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

Tengfei Liang®, Yi Jin®, Member, IEEE, Wu Liu®, Member, IEEE, and Yidong Li®, Senior Member, IEEE

Abstract—The visible-infrared person re-identification (VI-ReID) is a challenging ReID task, which aims to retrieve and match the same identity’s images between the heterogeneous visible and infrared modalities. Thus, the core of this task is to bridge the huge gap between these two modalities. The existing methods mainly face the problem of insufficient perception of modality information, and can not learn good discriminative modality-invariant embeddings for identities, which limits their performance. To solve these problems, we propose a new cross-modality transformer-based method (CMTR) for this visible-infrared person re-identification task, which can explicitly mine the information of each modality and generate better discriminative features based on it. Specifically, to capture inherent characteristics of modalities, we design the novel modality embeddings, which are fused with tokens embeddings to encode modality information directly. Moreover, to enhance representation of modality embeddings and adjust the distribution of embeddings, we further propose a modality-aware enhancement loss based on the learned modality information, which contains two components to reduce intra-class distance and enlarging inter-class distance simultaneously. To our knowledge, this is the first exploration of applying pure transformer network to the cross-modality re-identification task. We implement extensive experiments on the public SYSU-MM01 and RegDB datasets, and compared with previous methods, our method achieves good performance with more compact and powerful embeddings for the cross-modality retrieval.

Index Terms—Cross-modality person re-id, modality-aware loss, modality embedding, transformer, VI-ReID.

I. INTRODUCTION

PERSON Re-Identification (ReID) task [1] aims to retrieve images of the given person among multiple different cameras with viewpoint and illumination changes, pose variations, etc. It has been studied for many years, and the corresponding methods [2], [3], [4], [5], [6], [7] achieve good performance. However, existing ReID methods mainly focus on the person retrieval in the single visible modality under RGB cameras, which constrains these methods to be used only during the daytime. To achieve the full-time intelligent video surveillance, based on the mechanism of existing surveillance cameras to automatically switch to infrared mode at night, the visible-infrared cross-modality person re-identification (VI-ReID) task [8], [9] is recently proposed to expand the application scope and attracts increasing attention of researchers in this field. The VI-ReID requires methods that can match images of the same identity between visible and infrared modalities, which is more challenging because of the huge heterogeneous gap.

The visible and infrared images are generated by cameras that capture light in different wavelength ranges. The former consists of three channels (red, green, and blue) with the color information, while the latter only contains one channel with infrared light radiation. They are intrinsically heterogeneous and different. To reduce the huge modality gap, a natural strategy proposed by researchers is to transform images of one modality into another. There are some GAN-based methods [10], [11], [12], [13], [14], [15] that attempt to learn the modality translation mapping. However, due to the heterogeneous imaging process, the same gray in infrared images can be totally different colors in visible images. Therefore, there is no reliable mapping relationship to support the generative model. Some recent methods have turned attention to the structural design of the convolutional neural network (CNN) model for more effective feature extraction. Based on the two-stream architecture, some models [7], [16], [17], [18], [19], [20] are designed to use shallow layers with modality-specific weights to extract shared features for different modalities and use deep layers with shared weights to learn discriminative features under the guidance of multiple metric learning losses. However, this strategy of implicit learning cannot well guarantee the hypothetical function of these layers without sufficient perception and deeper mining of the built-in modality characteristics and information, resulting in limited performance in this cross-modality task.

To solve the aforementioned problems and bridge the modality gap, we attempt to explore the new architecture. Different from CNN, the transformer model can directly model the global relationship, and each layer has a global receptive field without pooling layers, which can effectively capture fine-grained spatial context information. Furthermore, the weights of aggregation operation in its self-attention module can be dynamically adjusted according to the input data. Compared with...
We introduce the learnable modality embeddings (ME) to further design an illumination estimation that uses generative adversarial networks. Extensive experiments are conducted on the SYSU-MM01 dataset, introducing a Feature Distilling Generative Adversarial Network (FD-GAN) guided by human pose information to extract pose-unrelated and identity-related embeddings. For pose variations, resolution, illumination, etc., researchers have proposed a variety of solutions in recent years. Zhu et al. [24] proposed a novel method called Viewpoint-Aware Loss with Angular Regularization (VA-reID), projecting the extracted embeddings of different viewpoints into a unified hypersphere to eliminate person representation differences caused by viewpoints. Ge et al. [25] introduced a Feature Distilling Generative Adversarial Network (FD-GAN) guided by human pose information to extract pose-unrelated and identity-related embeddings. For resolution changes caused by the type of cameras or distance between cameras and persons, Zhang et al. [26] proposed the deep high-resolution pseudo-siamese framework (PS-HRNet), which effectively solves the problem of cross-resolution matching. Then, to alleviate the impact of illumination on person re-ID, Zhang et al. [27] further designed an illumination Estimation and Restoring framework (IER) and a Mixed Training strategy with both Original and Reconstructed images (MTOR). Their method significantly reduced the adverse effects of illumination changes and achieved higher performance. These research works greatly promoted the development and application of the single-modality visible person re-ID.

B. Visible-Infrared Person Re-ID

The Visible-infrared person re-identification attempts to recognize visible and infrared images of a person under cameras of different modalities. For the first time, Wu et al. clearly defined the VI-ReID task [8]. They contributed the challenging large-scale SYSU-MM01 dataset and proposed the basic Zero-Padding method. After that, many researchers put forward some new methods. For GAN-based methods, Dai et al. designed the cmGAN [10] that uses generative adversarial networks to integrate information from different modalities and improve the performance of VI-ReID. In this paper, we propose a new cross-modality transformer (CMTR) model, which is the first pure transformer-based exploration for the visible-infrared cross-modality person re-identification task.

The main contributions of our paper are as follows:

- We propose a new VI-ReID method, the cross-modality transformer (CMTR) network, which is the first pure transformer-based exploration for the visible-infrared cross-modality person re-identification task.
- We introduce the learnable modality embeddings (ME) to the CMTR network, which directly encode modality information and can be used effectively to alleviate the gap between the heterogeneous images.
- We design a novel modality-aware enhancement (MAE) loss function that enforces the ME to capture more helpful characteristics of each modality and adjust the distribution of extracted embeddings.
- Extensive experiments are conducted on the SYSU-MM01 and RegDB benchmarks and demonstrate the effectiveness of our CMTR method, which has better performance against the previous VI-ReID methods.

