + \left\| \mathbf{F}_I - \widetilde{\mathbf{F}}_I \right\|_2^2 \\ L_{SAD} &= L_{dis} + L_{con} \end{aligned} \tag{5}$$

where $CE(\cdot)$ is the identity loss defined by cross-entropy, $\widetilde{p}_{V,i}/\widetilde{p}_{I,j}$ mean the prediction based on $\widetilde{\mathbf{f}}_{V,i}/\widetilde{\mathbf{f}}_{I,j}$, and y_i/y_j represent the corresponding identity labels.

3.3 Graph-based Dual Interaction (GDI)

Although a video-sequence contains multiple frames, the viewpoint and scene are still fixed, limiting its representation learning from other viewpoints. This fact motivates us that we can jointly utilize all videos of the same identity in a mini-batch to learn more comprehensive sequence-level representations. Furthermore, the cross-modal interaction among videos also offers an opportunity for the modal-specific features to learn the complementary knowledge from the other modality. Consequently, the modality-gap can be further reduced by a modality compensation manner. To this end, the Graph-based Dual Interaction (GDI) module builds the cross-view and cross-modal correlated graph for each identity.

Graph Construction. For the p-th person in a mini-batch, we build RGB cross-view, IR cross-view and cross-modal correlated graph as shown in Fig.(2), denoting as $\mathcal{G}_{cv}^{p,V}(\mathcal{V}_{cv}^{p,V}, \mathcal{E}_{cv}^{p,V})$, $\mathcal{G}_{cv}^{p,I}(\mathcal{V}_{cv}^{p,I}, \mathcal{E}_{cv}^{p,I})$ and $\mathcal{G}_{cm}^{p}(\mathcal{V}_{cm}^{p}, \mathcal{E}_{cm}^{p})$, respectively. Among three graphs, the $\mathcal{G}_{cv}^{p,*}(\mathcal{V}_{cv}^{p,*}, \mathcal{E}_{cv}^{p,*}), * \in \{V, I\}$ treat all RGB or IR frame-level features as nodes, while the $\mathcal{G}_{cm}^{p}(\mathcal{V}_{cm}^{p}, \mathcal{E}_{cm}^{p})$ treat all frame-level features of the p-th person as nodes. For the edges, the $\mathcal{G}_{cv}^{p,*}(\mathcal{V}_{cv}^{p,*}, \mathcal{E}_{cv}^{p,*})$ and $\mathcal{G}_{cm}^{p}(\mathcal{V}_{cm}^{p}, \mathcal{E}_{cm}^{p})$ establish pair-wise connections between frames from different views and modalities, respectively. Note that the three graphs of different people share the same topology (i.e., edge connections). Intuitively, the more similar the two frames are, the stronger connections they should hold. Therefore, we further construct the adjacency matrix $\mathbf{A}_{cv}^{p,*} \in \mathbb{R}^{MT \times MT}$ and $\mathbf{A}_{cm}^{p} \in \mathbb{R}^{2MT \times 2MT}$ associate to both kinds of graphs based on cosine similarity. The formulations of the graph construction can be found in Tab.1.

Correlations Exploration. Once the graphs established, two GCNs are utilized for correlations exploration. Following [14], let $\widetilde{\mathbf{A}}_{cv}^{p,*} = \mathbf{A}_{cv}^{p,*} + \mathbf{I_{MT}}$ and $\widetilde{\mathbf{A}}_{cm}^{p} = \mathbf{A}_{cm}^{p} + \mathbf{I_{2MT}}$, the re-normalization trick is applied to both adjacent matrices:

$$\begin{aligned} \hat{\mathbf{A}}_{cv}^{p,*} &= (\widetilde{\mathbf{D}}_{cv}^{p,*})^{-\frac{1}{2}} \widetilde{\mathbf{A}}_{cv}^{p,*} (\widetilde{\mathbf{D}}_{cv}^{p,*})^{-\frac{1}{2}} \\ \hat{\mathbf{A}}_{cm}^{p} &= (\widetilde{\mathbf{D}}_{cm}^{p})^{-\frac{1}{2}} \widetilde{\mathbf{A}}_{cm}^{p} (\widetilde{\mathbf{D}}_{cm}^{p})^{-\frac{1}{2}} \end{aligned} \tag{6}$$

where $\widetilde{\mathbf{D}}_{(i,i)} = \sum_j \widetilde{\mathbf{A}}_{(i,j)}$ is the corresponding degree matrix, and $\mathbf{I_{MT}}/\mathbf{I_{2MT}}$ are the identity matrices whose dimensions are $MT/2MT$, respectively.

