Cross-Modality Transformer for Visible-Infrared Person Re-Identification

Kongzhu Jiang[1], Tianzhu Zhang[1,2⋆], Xiang Liu[3], Bingqiao Qian[1], Yongdong Zhang[1], and Feng Wu[1]

[1] University of Science and Technology of China
[2] Deep Space Exploration Lab
[3] Dongguan University of Technology of China
{kzjiang,qbq}@mail.ustc.edu.cn {tzzhang,zhyd73,fengwu}@ustc.edu.cn
succeedpkmba2011@163.com

Abstract. Visible-infrared person re-identification (VI-ReID) is a challenging task due to the large cross-modality discrepancies and intra-class variations. Existing works mainly focus on learning modality-shared representations by embedding different modalities into the same feature space. However, these methods usually damage the modality-specific information and identification information contained in the features. To alleviate the above issues, we propose a novel Cross-Modality Transformer (CMT) to jointly explore a modality-level alignment module and an instance-level module for VI-ReID. The proposed CMT enjoys several merits. First, the modality-level alignment module is designed to compensate for the missing modality-specific information via a Transformer encoder-decoder architecture. Second, we propose an instance-level alignment module to adaptively adjust the sample features, which is achieved by a query-adaptive feature modulation. To the best of our knowledge, this is the first work to exploit a cross-modality transformer to achieve the modality compensation for VI-ReID. Extensive experimental results on two standard benchmarks demonstrate that our CMT performs favorably against the state-of-the-art methods.

Keywords: Person Re-identification, Transformer, Cross-modality

1 Introduction

Person re-identification (Re-ID) aims at matching person images captured from non-overlapping camera views [40,42]. In recent years, it has gained increasing attention due to its significant practical value in video surveillance. Most of the existing methods [16,21,13,26,44,11,43,24] focus on visible (RGB) cameras and formulate the Re-ID task as a single-modality matching problem. However, the visible cameras are incapable of capturing valid appearance information of persons under poor illumination conditions (e.g., at night). To image clearly in the dark, in addition to the visible cameras, infrared (IR) cameras that are

⋆ Corresponding Author

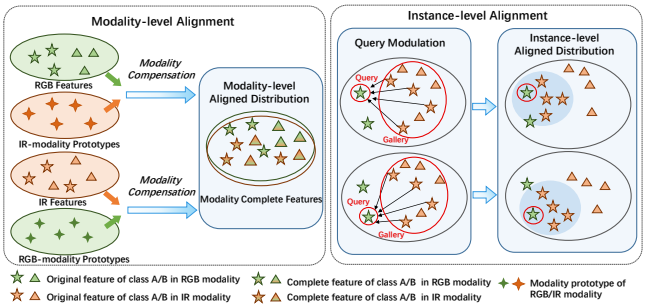


Fig. 1. Our motivation. The modality prototypes are introduced to store the global modality characteristics, which can be utilized to compensate for the missing modality features and thus contributes to the **modality-level alignment**. Then, by use of the query feature modulation, we can adaptively adjust the gallery sample features to activate query-related patterns and achieve **instance-level alignment**.

robust to illumination variants are also equipped in many surveillance scenarios. Hence, visible-infrared person re-identification (VI-ReID) [3,37,20] has recently been of great interest, which aims at retrieving IR person images of the same identity as the given RGB query and vice versa.

VI-ReID is challenging due to the cross-modal discrepancies between RGB and IR images, and the key issue is how to bridge the two modalities. To narrow the gap between two modalities, existing methods mainly focus on modality-level alignment. Some works are based on modality-shared feature learning [6,9,37,38,41], which decouple features into modality-specific and modality-shared features. Then they utilize the latter ones to align the modalities in the feature level while abandoning the modality-specific features. However, the modality-specific features also contain useful identity information that helps the final retrieval, such as colors. Therefore, with modality-shared cues only, the upper bound of the discrimination ability of the feature representation is limited. To address this limitation, modality compensation methods [32,20] have been proposed to compensate for the missing modality features. Specifically, in [20], the authors utilize the graph convolutional networks to obtain the compensated modality features based on the similarities between cross-modality samples in the current mini-batch.

By studying the previous VI-ReID methods based on modality compensation, we discover two characteristics that play an important role in achieving the robust VI-ReID. (1) **Modality-level alignment.** In previous modality compensation methods [20], the compensated features are produced solely based on the samples of the current mini-batch. This strategy suffers from a certain randomness, and thus causes the inconsistency of generated modality features when the samples are in the different mini-batches. To address this issue, an intuitive idea is to model several modality prototypes for representing global modality information. These modality prototypes can be used as the global basis for learning robust modality compensation for every sample. Therefore, it is necessary to model global modality prototypes to facilitate a better alignment between RGB and IR modalities. (2) **Instance-level alignment.** Due to intra-class variations

(e.g., viewpoint, illumination, and background clutter), the feature distribution of different samples with same ID varies greatly even under the same modality. Performing the modality-level alignment alone may lead to cases where the IR (RGB) instances are incorrectly aligned with the RGB (IR) instances of a different category. To align the cross-modal instances in the same class, most methods [9,37,38] utilize the supervised triplet loss to reduce the distances of the features of the same ID on the training set. However, in the open-set setting, because the categories of training and test sets have no overlap, the discriminative representations learned on the training set may not be optimal for the test images. Therefore, it is of vital importance to achieve dynamic instance-level alignment. In this way, the gallery instances can be adaptively refined according to the query features and be aligned to the query instances in the same class(as shown in Figure 1).