II. RELATED WORK

A. Person Re-Identification

The person re-identification is a fine-grained image retrieval task with the goal to match pedestrian images with the same identity under different cameras. To solve the intra-class discrepancy and inter-class confusion caused by factors such as viewpoint, pose variations, resolution, illumination, etc., researchers have proposed a variety of solutions in recent years. Zhu et al. [24] proposed a novel method called Viewpoint-Aware Loss with Angular Regularization (VA-reID), projecting the extracted embeddings of different viewpoints into a unified hypersphere to eliminate person representation differences caused by viewpoints. Ge et al. [25] introduced a Feature Distilling Generative Adversarial Network (FD-GAN) guided by human pose information to extract pose-unrelated and identity-related embeddings. For resolution changes caused by the type of cameras or distance between cameras and persons, Zhang et al. [26] proposed the deep high-resolution pseudo-siamese framework (PS-HRNet), which effectively solves the problem of cross-resolution matching. Then, to alleviate the impact of illumination on person re-ID, Zhang et al. [27] further designed an illumination Estimation and Restoring framework (IER) and a Mixed Training strategy with both Original and Reconstructed images (MTOR). Their method significantly reduced the adverse effects of illumination changes and achieved higher performance. These research works greatly promoted the development and application of the single-modality visible person re-ID.

B. Visible-Infrared Person Re-ID

The Visible-infrared person re-identification attempts to recognize visible and infrared images of a person under cameras of different modalities. For the first time, Wu et al. clearly defined the VI-ReID task [8]. They contributed the challenging large-scale SYSU-MM01 dataset and proposed the basic Zero-Padding method. After that, many researchers put forward some new methods. For GAN-based methods, Dai et al. designed the cmGAN [10] that uses generative adversarial networks to integrate information from different modalities and improve the performance of VI-ReID.
networks (GAN) to learn discriminative common representations for this cross-modality task. Wang et al. proposed the Dual-level Discrepancy Reduction Learning (D2RL) method [11], trying to handle the unified multi-spectral representation by image translation. Then Wang et al. introduced the AlignGAN method [13] to jointly exploit pixel alignment and feature alignment and reduce intra-modality variations. For the two-stream or multi-stream CNN methods, Ye et al. successively proposed the BDTR framework [16], the MAC model [28], the two-stream AGW [7], and the DDAG method [29], using special structure and loss constraints to make them learn discriminative features implicitly. Li et al. [30] designed a three-stream structure and introduced an auxiliary X modality to pull images from different modalities. Gao et al. [20] proposed a multi-feature space joint optimization (MSO) network based on the two-stream AGW, learning modality-sharable features in both the single-modality space and the common space. Liu et al. [19] further explored the setting of shared layers under the two-stream framework. Besides, some researchers introduced the relatively complicated strategy of blocking or pattern partition, such as the TSLFN+HC [31] and MPANet [32], trying to boost the performance. However, most existing methods learn the discriminative modality-invariant embeddings in an implicit way and ignore the direct mining and utilization of modality information. Different from these previous methods, our proposed CMTR method can effectively overcome this problem with the proposed modality embeddings and modality-aware enhancement loss.

C. Research of the Transformer

The Transformer method was first proposed by Vaswani et al. [23] to solve the machine translation task in the field of natural language processing (NLP). From then to now, it and its variants [33], [34] dominate multiple tasks in the NLP field for a long time. In the field of computer vision, Dosovitskiy et al. [35] successfully built a pure transformer model to handle the basic image classification task and got comparative performance against the CNN models. In recent years, some researchers applied it to lots of other visual tasks, showing its superiority to convolutional neural networks, such as object detection [36], [37], semantic segmentation [38], pose recognition [39], etc. Their work greatly promoted the development of these fields. Based on the vision transformer model, our method focuses on solving the core challenges of the cross-modality person re-identification, i.e., the huge heterogeneous modality gap between visible and infrared images, which is the first exploration in this task and achieves good results.

III. THE PROPOSED METHOD

In this section, the proposed visible-infrared cross-modality transformer (CMTR) network is explained in detail. We introduce the overall network structure in the first subsection. Then we focus on the designed modality embeddings and modality-aware enhancement loss and explain their definition and function in the next two subsections. In the last subsection, we give the overall formula of the objective function during the process of optimization.

A. Overall Network Structure

Our CMTR network is built with the vision transformer framework [35], and we adapt it to the VI-ReID task. For input images, we let vis and ir represent the visible modality and infrared modality. Thus, the visible image set is denoted as $X^{vis} = \{x^{vis} | x^{vis} \in \mathbb{R}^{C \times H \times W}\}$, and the infrared image set is denoted as $X^{ir} = \{x^{ir} | x^{ir} \in \mathbb{R}^{C \times H \times W}\}$. The $C$, $H$, $W$ denote the channel, height, and width of the image respectively. In a training batch, there are $B$ images with the same number of $x_i^{vis}$ and $x_i^{ir}$, where $i \in \{1, 2, \ldots, B/2\}$. As shown in Fig. 2(a), our

![Fig. 2. The framework of our proposed method. (a) The overall structure of visible-infrared cross-modality transformer (CMTR) network, which is based on the Vision Transformer (ViT) backbone with multiple loss constraints. (b) The conceptual illustration of the designed visible-infrared modality embeddings. (c) The diagram of modality-aware enhancement loss, which contains two components, the modality-aware center loss and the modality-aware ID loss.](image-url)
method mainly contains three stages from the bottom to the top: input embedding, feature extraction, and multi-loss constraint.

In the stage of input embedding, as illustrated at the bottom of Fig. 2(a), here is an example when the input is a visible image (it is similar when inputting an infrared image). The input image \(x_{vis}^{(1)}\) is first split into a sequence of patches with the shape of \(R^{N \times C \times P \times P}\), where \(P\) denotes the size and \(N\) denotes the length of this sequence. Besides, following the strategy of [21], [40], we generate the patches with overlapping by stride \(S \leq P\) (soft split) to enhance the correlation among adjacent patches. The patches are reshaped to flattened embeddings with shape of \(R^{N \times (C \times P^2)}\). Then through the linear projection, they are converted to a sequence of token embeddings (\(R^{N \times D}\), \(D\) denotes the embedding dimension). Same as the work in [35], an extra learnable [class] token embedding is merged into the sequence to capture the global attention of the whole image. In the CMTR network, before being sent to the transformer, the token embedding sequence is fused with position embeddings and the designed modality embeddings to generate input embeddings (More details will be introduced in the next subsection).

