Visible-Infrared Person Re-Identification With Modality-Specific Memory Network

Yulin Li, Tianzhu Zhang, *Member, IEEE*, Xiang Liu, Qi Tian, *Fellow, IEEE*, Yongdong Zhang, *Senior Member, IEEE*, and Feng Wu, *Fellow, IEEE*

***Abstract*— Visible-infrared person re-identification (VI-ReID) is challenging due to the large modality discrepancy between visible and infrared images. Existing methods mainly focus on learning modality-shared representations by embedding images from different modalities into a common feature space, in which some discriminative modality information is discarded. Different from these methods, in this paper, we propose a novel Modality-Specific Memory Network (MSMNet) to complete the missing modality information and aggregate visible and infrared modality features into a unified feature space for the VI-ReID task. The proposed model enjoys several merits. First, it can exploit the missing modality information to alleviate the modality discrepancy when only the single-modality input is provided. To the best of our knowledge, this is the first work to exploit the missing modality information completion and alleviate the modality discrepancy with the memory network. Second, to guide the learning process of the memory network, we design three effective learning strategies, including feature consistency, memory representativeness and structural alignment. By incorporating these learning strategies in a unified model, the memory network can be well learned to propagate identity-related information between modalities and boost the VI-ReID performance. Extensive experimental results on two standard benchmarks (SYSU-MM01 and RegDB) demonstrate that the proposed MSMNet performs favorably against state-of-the-art methods.**

***Index Terms*— Visible-infrared person re-identification, modality discrepancy, modality-specific memory network, missing modality information completion.**

I. Introduction

PERSON re-identification (ReID) aims to match images of a person captured from non-overlapping camera views [1], [2], [3]. It has gained increasing attention in computer vision area for both research and application. Most existing person ReID methods [4], [5], [6], [7], [8], [9] focus on matching pedestrian images captured by visible cameras during day time, which can be formulated as a single-modality matching problem. However, it is difficult for visible cameras to capture valid appearance characteristics of a person under poor illumination environments, e.g., at night, which limits the applicability of these methods in practical. To overcome this limitation, many modern surveillance cameras can automatically switch between visible and infrared modes when the lighting conditions change significantly. Therefore, it becomes essential to study the visible-infrared person re-identification (VI-ReID) in real-world scenarios, which is a cross-modality matching problem [10].

Compared to the widely studied conventional ReID task, VI-ReID encounters the additional modality discrepancy problem resulting from the different wavelength ranges used in the imaging process. Eventually, modality discrepancy leads to a situation where the intra-class distance is larger than the inter-class distance in VI-ReID. To overcome the modality discrepancy problem, various methods have been proposed. These methods can be generally categorized into two major categories: modality-shared feature learning methods [11], [12], [13], [14], [15], [16] and modality information completion methods [17], [18]. The modality-shared feature learning methods try to embed images of different modalities into a shareable feature space, then cross modality image retrieval is achieved based on the sharable feature representation. Wu et al. [10] propose a deep zero-padding framework for modality-shared feature learning. Some recent studies [12], [16], [19] exploit two-stream CNNs, including modality-specific shallow layers and shared deeper layers to learn a common feature space. However, since visible and infrared images have quite different appearances, how to directly embed images of different modalities into a shareable common feature space is still a difficult problem. Besides, the discriminability of the feature representation of these methods is limited for the reason that some discriminative modality information, such as colors of visible images, is regarded as redundant information. To address this limitation, the modality information completion methods [17], [18] have been proposed, and the goal is to make up the modality information from one modality to another. D^2RL [17] generates multi-spectral images to compensate for the missing modality information by utilizing generative adversarial network (GAN) [20]. Another representative work, cm-SSFT [18], tries

Manuscript received 31 March 2022; revised 29 August 2022 and 9 October 2022; accepted 17 October 2022. Date of publication 11 November 2022; date of current version 17 November 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 62022078, Grant 62021001, and Grant 12150007; in part by the National Key Research and Development Program of China under Grant 2018YFB0804204; and in part by the National Defense Basic Scientific Research Program under Grant JCKY2020903B002. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Clinton Fookes. *(Corresponding author: Tianzhu Zhang.)*

Yulin Li is with the School of Cyber Science and Technology, University of Science and Technology of China, Hefei 230027, China (e-mail: liyulin@mail.ustc.edu.cn).

Tianzhu Zhang is with the Department of Automation, School of Information Science and Technology, University of Science and Technology of China, Hefei 230027, China (e-mail: tzzhang@ustc.edu.cn).

Xiang Liu is with the School of Cyberspace Security, Dongguan University of Technology, Dongguan 523808, China (e-mail: liuxiang@dgut.edu.cn).

Qi Tian is with Cloud BU, Huawei, Shenzhen 518129, China (e-mail: tian.qi1@huawei.com).

Yongdong Zhang and Feng Wu are with the Department of Electronic Engineering and Information Science, School of Information Science and Technology, University of Science and Technology of China, Hefei 230027, China (e-mail: zhyd73@ustc.edu.cn; fengwu@ustc.edu.cn).

Digital Object Identifier 10.1109/TIP.2022.3220408

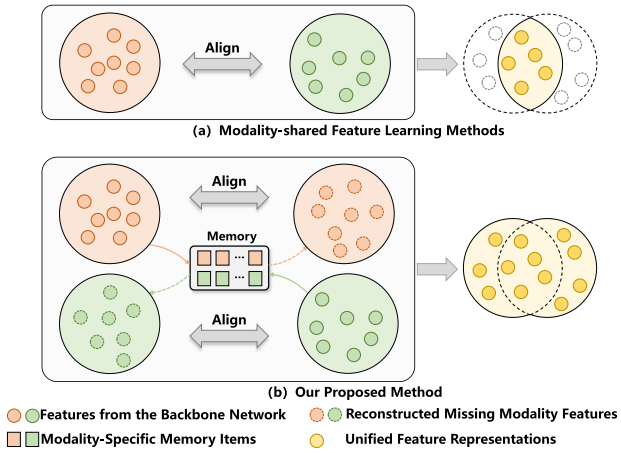

Fig. 1. Comparison of modality-shared feature learning methods and our method. Red circles refer to features of visible modality and green circles refer to features of infrared modality. (a) Modality-shared feature learning methods try to embed images from different modalities into a sharable feature space. (b) The proposed method alleviates the modality discrepancy by completing the missing modality information and aggregating features from both visible and infrared modalities into a unified feature space.

to complete the missing modality information of each sample based on the inter-modality and intra-modality affinity between different samples. However, due to the uncertainty of image generation, the generative model inevitably introduces noisy generated samples, which are harmful for the missing modality information completion. For cm-SSFT, it needs to utilize the information of other query samples to obtain the missing modality feature during the test stage, which is prohibited in formal ReID test protocol. Thus, with the single modality input only, it is difficult to complete the missing modality information and bridge the modality discrepancy.

Based on the above discussions, the key challenge in VI-ReID is how to bridge the modality discrepancy between the visible and infrared modalities. Previous methods [17], [18] have proven that an effective way to achieve this goal is to complete the missing modality information. Recently, Tang et al. [21] propose generalized deep transfer networks (DTNs) to transfer label information across heterogeneous domains. On the basis of these insight, as shown in Figure 1 (b), an intuitive idea is to design a memory network by storing prototypical features of each modality as a bridge to propagate information between visible and infrared modalities. Each item in the memory network corresponds to prototypical features of the visible/infrared modality. The missing modality features can be reconstructed by retrieving items from the other modality in the memory network. By aggregating the original and missing modality features and obtaining a unified feature space, the modality discrepancy can be well alleviated. To make the memory network serve as a bridge between visible and infrared modalities and complete modality information, We sum up the following three important characteristics that should be considered. (1) **Consistency**. Given an input visible (infrared) feature, we can reconstruct the corresponding infrared (visible) feature from the memory network ideally. That is, the reconstructed feature should be consistent with the feature obtained from the backbone network. (2) **Representativeness**. The memory items should be representative enough for each modality, so that they can be used as the proxy of each modality for modality information propagation. (3) **Alignment**. The memory items from visible and infrared modalities should be well aligned. That is, there should exist a one-to-one correspondence between paired memory items from different modalities. Only in this way can the memory network be used as the bridge to reconstruct the missing modality information.

Inspired by the above insights, we propose a novel **Modality-Specific Memory Network (MSMNet)** to achieve modality information completion for VI-ReID. Specifically, we propose a memory network containing memory items arranged in pairs: visible memory items and infrared memory items. To complete the missing modality feature, take the visible (infrared) feature as input, we calculate the similarity with the visible (infrared) memory items. The similarity is then used to aggregate infrared (visible) memory items to obtain feature representations of infrared (visible) modality. In this way, we can get the missing modality information and leverage information from both visible and infrared modalities even when the single-modality input is provided only. To learn the memory network well, we introduce three learning strategies to deal with the characteristics discussed above. For the **consistency** characteristic, two modality discriminators are adopted to make the reconstructed features from the memory network consistent with features extracted from the backbone network. For the **representativeness** characteristic, we introduce a feature reconstruction loss. The feature reconstruction loss encourages features can be reconstructed by memory items from the same modality, thus making the memory network save the prototypical representation. For the **alignment** characteristic, we introduce a structural alignment loss making the paired memory items from visible and infrared modalities save the corresponding prototypical features. In this way, memory items of two modalities can be well learned to propagate identity-related information between modalities and boost the VI-ReID performance.

The contributions of our method could be summarized into three-fold:

- We introduce a novel Modality-Specific Memory Network (MSMNet) to achieve missing modality information completion for the VI-ReID task. To the best of our knowledge, this is the first work by exploring the memory network as the bridge to reduce the modality discrepancy for VI-ReID task.
- To learn the memory network, we introduce three learning strategies, including feature consistency, memory representativeness, and structural alignment, to make paired memory items representative and structural alignment with each other. We conduct sufficient qualitative analysis and also explain the insight of the proposed modality-specific memory network.
- We achieve a new state-of-the-art on two standard VI-ReID benchmarks and demonstrate the effectiveness of our approach through extensive experiments.

II. Related Work

In this section, we briefly overview methods that are related to person ReID, visible-infrared person ReID and memory network respectively.

