Image-Specific Information Suppression and Implicit Local Alignment for Text-Based Person Search

Shuanglin Yan, Hao Tang, Liyan Zhang, and Jinhui Tang, Senior Member, IEEE

Abstract—Text-based person search (TBPS) is a challenging task that aims to search pedestrian images with the same identity from an image gallery given a query text. In recent years, TBPS has made remarkable progress, and state-of-the-art (SOTA) methods achieve superior performance by learning local fine-grained correspondence between images and texts. However, most existing methods rely on explicitly generated local parts to model fine-grained correspondence between modalities, which is unreliable due to the lack of contextual information or the potential introduction of noise. Moreover, the existing methods seldom consider the information inequality problem between modalities caused by image-specific information. To address these limitations, we propose an efficient joint multilevel alignment network (MANet) for TBPS, which can learn aligned image/text feature representations between modalities at multiple levels, and realize fast and effective person search. Specifically, we first design an image-specific information suppression (ISS) module, which suppresses image background and environmental factors by relation-guided localization (RGL) and channel attention filtration (CAF), respectively. This module effectively alleviates the information inequality problem and realizes the alignment of information volume between images and texts. Second, we propose an implicit local alignment (ILA) module to adaptively aggregate all pixel/word features of image/text to a set of modality-shared semantic topic centers and implicitly learn the local fine-grained correspondence between modalities without additional supervision and cross-modal interactions. Also, a global alignment (GA) is introduced as a supplement to the local perspective. The cooperation of global and local alignment modules enables better semantic alignment between modalities. Extensive experiments on multiple databases demonstrate the effectiveness and superiority of our MANet.

Index Terms—Image-specific information, implicit local alignment (ILA), information inequality, text-based person search (TBPS).

I. INTRODUCTION

PERSON reidentification (Re-ID) is a hot research topic in computer vision that aims to retrieve a given query from a gallery set collected across cameras. Re-ID can be broadly categorized into three subtasks based on the query data type: image-based Re-ID [1], [2], [3], [4], [5], [6], video-based Re-ID [7], [8], [9], [10], and text-based person search (TBPS) [11]. Over the past decade, significant progress has been made in image-based and video-based Re-ID. Compared with the above two subtasks, TBPS can search the target object with a simpler and more accessible text description. As a result of its practicality, TBPS has gained increasing attention in recent years.

TBPS is challenging as it involves processing information from two heterogeneous modalities, integrating the difficulties of image-based Re-ID and image–text matching. The key to TBPS lies in cross-modal alignment. Early methods [12], [13], [14], [15] aligned images and texts into a joint embedding space by designing loss functions or network models. However, these methods focus on aligning global image and text representations while overlooking local fine-grained alignment. Consequently, some later works [16], [17], [18], [19], [20] have been devoted to modeling local fine-grained correspondence between modalities. The general process involves generating local parts of images/texts, followed by establishing local correspondence through interactions between them, as shown in Fig. 1(c). The establishment of fine-grained correspondence depends on the quality of local parts. However, most existing methods directly split images/texts [18], [19], [20] or introduce external models [16], [17] to generate local parts explicitly. As shown in Fig. 1, complete and meaningful local parts are irregular, and this hard split way may destroy contextual information, leading to ambiguity when modeling fine-grained correspondence, while the introduction of external models may introduce noise and increase computational costs due to the domain gap between external data and TBPS data. Consequently, such explicitly generated local parts are unreliable for modeling local fine-grained correspondence between modalities. Also, the direct interactions between modalities may affect each other, which is not conducive to modeling fine-grained correspondence.

Furthermore, most existing methods directly encode images and texts in the same way to obtain image and text features,
between modalities, an implicit local alignment (ILA) and a
information and achieve the alignment of information volume
module to alleviate the influence of image-specific
TBPS. It comprises an image-specific information suppression
model that suppresses image-specific information and implic-
variations.

Our observation suggests that this problem is mainly caused
first large-scale person description dataset, the Chinese Uni-
work guided by global relations assigns higher weights to
nonpedestrian regions, thereby suppressing background noise.
In addition, we introduce a channel attention filtration (CAF)
module that performs instance normalization (IN) on image
features to filter out environmental factor noise followed by
recalling identity-related information from the filtered infor-
mation through channel attention, avoiding the loss of identity
information during filtering.

The ILA module implicitly learns locally aligned image and
text features by introducing a set of learnable modality-shared
semantic topic centers, as shown in Fig. 1(d). The relations
between each pixel (word) feature of the image (text) and
these topic centers are used as soft-assign weights to aggregate
all pixel (word) features of the image (text) to each semantic
center. This generates a set of local image (text) features for
a set of centers, which are locally aligned, since the
semantic centers are modality-shared. The introduction of the
modality-shared centers avoids the direct interactions between
modalities. Each locally aligned feature is generated based on
the complete context information of the image/text, thus avoid-
ing ambiguity and noise, and can better model fine-grained
correspondence. The GA module aggregates image salient
information spatially and text salient information temporally,
maximizing their similarity in joint embedding space for GA.
The above modules are integrated and jointly optimized in
an end-to-end manner. During inference, global and local
image/text features extracted by MANet are combined for
similarity measurement.

Our main contributions can be summarized as follows.
First, we design an ISS module (including RGL and CAF) to
address the information inequality between modalities caused
by image background and environmental factors. To the best
of our knowledge, we are the first to explicitly clarify the
presence of environmental factors for TBPS. Second, an ILA
module is proposed to implicitly learn locally aligned image
and text features by introducing a set of learnable modality-shared
semantic centers without additional supervision and cross-modal interactions.

The above observations motivate us to develop a lightweight
model that suppresses image-specific information and implic-
ity learns locally aligned image and text features. To this end,
we propose a joint multilevel alignment network (MANet) for
TBPS. It comprises an image-specific information suppression
(ISs) module to alleviate the influence of image-specific
information and achieve the alignment of information volume
between modalities, an implicit local alignment (ILA) and a
global alignment (GA) module, to implicitly learn semantically
locally and globally aligned features.

The ISS module includes a relation-guided localization
(RGL) submodule, which calculates the attention weight for
each pixel of the image feature map using global relations
between the pixel and all other pixels as a guide. With
the supervision of the cross-modal alignment loss, our net-
world guided by global relations assigns higher weights to
modality-shared human body regions and lower weights to
nonpedestrian regions, thereby suppressing background noise.
without considering the information inequality problem caused
by modality-specific information in images. Specifically, the
text description only includes the appearance information of a
pedestrian, while the image captured by the surveillance cam-
ner includes additional surrounding environment information
not described in the text. This environmental information acts
as noise and increases the difficulty of cross-modal alignment,
leading to information inequality between images and texts.
Our observation suggests that this problem is mainly caused
by two types of image-specific information: first, image back-
ground, as shown in Fig. 1(a). Cross-modal alignment only
focuses on the common information between modalities, while
image background that is not described in the text undoubtedly
enlarges the gap between modalities. Second, environmental
factors, such as illumination, weather, and so on, as shown in
Fig. 1(b), which not only act as noise but can even destroy
the semantics of the image itself, leading to large intraclass
variations.