Afterwards, the cross-view correlations $\widehat{\mathbf{F}}^p = [\widehat{\mathbf{F}}_V^p, \widehat{\mathbf{F}}_I^p]$ and cross-modal correlations $\overline{\mathbf{F}}^p$ for the p-th person can be obtained through Eq.(7) and Eq.(8):

$$\widehat{\mathbf{F}}^p = [\hat{\mathbf{A}}_{cv}^{p,V} \mathbf{F}_V^p \mathbf{W}_{cv}, \hat{\mathbf{A}}_{cv}^{p,I} \mathbf{F}_I^p \mathbf{W}_{cv}] \tag{7}$$

$$\overline{\mathbf{F}}^p = \hat{\mathbf{A}}_{cm}^{p} [\mathbf{F}_V^p, \mathbf{F}_I^p] \mathbf{W}_{cm} \tag{8}$$

where $\mathbf{F}_V^p/\mathbf{F}_I^p$ means all RGB/IR frame-level features for the p-th person, $[,]$ means the concatenation, and $\mathbf{W}_{cv}/\mathbf{W}_{cm}$ denotes the parameters for two GCNs.

Mutual Learning. Notably, the correlations exploration is infeasible in the testing stage. To solve it, a mutual learning manner is proposed to make the model adaptively transfer the above correlations to the sequence-level representations. Specifically, there are P different people in a mini-batch. The frame-level features $\mathbf{F} = \{[\mathbf{F}_V^p, \mathbf{F}_I^p]\}_{p=1}^P$, the cross-view correlations $\widehat{\mathbf{F}} = \{\widehat{\mathbf{F}}^p\}_{p=1}^P$, and the cross-modal correlations $\overline{\mathbf{F}} = \{\overline{\mathbf{F}}^p\}_{p=1}^P$ are supervised by the

Table 1: The formulations for cross-view and cross-modal graph constructions. In a mini-batch, there are M RGB and M IR videos for each person, and each video contains T frames.

Graph	Nodes $\mathcal{V}$	edges $\mathcal{E}$	Adjacency Matrix $\mathbf{A}$
$\mathcal{G}^{p,*}_{cv}(\mathcal{V}^{p,*}_{cv}, \mathcal{E}^{p,*}_{cv})$	$\mathcal{V}^{p,*}_{cv} = \{\mathbf{f}^{p,*}_{i,j}\}^{M,T}_{i=1,j=1}$	$\mathcal{E}^{p,*}_{cv} = \{(\mathbf{f}^{p,*}_{i,j}, \mathbf{f}^{p,*}_{m,n})\}_{i\neq m}$	$\mathbf{A}^{p,*}_{cv}(m,n) = \begin{cases} \frac{\mathbf{f}_m \cdot \mathbf{f}_n}{\|\mathbf{f}_m\|\|\mathbf{f}_n\|}, & (\mathbf{f}_m, \mathbf{f}_n) \in \mathcal{E}^{p,*}_{cv} \\ 0, & otherwise \end{cases}$
$\mathcal{G}^{p}_{cm}(\mathcal{V}^{p}_{cm}, \mathcal{E}^{p}_{cm})$	$\mathcal{V}^{p}_{cm} = \{\mathbf{f}^{p,V}_{i}\}^{MT}_{i=1} \cup \{\mathbf{f}^{p,I}_{j}\}^{MT}_{j=1}$	$\mathcal{E}^{p}_{cm} = \{(\mathbf{f}^{p,V}_{i}, \mathbf{f}^{p,I}_{j})\}$	$\mathbf{A}^{p}_{cm}(m,n) = \begin{cases} \frac{\mathbf{f}_m \cdot \mathbf{f}_n}{\|\mathbf{f}_m\|\|\mathbf{f}_n\|}, & (\mathbf{f}_m, \mathbf{f}_n) \in \mathcal{E}^{p}_{cm} \\ 0, & otherwise \end{cases}$

Table 2: Comparisons of our method with SOTA methods on HITSZ-VCM in terms of CMC(%) and mAP(%).

