IMG-Net: inner-cross-modal attentional multigranular network for description-based person re-identification

Zijie Wang,[a] Aichun Zhu,[a,b,*] Zhe Zheng,[a] Jing Jin,[a] Zhouxin Xue,[a] and Gang Hua[b]

[a]Nanjing Tech University, School of Computer Science and Technology, Nanjing, China
[b]China University of Mining and Technology, School of Information and Control Engineering, Xuzhou, China

Abstract. Given a natural language description, description-based person re-identification aims to retrieve images of the matched person from a large-scale visual database. Due to the existing modality heterogeneity, it is challenging to measure the cross-modal similarity between images and text descriptions. Many of the existing approaches usually utilize a deep-learning model to encode local and global fine-grained features with a strict uniform partition strategy. This breaks the part coherence, making it difficult to capture meaningful information from the within-part and semantic information among body parts. To address this issue, we proposed an inner-cross-modal attentional multigranular network (IMG-Net) to incorporate inner-modal self-attention and cross-modal hard-region attention with the fine-grained model for extracting the multigranular semantic information. Specifically, the inner-modal self-attention module is proposed to address the within-part consistency broken problem using both spatial-wise and channel-wise information. Following it is a multigranular feature extraction module, which is used to extract rich local and global visual and textual features with the help of group normalization (GN). Then a cross-modal hard-region attention module is proposed to obtain the local visual representation and phrase representation. Furthermore, a GN is used instead of batch normalization for the accurate batch statistics estimation. Comprehensive experiments with ablation analysis demonstrate that IMG-Net achieves the state-of-the-art performance on the CUHK-PEDES dataset and outperforms other previous methods significantly. © *2020 SPIE and IS&T* [DOI: 10.1117/1.JEI.29.4.043028]

Keywords: person re-identification; natural language description; multigranular matching.

Paper 200265 received Apr. 17, 2020; accepted for publication Aug. 6, 2020; published online Aug. 28, 2020.

1 Introduction

As person re-identification (Re-ID) is currently widely applied in activity analysis and video surveillance,[1–11] it has drawn remarkable attention. Large-scale videos are generated every second with surveillance cameras emerging rapidly. It seems infeasible to search for corresponding criminal suspects manually from large-scale video data, which makes it an urgent need to develop automatic approaches to efficiently handle this task. In general, existing methods can be classified into the ones with image-based query, attribute-based query, and description-based query according to the query modality. The major limitation of image-based person Re-ID methods is the requirement for at least one image of the queried person. Nevertheless, verbal descriptions of the suspect may be the only accessible information in many crime scenes. While attribute-based methods suffer from the limited capability to describe the appearance, description-based methods can provide much more detailed information about the queried person. Although description-based person Re-ID has the above advantages and has been studied from various perspectives,[3,12–14] it remains a challenging task to be addressed.

Given a sentence describing a certain person, description-based person Re-ID aims to retrieve images of this person from a large-scale image database. It is tough to measure the cross-modal

*Address all correspondence to Aichun Zhu, E-mail: aichun.zhu@njtech.edu.cn

1017-9909/2020/$28.00© 2020 SPIE and IS&T

similarity between images and text descriptions due to the existing modality heterogeneity. Currently, the major challenge in description-based person Re-ID is about how to effectively extract feature vectors from both image and text modalities. Information contained in feature vectors extracted the two different modalities should be well-aligned so as to make the following matching step more accurate. Ye et al.[15] conducted a specific attribute completion to enrich the original text query and generated a more expressive attribute vector, and then, a pairwise-based metric learning is introduced for completed attribute vectors. Recently, deep neural networks are commonly utilized in previous works to extract global and local representations. Based on the work of Ye et al.,[15] Shuang et al.[3] made the CUHK-PEDES dataset, which is currently the only benchmark dataset for description-based person Re-ID tasks, and further adopted a pretrained VGG-16 backbone to extract global visual features. Following their work, Chen et al.[12] proposed an efficient patch–word matching model to capture the local similarity between image and text. More recently, many researchers attempt to fuse local and global features to better handle this task. Some works employ hard partitioning methods to utilize visual local information. Niu et al.[14] proposed a multigranularity image–text alignments (MIA) model that extracts fine-grained features by partitioning the feature map horizontally into multiple nonoverlapping parts. Then, they adopted a cross-modal attention mechanism to determine affinities between local/global visual and textual components. However, strict uniform partitioning strategies, such as this one, can be rough and usually breaks within-part consistency. Some other works adopted approaches based on preprocessing with external cues. Jing et al.[13] employed pose information as inner-modal attention to provide soft partial image regions and aggregated more discriminative information for the following partitioning. Nevertheless, utilizing prior knowledge, such as poses, can suffer from the deviations of the pose estimation and large computation consumption.

Most of the state-of-the-art methods[3,13,14] aim to enable a deep-learning model to encode local and global fine-grained features with a strict uniform partition strategy. This breaks the part coherence, making it difficult for the deep neural network to capture meaningful information from the within-part and semantic information among body parts, thereby harming the performance. Thus, we attempt to enable the fine-grained deep-learning model to join the meaningful information and semantic information with a horizontally attentional partition (as shown in Fig. 1). We incorporate inner-modal self-attention and cross-modal hard-region attention with the fine-grained model to extract the multigranular semantic information. Therefore, this paper proposed an inner-cross-modal attentional multigranular network (IMG-Net), which extracts four representations of different granularities, including global visual representation, local visual representation, sentence representation, and phrase representation, and then employs three different cross-modal combinations of global and local features to match visual and textual information. Instead of taking the combination of visual local feature and phrase representation (local2phrase) into consideration as Niu et al.,[14] which can be time-consuming and further break the within-part consistency, only the global2sentence, local2sentence, and global2phrase combinations are utilized. IMG-Net contains three main modules: inner-modal self-attention module,

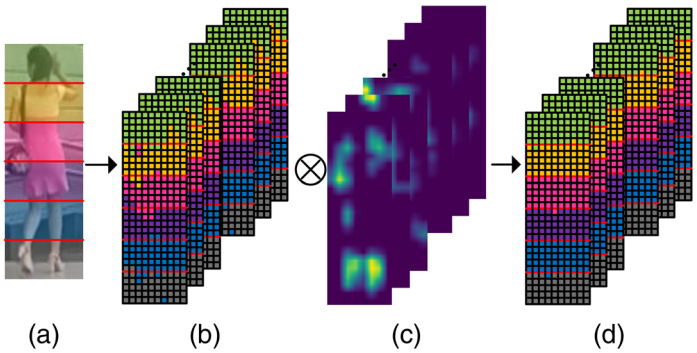

Fig. 1 (a) Illustration of the horizontal partitioning strategy of the input image. (b) Illustration of the output feature of the visual backbone, which suffers from within-part inconsistency problem. (c) Inner-modal self-attention masks are employed to denoise the feature map and enhance the within-part consistency. (d) Illustration of the final output feature map of the self-attention module.

multigranular feature extraction module, and cross-modal hard-region attention module. The inner-modal self-attention module is proposed to address the within-part consistency broken problem using both spatial-wise and channel-wise information. Following it is a multigranular feature extraction module, which is used to extract rich local and global visual and textual features with the help of group normalization (GN). Then a cross-modal hard-region attention module is proposed to obtain the local visual representation and phrase representation. Our proposed method is evaluated on a challenging dataset CUHK-PEDES,[3] which is currently the only available dataset for description-based person Re-ID tasks. Considering that there are over 10,000 person ID categories in the training set, simply employing batch normalization (BN) may suffer from inaccurate batch statistics estimation. To deal with this problem, we adopt the GN[16] that divides the channels into groups and computes the mean and variance within each group for normalization. GN is much more stable with a relatively small batch size (e.g., Ref. 7), as its computation has nothing to do with the batch size. Experimental results present that the proposed IMG-Net achieves the state-of-the-art performance on this dataset.

