Intermediary-Guided Bidirectional Spatial–Temporal Aggregation Network for Video-Based Visible-Infrared Person Re-Identification

Huafeng Li, Minghui Liu, Zhanxuan Hu, Feiping Nie, *Senior Member, IEEE*, and Zhengtao Yu

***Abstract*— This work focuses on the task of Video-based Visible-Infrared Person Re-Identification, a promising technique for achieving 24-hour surveillance systems. Two main issues in this field are modality discrepancy mitigating and spatial–temporal information mining. In this work, we propose a novel method, named Intermediary-guided Bidirectional spatial–temporal Aggregation Network (IBAN), to address both issues at once. Specifically, IBAN is designed to learn modality-irrelevant features by leveraging the anaglyph data of pedestrian images to serve as the intermediary. Furthermore, a bidirectional spatial–temporal aggregation module is introduced to exploit the spatial–temporal information of video data, while mitigating the impact of noisy image frames. Finally, we design an Easy-sample-based loss to guide the final embedding space and further improve the model's generalization performance. Extensive experiments on Video-based Visible-Infrared benchmarks show that IBAN achieves promising results and outperforms the state-of-the-art ReID methods by a large margin, improving the rank-1/mAP by 1.29%/3.46% at the Infrared to Visible situation, and by 5.04%/3.27% at the Visible to Infrared situation. The source code of the proposed method will be released at https://github.com/lhf12278/IBAN.**

***Index Terms*— Visible-infrared person re-identification, bidirectional spatial–temporal aggregation, anaglyph data, modality discrepancy.**

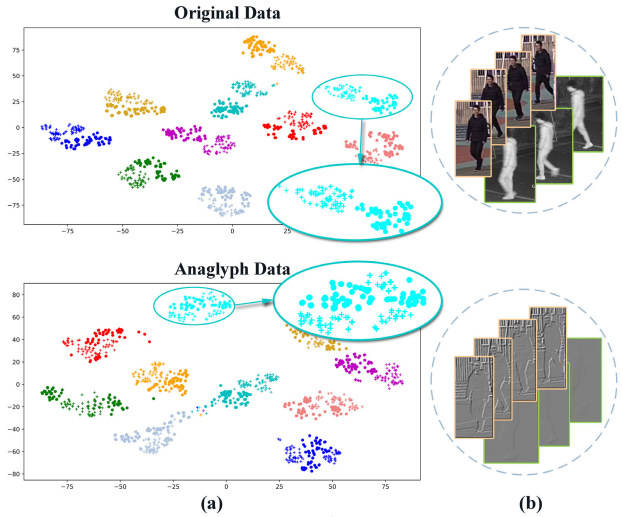


Fig. 1. Comparison between visible-infrared images and their anaglyph images. (a) T-sne visualization results in embedding space, '+' means the visible data, and '•' means the infrared data. Different colors represent different IDs. (b) Images of Visible-Infrared data and corresponding anaglyph data. The images with yellow boundaries are original visible data and their corresponding anaglyph data; The images with green boundaries are original infrared data and their corresponding anaglyph data. The anaglyph data are generated by an anaglyph operator.

I. Introduction

PERSON re-identification (ReID) aims at matching person data acquired from multiple disjoint cameras. It plays an important role in intelligent video surveillance and has seen enormous progress in recent years. The basic idea of recent studies is to seek a discriminative embedding space, in which features from the same ID have larger similarities. Most recent studies mainly focus on the image-based single-modality Re-ID [1], [2], [3], [4], [5], [6], [7], which captures the data via visible cameras and extracts discriminative features within a single image. Although impressive results have been made in recent years, there are still two obstacles that limit its application in 24-hour intelligent video surveillance. First, the visible cameras cannot capture clear data in a dark environment. Second, image-based feature extraction cannot effectively tackle some challenging issues, such as pose, appearance similarity, occlusions, and frame misalignment. An alternative solution to the above two problems is Video-based Visible-Infrared Person Re-Identification (VVI-ReID) [8], which leverages the visible and infrared data simultaneously and extracts discriminative representation within a video clip.

Two main challenges in VVI-ReID are modality discrepancy mitigating and spatial-temporal information mining. Modality discrepancy mitigating has been widely studied in Cross-Modality Person ReID. Examples include image translation-

Manuscript received 25 August 2022; revised 10 January 2023; accepted 13 February 2023. Date of publication 16 February 2023; date of current version 6 September 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 61966021, Grant 62276120, and Grant 62201453. This article was recommended by Associate Editor S. Gao. *(Huafeng Li and Minghui Liu contributed equally to this work.) (Corresponding author: Zhanxuan Hu.)*

Huafeng Li, Minghui Liu, and Zhengtao Yu are with the Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650500, China (e-mail: lhfchina99@kust.edu.cn; liuminghui@stu.kust.edu.cn).

Zhanxuan Hu is with the School of Computer Science and Technology, Xi'an University of Posts and Telecommunications, Xi'an, Shaanxi 710121, China (e-mail: huzhanxuan@mail.nwpu.edu.cn).

Feiping Nie is with the School of Artificial Intelligence, Optics and Electronics (iOPEN) and the Key Laboratory of Intelligent Interaction and Applications (Ministry of Industry and Information Technology), Northwestern Polytechnical University, Xian, Shaanxi 710072, China (e-mail: feipingnie@gmail.com).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCSVT.2023.3246091.

Digital Object Identifier 10.1109/TCSVT.2023.3246091

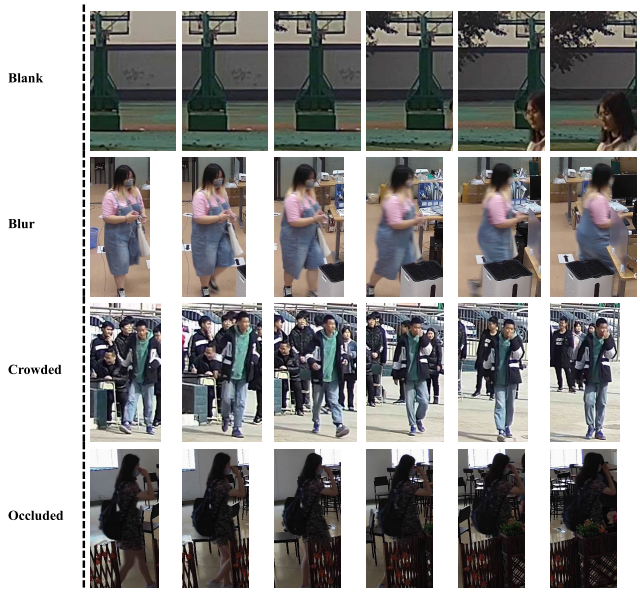


Fig. 2. Video clips with noisy frames.

based methods [9], [10], [11] and modality-invariant feature learning-based methods [12], [13], [14], [15]. Though driven by different motivations, these approaches can be thought of as mitigating modality discrepancy in different learning states. Spatial-temporal information mining is an inevitable problem in video-based vision tasks, such as ReID and action recognition [16], [17]. Approaches to this issue mainly differ in how to aggregate the features of frames in a video clip. Examples include temporal information aggregating-based methods [18], [19], [20] and complementary information mining-based methods [21], [22]. The former often regard each video clip as a directed sequence and introduces an RNN-based module or a 3D-CNN-based module to aggregate the features of all frames in order to extract temporal features. Differently, the latter aims at extracting discriminative features by exploring the interrelationship among unordered frames. Though promising results have been achieved on both Cross-Modality ReID and video-based ReID, how to solve the above two challenges simultaneously via a unified framework is still an open problem.

To this end, we present an Intermediary-guided Bidirectional spatial-temporal Aggregation Network (IBAN) for VVI-ReID. IBAN is inspired by an observation that color is one of the main elements causing the discrepancies between two modalities [12], while anaglyph images, generated from visible and infrared images, can serve as an intermediary to mitigate modality discrepancy. As shown in Figure 1, the T-sne visualization results of features in embedding space, the distribution gap between infrared features and visible features is large, while the features of heterogeneous anaglyph images with the same ID have been aggregated in embedding space. An intuitive idea is replacing the original infrared and visible images with anaglyph images. Nevertheless, the generalization performance is generally poor on unseen IDs due to the limited discriminative information of anaglyph images. From this perspective, we leverage anaglyph images to serve as an intermediary that mitigates modality discrepancy and guides the model to extract modality-irrelevant features. This mechanism differs IBAN from existing modality generation-based Cross-Modality ReID methods [23], [24], [25], where auxiliary modality is generated from only the visible images, and modality discrepancy mitigating is conducted only in the final feature space.