Inspired by the above discussions, we propose a novel Cross-Modality Transformer (CMT) by jointly exploring a modality-level alignment module and an instance-level alignment module for visible-infrared person re-identification. In the **modality-level alignment module**, we introduce an encoder-decoder architecture, which can achieve the modality feature enhancement and compensation. In the encoder, we adopt a self-attention mechanism to capture the interrelationship between local human parts. Then, we introduce two sets of learnable modality prototypes to represent the RGB and IR modalities respectively, and design a decoder to compensate for the missing modality. Taking a RGB sample as the example, we take the IR modality prototypes as queries and the part features of the RGB sample as keys and values of the transformer decoder. By use of the cross-attention between the part features and the modality prototypes, we can obtain the attention scores which can be regarded as the soft correspondences between IR modality prototypes and part features. Then we can compensate for IR modality-specific features by aggregating the related part features according to the attention matrix. Besides, to guide the learning of modality prototypes, we design a modality consistency loss to constrain the compensated IR features to be aligned with the real IR modality features. Similarly, the modality compensation for IR images can be achieved in the same way. In the **instance-level alignment module**, a feature modulator is proposed to adaptively adjust the representations of instances. Concretely, given the query sample x, we can utilize its feature to generate the channel-wise modulation parameters. These parameters reveal the most discriminative patterns of sample x, and can be employed to modulate other samples of the current mini-batch in the channel dimension. Thus, the crucial x-related channels of other samples can be strengthened, and the irrelevant channels can be suppressed. In this way, we can achieve query-adaptive feature modulation during the test, which can adaptively activate coherent patterns between query instances and gallery instances and facilitate a better instance-level alignment even for the unseen test categories.

The main contributions of this paper can be summarized as follows: (1) We propose a novel Cross-Modality Transformer for VI-ReID to jointly explore modality-level alignment and instance-level alignment. To the best of our knowl-

edge, this is the first work to exploit a cross-modality transformer to achieve the modality compensation for VI-ReID. (2) A modality-level alignment module is proposed to compensate for the missing modality-specific information via a Transformer encoder-decoder architecture. Also, we design an instance-level alignment module to adaptively adjust the sample features, which is achieved by query-adaptive feature modulation. (3) Extensive experimental results on two standard benchmarks demonstrate that the proposed model performs favorably against state-of-the-art VI-ReID methods.

2 Related Work

Single-Modality Person Re-ID. Single-modality person re-identification aims at matching pedestrian images across disjoint visible cameras. The considerable viewpoint changes and human pose variations under different visible cameras are the main challenges of single modality person Re-ID. Existing works mainly focuses on representation learning [42,16,21,26,30] and metric learning [43,24,36], and have achieved excellent performance on the widely-used datasets. However, due to the large cross-modality discrepancies, these methods may not be applicable for the VI-ReID task in the practical surveillance scenarios. Differently, our method proposes a modality-level alignment module and an instance-alignment module to learn a unified ReID framework for both RGB and IR modalities.

Visible-Infrared Person Re-ID. Visible-Infrared person Re-ID is challenging due to the cross-modal discrepancies between visible and infrared images. To address this challenge, existing methods [9,15,6,37,29,3,39,41] mainly focus on learning modality-shared feature representations to achieve modality-level alignment. Some image translation-based methods [3,14,28,29] are developed to firstly achieve modality unification and then learn modality-shared representations. Wang et al. [29] propose an end-to-end alignment generative adversarial network by exploiting pixel alignment and feature alignment jointly. [28] generates cross-modality paired-images and performs both global set-level and fine-grained instance-level alignments. Another line of works [37,7,8,18] attempts to learn modality-shared features by designing various two-stream architectures. Ye et al. [37] propose a novel modality-aware collaborative ensemble learning method with the middle-level sharable two-stream network. [7] exploits the optimal two-stream architecture by neural architecture search for VI-ReID. However, the modality-specific features are generally ignored by the above methods, which limits the upper bound of the discrimination ability of the feature representation. To address this limitation, modality compensation methods [32,20] are proposed to compensate for the missing modality features. [32] generates multi-spectral images to compensate for the lacking specific information by utilizing the generative adversarial network. In [20], a cross-modality shared specific feature transfer algorithm is proposed to explore the potential of both the modality-shared information and the modality-specific features. However, the compensated features extracted by [20] only depend on the current mini-batch, which causes the inconsistency of generated modality features when the samples are in the different

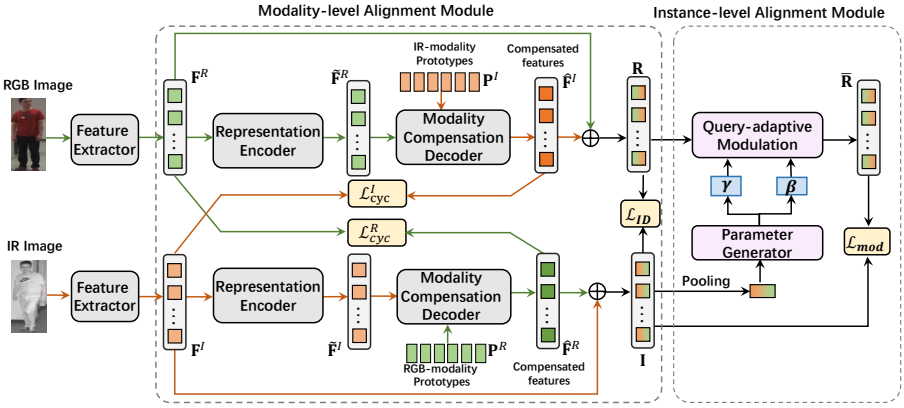


Fig. 2. Framework of our Cross-Modality Transformer. (1) The modality-level alignment module compensates for the missing modality information by the cross-attention between the modality prototypes $\mathbf{P}^I/\mathbf{P}^R$ and features $\tilde{\mathbf{F}}^R/\tilde{\mathbf{F}}^I$. A modality consistency loss $\mathcal{L}_{cyc}$ is proposed to make the modality prototypes focus on the corresponding global modality information. (2) The instance-level alignment module leverages the characteristics of the given query to automatically adapt the other instance features by the query-adaptive modulation, which helps align the query and gallery instances in the same class. For simplicity, we take an IR image $\mathbf{I}$ as an example and show the process of using $\mathbf{I}$ as the query for the modulation in the figure.

mini-batches. Differently, we introduce two sets of global modality prototypes to represent the RGB and IR modalities respectively, which can be used as the global basis to learn modality compensation for every sample.