In summary, the main contributions of this paper are fourfold. (1) An inner-cross-modal attention is proposed to address the within-part consistency broken problem in the description-based person Re-ID task. (2) An IMG-Net is proposed to incorporate inner-modal self-attention and cross-modal hard-region attention with the fine-grained model for extracting the multigranular semantic information. (3) GN is first used in the description-based person Re-ID task. (4) A comprehensive study is carried out to evaluate the proposed IMG-Net model. Experimental results demonstrate that the proposed IMG-Net outperforms the previous methods and achieve the state-of-the-art performance on the CUHK-PEDES dataset.

2 Related Works

2.1 *Person Re-Identification*

In recent years, person Re-ID has got more and more attention in the related field.[1–6,8–11,15,17,18] The current person Re-ID methods are mainly based on deep learning. Liu et al.[19] proposed the pose transferrable-person-Re-ID framework, which uses the pose transferred sample extension (i.e., with ID supervision) to enhance the Re-ID model training, and in the experiment, the superior performance improvement was achieved. In the case of not designing in detail and expanding the Re-ID model, the effect is better than other methods. Zhong et al.[20] recently proposed to study the intradomain variation of the target domain and proposed a view that the Re-ID model is dependent on sample invariance, camera invariance, and neighborhood invariance. Sun et al.[21] recently proposed a visibility-aware part model to significantly improve the learned representation and the achieving accuracy by considering a few parts of the Re-ID scenes combined with the self-supervising model of some feature observations to perceive the visibility of the region.

2.2 *Description-Based Person Re-Identification*

Description-based person Re-ID has been studied from various perspectives.[3,12–15] Due to the heterogeneity of the modes (cross-modality), it is difficult to directly measure the similarity between images and descriptions. Ye et al.[15] conducted a specific attribute completion to enrich the original text query and generated a more expressive attribute vector, and then, a pairwise-based metric learning is introduced for completed attribute vectors. On this basis, Shuang et al.[3] made the CUHK-PEDES dataset, which is currently the only benchmark dataset for description-based person Re-ID tasks, and further employed a VGG-16 to extract global visual features. Following this work, Chen et al.[12] proposed an efficient patch–word matching model to capture the local similarity between image and text. A text–image modality adversarial matching approach (TIMAM) is proposed by Sarafianos et al.[22] to learn modality-invariant feature representation by virtue of adversarial and cross-modal matching objectives. Recently, many researchers attempt to fuse local and global features in this task. Niu et al.[14] proposed an MIA model to extract fine-grained features by partitioning the feature map horizontally into

multiple nonoverlapping parts and then adopted a cross-modal attention mechanism to determine affinities between visual and textual components. This strict uniform partitioning strategy usually breaks within-part consistency. Despite some researchers[13] employ pose information as inner-modal attention to provide soft partial image regions and aggregate more discriminative information for the following partitioning, it still suffers from the deviations of the pose estimation and the large computation consumption.

2.3 *Attention for Person Search*

The attention mechanism has become an integral part of models in many deep-learning-based computer vision tasks,[23–29] such as image classification and segmentation,[30,31] object detection.[32,33] Wang et al.[34] proposed the nonlocal block to capture long-range dependencies by obtaining the weighted sum of the features at all positions. Recently, many attentional deep-learning methods have been proposed to handle the misalignment problems in person Re-ID.[34–38] A common solution of these methods is to incorporate an attentional subnetwork to a deep-Re-ID model. Nevertheless, these models usually consider only coarse region-level attention while ignoring the fine-grained saliency information. To address this problem, Li et al.[38] proposed a harmonious attention convolutional neural network (HA-CNN) model to combine the soft pixel attention and hard-region attention for person Re-ID. Jiang et al.[39] proposed a simple self-attention learning method to capture multilevel information from different layers with the help of the dilated convolution. For the recent work on the description-based person Re-ID, Li et al.[37] proposed a hard-matching co-attention method to extract word–image features using a spatial attention and a semantic attention. Jing et al.[13] employed pose information as inner-modal attention to provide soft partial image regions.

Different from the above methods, we are probably the first to incorporate inner-modal self-attention and cross-modal hard-region attention with the fine-grained model to extract the multigranular semantic information. During the feature vector matching stage, instead of taking the combination of visual local feature and phrase representation (local2phrase) into consideration as Niu et al.,[14] which can be time-consuming and further break the within-part consistency, only the global2sentence, local2sentence, and global2phrase combinations are utilized. By utilizing GN, IMG-Net can employ a relatively much smaller batch size than previous works using a batch size of 96[14] or even 128,[13,3] while reaching much better performance.

3 Proposed Method

In this section, we describe the proposed IMG-Net in detail. As shown in Fig. 2, it contains an inner-modal self-attention module, a multigranular feature extraction module, and a cross-modal hard-region attention module. In the following sections, we will detail these modules.

3.1 *Inner-Modal Self-Attention Module*

As discussed in Sec. 1, simply partitioning the image to extract local features can be a bad choice in some cases. This uniform partition strategy inevitably introduces partition noises and errors, hence compromising the within-part consistency and leading to a less discriminative learned feature. To handle this problem, we adopt the inner-modal self-attention module that combines a channel-wise attention and a spatial-wise attention as complementary to each other (see Fig. 3).

The channel-wise attention takes a three-dimensional feature map $X \in \mathbb{R}^{h\times w\times c}$ as input, where h, w, and c denote height, width, and channel number of the feature map, respectively. We first obtain a channel descriptor by a squeeze operation, which is realized as an average pooling operation to aggregate the feature map across the two spatial dimensions.

$$X_{\text{squ}}^{C} = \frac{1}{h \times w} \sum_{i=1}^{h} \sum_{j=1}^{w} X_{i,j,1:c}. \tag{1}$$

Then we perform an excitation operation to model the channel-wise relationships, which is defined as

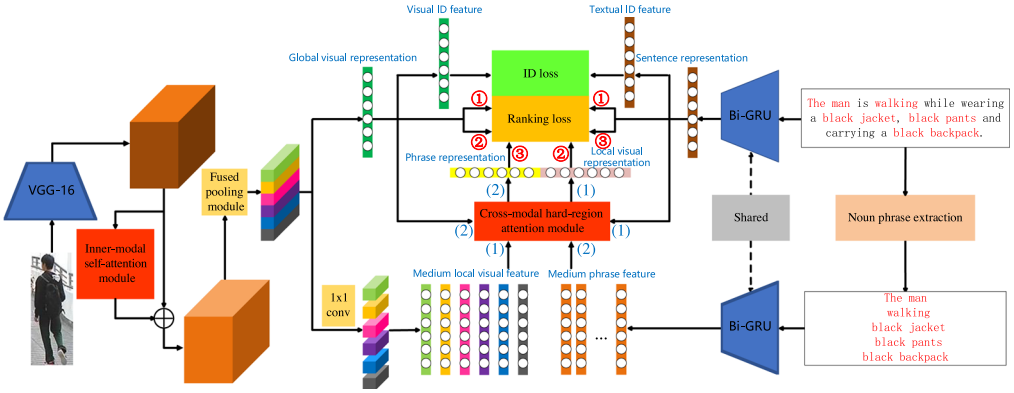

Fig. 2 The overall architecture of our proposed IMG-Net. It contains an inner-modal self-attention module, a multigranular feature extraction module, and a cross-modal hard-region attention module. IMG-Net extracts four representations of different granularities—global visual representation, local visual representation, sentence representation, and phrase representation—and matches visual and textual information via three different cross-modal combinations, including global2sentence, local2sentence, and global2phrase. The sequence numbers (1) and (2) denote the corresponding vectors processed and generated by the cross-modal hard-region attention module, while (1) to (3) denote the corresponding combinations when training IMG-Net.

