CycleTrans: Learning Neutral Yet Discriminative Features via Cycle Construction for Visible-Infrared Person Re-Identification

Qiong Wu, Jiaer Xia, Pingyang Dai, Yiyi Zhou, *Member, IEEE*, Yongjian Wu, and Rongrong Ji, *Senior Member, IEEE*

***Abstract*— Visible-infrared person re-identification (VI-ReID) is the task of matching the same individuals across the visible and infrared modalities. Its main challenge lies in the modality gap caused by the cameras operating on different spectra. Existing VI-ReID methods mainly focus on learning general features across modalities, often at the expense of feature discriminability. To address this issue, we present a novel cycle-construction-based network for neutral yet discriminative feature learning, termed *CycleTrans*. Specifically, CycleTrans uses a lightweight knowledge capturing module (KCM) to capture rich semantics from the modality-relevant feature maps according to pseudo anchors. Afterward, a discrepancy modeling module (DMM) is deployed to transform these features into neutral ones according to the modality-irrelevant prototypes. To ensure feature discriminability, another two KCMs are further deployed for feature cycle constructions. With cycle construction, our method can learn effective neutral features for visible and infrared images while preserving their salient semantics. Extensive experiments on SYSU-MM01 and RegDB datasets validate the merits of CycleTrans against a flurry of state-of-the-art (SOTA) methods, +1.88% on rank-1 in SYSU-MM01 and +1.1% on rank-1 in RegDB. Our code is available at https://github.com/DoubtedSteam/CycleTrans.**

***Index Terms*— Cross-modality retrieval, deep learning, person re-identification.**

Manuscript received 12 April 2023; revised 20 December 2023; accepted 20 March 2024. This work was supported in part by the National Key Research and Development Program of China under Grant 2022ZD0118202; in part by the National Science Fund for Distinguished Young Scholars under Grant 62025603; in part by the National Natural Science Foundation of China under Grant U21B2037, Grant U22B2051, Grant 62176222, Grant 62176223, Grant 62176226, Grant 62072386, Grant 62072387, Grant 62072389, Grant 62002305, and Grant 62272401; and in part by the Natural Science Foundation of Fujian Province of China under Grant 2021J01002 and Grant 2022J06001. *(Corresponding author: Pingyang Dai.)*

Qiong Wu and Yiyi Zhou are with the Key Laboratory of Multimedia Trusted Perception and Efficient Computing, Ministry of Education of China, Xiamen 361005, China, and also with the Institute of Artificial Intelligence, Xiamen University, Xiamen 361005, China (e-mail: qiong@stu.xmu.edu.cn; zhouyiyi@xmu.edu.cn).

Jiaer Xia and Pingyang Dai are with the Key Laboratory of Multimedia Trusted Perception and Efficient Computing, Ministry of Education of China, Xiamen University, Xiamen 361005, China (e-mail: xiajiaer@stu.xmu.edu.cn; pydai@xmu.edu.cn).

Yongjian Wu is with the Youtu Laboratory, Tencent Company Ltd., Shanghai 200233, China (e-mail: littlekenwu@tencent.com).

Rongrong Ji is with the Key Laboratory of Multimedia Trusted Perception and Efficient Computing, Ministry of Education of China, and Fujian Engineering Research Center of Trusted Artificial Intelligence Analysis and Application, Institute of Artificial Intelligence, Xiamen University, Xiamen 361005, China, and also with the Peng Cheng Laboratory, Shenzhen 518066, China (e-mail: rrji@xmu.edu.cn).

Digital Object Identifier 10.1109/TNNLS.2024.3382937

I. Introduction

VISIBLE-INFRARED person re-identification (VI-ReID) [1] aims at matching visible and infrared images of pedestrians with the same identity, which are captured by the cameras operating on different spectra. As more and more infrared cameras are deployed in real-world scenarios, the research of VI-ReID has attracted increasing attention from both academia and industry [1], [2], [3], [4], [5], [6], [7], [8]. In addition to the intrinsic challenges of traditional Re-ID tasks, such as the variations of viewpoints and body poses, VI-ReID also suffers from the obvious appearance difference between pedestrian images of different modalities [9], [10], [11], [12]. Meanwhile, besides the blur of image [13] and occlusion of human body [14], feature extraction is also hindered by the characteristics of cameras, e.g., the appearance of the same person in different modalities only has limited shared information.

This issue is also coined as *modality gap* [1], [15], [16], [17], as illustrated in Fig. 1(a). Specifically, under different types of cameras, the pedestrian will exhibit notable differences in visual characteristics, e.g., the color and texture of clothes. And this gap will be further reflected in the features extracted by deep neural networks, as shown in Fig. 1(b). In this case, the traditional Re-ID methods [18], [19], [20], [21], which identify pedestrians mainly based on the appearance, often fail to accomplish this task.

In recent years, a bunch of methods have been proposed for VI-ReID and achieved remarkable progress [2], [22], [23], [24], [25], [26], [27], [28], [29], [30], [31]. The prevalent solution [22], [23], [24], [25], [32] to modality gap is aligning the feature or pixel distributions of two modalities, which, however, usually comes at the expense of feature discriminability. To explain, the feature alignment needs to cluster the samples of the same modality in the joint semantic space. This optimization process also reduces the semantic distances between the samples of different identities, as shown in Fig. 1. Meanwhile, the salient semantics of pedestrian images tend to be lost during alignment, e.g., the details of cloths, which also greatly reduce the descriptive power of learned features. In this case, how to make a trade-off between the generality and discriminability of multi-modal features is the key to VI-ReID.

To address this issue, we propose a novel cycle-construction-based network (CycleTrans) for VI-ReID.

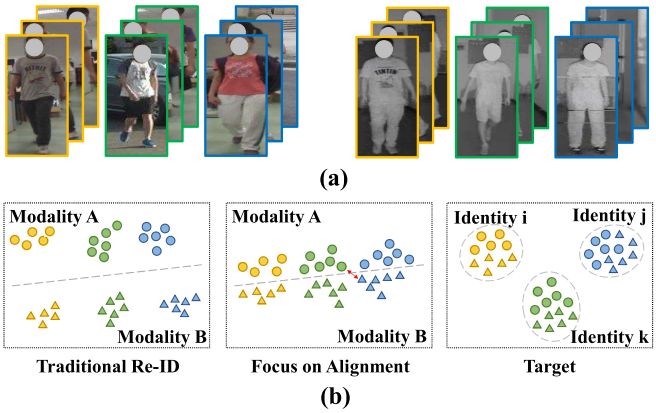

Fig. 1. Illustrations of the examples of VI-ReID and the feature spaces of different Re-ID methods. (a) In VI-ReID, pedestrians with the same identity exhibit notable appearance differences between visible and infrared images, which is often termed *modality gap*. (b) Traditional methods (left) often fail to match pedestrians across modalities, and only cross-modality alignment (middle) often narrows down the decision boundaries between samples of different identities. So, the idea semantic space for VI-ReID (right) should be neutral yet discriminative.

The main principle of CycleTrans is to enhance the descriptive power of transformed neutral features via semantical cycle reconstructions. As shown in Fig. 2, the proposed CycleTrans consists of three knowledge capturing modules (KCMs) sharing the same parameters, and a discrepancy modeling module (DMM). Specifically, the first KCM extracts discriminative semantics from convolution feature maps according to modality-specific anchors. Afterward, DMM is applied to transform these features into neutral ones for visible and infrared images, which is achieved by using modality-irrelevant prototypes as the transfer targets. To ensure discriminability, another two KCMs are further applied to reconstruct the modality-relevant features learned before. Based on this, two cycle constructions are built. The cycle construction ends with the original modality, which can benefit the discriminability. When cycle construction ends with another modality, it helps alleviate the modality gap. Through these cycle construction process, the proposed method can well model general features across modalities while preserving their salient semantics for fine-grained pedestrian identification.

To validate the proposed CycleTrans, we conduct extensive experiments on two benchmarks, namely SYSU-MM01 [1] and RegDB [17]. The experimental results not only show its obvious performance gains over the state-of-the-art (SOTA) methods, e.g., +1.88% Rank-1 on SYSU-MM01 and +1.1% Rank-1 on RegDB than DEEN [33], but also greatly confirm its effectiveness in bridging the modality gap.

Overall, our main contributions are threefold.

1) We propose a novel cycle-construction-based network for VI-ReID, termed CycleTrans. CycleTrans applies shared prototypes as transferring targets to mitigate the modality gap, and adopts the cycle construction to enhance feature discriminability.
2) To alleviate the modality gap while preserving salient semantics, two novel modules are proposed, namely KCM and DMM, which can help the model learn discriminative yet neural features.
3) The proposed CycleTrans achieves new SOTA performance on multiple benchmark datasets, e.g., 76.58% on Rank-1 in SYSU-MM01 under *all-search single-shot* setting. And the experimental results also well validate its effectiveness toward the modality gap.