Method	Sources	Type	Infrared to Visible (I2V)					Visible to Infrared (V2I)				
			R1	R5	R10	R20	mAP	R1	R5	R10	R20	mAP
LbA [26]	ICCV'21	Image	46.38	65.29	72.23	79.41	30.69	49.30	69.27	75.90	82.21	32.38
MPANet [38]	CVPR'21	Image	46.51	63.07	70.51	77.77	35.26	50.32	67.31	73.56	79.66	37.80
DDAG [44]	ECCV'20	Image	54.62	69.79	76.05	81.50	39.26	59.03	74.64	79.53	84.04	41.50
VCD [33]	CVPR'21	Image	54.53	70.01	76.28	82.01	41.18	57.52	73.66	79.38	83.61	43.45
CAJL [43]	ICCV'21	Image	56.59	73.49	79.52	84.05	41.49	60.13	74.62	79.86	84.53	42.81
SGIEL† [7]	CVPR'23	Image	67.65	80.32	84.73	-	52.30	70.23	82.19	86.11	-	52.54
MITML [20]	CVPR'22	Video	63.74	76.88	81.72	86.28	45.31	64.54	78.96	82.98	87.10	47.69
IBAN [15]	TCSVT'23	Video	65.03	78.34	82.98	87.19	48.77	69.58	81.51	85.43	88.78	50.96
Ours	-	Video	67.23	79.30	83.66	87.43	50.46	70.68	82.19	85.37	88.55	53.28
Ours†	-	Video	**69.22**	**80.61**	**85.03**	**88.66**	**53.77**	**73.13**	**83.47**	**86.86**	**89.72**	**56.09**

identity loss, which are jointly denoted as L_{gid}. Then, the cross-view and cross-modal mutual learning can be formulated as:

$$L_{ml} = \frac{1}{2NT}\left(\sum_{i=1}^{2NT} D_{KL}(p_i||p_i^{cv}) + \sum_{i=1}^{2NT} D_{KL}(p_i||p_i^{cm})\right) \quad (9)$$

where p_i, p_i^{cv}, p_i^{cm} denote the prediction based on the frame-level features $\mathbf{f}_i \in \mathbf{F}$ and the corresponding correlations $\mathbf{f}_i^{cv} \in \widehat{\mathbf{F}}, \mathbf{f}_i^{cm} \in \overline{\mathbf{F}}$, respectively. $D_{KL}(\cdot)$ means the Kullback-Leibler divergence between two distributions.

The objective of GDI is the combination of the identity loss and mutual learning loss: $L_{GDI} = L_{gid} + L_{ml}$.

3.4 Training Objectives

Finally, the sequence-level representations of each pedestrian can be obtained through temporal average pooling over all frame-level features, as shown in Fig.2.

The total objective function of our SAADG is defined as:

$$L = L_{id} + L_{tri} + L_{SAD} + \lambda L_{GDL} \quad (10)$$

where L_{id} and L_{tri} are identity loss and triplet loss based on the sequence-level representations, λ is utilized to balance the contribution of the SAD and GDI, which is set to 0.5.

4 EXPERIMENTS

4.1 Experiments Settings

Dataset and Evaluation Protocol. We evaluate our method mainly on the dataset **HITSZ-VCM** [20], which is specifically designed for **VVI-ReID**. It is captured by 6 visible and 6 infrared cameras and is officially divided into the training and testing sets. Specifically, the training set contains 500 identities with 6,142 visible and 4,919 infrared video sequences, while the testing set contains 427 identities with 5,643 visible and 5,159 infrared video sequences. The testing protocol contains both 'infrared-to-visible (I2V)' and 'visible-to-infrared (V2I)'. Notably, our SA and SAD can be treated as plug-in modules for existing methods. To further show their transferability, we also add them to existing image-based VI-ReID methods which are evaluated on the **SYSU-MM01** dataset. It contains images captured by 4 visible and 2 infrared cameras with 395 identities for training and another 96 identities for evaluation. We conduct experiments in both *all-search* and *indoor-search* modes under single-shot setting, where *all-search*(*indoor-search*) means that images from all visible cameras (indoor visible cameras) are utilized to form the gallery set. For quantitatively evaluation, we adopt the widely used Cumulative Matching Characteristic curve (CMC) and mean Average Precision (mAP) as metrics.

Implementation Details. We adopt the same backbone as that in [20]. Besides, a strong backbone is also adopted by following the [7] to make a fair comparison with [7]. The strong backbone replaces the average pooling layer with GEM-pooling similar to [45] and additionally takes the Channel-Level Random Erasing that proposed in [43] for further data augmentation. To fairly compare with the image-based and video-based methods, we follow the widely utilized settings to form a mini-batch. Specifically, 8 identities are randomly sampled, each of which contains 2(4) visible and 2(4) infrared videos(images) for HITSZ-VCM(SYSU-MM01). The SGD optimizer is utilized with the weight decay of 5×10^{-4} and the momentum of 0.9 for optimization. The learning rate is initialized to 0.1 with a linear warmup strategy for 10 epochs, and then decayed

at the 60-th and 120-th epochs with a decay factor of 0.1. Besides, being similar to [20], we set the learning rate of the backbone to be one-tenth of other components in the whole training process of 200 epochs.