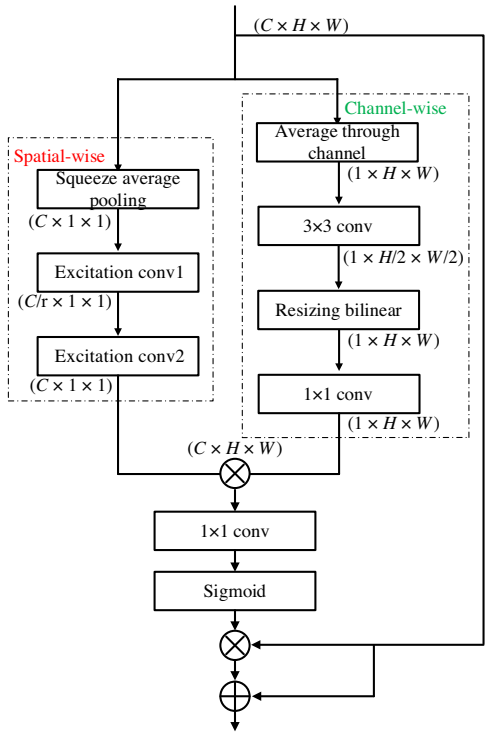

Fig. 3 The inner-modal self-attention module consists of a channel-wise attention and a spatial-wise attention. The two attention branches can properly complement each other to select the fine-grained important pixels and denoising the feature map, to enhance the within-part consistency.

$$C = \mathrm{ReLU}[W_{\mathrm{ex2}} \times \mathrm{ReLU}(W_{\mathrm{ex1}} \times X_{\mathrm{squ}}^{C})], \tag{2}$$

where $C \in \mathbb{R}^{1\times1\times c}$, $W_{\mathrm{ex01}} \in \mathbb{R}^{\frac{c}{r}\times c}$, and $W_{\mathrm{ex2}} \in \mathbb{R}^{c\times\frac{c}{r}}$. To reduce parameter number and computation cost, a bottleneck is adopted via a dimension-reducing fully-connected (FC) layer and a dimension-increasing FC layer. The term r denotes the reduction rate.

The spatial-wise attention takes the same $X \in \mathbb{R}^{h\times w\times c}$ as input. Considering that the spatial attention mask is shared by all channels, X is first averaged through the channel dimension.

$$X^S_{\text{avg}} = \frac{1}{c}\sum_{i=1}^{c} X_{1:h,1:w,i}. \tag{3}$$

After that, X^S_{avg} is subsequently passed through a 3×3 convolution layer with stride 2, a resizing bilinear layer, and a 1×1 convolution layer to obtain the final spatial attention mask $S \in \mathbb{R}^{h\times w\times 1}$.

After obtaining $S \in \mathbb{R}^{h\times w\times 1}$ and $C \in \mathbb{R}^{1\times 1\times c}$ through two separate branches, we form the final attention mask by a Hadamard product followed by a 1×1 convolution layer and a sigmoid layer to constrain the attention value between 0 and 1.

$$A = \text{sigmoid}[W^A \times (S * C)], \tag{4}$$

where W^A denotes the 1×1 convolution layer to further fuse the two attention masks after the Hadamard product.

Considering that the values of the attention mask A are below 1, directly fusing A with the input feature map X by a Hadamard product operation may degrade the discriminative information in X. Thus, we adopt the attention residual learning method.

$$X_{\text{atten}} = (1 + A) * X, \tag{5}$$

where $X_{\text{atten}} \in \mathbb{R}^{h\times w\times c}$ is the final output feature map of inner-modal attention module. With this attention residual learning method, the performance of our inner-modal attention module will be no worse than its counterpart without attention when each value in A is 0.

3.2 *Multigranular Feature Extraction*

IMG-Net extracts four representations of different granularities for the subsequent image–text matching part, including global visual representation, local visual representation, sentence representation, and phrase representation.

For visual representations extraction, we first use a conventional CNN backbone followed by the inner-modal attention module to extract the shared medium feature map $\varphi(I) \in \mathbb{R}^{w\times h\times c}$, where h, w, and c denote height, width, and channel number of the feature map, respectively. Then, $\varphi(I)$ is horizontally partitioned into k nonoverlapping stripes to fetch the fine-grained component features. This can be carried out simply by a pooling operation. In this work, we proposed a fused pooling module, which is described as follows:

$$\phi(I) = \text{AvgPooling}[\varphi(I)] + \text{MaxPooling}[\varphi(I)]. \tag{6}$$

Average pooling layer is adopted in many previous works for its ability to take contextual information into consideration while downsampling. Nevertheless, it may weaken some discriminative signals if signals surrounded are relatively weak (e.g., background). Therefore, we also separately pass the $\varphi(I)$ through a maximum pooling layer to catch those discriminative signals. The outputs of average pooling layer and maximum pooling layer are added to form the final output of the fused pooling module. The $\varphi(I)$ is so that transformed to $\phi(I) \in \mathbb{R}^{k\times 1\times c}$.

Then, we split the visual representation extraction part into two branches, namely, global visual branch and local visual branch.

As for the global visual branch, we just reshape the $\phi(I)$ into a $(k\times c)$-dimension vector $\phi^G_{\text{reshaped}}(I)$, then pass it through a GN layer followed by an FC layer to obtain the final global visual representation $V^G \in \mathbb{R}^P$.

In the local visual branch, a 1×1 convolutional layer is first used to reduce the dimensionality from c to $c/2$ to get $\phi^L_{\text{reduced}}(I) \in \mathbb{R}^{p\times c/2}$, which means there are k horizontal regions and each is represented by a $c/2$-dimensional vector. After the reduction, we separately passed these k vectors through a GN layer and then two FC layers with a ReLU layer between them to obtain

a medium local visual vector $M_i^V \in \mathbb{R}^P$, $i = 1, \cdots, k$. Then we concatenate them to form the medium local visual feature $M^V \in \mathbb{R}^{k \times P}$.

For textual representations extraction, each word $w \in \mathbb{R}^W$ in the text is embedded into a vector $x \in \mathbb{R}^E$ by

$$x = W_e \times w, \tag{7}$$

where $W_e \in \mathbb{R}^{E \times W}$. After that, a bi-directional gated recurrent unit (bi-GRU) is utilized to model the dependencies between adjacent words for both sentences and phrases.

$$\overrightarrow{h_t} = \overrightarrow{\mathrm{GRU}}(x, \overrightarrow{h_{t-1}}), \tag{8}$$

$$\overleftarrow{h_t} = \overleftarrow{\mathrm{GRU}}(x, \overleftarrow{h_{t-1}}), \tag{9}$$

where $\overrightarrow{\mathrm{GRU}}$ and $\overleftarrow{\mathrm{GRU}}$ represent the forward and backward GRUs, respectively, $t = 1, \cdots, n$, and n is the number of words in the input sentence.