II. Related Work

VI-ReID is an essential task that aims to match individuals across the visible and infrared modalities, effectively compensating for the deficiencies of visible cameras in low-light conditions. This task introduces unique challenges beyond those found in traditional ReID, such as varying viewpoints, illumination, and body poses, while also contending with the modality gap–the marked appearance differences when captured by different camera types [34], [35], [36], [37].

To address these initial challenges, Wu et al. [1] introduced the foundational SYSU-MM01 dataset and proposed a deep zero-padding network specifically designed for cross-modality matching. Building on this foundation, two-stream models were explored to process each modality independently, aiming to minimize variations at both the feature and prediction levels [38], [39], [40]. These methods set the stage for more advanced strategies like MSO [41] and CoAL [42], which further honed the capture of intra-modality information and enhanced feature discriminability. The integration of GANs marked a significant evolution in the field, with CmGAN [2] being the first to employ these networks for VI-ReID. Subsequent innovations followed, including AlignGAN [23] and JSIA [24], which leveraged GANs to generate images for the missing modality and align cross-modal distributions at multiple levels. In parallel, D^{2}RL [22] proposed a novel four-dimensional image space that encompasses both RGB and infrared data. As the field progressed, researchers introduced the concept of an intermediate modality with works like X-modality [25], cm-SSFT [5], SFANet [43], and MSA [44], which served as a bridge between the visible and infrared spectra. Besides, PartMix [45] generates the middle modality through a novel data augmentation way. However, a drawback emerged in that cm-SSFT required additional modality information even during the testing phase. In an effort to circumvent this issue, FBP-AL [46] and FMCNet [47] concentrated on extracting features that transcend modalities, with FMCNet utilizing a memory bank approach and MAUM [48] focusing on information aggregation from alternate-modality memory banks. Further refining this approach, MSCLNet [49] sought to combine representations from both modalities to increase discriminability and suppress noise. Meanwhile, MPANet [50] delved into the subtleties of inter-modality differences without supplementary supervision. Most recently, DEEN [51] and SMCL [33] have innovated by generating diverse embeddings and applying modality mixup constraints, respectively, to mitigate the modality gap while preserving discriminability. In the realm of part-based approaches, SCS+ [52] and MHSA-Net [53] brought new insights by focusing on the comparison of identical body parts, with the former using a clustering

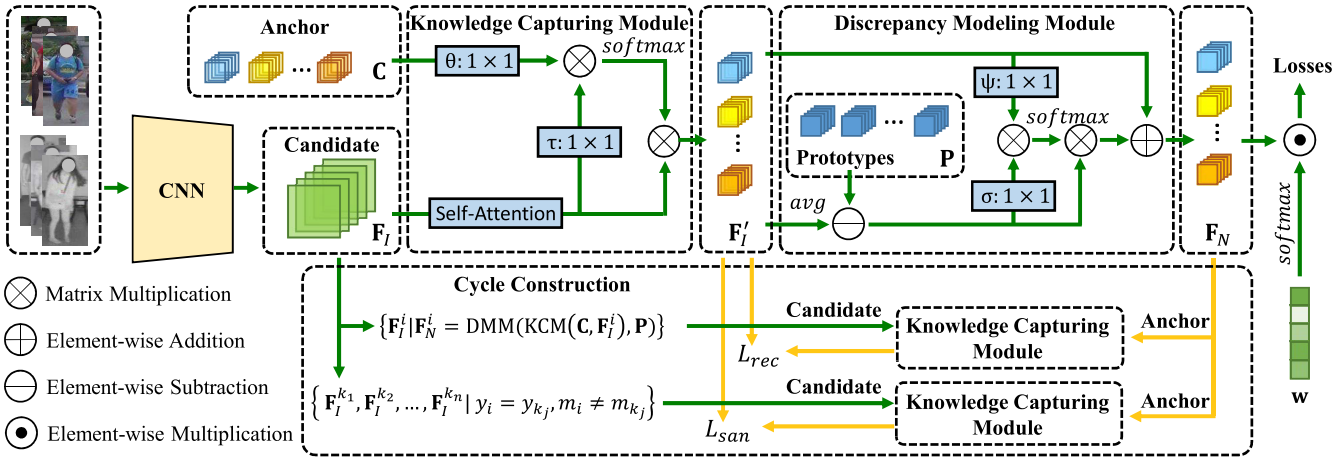

Fig. 2. Overview of the proposed CycleTrans. Given an image of arbitrary modality, CycleTrans first uses the proposed KCM to gather salient yet task-related semantics from convolution feature maps based on the modality-relevant pseudo anchors. Afterward, the DMM is deployed to transform these features into neutral ones via modeling the discrepancy to modality-irrelevant prototypes. To ensure feature discriminability, a cycle construction stage is implemented (bottom), where another two KCMs are used to transform neutral features into the original modality-relevant representations.

algorithm for part detection and the latter ensuring feature consistency within the same head. Distinguishing itself from the former methods, the proposed *CycleTrans* method sets a new precedent by transforming features of both modalities onto a shared distribution, guided by modality-irrelevant prototypes. It maintains discriminability through innovative semantic-cycle constructions, offering a novel perspective on the persistent challenge of the modality gap in VI-ReID.

III. Preliminary

Let $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i, \mathbf{m}_i)\}_{i=1}^{N}$ denotes the VI-ReID dataset which has N samples in total. For each example, denoted as $(\mathbf{x}_i, \mathbf{y}_i, \mathbf{m}_i)$, the image $\mathbf{x}_i$ has a corresponding identity label $\mathbf{y}_i \in \mathcal{Y} = \{\mathbf{y}_j\}_{j=1}^{N_p}$ and a modality label $\mathbf{m}_i \in \mathcal{M} = \{v, r\}$, where N_p is the number of identities, and v and r denote the visible and infrared modalities, respectively.

Given a pedestrian image, VI-ReID aims to match the same person in the other modality by ranking the similarity to instances in the gallery set,[1] and its objective can be defined as

$$\begin{aligned} &\underset{\Theta}{\operatorname{argmin}} \sum_{i,j,k} I\left(d_\Theta(\mathbf{x}_i, \mathbf{x}_j) > d_\Theta(\mathbf{x}_i, \mathbf{x}_r)\right), \\ &\text{where, } \mathbf{y}_i = \mathbf{y}_j, \mathbf{y}_i \neq \mathbf{y}_r, \mathbf{m}_i \neq \mathbf{m}_j, \mathbf{m}_i \neq \mathbf{m}_r. \end{aligned} \tag{1}$$

Here, $I(\cdot)$ is an indicator function that returns 1 if the condition is satisfied and 0 otherwise. $d_\Theta(\cdot,\cdot)$ measures distance between two features extracted by the model with parameters Θ.

IV. Method

A. Overview

The overall structure of the proposed cycle-construction-based network (*CycleTrans*) is depicted in Fig. 2. Its main principle is to maintain the descriptive power of the transformed neutral features via feature cycle constructions.

[1]In testing, gallery set contain a series of pedestrian images whose identity is known.

Specifically, for a visible or infrared image $\mathbf{x}_i$, we first apply a convolutional backbone to extract its feature map, denoted as $\mathbf{F}_I \in \mathbb{R}^{h\times w\times d}$, where $h \times w$ denotes the resolution and d is dimensionality. Afterward, we use the proposed KCM to mine rich semantics from $\mathbf{F}_I$

$$\mathbf{F}_I' = \mathrm{KCM}(\mathbf{F}_I, \mathbf{C}) \tag{2}$$

where $\mathbf{C} \in \mathbb{R}^{k\times d}$ denotes the trainable pseudo anchors of the corresponding modality. After the process of KCM, the obtained features $\mathbf{F}_I' \in \mathbb{R}^{k\times d}$ contain descriptive semantics for Re-ID, but it is still modality-relevant.

To this end, we further transform $\mathbf{F}_I'$ into neutral features via a novel DMM

$$\mathbf{F}_N = \mathrm{DMM}(\mathbf{P}, \mathbf{F}_I') \tag{3}$$

where $\mathbf{P} \in \mathbb{R}^{n\times d}$ are modality-irrelevant prototypes. Neutral features $\mathbf{F}_N \in \mathbb{R}^{k\times d}$ are further flattened to a representation vector and then used for cross-modal retrieval.

To ensure the discriminability of the transformed $\mathbf{F}_N$, we use it to reconstruct the modality-relevant features $\mathbf{F}_I'$ via another two KCMs. To keep the model compact, three KCMs share the same parameters.

Overall, through this cycle-construction paradigm, the proposed CycleTrans can well capture salient semantics from each modality, while learning effective neutral representations for cross-modal retrieval.

B. Knowledge Capturing Module

KCM is a novel and lightweight module for learning discriminative and task-related semantics from convolutional feature maps.