Transformer in Person Re-ID. Transformers have recently received increasing attention for computer vision tasks, including image classification [5,19], object detection [1,45], image segmentation [31,19], and so on. Most existing Re-ID methods apply Transformer to a single modality. For example, He et al. [11] utilize a pure-transformer with a side information embedding and a jigsaw patch module to learn discriminative features. Li et al. [17] exploit a transformer architecture to discover diverse parts for occluded person Re-ID. Different from the above methods, our CMT is designed for VI-ReID to compensate for the missing modality-specific information.

3 Our Method

In this section, we introduce the details of the proposed Cross-Modality Transformer (CMT) for the VI-ReID task. As shown in Figure 2, the proposed CMT mainly consists of two modules. (1) The modality-level alignment module aims at compensating for the missing modality-specific information via a Transformer encoder-decoder architecture. (2) The instance-level alignment module is responsible for aligning the gallery instances with the query instance in the same class by a query-adaptive feature modulation mechanism.

3.1 Modality-level Alignment Module

In order to achieve the modality-level alignment, we follow the architecture of Transformer [27] and design a representation encoder and a modality compensation decoder, which is able to adaptively compensate for the lacking modality-special information. Different from the previous modality compensation method [20] that depends on the information in the mini-batch, we design two set of learnable modality prototypes to provide the global modality information for more robust modality compensation.

Representation Encoder. Following existing works [8,25,40,39], we adopt a two-stream network based on ResNet-50 as our feature extractor for the RGB and IR modalities, where the first two stages are parameter-independent and the latter three stages are parameter-shared. We first use the feature extractor ϕ to extract the feature maps for the given visible images and infrared images. Then, following the practice in the part-based methods [39,18], we horizontally split the feature maps into p non-overlapping parts with a region pooling strategy. In this way, the RGB and IR images can be represented by the set of the part features: $\mathbf{F}^R = \left[f_1^R; f_2^R; \ldots; f_p^R\right] \in \mathbb{R}^{p\times d}$ and $\mathbf{F}^I = \left[f_1^I; f_2^I; \ldots; f_p^I\right] \in \mathbb{R}^{p\times d}$, where $f_i^R, f_i^I \in \mathbb{R}^d$ indicate the i^{th} part feature of two modalities. These part features are taken as the inputs of transformer encoder. For the simplicity of the description, we take the RGB image as an example.

In the representation encoder, we adopt a self-attention layer to capture the inter-relationship between the local human parts to refine the part representations. Specifically, we take part features as the query $\mathbf{Q}$, key $\mathbf{K}$ and value $\mathbf{V}$. We generate the ($\mathbf{Q}$, $\mathbf{K}$, $\mathbf{V}$) triplets by independent linear projection layers:

$$\mathbf{Q} = \mathbf{F}^R\mathbf{W}^Q, \quad \mathbf{K} = \mathbf{F}^R\mathbf{W}^K, \quad \mathbf{V} = \mathbf{F}^R\mathbf{W}^V, \tag{1}$$

where $\mathbf{W}^Q \in \mathbb{R}^{d\times d}, \mathbf{W}^K \in \mathbb{R}^{d\times d}, \mathbf{W}^V \in \mathbb{R}^{d\times d}$ are linear projections, and $\mathbf{Q}, \mathbf{K}, \mathbf{V} \in \mathbb{R}^{p\times d}$. Then, the attention weights between the query $\mathbf{Q}$ and the key $\mathbf{K}$ can be derived by the inner product with a scaling operation and a Softmax normalization. Based on the attention weights, we can obtain the refined part features as the weighted sum of values $V \in \mathbb{R}^{p\times d}$. Formally:

$$Attention\left(\mathbf{Q}, \mathbf{K}, \mathbf{V}\right) = softmax\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}}\right)\mathbf{V}. \tag{2}$$

Equation (2) is implemented with the multi-head attention mechanism, and a feed-forward network is also applied. For more details, please refer to the work [27].

Modality Compensation Decoder. In the modality compensation decoder, we introduce two sets of learnable modality prototypes to represent the global modality information of RGB and IR modalities respectively, which are denoted as $\mathbf{P}^R = \left[p_1^R; p_2^R; \ldots; p_p^R\right], \mathbf{P}^I = \left[p_1^I; p_2^I; \ldots; p_p^I\right] \in \mathbb{R}^{p\times d}$, where p_i^I is the modality prototype for the i-th part feature in the IR modality. Following the standard architecture of the transformer [27], we first use a self-attention layer to incorporate the local context information between prototypes. The implementation is the same as the self-attention layer in the representation encoder, but the

keys, queries and values arise from IR/RGB modality prototypes. Subsequently, we compensate for the missing modality features by the cross-attention between the modality prototypes and part features. Given the RGB/IR feature map of the encoder output $\tilde{\mathbf{F}}^R = \left[\tilde{f}_1^R; \tilde{f}_2^R; \ldots; \tilde{f}_p^R\right]$ / $\tilde{\mathbf{F}}^I = \left[\tilde{f}_1^I; \tilde{f}_2^I; \ldots; \tilde{f}_p^I\right] \in \mathbb{R}^{p\times d}$ ($\tilde{f}_i^R$ and $\tilde{f}_i^I$ represent the i^{th} part feature of RGB and IR samples, respectively), we take RGB features as an example to elaborate on the compensation process of the IR modality. Specifically, the IR modality prototypes $\mathbf{P}^I$ are taken as the queries $\mathbf{Q}^I$, and the RGB part features $\tilde{\mathbf{F}}^R$ are taken as keys $\mathbf{K}^R$ and values $\mathbf{V}^R$ of the modality compensation decoder. Formally:

$$\mathbf{Q}^I = \mathbf{P}^I\mathbf{W}^Q, \mathbf{K}^R = \tilde{\mathbf{F}}^R\mathbf{W}^K, \mathbf{V}^R = \tilde{\mathbf{F}}^R\mathbf{W}^V. \tag{3}$$