Table 3: The effectiveness of SA and SAD on image-based methods. The experiments are conducted on SYSU-MM01 under the single-shot setting.

Method	Sources	All Search		Indoor Search	
		R1	mAP	R1	mAP
AGW	TPAMI2021	47.50	47.65	55.89	62.76
AGW+SA+SAD	-	57.95	54.77	69.34	73.25
DDAG	ECCV2020	54.75	53.02	61.02	67.98
DDAG+SA+SAD	-	60.27	55.32	67.12	72.03
CAJL	ICCV2021	69.88	66.89	76.30	80.40
CAJL+SA+SAD	-	71.02	67.65	78.71	81.05

4.2 Comparison with Existing Methods

In this subsection, we compare our method with state-of-the-art VI-ReID methods, including image-based and video-based ones. Specifically, for the image-based methods, the DDAG[44], CAJL[43], LbA[26], MPANet[38], VCD[33], and SGIEL[7] are taken into consideration. For a fair comparison, the frame-level features are obtained through the image-based methods and a temporal average pooling layer is conducted to form the sequence-level representations. The video-based methods contain MITML [20] and IBAN [15].

The results are shown in Tab.2, in which the † means that the strong backbone is utilized. It can be found that our SAADG outperforms almost all VI-ReID methods in both I2V and V2I testing modes of HITSZ-VCM and is competitive to SGIEL without the strong backbone. When the strong backbone is adopted, our SAADG† achieves further improvement and is remarkable superior to SGIEL. Specifically, compared with the second best method (SGIEL), our method obtain (+1.57% R1,+1.47% mAP) for I2V and (+2.9% R1,+3.55% mAP) for V2I. Besides, note that the SGIEL [7] and IBAN [15] have utilized additional auxiliary data, i.e., the human parsing masks and anaglyph data. Compared with them, our method also reserves the superiority, which further substantiates its effectiveness.

In addition, we also demonstrate the transferability of SA and SAD modules on the image-based VI-ReID task. Specifically, we select three methods: AGW[45], DDAG[44], CAJL[43] and evaluate their performance on the SYSU-MM01 dataset with or without the SA and SAD modules. The results are summarized in Tab.3. As we can see, these two modules achieve performance improvement for all three methods, further demonstrating the general effectiveness of our method in both image-based and video-based VI-ReID tasks.

4.3 Ablation Study

In this subsection, we conduct ablation studies to show the contribution of our proposed SA, SAD and GDI modules. All experiments are performed on HITSZ-VCM under both I2V and V2I modes. The results are summarized in Tab.4. The 'base' (Exp 1) means the strong backbone, which is trained with identity and triplet loss.

Table 4: Ablation results for key components of our method. We conduct all experiments based on the strong baseline on HITSZ-VCM.

Exp	Components						I2V		V2I	
	SA	SAD		GDI			R1	mAP	R1	mAP
		Intra	Inter	CV	CM	ML				
1 (base)							57.46	44.04	61.19	46.24
2		✓					63.61	49.44	67.97	52.19
3		✓	✓				65.34	51.50	68.54	53.72
4	✓	✓	✓				67.84	52.75	71.17	55.26
5				✓		✓	61.65	47.16	64.82	48.59
6				✓	✓	✓	64.33	49.15	67.62	51.04
7		✓	✓	✓	✓	✓	66.58	52.30	70.21	54.15
8	✓	✓	✓	✓	✓		68.39	53.05	71.66	55.61
9(full)	✓	✓	✓	✓	✓	✓	69.22	53.77	73.13	56.09

Table 5: Comparison between our SA and different data augmentation methods. The SA in SAADG is replaced with other data augmentation methods and evaluated on HITSZ-VCM.

Method	Sources	I2V		V2I	
		R1	mAP	R1	mAP
ColorJitter	MM2022	67.58	52.40	71.15	54.83
Channel Augmentation	ICCV2021	68.17	52.55	72.05	55.16
Style Augmentation (ours)	-	69.22	53.77	73.13	56.09

Effects of SAD. We first illustrate the effectiveness of the SAD module. In Exp 2 and Exp 3 of Tab.4, the intra-modal style attack is added into the baseline model, followed by the addition of inter-modal attack. Both style attack strategies have achieved remarkable performance improvements. Besides, compared between the Exp 6 and Exp 7, the SAD module continuously boosts the matching accuracy with the GDI module. These facts indicate that the intra- and inter-modal style attacks do simulate the intra-modal distractors and the inter-modal discrepancies in real scenarios. By defending these attacks, the SAD module effectively improves the robustness of our model.