In the feature extraction stage, the basic vision transformer (ViT) [35] model is used as the backbone extractor. By using multi-layer self-attention modules, the model can perceive more effective global features than CNN-based methods. As the top of Fig. 2(a) shows, corresponding to the location of [class] token, we can get the image vector for each \(x_{m}^{i}\) from output of the backbone. Let \(I\) and \(F\) denote the process of input embedding and feature extraction. The image vector can be extracted as follows:

\[
v_{i}^{m} = F(I(x_{i}^{m})), \quad m \in \{vis, ir\}.
\]

During the last stage of multi-loss constraint, these image vectors \(v_{i}^{m}\) obtained from the training batch pass through batch normalization (BN) layer and fully connected (FC) layer in sequence. After these different layers, the CMTR method calculates multiple kinds of losses, the identity (ID) loss, weighted regularization triplet (WRT) loss, and our proposed modality-aware enhancement (MAE) loss, which jointly constrain the distribution of extracted vectors to generate more discriminative ID embeddings that are invariant to the visible and infrared modalities.

### B. Visible-Infrared Modality Embeddings

From our observations, perception of the modality characteristics can help to generate modality-invariant features. However, this key is ignored by many existing methods. To achieve this, we introduce the modality embeddings (ME) into our CMTR model, which directly aims to learn and capture inherent information and characteristics of each modality.

Inspired by the idea of the position embeddings in plain transformer [23] or the segmentation embeddings in BERT [33] that can learn positional information or segmented information, our modality embeddings are introduced in a similar way with the different purpose to encode modality information. This design can be naturally integrated into the transformer framework. CNN-based models do not have this advantage. As shown in Fig. 4, the input embeddings of our CMTR network are obtained with three components, i.e., token embeddings, position embeddings, and modality embeddings. The first two are consistent with previous methods. The token embeddings are generated from image patches, and the position embeddings are defined with learnable parameters. For modality embeddings, images in each modality share the same embeddings with all patches. Specifically, let \(\{x_{i, p1}^{m1}, x_{i, p2}^{m1}, x_{i, p3}^{m1}, \ldots, x_{i, pN}^{m1}\}\) denote the patch sequence of image \(x_{i}^{m}\). The sequence \(\{e_{pos}^{p1}, e_{pos}^{p2}, e_{pos}^{p3}, \ldots, e_{pos}^{pN}\}\) denotes position embeddings. The stage of input embedding \(I(x_{i}^{m})\) can be formulated as follows:

\[
I(x_{i}^{m}) = \mathcal{L}P(\{x_{i, p1}^{m1}, x_{i, p2}^{m1}, x_{i, p3}^{m1}, \ldots, x_{i, pN}^{m1}\}) + \{e_{pos}^{p1}, e_{pos}^{p2}, e_{pos}^{p3}, \ldots, e_{pos}^{pN}\}
\]

\[
+ \left\{e_{vis}^{vis}, e_{vis}^{vis}, \ldots, e_{vis}^{vis}\right\}, \quad \text{if } m \text{ is } vis.
\]

\[
+ \left\{e_{ir}^{ir}, e_{ir}^{ir}, e_{ir}^{ir}, \ldots, e_{ir}^{ir}\right\}, \quad \text{if } m \text{ is } ir.
\]

where the \(\mathcal{L}P\) denotes the linear projection in Fig. 2(a), converting information of patches into token embeddings. Specifically, as shown in Fig. 3, we first sample patches sequentially from the image, and each patch is flattened into a 1D vector. Following previous work [35], the linear projection is implemented with the fully connection (FC) layer and normalization (Norm) layer. We pass the patch vectors through these layers to get the token embeddings. The \(e_{vis}\) and \(e_{ir}\) denote the visible modality embedding and the infrared modality embeddings respectively.
As shown in (2), these different types of embeddings are fused together in an additive manner. The position embeddings \( e^{pos} \) vary among patches, while the modality embeddings \( e^m (m \in \{vis, ir\}) \) vary between image modalities, perceiving and encoding different types of information.

C. Modality-Aware Enhancement Loss

The above-mentioned approach of how to use the modality embeddings makes them capture modality characteristics semantically, but the constraint of this way is relatively weak. To further enhance their ability to learn the modality information, and let the learned ME assist in generating more effective modality-invariant embeddings, we propose the modality-aware enhancement (MAE) loss.

As shown at the top of Fig. 2, the MAE loss acts on the extracted features after batch normalization (BN), which are used as the matching embeddings during testing. We let \( f^m = BN(v^m) \) denote the extracted features. The MAE loss consists of two parts (Fig. 2(c)): the modality-aware center loss and the modality-aware ID loss, which are designed to pull intra-class features and push inter-class features based on the modality embeddings. Our MAE loss \( L_{\text{MAE}} \) is calculated by adding them as follows:

\[
L_{\text{MAE}} = L_{\text{MAC}} + L_{\text{MAID}},
\]

(3)

where \( L_{\text{MAC}} \) denotes the modality-aware center loss, and \( L_{\text{MAID}} \) denotes the modality-aware ID loss.

For the definition of \( L_{\text{MAC}} \), it focuses on reducing the gap between different modalities under the same identity, and utilizes the learned knowledge from ME to narrow the intra-class feature distance. During training, we randomly sample \( Q \) identities’ images in a batch, and each identity contains \( K \) images, including \( K/2 \) visible images and \( K/2 \) infrared images. Specifically, the \( L_{\text{MAC}} \) can be formulated by:

\[
L_{\text{MAC}} = \sum_{q=1}^{Q} \sum_{k=1}^{K} \log (1 + \exp^{D(f^m_{q,k} - \phi_m(e^m), f^m_{q,c}))},
\]

\[
f^m_{q,c} = \frac{1}{K} \sum_{k=1}^{K} (f^m_{q,k} - \phi_m(e^m)), \quad m \in \{vis, ir\},
\]

(4)

where \( f^m_{q,k} \) denotes the extracted feature from the \( q \)-th identity’s \( k \)-th image with modality \( m \). \( \phi_m(\cdot) \) represents the mapping function to mine the knowledge from the modality embeddings \( e^m \). Specifically, we use the full connection layer to implement this function, and it converts the learned ME to a new feature embedding with the same dimension, i.e., \( \phi_m(\cdot) = FC_{D \rightarrow D}(\cdot) \). In this formula, we let the \( f^m_{q,k} \) subtract the corresponding \( \phi_m(e^m) \) directly to remove the modality-specific information and filter out modality-invariant features. The \( f^m_{q,c} \) denotes center feature vector of the \( q \)-th identity, which is the mean value of the image features after the modality removal. The \( L_{\text{MAC}} \) pulls the distance between id’s image features and its center feature vector, and we use the cosine distance \( D(\cdot, \cdot) \) to measure their difference. Besides, \( L_{\text{MAC}} \) use the soft-margin constraint to avoid setting hyperparameter of the hard margin in traditional distance loss functions [5], [41]. Through the constraint of modality-aware center loss \( L_{\text{MAC}} \), our method extracts more compact cross-modality features for each identity.