Furthermore, IBAN introduces a novel bidirectional spatial-temporal aggregation module to exploit the spatial-temporal information of video data, while mitigating the impact of noisy image frames. The intuition behind this module is two-fold. First, as shown in Figure2, not all frames in a video sequence are beneficial for feature learning. The effect of noisy frames with detection errors or heavy occlusions should be weakened, while the effect of informative frames with partial occlusions or pose changes should be strengthened. Second, video data is bidirectional, and both forward direction (the arranged order of video frames) and backward direction (opposite to the forward direction) can provide useful spatial-temporal information [26], [27] To this end, an adaptively weighted mechanism is used in the bidirectional spatial-temporal aggregation module, which leverages bi-directional spatial-temporal information to assign a specific weight for each frame.

Finally, we pay attention to the final embedding space. The triplet loss has been widely used to guide embedding space learning in recent ReID approaches including Cross-Modality ReID. In practice, however, its effect in mitigating modality discrepancy is limited and even inverted. Triplet loss enforces that an anchor is closer to all other samples from the same ID than it is to any sample from the other ID. That is, it considers only the ID discrepancy and ignores the modality discrepancy between different identities. Consequently, different IDs are apart from each other, and one or multiple IDs can be considered a fake sub-modality. From this perspective, we hypothesize that the triplet loss should work on the condition that all samples locate in a common embedding space. To this end, we design an easy sample-based loss to assist the triplet loss. For example, given an anchor from modality $\mathcal{M}_1$, we first seek its easy sample from another modality $\mathcal{M}_2$, i.e., a sample with the largest similarity with the anchor. Then, we attract the residual samples from the modality $\mathcal{M}_2$ to close the easy sample. Easy sample-based loss is complementary to the intermediary-guided module, and they handle modality discrepancy in different feature levels.

In practice, IBAN integrates the Intermediary-guided feature learning module, bidirectional spatial-temporal aggregation module, and easy sample-based loss into a unified framework. Extensive experiments on cross-modality ReID show that without other bells and whistles, it achieves state-of-the-art results and provides a large advantage over existing ReID methods including video-based and image-based cross-modality methods. The main contributions of this work are as follows:

- We provide a novel method IBAN for VVI-ReID, which handles modality discrepancy in different feature levels and mines spatial-temporal information from different directions.

- IBAN leverages anaglyph data of the pedestrian images to serve as the intermediary, guiding the model to extract modality-invariant features. Besides, it achieves feature aggregation of frames via exploring the bidirectional spatial-temporal information of video data, while mitigating the impact of noisy image frames.
- We conduct extensive experiments to validate the effectiveness of IBAN, and the experimental results demonstrate that it significantly outperforms state-of-the-art competitors on the cross-modality ReID task. In addition, we provide a detailed discussion of the modules of IBAN via ablation studies.

The rest of this paper is arranged as follows. We review the related works in Section II, and provide the details of IBAN in Section III. In Section IV, we report the experimental results of IBAN on the VVI-ReID task. Finally, we conclude this work and discuss the potential improvements in Section V

II. Related Work

The main challenges faced in person ReID include pose variations, occlusions, illumination changes, and large intra-class variations caused by various camera views. A large number of approaches have been developed to address these challenges [28], [29], taking a comprehensive review beyond the scope of this work. Here, we focus only on the video-based person Re-ID and visible-infrared cross-modality person Re-ID.

A. Video-Based Person ReID

Video-based person ReID aims at matching the same ID from a gallery set of video clips, which is a powerful approach for tackling the problems of pose variations and occlusions in ReID. And, efficiently exploiting spatial-temporal information while generating discriminative features is pivotal for its success. The methods for processing video data can be roughly divided into two categories: temporal information aggregating-based methods and complementary information mining-based methods. The former regards each video clip as a directed sequence and introduces a temporal information mining module to aggregate the features of all frames, such as LSTM [18], [22], [30] and 3D-CNNs [19], [20]. Differently, the latter assumes that the frames of a video clip are unordered but complementary. To mine the complementary information, it often leverages an attention mechanism [31], [32], [33] or Graph Convolutional Network (GCN) [21], [34], [35] to explore the interrelationship among different frames. In practice, RNN-based modules are beneficial for learning temporal features, but they always neglect the importance of spatial clues. The networks used in 3D-CNN and GCN are often complex, computationally expensive, and difficult to optimize, which limits their application in tackling real-world challenges.

In addition, a number of temporal information mining methods have been proposed for other video-based vision tasks, such as action recognition tasks [16], [17], [36], [37]. However, all of them ignore two significant issues. First, video data is bidirectional, and both the forward direction (the arranged order of video frames) and backward direction (opposite to the forward direction) can provide useful spatial-temporal information. Second, not all frames in a video clip are equally beneficial for feature learning. The bidirectional problem is first exploited in [26], and then used in [27]. But, both of them ignore the frame difference of the video clip. In this work, we address these two issues simultaneously using a bidirectional spatial-temporal aggregation module, which exploits the spatial-temporal information of video data, while mitigating the impact of noisy image frames.

B. Visible-Infrared Cross-Modality Person ReID

Visible-infrared cross-modality person Re-ID [38] aims to achieve person matching across two different modalities, which can significantly improve the performance of the model in poor illumination environments. Meanwhile, it takes the researchers an additional issue, i.e., modality discrepancy mitigating. As mentioned above, the methods for this issue mainly include image translation-based methods [9], [10], [39] and modality-invariant feature learning-based methods [14], [15], [40], [41]. The former aims at mitigating modality discrepancy at the image level via generating data using Generative Adversarial Network (GAN) or other approaches. AlignGAN [10] is one of the pioneers, which generates fake-infrared images by generators and feeds them into a weight-share network with real-infrared images. Meanwhile, a discriminator is used to distinguish whether the input infrared images are real or not. Through the above min-max two-player game, sufficiently realistic fake-infrared images can be generated to mitigate the modality discrepancy. To make full use of the data from both of the modalities [11], several multi-generation-based methods have been developed. For example, GECNet [42] uses the gray data generated by visible data to enhance the infrared data, and then, fake visible data is generated using enhanced infrared data through a colorization Siamese GAN. In this way, the connection between visible data and infrared data is enhanced in the three times conversion-generation processes. Besides, considering that the inconsistent convergence direction between the generator and discriminator will lead to optimization puzzles, several methods propose to replace generation with transformation, transforming original images to a more appropriate media modality. For example, [42] employs gray images to fully replace visible images and infrared images for feature learning so that useful information can be retained furthest. Reference [25] mitigates modality discrepancy in the final feature space and pushes the features of heterogeneous images to close the syncretic features generated via the syncretic modality generative module. Differently, modality-invariant feature learning methods reduce the distribution gap between different modalities in feature space. For example, Ye et al. [43] leverage an elaborate dual-stream network to map features of different modalities into a common higher-dimensional embedding space. Shuai et al. [44] align the features of different modalities by combining local features and global features. Fu et al. [45] introduce a dual alignment