A. Person ReID

Person ReID aims to match images of a person captured from non-overlapping visible camera views [1], [2], [3]. Existing person ReID works can be summarized to hand-crafted descriptors [22], [23], [24], metric learning based methods [4], [25], [26], [27], [28] and deep learning based methods [5], [6], [7], [8], [29], [30], [31], [32]. In the early works based on hand-crafted descriptors, Li et al. [22] extract local color descriptors from patches and aggregate them using hierarchical Gaussianization [33] to capture spatial information. In [24], Liao et al.propose the local maximal occurrence (LOMO) descriptor, which includes the color and SILTP histograms. Metric learning based methods focus on proper loss functions, like the contrastive loss [34] and the triplet loss [27]. In [28], an improved triplet loss function is proposed to pull the instances of the same identity closer, and push the instances belonging to different identities farther from each other in the learned feature space. Recently, with the advancement of deep learning, person Re-ID has achieved inspiring performance. Wang et al. [35] introduce the problem of learning person ReID models from videos with weak supervision and present a new co-person attention mechanism to utilize relationships between videos with common person identities. Wu et al. [36] propose a novel multi-level Context-aware Part Attention (CPA) model to learn discriminative and robust local part features. In [37], a novel online label co-refining framework is proposed for robust Re-ID model learning. Among these works based on deep learning, leveraging local features extracted from human body parts has been the mainstream for robust feature learning. Several works [5], [8], [9], [31], [32] employ hand-crafted splitting, pose estimation models or attention mechanisms to obtain part features and perform part-level feature alignment. Sun et al. [5] uniformly partition the feature map and learn part-level features by multiple classifiers. Kalayeh et al. [9] extract several region parts with human parsing methods and assemble final discriminative representations with part-level features. Zhao et al. [8] and Liu et al. [38] extract part-level features by designing modules based on attention mechanisms. Although having achieved great success in the conventional person ReID, these methods are developed for visible modality and cannot perform well for the VI-ReID task, which limits the applicability in poor lighting surveillance scenarios.

B. Visible-Infrared Person ReID

Different from conventional person ReID, VI-ReID aims to match visible and infrared person images captured by disjoint cameras [10]. Current VI-ReID methods can be generally categorized into two major categories: modality-shared feature learning methods [10], [11], [12], [13], [14], [15], [16], [39], [40], [41], [42] and modality information completion methods [17], [18]. As for the modality-shared feature learning methods, some methods mainly focus on designing an effective feature extractor to extract modality shared and discriminative representation [10], [11], [13], [41]. Wu et al. [10] propose a deep zero-padding network to embed features in a common space, and build the first large-scale VI-ReID dataset named SYSU-MM01. Ye et al. [11] introduce a two-stream network to handle the two modalities respectively and a bi-directional dual-constrained top-ranking loss to learn discriminative feature representations. Chen et al. [43] introduce a structure-aware positional transformer network to learn semantic-aware sharable modality features by utilizing the structural and positional information. Recently, Fu et al. [13] and Chen et al. [41] exploit the neural architecture search technique to find the optimal network architecture to minimize the modality discrepancy. Some image translation based methods adopt GAN to reduce the distribution divergence of features of different modalities. Dai et al. [39] introduce a cross-modality Generative Adversarial Network (cmGAN) by exploiting a modality discriminator to reduce the distribution divergence of visible and infrared features. Wang et al. [44] propose an Alignment Generative Adversarial Network (AlignGAN) by exploiting pixel alignment and feature alignment jointly. Several studies [42], [45] try to generate intermediate modality to further eliminate differences across different modalities. Li et al. [45] introduce an auxiliary X modality and convert the cross-modality learning as an X-Infrared-Visible three-modality learning problem. Wei et al. [42] propose a novel syncretic modality collaborative learning (SMCL) model to bridge the cross modality gap. Some other studies exploit two-stream CNNs with deep metric learning [11], [12], [16], [40] or the attention mechanism [14], [15] to learn modality-shared feature representations. In [46], a novel bi-directional exponential angular triplet loss is proposed to address the difficulty in learning angularly discriminative feature embedding. Zhang et al. [47] design a Angle metric space for solving VI-ReID problem and propose a cyclic projection network (CPN) for implementing the Angular metric learning. Ye et al. [16] introduce the hierarchical inter-modality metric learning technique to learn modality-shared embedding. Liu et al. [12] propose the hetero-center triplet loss to constrain feature centers, which can effectively reduce inter-modality discrepancy. Ye et al. [15], [48] design a dynamic dual-attentive aggregation module by mining both intra-modality part-level and cross-modality graph-level contextual cues. However, the discriminability of the feature representation of these methods is limited for the reason that some discriminative information, such as colors of visible images are regarded as redundant information.

Recently, some modality information completion methods [17], [18] try to make up the modality information from one modality to another to boost the VI-ReID performance. Wang et al. [17] apply GANs to generate missing modality information and extend the input of the feature extractor to four dimensions. Lu et al. [18] try to complete the missing modality information of each sample based on the inter-modality and intra-modality affinity between different samples. Different from existing methods, we propose a modality-specific memory network as the bridge to complete the modality information and alleviate the modality discrepancy for the VI-ReID task.

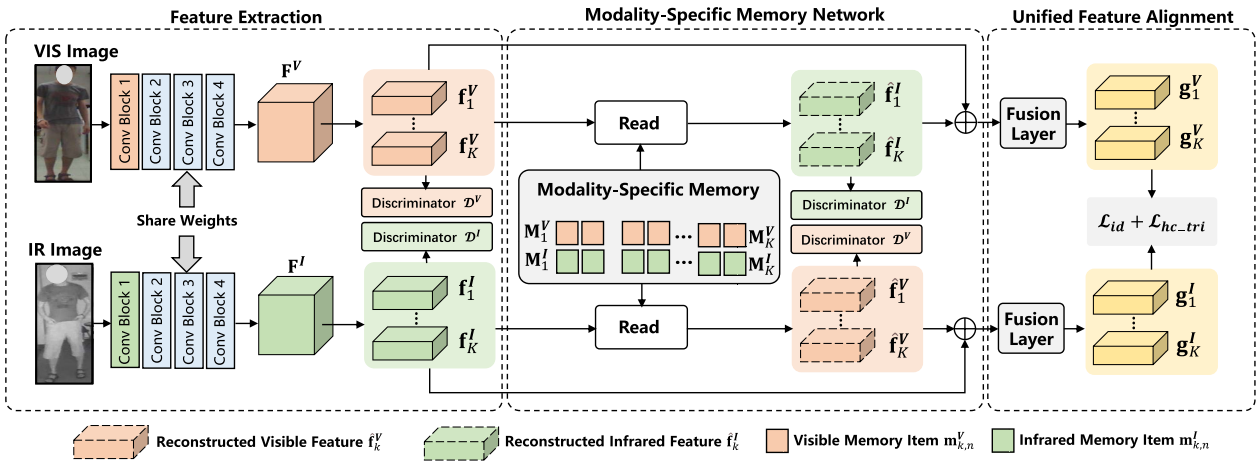

Fig. 2. Overall architecture of our proposed Modality-Specific Memory Network (MSMNet), which consists of three parts: feature extraction, modality-specific memory network and unified feature alignment. Given the input feature, the memory network is utilized to complete its corresponding missing modality feature. Then, we aggregate the input feature and its missing modality feature to obtain a unified feature representation. ID-based classification loss and hetero-center triplet loss are simultaneously adopted for unified feature alignment learning.

C. Memory Network

Memory network is a scheme to store information in external memory and read the relevant contents from the memory [49], [50]. It is first introduced by Weston et al. [49] to reason with an additional memory component for the task of question and answering. Miller et al. [51] further introduce key-value paired memory structure where they address relevant memory items by keys, and the corresponding values are subsequently returned. Due to its high flexibility of storing different knowledge in key-value pairs, the memory network has been widely adopted in solving various vision problems such as video object segmentation [52], [53], domain adaptation [54], image colorization [55], anomaly detection [56], [57] and video-based person ReID [53]. Seoung et al. [52] propose a space-time memory network that stores the past frames with object masks, then the current frame is segmented using the mask information in the memory network. Vibashan et al. [54] propose memory guided category-specific attention maps for category-aware domain adaption. Wang et al. [58] develop a structured and explicit memory architecture that allows agents to access to its past percepts and explore environment layouts for vision-language navigation. Lu et al. [59] integrate a novel graph memory mechanism to efficiently adapt the segmentation network to specific videos without catastrophic inference or finetune. Chanho et al. [53] introcude the spatial and temporal memories to refine frame-level person representions and to aggregate the refined frame-level features into a sequence-level person representation. Inspired by these works, our work utilizes the memory network for storing prototypical features of each modality. By using the memory network as a bridge, we can complete the missing modality information and leverage information from both visible and infrared modalities to obtain a unified feature space for VI-ReID task.

III. Our Approach

In this section, we introduce the proposed Modality-Specific Memory Network (MSMNet) in detail. We first give a brief introduction of the feature extraction process in section III-A. In section III-B, we describe each component of the memory network and the process of modality information completion in detail. Finally, we introduce the proposed learning strategies for the memory network in Section III-C and the training and inference process in section III-D.

A. Feature Extraction

Following previous works [12], [60], we adopt two-stream CNNs to extract feature maps, where the first two convolutional blocks are different to capture modality-specific low-level feature patterns and the parameters of the deep convolutional blocks are shared for two modalities. The architecture of the feature extractor is shown in Figure 2. Given a pair of images with the same identity, we can extract feature maps $\mathbf{F}^{\boldsymbol{V}} \in \mathbb{R}^{H\times W\times C}$ for the visible image and $\mathbf{F}^{\boldsymbol{I}} \in \mathbb{R}^{H\times W\times C}$ for the infrared image, where H, W, C are the height, width and number of channels, respectively. Then $\mathbf{F}^{\boldsymbol{V}}$ and $\mathbf{F}^{\boldsymbol{I}}$ are horizontally partitioned into K parts and each part is averagely pooled to obtain the part feature vectors $\mathbf{f}_k^{\boldsymbol{V}} \in \mathbb{R}^{1\times C}$ and $\mathbf{f}_k^{\boldsymbol{I}} \in \mathbb{R}^{1\times C}$ respectively, where $k = 1, 2, \cdots, K$. To make part features from two modalities discriminative, we add the classification loss $\mathcal{L}_{\text{sid}}$ with two modality-specific classifiers:

$$\mathcal{L}_{sid} = -\sum_{k=1}^{K}\left[\mathbf{y}^{\boldsymbol{V}} \log P\left(\mathbf{f}_k^{\boldsymbol{V}}; \theta_k^{\boldsymbol{V}}\right) + \mathbf{y}^{\boldsymbol{I}} \log P\left(\mathbf{f}_k^{\boldsymbol{I}}; \theta_k^{\boldsymbol{I}}\right)\right], \quad (1)$$

where $\mathbf{y}^{\boldsymbol{V}}$ and $\mathbf{y}^{\boldsymbol{I}}$ denote the identity labels of $\mathbf{f}_k^{\boldsymbol{V}}$ and $\mathbf{f}_k^{\boldsymbol{I}}$, and $P\left(\mathbf{f}_k^{\boldsymbol{V}}; \theta_k^{\boldsymbol{V}}\right)$ and $P\left(\mathbf{f}_k^{\boldsymbol{I}}; \theta_k^{\boldsymbol{I}}\right)$ are probability predictions of visible and infrared classifiers with parameters $\theta_k^{\boldsymbol{V}}$ and $\theta_k^{\boldsymbol{I}}$, respectively. The classification loss $\mathcal{L}_{sid}$ ensures that $\mathbf{f}_k^{\boldsymbol{V}}$ and $\mathbf{f}_k^{\boldsymbol{I}}$ can maintain discriminative identity information of visible modality and infrared modality respectively.