Concretely, given the feature map of an arbitrary modality $\mathbf{F}_I \in \mathbb{R}^{h\times w\times d}$, we first reshape it to a 2-d tensor $\hat{\mathbf{F}}_I \in \mathbb{R}^{hw\times d}$. Then, we apply a *dot-product* attention to refine the features by aggregating semantics from similar regions

$$\tilde{\mathbf{F}}_I = \mathrm{Softmax}\left(\mathrm{norm}(\hat{\mathbf{F}}_I)\mathrm{norm}(\hat{\mathbf{F}}_I)^T\right)\hat{\mathbf{F}}_I \tag{4}$$

where norm$(\cdot)$ denotes l-2 normalization. The obtained feature map $\tilde{\mathbf{F}}_I$ mainly represents the general semantics of a given image, while the relevant ones for pedestrian identification still need to be enhanced.

Then we implement a cross-attention operation to mine task-related semantics based on the learn-able pseudo anchors $\mathbf{C} \in \mathbf{R}^{k\times d}$

$$\mathbf{F}'_I = \text{Softmax}\left(\frac{\mathbf{C}\mathbf{W}_\theta\left(\tilde{\mathbf{F}}_I\mathbf{W}_\tau\right)^T}{\sqrt{c}}\right)\tilde{\mathbf{F}}_I \tag{5}$$

where $\mathbf{W}_\theta$ and $\mathbf{W}_\tau$ are weight matrices. To adaptively select the pattern that is most relevant to the task, we apply pseudo anchors to filter information from the original feature maps. Since the pseudo anchors $\mathbf{C}$ are highly task-related, they can well help the model mine useful semantics for VI-ReID via (4) and (5) from the candidate $\mathbf{F}_I$, resulting in more discriminative modality-relevant features.

Notably, in our CycleTrans, KCM first serves to extract modality-relevant features based on the pseudo anchors for VI-ReID. And two modalities share the pseudo anchors for the aligned semantic. During the cycle construction, KCM is used as a module to reconstruct modality-relevant features based on neutral features, which can be achieved by placing different feature maps as the candidate. In KCM, the process of gradient backward is very similar to that in *self-attention* [54].

C. Discrepancy Modeling Module

DMM acts to mitigate the modality gap of VI-ReID. Instead of directly embedding the modality-relevant features into a common semantic space, DMM learns the neutral features via aggregating information from a set of modality-irrelevant prototypes. The prototypes $\mathbf{P} \in \mathbb{R}^{n\times d}$ consist of a set of learnable vectors that represent the semantics of appearance [55]. Considering the actual appearance of the pedestrian is modality-independent, so the two modalities share the prototypes.

Concretely, given the discriminative modality-relevant features learned by KCM, denoted as $\mathbf{F}'_I$, we first calculate their discrepancy to the trainable prototypes, where n is the number of prototypes

$$\mathbf{P}' = \mathbf{P} - \hat{\mathbf{f}}_I. \tag{6}$$

Here, $\mathbf{P}' \in \mathbb{R}^{n\times d}$ refers to the obtained discrepancy tensor and $\hat{\mathbf{f}}_I$ denotes the averaged feature of $\mathbf{F}'_I$. The gradient will not be zero due to the difference between the average feature $\hat{\mathbf{f}}_I$ and the modality-dependent feature $\mathbf{F}'_I$. Afterward, the neutral features $\mathbf{F}_N \in \mathbb{R}^{k\times d}$ are obtained via a residual connection and a cross attention

$$\begin{aligned}\mathbf{F}_N &= \mathbf{F}'_I + \mathbf{A}\mathbf{P}',\\ \text{where } \mathbf{A} &= \text{Softmax}\left(\frac{\mathbf{F}'_I\mathbf{W}_\psi(\mathbf{P}\mathbf{W}_\sigma)^T}{\sqrt{c}}\right).\end{aligned} \tag{7}$$

Here, the attention weights $\mathbf{A} \in \mathbb{R}^{k\times n}$ are also the weighted adjacent matrix between $\mathbf{F}'_I$ and $\mathbf{P}$. The $\mathbf{W}_\psi$ and $\mathbf{W}_\sigma$ here are weight matrices. $\mathbf{A}$ can reformulate semantics in a general space according to the modality-relevant features $\mathbf{F}'_I$.

Note that the sum of each row in $\mathbf{A}$ equals to 1, and $\mathbf{P}' = \mathbf{P} - \hat{\mathbf{f}}_I$. Thus, (7) can be rewritten as

$$\mathbf{F}_N = \mathbf{F}'_I + \mathbf{A}\left(\mathbf{P} - \hat{\mathbf{f}}_I\right) \tag{8}$$

$$\mathbf{F}_N = \left(\mathbf{F}'_I - \hat{\mathbf{f}}_I\right) + \mathbf{A}\mathbf{P}. \tag{9}$$

Considering that $\hat{\mathbf{f}}_I$ is the averaged vector of $\mathbf{F}'_I$, the term of $(\mathbf{F}'_I - \hat{\mathbf{f}}_I)$ in (9) will result in an informative sparse tensor. In this case, $\mathbf{F}_N$ is mainly composed of the newly aggregated prototype features, i.e., $\mathbf{A}\mathbf{P}$, thereby achieving the alignment of cross-modality distributions.

To enhance the neutral features, we also place trainable weights to adaptively adjust the contribution of each pattern in $\mathbf{F}_N$, which is achieved by

$$\mathbf{F}_{N_i} \rightarrow \frac{e^{w_i}}{\sum_{j=1}^{k} e^{w_j}}\mathbf{F}_{N_i} \tag{10}$$

where w_j is the weight for jth pattern of neutral feature. And "$\rightarrow$" refers to weighting up the neutral feature $\mathbf{F}_{N_i}$ extracted according to the ith pseudo anchor. In this way, the path of gradient backward in DMM is similar to cross-attention [55].

From (8) and (9), we can see that $\mathbf{F}_N$ contains a certain amount of discriminative information form $\mathbf{F}'_I - \hat{\mathbf{f}}_I$, but it is still hard to ensure that they are discriminative enough for VI-ReID. In this case, we further implement *Cycle Constructions* to enhance their descriptive power.

D. Cycle Construction

The main assumption of Cycle Construction is that if the learned neutral features can recover modality-relevant information well, they are capable of both cross-modality alignment and prominent feature discrimination.

Specifically, the proposed cycle construction consists of two processes, which transform the neutral features into visible and infrared ones, respectively. Taking a visible image for example, of which feature maps are denoted as $\mathbf{F}^v_I \in \mathbb{R}^{hw\times c}$, we apply the proposed KCM to reconstruct its modality-relevant features

$$\mathbf{F}^v_{Re} = \text{KCM}\left(\mathbf{F}^v_I, \mathbf{F}_N\right) \tag{11}$$

where $\mathbf{F}^v_{Re}$ is the recovered features and $\mathbf{F}_N$ acts the role of pseudo anchors described in (5). During training, we will minimize the l-1 distance between the recover features $\mathbf{F}^v_{Re}$ and the modality-relevant ones $\mathbf{F}^{v'}_I$ defined in (5) for the discriminability of neutral features $\mathbf{F}^N_I$.

In the other stream, CycleTrans project the neutral features to the other modality through KCM, i.e., the infrared one here, defined as

$$\begin{aligned}\mathbf{F}^r_{Re} &= \text{KCM}\left(\left[\mathbf{F}^{k_1}_I, \mathbf{F}^{k_2}_I, \ldots, \mathbf{F}^{k_h}_I\right], \mathbf{F}_N\right)\\ \text{where } \mathbf{y}_i &= \mathbf{y}_{k_j}, \mathbf{m}_i \neq \mathbf{m}_{k_j}, j = 1, 2, \ldots, h.\end{aligned} \tag{12}$$

Here, $\mathbf{F}^r_{Re}$ denotes the recovered infrared features and $[\mathbf{F}^{k_1}_I, \mathbf{F}^{k_2}_I, \ldots, \mathbf{F}^{k_h}_I]$ denotes h feature maps that have the same identity but from the infrared modality in the batch.

In (12), the neutral features are regarded as the pseudo anchors for KCM to aggregate semantics from all feature maps that may provide valuable information. It can help to rule out

the factors that may affect appearance discrepancy between samples for more accurate reconstruction, e.g., viewpoints, body poses, and obstructions.

To ensure the reconstruction, we also minimize the semantic distance between two generated features, i.e., $\mathbf{F}_I^r$, and $\mathbf{F}_{Re}^r$. This objective is also beneficial for alleviating the modality gap. For an infrared image, the process of cycle construction is the same. To maintain the compactness of CycleTrans, we share the parameters of the three KCMs. The gradient is only backward through the anchor features to the backbone.

E. Optimization

During training, we apply the following objectives to optimize CycleTrans.