Then the last hidden states of the forward and backward GRUs $\overrightarrow{h_n}$ and $\overleftarrow{h_n}$ are concatenated to obtain the textual representation as

$$e = \mathrm{concat}(\overrightarrow{h_n}, \overleftarrow{h_n}), \tag{10}$$

where $e \in \mathbb{R}^P$.

As for sentence representation extraction, we pass the concatenated feature e^S through a GN layer and an FC layer to obtain the sentence representation $T^S \in \mathbb{R}^P$.

When it comes to phrases, the concatenated feature of each phrase $e_i^P \in \mathbb{R}^P$, $i = 1, \cdots, n$, is separately passed through a GN layer and two FC layers with a ReLU layer between them to obtain a medium phrase representation vector $M_i^P \in \mathbb{R}^P$. Then we concatenate them to form the medium phrase feature $M^P \in \mathbb{R}^{n \times P}$.

3.3 *Cross-Modal Hard-Region Attention Module*

To handle the two medium feature M^V and M^P, we adopt the cross-modal hard-region attention module to finally obtain the local visual representation and phrase representation.

For the local visual representation, we first compute how firmly each local part relates to the sentence representation T^S, which is represented as

$$\alpha_i^V = \frac{\exp[\cos(M_i^V, T^S)]}{\sum_{j=1}^{6} \exp[\cos(M_j^V, T^S)]}, \tag{11}$$

where α_i^V represents the relation between the i'th local visual part and the sentence, and $\cos(\cdot, \cdot)$ denotes the cosine similarity function between two feature vectors.

Then, we can obtain the local visual representation with a threshold-guided weighted summation

$$V^L = \sum_{\alpha_i^V > \frac{1}{k}} \alpha_i \cdot M_i^V, \tag{12}$$

where $V^L \in \mathbb{R}^P$ is the final local visual representation.

Following a similar method, we can obtain the final phrase representation by

$$\alpha_i^P = \frac{\exp[V^G, \cos(M_i^P)]}{\sum_{j=1}^{n} \exp[\cos(V^G, M_j^P)]}, \tag{13}$$

$$T^P = \sum_{\alpha_i^P > \frac{1}{n}} \alpha_i \cdot M_i^P. \tag{14}$$

3.4 *Image–Text Matching*

Now that the four representations of diverse modals and granularities have been obtained, the next step is to match visual and textual information, respectively. IMG-Net proposes three different cross-modal combinations of global and local features, including global2sentence, local2sentence, and global2phrase.

For each of the three parts, we can simply compute the similarity with relevant representations by

$$\mathrm{Simi}_{\mathrm{GS}} = \cos(V^G, T^S), \tag{15}$$

$$\mathrm{Simi}_{\mathrm{LS}} = \cos(V^L, T^S), \tag{16}$$

$$\mathrm{Simi}_{\mathrm{GP}} = \cos(V^G, T^P), \tag{17}$$

where $\cos(\cdot,\cdot)$ denotes the cosine similarity function between two feature vectors.

In the test stage, we fuse the three similarities as

$$\mathrm{Simi}_{\mathrm{IMGNet}} = \mathrm{Simi}_{\mathrm{GS}} + \frac{1}{2}(\mathrm{Simi}_{\mathrm{LS}} + \mathrm{Simi}_{\mathrm{GP}}). \tag{18}$$

3.5 *Loss Functions*

The identification loss is adopted to classify people into different groups according to their identifications. The proposed ID losses L_{ID}^V and L_{ID}^T for visual and textual part are defined as

$$L_{\mathrm{ID}}^V = -\log\{\mathrm{softmax}[W_{\mathrm{ID}} \times GN(V^G)]\}, \tag{19}$$

$$L_{\mathrm{ID}}^T = -\log\{\mathrm{softmax}[W_{\mathrm{ID}} \times GN(V^T)]\}, \tag{20}$$

where $W_{\mathrm{ID}} \in \mathbb{R}^{Q\times P}$ is a shared transformation matrix that denotes an FC layer with no bias and Q is the number of different people in the training set. GN denotes the group normalization layer. Only the two global representations V^G and T^S are used here allowing for global representations can provide more complete information for this classification part. The transformation matrix W_{ID} is shared to map visual and textual ID representation into a same feature space.

In addition, the triplet ranking loss is commonly used in person Re-ID and description-based person Re-ID tasks. Instead of using the furthest positive and closest negative sampled pairs in each minibatch, we follow visual-semantic embeddings to adopt the sum of all pairs when computing the hinge-based triplet ranking loss function.

$$\begin{aligned} L_{\mathrm{ranking}}^k = &\sum_{\hat{T}} \max[\alpha - \cos(V,T) + \cos(V,\hat{T})] \\ &+ \sum_{\hat{V}} \max[\alpha - \cos(V,T) + \cos(\hat{V},T)], \end{aligned} \tag{21}$$

where L_{ranking}^k denotes $L_{\mathrm{ranking}}^{GS}$, $L_{\mathrm{ranking}}^{LS}$, or $L_{\mathrm{ranking}}^{GP}$. V can be V^G or V^L while T can be T^S or T^P, respectively, according to L_{ranking}^k. (V,T) denotes the matched visual-textual pairs while $(V,\hat{T})$ or $(\hat{V},T)$ denotes the mismatched pairs and α is a margin.

3.6 *Training Strategy*

We carry out a two-stage training strategy to train the IMG-Net model.

Stage 1. We first fix parameters of the visual CNN backbone and train the left parts of IMG-Net including the textual branches and the FC layer in the global visual branch with the two ID losses. Thus, here, the loss function is

$$L_{\mathrm{stage1}} = L_{\mathrm{ID}}^V + L_{\mathrm{ID}}^T. \tag{22}$$

Stage 2. Then we fine-tune all parameters of IMG-Net including parameters of VGG-16 with both ID losses and ranking losses. The loss function in stage 2 is

$$L_{\text{stage2}} = L_{\text{ID}}^{V} + L_{\text{ID}}^{T} + L_{\text{ranking}}^{GS} + L_{\text{ranking}}^{LS} + L_{\text{ranking}}^{GP}. \tag{23}$$

As discussed by Ye et al.,[40] ID loss treats the training process of person Re-ID as a classification problem, i.e., each identity is a distinct class, while the ranking loss treats the Re-ID model training process as a retrieval ranking problem. As the identification loss concerns mainly the ID category of a given person, it functions more like a loose constraint that fails to catch enough detailed information for the fine-grained matching task. On the contrary, the ranking loss is stricter because it regards the description sentences annotated for a certain image as negative for any other images even with the same person ID. The ranking loss ensures the positive pairs are closer than negative ones with a margin of α. Owing to the diverse intrinsic properties of ID loss and ranking loss, they function differently during training. In stage 1, the ID losses perform as initialization for stage 2 by eliminating obvious mismatched pairs. In stage 2, ranking losses are adopted to catch more fine-grained information while fine-tuning the model, and in this stage, ID losses help to regularize the model.