B. Modality-Specific Memory Network

To accurately memorize and propagate information between visible and infrared modalities and further obtain the unified

feature representation, we introduce a modality-specific memory network keeping prototypical features of each modality. Given an input image, we can read from the memory network to reconstruct its missing modality feature. For example, given a visible image, we intend to reconstruct its corresponding infrared feature. To achieve this goal, we introduce paired modality-specific memory items, $\mathbf{M}_k^V \in \mathbb{R}^{C\times N}$ for visible modality and $\mathbf{M}_k^I \in \mathbb{R}^{C\times N}$ for infrared modality, corresponding to k-th part of visible and infrared images. Here, N represents the number of memory items for each part to model the part variation. The modality-specific memory items are arranged in pairs, and individual items correspond to prototypical features of visible or infrared modality.

1) Memory Read: Take visible feature $\mathbf{f}_k^V$ as input, we introduce how to read from the memory network and obtain the reconstructed missing modality feature. To read the appropriate infrared memory items, we first compute the similarity between $\mathbf{f}_k^V$ and each visible memory item:

$$\mathbf{s}_{k,n}^V = \frac{\mathbf{f}_k^V \cdot \mathbf{m}_{k,n}^V}{\left\|\mathbf{f}_k^V\right\|_2 \cdot \left\|\mathbf{m}_{k,n}^V\right\|_2}, \tag{2}$$

where $\mathbf{m}_{k,n}^V$ represents the n-th memory item from $\mathbf{M}_k^V$ and $n = 1, 2, \cdots, N$. Then, the matching probability is obtained using the Softmax function as follows,

$$\alpha_{k,n}^V = \frac{exp\left(\mathbf{s}_{k,n}^V/\tau\right)}{\sum_{n=1}^N exp\left(\mathbf{s}_{k,n}^V/\tau\right)}, \tag{3}$$

where τ is the temperature parameter. By calculating the matching probability over all visible memory items, the reading weight $\mathbf{A}_k^V = \left[\alpha_{k,1}^V; \alpha_{k,2}^V; \ldots; \alpha_{k,N}^V\right]$ for $\mathbf{f}_k^V$ can be obtained. Based on the reading weight, the reconstructed infrared feature for $\mathbf{f}_k^V$ can be calculated by taking a weighted aggregation of the infrared memory items:

$$\hat{\mathbf{f}}_k^I = \sum_{n=1}^N \alpha_{k,n}^V \mathbf{m}_{k,n}^I, \tag{4}$$

where $\hat{\mathbf{f}}_k^I$ is the reconstructed infrared feature based on $\mathbf{f}_k^V$ and the infrared memory items $\mathbf{M}_k^I$, and $\mathbf{m}_{k,n}^I$ represents the n-th memory item from $\mathbf{M}_k^I$.

The same procedure can be applied for the infrared feature $\mathbf{f}_k^I$ and we can get its corresponding reading weight $\mathbf{A}_k^I$ by calculating the similarity with $\mathbf{M}_k^I$. Its corresponding visible feature $\hat{\mathbf{f}}_k^V$ can be also reconstructed as:

$$\hat{\mathbf{f}}_k^V = \sum_{n=1}^N \alpha_{k,n}^I \mathbf{m}_{k,n}^V. \tag{5}$$

Figure 3 illustrates the whole process of the memory read operation.

2) Unified Feature Alignment: After obtaining the reconstructed missing modality features, we add the reconstructed missing modality features to the input features to obtain the unified feature representation:

$$\mathbf{g}_k^V = h\left(\mathbf{f}_k^V + \hat{\mathbf{f}}_k^I\right), \quad \mathbf{g}_k^I = h\left(\mathbf{f}_k^I + \hat{\mathbf{f}}_k^V\right), \tag{6}$$

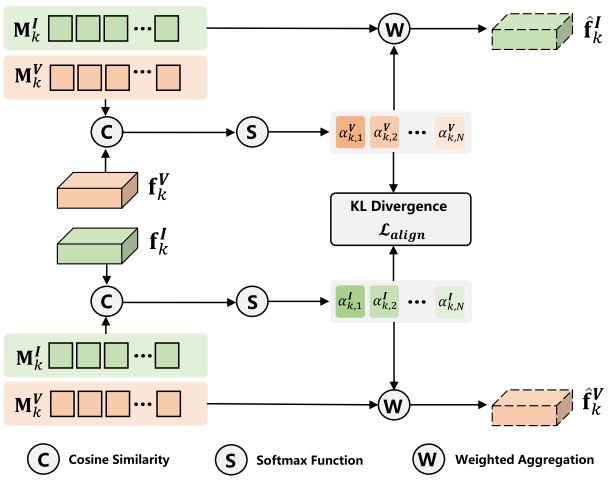

Fig. 3. Read operations for the modality-specific memory network and the explanation of the structural alignment loss. Note that the two $\mathbf{M}_k^V$/$\mathbf{M}_k^I$ in the figure are the same.

where $h(\cdot)$ is a fusion layer consisting of a linear layer and a batch normalization layer. By fusing the original features and the reconstructed modality features, the visible and infrared images are naturally embedded into a common feature space. We adopt the cross-entropy loss and the hetero-center triplet loss [12] simultaneously for unified feature alignment learning. For the classification loss, we adopt modality-shared classifiers to predict the identities:

$$\mathcal{L}_{id} = -\sum_{k=1}^K \left[\mathbf{y}^V \log P\left(\mathbf{g}_k^V; \theta_k\right) + \mathbf{y}^I \log P\left(\mathbf{g}_k^I; \theta_k\right)\right], \tag{7}$$

where $\mathbf{y}^V$ and $\mathbf{y}^I$ denote the identity label of $\mathbf{g}_k^V$ and $\mathbf{g}_k^I$, and $P\left(\mathbf{g}_k^V; \theta_k\right)$ and $P\left(\mathbf{g}_k^I; \theta_k\right)$ are probability predictions of modality-shared classifiers with parameters θ_k.

3) Hetero-Center Triplet Loss: With the proposed modality-specific memory network, we can reconstruct the missing modality features and obtain the unified feature representations $\mathbf{g}_{m,k}^V$ and $\mathbf{g}_{m,k}^I$, respectively. Here $m \in 1, 2, \ldots, M$ represent the m-th visible image and infrared image in the current mini-batch. Notice that the subscript m is **omitted** for simplicity in the other part of the paper. In order to alleviate the modality discrepancy in the unified feature space, we adopt the hetero-center triplet loss [12] to align $\mathbf{g}_{m,k}^V$ and $\mathbf{g}_{m,k}^I$. First, for all features in a mini-batch, the centers for the features of every identity from each modality are computed as:

$$\mathbf{c}_{i,k}^V = \frac{1}{|\mathcal{P}_i|} \sum_{\mathbf{g}_{m,k}^V \in \mathcal{P}_i} \mathbf{g}_{m,k}^V,$$
$$\mathbf{c}_{i,k}^I = \frac{1}{|\mathcal{Q}_i|} \sum_{\mathbf{g}_{m,k}^I \in \mathcal{Q}_i} \mathbf{g}_{m,k}^I, \tag{8}$$

where $\mathcal{P}_i$ denotes the visble image set with identity label i and $\mathcal{Q}_i$ denotes the infrared image set with identity label i in the current mini-batch, and $|\cdot|$ denotes the number of images in the set. Therefore, $\mathbf{c}_{i,k}^V$ and $\mathbf{c}_{i,k}^I$ denote the visible image center and infrared image center for k-th part with identity label i. The goal of hetero-center triplet loss is to make those features

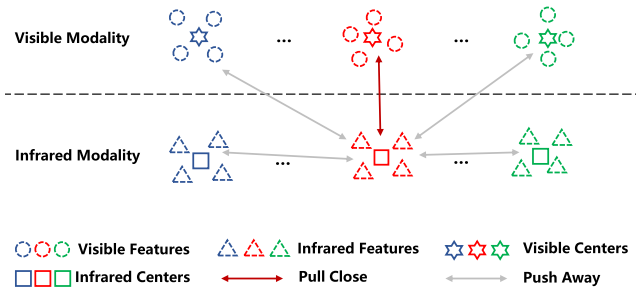

Fig. 4. The illustration of the hetero-center triplet loss, which aims at pulling close those centers with the same identity label from different modalities, while pushing away those centers with different identity labels regardless of which modality it is from. Different colors denote different identities.

from the same identity close to each other, while those features from different identities far from each other. Therefore, based on the calculated centers, the hetero-center triplet loss is defined as:

$$\begin{aligned}\mathcal{L}_{hc\_tri} = &\sum_{i=1}^{M}\sum_{k=1}^{K}\Big[\rho + \left\|\mathbf{c}_{i,k}^{\boldsymbol{V}}\right. \\ &\left.-\mathbf{c}_{i,k}^{\boldsymbol{I}}\right\|_2 - \min_{*\in\{\boldsymbol{V},\boldsymbol{I}\}}\left\|\mathbf{c}_{i,k}^{\boldsymbol{V}} - \mathbf{c}_{j,k}^{*}\right\|_2\Big]_+ \\ &+\sum_{i=1}^{M}\sum_{k=1}^{K}\Big[\rho + \left\|\mathbf{c}_{i,k}^{\boldsymbol{I}}\right. \\ &\left.-\mathbf{c}_{i,k}^{\boldsymbol{V}}\right\|_2 - \min_{*\in\{\boldsymbol{V},\boldsymbol{I}\}}\left\|\mathbf{c}_{i,k}^{\boldsymbol{I}} - \mathbf{c}_{j,k}^{*}\right\|_2\Big]_+, \end{aligned} \tag{9}$$

where $[\cdot]_+$ indicates the max function $max(0,\cdot)$, ρ is the margin for the triplet loss, and $\mathbf{c}_j^*$ denotes an image center with the different identity label j. For each identity, $\mathcal{L}_{hc\_tri}$ concentrates on cross-modality positive pair and the mined hardest negative pair in visible and infrared modalities. The illustration of the hetero-center triplet loss is shown in Figure 9.

The original triplet loss $\mathcal{L}_{tri}$ with batch hard stategy [27] computes the loss by comparison of the anchor to all the other samples, which is too strict about constraining the pairwise distance if there exist some outliers. These outliers would form the adverse triplet and destroy other pairwise distances in metric learning. Therefore, the center of each person is adopted as the identity agent to compute triplet loss in [12]. In this way, we can relax the strict constraint by replacing the comparison of the anchor to all the other samples with the anchor center to all the other centers. $\mathcal{L}_{hc\_tri}$ also preserves the property of handling both the intra-class and inter-class variations simultaneously on visible and infrared modalities in the common feature space. According to [12], the hetero-center triplet loss can also reduce the computational cost during training.