The above observations motivate us to develop a lightweight
model that suppresses image-specific information and implicit-
ly learns locally aligned image and text features. To this end,
we propose a joint multilevel alignment network (MANet) for
TBPS. It comprises an image-specific information suppression
(ISs) module to alleviate the influence of image-specific
information and achieve the alignment of information volume
between modalities, an implicit local alignment (ILA) and a

II. RELATED WORK

A. Text-Based Person Search

TBPS was first proposed by Li et al. [11] and released the
first large-scale person description dataset, the Chinese Uni-
versity of Hong Kong (CUHK) PErson DEscription (CUHK-
PEDES) dataset, which greatly promoted the development of
TBPS. The core challenge in TBPS is achieving cross-modal alignment. To tackle this problem, many methods have been put forward in recent years. Some early works [11], [12], [13], [15], [21] primarily focused on global matching between images and texts. For instance, Zhang and Lu [13] proposed a cross-modal projection matching loss and a cross-modal projection classification loss to learn modality-shared global features. Chen et al. [15] proposed a cross-modal knowledge adaptation method to solve the information inequality problem between images and texts by directly adapting the knowledge of images to the knowledge of texts at three different levels. Although these global-matching-based methods are straightforward and efficient, they may overlook distinctive local details and mix some noise information.

To mine locally discriminative detail information, a plethora of works [16], [17], [18], [19], [20], [22], [23], [24], [25], [26], [27] have focused on exploring local matching between images and texts. Some efforts work on more effective attention mechanisms to mine fine-grained correspondence by complex cross-modal pairwise interactions. For example, Niu et al. [23] designed a hierarchically multigranularity image–text-matching mechanism, i.e., global–global, global–local, and local–local, to achieve more accurate similarity evaluation. Afterward, some effective and lightweight models [20], [28], [29] have been proposed to employ local image parts as a guide to learn correspondingly aligned local text features through cross-modal alignment loss supervision, without complex pairwise interactions. For example, semantically self-aligned network (SSAN) [20] designed a word attention module to attend to part-relevant words with the image part feature as a guide. With the success of Transformers in various tasks, several Transformer-based methods [30], [31] have been proposed recently and achieved state-of-the-art (SOTA) performance.

However, most existing local-matching methods rely on explicitly generated local parts to model fine-grained correspondence between modalities. Such explicitly generated local parts may be ambiguous due to the lack of contextual information [32], [33], or noisy due to the domain gap between external and TBPS data [34], which are unreliable for modeling local fine-grained correspondence between modalities. Also, the direct interactions between modalities may affect each other. In contrast, our method introduces a set of modality-shared semantic centers to implicitly learn locally aligned features, avoiding the guidance or interactions between images and texts. Also, the generation of each locally aligned feature is based on the complete context information of the image/text. Moreover, several works [15], [29] also consider the information inequality problem between images and texts, but they only paid attention to reduce the impact of image background, ignoring environmental factors, while we suppress image background and filter out environmental factors simultaneously.

B. Cross-Modal Retrieval

For cross-modal retrieval tasks [12], [35], [36], [37], the primary challenge is to learn effective and modality-shared visual/textual embeddings in a joint embedding space. Early works [12], [13], [38] focus on designing networks and optimization loss to project images and texts into a joint embedding space. For example, Zheng et al. [12] proposed a dual-path convolutional neural network (CNN) network and an instance loss for image–text retrieval. Zhang and Lu [13] designed a cross-modal projection matching loss and a cross-modal projection classification loss to learn discriminative image–text embeddings. However, these methods only learn global embeddings from images and texts, ignoring local details and fine-grained correspondence between them. Recently, some advanced works [25], [39], [40], [41] have focused on exploring the fine-grained interaction between modalities in the embedding space. For instance, Lee et al. [25] employed the bottom-up attention and bidirectional gate recurrent unit (GRU) network to get image region features and word features, respectively, and presented a stacked cross attention to infer image–text similarity by aligning image region and word features. Liu et al. [39] proposed a graph-structured matching network to infer fine-grained correspondence between modalities. These methods provide valuable inspiration for TBPS. However, unlike general cross-modal retrieval, all samples for TBPS belong to a single category (person), making it more challenging to mine discriminative details and model fine-grained correspondence. Therefore, this article aims to explore an effective local alignment framework for TBPS to learn modality-shared local features.

C. Attention-Based Re-ID

Re-ID is a challenging task due to various factors, such as viewpoint variations, low resolutions, illumination changes, unconstrained poses, occlusions, background clutter, and heterogeneous modalities. The attention mechanism [36], [42], [43] has been employed in many Re-ID methods [6], [8], [44], [45], [46], [47], [48] to overcome the above challenges due to their ability to focus on important information related to the task and suppress irrelevant information. For instance, Yang et al. [44] designed intra-attention and interattention modules to learn robust and discriminative human features from global image and local parts, respectively. To introduce global structural information, Zhang et al. [6] and Li et al. [8] explored global scope relations for feature nodes at each location to learn attention for informative and discriminative feature learning. To eliminate the influence of background clutter on Re-ID, Song et al. [46] designed a mask-guided contrastive attention module by introducing external clues to guide the network to focus on the human body region. Inspired by these works, we design two attention modules to address the information inequality problem between images and texts. The attention modules guide the network to focus on the modality-shared human body region in images and filter out environmental factors while preserving identity information.

III. MANET FRAMEWORK

In this section, we elaborate on the implementation details of our MANet framework, and the overview is shown in Fig. 2. We begin by extracting visual and textual features using CNN
and RNN networks, respectively. Next, we feed the extracted image features to the ISS module, which includes RGL and CAF to eliminate the influence of image background and environmental factors. The processed image and text features are then fed to the global and local encoding branches for global and local alignments, respectively. Finally, the above modules are jointly optimized by identification loss, cross-modality ranking loss, and identity-relevant content consistency loss to learn modality-aligned features for text-to-image retrieval.

A. Visual and Textual Representations

For a batch of $N$ image–text pairs $\{I_i, T_i\}_{i=1}^N$, and corresponding ground-truth label $\{y_i\}_{i=1}^N$ drawn from a TBPS database, where each $y_i \in \mathbb{R}^{N_i}$ is a one-hot ground-truth label vector, $N_i$ is the number of identities. Given a text–image pair $(I_i, T_i)$, the goal is to encode it into a joint feature space to measure the similarity. For an image $I_i$, we utilize a ResNet-50 [49] pretrained on ImageNet [50] with the final fully connected layer removed as the visual backbone to extract the visual feature map $F_i \in \mathbb{R}^{C \times H \times W}$, where $C$ denotes the number of feature channel with height $H$ and width $W$. Our goal is to achieve both global and local alignments between images and texts. To obtain the global and local features for each image, we introduce the global image branch $F^{\text{glb}}_i$ and the local branch $F^{\text{loc}}_i$ to generate a global visual feature $v^g = F^{\text{image}}_{\text{glb}}(F) \in \mathbb{R}^{d_g}$ and a set of local visual features $\{v_1^l, v_2^l, \ldots, v_K^l\} = F^{\text{loc}}_i(F) \in \mathbb{R}^{K \times d_l}$, where $K$ denotes the number of local features.