Then, we can obtain the dot-production attention scores between queries $\mathbf{Q}^I$ and keys $\mathbf{K}^R$, which can be regarded as the soft correspondences between the modality prototypes and part features. To compensate for the missing modality features, we can project the part features into the corresponding modality space according to the attention weights. Concretely, the compensated IR part features $\hat{\mathbf{F}}^I = \left[\hat{f}_1^I; \hat{f}_2^I; \ldots; \hat{f}_p^I\right]$ for the RGB sample are derived as the weighted sum over all values $\mathbf{V}^R$:

$$\begin{aligned}\hat{\mathbf{F}}^I &= \mathit{Attention}\left(\mathbf{Q}^I, \mathbf{K}^R, \mathbf{V}^R\right) \\ &= \mathit{softmax}\left(\frac{\mathbf{Q}^I(\mathbf{K}^R)^T}{\sqrt{d}}\right)\mathbf{V}^R,\end{aligned} \tag{4}$$

where $\hat{\mathbf{F}}^I \in \mathbb{R}^{p\times d}$. Similarly, the compensated RGB part features $\hat{\mathbf{F}}^R$ for the samples with the IR modality can also be derived by Equation (3) and Equation (4). Finally, the complete modality representations can be acquired by combining the original features and the compensated modality features:

$$\mathbf{R} = \mathbf{F}^R + \hat{\mathbf{F}}^I, \quad \mathbf{I} = \mathbf{F}^I + \hat{\mathbf{F}}^R, \tag{5}$$

where $\mathbf{R}$ and $\mathbf{I}$ are the complete RGB and IR modality representations, respectively. These complete representations are in the shared embedding space, where the samples with different modalities can be aligned well. In this way, our modality compensation decoder can achieve a robust modality-level alignment and bridge the inter-modality discrepancies, which can facilitate a better cross-modality retrieval.

Modality Consistency Loss. As we have no ground truths for the compensated modality features, the learning of the decoder is difficult. To resolve this issue, we design a modality consistency loss to guide the learning of modality prototypes, which constrains the compensated RGB/IR features to be aligned with the real RGB/IR modality features. We first compute the two centroid features of each identity for two modalities in the mini-batch:

$$\mathbf{C}_i^R = \frac{1}{K}\sum_{j=1}^{K}\mathbf{F}_{i,j}^R, \quad \mathbf{C}_i^I = \frac{1}{K}\sum_{j=1}^{K}\mathbf{F}_{i,j}^I, \tag{6}$$

where $\mathbf{F}_{i,j}^R, \mathbf{F}_{i,j}^I$ denote the j^{th} RGB/IR image feature of the i^{th} person in the mini-batch, and $\mathbf{C}_i^R$, $\mathbf{C}_i^I$ represent the RGB/IR centroid features of the i^{th} person. Based on the centroids, the modality consistency loss $\mathcal{L}_{cyc}^R$ and $\mathcal{L}_{cyc}^I$ for RGB/IR modalities are defined as:

$$\mathcal{L}_{cyc}^R = \frac{1}{NK} \sum_{i=1}^{N} \sum_{j=1}^{K} \|\hat{\mathbf{F}}_{i,j}^R - \mathbf{C}_i^R\|_2, \tag{7}$$

$$\mathcal{L}_{cyc}^I = \frac{1}{NK} \sum_{i=1}^{N} \sum_{j=1}^{K} \|\hat{\mathbf{F}}_{i,j}^I - \mathbf{C}_i^I\|_2, \tag{8}$$

where $\hat{\mathbf{F}}_{i,j}^R, \hat{\mathbf{F}}_{i,j}^I$ are the compensated RGB, IR features. Constrained by $\mathcal{L}_{cyc}^R$ and $\mathcal{L}_{cyc}^I$, the modality prototypes are forced to learn the corresponding modality information to approach the real modality features, which consequently facilitates a more reliable modality compensation.

ID Loss. To guide the complete representations $\mathbf{R}$ and $\mathbf{I}$ to focus on the ID-related discriminative information, we design an ID loss consisting of an identity classification loss $\mathcal{L}_{cls}$ and a hetero-center based triplet loss $\mathcal{L}_{hc\_tri}$ following the practice in [18]. Concretely, the ID loss is formulated as:

$$\mathcal{L}_{ID} = \mathcal{L}_{cls} + \mathcal{L}_{hc\_tri} \tag{9}$$

$$\mathcal{L}_{cls} = E\left(-log\, p\left(R\right)\right) + E\left(-log\, p\left(I\right)\right) \tag{10}$$

$$\mathcal{L}_{hc\_tri} = E\left[\alpha + d_{c_a,c_p} - d_{c_a,c_n}\right]_+, \tag{11}$$

where $p\left(\right)$ is the probability of correct prediction, and E represents the expectation. In Equation (11), c_a denotes the centroid feature calculated by the RGB features $\mathbf{R}$ or IR features $\mathbf{I}$ in the current mini-batch. c_a and c_p form a positive pair of centroid features belonging to the same person but with different modalities, while c_a and c_n form a negative pair of centroid features belonging to different persons, and α is a margin parameter.

3.2 Instance-level Alignment Module

Due to the large intra-class variations like viewpoint changes and background clutter, the feature distribution of different samples with the same ID has large differences. Therefore, we propose an instance-level alignment module, where we leverage the characteristics of the given query to automatically adapt the instance features by the query-adaptive modulator. Specifically, the modulator employs an affine transformation to excite the query-related channels by the learned modulation parameters. Next we will give the details.

Parameter Generator. The Instance-level Alignment Module is symmetry for the visible and infrared modality. Given any sample feature $X \in \mathbb{R}^{p \times d}$ in the current mini-batch from RGB or IR modality, we take it as the query and transform the query characteristics into the modulation parameters. Concretely,

we propose two parameter generators g_γ and g_β to obtain the channel-wise modulation parameters, *i.e.*, the scaling parameter γ and the shifting parameter β. Each generator contains two linear layers, with the first layer followed by a ReLU activation function. Formally, the modulation parameters γ and β are generated by

$$\gamma = g_\gamma(GAP(X)), \beta = g_\beta(GAP(X)), \tag{12}$$

where $\gamma, \beta \in \mathbb{R}^d$, and GAP represents the global average pooling, which is used to aggregate the part features. After the end-to-end training, the parameter generators g_γ and g_β can extract key characteristics in the query feature, and project them into the modulation weights that indicate which channels could be useful in the instance-level alignment. Although sharing a similar network structure with SENet [12], the parameter generator is designed to modulate other samples rather than enhance the samples themselves.