1) Cross-Entropy Loss: As the main objective of VI-ReID, *cross-entropy loss* is used to learn the identities of samples with classifier $\mathrm{C}(\cdot)$ under the supervision of the label $\mathbf{y}_i$

$$\mathcal{L}_{id} = -\frac{1}{B}\sum_{i=1}^{B}\log P\left(\mathbf{y}_i|\mathrm{C}\left(\mathbf{f}_N^{(i)}\right)\right) \tag{13}$$

where $\mathrm{C}(\mathbf{f}_N^{(i)})$ is the predicted identity based on the flattened neutral feature $\mathbf{f}_N^{(i)} \in \mathbb{R}^{kd}$ of the sample $\mathbf{x}_i$.

2) Metric Loss: To semantically separate the obtained neutral features, we apply a *metric loss* to CycleTrans

$$\mathcal{L}_{me} = \frac{1}{B^2}\sum_{i=1}^{B}\sum_{j=1,\mathbf{y}_i\neq\mathbf{y}_j}^{B}\Big[\rho - d\left(\mathbf{f}_N^{(i)},\mathbf{f}_N^{(j)}\right) + d\left(\mathbf{f}_N^{(i)},\bar{\mathbf{f}}_N^{(i)}\right) + d\left(\mathbf{f}_N^{(j)},\bar{\mathbf{f}}_N^{(j)}\right)\Big]_+ \tag{14}$$

where $[\cdot]_+$ represents $\max\{\cdot,0\}$, B denotes the batch size. $d(\cdot)$ is the distance function, which is l-2 here. $\mathbf{c}^{(i)}$ denotes the center of class $\mathbf{f}_N^{(i)}$ belong to, which is calculated in each batch, and ρ is the least margin between two classes. Via (14), CycleTrans can well separate the neural features of different identities and minimize the distance between the example and its multi-modality anchor, i.e., the class center $\mathbf{c}$. In this case, it is much easier to obtain the general and cross-modality representation, which is critical in VI-ReID.

3) Separation Loss: To learn neutral features with more diverse patterns, we define the following regularization term:

$$\mathcal{L}_{\text{sep}} = \frac{1}{k^2}\sum_{i=1}^{k-2}\sum_{i=j+1}^{k-1}\frac{\mathbf{F}_{N_i}\mathbf{F}_{N_j}}{|\mathbf{F}_{N_i}|_2|\mathbf{F}_{N_j}|_2} \tag{15}$$

where $\mathbf{F}_{N_i}$ is the ith pattern of the neutral feature. Note that, the last pattern of neutral features $\mathbf{F}_N$ is not involved in the $\mathcal{L}_{sep}$, which plays the role of global representation.

4) Modality Fusion Loss: We also apply the *Multikernel Maximum Mean Discrepancy* (MMD) [56] with Gaussian kernel to make features following a similar distribution:

$$\mathcal{L}_{\text{MMD}} = ||\mathrm{E}_v\left[\phi\left(\mathbf{F}_N^v\right)\right] - \mathrm{E}_r\left[\phi\left(\mathbf{F}_N^r\right)\right]||_{\mathcal{H}_k}^2 \tag{16}$$

where $\phi(\cdot)$ is an implicit feature mapping function and $\mathcal{H}_k$ represents the *Reproducing Kernel Hilbert Space* (RKHS). $\mathbf{F}_N^v$ and $\mathbf{F}_N^r$ denote the neutral features of visible and infrared images, respectively. Equation (16) can ensure the consistency between the neural features of different modalities.

5) Reconstruction Loss: To ensure the discriminability of neutral features and the quality of reconstructions, we propose a distance-based reconstruction loss

$$\mathcal{L}_{\text{rec}} = |\mathbf{F}_{Re}^v - \mathbf{F}_I^{v'}|_1. \tag{17}$$

Here, the $|\cdot|_1$ represents the l-1 distance. By decreasing the distance between reconstructed features $\mathbf{F}_{Re}^v$ and modality-relevant features $\mathbf{F}_I^{v'}$, we can keep semantic consistency during transformation.

6) Alignment Loss: We also introduce an *Alignment loss* to ensure the quality of recovered cross-modality features, which is defined by

$$\mathcal{L}_{\text{aln}} = |\mathbf{F}_{Re}^r - \mathbf{F}_I^{v'}|_2 \tag{18}$$

where $|\cdot|$ denote the l-2 distance. Equation (18) can also serve to reduce the gap between visible and infrared images by aligning two types of features.

Notably, in (17) and (18), we use the reconstruction of visible features as an example. During training, these loss terms are also applied to infrared images.

In summary, the overall objective function of the proposed CycleTrans is defined as

$$\mathcal{L} = \mathcal{L}_{id} + \mathcal{L}_{me} + \lambda_1\mathcal{L}_{\text{sep}} + \lambda_2\mathcal{L}_{\text{MMD}} + \lambda_3\mathcal{L}_{\text{rec}} + \lambda_4\mathcal{L}_{\text{aln}} \tag{19}$$

where λ_1, λ_2, λ_3 and λ_4 are hype-parameters. They are set mainly based on our empirical knowledge and put the cross-entropy loss at the center. Specifically, both λ_3 and λ_4 are directly set to 0.1 based on their scales of gradients. Only λ_1 and λ_2 will be tuned during experiments.

V. Experiments

A. Datasets and Metrics

We validate the proposed CycleTrans on two VI-ReID benchmarks, namely SYSU-MM01 [1] and RegDB [17].

SYSU-MM01 is a large-scale dataset consisting of both indoor and outdoor images captured by four visible cameras and two near-infrared ones. The training set contains 395 identities with 22 258 visible images and 11 909 infrared ones. The query set has 3803 infrared images and the gallery set shares 96 identities. Under the *single-shot* and *multishot* setting, there are 301 and 3010 randomly sampled visible images in the gallery, respectively.

RegDB is a small-scale dataset with images captured by a pair of aligned cameras (one visible and one thermal). It contains 8240 images of 412 identities, each with ten visible and ten thermal images. The dataset is randomly divided into two splits, i.e., images of 206 identities for training and the rest of 206 identities for testing.

For two VI-ReID datasets, the cumulative matching characteristic (CMC) [57], including **Rank-1**, **Rank-10**, and **Rank-20** accuracies, and mean average precision (mAP) metrics are used as the evaluation metrics. All comparisons use the same metrics.

TABLE I
HYPERPARAMETER SELECTION ON SYSU-MM01 UNDER ALL-SEARCH SINGLE-SHOT SETTING

Number of anchors k	Number of prototypes n	SYSU-MM01 All-search Single-shot			
		Rank-1	Rank-10	Rank-20	*m*AP
7	1024	71.96	95.55	98.46	67.24
7	256	69.93	95.30	98.43	65.74
7	512	70.10	95.31	98.53	65.79
7	2048	69.89	95.16	98.19	65.61
4	1024	69.94	95.41	98.44	65.48
5	1024	70.21	95.18	98.22	65.88
6	1024	71.02	95.45	98.27	66.31
8	1024	69.97	95.43	98.39	65.46

TABLE II
ABLATION STUDY ON SYSU-MM01 UNDER ALL-SEARCH SINGLE-SHOT SETTING

Method	SYSU-MM01 All-search Single-shot			
	Rank-1	Rank-10	Rank-20	*m*AP
Baseline	58.99	91.18	96.06	54.29
+ KCM	62.32	91.53	96.39	57.49
+ $\mathcal{L}_{me}$	64.93	93.88	97.60	60.74
+ $\mathcal{L}_{MMD}$	65.73	93.09	96.99	61.99
+ DMM	67.40	94.76	98.15	63.01
+ $\mathcal{L}_{rec}$	69.76	95.09	98.23	65.15
+ $\mathcal{L}_{aln}$ (Full)	71.96	95.55	98.46	67.24

TABLE III
IMPACT OF DIFFERENT ALTERNATIVES OF DMM ON SYSU-MM01 UNDER ALL-SEARCH SINGLE-SHOT SETTING

Method	SYSU-MM01 All-search Single-shot			
	Rank-1	Rank-10	Rank-20	*m*AP
Baseline	58.99	91.18	96.06	54.29
$\mathbf{F}_N = \mathbf{AP}$	68.74	94.80	98.09	64.32
$\mathbf{F}_N = \mathbf{F}'_I + \mathbf{AP}$	66.56	94.38	98.07	62.66
Transformer	68.44	94.23	97.76	63.55
DMM	71.96	95.55	98.46	67.24

B. Implementation Details

For CycleTrans, we use ResNet-50 [64] as the backbone, and the stride of the last convolutional layer is set to 1 for more fine-grained information. The classifier C(·) consists of a BN neck [4] and an FC layer without bias. The input images are resized to 384×192 and are randomly flipped and erased [65] with 50% probability. The λ hyper-parameter sets are $[0.3, 0.7, 0.15, 0.2]$ for SYSU-MM01, and $[0.2, 0.8, 0.1, 0.1]$ for RegDB. The margin ρ in $\mathcal{L}_{me}$ is set to 0.5. The number of prototypes is set to 1024. We apply seven pseudoanchors for both SYSU-MM01 and RegDB to extract neutral features from both two modalities' images. Both anchors and prototypes are trainable vectors and are initialized by a normal distribution with a mean of 0.0 and a variance of 0.02. During training, each mini-batch contains 64 images of eight identities. We randomly sample four visible images and four infrared images for each identity. The proposed model is trained for a total of 140 epochs and optimized by *Adam* [66] with an initial learning rate of 3.5×10^{-4}. The learning rate decays at the 40th and 70th epoch with a decay factor of 0.1.