Given a textual caption $T$, we first encode each word as a vector, i.e., word embedding (WE). Two commonly used methods for TBPS are to generate WE through Word2Vector (W2V) [20] or bidirectional encoder representations from transformers (BERT) [51]. For a fair comparison, we generate WE in both ways. For W2V, we first create a vocabulary from the text database of the training set, only counting some popular words whose appearance frequency is more than 2 [20], and the size of the vocabulary is $V$. After that, we first encode each word in $T$ into a length-$V$ one-hot vector $v_j$. Then, we transform the one-hot vector $v_j$ by a WE matrix $W_{v} \in \mathbb{R}^{V \times d_{w}}$ into a $d_w$-dimensional WE $w_j = W_{v} v_j$. For BERT, we first use the lower cased byte pair encoding (BPE) with a 30,522 vocabulary size to tokenize the text $T$. Then, the textual token sequence is fed into the pretrained BERT model to generate the WE. Note that BERT is only used to generate WE, which is frozen during training.

The long short-term memory (LSTM) network as a special RNN has shown great capabilities in capturing the long-term dependencies among words. We adopt bidirectional LSTM (bi-LSTM) as a text backbone to extract text features. After the embedding stage, all WEs $\{w_1, w_2, \ldots, w_L\} \in \mathbb{R}^{d_w \times L}$ of the caption $T$ generated by W2V or BERT are fed into the bi-LSTM network in sequence to get the hidden
states of forward and backward directions for each word. By averaging the hidden states of two directions to generate the corresponding contextual-aware text representation $e_i \in \mathbb{R}^C$ for each word, all word representations for a caption are stacked as $E = [e_1, e_2, \ldots, e_L] \in \mathbb{R}^{C \times L}$, where $L$ is the number of words in each caption. It should be noted that the length of each caption is different. To ensure the consistency of caption length, we unify the length of the caption as $L$, select the first $L$ words when the length is greater than $L$, and fill zero at the end of the caption when the length is less than $L$.

Similarly, in order to perform global and local alignment, we also obtain the global text feature $t^g = F_{\text{glob}}(E) \in \mathbb{R}^d$ and local text features $\{t_1, t_2, \ldots, t_K\} = F_{\text{loc}}(E) \in \mathbb{R}^{K \times d}$, for each caption by the global text branch $F_{\text{glob}}$ and the local branch $F_{\text{loc}}$. The details of the global branches $F_{\text{image}}$, $F_{\text{text}}$ and the local branch $F_{\text{loc}}$ are presented in the following.

### B. Image-Specific Information Suppression

Considering the fact that the information contained in an image and a text is not equal, it is embodied in two aspects. First, the image is captured at a distance from the person and encompasses not only the pedestrian but also the surrounding background [Fig. 1(a)]. Second, the Re-ID task involves capturing images from multiple cameras at different times and places. Due to the differences in camera parameters and environment (lighting conditions, weather, viewpoint, and so on), the captured images encompass environmental factors that lead to significant intraclass variations, meaning multiple images of the same person are very different [Fig. 1(b)]. Texts are generally cleaner than images, as they mostly describe pedestrian appearance information. Therefore, the background and environmental factors [52], [53] in the image act as noise for text, which increases the modality gap between images and texts, leading to the information inequality problem. To address this issue, we propose an ISS module consisting of two submodules: RGL and CAF, to deal with background and environmental factors, respectively. The solution to this problem can effectively narrow the modality gap and achieve the alignment of information volume between modalities.

1) Relation-Guided Localization: The attention mechanism is widely used in various tasks due to its ability to enhance discriminative features and suppress irrelevant ones, which is in line with our goal. Zhang et al. [6] has proved that the relations among spatial positions in the image provide clustering-like information, which is helpful for the semantic understanding of the image. Therefore, an RGL module is introduced to guide attention learning through the relations among spatial positions. This enables the module to suppress the background and highlight the pedestrian region in the image that corresponds to the text description, as illustrated in Fig. 3.

It is crucial how to measure the relation between features effectively. The most common way is to use the Euclidean or cosine distance to calculate a relational scalar between different features, which indicates the degree of correlation between them while lacking detailed correspondence [54]. Unlike them, we capture more detailed associations among spatial positions by computing a relation vector instead of a scalar. The relation function between vectors $x \in \mathbb{R}^C$ and $y \in \mathbb{R}^C$ is formulated as follows:

$$r(x, y; W, W_\phi, W_\theta) = \text{ReLU}(\text{BN}(W(x - y_\phi)))$$

where $W \in \mathbb{R}^{(C/r_2) \times (C/r_2)}$ denotes a weight matrix of a fully connected (FC) layer, $\text{BN}(\cdot)$ is a batch normalization (BN) layer, and $\text{ReLU}(\cdot)$ is a rectified linear unit activation layer. Based on this, we can get a $(C/r_2)$-dimensional relation vector.

With the visual feature map $F \in \mathbb{R}^{C \times H \times W}$ from the visual backbone, we calculate the relations between each spatial location and all spatial locations to obtain the attention score of each spatial location. Specifically, for a feature vector $f_i \in \mathbb{R}^C$ at the $i$th position, we compute the relation between it and all $N(=H \times W)$ positions to get $N$ relation vectors. Then, the $N$ relation vectors are concatenated to denote the global relation vector $r_i \in \mathbb{R}^{NC/r_2}$ at the $i$th position:

$$r_i = \text{Concat}([r_{i,1}, r_{i,1}, \ldots, r_{i,N}])$$ (2)

To obtain the attention at the $i$th position, we combine the visual feature $f_i$ with its corresponding global relation vector $r_i$ to generate the attention by a projection function implemented by FC layer followed by BN and sigmoid function:

$$a_i = \text{Sigmoid}(\text{BN}(W_a[f_i, r_i]))$$ (4)

where $W_a \in \mathbb{R}^{C \times (1+N/C)}$ and $a_i \in \mathbb{R}^C$. For all positions, we get an attention map $A \in \mathbb{R}^{C \times H \times W}$. Finally, the attention map $A$ and the visual feature map $F$ perform elementwise product to get the enhanced feature map $\tilde{F}$. The enhanced feature map $\tilde{F}$ is subsequently processed in the later stage and fed into the joint space to align with the corresponding text. As the text does not provide any information about the image background, under the supervision of cross-modal alignment loss, the network guided by global relations assigns higher weights to modality-shared human body regions and lower weights to nonpedestrian regions to suppress the background noise.