4 Experiments

4.1 *Experiment Setup*

4.1.1 *Dataset and evaluation metrics*

The CUHK-PEDES dataset[41] is currently the only available dataset for description-based person Re-ID tasks. Following the data split approach of Li et al.,[41] the training set contains 34,054 images, 11,003 persons, and 68,126 textual descriptions. There are 3078 images, 1000 persons, and 6158 textual descriptions in the validation set, while 3074 images, 1000 persons, and 6156 textual descriptions in the testing set. Almost every image has two descriptions, and each sentence is generally no shorter than 23 words.

The top-k accuracy is adopted to evaluate the performance. Given a query description, all test images are ranked by their similarities with this sentence. If any image of the corresponding person is contained in the top-k images, we call this a successful retrieval. Top-1, top-5, and top-10 accuracies for all experiments are reported.

4.1.2 *Implementation details*

Considering that almost all of the previous methods on the description-based person Re-ID task employ VGG-16[42] and ResNet-50[43] as visual CNN backbone, we also choose the pretrained VGG-16 and ResNet-50 as the backbone for IMG-Net, to compare with previous works fairly. The dimensionality P is set to 1024 in our experiments. After dropping the words that appear less than twice, the word number W becomes 4984. The dimensionality E of embedded word vectors is set to 300. Noun phrases of each sentence are obtained with the Natural Language ToolKit by syntactic analysis, word segmentation, and part-of-speech tagging, whose number is kept flexible.

Parameters of the visual CNN backbone are initialized with weights pretrained on the ImageNet classification task. An Adam optimizer is adopted to train the model with a batch size of 32. With the help of GN, IMG-Net employs a relatively much smaller batch size than previous works using a batch size of 96[14] or even 128,[13,3] while reaching a much better performance. The margin α of ranking losses is set to 0.2. In training stage 1, we start the iteration with a learning rate of 1×10^{-3} for 10 epochs with all weights in the visual backbone fixed. In stage 2, we first initialize the learning rate to 2×10^{-4}. During the early 15 epochs, we just let the Adam optimizer to find its way down. After that, the initial learning rate for later epochs is defined as

$$lr = 2 \times 10^{-4} \times \left(\frac{1}{10}\right)^{\text{epoch}//10}, \tag{24}$$

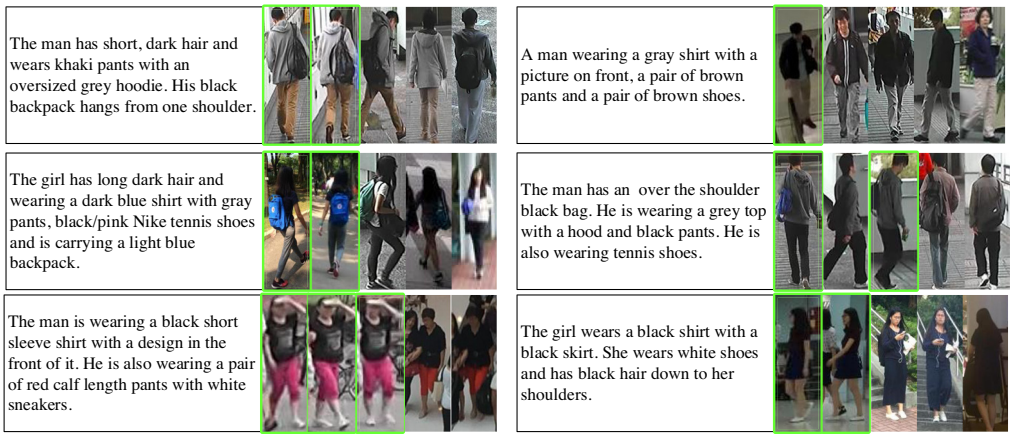

Fig. 4 Examples of top-5 description-based person Re-ID results by IMG-Net. Given a certain sentence, IMG-Net calculates the similarities between this sentence and each image in the person image database. Then the five images that are most similar with the query sentence are picked out and sorted by similarity. Images of the corresponding person are marked by green rectangles.

where lr means the learning rate and $\cdot\ //\ \cdot$ denotes a division operation only takes the integer part. We totally train the stage 2 for 20 epochs.

Several examples of top-5 retrieval results are shown in Fig. 4. Given a certain sentence, IMG-Net calculates the similarities between this sentence and each image in the person image database. Then the five images that are most similar with the query sentence are picked out and sorted by similarity. Images of the target person are marked by green rectangles.

4.2 *Ablation Study*

To further investigate the proposed components of IMG-Net, we carry out plenty of ablation experiments and the results have been shown in Tables 1–6. The highest values in each table are bolded. Some of the examples of top-5 description-based person Re-ID results by IMG-Net have been shown in Fig. 5.

4.2.1 *Inner-modal self-attention modules*

Ablation analysis is first employed to evaluate the effectiveness of each self-attention component in IMG-Net. As shown in Tables 1 and 2, "channel" and "spatial" denote whether the channel-wise attention and spatial-wise attention are employed in the model, respectively. The results show that either the channel-wise or spatial-wise attention can bring accuracy enhancement separately. When fused, the performance is further lifted, suggesting that these two attention modules can function as a complementation to each other to achieve the best performance.

Fig. 5 Visualization of attention maps extracted by the proposed inner-modal self-attention module. By employing the inner-modal self-attention masks, IMG-Net can catch more salient information and the within-part consistency can be enhanced.

Table 1 Ablation analysis of the IMG-Net model (VGG-16 backbone).

Channel	Spatial	GS	LS	GP	BN	GN	Backbone	Top-1	Top-5	Top-10
×	×	✓	✓	✓	✓	×	VGG-16	47.89	70.02	79.24
✓	✓	✓	✓	✓	✓	×	VGG-16	51.38	72.51	81.55
×	×	✓	×	×	×	✓	VGG-16	43.41	67.32	76.54
×	×	×	✓	×	×	✓	VGG-16	44.81	69.43	78.49
×	×	×	×	✓	×	✓	VGG-16	39.10	64.97	77.02
×	×	✓	✓	×	×	✓	VGG-16	48.72	72.97	81.94
×	×	✓	×	✓	×	✓	VGG-16	48.13	70.86	81.33
×	×	×	✓	✓	×	✓	VGG-16	49.08	73.33	82.96
×	×	✓	✓	✓	×	✓	VGG-16	50.11	75.11	83.03
✓	×	✓	×	×	×	✓	VGG-16	45.94	69.17	78.62
✓	×	×	✓	×	×	✓	VGG-16	45.32	70.11	79.01
✓	×	×	×	✓	×	✓	VGG-16	41.88	66.89	77.71
✓	×	✓	✓	×	×	✓	VGG-16	49.90	73.20	82.21
✓	×	✓	×	✓	×	✓	VGG-16	48.73	72.48	81.45
✓	×	×	✓	✓	×	✓	VGG-16	50.92	74.95	83.53
✓	×	✓	✓	✓	×	✓	VGG-16	51.66	75.47	83.72
×	✓	✓	×	×	×	✓	VGG-16	46.32	69.84	78.53
×	✓	×	✓	×	×	✓	VGG-16	47.02	70.96	79.65
×	✓	×	×	✓	×	✓	VGG-16	42.08	67.81	76.86
×	✓	✓	✓	×	×	✓	VGG-16	50.25	74.69	82.88
×	✓	✓	×	✓	×	✓	VGG-16	49.96	73.21	81.60
×	✓	×	✓	✓	×	✓	VGG-16	52.49	75.46	83.56
×	✓	✓	✓	✓	×	✓	VGG-16	53.64	75.88	83.92
✓	✓	✓	×	×	×	✓	VGG-16	47.01	70.01	78.82
✓	✓	×	✓	×	×	✓	VGG-16	47.14	71.28	79.95
✓	✓	×	×	✓	×	✓	VGG-16	42.14	67.58	77.89
✓	✓	✓	✓	×	×	✓	VGG-16	51.62	74.89	83.01
✓	✓	✓	×	✓	×	✓	VGG-16	50.09	74.39	81.77
✓	✓	×	✓	✓	×	✓	VGG-16	53.74	75.76	83.52
✓	✓	✓	✓	✓	×	✓	VGG-16	**54.32**	**75.93**	**84.21**