The modality-aware ID loss \( L_{\text{MAID}} \) aims at learning discriminative features among different identities, which is also based on the learned information from ME, designed to push the distance between ids’ image features. The equation of \( L_{\text{MAID}} \) can be formulated as follows:

\[
L_{\text{MAID}} = \sum_{q=1}^{Q} \sum_{k=1}^{K} \text{CrossEntropy}(p^m_{q,k}, t^m_{q,k}),
\]

\[
p^m_{q,k} = \text{Softmax}(FC_{id}(f^m_{q,k} - \phi_m(e^m))),
\]

(5)

where \( t^m_{q,k} \) denotes the one-hot target label for \( q \)-th identity. The predicted label \( p^m_{q,k} \) is calculated from image features \( f^m_{q,k} - \phi_m(e^m) \) with the same modality removal process as \( L_{\text{MAC}} \). We use the auxiliary FC layer \( FC_{id} \) to generate logits for classification, and the \( p^m_{q,k} \) is obtained through the Softmax operation on logits. The \( L_{\text{MAID}} \) calculates \( \text{CrossEntropy}(\cdot, \cdot) \) between predictions and targets, attempting to classify different identities’ input images. With the constraint of modality-aware ID loss \( L_{\text{MAID}} \), the features extracted by the model are given stronger distinguishing ability to achieve more accurate matching.

By optimizing the modality-aware enhancement loss \( L_{\text{MAE}} \), firstly, the network can utilize the modality removal process to enforce ME to mine more useful modality-specific characteristics, which is a more direct way to enhance the ME’s representation. Secondly, the ME-based loss functions can adjust the distribution of feature embeddings to be more discriminative for the image retrieval and less affected by the heterogeneous cross-modality gap.

D. Overall Objective Function

As shown at the top of Fig. 2(a), our CMTR network is constrained by three kinds of losses, and these constraints are jointly optimized. The ID loss \( L_{\text{ID}} \) and the WRT loss \( L_{\text{WRT}} \) are loss constraints frequently used in the field of ReID [5], [7]. The ID loss transforms the retrieval task of ReID into an image classification task, and it regards the person ID as the category of the corresponding images. Assume that the \( y_i \) represents the one-hot label of each image \( x_i \). By introducing the classifier (implemented with the FC layer), we can get the predicted logits \( p_i \). The \( L_{\text{ID}} \) is defined as follows:

\[
L_{\text{ID}} = -\frac{1}{B} \sum_{i=1}^{B} y_i \log p_i
\]

(6)

Note that \( B \) counts the number of images in a batch. In addition to the ID loss, the WRT loss proposed in [7] is an improved
version of Triplet loss as follows:

\[
\mathcal{L}_{\text{WRT}}(f_i, f_{i+}, f_{i-}) = \log(1 + \exp(w_i^+ d_{i,i+} - w_i^- d_{i,i-})),
\]

\[
w_i^+ = \frac{\exp(d_{i,i+})}{\sum_{d \in D_+} \exp(d)}, \quad w_i^- = \frac{\exp(-d_{i,i-})}{\sum_{d \in D_-} \exp(-d)}.
\]

(7)

The \( f_i \) denotes the extracted embedding of the anchor sample, and the \( f_{i+} \) and \( f_{i-} \) denote its positive and negative samples. The \( d_{i,i+} \) is the distance between positive sample pairs, while the \( d_{i,i-} \) represents the distance between negative sample pairs. The \( D_+ \) and \( D_- \) denote the set of positive pair distance and the set of negative pair distance respectively. By adding this weighted regularization, the WRT loss frees the user from setting a fixed margin, which is more flexible and adaptable.

The overall objective function is calculated as follows:

\[
\mathcal{L}_{\text{overall}} = \mathcal{L}_{\text{ID}} + \mathcal{L}_{\text{WRT}} + \lambda \cdot \mathcal{L}_{\text{MAE}},
\]

we adopt the \( \mathcal{L}_{\text{ID}} \) and \( \mathcal{L}_{\text{WRT}} \) in the locations similar to previous work [7]. As for our modality-aware enhancement loss (\( \mathcal{L}_{\text{MAE}} \)), it is added to the formula with weight, and the hyperparameter \( \lambda \) can control the proportion of this loss.

IV. EXPERIMENTS

A. Experimental Settings

1) Datasets: Our experiments are performed on two public datasets, SYSU-MM01 [8] and RegDB [9], which are the standard benchmarks and commonly used by existing methods in the VI-ReID task.

SYSU-MM01 is currently the largest and most challenging visible-infrared cross-modality person ReID dataset. It totally consists of 29,033 visible images and 15,712 infrared images of 491 identities, which are collected by 4 visible cameras and 2 infrared ones from indoors and outdoors. The training set contains 22,258 visible images and 11,909 infrared images of 395 identities, and the testing set contains 96 identities’ images. Following [8], 3,803 infrared images of these testing identities are used to form the query set. Corresponding to the single-shot or multi-shot setting, 1 or 10 images of each identity under each visible camera are randomly selected to form the gallery set. Besides, the dataset has two testing modes: the all-search mode is evaluated with the indoor and outdoor images, while the indoor-search mode is evaluated with only the indoor images.

RegDB is collected by dual aligned visible and far-infrared cameras, including 412 identities’ images. Each identity has 10 visible images and 10 far-infrared images. Consistent with previous methods [13], [20], [32], [42], [43], we equally divide the dataset into two parts as the training set and testing set by random selection. Each set contains 2,060 visible images and 2,060 far-infrared images. In testing set, when performing the Visible to Thermal/Thermal to Visible mode, all the 2,060 visible/far-infrared images are used as query set, and all the 2,060 far-infrared/visible images are used as gallery set to perform cross-modality retrieval.

2) Evaluation Metrics: We evaluate methods with two widely used metrics of this task: the Cumulative Matching Characteristics (CMC) curve and the mean Average Precision (mAP). The CMC is denoted as Rank-k (Rk for short in the table) to measure the correct rate in the k-nearest matching results, and we calculate R1, R10, R20 in the experiments. Besides, according to [7], the mean inverse negative penalty (mINP) is also used as an auxiliary metric in our ablation study. Following the official test protocol [8], we conduct repeated random selections of gallery on SYSU-MM01 for 10 times to get more stable average result. Similarly, experiments on RegDB average results from 10 times repeated random partition of training and testing sets.