Effects of GDI. The GDI module consists of cross-view (CM) and cross-modal (CM) interaction. Similar to the SAD, we apply them to the baseline model in order in the Exp 5 and Exp 6, which also gain performance enhancement, respectively. Moreover, as listed in Exp 7, the GDI module still demonstrates its effectiveness combined with the SAD module. It has been proved that the GDI module allows our model to extract more comprehensive but view- and modal-consistent features through the graph-based interactions. Eventually, the comparison between Exp 8 and Exp 9 illustrates the effectiveness of our mutual learning manner.

Effects of SA. The SA is treated as an auxiliary module for a richer variety of frame styles, and it is closely related to the SAD module. Therefore, we put it together with the SAD module to demonstrate its effectiveness. As shown in Exp 4 and Exp 9, the involvement of SA can further promote the performances in the presence of both 'SAD' and 'SAD+GDI'. It can be deduced that the more diverse frame styles enforce the model to defense more challenging style attacks in the training stage, improving the generalization capability for unseen scenarios. Furthermore, in the subsection 3.1, we have discussed the differences between

our SA to other data augmentation methods. Here, we replace the SA in SAADG with other methods to further demonstrate its superiority. In detail, since the official code of Patch Mixing [27] is temporarily unavailable, we adopt the channel augmentation [43] and the ColorJitter without the keypoint heatmap [19] to replace the SA, respectively. The results in Tab.5 show that the SA takes the lead, which indicates that the diverse frame styles are indeed more helpful in boosting the model robustness.

4.4 Visualization Analysis

In this subsection, we respectively visualize the features in the feature spaces and on the feature maps. A comprehensive analysis is conducted as well to illustrate why our proposed method is effective.

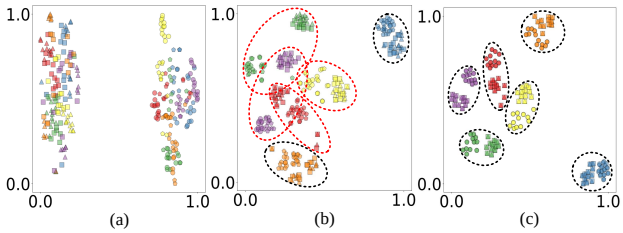

Figure 6: Visualization of feature spaces in the testing set by tSNE. The circles/triangles and pentagons/squares respectively refer to the original and turbulent frame-level features for visible/infrared modalities. The identity is distinguished by color. (a) baseline pre-trained on ImageNet; (b) baseline; (c) Our proposed SAADG.

Visualization of Feature Spaces. Here, we randomly select 6 people in the testing set and visualize their original and turbulent frame-level features by t-SNE [34]. As shown in Fig.6, the 'baseline' model that pre-trained on ImageNet fails to alleviate the modality-gap and separate different people. Nonetheless, the SA and SAD do generate frames with richer styles, which resulting in a more diverse feature space distribution. Besides, compared with the 'baseline' model, our proposed model has better robustness for the style attacks and achieves to keep the feature invariant before and after the attacks. Consequently, our proposed model learns a more accurate feature space, where features from both modalities are well grouped according to the identity.

Visualization on Feature Maps. Furthermore, we randomly choose a person and visualize the feature maps of both modalities in different scenarios (identified by different camera id). We utilize the last feature map of the backbone and visualize it through GradCAM [28]. The results are illustrated in Fig.7. It shows that the 'baseline' model tends to focus on specific local parts to identify a person. Besides, the focused parts are inconsistent among different scenarios, and even exist misalignments (e.g., the left parts in Cam 6 for the visible modality and the top left corner in Cam 4 for the infrared modality). By contrast, our proposed method concentrates on the whole person, and the attention regions are consistent among different scenarios and modalities. As a result, it can learn more comprehensive representations that are invariant to both intra- and inter modal discrepancies.

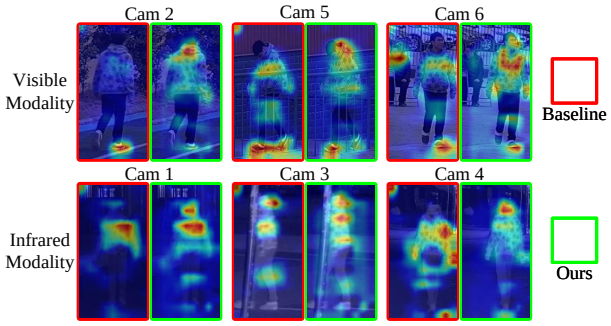

Figure 7: Visualization of feature maps in multiple scenarios (different Cam IDs) for both modalities. The red and green bounding boxes denote feature maps of the baseline and our SAADG.

