TIPCB: A Simple but Effective Part-based Convolutional Baseline for Text-based Person Search

Yuhao Chen[a], Guoqing Zhang[a,c,*], Yujiang Lu[a], Zhenxing Wang[b], Yuhui Zheng[a], Ruili Wang[c]

[a]*School of Computer and Software, Nanjing University of Information Science and Technology, Nanjing, 210044, China*
[b]*School of Mathematics and Statistics, Nanjing University of Information Science and Technology, Nanjing, 210044, China*
[c]*Institute of Natural and Mathematical Sciences, Massey University, Auckland, 4442, New Zealand*

Abstract

Text-based person search is a sub-task in the field of image retrieval, which aims to retrieve target person images according to a given textual description. The significant feature gap between two modalities makes this task very challenging. Many existing methods attempt to utilize local alignment to address this problem in the fine-grained level. However, most relevant methods introduce additional models or complicated training and evaluation strategies, which are hard to use in realistic scenarios. In order to facilitate the practical application, we propose a simple but effective end-to-end learning framework for text-based person search named **TIPCB** (i.e., **T**ext-**I**mage **P**art-based **C**onvolutional **B**aseline). Firstly, a novel dual-path local alignment network structure is proposed to extract visual and textual local representations, in which images are segmented horizontally and texts are aligned adaptively. Then, we propose a multi-stage cross-modal matching strategy, which eliminates the modality gap from three feature levels, including low level, local level and global level. Extensive experiments are conducted on the widely-used benchmark dataset (CUHK-PEDES) and verify that our method outperforms the state-

*Corresponding author
Email addresses: chinayhchen@gmail.com (Yuhao Chen), xiayang14551@163.com (Guoqing Zhang), yujiang_lu@163.com (Yujiang Lu), wangzx13142021@163.com (Zhenxing Wang), zheng_yuhui@nuist.edu.cn (Yuhui Zheng), Ruili.wang@massey.ac.nz (Ruili Wang)

of-the-art methods by **3.69%**, **2.95%** and **2.31%** in terms of Top-1, Top-5 and Top-10. Our code has been released in https://github.com/OrangeYHChen/TIPCB.

Keywords: Cross-modality, Person serach, Local representation

1. Introduction

Person search is a key technology in the field of image retrieval, aiming to find target person images from large databases with given retrieval conditions, including person images, relevant attributes or natural language descriptions. According to the modality of the query, this technology can be broadly divided into image-based search [1, 2, 3], attribute-based search [4, 5, 6] and text-based search [7, 8, 9]. In recent years, person search has gained increasing attention due to its wide applications in public security and video surveillance, e.g., searching for suspects and missing persons.

In this paper, we research on the task of text-based person search, as is illustrated in Figure 1. Specifically, it is required to sort all person images in a large gallery according to their similarity with the textual description of the query, and select the top person images as matching items [7]. Since textual descriptions are much more natural and accessible as retrieval queries, text-based person search has large potential values in conditions without target person images, e.g., searching for a suspect according to the description of the eyewitness.

Text-based person search is still a challenging task because it has the difficulties of both person re-identification and cross-modal retrieval. On the one hand, it is hard to extract robust visual representations due to the disturbance from occlusion, background clutter and pose/viewpoint variances. On the other hand, some images or descriptions of different persons have very similar high-level semantics, while the domains of image and text have significant differences, resulting in inter-modal feature variances much larger than intra-modal feature variances.

Therefore, a series of relevant methods have been proposed to reduce the

Figure 1: Illustration of text-based person search problem. Given a textual description, it is aimed to retrieve the corresponding person images from a large gallery database. Since some key information is hidden in the local details, matching local features is necessary for the enhancement of compatibility between person images and textual descriptions.

gap between image domain and text domain in recent years. We broadly categorize them into global-matching methods and local-matching methods. Global-matching methods mainly focus on the global visual and textual representation learning and obtain the unified feature space regardless of modality [7, 8, 10, 11, 12, 13, 14, 15]. However, images contain many distinctive local details which are hard to explore by global representation extraction. Besides, there are a few irrelevant regions in images, which bring noises to global information. In order to further mine discriminative and comprehensive information, some local-matching methods are proposed, which match person images and textual descriptions by local alignment [9, 16, 17, 18, 19, 20, 21].

However, most of the existing local-matching methods are not practical enough to meet the requirements of realistic scenarios due to their high complexity. Some of these methods introduce additional models or apply multi-task learning strategies, such as human pose estimation [17, 22, 23], semantic segmentation [18, 24] or attribute recognition [21, 25], which bring a huge amount of computation and make networks unable to perform end-to-end learning. Be-

sides, some of these methods adopt the multi-granularities similarity measure strategy [16, 19]. In the use phase, these method need to learn multiple local representations for each image or text, and repeatedly calculate the local similarity. Both additional models and complex similarity measure are time-consuming for practical applications. Thus, it is necessary to design a simple but effective framework for text-based person search problem.

In this paper, we propose a novel end-to-end learning framework named **TIPCB** (i.e., **T**ext-**I**mage **P**art-based **C**onvolutional **B**aseline) to facilitate the practical application. Firstly, a novel dual-path local alignment network structure is proposed to extract visual and textual local representations. Visual local representations are extracted by general PCB strategy [2], in which person images are horizontally segmented into several stripes. In the textual representation learning path, the word embeddings are learnt by a BERT model with pretrained and fixed parameters [26], and further processed by a multi-branch residual network. In each branch, the textual representation is learnt to adaptively match a corresponding visual local representation, so as to extract the aligned textual local representation. Besides, a multi-stage cross-modal matching strategy is proposed, which eliminates the modality gap from low-level, local-level and global-level features, and then the feature gap between image domain and text domain can be reduced in a progressive way.