4.2.2 *Combination of granularities*

We provide an analysis of the effect of each granularity and the way they are combined. Tables 1 and 2 show the results of models trained with varied combinations, while Table 3 shows performances of a fully trained IMG-Net model tested with diverse granularities. "GS," "LS," and "GP" denote the global2sentence, local2sentence, or global2phrase granularity, respectively, and

Table 2 Ablation analysis of the IMG-Net model (ResNet-50 backbone).

Channel	Spatial	GS	LS	GP	BN	GN	Backbone	Top-1	Top-5	Top-10
×	×	✓	✓	✓	✓	×	ResNet-50	49.91	70.83	80.00
✓	✓	✓	✓	✓	✓	×	ResNet-50	53.59	73.88	82.51
×	×	✓	×	×	×	✓	ResNet-50	45.66	68.36	77.43
×	×	×	✓	×	×	✓	ResNet-50	46.98	69.97	79.88
×	×	×	×	✓	×	✓	ResNet-50	41.28	66.11	78.82
×	×	✓	✓	×	×	✓	ResNet-50	51.01	73.91	81.84
×	×	✓	×	✓	×	✓	ResNet-50	50.34	71.96	82.30
×	×	×	✓	✓	×	✓	ResNet-50	51.38	74.40	84.00
×	×	✓	✓	✓	×	✓	ResNet-50	52.42	76.06	84.94
✓	×	✓	×	×	×	✓	ResNet-50	48.23	70.29	79.46
✓	×	×	✓	×	×	✓	ResNet-50	47.43	71.23	79.99
✓	×	×	×	✓	×	✓	ResNet-50	43.81	68.11	78.45
✓	×	✓	✓	×	×	✓	ResNet-50	51.39	74.80	82.91
✓	×	✓	×	✓	×	✓	ResNet-50	50.97	73.49	82.85
✓	×	×	✓	✓	×	✓	ResNet-50	52.82	75.91	84.23
✓	×	✓	✓	✓	×	✓	ResNet-50	53.69	76.88	84.93
×	✓	✓	×	×	×	✓	ResNet-50	48.72	70.94	79.32
×	✓	×	✓	×	×	✓	ResNet-50	49.23	72.00	80.33
×	✓	×	×	✓	×	✓	ResNet-50	44.60	68.81	77.60
×	✓	✓	✓	×	×	✓	ResNet-50	52.56	75.66	86.54
×	✓	✓	×	✓	×	✓	ResNet-50	51.90	74.33	82.63
×	✓	×	✓	✓	×	✓	ResNet-50	53.71	76.58	84.26
×	✓	✓	✓	✓	×	✓	ResNet-50	54.87	77.19	84.70
✓	✓	✓	×	×	×	✓	ResNet-50	49.53	71.21	79.66
✓	✓	×	✓	×	×	✓	ResNet-50	49.40	72.17	81.00
✓	✓	×	×	✓	×	✓	ResNet-50	44.14	68.47	78.56
✓	✓	✓	✓	×	×	✓	ResNet-50	53.66	76.13	83.61
✓	✓	✓	×	✓	×	✓	ResNet-50	52.11	75.77	82.54
✓	✓	×	✓	✓	×	✓	ResNet-50	55.00	76.74	84.45
✓	✓	✓	✓	✓	×	✓	ResNet-50	**56.48**	**76.89**	**85.01**

are utilized while training and testing the model. The results show that either training or testing with more than one single granularity brings performance gain, which indicates that the multigranular cross-modal matching can provide more comprehensive information, hence leading to a more accurate retrieval. Specifically, as for the three combinations that combine two granularities, the one that combines local2sentence and global2phrase outperforms the other two,

Table 3 Analysis of similarity combination while testing.

Method	Backbone	Top-1	Top-5	Top-10
S_{GS}	VGG-16	49.97	73.94	82.62
S_{LS}	VGG-16	49.81	74.27	82.46
S_{GP}	VGG-16	45.00	69.88	78.95
S_{GS+LS}	VGG-16	53.09	75.44	83.59
S_{GS+GP}	VGG-16	51.69	74.76	83.01
S_{LS+GP}	VGG-16	52.99	75.89	83.85
S_{ALL}	VGG-16	**54.32**	**75.93**	**84.21**
S_{GS}	ResNet-50	51.03	74.98	83.33
S_{LS}	ResNet-50	50.88	75.40	83.19
S_{GP}	ResNet-50	46.08	71.00	80.12
S_{GS+LS}	ResNet-50	55.99	76.77	84.92
S_{GS+GP}	ResNet-50	54.22	76.01	84.44
S_{LS+GP}	ResNet-50	55.64	76.61	84.70
S_{ALL}	ResNet-50	**56.48**	**76.89**	**85.01**

which proves that matching according to the crucial components while excluding the irrelevant ones can perform better than coarsely taking the whole global context into consideration. Thereby, the full IMG-Net model that employs both the coarse global and the fine-grained local information undoubtedly outperforms any other model that utilizes part of the three granularities.

4.2.3 *Group normalization*

We replace all GN layers in the IMG-Net model by BN layers to prove the strength of GN in the description-based person Re-ID task. Tables 1 and 2 show that GN outperforms BN with the same experiment setting by a nontrivial margin. "BN" means all GN layers in the model are replaced by BN layers, while others with "GN" still use GN.

As GN is proved to be able to improve the model performance by a nontrivial margin for description-based person Re-ID tasks, we carry out experiments to further study the performance of GN on the general single-modality person Re-ID task. As shown in Table 4, after all BN layers being replaced by GN layers, performance of methods on the general single-modality person Re-ID task are improved nontrivially, which proves that the effectiveness of GN in the general single-modality person Re-ID task.

Table 4 Ablation analysis of the GN on the general single-modality person Re-ID task on the Duke-MTMC dataset.