5 CONCLUSION

In this paper, we solve the two key issues for video-based visible-infrared person Re-ID (VVI-ReID). The one is learning invariant sequence-level representations against both intra- and inter-modal discrepancies. We achieve it from a novel style disturbance perspective and propose the Style Attack and Defense (SAD) module. The SAD simulates those discrepancies by applying different types of style attack during the feature extraction, and then it enforces the model to keep the feature consistent. By this adversarial strategy, the robustness of the model on both intra- and inter-modal discrepancies has been boosted. We also append a Style Augmentation (SA) module before the SAD to generate a richer range of frame styles, which provide the model with more challenging style attacks. The other issue is that the learned sequence-level representations should be as discriminative as possible. Different from existing methods, we extend the representation learning to multiple videos. The Graph-based Dual Interaction (GDI) is then proposed to explore the cross-view and cross-modal correlations. Thanks to the GDI, more comprehensive and consistent sequence-level representations can be achieved. Extensive experiments on both the image-based dataset SYSU-MM01 and video-based dataset HITSZ-VCM have shown the effectiveness of our method.

6 ACKNOWLEDGEMENTS

This work was supported by the NSFC under Grant (62272133, 61906162, 62276120, 62176077), in part by Shenzhen Colleges and Universities Stable Support Program No. GXWD20220811170100001, Yunnan Fundamental Research Projects (202301AV070004), Shenzhen Science and Technology Program (RCBS20200714114910193), Shenzhen Key Technical Project (2022N001, 2020N046), Shenzhen Fundamental Research Fund (JCYJ20210324132210025), and Guangdong Provincial Key Laboratory of Novel Security Intelligence Technologies (2022B1212010005).

REFERENCES

[1] Abhishek Aich, Meng Zheng, Srikrishna Karanam, Terrence Chen, Amit K Roy-Chowdhury, and Ziyan Wu. 2021. Spatio-temporal representation factorization for video-based person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV'21)*. 152–162.

[2] Di Chen, Andreas Doering, Shanshan Zhang, Jian Yang, Juergen Gall, and Bernt Schiele. 2022. Keypoint message passing for video-based person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI'22)*, Vol. 36. 239–247.

[3] Seokeon Choi, Sumin Lee, Youngeun Kim, Taekyung Kim, and Changick Kim. 2020. Hi-CMD: Hierarchical cross-modality disentanglement for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'20)*. 10257–10266.

[4] Neng Dong, Liyan Zhang, Shuanglin Yan, Hao Tang, and Jinhui Tang. 2023. Erasing, Transforming, and Noising Defense Network for Occluded Person Re-Identification. *arXiv preprint arXiv:2307.07187* (2023).

[5] Chanho Eom, Geon Lee, Junghyup Lee, and Bumsub Ham. 2021. Video-based person re-identification with spatial and temporal memory networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV'21)*. 12036–12045.

[6] Xing Fan, Hao Luo, Chi Zhang, and Wei Jiang. 2020. Cross-spectrum dual-subspace pairing for RGB-infrared cross-modality person re-identification. *arXiv preprint arXiv:2003.00213* (2020).

[7] Jiawei Feng, Ancong Wu, and Wei-Shi Zheng. 2023. Shape-Erased Feature Learning for Visible-Infrared Person Re-Identification. *arXiv preprint arXiv:2304.04205* (2023).

[8] Yajun Gao, Tengfei Liang, Yi Jin, Xiaoyan Gu, Wu Liu, Yidong Li, and Congyan Lang. 2021. MSO: Multi-feature space joint optimization network for rgb-infrared person re-identification. In *Proceedings of the 29th ACM International Conference on Multimedia (MM'21)*. 5257–5265.

[9] Zan Gao, Hongwei Wei, Weili Guan, Weizhi Nie, Meng Liu, and Meng Wang. 2022. Multigranular visual-semantic embedding for cloth-changing person re-identification. In *Proceedings of the 30th ACM International Conference on Multimedia (MM'22)*. 3703–3711.

[10] Xinqian Gu, Hong Chang, Bingpeng Ma, Hongkai Zhang, and Xilin Chen. 2020. Appearance-preserving 3d convolution for video-based person re-identification. In *Proceedings of the IEEE/CVF European Conference on Computer Vision (ECCV'20)*. Springer, 228–243.