C. Ablation Study

We first evaluate the influence of prototypes and anchors by adjusting their numbers. As shown in Table I, more or fewer prototypes both degrade the performance of our CycleTrans. The more prototypes make features from two modalities that have no common representation and cannot alleviate the modality discrepancy. While too few prototypes are missing to represent the necessary information about a person. Similarly, too much anchors cause the conflict in (15) and less anchors capture inadequate information for a person to decline the performance.

We then ablate our CycleTrans on SYSU-MM01 under *all-search single-shot* setting [1], of which results are given in Table II. Here, *baseline* denotes that the model only consists of the convolution backbone and is trained merely with the cross-entropy loss $\mathcal{L}_{id}$.

Table II shows the cumulative results of each design in CycleTrans. From this table, we can first observe that the proposed KCM and DMM can significantly improve model performance, achieving $+3.33\%$ and $+1.67\%$ gains on Rank-1, respectively. The use of cycle construction, i.e., $+\mathcal{L}_{rec}$ and $+\mathcal{L}_{aln}$, can also improve performance to a large extent, e.g., $+4.56\%$ on Rank-1 compared to "$+DMM$." Meanwhile, we also notice that the metric loss $\mathcal{L}_{me}$ can also bring improvements on all metrics, suggesting its benefits for neutral features. Lastly, combining all designs proposed in CycleTrans can improve the baseline by up to $+12.97\%$ Rank-1, strongly validating their effectiveness.

We also examine different alternatives of the proposed DMM, i.e., (7), of which results are given in Table III. The second block of Table III shows the different choices of DMM, including the one aggregating prototypes without residual connection, i.e., $\mathbf{F}_N = \mathbf{AP}$, and the one without discrepancy modeling, i.e., $\mathbf{F}_N = \mathbf{F}'_I + \mathbf{AP}$. We also use a Transformer layer [67] for comparison.

The first alternative only uses the aggregated prototypes as neutral features, which can strictly follow the distribution of prototype information. However, this alternative will make the convolution backbone hard to optimize, since the image features are not directly involved in the objective functions. Meanwhile, the lack of fine-grained image semantics from residual connection also limits its performance upper-bound. Compared to DMM, the second alternative does not include discrepancy modeling, which leads to obvious performance degradation. One hypothesis is that without discrepancy modeling, the obtained neutral features are still highly modality-relevant, making the model fail in cross-modal retrieval. The use of a Transformer layer is a good choice for neutral feature transformation, which takes the modality-relevant features as queries and the prototypes as keys and values. However, its performance is still inferior to our DMM, e.g., -3.52% Rank-1 and -3.69% *m*AP. Overall, these designs well confirm the effectiveness of our DMM in neutral feature learning for VI-ReID.

D. Comparison With SOTA Methods

We then compare our CycleTrans with a set of SOTAs on SYSU-MM01 and RegDB, of which results are given in Tables IV and V, respectively.

TABLE IV

COMPARISON BETWEEN CYCLETRANS AND THE SOTA METHODS ON SYSU-MM01. THE BEST PERFORMANCE IS **BOLD**, AND THE SECOND BEST IS UNDERLINED. THE METHODS THAT TAKE THE SAME BACKBONE AND SETTINGS AS AGW [17] ARE MARKED WITH "*"

Method	All-Search								Indoor-Search							
	Single-Shot				Multi-Shot				Single-Shot				Multi-Shot			
	R1	R10	R20	*m*AP	R1	R10	R20	*m*AP	R1	R10	R20	*m*AP	R1	R10	R20	*m*AP
Zero-Padding [1]	14.80	54.12	71.33	15.95	19.13	61.40	78.41	10.89	20.58	68.38	85.79	26.92	24.43	75.86	91.32	18.86
BDTR [58]	17.01	55.43	71.96	19.66	-	-	-	-	-	-	-	-	-	-	-	-
D-HSME [38]	20.68	62.74	77.95	23.12	-	-	-	-	-	-	-	-	-	-	-	-
cmGAN [2]	26.97	67.51	80.56	27.80	31.49	72.74	85.01	22.27	31.63	77.23	89.18	42.19	37.00	80.94	92.11	32.76
D^2RL [22]	28.90	70.60	82.40	29.20	-	-	-	-	-	-	-	-	-	-	-	-
Hi-CMD [59]	34.94	77.58	-	35.94	-	-	-	-	-	-	-	-	-	-	-	-
JSIA-ReID [24]	38.10	80.70	89.90	36.90	45.10	85.70	93.80	29.50	43.80	86.20	94.20	52.90	52.70	91.10	96.40	42.70
AlignGAN [23]	42.40	85.00	93.70	40.70	51.50	89.40	95.70	33.90	45.90	87.60	94.40	54.30	57.10	92.70	97.40	45.30
AGW [17]	47.50	-	-	47.65	-	-	-	-	54.17	-	-	62.97	-	-	-	-
DFE [60]	48.71	88.86	95.27	48.59	54.63	91.62	96.83	42.14	52.25	89.86	95.85	59.68	59.62	94.45	98.07	50.60
XIV-ReID [25]	49.92	89.79	95.96	50.73	-	-	-	-	-	-	-	-	-	-	-	-
CMM+CML [61]	51.80	92.72	97.71	51.21	56.27	94.08	98.12	43.39	54.98	94.38	99.41	63.70	60.42	96.88	99.50	53.52
FBP-AL [46]	54.14	86.04	93.03	50.20	-	-	-	-	-	-	-	-	-	-	-	-
SIM [62]	56.93	-	-	60.88	-	-	-	-	-	-	-	-	-	-	-	-
CoAL [42]	57.22	92.29	97.57	57.20	-	-	-	-	63.86	95.41	98.79	70.84	-	-	-	-
MSO [41]	58.70	92.06	-	56.42	65.85	94.37	-	49.56	63.09	96.61	-	70.31	72.06	91.77	-	61.69
DG-VAE [26]	59.49	93.77	-	58.46	-	-	-	-	-	-	-	-	-	-	-	-
MCLNet* [16]	65.40	93.33	97.14	61.98	-	-	-	-	72.56	96.98	99.20	76.58	-	-	-	-
SFANet [43]	65.74	92.98	97.05	60.83	-	-	-	-	71.60	96.60	99.45	80.05	-	-	-	-
FMCNet [47]	66.34	-	-	62.51	73.44	-	-	56.06	68.16	-	-	74.09	78.86	-	-	63.82
SMCL [33]	67.39	92.87	96.76	61.78	72.15	90.66	94.32	54.93	68.84	96.55	98.77	75.56	79.57	95.33	98.00	66.57
CAJ* [63]	69.88	95.71	98.46	66.89	-	-	-	-	76.26	97.88	99.49	80.37	-	-	-	-
MPANet [50]	70.58	96.21	98.80	68.24	75.58	97.91	99.43	62.91	76.74	98.21	99.57	80.95	84.22	99.66	99.96	75.11
MAUM [48]	71.68	-	-	68.79	-	-	-	-	76.97	-	-	81.94	-	-	-	-
DEEN [51]	74.70	**97.60**	99.20	71.80	-	-	-	-	80.30	99.00	99.80	83.30	-	-	-	-
CycleTrans(Ours)	71.96	95.55	98.46	67.24	79.36	97.64	99.18	62.53	82.55	99.58	99.95	80.46	89.28	99.83	99.99	76.25
CycleTrans*(Ours)	**76.58**	97.22	**99.28**	**72.62**	**82.82**	**98.59**	**99.71**	**68.52**	**87.22**	**99.64**	**99.98**	**84.92**	**91.21**	**99.81**	**100.00**	**81.41**

1) Comparisons on SYSU-MM01: As shown in Table IV, the proposed CycleTrans outperforms existing SOTAs by large margins on SYSU-MM01. Specifically, compared to the latest method, i.e., SMCL [33], CycleTrans can obviously improve the performance of all metrics under *All-Search* setting, e.g., +4.57% on Rank-1 and +5.46% on *m*AP. Under the setting of *Indoor-Search*, the advantages of CycleTrans are further expanded. For instance, the SOTA performance on *Single-shot* Rank-1 and *Multi-shot* Rank-1 is improved by +9.99% and +9.71% by our method, which is indeed very significant. When taking the same backbone and settings as AGW [17], our CycleTrans maintains its advantage. Specifically, compared to DEEN [51], CycleTrans improves performance under All-Search Single-Shot, e.g., +1.88% on Rank-1 and +0.82% on *m*AP.