3) Implementation Details: Our proposed method is implemented with the PyTorch deep learning framework. For the transformer backbone, we use the ViT-Base [35] model with pretrained weights on the ImageNet dataset. Before entering the network, the visible and infrared images are resized to \( 3 \times 256 \times 128 \) (\( C \times H \times W \)). Following most previous methods [7], [20], [32], [44], [45], we repeat the single channel of the infrared image three times to make it contain three channels. The patches are generated with 16 \times 16 size following [35], and the stride \( S \) is set to 8 for half overlap. During training, we adopt the common data augmentation strategies: the random horizontal flip and random erasing [46]. In a mini-batch, we randomly sample 4 identities’ images and each identity has 8 visible images and 8 infrared images (i.e., \( Q = 4, K = 16 \)). We use the AdamW [47] optimizer with weight decay set to 0.0005. The base learning rate is initialized at 0.001 and we make the learning rate of all pretrained layers to be 0.1 times of the base learning rate. Specifically, the whole model is totally trained for 80 epochs with decay by 0.1 at epoch 20 and 50 on the SYSU-MM01 dataset. For the RegDB dataset, we train the model for 50 epochs with decay at epoch 30. The trade-off hyperparameter \( \lambda \) in objective function is empirically set to 6. During testing, all the query and gallery images are sent into the model to extract embeddings with cosine distance to rank retrieval results.

B. Ablation Experiments

The ablation experiments are designed to evaluate the influence of our proposed modality embeddings (ME) and modality-aware enhancement (MAE) loss. We conduct these experiments on the SYSU-MM01 dataset with the difficult single-shot setting of all-search mode by default.

1) Effectiveness of the Designed Components: Our CMTR method consists of three main components, the vision transformer (ViT) network, the designed modality embeddings (ME), and the proposed modality-aware enhancement (MAE) loss. The MAE loss can be further divided into two components: the MAC loss and the MAID loss. During ablation experiments, we quantitatively verify the effects of these components in turn.

As shown in Table I, the “BASE” represents the baseline ViT network trained with the common \( \mathcal{L}_{\text{ID}} \) and \( \mathcal{L}_{\text{WRT}} \) (Index-1), and it achieves 57.58% in Rank-1 accuracy and 55.21% in mAP. By introducing modality embeddings (ME) into this network, the model (Index-2) achieves 3.12% Rank-1 and 2.96%
5(d) shows visible and insets to and is higher than that of Fig. S×P.

To better understand (Index-3 & 4). Compared with =1 6 EII (b)

and represents the length of and method. (a) A visible image. (b)

\( S = 6 \) (c) and 6(a) captures more identity’s profile and texture

method to generate attention maps

\( C \) (d) \( S \) method. (a) Feature distribution of BASE method. (b) Feature distribution of BASE+MAC method. Different colors represent different IDs, and (e) CAM of BASE+ME on the visible image. (f) CAM of BASE+ME on the infrared image.

\( mAP \) improvements. For the study of modality-aware enhancement (MAE) loss, we verify the effects of its components \( \mathcal{L}_{\text{MAC}} \) and \( \mathcal{L}_{\text{MAID}} \) respectively (Index-3 & 4). Compared with “BASE”+“ME,” the MAC loss brings \( +1.36\% \) Rank-1 and +2.56\% mAP and the MAID loss brings \( +1.48\% \) Rank-1 and +2.79\% mAP. What’s more, through jointly optimizing the MAC loss and the MAID loss, i.e., the MAE loss, our method’s performance is further improved \( (+4.75\% \text{ Rank-1 and } +4.73\% \text{ mAP}) \), which presents the complementarity between these two kinds of losses. These experimental results demonstrate the effectiveness of proposed components.

2) Analysis of the Modality Embeddings: To better understand the influence of modality embeddings, we visualized attention maps of the BASE and BASE+ME models. As illustrated in Fig. 6, the Fig. 6(a) and (b) show visible and infrared images of an identity in the SYSU-MM01 dataset. We use the Grad-CAM [49] method to generate attention maps (Fig. 6(c)–(f)) on them. Compared with Fig. 6(c), we can notice that Fig. 6(e) captures more identity’s profile and texture information (such as the pattern on this man’s shirt), which are modality-invariant. For the infrared image (Fig. 6(b)), although it is difficult for human eyes to distinguish the pattern on the shirt, the BASE+ME model can still perceive the information of this position in Fig. 6(f). Besides, the consistency of Fig. 6(e) and (f) is higher than that of Fig. 6(c) and (d), which demonstrates the auxiliary ability of ME to reduce the impact of modality discrepancy.

3) Analysis of the Proposed MAE Loss: In order to understand the effect of MAE loss intuitively, we show feature distributions of the final extracted embeddings with the t-SNE [48] method. As illustrated in Fig. 5, we randomly select visible images and infrared images of some identities, and get their embeddings through our network. From Fig. 5(a) to (b), we can notice that the feature embeddings with the same ID (in the circles) are more compact, i.e., the introduction of MAC loss helps reduce the intra-class differences. By comparing Fig. 5(a) and (c), the distance between circles with different IDs is increased. Thus, the MAID loss helps expand the inter-class differences. It is worth noting that when only the MAC loss is used, the embeddings with different IDs are gathered into some large clusters (divided by the red line in Fig. 5(b)) due to the lack of inter-class constraints. When using only the MAID loss (Fig. 5(c)), intra-class consistency is not well constrained. Finally, as shown in Fig. 5(d), when using the MAE (MAC&MAID) loss, embeddings have better distribution with compact intra-class distances and uniformly larger inter-class distances. The MAE loss helps to generate more effective embeddings for retrieval.

C. Module Settings and Discussion

1) Comparison on Stride of Patch Generation: In the stage of input embedding, we generate the patch sequences in an overlapping manner, as suggested in the previous research [21], [40]. Specifically, consistent with the input of the base ViT model [35], the patch size used in our method is \( 16 \times 16 \) (\( P = 16 \)), and we conduct experiments with different strides in our proposed CMTR network.

In Table II, we conduct experiments with stride \( S \) set to \{16, 14, 12, 10, 8, 6, 4\}. Note that \( N \) represents the length of patch sequence. When \( S \) is set to 16 (\( S = P \)), there is no overlap between patches. From Table II, we can notice that this method (Index-1) achieves relatively low performance with

| Table I: Ablation Experiments of Designed Components on the SYSU-MM01 Dataset With All-Search Single-Shot Mode |
|---|
| Index | BASE | ME | MAE | MAC | MAID | R1 | R10 | R20 | mAP | mNP |
| 1 | ✓ | ✗ | ✗ | ✔ | ✔ | 57.58 | 89.31 | 95.15 | 55.21 | 42.15 |
| 2 | ✓ | ✗ | ✗ | ✔ | ✗ | 60.70 | 91.11 | 96.28 | 58.17 | 44.94 |
| 3 | ✓ | ✗ | ✗ | ✔ | ✗ | 62.06 | 92.38 | 96.82 | 60.73 | 48.97 |
| 4 | ✓ | ✗ | ✗ | ✔ | ✗ | 62.18 | 91.80 | 96.60 | 60.96 | 49.46 |
| 5 | ✓ | ✓ | ✓ | ✓ | ✓ | 65.45 | 94.47 | 98.16 | 62.90 | 49.64 |
59.83% Rank-1 and 57.68% mAP. As the stride $S$ decreases, the overlapping area between adjacent patches will gradually increase. We can notice that the model performs better when using a relatively small stride. Specifically, compared with the Index-1 model, other models (Index-2,3,4,5,6,7) gain +1.05%, +1.79%, +3.33%, +5.62%, +5.65%, +5.49% Rank-1 and +0.56%, +1.16%, +3.34%, +5.22%, +5.16%, +5.08% mAP, respectively. The results demonstrate the effectiveness of the overlapping patch strategy in the VI-ReID task.