The main contributions of this paper can be summarized as follows:

- A novel dual-path local alignment network is proposed to jointly learn visual and textual representations, which can align local features in a simple but effective way.
- A multi-stage cross-modal matching strategy is designed to reduce the gap between two modalities in a progressive way. The whole framework can be trained in end-to-end manner.
- Extensive experiments are conducted on CUHK-PEDES dataset [7], and the results clearly verify that our proposed TIPCB framework achieves the state-of-the-art performance.

2. Related Work

2.1. Person Re-identification

In the past decade, a variety of person re-identification (image-based person search) methods have sprung up, attempting to extract discriminative and robust representations for person images, and overcome the difficulties of occlusion, background clutter and so on [27, 28]. Among them, local representation learning is an efficient and widely-used technology, which can explore some detailed and distinctive features. Specifically, PCB [2] applies horizontal segmentation on the feature map of each person image, and extract local representations for obtained stripes independently. MGN [29] adopts a multi-granularities representation learning strategy which cuts each feature map into several parts with different scales. Besides, spatial attention mechanism is introduced to align human parts and further improves the robustness of local representation [30]. Furthermore, some methods utilize additional models to assist local segmentation and mine detailed information, such as pose estimation [31, 32] and human semantic segmentation [33].

Extensive works have also achieved great progress in video-based re-ID [34, 35, 36], occluded re-ID [37], unsupervised re-ID [38, 39, 40], cross-resolution re-ID [41, 42, 43], RGB-IR cross-modality re-ID [44, 45, 46] and so on. Person re-identification has obtained rapid development, but this technology cannot be applied in the scenario without query images.

2.2. Text-based Person Search

A series of high-performance methods have been proposed for text-based person search in recent years, which can be broadly categorized into global-matching methods and local-matching methods. GNA-RNN [7] is the first designed framework for this task, which combines a recurrent neural network with the proposed gated neural attention mechanism to learn affinities between textual descriptions and person images. Later, an identity-aware two-stage network [13] is proposed to jointly minimize the intra-identity distance and the

cross-modal distance. Besides, in order to embed images and texts to a shared visual-textual space, Zheng et al. [12] design an end-to-end trainable model with a CNN+CNN dual-path structure, and Zhang et al. [14] introduce the cross-modal projection learning in objective functions. Furthermore, TIMAM [15] attempts to learn modality-invariant representations in a shared space by adversarial learning. However, the above methods only focus on global representations, which may miss some distinctive local details or mix a little noise information.

Therefore, some local-matching methods are explored to overcome this shortcoming. Aggarwal et al. [21] introduce human attribute recognition to help bridge the modality gap between the image-text inputs. Wang et al. [18] design a light auxiliary attribute segmentation layer to guide the alignment between visual local representations with parsed textual attributes. Jing et al. [17] propose a multi-granularities attention network to align visual and textual local representations with the aid of the human pose information. These methods apply additional models or multi-task learning strategies to enhance the local alignment but bring a huge amount of computation. In addition, MIA [19] aligns the local representations from multiple granularities, including global-global, global-local and local-local levels. NAFS [16] conducts joint alignments over full-scale representations with a novel staircase CNN and a locality-constrained BERT. Nevertheless, such methods are still time-consuming in the use phase due to their complex similarity measure strategies. By comparison, our proposed method utilizes an end-to-end trainable dual-path network to learn local aligned representations simply and effectively.

3. Proposed Method

In this section, we will detailly explain our proposed Text-Image Part-based Convolutional Baseline (TIPCB) for text-based person search problem. We first illustrate the dual-path local alignment network structure, including the visual CNN branch and the textual CNN branch, and then the multi-stage cross-modal

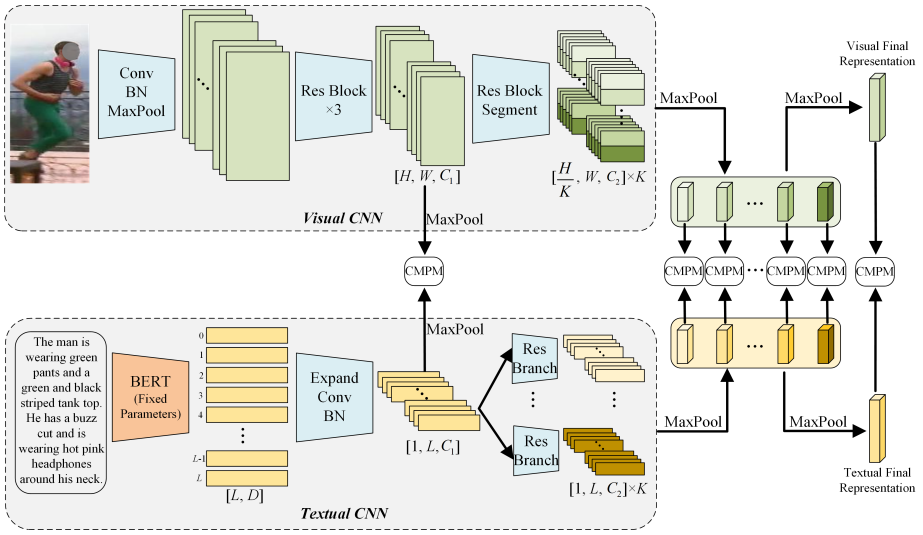

Figure 2: The architecture of the proposed TIPCB. This framework consists of a dual-path local alignment structure, where the visual CNN branch applies the PCB after the backbone network, and the textual CNN branch applies a multi-branch residual network after a pre-trained BERT model. This framework adopts a multi-stage cross-modal matching strategy, which conducts projection matching on low-level, local-level and global-level representations.

matching strategy is introduced to eliminate the modality gap.