2) Comparisons on RegDB: Similar advantages of CycleTrans can be also witnessed on RegDB in Table V, which is a smaller-scale dataset. Under two cross-modality settings, our method achieves new SOTA performance on all metrics. Notably, the latest method FMCNet [47] has already achieved obvious gains over previous VI-ReID methods, but our CycleTrans can further improve performance, e.g., +2.2% and +1.9% Rank-1 on two settings. When taking the same backbone and settings as AGW [17], the proposed CycleTrans achieves competitive performance. Under both *Infrared to Visible* and *Visible to Infrared* settings, the proposed CycleTrans improve *m*AP by +2.2% and +1.9%, compare to DEEN [51].

Considering SYSU-MM01 and RegDB are two highly competitive benchmarks, these significant performance gains strongly validate the effectiveness of the proposed CycleTrans and our motivation about the modality gap.

E. Quantitative Analysis

1) Impact of Hyper-Parameters: In Fig. 3, we report the impact of hyper-parameters in the proposed CycleTrans. We can first observe that CycleTrans is reasonably robust to the values used to control the impact of different modules in cycle construction and identification (i.e., λ_1, λ_2, λ_3, and λ_4). In particular, the most significant difference appears in λ_1, when the difference between 0.3 and 0.5 reaches 3.23% in Rank-1. Especially when the value is too large, the performance reduction is particularly obvious. Since the model gives up some valuable information, it ensures that the information captured by each anchor is not duplicated. We can also observe that the loss function used to eliminate modality differences is not sensitive to their weights, i.e., λ_2 and λ_4. This shows that the proposed KCM and DMM play an effective role in alleviating modality gap. As for

TABLE V
COMPARISON WITH SOTA METHODS ON REGDB. THE BEST PERFORMANCE IS **BOLD**, AND THE SECOND BEST IS UNDERLINED. THE METHODS THAT TAKE THE SAME BACKBONE AND SETTINGS AS AGW [17] ARE MARKED WITH "*"

Method	Infrared to Visible		Visible to Infrared	
	Rank-1	*m*AP	Rank-1	*m*AP
Zero-Padding [1]	16.7	17.9	17.8	18.9
BDTR [58]	32.7	31.1	33.5	31.8
D^2RL [22]	-	-	43.4	44.1
JSIA-ReID [24]	48.1	48.9	48.5	49.3
D-HSME [38]	50.2	46.2	50.9	47.0
AlignGAN [23]	56.3	53.4	57.9	53.6
CMM+CML [61]	59.8	60.9	-	-
XIV-ReID [25]	62.3	60.2	-	-
AGW [17]	-	-	70.0	66.4
DFE [60]	68.0	66.7	70.2	69.2
DG-VAE [26]	-	-	73.0	71.8
CoAL [42]	74.1	69.9	-	-
FBP-AL [46]	-	-	73.4	68.2
MSO [41]	74.6	67.5	73.6	66.9
SIM [62]	75.2	78.3	74.7	75.2
SFANet [43]	70.2	63.8	76.3	68.0
MCLNet* [16]	75.9	69.5	80.3	73.1
MPANet [50]	82.8	80.7	83.7	80.9
SMCL [33]	83.1	78.6	83.9	79.8
CAJ* [63]	84.8	77.8	85.0	79.1
FMCNet [47]	89.1	84.4	88.4	83.9
MAUM [48]	87.0	84.3	87.9	85.1
DEEN [51]	89.5	83.4	**91.1**	85.1
CycleTrans(Ours)	**91.3**	84.9	90.3	84.9
CycleTrans*(Ours)	90.6	**85.6**	91.5	**87.0**

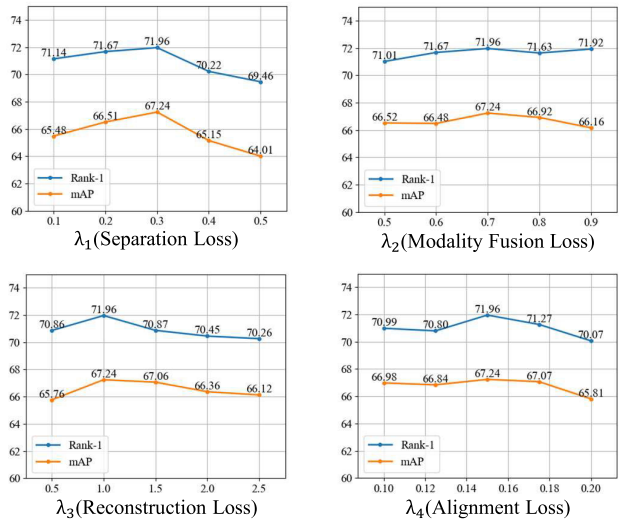

Fig. 3. Impact of hyper-parameters λ_1, λ_2, λ_3 and λ_4. The performances are evaluated on SYSU-MM01 under all-search single-shot.

reconstruction loss, the λ_2 is limited to a small range to ensure that the model can be optimized normally. Experimental results well confirm the effectiveness of the proposed CycleTrans in alleviating modality gap and extracting discriminative features.

2) Inference Efficiency: We further compare the actual inference efficiencies of CycleTrans and representative methods. The computation overhead during both the training and testing stages are reported in Table VI. We can first observe that the proposed CycleTrans significantly improves the performance, i.e., +17.59% and +18.33% on Rank-1 and *m*AP, with limited increase in the computation overhead, i.e., +0.6h in training. As for the representative method, e.g., DEEN [51], CycleTrans* achieves +1.88% and +0.82% on Rank-1 and *m*AP with 33.3% training time. Overall, the proposed CycleTrans method is an effective and efficient way to address the visible-infrared person ReID.

TABLE VI
COMPUTATION OVERHEAD OF THE PROPOSED CYCLETRANS. THE METHODS THAT TAKE THE SAME BACKBONE AND SETTINGS AS AGW [17] ARE MARKED WITH "*"

Method	SYSU-MM01			
	All-search Single-shot			
	Rank-1	*m*AP	Training Time	Inference Time
Baseline	58.99	54.29	3.0h	0.019 s/sample
DEEN [51]	74.70	71.80	12.9h	0.031 s/sample
CycleTrans	71.96	67.24	3.6h	0.023 s/sample
CycleTrans*	76.58	72.62	4.3h	0.030 s/sample

F. Qualitative Analysis

To gain deep insight into the proposed CycleTrans, we further visualize the distributions of different features extracted by the baseline and our CycleTrans in Fig. 4. We randomly visualize samples of ten identities from the testing set via t-SNE [68]. Fig. 4(a) shows the feature distribution of the baseline. We can see that although these features can be mapped to different clusters, images of the same identity but different modalities are still hard to distinguish. For instance, the blue and yellow features of the same modalities are closely distributed in this space and hard to identify. Fig. 4(b) shows the results of CycleTrans without Cycle Construction. It illustrates that CycleTrans can well transform these modality-relevant features into neutral ones with the help of the proposed DMM, resulting in better clusters than Fig. 4(a). However, due to the lack of enough feature discriminability, the cross-modality features of some identities still do not exhibit clear semantic margins, e.g., the yellow and blue examples. With cycle construction, this problem is greatly alleviated, as shown in Fig. 4(c). From this figure, we can see that our CycleTrans can learn clear margins between features of different identities. Meanwhile, the better clustering result of CycleTrans than the other two methods suggests a stronger descriptive power.