Query-adaptive Modulation. The modulation parameters reveal the most discriminative patterns of X, and can be employed to perform the query-adaptive modulation on the other sample features Y in the current mini-batch to achieve the instance-level alignment. Specifically, the query-adaptive modulation layer employs an affine transformation by the scaling parameter γ and the shifting parameterβ on Y:

$$\bar{Y_i} = Y_i \odot \gamma + \beta, \tag{13}$$

where $\odot$ denotes a point-wise vector multiplication, and Y_i represents the i^{th} part features of the sample Y, $\bar{Y_i}$ is the modulated feature. In the modulation, the crucial query-related channels of the Y can be strengthened and the irrelevant channels can be suppressed based on the modulation weights of γ and β. In this way, the instances that have the same ID with the query can be better aligned together. During the testing, the query-adaptive feature modulation will adjust the gallery representations according to the query features, which promotes the alignment between the query and gallery with the same ID, and contributes to a better retrieval.

Modulation Discriminative Loss. Without the constraints, the modulation on the channels may cause some disturbances to the representations, which will undermine the discrimination power of each instance. To help the modulated features preserve the discriminative ability, we propose a modulation discriminative loss to restrain the modulated features, which takes the form of the triplet loss:

$$\mathcal{L}_{mod} = E\left[\alpha + d_{X,\bar{Y}_p} - d_{X,\bar{Y}_n}\right]_+, \tag{14}$$

where X and $\bar{Y}_p$ form a positive pair of feature vectors belonging to the same person, X and $\bar{Y}_n$ form a negative pair of feature vectors belonging to different persons, α is a margin parameter.

3.3 Training and Inference

For the VI-ReID task, our proposed CMT is trained by minimizing the overall objective with identity labels as defined in

$$\mathcal{L}_{CMT} = \mathcal{L}_{ID} + \mathcal{L}_{cyc}^{R} + \mathcal{L}_{cyc}^{I} + \lambda \mathcal{L}_{mod}. \tag{15}$$

During the testing stage, we first extract query features, and then generate modulation parameters according to query features to adjust the feature embedding of galleries. Finally, we reshape the feature dimension to $\mathbb{R}^{pd}$ for the feature retrieval.

4 Experiments

In this section, we first introduce datasets and implementation details. Then, we show experimental results and some visualizations.

4.1 Dataset and Evaluation Protocol

SYSU-MM01 [34] is the first large-scale benchmark dataset for VI-ReID collected by 6 cameras, including 4 visible and 2 infrared cameras. Specially, four cameras are deployed in the outdoor environments and two are deployed in the indoor environments. SYSU-MM01 contains 491 persons with a total of 287,628 visible images and 15,792 infrared images. The training set contains 395 persons, including 22258 visible images and 11909 infrared images. The test set contains 96 persons, with 3,803 IR images for query and 301/3010 (one-shot/multi-shot) randomly selected RGB images as the gallery. Meanwhile, it contains two different testing settings, all-search and indoor-search settings. Detailed descriptions of the experimental settings can be found in [34].

RegDB [23] is collected by a dual-camera system, including one visible and one infrared camera. There are 412 identities and 8,240 images in total, with 206 identities for training and 206 identities for testing. For each person, there are 10 visible images and 10 infrared images. The testing stage also contains two evaluation settings. One is Visible to Infrared to search IR images from a RGB image. The other setting is Infrared to Visible to search RGB images from a IR image. The evaluation procedure is repeated for 10 trials to record the mean values.

Evaluation Protocol. Two evaluation metrics are used to measure the performance. The first one is the Cumulative Matching Characteristic (CMC) curves. The CMC represents the probability that a query identity appears in different sized candidate lists. We report the rank-1,10,20 accuracy in experiments. The other is the Mean Average Precision (mAP).

4.2 Implementation Details

The proposed method is implemented with the PyTorch framework on a single RTX3090Ti GPU. Following the existing methods [33,25,20], we choose ResNet-50 [10] pretrained on ImageNet as the backbone network and reduce the stride of

Table 1. Performance comparison with state-of-the-art methods on SYSU-MM01 dataset. Rank-k accuracy (%) and mAP (%) are reported.

Method	Venue	All-Search								Indoor-Search							
		Single-Shot				Multi-Shot				Single-Shot				Multi-Shot			
		R1	R10	R20	mAP	R1	R10	R20	mAP	R1	R10	R20	mAP	R1	R10	R20	mAP
Zero-Padding [34]	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	18.86
cmGAN [4]	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
D^2RL [32]	CVPR-19	28.90	70.60	82.40	29.20	-	-	-	-	-	-	-	-	-	-	-	-
Hi-CMD [3]	CVPR-20	34.94	77.58	-	35.94	-	-	-	-	-	-	-	-	-	-	-	-
JSIA-ReID [28]	AAAI-20	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 [29]	ICCV-19	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
cm-SSFT(sq) [20]	CVPR-20	47.70	-	-	54.10	-	-	-	-	57.40	-	-	59.10	-	-	-	-
XIV [15]	AAAI-20	49.92	89.79	95.96	50.73	-	-	-	-	-	-	-	-	-	-	-	-
DDAG [39]	ECCV-20	54.75	90.39	95.81	53.02	-	-	-	-	61.02	94.06	98.41	67.98	-	-	-	-
LbA [25]	ICCV-21	55.41	-	-	54.1	57.4	-	-	59.1	-	-	-	-	-	-	-	-
NFS [2]	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	97.7	99.51	61.45
HCT [18]	TMM-20	61.68	93.1	97.17	57.51	-	-	-	-	63.41	91.69	95.28	68.17	-	-	-	-
CM-NAS [7]	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
MCLNet [8]	ICCV-21	65.40	93.33	97.14	61.98	-	-	-	-	72.56	96.98	99.20	76.58	-	-	-	-
SMCL [33]	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 [35]	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**
CMT(our)	ECCV-22	**71.88**	**96.45**	**98.87**	**68.57**	**80.23**	**97.91**	**99.53**	**63.13**	**76.9**	97.68	**99.64**	79.91	**84.87**	99.41	**99.97**	74.11

the last convolutional block from 2 to 1. For each mini-batch, we randomly choose 8 identities from each modality and sample 8 person images for each identity. The input images are first resized to 384×144, then we adopt random cropping with zero-padding, random horizontal flipping, and random erasing for data augmentation. In addition, we use the Adam optimizer for optimization with an initial learning rate of 3.5×10^{-4}, and the weight decay is set to 5×10^{-4}. We decay the learning rate by 0.1 and 0.01 at 60 and 90 epochs. The whole training process consists of 120 epochs. The number of part features p is set to 6. The hype-parameters λ is set to 0.2.