[11] Xin Hao, Sanyuan Zhao, Mang Ye, and Jianbing Shen. 2021. Cross-modality person re-identification via modality confusion and center aggregation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV'21)*. 16403–16412.

[12] Xun Huang and Serge Belongie. 2017. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV'17)*. 1501–1510.

[13] Xin Jin, Cuiling Lan, Wenjun Zeng, Zhibo Chen, and Li Zhang. 2020. Style normalization and restitution for generalizable person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'20)*. 3143–3152.

[14] Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907* (2016).

[15] Huafeng Li, Minghui Liu, Zhanxuan Hu, Feiping Nie, and Zhengtao Yu. 2023. Intermediary-guided Bidirectional Spatial-Temporal Aggregation Network for Video-based Visible-Infrared Person Re-Identification. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)* (2023).

[16] Jianing Li, Shiliang Zhang, and Tiejun Huang. 2019. Multi-scale 3d convolution network for video based person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI'19)*, Vol. 33. 8618–8625.

[17] Yulin Li, Tianzhu Zhang, Xiang Liu, Qi Tian, Yongdong Zhang, and Feng Wu. 2022. Visible-Infrared Person Re-Identification With Modality-Specific Memory Network. *IEEE Transactions on Image Processing (TIP)* 31 (2022), 7165–7178.

[18] Zechao Li, Hao Tang, Zhimao Peng, Guo-Jun Qi, and Jinhui Tang. 2023. Knowledge-guided semantic transfer network for few-shot image recognition. *IEEE Transactions on Neural Networks and Learning Systems (TNNLS)* (2023).

[19] Tengfei Liang, Yi Jin, Wu Liu, Songhe Feng, Tao Wang, and Yidong Li. 2022. Keypoint-Guided Modality-Invariant Discriminative Learning for Visible-Infrared Person Re-identification. In *Proceedings of the 30th ACM International Conference on Multimedia (MM'22)*. 3965–3973.

[20] Xinyu Lin, Jinxing Li, Zeyu Ma, Huafeng Li, Shuang Li, Kaixiong Xu, Guangming Lu, and David Zhang. 2022. Learning Modal-Invariant and Temporal-Memory for Video-based Visible-Infrared Person Re-Identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'22)*. 20973–20982.

[21] Xuehu Liu, Pingping Zhang, Chenyang Yu, Huchuan Lu, and Xiaoyun Yang. 2021. Watching you: Global-guided reciprocal learning for video-based person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'21)*. 13334–13343.

[22] Yiheng Liu, Zhenxun Yuan, Wengang Zhou, and Houqiang Li. 2019. Spatial and temporal mutual promotion for video-based person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI'19)*, Vol. 33. 8786–8793.

[23] Zhongxing Ma, Yifan Zhao, and Jia Li. 2021. Pose-guided inter-and intra-part relational transformer for occluded person re-identification. In *Proceedings of the 29th ACM International Conference on Multimedia (MM'21)*. 1487–1496.

[24] Dat Tien Nguyen, Hyung Gil Hong, Ki Wan Kim, and Kang Ryoung Park. 2017. Person recognition system based on a combination of body images from visible light and thermal cameras. *Sensors* 17, 3 (2017), 605.

[25] Jiaqi Ning, Fei Li, Rujie Liu, Shun Takeuchi, and Genta Suzuki. 2022. Temporal Extension Topology Learning for Video-based Person Re-Identification. In *Proceedings of the Asian Conference on Computer Vision (ACCV'22)*. 207–219.

[26] Hyunjong Park, Sanghoon Lee, Junghyup Lee, and Bumsub Ham. 2021. Learning by aligning: Visible-infrared person re-identification using cross-modal correspondences. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV'21)*. 12046–12055.

[27] Zhihao Qian, Yutian Lin, and Bo Du. 2023. Visible-Infrared Person Re-Identification via Patch-Mixed Cross-Modality Learning. *arXiv preprint arXiv:2302.08212* (2023).

[28] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV'17)*. 618–626.

[29] Tongzhen Si, Fazhi He, Penglei Li, and Xiaoxin Gao. 2023. Tri-modality consistency optimization with heterogeneous augmented images for visible-infrared person re-identification. *Neurocomputing* 523 (2023), 170–181.