4) Initialization and Update of the Memory Network: We set memory items $\mathbf{M}_k^{\boldsymbol{V}}$ and $\mathbf{M}_k^{\boldsymbol{I}}$ aas learnable parameters and initialize them with Kaiming initialization. During the training stage, the following proposed learning strategies can guide the updating process of memory items through back-propagation.

C. Memory Learning Strategies Designing

The proposed MSMNet explores a new way to utilize both visible and infrared features to generate more discriminative feature representation. However, with the loss functions discussed above, it is not sufficient for learning the memory network serving as the bridge to propagate modality information. Firstly, if the reconstructed missing modality features are inconsistent with the features obtained from the backbone network, they will become interference information for unified feature learning. Secondly, if the memory items are not representative enough and do not save prototypical features of each modality, the reconstructed missing modality features cannot well represent the missing modality information. Lastly, if there is no correspondence between visible and infrared memory items, in Eq. (4) the reading weight $\mathbf{A}_i^{\boldsymbol{V}}$ cannot be used for weighted aggregation of infrared memory items. To deal with these problems, we design the following three learning strategies for the memory network.

1) Feature Consistency: To make the features reconstructed from the memory network consistent with features extracted from the backbone network, we utilize two modality discriminators $\mathcal{D}^{\boldsymbol{V}}$ and $\mathcal{D}^{\boldsymbol{I}}$ to classify the modality of reconstructed features $\hat{\mathbf{f}}_k^*$ and $\mathbf{f}_k^*$:

$$\mathcal{L}_{adv} = \sum_{k=1}^{K}\left[\log\mathcal{D}^*\left(\hat{\mathbf{f}}_k^*\right) + \log\left(1-\mathcal{D}^*\left(\mathbf{f}_k^*\right)\right)\right], \tag{10}$$

where $* \in \{\boldsymbol{V}, \boldsymbol{I}\}$. The modality discriminators $\mathcal{D}^{\boldsymbol{V}}$ and $\mathcal{D}^{\boldsymbol{I}}$ are classifiers consisting of two fully connected layers stacked with a Sigmoid function. To match the distribution of the input and the reconstructed features from memory network, we utilize the gradient reversal layer as proposed in [61]. The gradient reversal layer flips the gradient sign before propagating the gradients back to the backbone network. The discriminators $\mathcal{D}^{\boldsymbol{V}}$ and $\mathcal{D}^{\boldsymbol{I}}$ are trained to minimize Eq. (10) while the backbone network is trained to maximize Eq. (10). The adversarial training between the backbone network and discriminator helps to reduce the distribution gap between modality features reconstructed from the memory network and those from the backbone network. Thus, the reconstructed modality features $\hat{\mathbf{f}}_k^{\boldsymbol{V}}$ and $\hat{\mathbf{f}}_k^{\boldsymbol{I}}$ can well provide the missing modality information.

2) Memory Representativeness: Since we expect the memory items to store prototypical features of each modality, they should be representative enough for each modality. If each input sample can be well reconstructed using memory items from the same modality, then the memory items can be used as the proxy of the modalities for modality information propagation. Inspired by this, we propose a reconstruction loss to make sure that we can reconstruct the input features with memory items from the same modality. For example, with the reading weight $\mathbf{A}_k^{\boldsymbol{V}}$, we can reconstruct the input feature $\mathbf{f}_k^{\boldsymbol{V}}$ from visible memory items $\mathbf{M}_k^{\boldsymbol{V}}$. we first obtain the reconstructed input features as follows:

$$\widetilde{\mathbf{f}}_k^{\boldsymbol{V}} = \sum_{n=1}^{N}\alpha_{k,n}^{\boldsymbol{V}}\mathbf{m}_{k,n}^{\boldsymbol{V}}, \quad \widetilde{\mathbf{f}}_k^{\boldsymbol{I}} = \sum_{n=1}^{N}\alpha_{k,n}^{\boldsymbol{I}}\mathbf{m}_{k,n}^{\boldsymbol{I}}. \tag{11}$$

Then we minimize the Euclidean distance between the input features and the reconstructed input features:

$$\mathcal{L}_{rec} = \sum_{k=1}^{K} \left[\left\| \mathbf{f}_k^V - \widetilde{\mathbf{f}}_k^V \right\|_2 + \left\| \mathbf{f}_k^I - \widetilde{\mathbf{f}}_k^I \right\|_2 \right]. \quad (12)$$

With the reconstruction loss, the modality-specific memory items $\mathbf{M}_k^V$ and $\mathbf{M}_k^I$ are guided to save the representative features of each modality.

3) Structural Alignment: As shown in Eq. (4), to reconstruct the missing modality feature from the memory network by using the single modality input and memory items from the other modality, we utilize memory items from the same modality as the bridge. For example, to reconstruct $\hat{\mathbf{f}}_i^I$ from $\mathbf{f}_i^V$, the visible memory items are regarded as the bridge to aggregate infrared memory items, as shown in Eq. (4). That is, with the reading weight $\mathbf{A}_k^V$, the corresponding infrared memory items in $\mathbf{M}_k^I$ are aggregated. This process requires a correspondence between visible and infrared memory items. To achieve this goal, we introduce the following structural alignment loss to align visible and infrared memory items:

$$\mathcal{L}_{align} = \sum_{k=1}^{K} \left[D_{KL}\left(\mathbf{A}_k^V \| \mathbf{A}_k^I\right) + D_{KL}\left(\mathbf{A}_k^I \| \mathbf{A}_k^V\right) \right], \quad (13)$$

where $D_{KL}(\cdot)$ represents the KL divergence [62], and the explanation of the structural alignment loss is shown in Figure 3. With the proposed structural alignment loss, the visible modality memory items save the visible prototypical features in the same location where the infrared memory items save the corresponding infrared prototypical features.

To sum up, the losses for memory network learning is formulated as:

$$\mathcal{L}_{mem} = \lambda_{adv}\mathcal{L}_{adv} + \lambda_{rec}\mathcal{L}_{rec} + \lambda_{align}\mathcal{L}_{align}, \quad (14)$$

where λ_{adv}, λ_{rec} and λ_{align} are hyper-parameters to balance the contribution of each loss function. Through the above three learning strategies, memory items of two modalities can be well learned to serve as a bridge to propagate discriminative information between visible and infrared modalities.

D. Training and Inference

For the VI-ReID task, our proposed MSMNet is trained by minimizing the overall objective with identity labels as defined in Eq. (15).

$$\mathcal{L}_{all} = \mathcal{L}_{sid} + \mathcal{L}_{id} + \mathcal{L}_{hc\_tri} + \mathcal{L}_{mem}. \quad (15)$$

During the testing stage, for each image from visible or infrared modality, we concatenate the unified part features $\{\mathbf{g}_k^*\}_{k=1}^K$ in Eq. (6) together as its final representation:

$$\mathbf{g}^* = \left[\mathbf{g}_1^*, \mathbf{g}_2^*, \cdots, \mathbf{g}_K^*\right], \quad (16)$$

where $* \in \{\boldsymbol{V}, \boldsymbol{I}\}$ and $[\cdot]$ denotes the concatenation operator. Finally, cross modality matching is conducted by computing cosine similarities of feature vectors $\mathbf{g}^{\boldsymbol{V}}$ or $\mathbf{g}^{\boldsymbol{I}}$ between the probe images and gallery images.

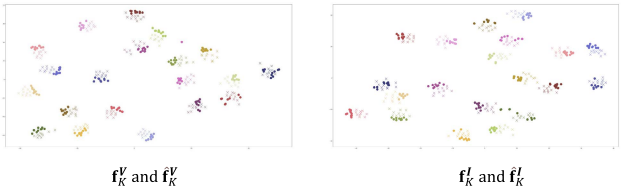

Fig. 5. The t-SNE visualization of original features and reconstructed features. Different colors represent different identities. Here circle represent original visible/infrared features and cross represent reconstructed modality features.

1) The Insight of the Memory Network: In contrast to the memory mechanism in MoCo [63], in which the goal is to store more samples to increase the number of negative samples during contrastive learning, our modality-specific memory has the goal of aiding modality information completion by saving all the potential visual patterns for visible/infrared modality within the whole dataset and allowing access through the memory read operation. Since the the potential visual patterns are shared across training and test set, we can direct use the memory items to reconstruct missing modality information during the testing. The modality-specific memory items are expected to reconstruct the initial features from the same modality and save **representative visual patterns** of each modality. To sum up, we intend to represent the feature distriubution of each modality with $K \times N$ trainable memory items and further achieve missing modality information reconstruction with these representative memory items.

We also alalyse the feature distribution of the reconstructed feature representations using t-SNE, and results are shown in Figure 5. Here circle represent original visible/infrared features $\mathbf{f}_k^{\boldsymbol{V}}$ and $\mathbf{f}_k^{\boldsymbol{I}}$ and cross represent reconstructed modality features $\hat{\mathbf{f}}_k^V$ and $\hat{\mathbf{f}}_k^I$. The reconstructed visible/infrared features are aligned with original visible/infrared features of the same ID in the feature space.

IV. Experiments

In this section, we first introduce implementation details and datasets. Then, we conduct comprehensive experiments to validate the effectiveness of the proposed modality-specific memory network as well as each of its components. Finally, we provide more analysis and visualization results to better understand our method.