3.1. Visual Representation Learning

As illustrated in Figure 2, our proposed TIPCB contains two CNN branches which aim to learn discriminative and compatible visual and textual representations from the input person images and descriptions, respectively. In the training phase, we assume the training data as $D=\{I_i, T_i\}_{i=1}^{N}$ where N represents the number of image-text pair in each batch, and each pair consists of an image I and a corresponding description T. (The subscript i is omitted in the following for simplicity unless necessary.) In visual CNN branch, ResNet-50 [47] is adopted as the backbone to extract visual features, which mainly consists of four residual blocks. Different residual blocks can capture semantic information from different level [48]. For each image I, we define the feature generated by the 3$^{\text{rd}}$ and 4$^{\text{th}}$ residual block as its low-level feature map $f_l^I \in \mathbb{R}^{H\times W\times C_1}$ and

high-level feature map $f_h^I \in \mathbb{R}^{H \times W \times C_2}$, where H, W and C_1/C_2 represent the dimension of height, width and channel in the above feature maps. Then we obtain its visual low-level representation $v_l^I \in \mathbb{R}^{C_1}$ by:

$$v_l^I = \mathrm{GMP}(f_l^I) \tag{1}$$

where GMP represents a global max-pooling layer as a filter to mine salient information.

Here, we adopt the PCB [2] strategy to obtain the local regions. Specially, the high-level feature map f_h^I is segmented into K horizontal stripes which are denoted as $\left\{f_{p1}^I, f_{p2}^I, ..., f_{pK}^I\right\}$ where $f_{pi}^I \in \mathbb{R}^{\frac{H}{K} \times W \times C_2}$. For each stripe, similar to Formula (1), we still adopt a global max-pooling layer to extract the visual local representation $v_{pi}^I \in \mathbb{R}^{C_2}$. In order to fuse all local representations, we select the maximum value of each element in channel dimension, and get the visual global representation $v_g^I \in \mathbb{R}^{C_2}$:

$$v_g^I = \mathrm{Max}\left(v_{p1}^I, v_{p2}^I, ..., v_{pK}^I\right) \tag{2}$$

Therefore, we get the visual feature set $V^I = \left\{v_l^I, v_{p1}^I, ..., v_{pK}^I, v_g^I\right\}$ containing low-level, local-level and global-level representations. In the testing phase, only the global-level representation is adopted to measure similarity.

3.2. Textual Representation Learning

In textual CNN branch, a high-performance language representation model BERT [26] is applied to extract discriminative word embeddings, which can learn the contextual relations between the words by the bi-directional training of Transformer [49]. Specifically, we break each textual description T up into a list of words, and insert [CLS] and [SEP] into the beginning and end of each sentence. Then this list is embedded into tokens by a pretrained tokenizer. To ensure the consistency of text length, we select the first L tokens when the text is longer than L, and apply zero-padding in the end of the text when the text is shorter than L. After that, each tokenized textual description is input into the BERT model, which is pretrained and parameter-fixed, to extract word

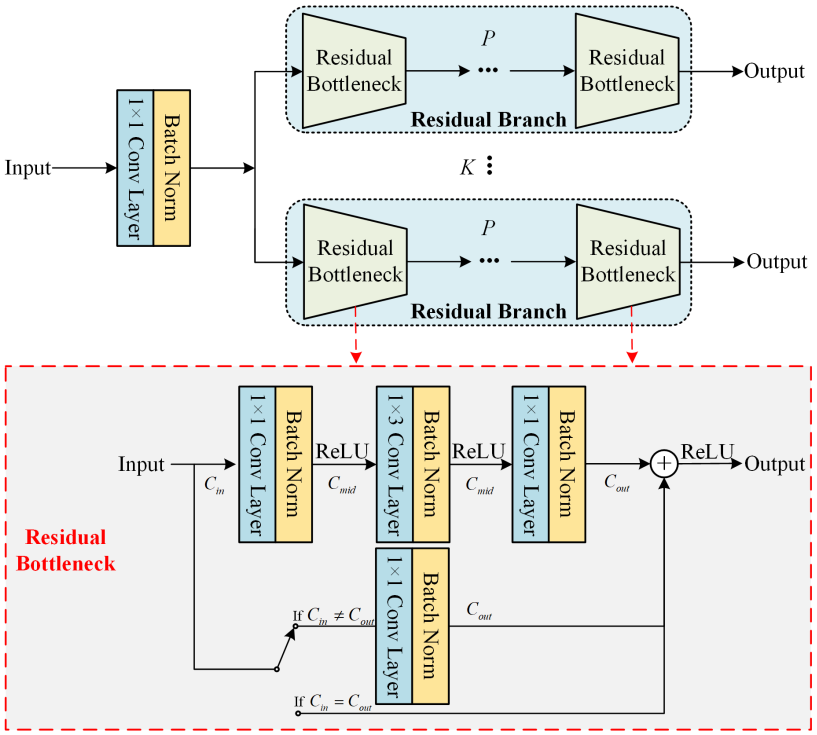

Figure 3: The details of the multi-branch textual CNN.

embeddings $t \in \mathbb{R}^{L \times D}$, where D represents the dimension of each word embedding. Here we "freeze" the weights of the BERT model for the following three reasons: 1) the pretrained BERT itself has the strong semantic representation ability, and we only use it as the word embedding layer, 2) the following CNN structure is capable to further process the word embeddings, and 3) only training the CNN structure can significantly reduce the amount of training parameters and accelerate the convergence of the model.