4.3 Comparison with the State-of-the-art Methods

Comparisons on SYSU-MM01. We compare our CMT with various state-of-the-art methods under both all-search and single-search settings. As shown in Table 1, our CMT ranks either the first or the second among all settings, and sets the new state-of-the-art results in all-search setting, which strongly proves the effectiveness of our method. In the indoor-search setting, our method also performs comparably with the state-of-the-art methods. Based on the results, we have the following observations. (1) Compared with the methods (cmGAN [4], Hi-CMD [3],AlignGAN [29]) that only focus on learning modality-shared features by feature disentanglement, our method achieves much better performance on all settings. This is because the modality-shared features lose some useful identity information, such as colors. Therefore, with modality-shared cues only, the upper bound of the discrimination ability of the feature representation is limited. Differently, we design a modality-level alignment module to adaptively compensate for the lacking modality-special information via a transformer encoder-decoder architecture. (2) Compared with the best modality compensation method (i.e., cm-SSFT [20] in a multi-query setting), our method improves the Rank-1 accuracy and mAP by 10.28% and 5.37% in the all-search single shot setting. The reason is that the compensated features in [20] are produced solely based on

Table 2. Comparison of the Rank-k accuracy (%) and mAP (%) performances with state-of-the-art methods on RegDB.

Method	Venue	Visible to Infrared				Infrared to Visible			
		R1	R10	R20	mAP	R1	R10	R20	mAP
Zero-Padding [34]	ICCV-17	17.75	34.21	44.35	18.90	16.63	34.68	44.25	17.82
D^2RL [32]	CVPR-19	43.4	66.1	76.3	44.1	-	-	-	-
JSIA-ReID [28]	AAAI-20	48.50	-	-	48.90	-	-	-	-
AlignGAN [29]	ICCV-19	57.90	-	-	53.60	56.30	-	-	53.40
XIV [15]	AAAI-20	-	-	-	-	62.21	83.13	91.72	60.18
cm-SSFT(sq) [20]	CVPR-20	65.4	-	-	65.6	63.8	-	-	64.2
DDAG [39]	ECCV-20	69.34	86.19	91.49	63.46	68.06	85.15	90.31	61.80
Hi-CMD [3]	CVPR-20	70.93	86.39	-	66.04	-	-	-	-
LbA [25]	ICCV-21	74.17	-	-	67.64	72.43	-	-	65.46
MCLNet [8]	ICCV-21	80.31	92.70	96.03	73.07	75.93	90.93	94.59	69.49
NFS [2]	CVPR-21	80.54	91.96	95.07	72.1	77.95	90.45	93.62	69.79
MPANet [35]	CVPR-21	83.7	-	-	80.9	82.8	-	-	80.7
SMCL [33]	ICCV-21	83.93	-	-	79.83	83.05	-	-	78.57
CM-NAS [7]	ICCV-21	84.54	95.18	97.85	80.32	82.57	94.51	97.37	78.31
HCT [18]	TMM-20	91.05	97.16	98.57	83.28	89.3	96.41	98.16	81.46
CMT(our)	ECCV-22	**95.17**	**98.82**	**99.51**	**87.3**	**91.97**	**97.92**	**99.07**	**84.46**

the samples of the current mini-batch. This strategy suffers from a certain randomness, and does not match the default single query settings of most methods. Notably, we introduce several modality prototypes to store the global modality characteristics without relying on the current mini-batch. (3) Compared with JSIA-ReID [28] that is based on instance-level alignment between the cross-modality paired images generated by the GAN, our method acquires a better performance in all results. This is because different from JSIA-ReID, we exploit query-adaptive feature modulation to conduct more diverse and flexible instance-level alignment. In our method, the gallery instances can be adaptively refined according to the query features, while other methods do not take this into account.

Comparisons on RegDB. As shown in Table 2, it can be seen that our CMT has distinct advantages over the state-of-the-art methods on RegDB. Under the Visible to Infrared setting, compared with the state-of-the-art HCT [18], our method improves the Rank-1 accuracy and mAP by 4.12% and 4.02%. When switching to the Infrared to Visible setting, our method surpasses the HCT by 4.02% and 3% in terms of the Rank-1 accuracy and the mAP, respectively. Hence, it can be proved that our proposed method is robust against different query settings.

4.4 Ablation Study

In this section, we perform detailed ablation studies on SYSU-MM01 dataset under the all-search setting to evaluate each component of our CMT. We denote the Modality-level Alignment Module as MAM and the Instance-level Alignment Module as IAM. The results are shown in Table 3.

Baseline. We adopt the HCT [18] as our baseline method, which explores the two-stream network with shared parameters and uses a hetero-center based

Table 3. Analysis of the effectiveness of different components on SYSU-MM01 dataset under the all-search setting. Rank-k accuracy (%)and mAP (%) are reported.