A. Datasets and Evaluation Protocol

1) SYSU-MM01: [10] is the first large-scale benchmark for VI-ReID. This dataset consists of a total of 287,628 visible images taken by 4 visible cameras in the daytime, and 15,792 infrared images taken by 2 infrared cameras in the dark environment. These images are captured in both indoor and outdoor scenarios. The training set and the testing set have 395 and 96 person identities, respectively. Following [10], there are two testing modes: *all-search* and *indoor-search*. For the *all-search* mode, the gallery set contains visible images in both indoor and outdoor scenarios. For the *indoor-search* mode, the gallery set merely contains visible images in the indoor scenario. For both modes, there are also two settings:

TABLE I

COMPARISONS WITH STATE-OF-THE-ART METHODS ON THE SYSU-MM01 DATASET. CM-SSFT* DENOTES THAT WE REPORT THE RESULTS OF CM-SSFT UNDER THE SINGLE-QUERY SETTING [18] FOR FAIR COMPARISONS WITH OTHER METHODS. RANK-K ACCURACY (%) AND MAP (%) ARE REPORTED

Method	Venue	All-Search								Indoor-Search							
		Single-Shot				Multi-Shot				Single-Shot				Multi-Shot			
		R-1	R-10	R-20	mAP	R-1	R-10	R-20	mAP	R-1	R-10	R-20	mAP	R-1	R-10	R-20	mAP
Zero-Pad [10]	ICCV-17	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	18.64
TONE [16]	AAAI-18	12.52	50.72	68.60	14.42	-	-	-	-	20.82	68.86	84.46	26.38	-	-	-	-
HCML [16]	AAAI-18	14.32	53.16	69.17	16.16	-	-	-	-	24.52	73.25	86.73	30.08	-	-	-	-
cmGAN [39]	IJCAI-18	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
BDTR [11]	IJCAI-18	27.32	66.96	81.07	27.32	-	-	-	-	31.92	77.18	89.28	41.86	-	-	-	-
D^{2}RL [17]	CVPR-19	28.9	70.6	82.4	29.2	-	-	-	-	-	-	-	-	-	-	-	-
AlignGAN [44]	ICCV-19	42.4	85.0	93.7	40.7	51.5	89.4	95.7	33.9	45.9	87.6	94.4	54.3	57.1	92.7	97.4	45.3
JSIA-ReID [64]	AAAI-20	38.1	80.7	89.9	36.9	45.1	85.7	93.8	29.5	43.8	86.2	94.2	52.9	52.7	91.1	96.4	42.7
Xmodal [45]	AAAI-20	49.92	89.79	95.96	50.73	-	-	-	-	-	-	-	-	-	-	-	-
cm-SSFT* [18]	CVPR-20	47.7	-	-	54.1	57.4	-	-	59.1	-	-	-	-	-	-	-	-
MACE [19]	TIP-20	51.64	87.25	94.44	50.11	-	-	-	-	57.35	93.02	97.47	64.79	-	-	-	-
DDAG [15]	ECCV-20	54.75	90.39	95.81	53.02	-	-	-	-	61.02	94.06	98.41	67.98	-	-	-	-
HAT [40]	TIFS-20	55.29	92.14	97.36	53.89	-	-	-	-	62.10	95.75	99.20	69.37	-	-	-	-
HCT [12]	TMM-20	61.68	93.10	97.17	57.51	-	-	-	-	63.41	91.69	95.28	68.17	-	-	-	-
NFS [41]	CVPR-21	56.91	91.34	96.52	55.45	63.51	94.42	97.81	48.56	62.79	96.53	99.07	69.79	70.03	97.70	99.51	61.45
MID [65]	AAAI-22	60.27	92.90	-	59.40	-	-	-	-	64.86	96.12	-	70.12	-	-	-	-
CM-NAS [13]	ICCV-21	61.99	92.87	97.25	60.02	68.68	94.92	98.36	53.45	67.01	97.02	99.32	72.95	76.48	98.68	99.91	65.11
SPOT [43]	TIP-22	65.34	92.73	97.04	62.25	-	-	-	-	69.42	96.22	99.12	74.63	-	-	-	-
MCLNet [66]	ICCV-21	65.40	93.33	97.14	61.98	-	-	-	-	72.56	96.98	99.20	76.58	-	-	-	-
SMCL [42]	ICCV-21	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
MPANet [67]	CVPR-21	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
MSMNet (Ours)	This paper	**73.46**	**96.27**	**98.82**	**69.58**	**78.59**	97.61	**99.46**	**64.14**	**78.92**	97.50	98.88	**81.22**	**87.31**	99.25	99.85	**75.84**

single-shot and *multi-shot*, where 1 or 10 images of a person are randomly selected to form the gallery set.

2) RegDB: [68] is constructed by a dual-camera system that includes a visible camera and a thermal-infrared camera. It contains 412 persons, where each person has 10 images from the visible camera and 10 images from the thermal-infrared camera. Following [16], 2,060 images from 206 person identities are randomly chosen as the training set and the remaining 2,060 images from 206 identities make up the testing set. There are two evaluation settings: *Visible to Infrared* and *Infrared to Visible*. Taking *Visible to Infrared* for example, it denotes using the visible images as the probe set and the infrared images as the gallery set.

3) Evaluation Metrics: Following standard evaluation protocols for VI-ReID [10], [60], we adopt Cumulative Matching Characteristic (CMC) and mean Average Precision (mAP) for performance evaluation. The reported results on the SYSU-MM01 dataset are an average performance of 10 times repeated random probe/gallery splits [10], while that on the RegDB dataset are an average performance of 10 trials with different splits of training/testing sets [16], [17].

B. Implementation Details

Following the previous VI-ReID methods [12], [18], we adopt ResNet50 [69] pretrained on ImageNet as our backbone network. Following [12], [15], [60], we also set the stride of the last convolutional block as 1. The number of part features K is set to 6 and the number of memory items N for each part is set to 20. We resize each person image to the size of 384×144, and apply horizontal flipping and random erasing [70] for data augmentation. For a mini-batch, we randomly choose 8 identities and each identity has 4 visible and 4 infrared images. We use the Adam optimizer to train our model for 120 epochs with a batch size of 64. The learning rate is initialized to 3.5×10^{-4} and decayed to its 0.1 and 0.01 at the 60-th and 90-th epochs. The hyper-parameters λ_{adv}, λ_{rec} and λ_{align} are set to be 1, 0.2 and 2, respectively. We implement our model with PyTorch and train it on one Geforce RTX 3090 GPU.

C. Comparison With State-of-the-Art Methods

In this section, we compare the proposed MSMNet with the state-of-the-art VI-ReID methods on the SYSU-MM01 and RegDB datasets. The compared methods include image translation based methods (cmGAN [39], AlignGAN [44], JSIA-ReID [64]), intermediate modality generation based methods (Xmodal [45], SMCL [42], MID [65]), deep metric learning based methods (BDTR [11], TONE [16], HCML [16], MCLNet [66]), modality-shared feature learning with two-stream CNNs (DDAG [15], HAT [40], HCT [12]), modality-shared feature learning with neural architecture search technique (NFS [41], CM-NAS [13]) and modality information completion methods (D^{2}RL [17], cm-SSFT* [18]). We directly use the original results from published papers for comparison.

1) Results on the SYSU-MM01 Dataset: Table I shows the performance of our method and previous methods on the SYSU-MM01 dataset, which shows that MSMNet achieves competitive performance compared with the state-of-the-arts. Specifically, the proposed model achieves 73.46% Rank-1 accuracy and 69.58% mAP in the most challenging *single-shot* and *all search* modes, improving the Rank-1 accuracy by 2.88% and mAP by 1.34% over the best SOTA MPANet, which verifies the superiority of the proposed method. According to the experimental results, we make the following observations. 1) Our method performs much better than the image translation based methods. For example, compared with AlignGAN [44], our method surpasses them by 31.06% in Rank-1 accuracy and 28.88% in mAP in the *single-shot* and *all search* modes. Meanwhile, our MSMNet does not require a time-consuming image generation process and avoids introducing

TABLE II
COMPARISONS WITH STATE-OF-THE-ART METHODS ON THE REGDB DATASET.RANK-K ACCURACY (%) AND MAP (%) ARE REPORTED. OUR METHOD ACHIEVES THE BEST PERFORMANCE

Method	Venue	Visible to Infrared				Infrared to Visible			
		R-1	R-10	R-20	mAP	R-1	R-10	R-20	mAP
Zero-Pad [10]	ICCV-17	17.74	34.21	44.35	18.90	16.63	34.68	44.25	17.82
HCML [16]	ICCV-17	24.44	47.53	56.78	20.08	21.70	45.02	55.58	22.24
BDTR [11]	IJCAI-18	33.56	58.61	67.43	32.76	32.92	58.46	68.43	31.96
D^2RL [17]	CVPR-19	43.4	66.1	76.3	44.1	-	-	-	-
AlignGAN [44]	ICCV-19	57.9	-	-	53.6	56.3	-	-	53.4
JSIA-ReID [64]	AAAI-20	48.5	-	-	49.3	48.1	-	-	48.9
Xmodal [45]	AAAI-20	62.21	83.13	91.72	60.18	-	-	-	-
cm-SSFT* [18]	CVPR-20	65.4	-	-	65.6	63.8	-	-	64.2
MACE [19]	TIP-20	72.37	88.40	93.59	69.09	72.12	88.07	93.07	68.57
DDAG [15]	ECCV-20	69.34	86.19	91.49	63.46	68.06	85.15	90.31	61.80
HAT [40]	TIFS-20	71.83	87.16	92.16	67.56	70.02	86.45	91.61	66.30
SPOT [43]	TIP-22	80.35	93.48	96.44	72.46	79.37	92.79	96.01	72.26
NFS [41]	CVPR-21	80.54	91.96	95.07	72.10	77.95	90.45	93.62	69.79
MPANet [67]	CVPR-21	83.70	-	-	80.90	82.80	-	-	80.70
MCLNet [66]	ICCV-21	80.31	92.7	96.03	73.07	75.93	90.93	94.59	69.49
SMCL [42]	ICCV-21	83.93	-	-	79.83	83.05	-	-	78.57
CM-NAS [13]	ICCV-21	84.54	95.18	97.85	80.32	82.57	94.51	97.37	78.31
MID [65]	AAAI-22	87.45	95.73	-	84.85	84.29	93.44	-	81.41
HCT [12]	TMM-20	91.05	97.16	98.57	83.28	89.30	96.41	98.16	81.46
MSMNet (Ours)	This paper	**95.98**	**99.05**	**99.72**	**88.55**	**93.44**	**98.30**	**99.36**	**86.66**

noisy generated samples. 2) Compared with the modality-shared feature learning with two-stream CNNs, our method exceeds HCT [12] by 11.78% in Rank-1 accuracy and 12.07% in mAP in the *single-shot* and *all search* modes. The proposed MSMNet can utilize information from visible and infrared modalities and boost the VI-ReID performance. 3)For the modality information completion method cm-SSFT, we compare the performance of cm-SSFT with single query for fair comparisons. The Rank-1 accuracy and mAP of our method are 25.76% and 15.48% higher than cm-SSFT, which shows the effectiveness of our modality-specific memory network for modality information completion.

2) Results on the RegDB Dataset: The comparison results on the RegDB Dataset are shown in Table II. The performance of MSMNet outperforms existing state-of-the-art methods by large margins under both evaluation settings. Specially, on the *visible-to-infrared* setting, MSMNet makes a significant improvement of 4.93% in Rank-1 accuracy and 5.27% in mAP compared to the top-performing method HCT [12]. The similar improvement also presents in the *infrared-to-visible* mode, which shows that our method is robust to visible and infrared query settings. Compared to the current best state-of-the-art method CM-NAS [13] punlished in ICCV 2021, our method outperforms CM-NAS by 11.44% in Rank-1 accuracy and 8.32% in mAP on the *visible-to-infrared* setting, and by 10.87% in Rank-1 accuracy and 8.35% in mAP on the *infrared-to-visible* setting. In conclusion, the above results clearly indicate that the proposed MSMNet leads a new state-of-the-art performance on the RegDB dataset.