In order to meet the inputting requirement of convolutional layers, we expand the dimension of word embedding from $t \in \mathbb{R}^{L \times D}$ to $t^* \in \mathbb{R}^{1 \times L \times D}$, where 1, L and D are regarded as the height, width and channel dimension of convolutional input, respectively. Motivated by the residual network [47] and the deep textual CNN [12], we design the multi-branch textual CNN, as shown in Figure 3. In the textual CNN, in order to map the word embeddings to the same channel

dimension as the visual low-level feature map $f_l^I \in \mathbb{R}^{H \times W \times C_1}$, the filter size of the first convolutional layer is set to $1 \times 1 \times D \times C_1$, which can be viewed as a lookup table. Then, we can obtain the textual low-level feature map $f_l^T \in \mathbb{R}^{1 \times L \times C_1}$.

The multi-branch textual CNN contains K residual branches, which correspond to the K stripes of person images. For each branch, it contains P textual residual bottlenecks and is aimed to adaptively learn the textual representations which can match the visual local representations. The textual residual bottleneck has the similar structure as the modules in ResNet, consisting of several convolutional layers and batch normalization layers. The skip connection is applied to transmit information from low layers to high layers, which can effectively restrain the network degradation problem and speed up the model training. Specifically, in order to keep the textual information uncompressed, the strides of all convolutional layers in bottlenecks are set to 1×1. For the first bottleneck of each branch, we modify the channel dimension of the textual feature map to C_2, which is consistent with the visual high-level feature map $f_h^I \in \mathbb{R}^{H \times W \times C_2}$, and then we keep the channel dimension unchanged in the following bottlenecks. After the multi-branch textual CNN, we obtain the textual local feature maps. Similar to visual CNN branch, we adopt a global max-pooling layer to extract the textual local representations and select the maximum value of each element in channel dimension to fuse these local representations. Then, we get the textual feature set $V^T = \{v_l^T, v_{p1}^T, ..., v_{pK}^T, v_g^T\}$ containing low-level, local-level and global-level representations.

Different from the deep textual CNN [12], we only stack a few bottlenecks rather than use a very deep residual network to extract textual representations for the following two reasons: 1) the downsampling between different stages in deep textual CNN brings obvious information loss, and 2) deep network does not bring obvious improvement compared with shallow network, which is contrary to the experience in image area and has been proven in [50]. In our experiment section, we will further verify the above viewpoints.

3.3. Multi-stage Cross-modal Matching

In order to eliminate the feature gap between image modality and text modality, we adopt the Cross-Modal Projection Matching (CMPM) loss [14] on low-level, local-level and global-level representations, which can associate representations across different modalities by incorporating the cross-modal projection into KL divergence. For each visual representation v_i^I, we assume that the set of image-text representation pairs is $\left\{\left(v_i^I, v_j^T\right), y_{i,j}\right\}_{j=1}^N$, where $y_{i,j} = 1$ represents that v_i^I and v_j^T are from the same person, while $y_{i,j} = 0$ means that they are not a matched pair. (The subscript used to indicate the representation level is omitted, because it is applicable for any representation pair from V^I and V^T.)

The probability that v_i^I and v_j^T are a matched pair can be calculated by:

$$p_{i,j} = \frac{\exp\left(\left(v_i^I\right)^\top \bar{v}_j^T\right)}{\sum_{k=1}^{N} \exp\left(\left(v_i^I\right)^\top \bar{v}_k^T\right)} \tag{3}$$

where $\bar{v}_j^T$ is the normalized textual representation and denoted as $\bar{v}_j^T = \frac{v_j^T}{\|v_j^T\|}$. In CMPM, the scalar projection of v_i^I on v_j^T is regarded as their similarity, and matching probability $p_{i,j}$ is the proportion of the similarity between v_i^I and v_j^T to the sum of similarity between v_i^I and $\left\{v_j^T\right\}_{j=1}^N$ in a batch. Then the CMPM loss can be calculated by:

$$L_{I2T} = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{N} p_{i,j} \log\left(\frac{p_{i,j}}{q_{i,j} + \varepsilon}\right) \tag{4}$$

where ε is a small number to avoid numerical problems, and $q_{i,j}$ is the normalized true matching probability between v_i^I and v_j^T since there might be more than one matched text descriptions in a batch, denoted as $q_{i,j} = \frac{y_{i,j}}{\sum_{k=1}^{N} y_{i,k}}$. The above procedure reduces the distance between each visual representation and its matched textual representations in a single direction, and we reversely conduct the similar procedure to draw each textual representation and its matched visual representations closer. Therefore, the bi-directional CMPM loss is computed by:

$$L_{CMPM} = L_{I2T} + l_{T2I} \tag{5}$$

The objective in our framework contains the cross-modal representation matching from three levels. The CMPM loss in low-level representations is to reduce the modality gap in an early stage. The CMPM loss in local-level representations can realize the local alignment between images and texts. The CMPM loss in global-level representations ensure that the final representations for evaluation have the stronger modal compatibility. Through the multiple stages of CMPM loss, the matching degree of image-text representations can be gradually improved, which will be further verified in ablation study. Finally, according to the visual and textual representation sets V^I and V^T, the overall objective function is calculated by:

$$L = \lambda_1 L^l_{CMPM} + \lambda_2 \sum_{k=1}^{K} L^{pk}_{CMPM} + \lambda_3 L^g_{CMPM} \tag{6}$$

where λ_1, λ_2, λ_3 are hyper-parameters to control the importance of different CMPM losses, and L^l_{CMPM}, $\left\{L^{pk}_{CMPM}\right\}_{k=1}^{K}$, L^g_{CMPM} indicate the CMPM loss of low-level, local-level and global-level representations, respectively.