Furthermore, cycle construction can also improve discriminability by effectively expanding the model's attention scopes. We visualize the attention results of CycleTrans and its alternatives with Grad-CAM [69] in Fig. 5. Fig. 5(a) shows the heat maps of the baseline. It only focuses on the information of a small region that can be generalized across modalities. However, such information is not sufficient for ReID. Fig. 5(b) and (c) show the results of CycleTrans without and with cycle construction. Benefiting from the first KCM, alternative (b) can capture more information for VI-ReID. But without cycle construction, its attention is likely to become

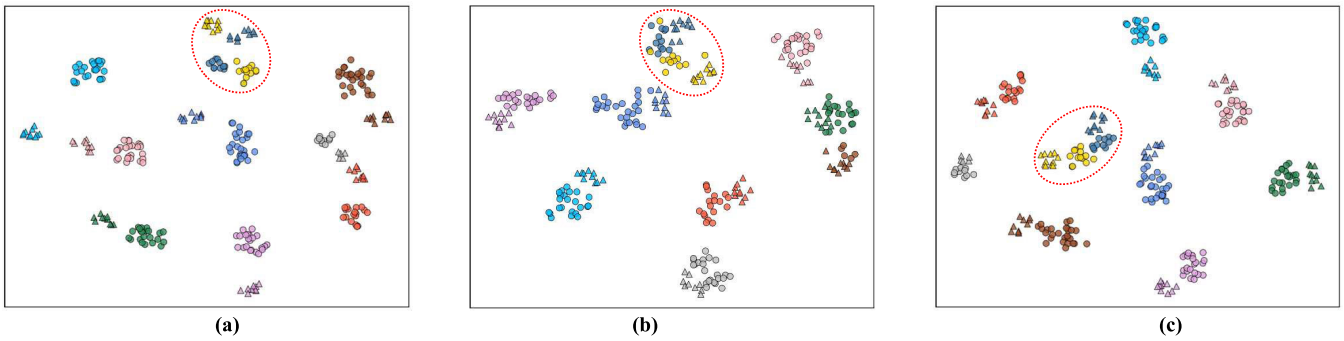

Fig. 4. Feature visualizations. Circles and triangles denote the features of visible and infrared images, respectively, and the colors represent different identities. (a) Baseline refers to the basic setting described in Table II. The middle plot shows the results of our CycleTrans (b) without cycle construction. Compared to the other two models, our (c) CycleTrans can well cluster features of different modalities but with the same identity. It also exhibits more clear semantic margins between identities.

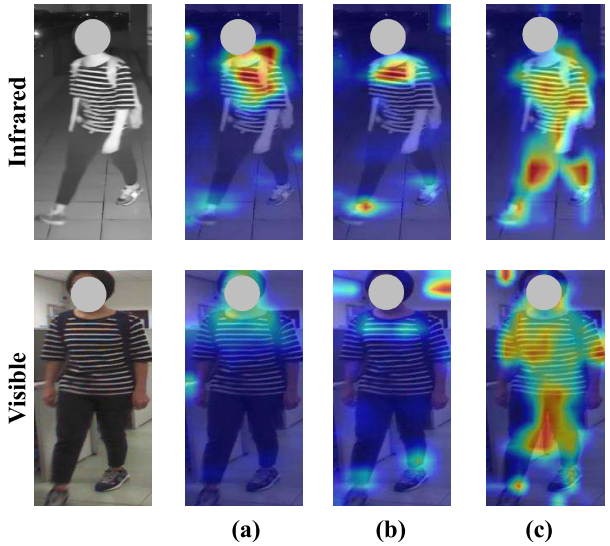

Fig. 5. Visualizations of attention results. (a) Baseline refers to the basic setting described in Table II. (b) w/o Cycle refers to the CycleTrans without cycle construction. Compared to the other two alternatives, (c) CycleTrans can grasp more details for VI-ReID.

noisy and sparse, e.g., attending to the background areas. In stark contrast, CycleTrans [Fig. 5(c)] can grasp more comprehensive and salient visual cues, and this visual information can also be well aligned across modalities. Furthermore, we can observe from Fig. 5(c) that the source of the knowledge captured by CycleTrans is the appearance of a pedestrian. According to (7), prototypes are aggregated based on their correlation with modality-relevant features. To this end, heat maps can well illustrate where the knowledge in a prototype comes from.

Overall, the visualization results well confirm the effectiveness of the proposed CycleTrans toward neutral yet discriminative feature learning for VI-ReID.

VI. Conclusion

In this article, we aim to address the modality gap in VI-ReID via learning neutral yet discriminative features. To approach this target, we propose a cycle-construction-based model for VI-ReID, termed CycleTrans. Specifically, CycleTrans first use a novel KCM to mine salient semantics from convolution feature maps based on pseudo anchors. Afterward, we propose a DMM to transform these semantics into neutral features based on the modality-irrelevant prototypes. To ensure the descriptive power of the neutral features, feature cycle constructions are performed via another two KCMs sharing the same parameters. To validate our CycleTrans, we conduct extensive experiments on two highly competitive benchmarks, namely SYSU-MM01 and RegDB. The experimental results not only report the new SOTA performance achieved by CycleTrans with great advantages to existing methods, e.g., $+1.88\%$ Rank-1 and $+1.1\%$ Rank-1 on SYSU-MM01 and RegDB, but also greatly validate the effectiveness of our method toward the modality gap.

References

[1] A. Wu, W. Zheng, H. Yu, S. Gong, and J. Lai, "RGB-infrared cross-modality person re-identification," in *Proc. ICCV*, 2017, pp. 5390–5399.

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

[3] M. Ye, X. Lan, and Q. Leng, "Modality-aware collaborative learning for visible thermal person re-identification," in *Proc. ACM Multimedia*, L. Amsaleg et al., Eds. 2019, pp. 347–355.

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

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

[6] C. Fu, Y. Hu, X. Wu, H. Shi, T. Mei, and R. He, "CM-NAS: Cross-modality neural architecture search for visible-infrared person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 11803–11812.

[7] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. H. Hoi, "Deep learning for person re-identification: A survey and outlook," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 6, pp. 2872–2893, Jun. 2022.

[8] Y. Huang, Q. Wu, J. Xu, Y. Zhong, P. Zhang, and Z. Zhang, "Alleviating modality bias training for infrared-visible person re-identification," *IEEE Trans. Multimedia*, vol. 24, pp. 1570–1582, 2022.

[9] Y. Ling et al., "Cross-modality Earth Mover's distance for visible thermal person re-identification," 2022, *arXiv:2203.01675*.

[10] Z. Huang, J. Liu, L. Li, K. Zheng, and Z. Zha, "Modality-adaptive mixup and invariant decomposition for RGB-infrared person re-identification," in *Proc. AAAI*, 2022, pp. 1034–1042.

[11] Z. Zhang, H. Zhang, and S. Liu, "Person re-identification using heterogeneous local graph attention networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 12136–12145.

[12] Y. Yan et al., "Weakening the influence of clothing: Universal clothing attribute disentanglement for person re-identification," in *Proc. IJCAI*, L. D. Raedt, Ed. 2022, pp. 1523–1529.

[13] X.-Y. Jing et al., "Super-resolution person re-identification with semi-coupled low-rank discriminant dictionary learning," *IEEE Trans. Image Process.*, vol. 26, no. 3, pp. 1363–1378, Mar. 2017.

[14] Y. Li, J. He, T. Zhang, X. Liu, Y. Zhang, and F. Wu, "Diverse part discovery: Occluded person re-identification with part-aware transformer," in *Proc. CVPR*, 2021, pp. 2898–2907.

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

[16] X. Hao, S. Zhao, M. Ye, and J. Shen, "Cross-modality person re-identification via modality confusion and center aggregation," in *Proc. ICCV*, 2021, pp. 16383–16392.

[17] D. Nguyen, H. Hong, K. Kim, and K. Park, "Person recognition system based on a combination of body images from visible light and thermal cameras," *Sensors*, vol. 17, no. 3, p. 605, Mar. 2017.

[18] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, "Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 501–518.

[19] Z. Zhong, L. Zheng, S. Li, and Y. Yang, "Generalizing a person retrieval model hetero- and homogeneously," in *Proc. ECCV*, 2018, pp. 176–192.

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

[21] C. Tay, S. Roy, and K. Yap, "AANet: Attribute attention network for person re-identifications," in *Proc. CVPR*, 2019, pp. 7134–7143.

[22] 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.*, Jun. 2019, pp. 618–626.

[23] G. Wang, T. Zhang, J. Cheng, S. Liu, Y. Yang, and Z. Hou, "RGB-infrared cross-modality person re-identification via joint pixel and feature alignment," in *Proc. ICCV*, 2019, pp. 3622–3631.

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

[25] D. Li, X. Wei, X. Hong, and Y. Gong, "Infrared-visible cross-modal person re-identification with an X modality," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, 2020, pp. 4610–4617.

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

[27] M. Ye, J. Shen, D. J. Crandall, L. Shao, and J. Luo, "Dynamic dual-attentive aggregation learning for visible-infrared person re-identification," 2020, *arXiv:2007.09314*.

[28] H. Liu, X. Tan, and X. Zhou, "Parameter sharing exploration and hetero-center triplet loss for visible-thermal person re-identification," *IEEE Trans. Multimedia*, vol. 23, pp. 4414–4425, 2021.

[29] P. Wang et al., "Deep multi-patch matching network for visible thermal person re-identification," *IEEE Trans. Multimedia*, vol. 23, pp. 1474–1488, 2021.

[30] L. Tan, P. Dai, R. Ji, and Y. Wu, "Dynamic prototype mask for occluded person re-identification," in *Proc. 30th ACM Int. Conf. Multimedia*, Oct. 2022, pp. 531–540.

[31] L. Tan et al., "Exploring invariant representation for visible-infrared person re-identification," 2023, *arXiv:2302.00884*.