D. Ablation Studies

In this section, we perform extensive ablation studies to investigate the effectiveness of the components of our model. Results are evaluated on the SYSU-MM01 dataset in the *all-search* and *single-shot* modes to analyze each component of our method. The baseline model is based on HCT [12] with a few modifications, which the optimizer is set to Adam and the learning rate is initialized to 3.5×10^{-4}, while they use SGD to optimize the model in their source code. We find that such training strategy yields better results than in their paper. The baseline model uses ResNet-50 as the backbone and uses part features $\mathbf{f}_k^V$ for visible images and $\mathbf{f}_k^I$ for infrared images without the proposed reconstructed missing modality features and three losses. The results are shown in Table III.

TABLE III
PERFORMANCE COMPARISON WITH DIFFERENT COMPONENTS OF OUR METHOD ON THE **SYSU-MM01** DATASET IN THE *All-Search* AND *Single-Shot* MODES

Index	$\mathcal{M}$	$\mathcal{L}_{align}$	$\mathcal{L}_{rec}$	$\mathcal{L}_{adv}$	SYSU-MM01			
					R-1	R-10	R-20	mAP
1					65.40	93.33	97.14	61.98
2	✓				68.87	95.05	98.13	66.34
3	✓	✓			70.92	95.48	98.24	67.45
4	✓		✓		70.69	95.62	98.33	67.82
5	✓			✓	69.88	95.24	98.18	67.26
6	✓	✓	✓		72.53	95.93	98.72	68.87
7	✓	✓	✓	✓	**73.46**	**96.27**	**98.82**	**69.58**

1) Effectiveness of the Memory Network: We first evaluate the effectiveness of the proposed modality-specific memory network. The memory network is denoted as $\mathcal{M}$ in Table. III. As shown in index-1 and index-2, compared with the baseline model, when only the modality-specific memory network is adopted, the performance is significantly improved by +3.47% mAP and +4.36% in Rank-1 accuracy. This is because the baseline model only considers the modality shared features, and some identity-related information is regarded as redundant information. In contrast, our proposed MSMNet adopts modality-specific memory items serving as a bridge to propagate identity-related information between two modalities to boost the VI-ReID performance.

2) Effectiveness of the Structural Alignment: From index-2 and index-3, when the structural alignment loss is added, the performance is greatly improved by +2.05% and up to 70.92%

TABLE IV

PERFORMANCE OF DIFFERENT WAYS OF COMPLETING THE MISSING MODALITY INFORMATION ON TWO DATASETS

Method	SYSU-MM01			RegDB		
	R-1	R-10	mAP	R-1	R-10	mAP
Base	65.40	93.33	61.98	91.46	95.92	84.02
Linear Layers	66.96	95.02	64.84	92.24	97.63	85.14
MSMNet(Ours)	73.46	96.27	69.58	95.98	99.05	88.55

Rank-1 accuracy. This shows that the correspondence between memory items from visible and infrared modalities is of significant importance to the learning of the memory network. The structural alignment loss makes the paired memory items from different modalities save the corresponding prototypical features. When memory items are aligned with each other, the designed memory read operation can retrive proper memory items for the missing modality information completion.

3) Effectiveness of the Memory Representativeness: From index-2 and index-4, when the reconstruction loss is added, the performance is improved by +1.82% in Rank-1 accuracy and +1.48% mAP. From index-3 and index-6, we can see that when the structural alignment loss has been added, the reconstruction loss can still improve +1.61% in Rank-1 accuracy and +1.42% mAP. This is because with the proposed reconstruction loss, modality-specific memory items are guided to reconstruct input features from the same modality and try to store prototypical features of each modality. Thus, the learned modality-specific memory network would be representative enough to serve as the proxy for modality information reconstruction and further boost the performance.

4) Effectiveness of the Feature Consistency: From index-2 and index-5, we can see that the modality discriminators bring in +1.01% Rank-1 accuracy and +0.92% mAP improvements. The modality discriminators make the reconstructed features from the memory network be consistent with the features extracted from the backbone, and can alleviate the modality discrepancy. From index-7, we can see that when the proposed three losses work together, our method achieves the best performance. The results demonstrate that each proposed loss plays an important role in modality-specific memory network learning and alleviating modality discrepancy.

5) Effectiveness of the Two Modality Discriminators: **(1)** The two modality discriminators are designed to make the features reconstructed from the memory network consistent with features extracted from the backbone network. Take the visible discriminator $\mathcal{D}^{\boldsymbol{V}}$ as an example. The reconstructed visible modality feature $\hat{\mathbf{f}}_k^{\boldsymbol{V}}$ tries to fool the discriminator $\mathcal{D}^{\boldsymbol{V}}$ by approximating original visible feature $\mathbf{f}_k^{\boldsymbol{V}}$. The visible discriminator tries to distinguish the original visible feature $\mathbf{f}_k^{\boldsymbol{V}}$ and the reconstructed visible feature $\hat{\mathbf{f}}_k^{\boldsymbol{V}}$ as accurate as possible. Finally, the reconstructed visible feature $\hat{\mathbf{f}}_k^{\boldsymbol{V}}$ will be eventually consistent with the original visible feature $\mathbf{f}_k^{\boldsymbol{V}}$. Similarly, $\mathcal{D}^{\boldsymbol{I}}$ is to make the reconstructed infrared feature $\hat{\mathbf{f}}_k^{\boldsymbol{I}}$ consistent with $\mathbf{f}_k^{\boldsymbol{I}}$. **(2)** If we use a single discriminator for both modalities, the orignal features $\mathbf{f}_k^{\boldsymbol{V}}$, $\mathbf{f}_k^{\boldsymbol{I}}$ and the reconstructed features $\hat{\mathbf{f}}_k^{\boldsymbol{V}}$, $\hat{\mathbf{f}}_k^{\boldsymbol{I}}$ will be directly aligned with a single discriminator $\mathcal{D}$. As the result, the visible feature $\mathbf{f}_k^{\boldsymbol{V}}$ and infrared feature $\mathbf{f}_k^{\boldsymbol{I}}$ will be directly aligned. That is, the process of modality information completion could provide little discriminative information. We conduct experiments to show the effectiveness of two modality discriminators, and the results are shown in Table V. From the results, we can see that two modality discriminators perform the best, e.g., improve the mAP by 2.61% as compared to that of using one discriminator on the SYSU-MM01 dataset. The use of one discriminator for both modalities even has a negative impact on performance.

TABLE V

PERFORMANCE OF USING ONE OR TWO MODALITY DISCRIMINATORS

Method	SYSU-MM01			RegDB		
	R-1	R-10	mAP	R-1	R-10	mAP
MSMNet w/o $\mathcal{L}_{adv}$	72.53	96.06	68.87	93.37	98.23	86.76
+ one discriminator $\mathcal{D}$	70.85	95.64	68.01	92.94	98.02	86.21
+ $\mathcal{D}^{\boldsymbol{V}}$ and $\mathcal{D}^{\boldsymbol{I}}$	73.46	96.27	69.58	95.98	99.05	88.55

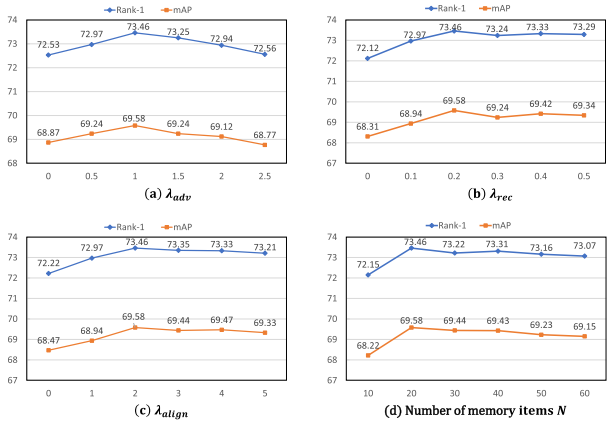

Fig. 6. Parameter analyses of λ_{adv}, λ_{rec}, λ_{align} and the number of memory items N in Eq. (14) on the SYSU-MM01 dataset in the *all-search* and *single-shot* modes.

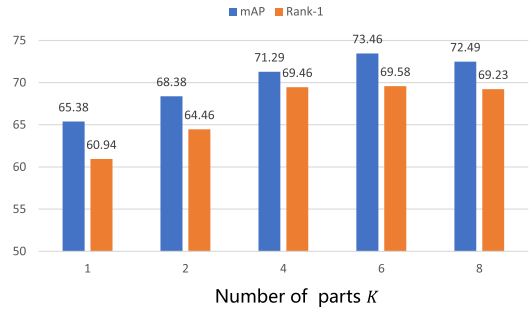

Fig. 7. Hyperparameter evaluations about the part number K on the SYSU-MM01 dataset.

6) Parameter Analysis: The proposed MSMNet involves three trade-off parameters λ_{adv}, λ_{rec} and λ_{align} in Eq. (14). Here, λ_{adv} controls the relative importance of the feature consistency, λ_{rec} controls the relative importance of memory representativeness and λ_{rec} controls the relative importance of memory structural alignment. We analyze the three hyperparameters on the SYSU-MM01 dataset in the *all-search* and *single-shot* modes, and we keep the other hyperparameters at the best choice when one hyperparameter varies. The Rank-1 and the mAP results of MSMNet with different λ_{adv}, λ_{rec} and λ_{align} are shown in Figure 6. We can see that our method is

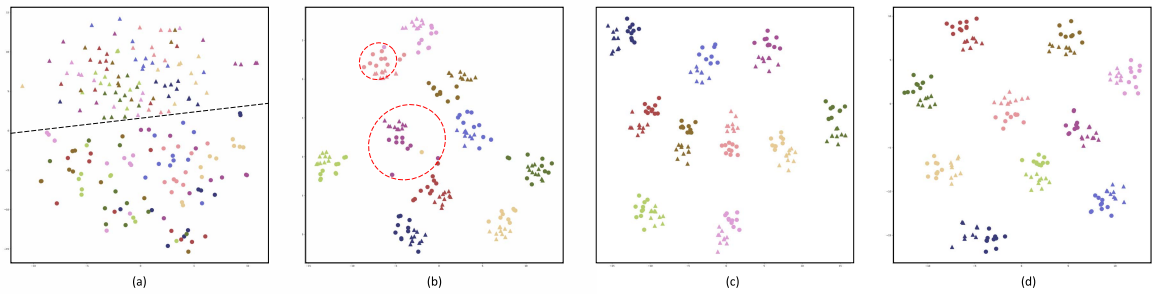

Fig. 8. Visualization of the distribution of learned representions from MSMNet and the baseline method by t-SNE. Different colors represent features of different ideneities, and triangle and circle symbols refer to features of visible and infrared images, respectively. (a) Features extracted by the ResNet-50 pretrained on ImageNet. (b) Features extracted by the baseline model. (c) Features extracted by MSMNet without three learning strategies. (d) Features extracted by the proposed MSMNet. It is obvious that the MSMNet better alleviates the modality discrepancy and improves the discriminability.

robust to the hyperparameters, and the most suitable parameter setting is that $\lambda_{adv} = 1$, $\lambda_{rec} = 0.2$ and $\lambda_{align} = 2$.