4. Experiment

4.1. Experimental Settings

Dataset. We evaluate our proposed TIPCB framework on the CUHK-PEDES dataset [7], which is the only large scale available benchmark for text-based person search problem, as is shown in Figure 4. It totally contains 40,206 person images of 13,003 identities by collecting samples from several re-ID datasets. Each person image has two corresponding textual descriptions on average, and each textual description has more than 23 words. These textual descriptions have a vocabulary of 9,408 different words. We adopt the same data split as [7]. The training set has 34,054 images of 11,003 identities. The validation and testing set have 3,078 images and 3,074 images of 1,000 identities, respectively.

Evaluation Protocol. We follow the standard evaluation metrics, and report the top-k ($k = 1, 5, 10$) accuracy to evaluate the performance. Specifically,

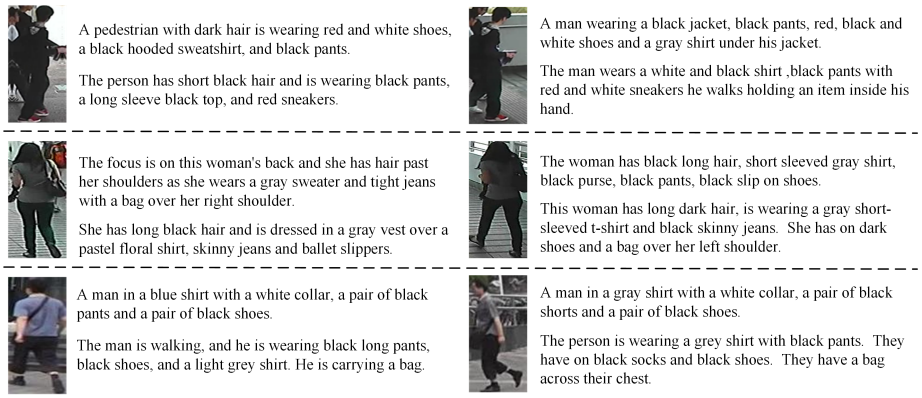

Figure 4: Samples in the CUHK-PEDES dataset. The person images in each line are of the same identity, and each image has two corresponding textual descriptions.

given a query text description, all gallery images are ranked according to their similarity values. A successful search means that a matched person image is existed among the top-k images.

Implementation Details. In the visual CNN branch, we adopt ResNet-50 pretrained on ImageNet [51] as the backbone to extract visual feature maps, and we modify the stride of $conv5\_1$ to 1 instead of 2 for larger feature maps. In the textual CNN branch, we extract word embeddings by the language model BERT-Base-Uncase pretrained on a large corpus including Toronto Book Corpus and Wikipedia. All the input images are resized to 384×128, and the text length is unified to $L = 64$. The number of local regions is set to K=6. In the multi-branch textual CNN, the number of bottlenecks in each residual branch is set to P= 3. In the visual and textual features, some parameters of dimensions are set to $H = 24$, $W = 8$, $D = 768$, $C_1 = 1024$ and $C_2 = 2048$. Each batch contains $N = 64$ image-text pairs.

In the training phase, Adam is selected to optimize our model with weight decay 4×10^{-5}. The model is trained for 80 epochs in total. The base learning rate is set to 3×10^{-3} and decreased by 0.1 after 50 epochs. Besides, we initialize the learning rate by the warm-up trick in first 10 epochs. We adopt the trick of horizontally flipping to augment data, where each image has 50% chance

to flip randomly. The hyper-parameters in the objective function are set to $\lambda_1 = 1$, $\lambda_2 = 1$ and $\lambda_3 = 1$. In the testing phase, the cosine distance is used to measure the similarity value of image-text pairs. We perform our experiments with PyTorch on a single Tesla V100 GPU.

4.2. Comparison with State-of-The-Art Methods

We compare our method with the existing text-based person search methods, and the comparable results are reported in Table 1. We divide these methods into two categories. Global-matching methods (type column is marked as "G") consist of GNA-RNN, IATV, Dual Path, CMPM + CMPC and TIMAM, and local-matching methods (type column is marked as "L") contain PWM + ATH, GLA, MIA, PMA, ViTAA, CMAAM and NAFS. We can find that the methods based on local alignment have become a hot topic and relatively achieved better performance in recent years, which can prove the importance of the local fine-grained alignment to strengthen the compatibility and discrimination of image-text features.

4.3. Ablation Studies

Effects of local representations. In order to verify the effectiveness of local representations, we conduct ablation studies to compare the performance of global representations and local representations of different region scales, and the results are listed in Figure 5(a). We can obtain the following two findings. On the one hand, compared with the global representations, the local representations have the much stronger discriminative ability, and the method with 6 local regions gains 3.25% improvement in Top-1 accuracy. One important reason is that global representations are hard to capture some distinctive local details. On the other hand, it can be observed that the Top-1 accuracy increases first and then decreases with the increase of the number of local regions, which can be explained by the local misalignment among person images. Due to the unstable vibration of person detection boxes and variances of viewpoints, the spatial distribution of head, body and limbs exists significant differences. We

Table 1: Comparison with state-of-the-art methods on CUHK-PEDES dataset. Top-1, Top-5 and Top-10 accuracies (%) are reported. The 1st, 2nd and 3rd top results are indicated by **red**, **blue** and **green** bold numbers, respectively. In the second column, "G" represents the methods only using global features, and "L" indicates the methods aligning local features. All listed methods do not use post-processing including re-ranking, to ensure the fairness of comparison.