2) Channel Attention Filtration: The presence of environmental factors in images not only widens the gap between modalities but also results in significant intraclass variations within images. To eliminate the influence of environmental factors, we introduce a CAF module, as depicted in Fig. 4.
Previous studies [55], [56], [57] have shown that IN can effectively reduce the discrepancy among instances by filtering out instance-specific style information. However, directly applying IN can potentially remove some identity-relevant information, thereby damaging the discrimination capability of features. To avoid this issue, we further mine the identity-relevant information from the IN-removed information after IN. Specifically, we apply IN to the visual feature $\hat{F}$ processed by RGL to remove the style information

$$\hat{F}_{IN} = IN(\hat{F}) = \gamma \left( \frac{\hat{F} - \mu(\hat{F})}{\sigma(\hat{F})} \right) + \beta$$

(5)

where $\mu(\cdot)$ and $\sigma(\cdot)$ denote the mean and standard deviation, and $\gamma, \beta \in \mathbb{R}$ are parameters learned from data. Next, we substitute the identity-relevant information by distilling it from the IN-removed information. The IN-removed information is defined as follows:

$$R = \hat{F} - \hat{F}_{IN}.$$

(6)

With the IN-removed information $R$, we mine the identity-relevant information from it through channel attention and then add it to the instance normalized feature

$$\hat{F} = \hat{F}_{IN} + w_C \odot R$$

(7)

where $\odot$ denotes the elementwise product operation and $w_C \in \mathbb{R}^C$ is the channelwise attention weight indicating the identity-relevant channels, which is generated in the same way as [58].

With the CAF module, we expect that the output feature can fully retain effective identity information while discarding redundant noise information. To ensure this, we introduce an identity-relevant content consistency loss, which is explained in detail in III-E.

After text encoding, image encoding, localization, and filtration, we alleviate the information inequality problem between images and texts, resulting in obtaining the text feature map $E \in \mathbb{R}^{C \times L}$ and image feature map $\hat{F} \in \mathbb{R}^{C \times H \times W}$. Although $\hat{F}$ and $E$ achieve the alignment of information volume, images are spatially structured while texts are orderless. Directly comparing these two heterogeneous features is not feasible as they are not well aligned at the semantic level. Therefore, we perform global and local semantic alignment between images and texts in Sections III-C and III-D.

C. Implicit Local Alignment

Local-level alignment plays a crucial role in TBPS, as it allows for the mining of fine-grained correspondences between images and texts. Pedestrian expressions can take various forms, such as image, text, and voice, but their underlying semantic information remains the same. Therefore, we hypothesize that there exists a set of latent semantic topic centers in a shared semantic space, which can completely express the semantic information of pedestrians and are shared across modalities. These centers can be compared with atoms in a dictionary, where different combinations of atoms can express anything. Building on this idea, we propose an ILA module that adaptively aggregates all pixel/word features of image/text to the same centers, resulting in multiple locally aligned image/text features.

Assuming the existence of a set of semantic topic centers $\{c_1, c_2, \ldots, c_K\} \in \mathbb{R}^{K \times d_c}$ in a shared semantic space, where $K$ denotes the number of centers and $c_i \in \mathbb{R}^{d_c}$, $d_c$ is the dimension of shared semantic space. These centers can be learned jointly with the entire network. To assign pixel/word features of image/text to their respective topic centers, we calculate the relations between the pixel (word) features of the image (text) and the centers, similar to (1). We provide a detailed explanation of the assignment on visual feature map, followed by a brief description of the text feature map, as the operation is the same for both feature maps. Given the visual feature map $\hat{F} \in \mathbb{R}^{C \times H \times W}$, first project it into the shared semantic space by a $1 \times 1$ convolution layer $W_{s} \in \mathbb{R}^{d_c \times C}$ after L2 normalization

$$Z^{image} = W_{s} \text{Normalize}(\hat{F})$$

(8)

where $Z^{image} \in \mathbb{R}^{d_c \times H \times W}$. Next, we compute assignments on the corresponding centers for each pixel feature. Specifically, for the feature vector $z^{image}_j \in \mathbb{R}^{d_c}$ at the $j$th pixel in $Z^{image}$, its assignment to the $j$th topic center $c_j$ can be generated in the same way as (1) as follows:

$$a_{i,j}^{image} = r(z^{image}_i, c_j, W^l, W^p, W_p)$$

(9)

where $W^l \in \mathbb{R}^{d_c \times (d_c/r_l)}$ and $W^p, W_p \in \mathbb{R}^{(d_c/r_p) \times d_c}$ are all learnable parameters matrices. Then, the aggregated local image feature for each center can be obtained

$$v^l_j = \sum_{i=1}^{N_v} a_{i,j}^{image} \odot z^{image}_i$$

(10)

where $\odot$ denotes the elementwise product operation and $N_v = H \times W$ denotes the number of pixels. Thus, we can generate a set of aggregated local image features $\{v^l_1, v^l_2, \ldots, v^l_K\}$ for all semantic topic centers.

Similarly, for the textual feature map $E$, we first extend $E \in \mathbb{R}^{C \times L}$ to $\hat{E} \in \mathbb{R}^{C \times L \times 1}$ to meet the input requirement of convolutional layer and then generate a set of aggregated local text features $\{t^l_1, t^l_2, \ldots, t^l_K\}$ in the same way using the share semantic topic centers

$$Z^{text} = W_{s} \text{Normalize}(\hat{E})$$

(11)

$$a_{i,j}^{text} = r(z^{text}_i, c_j, W^l, W^p, W_p)$$

(12)

$$t^l_j = \sum_{i=1}^{L} a_{i,j}^{text} \odot z^{text}_i$$

(13)

where $L$ denotes the length of caption and $z^{text}_i \in \mathbb{R}^{d_c}$ is the feature vector of the $i$th word in $Z^{text}$. Finally, we, respectively, concatenate all local image and text features to get the output.
feature $v^i \in \mathbb{R}^{Kd_i}$ and $t^i \in \mathbb{R}^{Kd_t}$, which are aligned effectively, since the learned local features for images and texts share the same semantic topic centers.

### D. Global Alignment

Moreover, we also introduce GA, since the global features contain more comprehensive information and are complementary to local features. For visual feature map $\hat{F} \in \mathbb{R}^{C \times H \times W}$, we first project it into a joint semantic embedding space through an FC layer $W_g^{\text{image}}$. Because only part of the image region is related to the text description, considering that the global max-pooling (GMP) layer can mine salient information and filter out noise, thus, perform GMP to obtain the global visual feature

$$v^g = \text{GMP}(W_g^{\text{image}}) \hat{F}.$$  \hspace{1cm} (14)

For text feature map $\hat{E} \in \mathbb{R}^{C \times L \times 1}$, we first perform GMP to get the feature vector, followed by projecting the feature vector into the joint semantic embedding space by an FC layer $W_g^{\text{text}}$ to obtain the global text feature:

$$t^g = W_g^{\text{text}} \text{GMP}(\hat{E})$$ \hspace{1cm} (15)

where $v^g, t^g \in \mathbb{R}^{d_g}$. For a batch of $N$ image–text pairs $(I_i, T_i)\text{'}s^{N}_{i=1}$, we generate the global image and text feature set, $V^g, T^g \in \mathbb{R}^{N \times d_g}$, respectively.

### E. Training Loss and Inference

After obtaining the global and local features for images and texts using the proposed model, we train the model with identification loss, cross-modality triplet ranking loss, and identity-relevant content consistency loss. The detailed descriptions of each loss are presented in the following.