We also evaluate the influence of different number of memory items N in Figure 6 (d). From the results, we can observe that the performance keeps improving before N arrives at 20, and when N is greater than 20, the performance of the model hardly changes. We conclude that 20 memory items are enough to complete the missing modality information. And when N is greater than 20, there are some duplicate memory items in the memory network that cannot bring in further performance improvements.

We perform hyperparameter evaluations about the part number K on the SYSU-MM01 dataset. The results are shown in Figure 7. When $K = 1$, our model learns the global feature for each image. From the results, we can observe that the performance continues to grow until $K = 6$, which means that it is enough to divide the image into six parts for discriminative local feature learning.

E. Further Analysis

In this section, we provide more analysis and visualization results to better understand our method.

1) The Superiority of Using the Memory Network: The key challenge in VI-ReID is how to bridge the modality discrepancy between visible and infrared modalities. An effective way to achieve this goal is to complete the missing modality information. However, it is hard to infer the missing modality information only with the single modality image as input. Thus, we intend to use the memory network which saves prototypical features of two modalities to provide the missing modality information. To verify the superiority of using memory networks, We replace the memory network with learnable linear layers to complete the missing information for comparison, which means $\hat{\mathbf{f}}_k^I = FC\left(\mathbf{f}_k^I\right)$ in Eq. 4 and $\hat{\mathbf{f}}_k^V = FC\left(\mathbf{f}_k^V\right)$ in Eq. 5. The experiments are conducted in the *all-search* and *single-shot* modes on the SYSU-MM01 dataset and on the *visible-to-infrared* setting on the RegDB dataset. The comparison results are shown in Table IV, and "Base" represents the baseline model. On both datasets, the performance of using linear layers is better than that of the baseline model, which proves the effectiveness of missing modality information completion. Moreover, the proposed MSMNet with memory network improves Rank-1 accuracy by 6.50% as compared to that with learnable linear layers on the SYSU-MM01 dataset, and boost the mAP by 3.41% compared with using learnable linear layers on the RegDB dataset. Therefore, our proposed modality-specific memory network can effectively complete the missing modality information completion for alleviating the modality discrepancy.

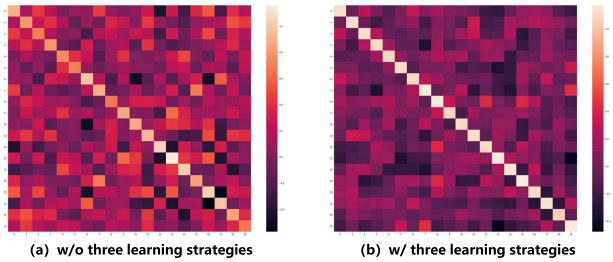

Fig. 9. The visualization of the correlation matrix between memory items in $\mathbf{M}_1^V$ when the model is trained with and without the proposed three learning strategies. The darker the color, the smaller the value. Best viewed in color.

2) Visualization of Unified Feature Distributions: In order to inspect the impact of the proposed method, we randomly select 10 identities from the test set and visualize the distributions of the learned features by t-SNE [71] in Figure 8. As shown in Figure 8 (a), the features extracted by the ResNet-50 pretrained on ImageNet have considerable modality discrepancy and cannot perform well for the cross modality matching. In Figure 8 (b), although most features extracted by the baseline model can be clustered well, there are still some identities that remain large inter-identity modality discrepancy, such as the purple ones. In Figure 8 (c), with the proposed memory network, the features are better clustered together than the baseline model. In Figure 8 (d), the unified feature representations extracted by the proposed MSMNet well alleviate the modality discrepancy and improves the discriminability. The features of different identities are separated into disjoint clusters with larger inter-class margins. Thus, with the reconstructed missing modality features, the discriminability of the unified feature representation obtained by MSMNet is significantly improved.

3) The Non-Redundancy of Memory Items: To validate the non-redundancy of N memory items in our method, we compare the correlation matrix between memory items in $\mathbf{M}_1^V$ when the model is trained with and without the proposed three learning strategies, and the results are shown in Figure 9.

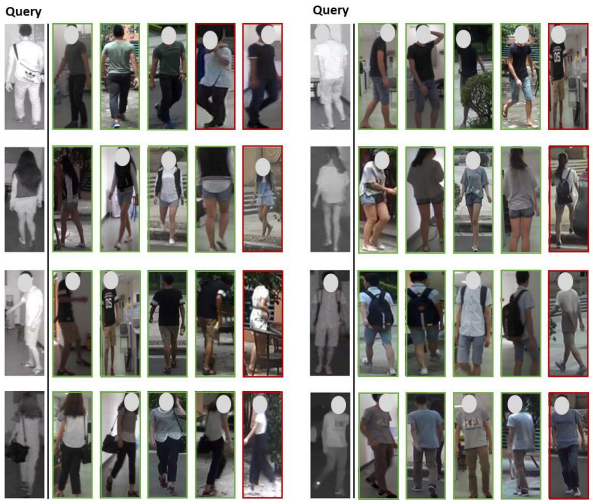

Fig. 10. Top-5 retrieval results on the SYSU-MM01 dataset. The green rectangles indicate correct retrieval results and red rectangles denote false retrieval results (best viewed in color).

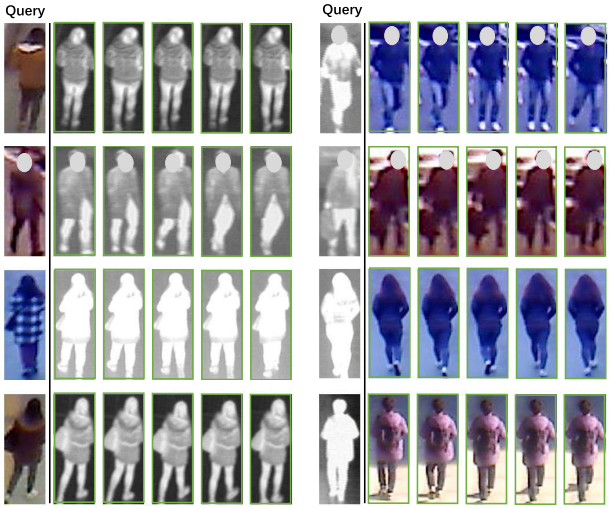

Fig. 11. Top-5 retrieval results on the RegDB dataset. The green rectangles indicate correct retrieval results and red rectangles denote false retrieval results (best viewed in color).

From the results, we can see that the correlation between memory items would be higher when the model is trained without the proposed three learning strategies. We conclude that the proposed feature consistency, memory representativeness, and structural alignment strategies would make memory items store diverse and representative modality features and force the memory items to be different from each other. From the parameter analysis result in Figure 6, we can also observe that finding a suitable N is also helpful in reducing the redundancy of N memory items. When N is greater than 20, there are some duplicate memory items in the memory network that cannot bring in further performance improvements.

4) Visualization of Retrieval Results: To better reflect the effectiveness of our methods, we visualize the top-5 retrieval results on two datasets. The ranking results on the SYSU-MM01 dataset are shown in Figure 10 and ranking results on the RegDB dataset are shown in Figure 11. The green rectangles denote the correct retrieval results, and the red rectangles denote the false retrieval results. It is obvious that the proposed MSMNet can retrieve most correct images, which illustrates the effectiveness of our method. Interestingly, in the first line of Figure 10, some images are even difficult for humans, but the proposed method can still retrieve the correct results. It can also be observed that due to large pose and viewpoint variations on the SYSU-MM01 dataset, different pedestrians with similar pose or body shape are likely to be identified as the same person, thus affecting the performance. While for the RegDB dataset, since images captured by the dual-camera system have achieved pose alignment, our method can retrieve almost all the correct pedestrian images. The visualization results prove that the proposed MSMNet can effectively alleviate the modality discrepancy and retrieve the correct pedestrian images in poor lighting surveillance scenarios.

V. Conclusion

In this paper, we propose a Modality-Specific Memory Network for the challenging VI-ReID. To alleviate the modality discrepancy between the visible and infrared modalities, we design a modality-specific memory network as a bridge to complete the missing modality information. To learn the memory network, we introduce three learning strategies to make paired memory items from two modalities representative and structural alignment with each other. Extensive experiments on two popular datasets demonstrate the superiority of the proposed MSMNet, as well as the effectiveness of each component of our method.

References

[1] S. Gong and T. Xiang, "Person re-identification," in *Visual Analysis of Behaviour*. Springer, 2011, pp. 301–313.
[2] F. Xiong, M. Gou, O. Camps, and M. Sznaier, "Person re-identification using kernel-based metric learning methods," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 1–16.
[3] L. Zheng, Y. Yang, and A. G. Hauptmann, "Person re-identification: Past, present and future," 2016, *arXiv:1610.02984*.
[4] M. Kostinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2288–2295.
[5] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, "Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)," in *Proc. ECCV*, May 2018, pp. 480–496.
[6] J. Liu, B. Ni, Y. Yan, P. Zhou, S. Cheng, and J. Hu, "Pose transferrable person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4099–4108.
[7] J. Xu, R. Zhao, F. Zhu, H. Wang, and W. Ouyang, "Attention-aware compositional network for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2119–2128.
[8] L. Zhao, X. Li, Y. Zhuang, and J. Wang, "Deeply-learned part-aligned representations for person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3219–3228.
[9] M. M. Kalayeh, E. Basaran, M. Gokmen, M. E. Kamasak, and M. Shah, "Human semantic parsing for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1062–1071.
[10] 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.
[11] M. Ye, Z. Wang, X. Lan, and P. C. Yuen, "Visible thermal person re-identification via dual-constrained top-ranking," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, vol. 1, Jul. 2018, p. 2.

[12] 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.

[13] 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," 2021, *arXiv:2101.08467*.

[14] 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.

[15] M. Ye, J. Shen, D. J. Crandall, L. Shao, and J. Luo, "Dynamic dual-attentive aggregation learning for visible-infrared person re-identification," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 229–247.

[16] M. Ye, X. Lan, J. Li, and P. Yuen, "Hierarchical discriminative learning for visible thermal person re-identification," in *Proc. AAAI Conf. Artif. Intell.*, vol. 32, no. 1, 2018, pp. 1–8.

[17] 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.

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

[19] 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.

[20] I. Goodfellow et al., "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, 2014, pp. 1–10.