Method	Type	Ref	Top-1	Top-5	Top-10
GNA-RNN [7]	G	CVPR17	19.05	-	53.64
IATV [17]	G	ICCV17	25.94	-	60.48
PWM+ATH [19]	L	WACV18	27.14	49.45	61.02
GLA [21]	L	ECCV18	43.58	66.93	76.26
Dual Path [15]	G	TOMM20	44.40	66.26	75.07
CMPM+CMPC [19]	G	ECCV18	49.37	-	79.27
MCCL [10]	G	ICASSP19	50.58	-	79.06
MIA [16]	L	TIP20	53.10	75.00	82.90
PMA [12]	L	AAAI20	53.81	73.54	81.23
ViTAA [13]	L	ECCV20	55.97	75.84	83.52
CMAAM [22]	L	WACV20	**56.68**	77.18	**84.86**
ResNet+BERT Structure					
TIMAM [20]	G	ICCV19	54.51	**77.56**	84.78
NAFS [11]	L	arXiv21	**59.94**	**79.86**	**86.70**
TIPCB (ours)	L	PR21	**63.63(↑3.69)**	**82.82(↑2.95)**	**89.01(↑2.31)**

apply the hard segmentation strategy based on PCB [2] in the visual CNN branch, which brings the noises in local region set. When the granularity of the local regions is too small, abundant noises in local region set will bring difficulties for the network to extract the common features of this region, as is shown in Figure 5(b).

Effects of shallow network for textual representation learning. Motivated by the conclusion that deep network does not bring obvious improvement for text classification compared with shallow network in [50], we design a much shallower network than the deep textual CNN [12] to extract textual representations from word embeddings. We conduct a series of ablation experiments to prove the superiority of shallow network in textual representation learning. As is shown in Figure 6(a), we test the multi-branch textual CNNs with dif-

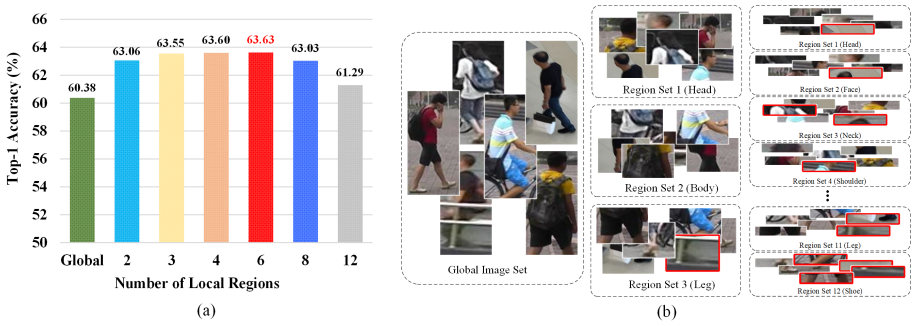

Figure 5: (a) Comparable results of different number of regions. Top-1 accuracy (%) is reported. (b) The region sets of different scales. The noises inside are marked by red bounding boxes.

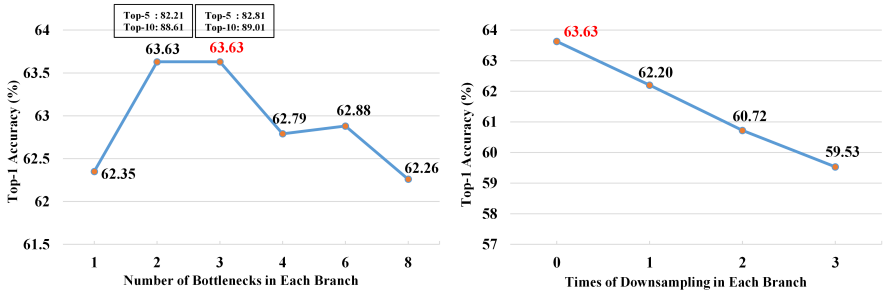

Figure 6: Comparable results of different number of bottlenecks (in Figure 6(a)) and different times of downsampling in each branch (in Figure 6(b)). Top-1 accuracy (%) is reported.

ferent number of bottlenecks in each branch. The result presents an overall trend of first increasing and then decreasing with the increase of the number of bottlenecks, and the network achieves the best performance when each branch has 3 bottlenecks. When it has only one residual bottleneck, the network has insufficient ability to extract discriminative features. Besides, the network for textual representation learning does not need too deep layers since text data is generally discrete and sparse, which is significantly different from image data. Therefore, our multi-branch textual CNN performs better in the condition that each branch has only 2$\sim$3 bottlenecks.

In addition, we keep the size of textual feature maps rather than use multi-stage downsampling in [12]. We compare the residual branches with different times of downsampling and the results are shown in Figure 6(b). Specifically, each residual branch has 3 residual bottlenecks by default, and the branch without downsampling outputs the feature map of $1 \times 64 \times 2048$. The spatial size of feature map will be reduced by half when each bottleneck applies the downsampling strategy. That is to say, the branch with 1, 2 and 3 times of downsampling output the feature maps of $1 \times 32 \times 2048$, $1 \times 16 \times 2048$ and $1 \times 8 \times 2048$, respectively. It can be observed that the Top-1 accuracy decreases significantly with the increase of times of downsampling. One of the important reasons is that the downsampling strategy brings obvious textual information loss. Therefore, we do not adopt downsampling and keep the size of textual feature maps unchanged in residual branches.

Effects of multi-stage cross-modal matching strategy. During the training phase, in order to stimulate the modality gap step by step, we apply a multi-stage cross-modal matching strategy, which applies the CMPM loss on the representations of three stages, including low-level and high-level representations. Note that both local-level and global-level representations belong to high-level representations. We conduct the following ablation experiments to verify the CMPM loss in each stage, and the results are reported in Table 2. (The accuracy of the method with only low-level CMPM loss is not listed here, because we only select the high-level representations during the testing phase, and the result will be not referential if the high-level CMPM loss is ignored.)