With a smaller stride $S$, the model can generate a longer patch sequence with larger $N$ (shown in the second and third columns of the Table II). Through overlapping patches and long sequences, the multi-head self-attention modules of the transformer model can learn richer and more powerful features. In practice, longer sequences will lead to more consumption of computing and memory resources. From Table II, we can observe that the performance gain brought by too small $S$ (i.e. $S = 6, 4$) is negligible. We think that the reason for this phenomenon is that when $S$ is too small, the overlapping area of adjacent patches is relatively large, resulting in the high similarity between patches and excessive redundant information. In the final settings of our CMTR method, we adopt the stride set to 8, so that half of the patches overlap each other, ensuring better experimental performance, while providing a simple and effective solution for VI-ReID.

2) Study of the Parameter $\lambda$ in the Objective Function: In the method of multi-loss optimization, weights are usually added as hyperparameters to the sub losses to adjust their magnitude. Generally, the sub loss that uses a larger weight will dominate the optimization process, and the model will be more inclined to optimize this loss first. Here, in the definition of our $L_{\text{overall}}$ (8), the weight $\lambda$ is used as a hyperparameter to balance the proportion between losses. During the ablation experiment, we evaluate the effect of this hyperparameter $\lambda$.

As illustrated in Fig. 7, the hyperparameter $\lambda$ is set to different values $\{0, 1, 2, 3, 4, 5, 6, 7\}$. With $\lambda$ set to zero, the network is optimized without (w/o) the proposed modality-aware enhancement loss, and is only optimized by the common $L_{\text{ID}}$ and $L_{\text{WT}}$. We plot its Rank-1,10,20 and mAP line as a benchmark. By introducing the modality-aware enhancement loss with hyperparameter $\lambda$ greater than zero, we can observe a significant performance improvement on both Rank-1 and mAP, and it is relatively stable with different $\lambda$ values. Thus, the performance of our method is insensitive to these different values, indicating that the feature extraction ability of our method does not depend too much on the specific order of multi-loss optimization. The experimental results demonstrate the stability and robustness of the proposed MAE loss.

3) Discussion on the Design of MAE Loss: During the experiment, we compare different implementation schemes of modality-aware enhancement loss, and explore the influence of several designs for the mapping function $\phi_m(\cdot)$, distance measurement method $D(\cdot, \cdot)$ in (4) and (5) and form of constraint in the MAC loss.

The $\phi_m(\cdot)$ is the function to mine the information from modality embeddings, which is necessary to utilize the inherent characteristics of each modality. As shown in Table III, we compare two implementations, the identity mapping and fully connection. As for the implementation of identity mapping mode, we directly adopt the learned ME without any further processing and operation. With the plain identity mapping mode, the modality removal operation uses the extracted features to directly subtract the ME of the corresponding modality, which is a hard way to eliminate the learned characteristics of modalities. Different
from this way, the introduction of fully connection layer can fur-
ther mine helpful information and knowledge in the ME, which
can support modality removal in a soft way. The experimental
results in the Table III show the positive effect of post-mining
mechanism in fully connection mode with improvements on the
Rank accuracy and mAP.

As for \( D(\cdot, \cdot) \) in the definition of loss function (4), we compare
various different distance calculation methods, including \( L^1, \)
\( L^2, \) Smooth-\( L^1 \) and cosine distance. Our designed MAE loss
acts on the features extracted by the model after the BN layer, and
the features at this location are constrained by the common ID
loss at the same time. As discussed in previous work [5], the ID
loss mainly optimizes the cosine distance. Thus, our MAE loss
should also act as a constraint with distance of the same kind.
The middle part of Table III shows the experimental results with
different distance metrics. Among all these settings, the cosine
distance has a significant advantage, which is also consistent
with the analysis and conclusion of the research [5].

When defining the modality-aware center loss \( \mathcal{L}_{MAC} \), we com-
pare two kinds of constraints on the embeddings after modality
removal process. One is to calculate the center of the modality
set like the CMCC loss [20], and then pull the center of different
modalities under the same ID, the other is to directly restrict the
sample of each ID to be close to its center (4). The former acts on
the set level, and the latter is carried out from the sample level.
In the lower part of Table III, we show the experimental results
under these different settings. In our CMTR network, the way of
constraining samples is more effective, which also reflects the
positive effect of knowledge from the learned ME on eliminat-
ing the influence of modality. Thus, we use these settings in final
version of the CMTR method.

We also conduct extra experiments on the weight of MAC
loss and MAID loss. Specifically, we add a hyperparameter \( \beta \)
_to adjust the proportion of MAC loss and MAID loss. In this case,
\( \mathcal{L}_{MAE} = (\beta \cdot \mathcal{L}_{MAC} + (1 - \beta) \cdot \mathcal{L}_{MAID}) \cdot 2, \) and \( \beta \in (0, 1) \). We
multiply the loss by 2 just so that when \( \beta \) is 0.5, the weight of each
sub-loss is 1 to align with the original setting for comparison.
Then, we change the hyperparameter \( \beta \) with different values in
\{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}. Here, we provide cor-
responding illustrations to show the performance of our method
with these settings of hyperparameter \( \beta \). As shown in Fig. 8,
we can observe that the performance is better when the weights
of the two sub-losses are similar (the hyperparameter \( \beta \) is close
to 0.5), which verifies the effectiveness of balancing intra-class
compactness and inter-class separation. It is worth noting that
although the performance when \( \beta \) is set to 0.4 is slightly higher
than our original setting with the same weights, we still adopt
1:1 weights in our final method to reduce unnecessary hyperpa-
rameters for brevity.