[30] Hanzhe Sun, Jun Liu, Zhizhong Zhang, Chengjie Wang, Yanyun Qu, Yuan Xie, and Lizhuang Ma. 2022. Not All Pixels Are Matched: Dense Contrastive Learning for Cross-Modality Person Re-Identification. In *Proceedings of the 30th ACM International Conference on Multimedia (MM'22)*. 5333–5341.

[31] Hao Tang, Zechao Li, Zhimao Peng, and Jinhui Tang. 2020. Blockmix: meta regularization and self-calibrated inference for metric-based meta-learning. In *Proceedings of the 28th ACM International Conference on Multimedia (MM'20)*. 610–618.

[32] Hao Tang, Chengcheng Yuan, Zechao Li, and Jinhui Tang. 2022. Learning attention-guided pyramidal features for few-shot fine-grained recognition. *Pattern Recognition (PR)* 130 (2022), 108792.

[33] Xudong Tian, Zhizhong Zhang, Shaohui Lin, Yanyun Qu, Yuan Xie, and Lizhuang Ma. 2021. Farewell to mutual information: Variational distillation for cross-modal person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'21)*. 1522–1531.

[34] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research (JMLR)* 9, 11 (2008).

[35] Guan'an Wang, Tianzhu Zhang, Jian Cheng, Si Liu, Yang Yang, and Zengguang Hou. 2019. RGB-infrared cross-modality person re-identification via joint pixel and feature alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV'19)*. 3623–3632.

[36] Yang Wang, Jinjia Peng, Huibing Wang, and Meng Wang. 2022. Progressive learning with multi-scale attention network for cross-domain vehicle re-identification. *Science China Information Sciences (SCIS)* 65, 6 (2022), 160103.

[37] Ancong Wu, Wei-Shi Zheng, Hong-Xing Yu, Shaogang Gong, and Jianhuang Lai. 2017. RGB-infrared cross-modality person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV'17)*. 5380–5389.

[38] Qiong Wu, Pingyang Dai, Jie Chen, Chia-Wen Lin, Yongjian Wu, Feiyue Huang, Bineng Zhong, and Rongrong Ji. 2021. Discover cross-modality nuances for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'21)*. 4330–4339.

[39] Shuangjie Xu, Yu Cheng, Kang Gu, Yang Yang, Shiyu Chang, and Pan Zhou. 2017. Jointly attentive spatial-temporal pooling networks for video-based person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV'17)*. 4733–4742.

[40] Shuanglin Yan, Hao Tang, Liyan Zhang, and Jinhui Tang. 2022. Image-specific information suppression and implicit local alignment for text-based person search. *arXiv preprint arXiv:2208.14365* (2022).

[41] Jinrui Yang, Wei-Shi Zheng, Qize Yang, Ying-Cong Chen, and Qi Tian. 2020. Spatial-temporal graph convolutional network for video-based person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'20)*. 3289–3299.

[42] Mouxing Yang, Zhenyu Huang, Peng Hu, Taihao Li, Jiancheng Lv, and Xi Peng. 2022. Learning with twin noisy labels for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'22)*. 14308–14317.

[43] Mang Ye, Weijian Ruan, Bo Du, and Mike Zheng Shou. 2021. Channel augmented joint learning for visible-infrared recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV'21)*. 13567–13576.

[44] Mang Ye, Jianbing Shen, David J. Crandall, Ling Shao, and Jiebo Luo. 2020. Dynamic dual-attentive aggregation learning for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF European Conference on Computer*

Vision (ECCV'20). Springer, 229–247.

[45] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. 2021. Deep learning for person re-identification: A survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 44, 6 (2021), 2872–2893.

[46] Zican Zha, Hao Tang, Yunlian Sun, and Jinhui Tang. 2023. Boosting few-shot fine-grained recognition with background suppression and foreground alignment. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)* (2023).

[47] Qiang Zhang, Changzhou Lai, Jianan Liu, Nianchang Huang, and Jungong Han. 2022. Fmcnet: Feature-level modality compensation for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'22)*. 7349–7358.

[48] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, Xin Jin, and Zhibo Chen. 2020. Relation-aware global attention for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'20)*. 3186–3195.

[49] Kecheng Zheng, Cuiling Lan, Wenjun Zeng, Jiawei Liu, Zhizheng Zhang, and Zheng-Jun Zha. 2021. Pose-guided feature learning with knowledge distillation for occluded person re-identification. In *Proceedings of the 29th ACM International Conference on Multimedia (MM'21)*. 4537–4545.

[50] Liang Zheng, Hengheng Zhang, Shaoyan Sun, Manmohan Chandraker, Yi Yang, and Qi Tian. 2017. Person re-identification in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'17)*. 1367–1376.