Base	MAM	IAM	Rank-1	Rank-10	Rank-20	mAP
✓	✗	✗	65.35	93.57	97.58	64.27
✓	✓	✗	70.55	95.27	98.21	66.50
✓	✗	✓	68.5	94.21	98.34	67.25
✓	✓	✓	**71.88**	**96.45**	**98.87**	**68.57**

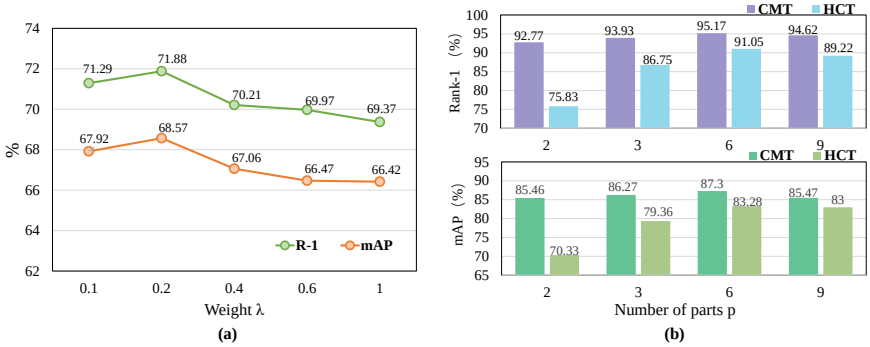
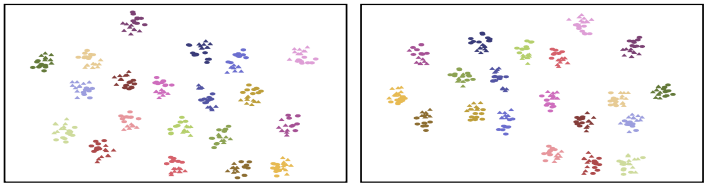


Fig. 3. The effect of weight λ in Equation (15) on SYSU-MM01 dataset under the all-search setting and the number of parts p on RegDB dataset. Rank-1 and mAP (%) are reported.

triplet loss to improve the traditional triplet loss. In addition, we replace the optimizer with the Adam optimizer and add random erasing as extra data augmentation. The details of the implementation can be found in Section 4.2.

Effectiveness of the Modality-level Alignment. Compared with the baseline model, the modality-level alignment module improves the Rank-1 accuracy and mAP by 5.2% and 2.23%. The improvements can be mainly ascribed to two reasons. For one thing, we automatically explore the modality prototypes by the modality consistency constraint, which can adaptively learn the modality-related information. The other reason is that, we conduct the modality feature compensation by the transformer, which can project the features of different modalities into a common complete space to achieve a better modality-level alignment.

Effectiveness of the Instance-level Alignment. Compared with the baseline model, adding the instance-level alignment, the performance is greatly improved by 2.98% and up to 67.25% mAP. Besides, on top of the modality-level alignment, the instance-level alignment can still achieve 2.07% improvements in mAP. This shows that the instance-level alignment is useful to reduce the distances of the samples in the same class. The complete version of our CMT gives the best results on SYSU-MM01 dataset under all-search setting, achieving a whopping accuracy gain of 6.53% and 4.3% on in Rank-1 and mAP, which proves the effectiveness of CMT.



(a) RGB modality ($\boldsymbol{F}^R$ and $\widetilde{\boldsymbol{F}}^R$) (b) IR modality ($\boldsymbol{F}^I$ and $\widetilde{\boldsymbol{F}}^I$)

Fig. 4. The t-SNE visualization of features on SYSU-MM01 dataset.The colors represent different categories. Circles represent original RGB/IR features and triangles represent compensated RGB/IR features.

4.5 Model Analysis

Parameters Analysis. We first evaluate the effect of the weight λ in Equation (15) on SYSU-MM01 dataset under the all-search setting. The Rank-1 and the mAP results of CMT with different λ are exhibited in Figure 3 (a). The most suitable parameter setting is to set λ as 0.2. Then, we compare the performance of CMT and our baseline model HCT [18] with different number of parts p. As shown in Figure 3 (b), with p increasing, the performance keeps improving before p arrives 6 on RegDB dataset. This is because a bigger p allows the network to pay more attention to the details. Besides, we can observe that CMT shows surprisingly powerful results and significant improvements over the baseline under the same p setting. Compared with the HCT, CMT is more robust to p, which further verifies the effectiveness of our method.

Visualization Analysis. To further verify the effectiveness of our modality-level alignment module, we use t-SNE [22] to visualize the original modality features ($\mathbf{F}^R$ and $\mathbf{F}^I$) and the compensated modality features ($\hat{\mathbf{F}}^R$ and $\hat{\mathbf{F}}^I$). As shown in Figure 4, compensated RGB/IR features are aligned with original RGB/IR features of the same ID in the feature space. It proves that our work can compensate for the lacking modality information to achieve a better modality-level alignment.

5 Conclusion

In this paper, we propose a novel Cross-Modality Transformer (CMT) to jointly explore a modality-level alignment module and an instance-level module for VI-REID. The proposed modality-level alignment module is able to compensate for the missing modality-specific information via a Transformer encoder-decoder architecture. We have also designed an instance-level alignment module to adaptively adjust the sample features, which is achieved by query-adaptive feature modulation. Extensive experimental results on two standard benchmarks demonstrate that our model performs favorably against state-of-the-art methods.

Acknowledgments This work was partially supported by the National Nature Science Foundation of China (62022078, 12150007, 62021001), National Defense Basic Scientific Research Program (JCKY2020903B002), and University Synergy Innovation Program of Anhui Province No. GXXT-2019-025.

References

1. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European conference on computer vision. pp. 213–229. Springer (2020)
2. Chen, Y., Wan, L., Li, Z., Jing, Q., Sun, Z.: Neural feature search for rgb-infrared person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 587–597 (2021)
3. Choi, S., Lee, S., Kim, Y., Kim, T., Kim, C.: Hi-cmd: Hierarchical cross-modality disentanglement for visible-infrared person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 10257–10266 (2020)
4. Dai, P., Ji, R., Wang, H., Wu, Q., Huang, Y.: Cross-modality person re-identification with generative adversarial training. In: International Joint Conference on Artificial Intelligence. vol. 1, p. 2 (2018)
5. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
6. Feng, Z., Lai, J., Xie, X.: Learning modality-specific representations for visible-infrared person re-identification. IEEE Transactions on Image Processing **29**, 579–590 (2019)
7. Fu, C., Hu, Y., Wu, X., Shi, H., Mei, T., He, R.: Cm-nas: Cross-modality neural architecture search for visible-infrared person re-identification. arXiv preprint arXiv:2101.08467 (2021)
8. Hao, X., Zhao, S., Ye, M., Shen, J.: Cross-modality person re-identification via modality confusion and center aggregation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 16403–16412 (2021)
9. Hao, Y., Wang, N., Li, J., Gao, X.: Hsme: hypersphere manifold embedding for visible thermal person re-identification. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 8385–8392 (2019)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016)
11. He, S., Luo, H., Wang, P., Wang, F., Li, H., Jiang, W.: Transreid: Transformer-based object re-identification. arXiv preprint arXiv:2102.04378 (2021)
12. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7132–7141 (2018)
13. Jiang, K., Zhang, T., Zhang, Y., Wu, F., Rui, Y.: Self-supervised agent learning for unsupervised cross-domain person re-identification. IEEE Transactions on Image Processing **29**, 8549–8560 (2020)
14. Kniaz, V.V., Knyaz, V.A., Hladuvka, J., Kropatsch, W.G., Mizginov, V.: Thermalgan: Multimodal color-to-thermal image translation for person re-identification in multispectral dataset. In: Proceedings of the European Conference on Computer Vision Workshops. pp. 0–0 (2018)
15. Li, D., Wei, X., Hong, X., Gong, Y.: Infrared-visible cross-modal person re-identification with an x modality. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 4610–4617 (2020)
16. Li, W., Zhu, X., Gong, S.: Harmonious attention network for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2285–2294 (2018)