1) **Identification Loss:** We adopt the identification loss to classify persons into different groups by their identities, which ensures identity-level matching. Specifically, for image–text pair $(I_i, T_i)$, the cross-entropy loss is imposed on the global feature $(v^g, t^g)$ and each of the $K$ local features $(\{v^g_k\}_{k=1}^{K}, \{t^g_k\}_{k=1}^{K})$. Therefore, the total identification loss $L_{\text{id}}$ is

$$L_{\text{id}} = L_{\text{id}}^{g} + \sum_{k=1}^{K} L_{\text{id}}^{l_k}$$ \hspace{1cm} (16)

$$L_{\text{id}}^{g} = -y_i \log \text{softmax}\left(\left(W^g_{ld}T v^g\right)\right)$$ \hspace{1cm} (17)

$$L_{\text{id}}^{l_k} = -y_i \log \text{softmax}\left(\left(W^l_{ld}^{k}T v^g\right)\right)$$ \hspace{1cm} (18)

where $W_{ld}^g \in \mathbb{R}^{d_g \times N_g}$ and $W_{ld}^{k} \in \mathbb{R}^{d_g \times N_g}$ are the parameters of the FC layer for classification, and $y_i$ is the one-hot ground-truth label vector of image $I_i$, (text $T_i$).

2) **Cross-Modality Ranking Loss:** We employ the common cross-modality bidirectional dual-constrained ranking loss to supervise the modality-shared global and local feature learning. In addition, we also introduce text descriptions of other images with the same identity as [20] as weak supervision signals to overcome the intraclass variance in the text similarity calculation.

$$L_{\text{rank}} = \max(\alpha_1 - S(v^g, t^g_z) + S(v^g, t^g_n), 0)$$
$$+ \max(\alpha_1 - S(v^g, t^g_z) + S(v^g, t^g_p), 0)$$
$$+ \lambda_1 \max(\alpha_2 - S(v^g, \hat{t}^g_z) + S(v^g, \hat{t}^g_n), 0)$$
$$+ \lambda_1 \max(\alpha_2 - S(v^g, \hat{t}^g_z) + S(v^g, \hat{t}^g_p), 0)$$ \hspace{1cm} (19)

where $v^g$ and $t^g$ ($z = g, l$) denote the global or local feature for image and text, respectively. $\hat{t}^g$ denotes the text feature corresponding to another image with the same identity as $v^g$. $(v^g, t^g_z)$ denotes the matched image–text feature pairs, and $(v^g, t^g_z)$ and $(v^g, t^g_p)$ denote the mismatched pairs. $\alpha_1$ and $\alpha_2$ indicate the margins, and the setting strategy of $\alpha_2$ keeps the same with [20]; $\lambda_1$ is a hyperparameter

$$L_{\text{rank}} = L_{\text{rank}}^{g} + L_{\text{rank}}^{l_k}.$$ \hspace{1cm} (20)

3) **Identity-Relevant Content Consistency Loss:** It is introduced to ensure that CAF can fully retain useful identity information while discarding redundant noise information. Specifically, we hope that the identity-relevant information can remain consistent before and after passing through CAF, while only filtering out noise information. A triplet loss is adapted to achieve the goal

$$L_{\text{cons}} = \max(\alpha_3 - S(\tilde{f}, \tilde{f}) + S(\tilde{f}, \tilde{f}_n), 0)$$
$$+ \max(\alpha_3 - S(\tilde{f}, \tilde{f}) + S(\tilde{f}_n, \tilde{f}), 0)$$ \hspace{1cm} (21)

where $\tilde{f} = \text{GAP}(\tilde{F})$ and $\tilde{f} = \text{GAP}(\hat{F})$, respectively, represent the feature vectors before and after $F$ pass through CAF, and $\text{GAP}()$ denotes the global average pooling. $\tilde{f}_n$ and $\tilde{f}$ denote the hardest negative sample for $F$ before and after through CAF in a mini-batch, respectively. $\alpha_3$ is the corresponding margin. $S(\cdot, \cdot)$ denotes the similarity function. The loss retains the identity information by pulling the distance between $\tilde{f}$ and $\tilde{f}$ as close as possible and ensures the discrimination ability of features by pulling distances between $\tilde{f}$ and $\tilde{f}_n$ and $\tilde{f}_n$ and $\tilde{f}$ as far as possible.

Finally, the total loss $L_{\text{total}}$ is utilized to train our model in an end-to-end manner as follows:

$$L_{\text{total}} = L_{\text{id}} + L_{\text{rank}} + \lambda_2 L_{\text{cons}}$$ \hspace{1cm} (22)

where $\lambda_2$ is a hyperparameter to balance the contributions of individual loss terms.

During inference, we compute the similarity scores $S^g$ and $S'$ of global and local features between each text query and image gallery, respectively, using the cosine similarity function. Based on this score $S = S^g + S'$, we rank the person images in the image gallery and retrieve the top matches corresponding to the text query.

### IV. EXPERIMENTS

In this section, we conduct extensive experiments on two TBPS databases to verify the effectiveness and superiority.
of our MANet. We first introduce two large-scale public databases, the implementation and training details of our proposed model. Next, we compare our method with SOTA methods on both databases. Finally, we present some ablation studies to demonstrate the effectiveness of each component of our MANet.

### A. Experiment Settings

1) Datasets: CUHK-PEDES [11] comprises 40,206 images and 80,412 text descriptions of 13,003 persons. Each image is manually annotated with two text descriptions with an average length of not less than 23 words. The database constitutes a vocabulary with 9,408 unique words. To enable a fair comparison with the existing methods, we follow the official data split protocol in [11]. Specifically, 34,054 images of 11,003 persons and corresponding 68,108 text descriptions are used as the training set. The remaining 2000 persons are equally divided into the validation and testing sets, with the validation set comprising 3,078 images and 6,156 text descriptions, and the testing set comprising 3,074 images and 6,148 text descriptions.

Identity-centric and fine-grained person description dataset (ICFG-PEDES) [20] contains 54,522 images of 4,102 persons collected from the multi-scene multi-time person ReID dataset (MSMT17) [72] database, with each image having a corresponding text description of an average length of 37 words. The vocabulary contains 5,554 unique words. Following the data split protocol in [20], the database is divided into training and testing sets. The training set contains 34,674 image–text pairs of 3,102 persons, while the testing set consists of 19,848 image–text pairs of the remaining 1000 persons. For both databases, we use Rank-K (R@K, K = 1, 5, and 10, higher is better) to evaluate the performance of different methods.

2) Implementation Details: For images, we use ResNet-50 [49] pretrained on ImageNet [50] as the visual backbone with minor modifications: remove the average pooling layer and the FC layer, and set the stride of the last convolution layer to 1 for larger feature map. All images are resized to 384 × 128 before being fed into the image backbone, followed by horizontal flipping for data augmentation. For texts, Bi-LSTM is used as the textual backbone, and the text length is unified to $L = 100$ before all texts are sent into the textual backbone. The dimension of WE is $d = 768$ for BERT ($d = 512$ for W2V). The model contains some parameters with dimensions $H = 24$, $W = 8$, $C = 2048$, $d_e = 2048$, and $d_c = 512$. The scale factors $r_1$, $r_2$, and $r_3$ are set to 32, 256, and 4, respectively. The size of the vocabulary in the training set varies for different databases, with CUHK-PEDES set to $V = 5000$ and ICFG-PEDES set to $V = 3000$.