Method	BN-Rank-1	BN-mAP	GN-Rank-1	GN-mAP
PAN[44]	71.6	51.5	74.4	55.3
SVDNet[45]	76.7	56.8	81.2	58.7
Part-based convolutional baseline[5]	83.3	69.2	85.6	73.8
SNR[46]	84.4	72.9	87.0	74.2

Table 5 Ablation analysis of the partitioning number k.

k	Backbone	Top-1	Top-5	Top-10
2	VGG-16	52.89	74.69	82.48
4	VGG-16	53.21	75.19	83.33
6	VGG-16	**54.32**	**75.93**	**84.21**
8	VGG-16	54.17	75.86	84.13
10	VGG-16	53.96	75.77	83.97
12	VGG-16	53.42	75.29	84.08
2	ResNet-50	44.82	75.58	83.99
4	ResNet-50	55.51	76.17	84.66
6	ResNet-50	56.28	76.45	84.87
8	ResNet-50	**56.48**	**76.89**	**85.01**
10	ResNet-50	56.33	76.44	84.90
12	ResNet-50	55.80	76.23	84.78

4.2.4 *Partitioning number k*

As shown in Table 5, profound ablation experiments are carried out to analyze the effect of the partitioning number k. Initially, it can be observed that the performance of IMG-Net keeps improving with the increase of k before it reaches a peak. The optimum partitioning number k for IMG-Net with a VGG-16 backbone is 6, while the optimal k for IMG-Net with a ResNet-50 backbone is 8. Then the performance starts to get worse with k keeps growing. It is reasonable that the more stripes the image is partitioned into, the more local details can be caught by IMG-Net. Nevertheless, if partitioned into too many stripes, each stripe can be too small to provide useful information, some of which even can introduce noise.

4.2.5 *Fused pooling method*

Table 6 shows the comparison of different pooling methods. We find that the model with the maximum pooling method performs slightly better than the one with the average pooling method. This is reasonable, as the average pooling method takes all contextual information into consideration while downsampling, it may not perform properly if the discriminative signal is surrounded by unrelated signals. In contrast, the maximum pooling method catches the most salient signals for a local view. By fusing the two methods together to take advantage of both contextual information and the most salient signals, IMG-Net achieves the result better than model with either of them.

Table 6 Comparison of pooling methods.

Method	Backbone	Top-1	Top-5	Top-10
Average pooling	VGG-16	53.36	75.09	83.97
Max pooling	VGG-16	53.51	75.21	84.01
Fused pooling	VGG-16	**54.32**	**75.93**	**84.21**
Average pooling	ResNet-50	55.12	76.28	84.66
Max pooling	ResNet-50	55.74	76.44	84.78
Fused pooling	ResNet-50	**56.48**	**76.89**	**85.01**

Table 7 Comparison with the state-of-the-art.

Method	Backbone	Top-1	Top-5	Top-10
CNN-recurrent neural network (RNN)[47]	VGG-16	8.07	—	32.47
Neural talk[48]	VGG-16	13.66	—	41.72
GNA-RNN[3]	VGG-16	19.05	—	53.64
IATV[37]	VGG-16	25.94	—	60.48
PWM-ATH[12]	VGG-16	27.14	49.45	61.02
Dual path[49]	VGG-16	32.15	54.42	64.30
GALM[13]	VGG-16	47.82	69.83	78.31
MIA[14]	VGG-16	48.00	70.70	79.30
IMG-Net (ours)	VGG-16	**54.32**	**75.93**	**84.21**
Dual path[49]	ResNet-50	44.40	66.26	75.07
GLA[12]	ResNet-50	43.58	66.93	76.26
MIA[14]	ResNet-50	53.10	75.00	82.90
GALM[13]	ResNet-50	54.12	75.45	82.97
TIMAM[22]	ResNet-101	54.51	77.56	84.78
IMG-Net (ours)	ResNet-50	**56.48**	**76.89**	**85.01**

4.3 *Comparison with the State-of-the-Art*

Table 7 shows comparisons with other state-of-the-art methods. It can be observed that our proposed IMG-Net model outperforms all other state-of-the-art methods under top-1, top-5, and top-10 metrics. Patch-word matching model with adaptive threshold mechanism (PWM-ATH) proposes an efficient patch–word matching model to capture the local similarity between image and text but ignores the global–local relations. IMG-Net outperforms PWM-ATH by over 27% under top-1 metric, which validates the significance of the local2sentence and global2phrase granularities in our method. Compared with the best competitor MIA using VGG-16 as visual backbone, the IMG-Net model significantly outperforms it by 6.32% under top-1 metric, indicating the superiorities of the combination of inner-modal self-attention and cross-modal hard-region attention. Pose-guided joint global and attentive local matching network (GALM) utilizes pose information to help localize the discriminative regions. With either VGG-16 or ResNet-50 as visual backbone, IMG-Net outperforms GALM by a margin without suffering from the deviations of the pose estimation and the large computation consumption, which proves the effectiveness of inner-modal self-attention and GN without utilizing extra pretrained cues (e.g., pose).

4.4 *Discussion and Future Work*

Currently, the major challenge in description-based person Re-ID is still about how to effectively extract discriminative feature vectors from both image and text modalities. Due to the modality heterogeneity, information carried by feature vectors from the two different modalities may not be aligned perfectly. Thus, many previous methods[12,13,14] attempt to extract fine-grained information from both visual and textual modalities to provide more detailed information for the following matching step. By employing the inner-modal self-attention module, IMG-Net addresses the within-part consistency broken problem while extracting local visual cues and achieves state-of-the-art performance on the CUHK-PEDES dataset.

To better address the above-mentioned problems, we consider further improving our method in the future work. More specifically, we may refine our local feature extracting method. As for local visual information extraction, more detailed part feature can be extracted using the idea based on semantic soft segmentation, which may be more efficient than the commonly used strict uniform partition strategy.

5 Conclusion

In this work, we design an IMG-Net to address the problems in the field of the description-based person Re-ID, which incorporates inner-modal self-attention and cross-modal hard-region attention with the fine-grained model for extracting the multigranular semantic information. Specifically, the inner-modal self-attention module employs a channel-wise attention and a spatial-wise attention, which is proposed to address the within-part consistency broken problem. Then, the multigranular feature extraction module extracts rich local/global visual and textual features with the help of GN. Finally, a cross-modal hard-region attention module provides the local visual representation and phrase representation. Furthermore, we evaluate our approach on the CUHK-PEDES dataset and the results indicate that the proposed IMG-Net improves the performance with a large margin.

Acknowledgments

This work was partially supported the National Natural Science Foundation of China (Grant No. 61503017), China Postdoctoral Science Foundation (Grant No. 2019M661999), and the Natural Science Foundation of Jiangsu Higher Education Institutions of China (Grant No. 19KJB520009).

References

1. Y.-J. Cho and K.-J. Yoon, "PaMM: pose-aware multi-shot matching for improving person re-identification," *IEEE Trans. Image Process.* **27**(8), 3739–3752 (2018).
2. J. Dai et al., "Video person re-identification by temporal residual learning," *IEEE Trans. Image Process.* **28**(3), 1366–1377 (2019).
3. L. Shuang et al., "Person search with natural language description," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 5187–5196 (2017).
4. C. Song et al., "Mask-guided contrastive attention model for person re-identification," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 1179–1188 (2018).
5. Y. Sun et al., "Beyond part models: person retrieval with refined part pooling (and a strong convolutional baseline)," *Lect. Notes Comput. Sci.* **11208**, 501–518 (2018).
6. H. Yao et al., "Deep representation learning with part loss for person re-identification," *IEEE Trans. Image Process.* **28**(6), 2860–2871 (2019).
7. X. Zhang et al., "Background-modeling-based adaptive prediction for surveillance video coding," *IEEE Trans. Image Process.* **23**(2), 769–784 (2014).
8. R. Hou et al., "VRSTC: occlusion-free video person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vision and Pattern Recognit.*, pp. 7183–7192 (2019).
9. A. Subramaniam, A. Nambiar, and A. Mittal, "Co-segmentation inspired attention networks for video-based person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vision*, pp. 562–572 (2019).
10. J. Dai et al., "Video person re-identification by temporal residual learning," *IEEE Trans. Image Process.* **28**(3), 1366–1377 (2019).
11. Z. Liu, D. Wang, and H. Lu, "Stepwise metric promotion for unsupervised video person re-identification," in *Proc. IEEE Int. Conf. Comput. Vision*, pp. 2429–2438 (2017).
12. T. Chen, C. Xu, and J. Luo, "Improving text-based person search by spatial matching and adaptive threshold," in *IEEE Winter Conf. Appl. Comput. Vision*, pp. 1879–1887 (2018).