4) Comparison of Loss Constraints: There exists some studies
that focus on adjusting the distribution of extracted em-
bodiments through loss constraints, such as the center loss [50],
Hetero-Center loss [31], and Cross-Modality Contrastive-Center
loss [20]. During experiments, we compare our proposed
modality-aware enhancement (MAE) loss with these losses in
our cross-modality transformer (CMTR) network to check and
discuss their performance differences.

As shown in Table IV, the “Baseline+ME” denotes the
baseline ViT network with modality embeddings (ME) opti-
mized by the common \( \mathcal{L}_{ID} \) and \( \mathcal{L}_{MRT} \) losses. We add Center
loss (“Center” for short), Hetero-Center loss (“HC”), Cross-
Modality Contrastive-Center Loss (“CMCC”) and the proposed
modality-aware enhancement (“MAE”) losses to this basic net-
work (Index-2,3,4,5 of Table IV). From quantitative evaluation
results, the center loss and HC loss bring little performance
improvement. The former brings +0.95% Rank-1 and +1.60% mAP,
the latter brings +1.03% Rank-1 and +1.61% mAP. The latest
CMCC loss brings +2.01% Rank-1 and +2.78% mAP. Compared with “Baseline+ME,” our MAE loss shows a promi-
nent effect (Index-5), and its Rank-1 and mAP results are greatly
improved.

The traditional loss of metric learning constraints directly act
on extracted embeddings. However, they do not consider the
effective mining and rational utilization of modality character-
istics and information, which limits their performance. In con-
trast, by using the learned modality embeddings, our designed
modality-aware enhancement loss overcomes the shortcomings
of existing losses and provides better guidance for the adjust-
ment of embedding distribution.

5) Visualization of the Semantic Embeddings: Our CMTR
model contains some learnable semantic embedding, including
the position embeddings ($e^{pos}$) and our designed modality embeddings ($e^{vis}$ & $e^{ir}$). In order to intuitively see what they look like, we adopt the similar visualization strategy of the previous research [51] and directly plot the heat maps of these semantic embeddings to show their values. As illustrated in Fig. 9, the position embeddings are shown as a matrix with the shape of $R^{N \times D}$ ($N = 465, D = 768$). They correspond to positions of all the patch tokens, and each position is represented by a $D$-dimensional embedding. The modality embeddings are represented as two vectors with the shape of $R^{D}$ ($D = 768$). Although they are all hidden high-level features inside the model and are difficult to be directly interpreted, we can still notice that the values learned in $e^{pos}$ vary according to different positions, and they encode different position information. The $e^{vis}$ and $e^{ir}$ appear to be quite different, and they are designed to perceive the characteristics of the two heterogeneous modalities.

D. Comparison With Existing VI-ReID Methods

In experiments, we compare our proposed CMTR model with the existing outstanding methods in the cross-modality VI-ReID field. We roughly divide the existing methods into two categories, (a) GAN-based methods: cmGAN [10], D$^2$RL [11], Hi-CMD [61], JSIA [12], AlignGAN [13], TS-GAN [14], GECNet [52], TSME [53] and (b) two/multi-stream methods without generative strategy: SDL [62], MSR [54], FMSP [63], AGW [7], DEF [56], AGT($\gamma$)+dDeS [15], X-Modality [30], CMAlign [58], DDAG [29], DML [59], MSO [20], VCD+VM-L [57], cm-SSFT [42], CM-NAS [45], SPOT [60]. Note that our CMTR method only uses the global feature extraction to obtain 768-dim embeddings, so some method designed with extra information or complicated local partition (such as the MPANet [32] with 14336-dim concatenated embeddings) are not listed here for fair comparison.

1) Comparisons on the RegDB Dataset: In Table V, we show the comparison results on the RegDB dataset. Our CMTR method shows great performance in all evaluation modes, which achieves 88.11% Rank-1 and 81.66% mAP under the Visible to Thermal (V to T) mode and 84.92% Rank-1 and 80.79% mAP under the Thermal to Visible (T to V) mode. Compared with the advanced CM-NAS [45] method that uses resource-consuming neural architecture search strategy, the CMTR method shows improvements on performance, demonstrating the effectiveness of our proposed method.

2) Comparisons on the SYSU-MM01 Dataset: As shown in Table VI, we also present the quantitative comparison on the
TABLE VI
COMPARISON WITH EXISTING VI-ReID METHODS ON THE SYSU-MM01 DATASET. WE ROUGHLY DIVIDE THE EXISTING METHODS INTO TWO CATEGORIES (FROM THE TOP TO THE BOTTOM) : GAN-BASED METHODS, TWO/MULTI-STREAM METHODS WITHOUT GENERATIVE STRATEGY

| Methods         | Venue         | All-Search | Indoor-Search |
|-----------------|---------------|------------|--------------|
|                 | R1  | R10 | R20 | mAP | R1  | R10 | R20 | mAP | R1  | R10 | R20 | mAP | R1  | R10 | R20 | mAP |
| cmGAN [10]      | JCAI 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 [11]      | CVPR 19       | 28.90  | 70.60 | 82.40 | 29.20 | -    | -    | -    | -    | -    | -    | -    | -    | -    | -    | -    |
| Hi-CMD [61]     | CVPR 20       | 34.94  | 77.58 | 35.94 | -    | -    | -    | -    | -    | -    | -    | -    | -    | -    | -    | -    |
| JSIA [12]       | 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 |
| AllgeGAN [13]   | 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 |
| TS-GAN [14]     | PRL. 21       | 49.80  | 87.30 | 93.80 | 47.40 | 56.10 | 90.20 | 96.30 | 38.50 | 50.40 | 90.80 | 96.80 | 63.10 | 59.30 | 91.20 | 97.80 | 50.20 |
| GECNet [52]     | TCSVT 22      | 53.37  | 89.86 | 95.66 | 51.83 | -    | -    | -    | -    | 60.60 | 94.29 | 98.10 | 62.89 | -    | -    | -    | -    |
| TSM [53]        | TCSVT 22      | 64.23  | 95.19 | 98.73 | 61.21 | 70.34 | 96.75 | 99.26 | 54.36 | 64.80 | 96.92 | 99.31 | 71.53 | 76.83 | 98.84 | 99.89 | 65.02 |

challenging SYSU-MM01 dataset, and we can get some inspiring observations in the table. From the top part of Table VI, we can notice that our CMTR method obviously outperforms the GAN-based methods, which reflects that the GAN-based methods do suffer from unstable mapping relationship between visible and infrared modalities with limited performance. This is also the reason why researchers recently turn to non generative methods. When compared with the two/multi-stream methods at the bottom part of the table, the performance of our CMTR method still exceeds the strong CM-NAS [45] method. Even compared with the all queries (AQ) version of cm-SSPT [42] that utilizes auxiliary set for testing and changes the formal test protocol, our method also shows improvement. It is worth noting that although the DFLN-ViT [22] method introduce some transformer blocks into the traditional CNN backbone, this design only increases the capacity of the model and does not solve the modal gap. In summary, different from these methods, our method pays more attention to the effective mining of modal invariance features and shows superiority in performance.