17. Li, Y., He, J., Zhang, T., Liu, X., Zhang, Y., Wu, F.: Diverse part discovery: Occluded person re-identification with part-aware transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2898–2907 (2021)
18. Liu, H., Tan, X., Zhou, X.: Parameter sharing exploration and hetero-center triplet loss for visible-thermal person re-identification. IEEE Transactions on Multimedia (2020)
19. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10012–10022 (2021)
20. Lu, Y., Wu, Y., Liu, B., Zhang, T., Li, B., Chu, Q., Yu, N.: Cross-modality person re-identification with shared-specific feature transfer. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 13379–13389 (2020)
21. Luo, H., Gu, Y., Liao, X., Lai, S., Jiang, W.: Bag of tricks and a strong baseline for deep person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 0–0 (2019)
22. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**(11) (2008)
23. Nguyen, D.T., Hong, H.G., Kim, K.W., Park, K.R.: Person recognition system based on a combination of body images from visible light and thermal cameras. Sensors **17**, 605 (2017)
24. Oh Song, H., Xiang, Y., Jegelka, S., Savarese, S.: Deep metric learning via lifted structured feature embedding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4004–4012 (2016)
25. Park, H., Lee, S., Lee, J., Ham, B.: Learning by aligning: Visible-infrared person re-identification using cross-modal correspondences. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 12046–12055 (2021)
26. Sun, Y., Zheng, L., Yang, Y., Tian, Q., Wang, S.: Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In: Proceedings of the European Conference on Computer Vision. pp. 480–496 (2018)
27. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems. pp. 5998–6008 (2017)
28. Wang, G.A., Zhang, T., Yang, Y., Cheng, J., Chang, J., Liang, X., Hou, Z.G.: Cross-modality paired-images generation for rgb-infrared person re-identification. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 12144–12151 (2020)
29. Wang, G., Zhang, T., Cheng, J., Liu, S., Yang, Y., Hou, Z.: Rgb-infrared cross-modality person re-identification via joint pixel and feature alignment. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3623–3632 (2019)
30. Wang, G., Yuan, Y., Chen, X., Li, J., Zhou, X.: Learning discriminative features with multiple granularities for person re-identification. In: Proceedings of the ACM International Conference on Multimedia. pp. 274–282 (2018)
31. Wang, W., Xie, E., Li, X., Fan, D.P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L.: Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 568–578 (2021)

32. Wang, Z., Wang, Z., Zheng, Y., Chuang, Y.Y., Satoh, S.: Learning to reduce dual-level discrepancy for infrared-visible person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 618–626 (2019)
33. Wei, Z., Yang, X., Wang, N., Gao, X.: Syncretic modality collaborative learning for visible infrared person re-identification. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 225–234 (2021)
34. Wu, A., Zheng, W.S., Yu, H.X., Gong, S., Lai, J.: Rgb-infrared cross-modality person re-identification. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 5380–5389 (2017)
35. Wu, Q., Dai, P., Chen, J., Lin, C.W., Wu, Y., Huang, F., Zhong, B., Ji, R.: Discover cross-modality nuances for visible-infrared person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4330–4339 (2021)
36. Yang, X., Zhou, P., Wang, M.: Person reidentification via structural deep metric learning. IEEE Transactions on Neural Networks and Learning Systems **30**(10), 2987–2998 (2018)
37. Ye, M., Lan, X., Leng, Q., Shen, J.: Cross-modality person re-identification via modality-aware collaborative ensemble learning. IEEE Transactions on Image Processing **29**, 9387–9399 (2020)
38. Ye, M., Lan, X., Li, J., Yuen, P.: Hierarchical discriminative learning for visible thermal person re-identification. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 32 (2018)
39. Ye, M., Shen, J., J. Crandall, D., Shao, L., Luo, J.: Dynamic dual-attentive aggregation learning for visible-infrared person re-identification. In: Proceedings of the European Conference on Computer Vision. pp. 229–247. Springer (2020)
40. Ye, M., Shen, J., Lin, G., Xiang, T., Shao, L., Hoi, S.C.: Deep learning for person re-identification: A survey and outlook. IEEE Transactions on Pattern Analysis and Machine Intelligence (2021)
41. Ye, M., Wang, Z., Lan, X., Yuen, P.C.: Visible thermal person re-identification via dual-constrained top-ranking. In: International Joint Conference on Artificial Intelligence. vol. 1, p. 2 (2018)
42. Zheng, L., Yang, Y., Hauptmann, A.G.: Person re-identification: Past, present and future. arXiv preprint arXiv:1610.02984 (2016)
43. Zheng, W.S., Gong, S., Xiang, T.: Reidentification by relative distance comparison. IEEE Transactions on Pattern Analysis and Machine Intelligence **35**(3), 653–668 (2012)
44. Zhong, Z., Zheng, L., Cao, D., Li, S.: Re-ranking person re-identification with k-reciprocal encoding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1318–1327 (2017)
45. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable detr: Deformable transformers for end-to-end object detection. arXiv preprint arXiv:2010.04159 (2020)