During training, we optimize our model using the Adam optimizer with an initial learning rate of 0.001 for the visual backbone and 0.01 for the rest of the network. The learning rate is decayed at the 30th and 50th epoch with a decay factor of 0.1. To facilitate training, we utilize the linear warm-up strategy for the first ten epochs. The model is trained for 70 epochs with a batch size of 64. Both loss margins, $\alpha_1$ and $\alpha_2$, are set to 0.2. The hyperparameters $\lambda_1$ and $\lambda_2$ are set to 0.1 and 1. Our MANet is implemented in PyTorch and trained on a single RTX3090 24G GPU.
B. Comparison With SOTA

In this section, we compare our method with SOTA methods on two public databases. The results on CUHK-PEDES and ICFG-PEDES are shown in Tables I and II, respectively. Our method consistently outperforms the SOTA methods on these two databases.

1) Results on CUHK-PEDES: We first evaluate the proposed method on the popular database, CUHK-PEDES, as shown in Table I. For a fair comparison, we classify the methods into two types based on the WE ways they are used. Early methods primarily used W2V for WE, and the best-performing method SSAN [20] achieves 61.37% R@1 accuracy. SSAN explicitly splits an image into multiple parts and uses these parts to guide the generation of part-related local textual features. However, this method relies heavily on the quality of the image parts, which may be ambiguous due to the lack of contextual information. As a result, it is not always reliable to guide the generation of aligned text parts using such image parts. In contrast, our method introduces a set of modality-shared semantic centers to implicitly learn locally aligned features, avoiding hard split and interactions between images and texts, and achieves 63.92% R@1 accuracy, outperforming SSAN [20] by 2.55%. The results highlight the superiority of implicit local feature learning for modality alignment.

With the popularity of Transformer, most subsequent works use the pretrained BERT model for WE. Benefiting from the rich language priors in BERT, such methods achieve superior performance. The typical methods include the following: 1) Simple and robust correlation filtering (SRCF) [29] designs two types of filters to effectively extract the key clues and adaptively align the local features, achieving 64.04% R@1 accuracy; 2) Text-image part-based convolutional baseline (TIPCB) [28] learns local features with a similar idea to SSAN [20] and achieves 64.26% R@1 accuracy; 3) Semantically aligned cross-modal network (AXM-Net) aligns images and texts by designing an AXM block for complex cross-modal interactions, achieving 64.44% R@1 accuracy; and 4) Learning granularity-unified representations (LGUR) [31] is based on a Transformer decoder to project textual and visual features into a set of query prototypes to learn aligned local features, which is currently the best method (65.25% R@1 accuracy). Compared with them, our MANet achieves 65.64% R@1 accuracy and surpasses the existing SOTA method LGUR [31] by 0.39%, achieving new SOTA performance on CUHK-PEDES without explicit split and cross-modal interactions. Notably, LGUR is a work of the same period as ours, which shares a similar idea to ours in projecting pixel/word features of image/text to a set of shared prototypes for generating local features. This further confirms the superiority of this implicitly locally aligned feature learning. Furthermore, these works [15], [29], [31] also consider the influence of image background on cross-modal alignment, but ignore the environmental factors, which is a crucial factor contributing to the information inequality problem between modalities. Our MANet simultaneously considers these two kinds of noises and tries our best to mitigate their impact on modality alignment.

2) Results on ICFG-PEDES: To prove the generality of the proposed MANet, we also evaluate the performance on another newly released large-scale database, ICFG-PEDES. The comparison results are presented in Table II. Since this database is up to date, only a few methods are available for comparison. Our MANet outperforms all existing methods reported on ICFG-PEDES in all metrics under both settings (WE via W2V or BERT). Specifically, using W2V for WEs, our MANet achieves 57.73%, 75.42%, and 81.72% in R@1, R@5, and R@10, respectively. When using BERT for WEs, our MANet achieves 59.44%, 76.80%, and 82.75% in R@1, R@5, and R@10, respectively, which significantly surpasses the existing SOTA method LGUR [31] by 2.02%, 1.83%, and 1.30% on all metrics.

The above results demonstrate the superiority and generality of our MANet for TBPS. This is attributed to its capacity to suppress image-specific information and learn implicitly aligned local features, effectively narrowing the modality gap between images and texts on multiple levels, which will be further demonstrated in Section IV-C.

C. Ablation Studies

In this section, we investigate the contribution of each module in the proposed MANet and the effect of important parameters on CUHK-PEDES, including GA, ILA, ISS module (which includes RGL and CAF), and the number K of topic centers. R@1, R@5, and R@10 accuracies (%) are reported.

1) Effectiveness of Components: To clearly illustrate the effectiveness of each component, we incrementally add them to the model and report the experimental results in Table III. Following SSAN [20], baseline refers to a basic GA, which means to perform max pooling on image and text feature maps (F and E), followed by embedding image and text features into a common space for GA through a modality-shared 1 × 1 convolutional layer. We train baseline using the same training strategy as ours. The first row shows the results of BASELINE.

GA represents replacing the GA in baseline as our GA module, and the result is shown in the second line. Compared with baseline, GA improves R@1, R@5, and R@10 by 2.35%, 2.23%, and 1.68%, respectively. When the ILA module is added to baseline, R@1, R@5, and R@10 accuracies are significantly improved by 3.88%, 2.93%, and 2.47% over baseline, which demonstrates that this method of implicitly learning locally aligned features by a set of modality-shared semantic topic centers can effectively model local fine-grained correspondence between images and texts, and narrow the semantic gap across modalities. The joint deployment of GA and ILA can further improve performance, with R@1 accuracy on CUHK-PEDES achieving 62.43%, which has surpassed all methods of the same type in Table I, further proving the superiority of our method. Moreover, the result also demonstrates that GA and ILA can work cooperatively with each other to align images and texts from different granularities.

Although the combination of GA and ILA has achieved superior performance, RGL and CAF can further improve R@1 accuracy by 1.12% and 1.22%, respectively, based on the above results, since RGL and CAF consider the influence
of image background and environmental factors, and alleviate the information inequality problem between modalities. The joint deployment of RGL and CAF can further improve performance and achieve the best. All the above results show that each module in the proposed MANet can play a positive role effectively in alleviating the modality gap.

2) Effectiveness of ISS: The attention mechanism is crucial in the ISS module. To prove the advantages of RGL and CAF in suppressing image-specific information, we conduct a series of comparative experiments, including directly using IN to eliminate environmental factors, removing the identity-relevant content consistency loss $L_{cons}$ designed in CAF, and replacing RGL with existing attention methods, such as convolutional block attention module (CBAM) [73], nonlocal (NL) [74], and relation-aware global attention (RAG) [6]. $A \rightarrow B$ means replacing $A$ with $B$, leaving the rest unchanged. The results are presented in Table IV.