[21] J. Tang, X. Shu, Z. Li, G.-J. Qi, and J. Wang, "Generalized deep transfer networks for knowledge propagation in heterogeneous domains," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 12, no. 4s, pp. 1–22, Nov. 2016.

[22] Z. Li, S. Chang, F. Liang, T. S. Huang, L. Cao, and J. R. Smith, "Learning locally-adaptive decision functions for person verification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3610–3617.

[23] Y. Yang, J. Yang, J. Yan, S. Liao, D. Yi, and S. Z. Li, "Salient color names for person re-identification," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 536–551.

[24] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 2197–2206.

[25] W. Zheng, S. Gong, and T. Xiang, "Re-identification by relative distance comparison," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 3, pp. 653–668, Mar. 2013.

[26] S. Liao and S. Z. Li, "Efficient PSD constrained asymmetric metric learning for person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3685–3693.

[27] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," 2017, *arXiv:1703.07737*.

[28] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng, "Person re-identification by multi-channel parts-based CNN with improved triplet loss function," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1335–1344.

[29] M. S. Sarfraz, A. Schumann, A. Eberle, and R. Stiefelhagen, "A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 420–429.

[30] C. Su, J. Li, S. Zhang, J. Xing, W. Gao, and Q. Tian, "Pose-driven deep convolutional model for person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3960–3969.

[31] G. Wang et al., "High-order information matters: Learning relation and topology for occluded person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 6449–6458.

[32] 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. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 2898–2907.

[33] X. Zhou, N. Cui, Z. Li, F. Liang, and T. S. Huang, "Hierarchical Gaussianization for image classification," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep. 2009, pp. 1971–1977.

[34] R. R. Varior, M. Haloi, and G. Wang, "Gated Siamese convolutional neural network architecture for human re-identification," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 791–808.

[35] X. Wang, M. Liu, D. S. Raychaudhuri, S. Paul, Y. Wang, and A. K. Roy-Chowdhury, "Learning person re-identification models from videos with weak supervision," *IEEE Trans. Image Process.*, vol. 30, pp. 3017–3028, 2021.

[36] D. Wu, M. Ye, G. Lin, X. Gao, and J. Shen, "Person re-identification by context-aware part attention and multi-head collaborative learning," *IEEE Trans. Inf. Forensics Security*, vol. 17, pp. 115–126, 2022.

[37] M. Ye, H. Li, B. Du, J. Shen, L. Shao, and S. C. H. Hoi, "Collaborative refining for person re-identification with label noise," *IEEE Trans. Image Process.*, vol. 31, pp. 379–391, 2021.

[38] X. Liu et al., "HydraPlus-Net: Attentive deep features for pedestrian analysis," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 350–359.

[39] 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.*, vol. 1, Jul. 2018, p. 2.

[40] 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.

[41] Y. Chen, L. Wan, Z. Li, Q. Jing, and Z. Sun, "Neural feature search for RGB-infrared person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 587–597.

[42] 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.

[43] C. Chen, M. Ye, M. Qi, J. Wu, J. Jiang, and C.-W. Lin, "Structure-aware positional transformer for visible-infrared person re-identification," *IEEE Trans. Image Process.*, vol. 31, pp. 2352–2364, 2022.

[44] G. Wang, T. Zhang, J. Cheng, S. Liu, Y. Yang, and Z. Hou, "RGB-infrared cross-modality person re-identification via joint pixel and feature alignment," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3623–3632.

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

[46] H. Ye, H. Liu, F. Meng, and X. Li, "Bi-directional exponential angular triplet loss for RGB-infrared person re-identification," *IEEE Trans. Image Process.*, vol. 30, pp. 1583–1595, 2020.

[47] Q. Zhang, J. Lai, and X. Xie, "Learning modal-invariant angular metric by cyclic projection network for VIS-NIR person re-identification," *IEEE Trans. Image Process.*, vol. 30, pp. 8019–8033, 2021.

[48] M. Ye, C. Chen, J. Shen, and L. Shao, "Dynamic tri-level relation mining with attentive graph for visible infrared re-identification," *IEEE Trans. Inf. Forensics Security*, vol. 17, pp. 386–398, 2021.

[49] J. Weston, S. Chopra, and A. Bordes, "Memory networks," 2014, *arXiv:1410.3916*.

[50] S. Sukhbaatar, A. Szlam, J. Weston, and R. Fergus, "End-to-end memory networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 2440–2448.

[51] A. Miller, A. Fisch, J. Dodge, A.-H. Karimi, A. Bordes, and J. Weston, "Key-value memory networks for directly reading documents," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2016, pp. 1–10.

[52] S. W. Oh, J.-Y. Lee, N. Xu, and S. J. Kim, "Video object segmentation using space-time memory networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9226–9235.

[53] C. Eom, G. Lee, J. Lee, and B. Ham, "Video-based person re-identification with spatial and temporal memory networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 12036–12045.

[54] V. Vs, V. Gupta, P. Oza, V. A. Sindagi, and V. M. Patel, "MeGA-CDA: Memory guided attention for category-aware unsupervised domain adaptive object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 4516–4526.

[55] S. Yoo, H. Bahng, S. Chung, J. Lee, J. Chang, and J. Choo, "Coloring with limited data: Few-shot colorization via memory augmented networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 11283–11292.

[56] D. Gong et al., "Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1705–1714.

[57] H. Park, J. Noh, and B. Ham, "Learning memory-guided normality for anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 14372–14381.

[58] H. Wang, W. Wang, W. Liang, C. Xiong, and J. Shen, "Structured scene memory for vision-language navigation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 8455–8464.

[59] X. Lu, W. Wang, M. Danelljan, T. Zhou, J. Shen, and L. V. Gool, "Video object segmentation with episodic graph memory networks," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 661–679.
[60] 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. 2021.
[61] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 1180–1189.
[62] S. Kullback and R. A. Leibler, "On information and sufficiency," *Ann. Math. Statist.*, vol. 22, no. 1, pp. 79–86, 1951.
[63] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9729–9738.
[64] G.-A. Wang et al., "Cross-modality paired-images generation for RGB-infrared person re-identification," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 7, 2020, pp. 12144–12151.
[65] Z. Huang, J. Liu, L. Li, K. Zheng, and Z.-J. Zha, "Modality-adaptive mixup and invariant decomposition for RGB-infrared person re-identification," 2022, *arXiv:2203.01735*.
[66] 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.
[67] Q. Wu et al., "Discover cross-modality nuances for visible-infrared person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 4330–4339.
[68] D. T. Nguyen, H. G. Hong, K. W. Kim, and K. R. Park, "Person recognition system based on a combination of body images from visible light and thermal cameras," *Sensors*, vol. 17, no. 3, p. 605, 2017.
[69] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[70] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 7, 2020, pp. 13001–13008.
[71] L. V. D. Maaten and G. Hinton, "Visualizing data using T-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 2579–2605, 2008.

Yulin Li received the bachelor's degree in electronic and information engineering from the Dalian University of Technology, Dalian, China, in 2019. He is currently pursuing the Ph.D. degree with the School of Cyber Science and Technology, University of Science and Technology of China, Hefei, China. His research interests include computer vision and machine learning, especially person re-identification, visual object tracking, and unsupervised representation learning.

Tianzhu Zhang (Member, IEEE) received the bachelor's degree in communications and information technology from the Beijing Institute of Technology, Beijing, China, in 2006, and the Ph.D. degree in pattern recognition and intelligent systems from the Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing, in 2011. He was an Associate Professor at CASIA. Currently, he is a Professor at the Department of Automation, School of Information Science and Technology, University of Science and Technology of China. His current research interests include pattern recognition, computer vision, multimedia computing, and machine learning. He served/serves as the Area Chair for CVPR 2020, ECCV 2020, ICCV 2019, ACM MM 2019, WACV 2018, ICPR 2018, and MVA 2017, and an Associate Editor for IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY and *Neurocomputing*.

Xiang Liu received the Ph.D. degree in electronics science and technology from the Beijing Institute of Technology in 2019. He had completed a general project of NFSC in 2018 and got a new general project of NFSC in 2020. He is currently a Teacher with the Dongguan University of Technology. His research interests are in artificial intelligence, machine learning, video coding and communication, multimedia information retrieval, visual information processing, and pattern recognition.

Qi Tian (Fellow, IEEE) received the B.E. degree in electronic engineering from Tsinghua University, the M.S. degree in ECE from Drexel University, and the Ph.D. degree in ECE from the University of Illinois at Urbana-Champaign (UIUC). He is currently a Chief Scientist in artificial intelligence at Cloud BU, Huawei. From 2018 to 2020, he was the Chief Scientist in computer vision at Huawei Noah's Ark Laboratory. He was also a Full Professor with the Department of Computer Science, The University of Texas at San Antonio (UTSA), from 2002 to 2019. His research interests include computer vision, multimedia information retrieval, and machine learning and published over 550 refereed journals and conference papers.

Yongdong Zhang (Senior Member, IEEE) received the Ph.D. degree in electronic engineering from Tianjin University, Tianjin, China, in 2002. He is currently a Professor at the University of Science and Technology of China. He has authored more than 100 refereed journals and conference papers. His current research interests include multimedia content analysis and understanding, multimedia content security, video encoding, and streaming media technology. He serves as an Editorial Board Member for *Multimedia Systems* journal and *Neurocomputing*. He was a recipient of the Best Paper Award in PCM2013, ICIMCS 2013, and ICME 2010, and the Best Paper Candidate in ICME 2011.

Feng Wu (Fellow, IEEE) received the B.S. degree in electrical engineering from Xidian University in 1992 and the M.S. and Ph.D. degrees in computer science from the Harbin Institute of Technology in 1996 and 1999, respectively. He is currently a Professor with the University of Science and Technology of China and the Dean of the School of Information Science and Technology. Before that, he was a Principle Researcher and a Research Manager with Microsoft Research Asia. His research interests include image and video compression, media communication, and media analysis and synthesis. He has authored or coauthored over 200 high quality papers (including several dozens of IEEE TRANSACTIONS papers) and top conference papers on MOBICOM, SIGIR, CVPR, and ACM MM. He has 77 granted U.S. patents. His 15 techniques have been adopted into international video coding standards. As the coauthor, he got the best paper award in IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY in 2009, PCM in 2008, and SPIE VCIP in 2007. He serves as an Associate Editor for IEEE TRANSACTIONS ON CIRCUITS AND SYSTEM FOR VIDEO TECHNOLOGY, IEEE TRANSACTIONS ON MULTIMEDIA, and several other international journals. He got IEEE Circuits and Systems Society 2012 Best Associate Editor Award. He also serves as the TPC Chair for MMSP 2011, VCIP 2010, and PCM 2009, and the Special Sessions Chair for ICME 2010 and ISCAS 2013.