From the experimental results, we can observe the following findings:

1) According to the results of variants (a) and (b) which only use one stage of CMPM loss, we find that their Top-1 accuracies of more than 57% have already surpasses most of the existing methods except for NAFS. It proves the high efficiency of our dual-path local alignment network.

2) According to the results of variants (a), (b), (c) and (d), we observe that the addition of low-level CMPM loss brings 0.91% and 1.37% improvements in Top-1 accuracy for variants (a) and (b), respectively. It proves the positive

Table 2: Comparable results of combinations of CMPM loss in different stages. Top-1, Top-5 and Top-10 accuracies (%) are reported.

Variant	Low-level	Local-level	Global-level	Top-1	Top-5	Top-10
(a)		✓		57.16	78.03	84.94
(b)			✓	58.47	80.19	86.97
(c)	✓	✓		58.07	79.12	85.70
(d)	✓		✓	59.84	81.39	87.95
(e)		✓	✓	62.39	81.93	88.71
(f)	✓	✓	✓	**63.63**	**82.81**	**89.01**

effect of matching low-level representations, which can reduce the modality gap in an early stage.

3) From the variants (a), (b) and (e), it can be found that the combination of local-level and global-level CMPM losses has 5.23% and 3.92% improvements in Top-1 accuracy for single local-level and single global-level CMPM loss. The reason is that the local-level CMPM loss can realize the local alignment between images and texts, and the global-level CMPM loss can ensure that the final representations have the stronger modal compatibility. This group of experiments can simultaneously verify the validity of matching local-level and global-level representations.

4) From the variant (f), we can find that the combination of low-level, local-level and global-level CMPM losses achieves the best performance, which further shows the high efficiency of our proposed multi-stage cross-modal matching strategy.

Effects of fusion strategy. We conduct a series of ablation experiments to compare the performance of different fusion strategies, including avg-pooling, max-pooling and the addition of them, and the results are listed in Table 3. It can be observed that the accuracy of the strategy with max-pooling is far higher than the strategy with avg-pooling, and sightly higher than the strategy with the addition of them. Here, global max-pooling layer can filter salient information to extract more discriminative representation, while global avg-pooling may mix irrelevant noises in representations since it integrates each element in the feature

Table 3: Comparable results of different fusion strategies. Top-1, Top-5 and Top-10 accuracies (%) are reported.

Fusion Strategy	Top-1	Top-5	Top-10
Avg-pooling	55.63	76.59	84.14
Max-pooling	**63.63**	**82.81**	**89.01**
Max-pooling+Avg-pooling	63.22	82.73	88.21

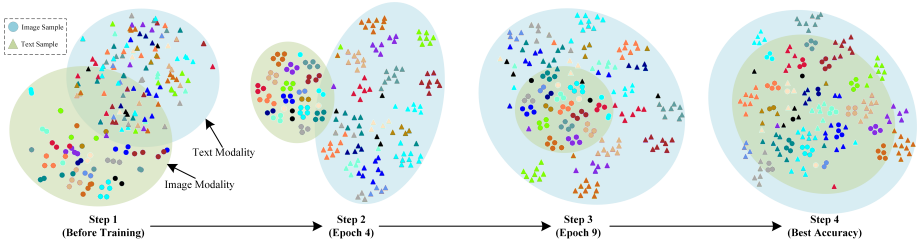

Figure 7: Visualization of features via t-SNE. We show the changing process of cross-modal feature distributions with training. There are shown 64 images with corresponding 128 textual descriptions. The feature of each image and text is marked as a circle and a rectangle, respectively. Each identity is indicated in a specific color.

map.

Visualizaton of features. We apply the t-SNE [52] to visualize the features in Figure 7 and show the changing process of feature distributions in four steps. Before training, there is a significant gap between text modality and image modality, and the distributions inside the modalities are disordered. After several training epochs, it can be observed that samples of the same identity begin to cluster, but the two modalities still have a large gap. Then, the distributions of two modalities begin to converge gradually until their centers are close. Finally, the feature distributions of two modalities coincide well to some extent, and the samples from the same identity can have a good clustering performance. This demonstrates that our TIPCB is capable to learn discriminative visual and textual representations, and eliminate the cross-modal distribution gap.

Sample analysis. As is shown in Figure 8, we visualize and analyze several examples of text-based person search by our proposed TIPCB. For each success-

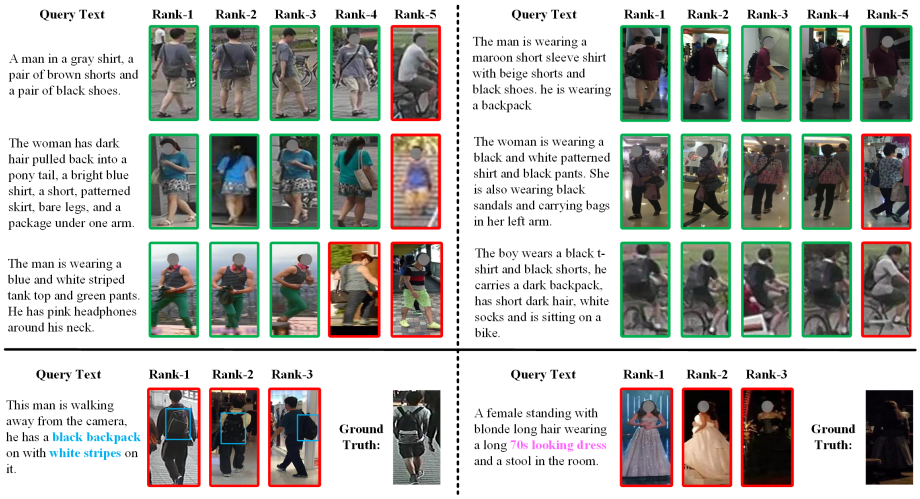

Figure 8: Visualization of text-based person search results by our proposed TIPCB. The green and red bounding boxes indicate correct and incorrect matches, respectively. The top 6 groups are successful searches, while the bottom 2 groups are failure searches.

ful case, we can find that the listed top-5 person images have multiple regions that can well match parts of the corresponding textual description. Note that our proposed TIPCB is capable to distinguish some hard samples, which only partly match the textual description. For example, in the case 6, our method successfully finds the perfect matched images which simultaneously meet the characteristics of "black short", "dark backpack", "short dark hair" and "sitting on a bike", and lists the image that only mismatches the condition of "black short" behind.