[32] J. Lu, W. Zhang, and H. Yin, "Generate and purify: Efficient person data generation for re-identification," *IEEE Trans. Multimedia*, vol. 24, pp. 558–566, 2022.

[33] Z. Wei, X. Yang, N. Wang, and X. Gao, "Syncretic modality collaborative learning for visible infrared person re-identification," in *Proc. ICCV*, 2021, pp. 225–234.

[34] L. Chen, H. Yang, Q. Xu, and Z. Gao, "Harmonious attention network for person re-identification via complementarity between groups and individuals," *Neurocomputing*, vol. 453, pp. 766–776, Sep. 2021.

[35] F. Wang, W. Zuo, L. Lin, D. Zhang, and L. Zhang, "Joint learning of single-image and cross-image representations for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1288–1296.

[36] B. Bryan, Y. Gong, Y. Zhang, and C. Poellabauer, "Second-order non-local attention networks for person re-identification," in *Proc. ICCV*, 2019, pp. 3759–3768.

[37] S. Zhou, F. Wang, Z. Huang, and J. Wang, "Discriminative feature learning with consistent attention regularization for person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8039–8048.

[38] Y. Hao, N. Wang, J. Li, and X. Gao, "HSME: Hypersphere manifold embedding for visible thermal person re-identification," in *Proc. AAAI*, 2019, pp. 8385–8392.

[39] 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.

[40] S. Zhang, Y. Yang, P. Wang, G. Liang, X. Zhang, and Y. Zhang, "Attend to the difference: cross-modality person re-identification via contrastive correlation," *IEEE Trans. Image Process.*, vol. 30, pp. 8861–8872, 2021.

[41] Y. Gao et al., "MSO: Multi-feature space joint optimization network for RGB-infrared person re-identification," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 5257–5265.

[42] X. Wei, D. Li, X. Hong, W. Ke, and Y. Gong, "Co-attentive lifting for infrared-visible person re-identification," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 1028–1037.

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

[44] Z. Miao, H. Liu, W. Shi, W. Xu, and H. Ye, "Modality-aware style adaptation for RGB-infrared person re-identification," in *Proc. IJCAI*, 2021, pp. 916–922.

[45] M. Kim, S. Kim, J. Park, S. Park, and K. Sohn, "PartMix: Regularization strategy to learn part discovery for visible-infrared person re-identification," in *Proc. CVPR*, 2023, pp. 18621–18632.

[46] Z. Wei, X. Yang, N. Wang, and X. Gao, "Flexible body partition-based adversarial learning for visible infrared person re-identification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 9, pp. 4676–4687, Sep. 2022.

[47] Q. Zhang, C. Lai, J. Liu, N. Huang, and J. Han, "FMCNet: Feature-level modality compensation for visible-infrared person re-identification," in *Proc. CVPR*, 2022, pp. 7349–7358.

[48] J. Liu, Y. Sun, F. Zhu, H. Pei, Y. Yang, and W. Li, "Learning memory-augmented unidirectional metrics for cross-modality person re-identification," in *Proc. CVPR*, 2022, pp. 19344–19353.

[49] Y. Zhang, S. Zhao, Y. Kang, and J. Shen, "Modality synergy complement learning with cascaded aggregation for visible-infrared person re-identification," in *Computer Vision—ECCV* (Lecture Notes in Computer Science), vol. 13674. Tel Aviv-Yafo, Israel: Springer, 2022, pp. 462–479.

[50] Q. Wu et al., "Discover cross-modality nuances for visible-infrared person re-identification," in *Proc. CVPR*, 2021, pp. 4330–4339.

[51] Y. Zhang and H. Wang, "Diverse embedding expansion network and low-light cross-modality benchmark for visible-infrared person re-identification," in *Proc. CVPR*, 2023, pp. 2153–2162.

[52] K. Zhu, H. Guo, S. Liu, J. Wang, and M. Tang, "Learning semantics-consistent stripes with self-refinement for person re-identification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 11, pp. 8531–8542, Nov. 2023.

[53] H. Tan, X. Liu, B. Yin, and X. Li, "MHSA-Net: Multihead self-attention network for occluded person re-identification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 11, pp. 8210–8224, Nov. 2023.

[54] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, Long Beach, CA, USA, 2017, pp. 5998–6008.

[55] A. van den Oord, O. Vinyals, and K. Kavukcuoglu, "Neural discrete representation learning," in *Proc. NIPS*, I. Guyon et al., Eds., 2017, pp. 6306–6315.

[56] A. Gretton et al., "Optimal kernel choice for large-scale two-sample tests," in *Proc. NIPS*, 2012, pp. 1214–1222.

[57] X. Wang, G. Doretto, T. Sebastian, J. Rittscher, and P. H. Tu, "Shape and appearance context modeling," in *Proc. ICCV*, 2007, pp. 1–8.

[58] M. Ye, Z. Wang, X. Lan, and P. C. Yuen, "Visible thermal person re-identification via dual-constrained top-ranking," in *Proc. IJCAI*, 2018, pp. 1092–1099.

[59] 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. CVPR*, 2020, pp. 10254–10263.

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

[61] Y. Ling, Z. Zhong, Z. Luo, P. Rota, S. Li, and N. Sebe, "Class-aware modality mix and center-guided metric learning for visible-thermal person re-identification," in *Proc. ACM Multimedia*, 2020, pp. 889–897.
[62] M. Jia, Y. Zhai, S. Lu, S. Ma, and J. Zhang, "A similarity inference metric for RGB-infrared cross-modality person re-identification," in *Proc. IJCAI*, 2020, pp. 1026–1032.
[63] 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.*, Montreal, QC, Canada, Oct. 2021, pp. 13547–13556.
[64] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, 2016, pp. 770–778.
[65] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," in *Proc. AAAI*, 2020, pp. 13001–13008.
[66] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, Y. Bengio and Y. LeCun, Eds. 2015, pp. 1–15.
[67] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. ICLR*, 2021, pp. 1–22.
[68] V. der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 73–77, 2008.
[69] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," *Int. J. Comput. Vis.*, vol. 128, no. 2, pp. 336–359, Feb. 2020.

Qiong Wu received the B.E. degree in computer science and technology from Xiamen University, Xiamen, China, in 2020. He is currently pursuing the Ph.D. degree with the Institute of Artificial Intelligence and the Key Laboratory of Multimedia Trusted Perception and Efficient Computing, Ministry of Education of China, Xiamen.

His research interests include pattern recognition, machine learning, and computer vision.

Jiaer Xia received the B.E. degree in electrical engineering and automation from Donghua University, Shanghai, China, in 2021. He is currently pursuing the master's degree with Xiamen University, Xiamen, China, and the Key Laboratory of Multimedia Trusted Perception and Efficient Computing, Ministry of Education of China, Xiamen.

His research interests include visual perception, multimodal models, and domain adaptation.

Pingyang Dai received the M.S. degree in computer science and the Ph.D. degree in automation from Xiamen University, Xiamen, China, in 2003 and 2013, respectively.

He is currently a Senior Engineer with the Key Laboratory of Multimedia Trusted Perception and Efficient Computing, Ministry of Education of China, Xiamen, and the School of Informatics, Xiamen University. His research interests include computer vision and machine learning.

Yiyi Zhou (Member, IEEE) received the Ph.D. degree from Xiamen University, Xiamen, China, in 2019, under the supervision of Prof. Rongrong Ji.

He was a Post-Doctoral Research Search Fellow with Xiamen University from 2019 to 2022. He is currently an Associate Professor with the School of Informatics and the Institute of Artificial Intelligence, Xiamen University. His research interests include vision-language learning and computer vision.

Yongjian Wu received the master's degree in computer science from Wuhan University, Wuhan, China, in 2008.

He is currently an Expert Researcher and the Director of the Youtu Laboratory, Tencent Company Ltd., Shanghai, China. His research interests include face recognition, image understanding, and large-scale data processing.

Rongrong Ji (Senior Member, IEEE) is currently a Nanqiang Distinguished Professor with Xiamen University, Xiamen, China, the Deputy Director of the Office of Science and Technology, Xiamen University, and the Director of the Media Analytics and Computing Laboratory, Xiamen. He has published more than 50 papers in ACM/IEEE TRANSACTIONS, including IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE and *IJCV* and more than 100 full papers on top-tier conferences, such as CVPR and NeurIPS. His publications have got over 20K citations in Google Scholar. His research interests include computer vision, multimedia analysis, and machine learning.

Prof. Ji is an Advisory Member of Artificial Intelligence Construction in the Electronic Information Education Committee of the National Ministry of Education. He was a recipient of the Best Paper Award from ACM Multimedia 2011. He was awarded as the National Science Foundation for Excellent Young Scholars in 2014, the National Ten Thousand Plan for Young Top Talents in 2017, and the National Science Foundation for Distinguished Young Scholars in 2020. He has served as an Area Chair for top-tier conferences, such as CVPR and ACM Multimedia.