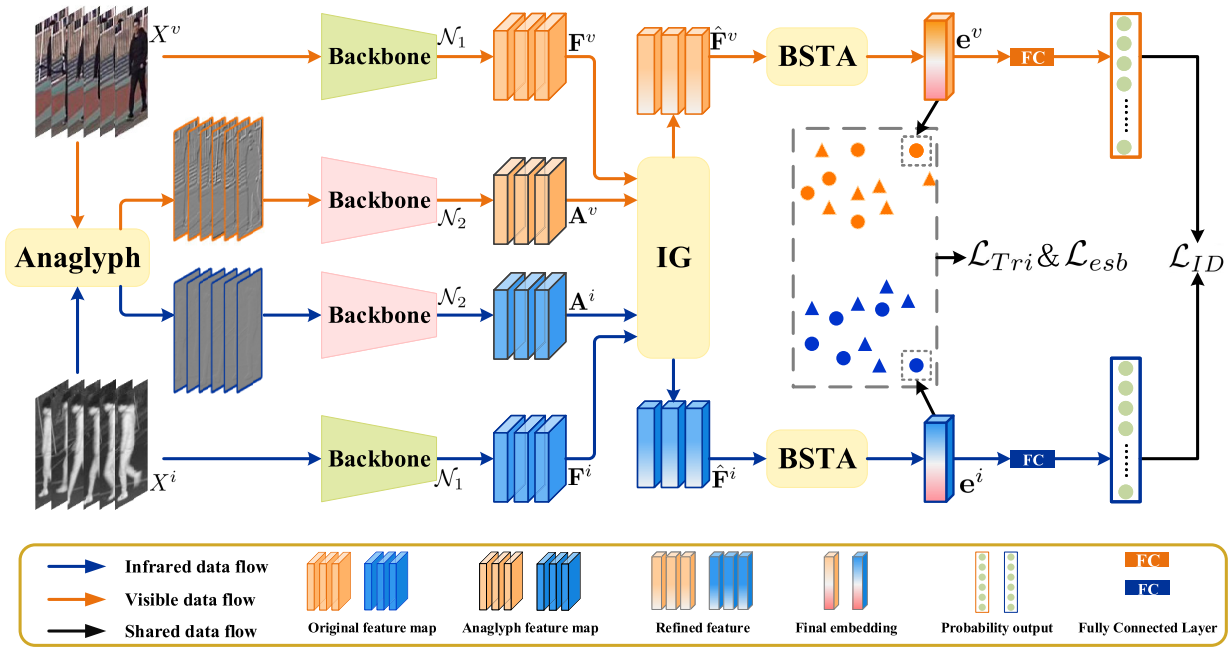


Fig. 3. An overview of our IBAN framework. First, we leverage an Intermediary-Guided module (IG) to mitigate the modality discrepancy. Second, we design a Bidirectional Spatial-Temporal Aggregation module (BSTA) to mine spatial-temporal information of video data. Finally, we introduce an additional Easy-Sample-Based (ESB) loss to constrain the final embedding space.

network(DAN) to solve the visible-infrared cross-modal cross-domain person Re-ID problem. Wang et al. [46] embed modality mitigation modules in a two-stream network to map different modal features to the same space. Meanwhile, some methods propose to mitigate modality discrepancy by strengthening the higher-order relations that are robust to modality variation, such as the structure of the human body [47].

It is worth noting that the above two approaches are not mutually exclusive, and many cross-modality works tend to combine them for performance improvement. Reference [48] leverages image transformation to mitigate the modality discrepancy of input data and design a shared-backbone-based network to align features. Reference [49] alleviates the modal differences via pixel-level mixing while adjusting the relative distance across multi-modality through a Dynamic Center Aggregation(DCA) loss. The method proposed in this paper also follows this idea, but differs from existing approaches [25], [42], we mitigate the modality discrepancy in different feature levels. The details will be presented in the next section.

III. Proposed Method

A. Overview

In this work, we propose a novel method, named Intermediary-guided Bidirectional spatial-temporal Aggregation Network (IBAN), to tackle the issues of modality discrepancy mitigating and spatial-temporal information mining in a unified framework. An overall pipeline of IBAN is presented in Figure 3, where the backbone denotes a dual-stream widely used in previous Visible-Infrared ReID [8]. Algorithm 1 describes the procedure of IBAN in detail. In the next subsection, we will elaborate on each step of IBAN.

Algorithm 1 An Overview to Our Proposed Method IBAN

Input: A mini-batch that consists of b visible video clips $\{X^v\}_b$ and infrared video clips $\{X^i\}_b$

while *Not Converged* **do**

Step.1: #Anaglyph data generation;

$A^v = \mathcal{A}(X^v)$ and $A^i = \mathcal{A}(X^i)$;

Step.2: #Feature extraction;

$\mathbf{F}^v = \mathcal{N}_1(X^v)$,$\mathbf{F}^i = \mathcal{N}_1(X^i)$;

$\mathbf{A}^v = \mathcal{N}_2(A^v)$,$\mathbf{A}^i = \mathcal{N}_2(A^i)$;

Step.3: #Feature discrepancy mitigation;

$\hat{\mathbf{F}}^v = \mathrm{IG}(\mathbf{F}^v, \mathbf{A}^v, \mathbf{A}^i)$, $\hat{\mathbf{F}}^i = \mathrm{IG}(\mathbf{F}^i, \mathbf{A}^i, \mathbf{A}^v)$;

Step.4: #Spatial-temporal aggregation;

$\mathbf{e}^v = \mathrm{BSTA}(\hat{\mathbf{F}}^v)$, $\mathbf{e}^i = \mathrm{BSTA}(\hat{\mathbf{F}}^i)$;

Step.5: #Calculate loss and BP;

$\mathcal{L} = \mathcal{L}_{ID} + \lambda_1\mathcal{L}_{Tri} + \lambda_2\mathcal{L}_{esb} + \lambda_3\mathcal{L}_{cr}$.;

Update the model using backpropagation.

end

B. IBAN Framework

Step#1 Anaglyph Data Generation: Given two cross-modality video clips $X^v = \{x_1^v, x_2^v, \ldots, x_k^v\}$ and $X^i = \{x_1^i, x_2^i, \ldots, x_k^i\}$, IBAN first generates their anaglyph data $A^v = \{a_1^v, a_2^v, \ldots, a_k^v\}$ and $A^i = \{a_1^i, a_2^i, \ldots, a_k^i\}$ via an anaglyph operator $\mathcal{A}$. Specifically, for each frame a^v or a^i, we have

$$a(i, j) = \sum_m \sum_n x(i + m, j + n)\mathcal{A}(m, n) + k, \tag{1}$$

where $\mathcal{A}$ denotes a edge detection operator, $a(i, j)$ denotes the (i, j)-th pixel of anaglyph image a, k is an offset value.

Step#2 Feature Extraction: Subsequently, we extract the features of both visible-infrared data and anaglyph data via two

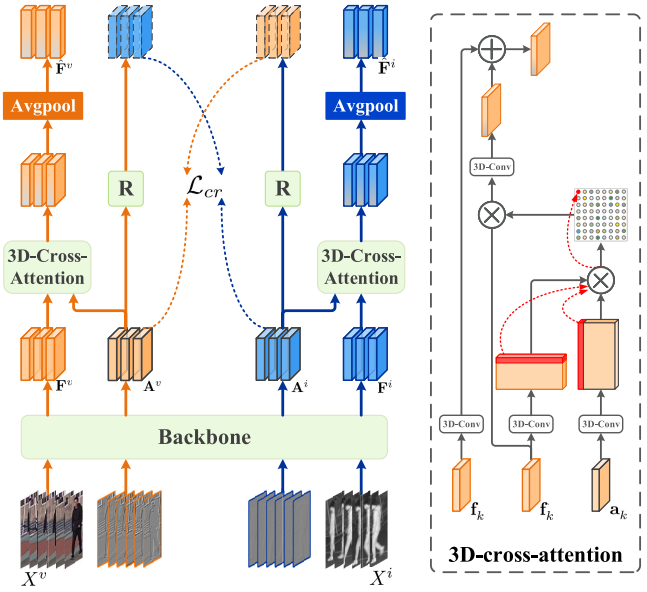


Fig. 4. An overall pipeline of IG module in IBAN.

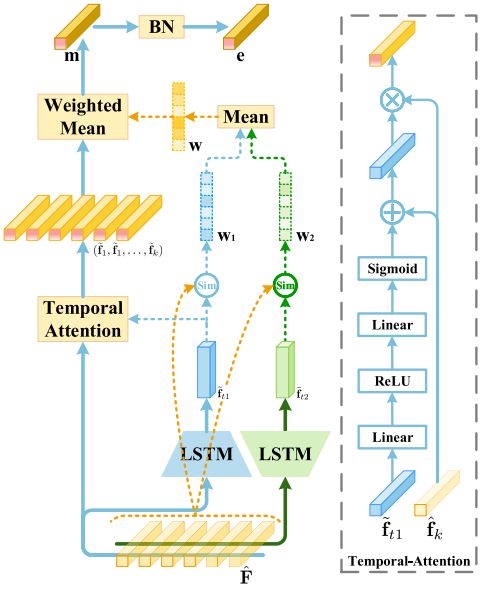


Fig. 5. An overall pipeline of BSTA module in IBAN.

parallel part-shared dual-stream networks $\mathcal{N}_1$ and $\mathcal{N}_2$. That is,

$$\mathbf{F}^v = \mathcal{N}_1(X^v), \mathbf{F}^i = \mathcal{N}_1(X^i), \tag{2}$$

and

$$\mathbf{A}^v = \mathcal{N}_2(A^v), \mathbf{A}^i = \mathcal{N}_2(A^i) \tag{3}$$

where $\mathbf{F}^v$ ($\mathbf{F}^i$) and $\mathbf{A}^v$ ($\mathbf{A}^i$) are extracted features of original data and anaglyph data respectively. Then we propose to leverage an Intermediary-Guided module (IG) to mitigate the modality discrepancy (between $\mathbf{F}^v$ and $\mathbf{F}^i$).