When replacing CAF with IN, there is a drop in R@1 by 1.66% over MANet as loss of some discriminative identity-relevant information along with the elimination of environmental factors. CAF overcomes this problem by using channel attention to restore useful identity information from the IN-removed information. Furthermore, CAF incorporates an identity-relevant content consistency loss to promote information restitutions and enhance feature discrimination, resulting in an improvement of nearly 0.5% in R@1 accuracy. In addition, we compare RGL with some existing attention methods, including the following: 1) CBAM [73], which directly feeds feature maps into a standard convolutional layer to generate weights for each spatial location; 2) NL [74], which strengthens the features of the target position by computing a weighted summation of features of all positions (sources), where the weights are the pairwise relations between the target position and all other positions; and 3) RAG [6], which models the global structural relations with other locations for each location and then infers the attention of the current location based on the feature of the current location and corresponding structural relations.

The results indicate a certain level of degradation in performance. As shown in Fig. 1, the image regions that correspond to textual descriptions are structured yet irregular. It is not reasonable for CBAM [73] to use a regular local context, i.e., a convolution operation of filter size $7 \times 7$, to determine the attention of a feature position. While NL [74] introduces global relations to strengthen the features of the target position by aggregating information from all positions, Cao et al. [75] have shown that the aggregation weights are invariant to the target position and only locally determined by source feature itself, which does not fully utilize the global structural information. RAG [6] overcomes these shortcomings and achieves better performance. In RAG, the measurement of the relation between features is crucial. RAG obtains a relation scalar by calculating the inner product (IP) between feature vectors, which only reflects the degree of correlation between features and lacks detailed correspondence. In contrast, our RGL computes a relation vector by the subtraction between feature vectors, which results in a 0.75% improvement in R@1 over RAG. Guided by these detailed relations, the network can better understand images and highlight modality-shared body regions to better align with texts under the supervision of the cross-modal ranking loss.

3) Different Aggregation Strategies: In the proposed ILA module, image and text features are aligned by adaptively aggregating them into a set of shared semantic topic centers. The feature aggregation strategy plays a crucial role in this process. We evaluated different aggregation strategies and presented the results in Table V. Generalized Vector of Locally Aggregated Descriptors layer (NetVLAD) [76] is a classic feature aggregation method that generates a set of local features by computing the weighted sum of the residuals between each pixel (word) feature and the centers. However, it introduces too much uncertainty, leading to severe performance degradation (2.83% drop over our ILA). In contrast, we compute the relations of all pixel (word) features to the center as weights and the weighted sum of all pixel features to generate the local feature for the center, rather than residuals. LGUR [31] implements similar ideas to ours based on Transformer’s decoder. For a fair comparison, we replace our ILA with the decoder in LGUR. The main difference between ILA and decoder is the weight calculation method when aggregating features. Decoder generates a weight scalar by calculating the IP between the

| Method                  | GA | ILA | RGL | CAF | R@1   | R@5   | R@10  | Params | FLOPs |
|-------------------------|----|-----|-----|-----|-------|-------|-------|--------|-------|
| Baseline                |    |     |     |     | 57.80 | 77.32 | 84.45 | 66.50M | 8.40  |
| GA                      | ✓  |     |     |     | 60.15 | 79.55 | 86.13 | 73.84M | 9.21  |
| ILA                     |    | ✓   |     |     | 61.68 | 80.25 | 86.92 | 67.75M | 8.84  |
| GA+ILA                  | ✓  |     | ✓   |     | 62.43 | 81.14 | 87.35 | 75.09M | 9.65  |
| GA+ILA+RGL              | ✓  | ✓   | ✓   | ✓   | 63.55 | 81.56 | 87.61 | 82.70M | 11.13 |
| GA+ILA+CAF              | ✓  | ✓   | ✓   | ✓   | 63.65 | 81.22 | 87.54 | 76.14M | 9.65  |
| GA+ILA+RGL+CAF (Ours)   | ✓  | ✓   | ✓   | ✓   | 63.92 | 82.15 | 87.69 | 83.75M | 11.14 |

| Method                  | R@1   | R@5   | R@10  | Params | FLOPs |
|-------------------------|-------|-------|-------|--------|-------|
| MANet                   | 63.92 | 82.15 | 87.69 | 83.75M | 11.14 |
| CAF→IN                  | 62.26 | 80.54 | 87.02 | 82.71M | 11.13 |
| CAF (w/o $L_{cons}$)    | 63.47 | 81.38 | 87.91 | 83.75M | 11.14 |
| RGL→CBAM [73]           | 63.03 | 81.11 | 87.90 | 76.67M | 9.66  |
| RGL→NL [74]             | 62.09 | 81.06 | 87.61 | 84.54M | 11.27 |
| RGL→RAG [6]             | 63.17 | 81.47 | 87.56 | 77.73M | 9.96  |
TABLE V
EFFECTS OF DIFFERENT FEATURE AGGREGATIONS ON CUHK-PEDES.
“IP” REPRESENTS AN INNER-PRODUCT OPERATION

| Method            | R@1  | R@5  | R@10 | Params | FLOPs |
|-------------------|------|------|------|--------|-------|
| NetVLAD [76]      | 61.09| 80.99| 86.81| 82.52M | 10.69 |
| Decoder [31]      | 63.63| 81.86| 87.83| 126.57M| 15.86 |
| Decoder (H=1) [31]| 61.98| 80.19| 87.15| 101.41M| 12.03 |
| ILA (IP)          | 61.19| 80.04| 87.04| 83.55M | 10.99 |
| ILA (Ours)        | 63.92| 82.15| 87.69| 83.75M | 11.14 |

4) Center Initialization: In our experiments, we randomly initialize the semantic centers \(\{c_1, c_2, \ldots, c_K\} \in \mathbb{R}^{K \times d}\) in the ILA module using a standard normal distribution. To evaluate the impact of center initialization on performance, we also consider several common initialization schemes, such as uniform distribution initialization, Kaiming_normal distribution, Xavier_normal distribution, all-zero initialization, all-one matrix initialization, identity matrix initialization, constant initialization, and orthogonal initialization. Table VI summarizes the results, showing that our method achieves superior performance under different initialization schemes. This demonstrates the robustness of our method to center initialization. While reasonable weight initialization can accelerate model convergence and facilitate global optimal solution search, especially for models with a large number of parameters, our center is only a simple parameter matrix with limited dependence on initialization.

5) Number of Topic Centers: The value of \(K\) determines the granularity of the local feature, which is crucial for establishing fine-grained correspondence between modalities. We select the most appropriate value of \(K\) from the range of 1–32, as illustrated in Fig. 5. The results indicate that if \(K\) is too small, the granularity of the obtained local features is insufficient, and establishing fine-grained correspondence becomes difficult. When \(K = 1\), the learned features represent a global one. As \(K\) increases, the R@1 accuracy gradually improves but begins to decrease after it is greater than 6. We believe that too large granularity will lead to semantic fragmentation and severely destroy the semantics of images and texts. When \(K = 6\), our proposed method achieves the best results.

6) Cost Complexity: Table III presents the cost complexity of our method, reporting the number of model parameters (Params) and the number of floating-point operations (FLOPs) for an input image. Compared with baseline, our MANet incurs higher computational cost primarily due to the GA (FC layer) and RGL modules (attention mechanism), while the CAF and ILA modules, which are the main contributions of our paper, require very low Params and FLOPs, leading to significant performance improvement at only a very low computational cost. Furthermore, Table VII compares the inference times (across all test queries) of different methods. Both baseline and GA are simple GAs. On this basis, we introduce the local branch, but just one more local vector (by concatenating multiple local features) for similarity calculation, so only a small increase in inference time.