V. CONCLUSION

In this paper, we propose a novel method for the VI-ReID task, the cross-modality transformer (CMTR) network. By introducing the modality embeddings (ME), the model can directly perceive characteristics of each modality. Furthermore, we design the modality-aware enhancement (MAE) Loss, which can enhance the learning ability of modality embeddings and help to generate better discriminative modality-invariant embeddings. The method shows good experimental performance against existing methods on the SYSU-MM01 and RegDB datasets. For the potential directions of future research, considering that the backbone networks of the Transformer have developed rapidly in recent years, new structures based on them can be explored...
to expand the framework of our method and further solve the VI-ReID task. Besides, lots of research work has verified the advantages of the Transformer model in processing multi-modal data. Therefore, in addition to the visible-infrared ReID task studied in our paper, researchers can explore the ability of this method in other related tasks, such as text-image or sketch-image person ReID. We believe that the proposed strategy in the CMTR method can be enlightening for other cross-modality tasks.

**REFERENCES**

[1] L. Zheng, Y. Yang, and A. G. Hauptmann, “Person re-identification: Past, present and future,” 2016, arXiv:1610.02984.

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

[3] Z. Zeng et al., “Illmination-adaptive person re-identification,” *IEEE Trans. Multimedia*, vol. 22, no. 12, pp. 3064–3074, Dec. 2020.

[4] C. Yan et al., “Beyond triplet loss: Person re-identification with fine-grained difference-aware pairwise loss,” *IEEE Trans. Multimedia*, vol. 24, pp. 1665–1677, 2022.

[5] H. Luo, Y. Gu, X. Liao, S. Lai, and W. Jiang, “Bag of tricks and a strong baseline for deep person re-identification,” in *Proc. IEEE Comput. Vis. Pattern Recognit.* 2019, pp. 1487–1495.

[6] G. Wang, Y. Yuan, X. Chen, J. Li, and X. Zhou, “Learning discriminative features with multiple granularities for person re-identification,” in *Proc. 26th ACM Int. Conf. Multimedia*, 2018, pp. 274–282.

[7] M. Ye et al., “Deep learning for person re-identification: A survey and outlook,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 6, pp. 2872–2893, Jun. 2022.

[8] A. Wu, W. Zheng, H. Yu, S. Gong, and J. Lai, “RGB-infrared cross-modality person re-identification,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 5390–5399.

[9] D. T. Nguyen, H. G. Hong, K. 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, 2017, Art. no. 605.

[10] P. Dai, R. J., H. Wang, Q. Wu, and Y. Huang, “Cross-modality person re-identification with generative adversarial training,” in *Proc. Int. Joint Conf. Artif. Intell.*, 2018, pp. 677–683.

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

[12] G. Wang et al., “Cross-modality paired-images generation for RGB-infrared person re-identification,” in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 1214–1215.

[13] G. Wang et al., “RGB-infrared cross-modality person re-identification via joint pixel and feature alignment,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 3622–3631.

[14] Z. Zhang, S. Jiang, C. Huang, Y. Li, and R. Y. D. Xu, “RGB-BR cross-modality person ReID based on teacher-student GAN model,” *Pattern Recognit. Lett.*, vol. 150, pp. 155–161, 2021.

[15] Y. Huang et al., “Alleviating modality bias training for infrared-visible person re-identification,” *IEEE Trans. Multimedia*, vol. 24, pp. 1570–1582, 2021.

[16] M. Ye, Z. Wang, X. Lan, and P. C. Yuen, “Visible thermal person re-identification via dual-constrained top-ranking,” in *Proc. Int. Joint Conf. Artif. Intell.*, 2018, pp. 1092–1099.

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

[18] H. Liu, J. Cheng, W. Wang, Y. Su, and H. Bai, “Enhancing the discriminative feature learning for visible-thermal cross-modality person re-identification,” *Neurocomputing*, vol. 398, pp. 11–19, 2020.

[19] 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.

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

[21] S. He et al., “TransReID: Transformer-based object re-identification,” in *Proc. Int. Conf. Comput. Vis.*, 2021, pp. 14993–15002.
[49] R. R. Selvaraju et al., “Grad-CAM: Visual explanations from deep networks via gradient-based localization,” in Proc. IEEE Int. Conf. Comput. Vis., 2017, pp. 618–626.

[50] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, “A discriminative feature learning approach for deep face recognition,” in Proc. Eur. Conf. Comput. Vis., 2016, pp. 991–995.

[51] X. Liu, H. Yu, I. S. Dhillon, and C. Hsieh, “Learning to encode position for transformer with continuous dynamical model,” in Proc. Int. Conf. Mach. Learn., 2020, vol. 119, pp. 6327–6336.

[52] X. Zhong et al., “Grayscale enhancement colorization network for visible-infrared person re-identification,” IEEE Trans. Circuits Syst. Video Technol., vol. 32, no. 3, pp. 1148–1158, Mar. 2022.

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

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

[55] Y. Ling et al., “Class-aware modality mix and center-guided metric learning for visible-thermal person re-identification,” in Proc. 28th ACM Int. Conf. Multimedia, 2020, pp. 889–897.

[56] Y. Hao, N. Wang, X. Gao, J. Li, and X. Wang, “Dual-alignment feature embedding for cross-modality person re-identification,” in Proc. 27th ACM Int. Conf. Multimedia, 2019, pp. 57–65.

[57] X. Tian et al., “Farwell to mutual information: Variational distillation for cross-modal person re-identification,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2021, pp. 1522–1531.

[58] H. Park, S. Lee, J. Lee, and B. Ham, “Learning by aligning: Visible-infrared person re-identification using cross-modal correspondences,” in Proc. IEEE Int. Conf. Comput. Vis., 2021, pp. 12046–12055.

[59] D. Zhang et al., “Dual mutual learning for cross-modality person re-identification,” IEEE Trans. Circuits Syst. Video Technol., vol. 32, no. 8, pp. 5361–5373, Aug. 2022.

[60] C. Chen et al., “Structure-aware positional transformer for visible-infrared person re-identification,” IEEE Trans. Image Process., vol. 31, pp. 2352–2364, 2022.

[61] S. Choi, S. Lee, Y. Kim, T. Kim, and C. Kim, “Hi-CMD: Hierarchical cross-modality disentanglement for visible-infrared person re-identification,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2020, pp. 10254–10263.

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

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