Step#3 Modality Feature Discrepancy Mitigation Using IG: Inspired by the learning behavior of humans, we propose to mitigate the modality discrepancy by setting a proper reference for the network. Figure 1 shows that the anaglyph data has two characteristics compared with the original Visible-Infrared data. First, the modality discrepancy of the anaglyph data is small. Second, anaglyph data loses some discriminative information. Hence, the basic idea of Intermediary-guided module (IG) is leveraging the features of anaglyph images ($\mathbf{A}^v$, $\mathbf{A}^i$) to refine the features of visible-infrared images ($\mathbf{F}^v$, $\mathbf{F}^i$), reducing the gap between $\mathbf{F}^v$ and $\mathbf{F}^i$. Figure 4 illustrates the overall pipeline of the IG module. Next, we provide the details of IG. Suppose $\mathbf{F}^v = \{\mathbf{f}_1^v, \mathbf{f}_2^v, \ldots, \mathbf{f}_k^v\}$ and $\mathbf{F}^i = \{\mathbf{f}_1^i, \mathbf{f}_2^i, \ldots, \mathbf{f}_k^i\}$ are feature maps of the original video clips X^v and X^i, respectively, where k is the number of frames in each video clip. Correspondingly, $\mathbf{A}^v = \{\mathbf{a}_1^v, \mathbf{a}_2^v, \ldots, \mathbf{a}_k^v\}$ and $\mathbf{A}^i = \{\mathbf{a}_1^i, \mathbf{a}_2^i, \ldots, \mathbf{a}_k^i\}$ are feature maps of the anaglyph data. IG uses the 3D Cross-Attention (CA) mechanism to strengthen the similar information between $\mathbf{F}^v(\mathbf{F}^i)$ and $\mathbf{A}^v(\mathbf{A}^i)$. Specifically, for each $\mathbf{f}_k$ we have

$$\begin{aligned}\hat{\mathbf{f}}_k &= \mathrm{CA}(\mathbf{f}_k, \mathbf{a}_k) \\ &= \mathbf{W}(\mathrm{Softmax}(\mathbf{f}_k \cdot \mathbf{a}_k^T) \otimes \mathbf{a}_k) + \mathbf{f}_k,\end{aligned} \tag{4}$$

where $\otimes$ is dot product, $\mathbf{W}$ are learnable parameters, and $\hat{\mathbf{f}}_k$ denotes the refined feature map that contains modality-irrelevant information of $\mathbf{f}_k$. Note that we omit the superscript v and i, which demonstrates that IG works on both $\mathbf{f}^v$ and $\mathbf{f}^i$. Further, we use average pooling to translate each feature map as a vector.[1]

A potential problem of Eq. (4) is that the modality-irrelevant feature learning procedures of different modalities are separate, and it does not consider the modality discrepancy between anaglyph images with partial modal-specific information. To this end, we introduce a cross-reconstruction constraint to reduce the discrepancy between $\mathbf{A}^i$ and $\mathbf{A}^v$. Specifically, given a pair of features $\mathbf{a}_k^v$ and $\mathbf{a}_k^i$ from the same ID, we expect that $\mathbf{a}_k^v$ and $\mathbf{a}_k^i$ can reconstruct each other. Mathematically, the cross-reconstruction constraint is:

$$\mathcal{L}_{cr} = \sum_{t=1}^{b\times k} \|\mathbf{a}_t^v - \mathcal{R}(\mathbf{a}_t^i)\|_2 + \sum_{t=1}^{b\times k} \|\mathbf{a}_t^i - \mathcal{R}(\mathbf{a}_t^v)\|_2, \tag{5}$$

where b is the size of the mini-batch, and $\mathcal{R}$ denotes the reconstruction network that consists of a set of convolution layers. During training, the gradients of the original features $\mathbf{a}_t^v$ and $\mathbf{a}_t^i$ are detached. Subsequently, we input the refined features $\hat{\mathbf{F}}^v$ and $\hat{\mathbf{F}}^i$ into a Bidirectional Spatial-Temporal Aggregation (BSTA) module, obtaining the final embedding of each video clip.

Step#4 Bidirectional Spatial-Temporal Aggregation Using BSTA: An advantage of VVI-ReID is that the video data can provide rich temporal information(e.g., pedestrian gait characteristics), enhancing the robustness of the appearance feature. Suppose $\hat{\mathbf{F}} = \{\hat{\mathbf{f}}_1, \hat{\mathbf{f}}_2, \ldots, \hat{\mathbf{f}}_k\}$ is the refined feature of a video clip.[2] Existing video-based ReID methods often first use an LSTM network to extract its temporal information, and then introduce a Temporal-Attention (TA) module to aggregate them. Finally, a mean operation is conducted to obtain the

[1]To avoid confusability of notation, we also use $\hat{\mathbf{f}}_k$ to denote the generated feature vector.

[2]If not noted otherwise, we omit the superscript v and i in the rest of this work.

embedding vector. Specifically,

$$\begin{aligned}\tilde{\mathbf{f}}_t &= \text{LSTM}(\hat{\mathbf{f}}_1, \hat{\mathbf{f}}_2, \ldots, \hat{\mathbf{f}}_k)\\ \tilde{\mathbf{f}}_k &= \text{TA}(\tilde{\mathbf{f}}_t, \hat{\mathbf{f}}_k) = \text{FC}(\tilde{\mathbf{f}}_t) \otimes \hat{\mathbf{f}}_k + \hat{\mathbf{f}}_k\\ \mathbf{m} &= \text{mean}(\tilde{\mathbf{f}}_1, \tilde{\mathbf{f}}_2, \ldots, \tilde{\mathbf{f}}_k),\end{aligned} \tag{6}$$

where $\tilde{\mathbf{f}}_t$ denotes a sequence-level temporal feature, $\tilde{\mathbf{f}}_k$ denotes a frame-level spatial-temporal aggregation feature, $\mathbf{m}$ denotes a sequence-level spatial-temporal aggregation feature, and FC is a fully connected network that consists of a series of ReLU and Sigmoid layers.

Although such a mechanism is useful for mining and aggregating temporal information of data, it ignores two significant issues. First, video data is bidirectional, and both forward direction (the arranged order of video frames) and backward direction (opposite to the forward direction) can provide useful spatial-temporal information. Second, not all frames in a video clip are equally beneficial for feature learning. To this end, in this work, we propose a Bidirectional Spatial-Temporal Aggregation module (BSTA). An overall pipeline of BSTA is presented in Figure 5. Specifically, to address the first issue, we propose to extract the bidirectional spatial-temporal information of data. That is,

$$\begin{aligned}\tilde{\mathbf{f}}_{t1} &= \text{LSTM}(\hat{\mathbf{f}}_1, \hat{\mathbf{f}}_2, \ldots, \hat{\mathbf{f}}_k)\\ \tilde{\mathbf{f}}_{t2} &= \text{LSTM}(\hat{\mathbf{f}}_k, \hat{\mathbf{f}}_{k-1}, \ldots, \hat{\mathbf{f}}_1)\end{aligned} \tag{7}$$

where $\tilde{\mathbf{f}}_{t1}$ denotes the forward temporal feature and $\tilde{\mathbf{f}}_{t2}$ denotes the backward temporal feature. To further address the second issue, we leverage $\tilde{\mathbf{f}}_{t1}$ and $\tilde{\mathbf{f}}_{t2}$ to assign a specific weight for each frame. Specifically,