7) Qualitative Results: Fig. 6 visualizes some typical image response maps generated by the RGL module, which
Fig. 6. Visualization of image response maps in the RGL module.

Fig. 7. Visualization of image response maps and top-5 response words for each center in the ILA module.

Fig. 8. Comparison of top-7 retrieval results on CUHK-PEDES between baseline (the first row) and our MANet (the second row) for the same query texts. The correct results are marked by red rectangles.

V. CONCLUSION

In this article, we introduce MANet, an efficient joint MANet designed to alleviate the modality gap between images and texts for TBPS. Our method aims to implicitly learn the local fine-grained correspondence between images and texts while addressing the information inequality problem between modalities. Specifically, MANet leverages an ISS module to alleviate the information inequality problem. This module includes an RGL submodule that models global relation information to help localize the human body region, as well as a CAF submodule that applies IN and channel attention to eliminate environmental factors while preserving identity information. Next, in the ILA module, image and text features are adaptively aggregated into a set of shared semantic topic centers, which enables the implicit learning of locally aligned features without extra supervision and cross-modal interactions. Finally, a GA module is introduced to provide a global cross-modal measurement that complements the local perspective. The proposed MANet is optimized in an end-to-end manner. We demonstrate the superiority and effectiveness of MANet through both qualitative and quantitative experimental results, as well as extensive ablation studies.

REFERENCES

[1] L. Zhang, G. Du, F. Liu, H. Tu, and X. Sha, “Global-local multiple granularity learning for cross-modality visible-infrared person reidentification,” IEEE Trans. Neural Netw. Learn. Syst., early access, Jan. 17, 2021, doi: 10.1109/TNNLS.2021.3085978.
[2] H. Li, S. Yan, Z. Yu, and D. Tao, “Attribute-identity embedding and self-supervised learning for scalable person re-identification,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 10, pp. 3472–3485, Oct. 2020.

[3] Y. Wang, J. Peng, H. Wang, and M. Wang, “Progressive learning with multi-scale attention network for cross-domain vehicle re-identification,” *Sci. China Inf. Sci.*, vol. 65, no. 6, Jun. 2022, Art. no. 160103.

[4] L. Wu et al., “Pseudo-pair based self-similarity learning for unsupervised person re-identification,” *IEEE Trans. Image Process.*, vol. 31, pp. 4803–4816, Oct. 2021.

[5] D. Liu et al., “Generative metric learning for adversarially robust open-world person re-identification,” *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 19, no. 1, pp. 1–19, Jan. 2023.

[6] Z. Zhang, C. Lan, W. Zeng, X. Jin, and Z. Chen, “Relation-aware global attention for person re-identification,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3186–3195.

[7] R. Hou, H. Chang, B. Ma, R. Huang, and S. Shan, “BiCnet-TKS: Learning efficient spatial–temporal representation for video person re-identification,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 2014–2023.

[8] X. Li, W. Zhou, Y. Zhou, and H. Li, “Relation-guided spatial attention and temporal refinement for video-based person re-identification,” in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2020, vol. 34, no. 7, pp. 11434–11441.

[9] R. Yan, J. Tang, X. Shu, Z. Li, and Q. Tian, “Participation-contributed temporal dynamic model for group activity recognition,” in *Proc. 26th ACM Int. Conf. Multimedia*, Oct. 2018, pp. 1292–1300.

[10] W. Zhang, X. He, W. Lu, H. Qiao, and Y. Li, “Feature aggregation with reinforcement learning for video-based person re-identification,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 12, pp. 3847–3852, Dec. 2019.

[11] S. Li, T. Xiao, H. Li, B. Zhou, D. Yue, and X. Wang, “Person search with natural language description,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1970–1979.

[12] Z. Zheng, L. Zheng, M. Garrett, Y. Yang, M. Xu, and Y. Shen, “Dual-path convolutional image-text embeddings with instance loss,” *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 16, no. 2, pp. 51:1–51:23, 2020.

[13] Y. Zhang and H. Lu, “Deep cross-modal projection learning for image-text matching,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 568–580.

[14] Z. Li, H. Tang, Z. Zhang, and K. Huang, “Learning deep context-aware features over body and latent parts for person re-identification,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 384–393.

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

[16] P. Wang, C. Ding, Z. Shao, H. Hong, S. Zhang, and D. Tao, “Quality-aware part models for occluded person re-identification,” *IEEE Trans. Multimedia*, vol. 25, pp. 3154–3165, 2023.

[17] D. Chen, M. Wang, H. Chen, L. Wu, and J. Q. W. Peng, “Cross-modal retrieval with heterogeneous graph embedding,” in *Proc. ACM Int. Conf. Multimedia (MM)*, 2022, pp. 3291–3300.

[18] D. Liu, L. Wu, F. Zheng, L. Liu, and M. Wang, “Verbal-person nets: Pose-guided multi-granularity language-to-person generation,” *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Mar. 9, 2022, doi: 10.1109/TNNLS.2022.3151631.

[19] L. Zhang, X. Dong, D. V. Kalashnikov, S. Mehrotra, and D. Ramanan, “Query-driven approach to face clustering and tagging,” *IEEE Trans. Image Process.*, vol. 25, no. 10, pp. 4504–4513, Oct. 2016.

[20] F. Faghri, D. J. Fleet, J. R. Kiros, and S. Fidler, “VSE++: Improving visual-semantic embeddings with hard negatives,” in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, 2018, pp. 1–13.

[21] C. Liu, Z. Mao, T. Zhang, H. Xie, B. Wang, and Y. Zhang, “Graph structured network for image-text matching,” in *Proc. IEEE/CVF Conf. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10921–10930.

[22] S. Aggarwal, R. V. Babu, and A. Chakraborty, “Text-based person search via attribute-aligned matching,” in *Proc. IEEE Winter Conf. Comput. Vis. (WACV)*, Mar. 2020, pp. 2617–2625.

[23] C. Gao et al., “Contextual non-local alignment over full-scale representation for text-based person search,” 2021, arXiv:2101.03036.

[24] A. Zhu et al., “DSSL: Deep surroundings-person separation learning for text-based person retrieval,” in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 209–217.

[25] Z. Ding, C. Ding, Z. Shao, and D. Tao, “Semantically self-aligned network for text-to-image part-aware person re-identification,” 2021, arXiv:2107.12666.

[26] S. Li, T. Xiao, H. Li, W. Yang, and X. Wang, “Identity-aware textual-visual matching with latent co-attention,” in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1890–1899.

[27] D. Chen et al., “Improving deep visual representation for person re-identification by global and local image-language association,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 54–70.

[28] K. Niu, Y. Huang, W. Ouyang, and L. Wang, “Improving description-based person re-identification by multi-granularity image-text alignments,” *IEEE Trans. Image Process.*, vol. 29, pp. 5542–5556, 2020.

[29] Y. Jing, C. Si, J. Wang, W. Wang, L. Wang, and T. Tan, “Pose-guided multi-granularity attention network for text-based person search,” in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2020, vol. 34, no. 7, pp. 11189–11196.