13. Y. Jing et al., "Pose-guided joint global and attentive local matching network for text-based person search," in *Association for the Advance of Artificial Intelligence (AAAI)* (2020).
14. K. Niu et al., "Improving description-based person re-identification by multi-granularity image–text alignments," in *IEEE Trans. Image Process.* Vol. **29**, pp. 5542–5556 (2020).
15. M. Ye et al., "Specific person retrieval via incomplete text description," in *Proc. 5th ACM Int. Conf. Multimedia Retrieval*, pp. 547–550 (2015).
16. Y. Wu and K. He, "Group normalization," in *Eur. Conf. Comput. Vision* (2018).
17. D. Chang et al., "Person re-identification by multi-channel parts-based CNN with improved triplet loss function," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 1335–1344 (2016).
18. C. Su et al., "Pose-driven deep convolutional model for person re-identification," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 3960–3969 (2017).
19. J. Liu et al., "Pose transferrable person re-identification," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 4099–4108 (2018).
20. Z. Zhong et al., "Invariance matters: exemplar memory for domain adaptive person re-identification," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 598–607 (2019).
21. Y. Sun et al., "Perceive where to focus: learning visibility-aware part-level features for partial person re-identification," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 393–402 (2019).
22. N. Sarafianos, X. Xu, and I. A. Kakadiaris, "Adversarial representation learning for text-to-image matching," in *IEEE/CVF Int. Conf. Comput. Vision* (2019).
23. D. Chen et al., "Improving deep visual representation for person re-identification by global and local image-language association," *Lect. Notes Comput. Sci.* **11220**, 54–70 (2018).
24. Z. Lin et al., "A structured self-attentive sentence embedding," in *Int. Conf. Learn. Represent.* (2017).
25. K. Xu et al., "Show, attend and tell: neural image caption generation with visual attention," in *Int. Conf. Mach. Learn.*, pp. 2048–2057 (2015).
26. T. Wang et al., "Generative neural networks for anomaly detection in crowded scenes," *IEEE Trans. Inf. Forensics Secur.* **14**(5), 1390–1399 (2019).
27. T. Wang et al., "Data-driven prognostic method based on self-supervised learning approaches for fault detection," *J. Intell. Manuf.* 1–9 (2018).
28. T. Wang et al., "A reinforcement learning approach for UAV target searching and tracking," *Multimedia Tools Appl.* **78**, 4347–4364 (2019).
29. T. Wang and H. Snoussi, "Detection of abnormal visual events via global optical flow orientation histogram," *IEEE Trans. Inf. Forensics Secur.* **9**(6), 988–998 (2014).
30. C. Cao et al., "Look and think twice: capturing top-down visual attention with feedback convolutional neural networks," in *Proc. Int. Conf. Comput. Vision* (2016).
31. L. C. Chen et al., "Attention to scale: scale-aware semantic image segmentation," in *IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 1879–1887 (2016).
32. H. Li et al., "Pyramid attention network for semantic segmentation," in *Br. Mach. Vision Conf.* (2018).
33. D. Yoo et al., "AttentionNet: aggregating weak directions for accurate object detection," in *Proc. Int. Conf. Comput. Vision* (2015).
34. X. Wang et al., "Non-local neural networks," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 7794–7803 (2018).
35. X. Lan et al., "Deep reinforcement learning attention selection for person re-identification," in *Br. Mach. Vision Conf.* (2017).
36. D. Li et al., "Learning deep context-aware features over body and latent parts for person re-identification," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 384–393 (2017).
37. S. Li et al., "Identity-aware textual-visual matching with latent co-attention," in *Proc. IEEE Int. Conf. Comput. Vision*, pp. 1890–1899 (2017).
38. W. Li, X. Zhu, and S. Gong, "Harmonious attention network for person re-identification," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 2285–2294 (2018).
39. M. Jiang, Y. Yuan, and Q. Wang, "Self-attention learning for person re-identification," in *Br. Mach. Vision Conf.* (2018).

40. M. Ye et al., "Deep learning for person re-identification: a survey and outlook," arXiv: 2001.04193 (2020).
41. S. Li et al., "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.* (2017).
42. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *ICRL* (2015).
43. K. He et al., "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.* (2015).
44. Z. Zheng, L. Zheng, and Y. Yang, "Pedestrian alignment network for large-scale person re-identification," *IEEE Trans. Circuits Syst. Video Technol.* **29**(10), 3037–3045 (2019).
45. Y. Sun et al., "SVDNet for pedestrian retrieval," in *Proc. IEEE Int. Conf. Comput. Vision*, pp. 3800–3808 (2017).
46. X. Jin et al., "Style normalization and restitution for generalizable person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vision and Pattern Recognit.*, pp. 3143–3152 (2020).
47. S. Reed et al., "Learning deep representations of fine-grained visual descriptions," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 49–58 (2016).
48. O. Vinyals et al., "Show and tell: a neural image caption generator," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 3156–3164 (2015).
49. Z. Zheng et al., "Dual-path convolutional image–text embedding with instance loss," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, arXiv:1711.05535 (2017).

Zijie Wang is an undergraduate at the School of Computer Science and Technology, Nanjing Tech University. He majors in computer science and technology. His current research interests include person Re-ID, human pose estimation, and human action recognition.

Aichun Zhu received his MS degree from China University of Mining and Technology, China, in 2012, and his PhD from the University of Technology of Troyes, France, in 2016. He is an assistant professor at the School of Computer Science and Technology, Nanjing Tech University, Nanjing, China. His research interests include computer vision, machine learning, person Re-ID, human pose estimation, and action recognition.

Zhe Zheng is an undergraduate student at the School of Computer Science and Technology, Nanjing Tech University. He majors in computer science and technology. His current research interests include dense crowd counting.

Jing Jin received her BE degree from the School of Electronic Science and Engineering at Nanjing University, China, in 2009, and her PhD from the School of Electronic Science and Engineering at Nanjing University, China, in 2016. She is an assistant professor at the School of Computer Science and Technology, Nanjing Tech University, Nanjing, China. Her research interests include pattern recognition and computer vision.

Zhouxin Xue is an undergraduate student at the School of Computer Science and Technology, Nanjing Tech University. He majors in software engineering. His current research interests include machine learning and computer vision.

Gang Hua received his BS degree from Southeast University, China, in 1984. He received his MS and PhD degrees from China University of Mining and Technology, China, in 1992 and 2002, respectively. He is currently a professor at the School of Information and Control, China University of Mining and Technology, Xuzhou, China. His research interests include the control and supervision of mining safety, signal processing, compressed sensing, and computer vision.