$$\begin{aligned}\mathbf{W}_1 &= (\cos(\hat{\mathbf{f}}_1, \tilde{\mathbf{f}}_{t1}), \cos(\hat{\mathbf{f}}_2, \tilde{\mathbf{f}}_{t1}), \ldots, \cos(\hat{\mathbf{f}}_k, \tilde{\mathbf{f}}_{t1}))\\ \mathbf{W}_2 &= (\cos(\hat{\mathbf{f}}_1, \tilde{\mathbf{f}}_{t2}), \cos(\hat{\mathbf{f}}_2, \tilde{\mathbf{f}}_{t2}), \ldots, \cos(\hat{\mathbf{f}}_k, \tilde{\mathbf{f}}_{t2}))\\ \mathbf{W} &= \frac{\mathbf{W}_1 + \mathbf{W}_2}{2},\end{aligned} \tag{8}$$

where $\cos(\bullet)$ denotes the cos similarity function. Both $\cos(\hat{\mathbf{f}}_k, \tilde{\mathbf{f}}_{t1})$ and $\cos(\hat{\mathbf{f}}_k, \tilde{\mathbf{f}}_{t2})$ reflect the importance of the k-th frame. The final weighted vector $\mathbf{W} = \{w_1, w_2, \ldots, w_k\}$ is the mean of $\mathbf{W}_1$ and $\mathbf{W}_2$. Subsequently, we use the same TA module to aggregate the spatial-temporal information and generate a series of frame-level spatial-temporal aggregation features $\{\tilde{\mathbf{f}}_1, \tilde{\mathbf{f}}_1, \ldots, \tilde{\mathbf{f}}_k\}$. The final sequence-level embedding vector of a video clip is obtained by using a weighted mean mechanism, i.e.,

$$\mathbf{m} = \frac{w_1\tilde{\mathbf{f}}_1 + w_2\tilde{\mathbf{f}}_2 + \ldots, w_k\tilde{\mathbf{f}}_k}{k}. \tag{9}$$

Finally, BSTA introduces a Batch Normalization (BN) layer to normalize the embedding vectors of the current mini-batch of size b, and obtain the final embedding vectors:

$$\mathbf{E} = \text{BN}(\mathbf{M}), \tag{10}$$

where $\mathbf{E} = (\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_b)$ and $\mathbf{M} = (\mathbf{m}_1, \mathbf{m}_2, \ldots, \mathbf{m}_b)$. And, $\mathbf{E}$ is guided by an objective function discussed below.

Step#5 Loss Calculation and Back Propagation (BP): In general, triplet loss is used to guide the network to learn discriminative embedding. It maximizes the distance between anchor $\mathbf{e}^a$ and its negative instance $\mathbf{e}^n$, both of which have different IDs, and minimizes the distance between the anchor and its positive instance $\mathbf{e}^p$, both of which have the same ID. Mathematically, it is formulated as:

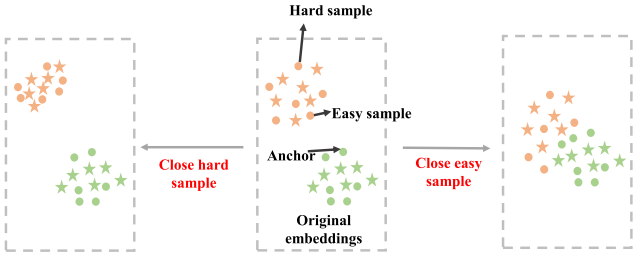


Fig. 6. Motivation of using Easy-sample-based (ESB) loss rather than Hard-sample-based (HSB) loss. Here, color represents modality, while shape represents ID.

$$\mathcal{L}_{Tri} = \frac{1}{b}\sum_{i=1}^{b}[0, m + \|\mathbf{e}_i^a - \mathbf{e}^p\|_2 - \|\mathbf{e}_i^a - \mathbf{e}^n\|_2]_+ \tag{11}$$

where m is a predefined margin. Although triplet loss has achieved great success in existing person ReID methods, including cross-modality ReID, we find that it considers only the classification differences and ignores the modality discrepancies between different identities. Consequently, the negative samples are often apart from the anchor, causing modality separation between different IDs. That is, many sub-modalities between different IDs will be generated. For labels of the training data and test data are unshared, sub-modalities cannot maintain identity consistency during the testing phase. For example, the sub-modality of person A may be aligned with the sub-modality of person B. As a result, the performance of the model is degraded in the test data.

Furthermore, in order to make the network learn better embeddings, previous works propose to use hard-positive and hard-negative samples only for relationship mining and neglect the fact that the easy samples are most suitable for cross-modality retrieval tasks and contain the worthiest information to be learned. To this end, in this work, we introduce an additional Easy-Sample-Based (ESB) loss to assist the triplet loss. Particularly, we expect that the ESB loss can reduce the modality gap between different IDs, while triplet loss can improve the discrimination of the embeddings.

As shown in Figure 6, given an anchor from modality $\mathcal{M}_1$, ESB first seeks its easy sample from another modality $\mathcal{M}_2$, i.e., a sample with the largest similarity with the anchor,[3] and then attracts the residual samples from the modality $\mathcal{M}_2$ to close the easy sample. Obviously, the easy sample has minimal modality discrepancy with heterogeneous data, while the hard sample has maximum modality discrepancy. And, enforcing the residual samples from $\mathcal{M}_2$ to close the hard samples will enlarge the modality discrepancy. Besides, a potential problem is that the hard sample may be outliers. Mathematically, ESB loss is formulated as:

$$\mathcal{L}_{esb} = \sum_{b_1=1}^{b/2}\sum_{b_2=1}^{b/2}\text{mse}(\mathbf{e}_{b_2}^v, \mathcal{E}(\mathbf{e}_{b_1}^i)) + \sum_{b_2=1}^{b/2}\sum_{b_1=1}^{b/2}\text{mse}(\mathbf{e}_{b_1}^i, \mathcal{E}(\mathbf{e}_{b_2}^v)), \tag{12}$$

[3]We ignore the ID differences between instance.

where $\mathcal{E}(\mathbf{e}^i_{b_1})$ denotes the easy sample of $\mathbf{e}^i_{b_1}$, sampled from the visible modality, and vice versa. That is,

$$\mathcal{E}(\mathbf{e}^i_{b_1}) = \text{argmax}\cos(\mathbf{e}^v_t, \mathbf{e}^i_{b_1}), t = 1, 2, \ldots, b/2, \quad (13)$$

and

$$\mathcal{E}(\mathbf{e}^v_{b_2}) = \text{argmax}\cos(\mathbf{e}^i_t, \mathbf{e}^v_{b_2}), t = 1, 2, \ldots, b/2, \quad (14)$$

The gradient of $\mathcal{E}(\mathbf{e})$ and $\mathcal{E}(\mathbf{e})$ are detached during training. ESB loss breaks the modal barriers between different IDs, and assists the triplet loss to learn modality-irrelevant and discriminative embedding vectors. In other words, it is complementary to the intermediary-guided module, and they handle modality discrepancy in different feature levels.

Taking all modules mentioned above into consideration, we obtain the overall loss of IBAN, as follows:

$$\mathcal{L} = \mathcal{L}_{ID} + \lambda_1\mathcal{L}_{Tri} + \lambda_2\mathcal{L}_{esb} + \lambda_3\mathcal{L}_{cr}, \quad (15)$$

where λ_1, λ_2 and λ_3 are hyper-parameters, $\mathcal{L}_{ID}$ is the cross-entropy loss. During training $\mathcal{L}_{ID}$, $\mathcal{L}_{Tri}$ and $\mathcal{L}_{esb}$ work on the visible data, infrared data, and anaglyph data, $\mathcal{L}_{cr}$ works only on the anaglyph data. Besides, more experiments of hyper-parameters analysis can be found in Fig. 7.

IV. Experiments

A. Dataset and Evaluation Protocols

In this work, all experiments are conducted on HITSZ-VCM [8], a unique video-based cross-modality re-identification dataset. HITSZ-VCM includes 251452 RGB images and 211807 IR images of 927 IDs captured by 12 non-overlapped cameras. Each track of a person consists of 24 consecutive frames, and all images are divided into 11785 RGB tracks and 10078 IR tracks, respectively. We follow [8] and divide the VCM-HIT into two parts: a training set and a test set. The training set contains 232, 496 images of 500 IDs and 11, 061 tracks, while the test set contains 230, 763 images of 427 IDs and 10, 802 tracks. Before training, we resize all images to the size of 288 × 144 and use random cropping and random flipping to achieve data augmentation.

Besides, we use rank-k and mAP (mean Average Precision)[4] to evaluate the performance of IBAN, both of them are widely used in cross-modality person re-identification. Suppose $p(k)$ is the precision at cut-off k in the returned list, $I(k)$ is an indicator function which is equal to 1 if the image at rank k is a correct image, zero otherwise. We have

$$\text{AP} = \frac{\sum_{k=1}^{n} p(k) * I(k)}{\text{number of correct images}} \quad (16)$$

where n denotes the number of retrieved images. Further, the mAP is defined as:

$$\text{mAP} = \frac{1}{q}\sum_{i=1}^{q}\text{AP}_i, \quad (17)$$

where q denotes the total number of queries.

[4]https://en.wikipedia.org/wiki/Evaluation_measures_(information_retrieval)

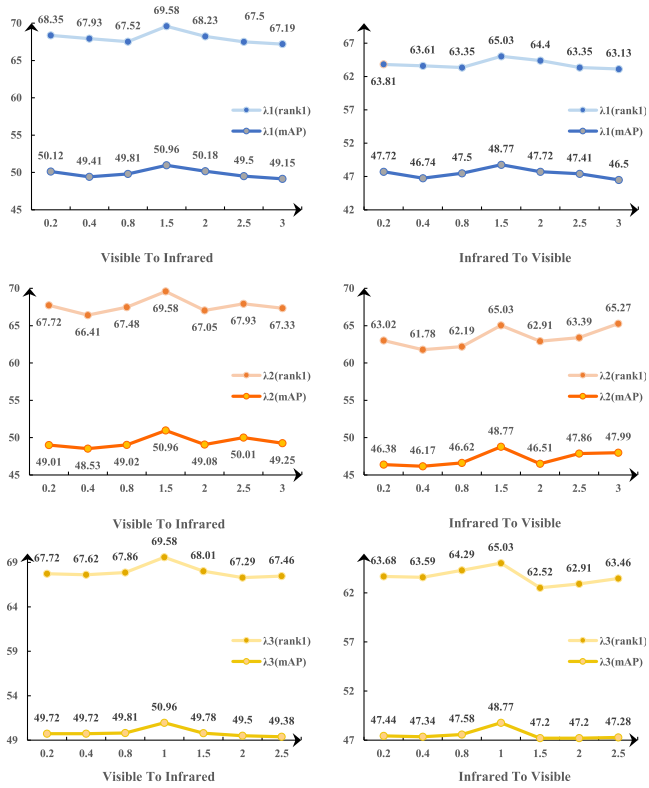


Fig. 7. Hyper-parameters Analysis of IBAN.

B. Experiment Setup

All experiments are implemented on a single RTX3090 GPU. Optimization is performed using an SGD optimizer. The learning rate is initialized as 0.1 and adjusted linearly by the warrior-up strategy during the first 10 epochs. Besides, we decrease it to 0.01 and 0.001 at 35-th and 80-th epochs, respectively. The weight decay is 0.0005, and batch size is 16 (each mini-batch includes 4 IDs and each ID includes 4 video clips of 2 modalities). Each training is run over 200 epochs.

C. Hyper-Parameters Setup

There are three main hyper-parameters in IBAN, i.e., λ_1, λ_2, and λ_3. In practice, we first determine their variation range by observing the values of corresponding loss terms. Second, we select the hyper-parameters via grid search. Besides, we find that the network mainly focuses on the shallow features of data (such as environment, posture, clothing, and other large-area features) at the early training stage. Hence, the reconstruction loss used in the IG module will prevent the network from extracting discriminative features, resulting in the collapse of network performance. To avoid this issue, we introduce the reconstruction loss in the interim of training. Experimental results demonstrate that the performance of IBAN is not sensitive to the epoch of introducing the reconstruction loss. Hence, we fix it as 40. That is, $\lambda_3 = 0$ when epoch$<$ 40. In Figure7, we report the performance of IBAN with different parameter configurations of λ_1, λ_2, and λ_3. We observe that the trends of visible-to-infrared and infrared-to-visible are consistent on all hyper-parameters.

TABLE I
COMPARISON WITH SEVERAL STATE-OF-THE-ART PERSON ReID METHODS ON HITSZ-VCM

Method	Venue	Infrared to Visible					Visible to Infrared				
		rank-1	rank-5	rank-10	rank-20	mAP	rank-1	rank-5	rank-10	rank-20	mAP
TCLNet	ECCV'20	48.32	64.44	72.03	78.34	36.38	52.13	68.60	75.29	80.74	38.52
DDAG	ECCV'20	54.62	69.79	76.05	81.50	39.26	59.03	74.64	79.53	84.04	41.50
LbA	ICCV'21	46.38	65.29	72.23	79.41	30.69	49.30	69.27	75.90	82.21	32.38
MPANet	CVPR'21	46.51	63.07	70.51	77.77	35.26	50.32	67.31	73.56	79.66	37.80
VSD	CVPR'21	54.53	70.01	76.28	82.01	41.18	57.52	73.66	79.38	83.61	43.45
CAJL	ICCV'21	56.59	73.49	79.52	84.05	41.49	60.13	74.62	79.86	84.53	42.81
MITML	CVPR'22	63.74	76.88	81.72	86.28	45.31	64.54	78.96	82.98	87.10	47.69
IBAN* (ours)	∼	**62.70**	**76.27**	**81.65**	**85.41**	**47.06**	**67.05**	**79.70**	**83.86**	**87.83**	**49.07**
IBAN (ours)	∼	**65.03**	**78.34**	**82.98**	**87.19**	**48.77**	**69.58**	**81.51**	**85.43**	**88.78**	**50.96**

TABLE II
ABLATION STUDY OF THE COMPONENTS IN IBAN

Method	Infrared to Visible					Visible to Infrared				
	rank-1	rank-5	rank-10	rank-20	mAP	rank-1	rank-5	rank-10	rank-20	mAP
Baseline	58.05	72.86	78.80	83.49	43.14	63.46	76.88	81.94	86.47	45.45
Baseline+IG^{-r}	59.27	73.65	79.06	83.62	44.14	65.29	78.60	82.90	87.12	47.09
Baseline+IG	61.04	75.04	80.41	85.10	45.61	66.33	79.00	83.19	87.08	47.97
Baseline+BSTA	59.71	74.56	80.06	84.90	43.27	65.15	78.21	83.02	87.29	45.57
Baseline+ESB	61.65	75.04	80.04	84.53	44.93	65.54	78.51	83.51	87.49	47.43
Baseline+IG+ESB+BSTA^{-b}	64.20	77.07	81.89	86.15	47.89	68.39	81.09	85.10	88.06	50.18
Baseline+IG+ESB+BSTA^{-l}	61.63	75.68	81.06	85.93	45.50	66.23	79.70	84.11	87.96	47.85
Baseline+IG+ESB+BSTA^{-a}	63.55	76.66	82.00	86.41	47.29	67.41	79.47	83.84	87.25	49.16
Baseline+IG+ESB+BSTA	65.03	78.34	82.98	87.19	48.77	69.58	81.51	85.43	88.78	50.96

Besides, the gap between maximum and minimum maintains at 0.2 ∼ 2.5, which demonstrates the robustness of IBAN to parameter selection. Based on these results, we fix $\lambda_1 = 1.5$, $\lambda_2 = 1.5$, and $\lambda_3 = 1$.

D. Comparison With Several Representative ReID Methods

To validate the effectiveness of IBAN, we compare it with several recent person ReID methods, including 5 cross-modality person ReID methods (LbA [50], MPANet [51], DDAG [52], VSD [53], CAJL [54]), 1 video-based person ReID method (TCLNet [55]), and the unique video-based Visible-Infrared Person Re-Identification method (MITML [8]). Note that we use an average pooling layer for all image-based visible-infrared cross-modal person Re-ID methods. We also report the results of IBAN without the temporal information mining module(BSTA) for a fair comparison with image-based cross-modality ReID methods.

The comparison results are reported in Table I, where IBAN* represents the model of IBAN without BSTA module. Observing Table I, we can find that both MITML and IBAN outperform other methods significantly, which demonstrates the advantage of using video-based Visible-Infrared Person Re-Identification. Further, IBAN improves MITML by a large margin, 5.04% Rank-1 and 3.27% mAP at the Visible to Infrared setting. Such a result demonstrates the effectiveness of IBAN. Next, we conduct an ablation study to validate the components of IBAN.

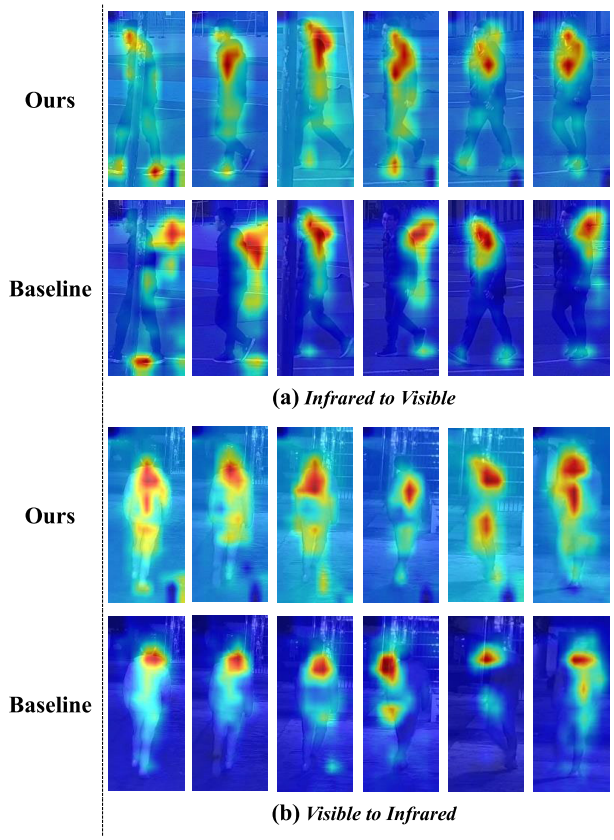


Fig. 8. CAM heat maps comparison between baseline and IBAN.

E. Ablation Study

The ablation study is based on a baseline model, a simple two-stream network. As mentioned above, we improve the baseline model by introducing an IG module for modality discrepancy mitigating, a BSTA module for spatial-temporal information mining, and an additional ESB loss to guide the

Fig. 9. ReID results of IBAN with different modules. 'B' denotes the baseline model. 'B+X' denotes the improved method with module 'X'.

final embedding space. Besides, we build an IG^{-r} module that removes the reconstruction part from IG, a BSTA^{-b} module that replaces the bidirectional weighted mechanism with a unidirectional weighted mechanism, a BSTA^{-a} module that removes the attention mechanism, and a BSTA^{-l} module that replaces LSTM with average pooling. The experimental results are shown in Table II. It reveals that all components are useful for improving the performance of IBAN. Further, IG contributes most to the performance and improves the accuracy of rank-1/mAP by 2.99%/2.47% in infrared-to-visible retrieval situations, and by 2.87%/2.52% in visible-to-infrared retrieval situations. Meanwhile, BSTA has the minimum impact, which improves the accuracy of rank-1/mAP by 1.66%/0.13% in infrared-to-visible retrieval situations, and by 1.69%/0.12% in visible-to-infrared retrieval situations. The advantage of BSTA over BSTA^{-b} supports our hypothesis that the benefits of BSTA lie in leveraging bidirectional spatial-temporal information of video data. Besides, the advantage of IG over IG^{-} supports our hypothesis that the cross-reconstruction module is beneficial for learning discriminative embeddings.

F. Visualization

To provide a deep analysis of IBAN, we compare it with the baseline using CAM heat maps. The experimental results, reported in Figure 8, demonstrate that IBAN can effectively weaken the interference of background information and focus on discriminative pedestrian information. Meanwhile, the features returned by IBAN come from almost the complete body of the pedestrian, while the features returned by the baseline model come from only a certain part. This indicates that the retrieval behavior of IBAN is more similar to that of humans, and it extracts and analyzes the features from the pedestrians themselves, instead of incorrectly extracting environmental features, such as pedestrian clothing features and pedestrian posture features.

Besides, we report the ReID results of IBAN with the different modules in Figure 9, where images with green boundaries are correct results that are of the same ID as the query, and images with red boundaries are incorrect results of a different ID from the query. Note that, we randomly sample a frame from each video clip for visualization. Observing Figure 9, we can find that

- Color plays a significant role in infrared-to-visible, while surface information greatly affects the results in visible-to-infrared. Besides, both infrared-to-visible and visible-to-infrared cannot effectively address the problem of the pose. Such a result demonstrates that the baseline model cannot address the modality discrepancy and leverage the spatial-temporal information of data.
- Observing the results of baseline+IG, one can find that the color and surface feature is relatively diverse in infrared-to-visible and visible-to-infrared, respectively. This result demonstrates that the IG module can partially mitigate the modality discrepancy.
- The result difference between baseline+IG and baseline+IG+BSTA reveals that BSTA can improve the robustness of the model to pose via leveraging the spatial-temporal information of video data.
- Observing the results of infrared-to-visible in baseline+IG+BSTA+ESB, i.e., IBAN, one can find that the quality of some returned images is very poor, such as the 3-th, 6-th, and 8-th images obtained under poor illumination environments. We can conclude that the infrared data is beneficial for improving the model's generalization performance to visible data with poor quality. And, the success is based on the hypothesis that ESB loss can mitigate modality discrepancy significantly in embedding space.

V. Conclusion

This paper provides a novel IBAN method to tackle the challenging task of VVI-ReID, which could handle modality discrepancy and mine spatial-temporal information simultaneously. Different from existing cross-modality methods, IBAN handles modality discrepancy in different feature levels and leverages a bidirectional spatial-temporal aggregation module to exploit the spatial-temporal information of video data. Extensive experimental results demonstrate the superiority of IBAN, and ablation studies validate the effectiveness of each module. In addition, note that although an intermediary-guided module is beneficial for reducing the gap between different modalities, Figure. 1 demonstrates that the features

of some classes still have clear modality discrepancy. Besides, a main disadvantage of IBAN is computing and storage costs during training. We have to train an additional model for the generated anaglyph data. But, note that IBAN retains the test time cost because it will remove this model during the test stage. In future work, we prefer to develop an approach that can adaptively select the classes or samples with significant modality discrepancies.

REFERENCES

[1] H. Li, N. Dong, Z. Yu, D. Tao, and G. Qi, "Triple adversarial learning and multi-view imaginative reasoning for unsupervised domain adaptation person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 5, pp. 2814–2830, May 2022.

[2] S. Liao and L. Shao, "Graph sampling based deep metric learning for generalizable person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 7359–7368.

[3] H. Li, Y. Chen, D. Tao, Z. Yu, and G. Qi, "Attribute-aligned domain-invariant feature learning for unsupervised domain adaptation person re-identification," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 1480–1494, 2021.

[4] T. Isobe, D. Li, L. Tian, W. Chen, Y. Shan, and S. Wang, "Towards discriminative representation learning for unsupervised person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 8526–8536.

[5] X. Sun and L. Zheng, "Dissecting person re-identification from the viewpoint of viewpoint," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 608–617.

[6] H. Li, J. Xu, Z. Yu, and J. Luo, "Jointly learning commonality and specificity dictionaries for person re-identification," *IEEE Trans. Image Process.*, vol. 29, pp. 7345–7358, 2020.

[7] Z. Zhuang, L. Wei, L. Xie, H. Ai, and Q. Tian, "Camera-based batch normalization: An effective distribution alignment method for person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, pp. 374–387, 2021.

[8] X. Lin et al., "Learning modal-invariant and temporal-memory for video-based visible-infrared person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 20973–20982.

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

[10] G.-A. Wang et al., "Cross-modality paired-images generation for rgb-infrared person re-identification," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 7, pp. 12144–12151.

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

[12] E. Basarana, M. Gökmenb, and M. E. Kamasaka, "An efficient framework for visible–infrared cross modality person re-identification," *Signal Process., Image Commun.*, vol. 87, Sep. 2020, Art. no. 115933.

[13] X. Hao, S. Zhao, M. Ye, and J. Shen, "Cross-modality person re-identification via modality confusion and center aggregation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 16403–16412.

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

[15] Z. Feng, J. Lai, and X. Xie, "Learning modality-specific representations for visible-infrared person re-identification," *IEEE Trans. Image Process.*, vol. 29, pp. 579–590, 2020.

[16] Q. Fan, C.-F. R. Chen, H. Kuehne, M. Pistoia, and D. Cox, "More is less: Learning efficient video representations by big-little network and depthwise temporal aggregation," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–10.

[17] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7794–7803.

[18] Y. Liu, Z. Yuan, W. Zhou, and H. Li, "Spatial and temporal mutual promotion for video-based person re-identification," in *Proc. AAAI Conf. Artif. Intell.*, 2019, vol. 33, no. 1, pp. 8786–8793.

[19] X. Gu, H. Chang, B. Ma, H. Zhang, and X. Chen, "Appearance preserving 3D convolution for video-based person re-identification," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Aug. 2020, pp. 228–243.

[20] A. Aich, M. Zheng, S. Karanam, T. Chen, A. K. Roy-Chowdhury, and Z. Wu, "Spatio-temporal representation factorization for video-based person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 152–162.

[21] D. Chen, A. Doering, S. Zhang, J. Yang, J. Gall, and B. Schiele, "Keypoint message passing for video-based person re-identification," in *Proc. AAAI Conf. Artif. Intell.*, 2022, vol. 36, no. 1, pp. 239–247.

[22] Y. Yan et al., "Learning multi-granular hypergraphs for video-based person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 2899–2908.

[23] M. Ye, J. Shen, and L. Shao, "Visible-infrared person re-identification via homogeneous augmented tri-modal learning," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 728–739, 2020.

[24] 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.*, 2020, vol. 34, no. 4, pp. 4610–4617.

[25] Z. Wei, X. Yang, N. Wang, and X. Gao, "Syncretic modality collaborative learning for visible infrared person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 225–234.

[26] J. Dai, P. Zhang, D. Wang, H. Lu, and H. Wang, "Video person re-identification by temporal residual learning," *IEEE Trans. Image Process.*, vol. 28, no. 3, pp. 1366–1377, Mar. 2019.

[27] X. Liu, P. Zhang, C. Yu, H. Lu, and X. Yang, "Watching you: Global-guided reciprocal learning for video-based person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 13334–13343.

[28] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. 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.

[29] 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)*, Oct. 2021, pp. 15013–15022.

[30] Z. Zhou, Y. Huang, W. Wang, L. Wang, and T. Tan, "See the forest for the trees: Joint spatial and temporal recurrent neural networks for video-based person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4747–4756.

[31] Y. Fu, X. Wang, Y. Wei, and T. Huang, "STA: Spatial-temporal attention for large-scale video-based person re-identification," in *Proc. AAAI Conf. Artif. Intell.*, 2019, vol. 33, no. 1, pp. 8287–8294.

[32] T. He, X. Jin, X. Shen, J. Huang, Z. Chen, and X.-S. Hua, "Dense interaction learning for video-based person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 1490–1501.

[33] J. Li, S. Zhang, J. Wang, W. Gao, and Q. Tian, "Global-local temporal representations for video person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3958–3967.

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

[35] 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)*, Jun. 2021, pp. 4370–4379.

[36] B. Jiang, M. Wang, W. Gan, W. Wu, and J. Yan, "STM: Spatiotemporal and motion encoding for action recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 2000–2009.

[37] Z. Liu, L. Wang, W. Wu, C. Qian, and T. Lu, "TAM: Temporal adaptive module for video recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 13708–13718.

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

[39] J. Liu, J. Wang, N. Huang, Q. Zhang, and J. Han, "Revisiting modality-specific feature compensation for visible-infrared person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 10, pp. 7226–7240, Oct. 2022.

[40] W. Hu, B. Liu, H. Zeng, Y. Hou, and H. Hu, "Adversarial decoupling and modality-invariant representation learning for visible-infrared person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 8, pp. 5095–5109, Aug. 2022.

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

[42] X. Zhong, T. Lu, W. Huang, M. Ye, X. Jia, and C.-W. Lin, "Grayscale enhancement colorization network for visible-infrared person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 3, pp. 1418–1430, Mar. 2022.

[43] M. Ye, X. Lan, J. Li, and P. Yuen, "Hierarchical discriminative learning for visible thermal person re-identification," in *Proc. AAAI Conf. Artif. Intell.*, Apr. 2018, vol. 32, no. 25, pp. 1–8. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/12293

[44] Z. Shuai, S. Li, Y. Gao, and F. Wu, "Adversarial learning based on global and local features for cross-modal person re-identification," in *Proc. 2nd Int. Conf. Artif. Intell. Comput. Eng. (ICAICE)*, Nov. 2021, pp. 1–4.

[45] X. Fu, F. Huang, Y. Zhou, H. Ma, X. Xu, and L. Zhang, "Cross-modal cross-domain dual alignment network for RGB-infrared person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 10, pp. 6874–6887, Oct. 2022.

[46] X. Wang, C. Chen, Y. Zhu, and S. Chen, "Feature fusion and center aggregation for visible-infrared person re-identification," *IEEE Access*, vol. 10, pp. 30949–30958, 2022.

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

[48] J. Hu and X. Li, "Cross-modal pedestrian re-identification based on generative confrontation network," in *Proc. Int. Conf. Electron., Circuits Inf. Eng. (ECIE)*, Jan. 2021, pp. 372–375.

[49] J. Kong, Q. He, M. Jiang, and T. Liu, "Dynamic center aggregation loss with mixed modality for visible-infrared person re-identification," *IEEE Signal Process. Lett.*, vol. 28, pp. 2003–2007, 2021.

[50] H. Park, S. Lee, J. Lee, and B. Ham, "Learning by aligning: Visible-infrared person re-identification using cross-modal correspondences," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 12026–12035.

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

[52] M. Ye, J. Shen, D. J. Crandall, L. Shao, and J. Luo, "Dynamic dual-attentive aggregation learning for visible-infrared person re-identification," in *Computer Vision–(ECCV)*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds. Cham, Switzerland: Springer, 2020, pp. 229–247.

[53] 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.

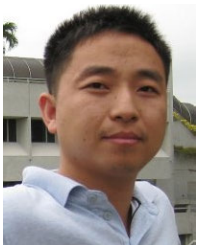
[54] M. Ye, W. Ruan, B. Du, and M. Z. Shou, "Channel augmented joint learning for visible-infrared recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 13547–13556.

[55] R. Hou, H. Chang, B. Ma, S. Shan, and X. Chen, "Temporal complementary learning for video person re-identification," in *Computer Vision–(ECCV)*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds. Cham, Switzerland: Springer, 2020, pp. 388–405.

Huafeng Li received the M.S. degree in applied mathematics and the Ph.D. degree in control theory and control engineering from Chongqing University, Chongqing, China, in 2009 and 2012, respectively. He is currently a Professor with the School of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, China. His research interests include image processing, computer vision, and information fusion.

Minghui Liu is currently pursuing the degree with the Faculty of Information Engineering and Automation, Kunming University of Science and Technology. Her research interests include computer vision and deep learning.

Zhanxuan Hu is currently an Associate Professor with the School of Computer Science and Technology, Xi'an University of Posts and Telecommunications, Xi'an, Shaanxi, China. His research interests include clustering, image representation learning, and ReID.



Feiping Nie (Senior Member, IEEE) received the Ph.D. degree in computer science from Tsinghua University, China, in 2009. He is currently a Full Professor with Northwestern Polytechnical University, China. He has authored more than 100 papers in the following top journals and conferences, including IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE (TPAMI), *International Journal of Computer Vision* (IJCV), IEEE TRANSACTIONS ON IMAGE PROCESSING (TIP), IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS (TNNLS/TNN), IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING (TKDE), *Bioinformatics*, ICML, NIPS, KDD, IJCAI, AAAI, ICCV, CVPR, and ACM MM. His papers have been cited more than 7000 times. His research interests include machine learning and its applications, such as pattern recognition, data mining, computer vision, image processing, and information retrieval. He is an associate editor or a PC member for several prestigious journals and conferences in the related fields.

Zhengtao Yu received the Ph.D. degree in computer application technology from the Beijing Institute of Technology, Beijing, China, in 2005. He is currently a Professor with the School of Information Engineering and Automation, Kunming University of Science and Technology, China. His main research interests include natural language process, image processing, and machine learning.