In terms of failure cases, we list the following two representative examples. In the first case, the textual description is too ambiguous, which only contains useful information about backpack. We can find that the top-3 person images all have the "black backpack", but they are not the matched ones. Besides, the detailed description of "white stripes" is too subtle to extract discriminative features from it. In the second case, the text uses "70s looking" to describe the dress, which is a rare phase and is difficult for network to learn. Thus, the top-3 person images have the dresses with different styles.

5. Conclusion

In this paper, in order to facilitate the practical application, we propose a simple but effective end-to-end learning framework for text-based person search named **TIPCB** (i.e., **T**ext-**I**mage **P**art-based **C**onvolutional **B**aseline). In contrast to the existing local-matching methods, TIPCB applies an end-to-end trainable structure without additional models and complex evaluation strategies. We design a novel dual-path local alignment network to learn visual and textual local representations, in which images are segmented horizontally and texts are aligned adaptively. Besides, we introduce a multi-stage cross-modal matching strategy to match the visual and textual representations from three levels and eliminate the modality gap step by step. The outstanding experimental results verify the superiority of our proposed TIPCB method.

Acknowledgment

This research is supported in part by the National Natural Science Foundation of China under Grant 61806099, U20B2065; and by the Natural Science Foundation of Jiangsu Province of China under Grant BK20180790.

References

[1] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, Q. Tian, Scalable person re-identification: A benchmark, in: The IEEE International Conference on Computer Vision (ICCV), 2015, pp.1116-1124.

[2] Y. Sun, L. Zheng, Y. Yang, Q. Tian, S. Wang, Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline), in: The European Conference on Computer Vision (ECCV), 2018, pp. 480–496.

[3] W. Zheng, S. Gong, T. Xiang, Person re-identification by probabilistic relative distance comparison, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2011, pp. 649–656.

[4] A. Schumann, R. Stiefelhagen, Person re-identification by deep learning attribute-complementary information, in: The IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2017, pp. 20–28.

[5] C. Su, S. Zhang, J. Xing, W. Gao, Q. Tian, Deep attributes driven multi-camera person re-identification, in: The European Conference on Computer Vision (ECCV), 2016, pp. 475–491.

[6] D. A. Vaquero, R. S. Feris, D. Tran, L. Brown, A. Hampapur, M. Turk, Attribute-based people search in surveillance environments, in: The IEEE Winter Conference on Applications of Computer Vision (WACV), 2009, pp. 1–8.

[7] S. Li, T. Xiao, H. Li, B. Zhou, D. Yue, X. Wang, Person search with natural language description, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1970–1979.

[8] K. Niu, Y. Huang, L. Wang, Fusing two directions in cross-domain adaption for real life person search by language, in: The IEEE International Conference on Computer Vision Workshops (ICCVW), 2019, pp. 0–0.

[9] T. Chen, C. Xu, J. Luo, Improving text-based person search by spatial matching and adaptive threshold, in: The IEEE Winter Conference on Applications of Computer Vision (WACV), 2018, pp. 1879–1887.

[10] Y. Wang, C. Bo, D. Wang, S. Wang, Y. Qi, H. Lu, Language person search with mutually connected classification loss, in: The IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019, pp. 2057–2061.

[11] J. Ge, G. Gao, Z. Liu, Visual-textual association with hardest and semi-hard negative pairs mining for person search, arXiv preprint arXiv:1912.03083.

[12] Z. Zheng, L. Zheng, M. Garrett, Y. Yang, M. Xu, Y.-D. Shen, Dual-path convolutional image-text embeddings with instance loss, ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) 16 (2) (2020) 1–23.

[13] S. Li, T. Xiao, H. Li, W. Yang, X. Wang, Identity-aware textual-visual matching with latent co-attention, in: The IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1890–1899.

[14] Y. Zhang, H. Lu, Deep cross-modal projection learning for image-text matching, in: The European Conference on Computer Vision (ECCV), 2018, pp. 686–701.

[15] N. Sarafianos, X. Xu, I. A. Kakadiaris, Adversarial representation learning for text-to-image matching, in: The IEEE International Conference on Computer Vision (ICCV), 2019, pp. 5814–5824.

[16] C. Gao, G. Cai, X. Jiang, F. Zheng, J. Zhang, Y. Gong, P. Peng, X. Guo, X. Sun, Contextual non-local alignment over full-scale representation for text-based person search, arXiv preprint arXiv:2101.03036.

[17] Y. Jing, C. Si, J. Wang, W. Wang, L. Wang, T. Tan, Pose-guided multi-granularity attention network for text-based person search, in: The AAAI Conference on Artificial Intelligence (AAAI), Vol. 34, 2020, pp. 11189–11196.

[18] Z. Wang, Z. Fang, J. Wang, Y. Yang, Vitaa: Visual-textual attributes alignment in person search by natural language, in: The European Conference on Computer Vision (ECCV), 2020, pp. 402–420.

[19] K. Niu, Y. Huang, W. Ouyang, L. Wang, Improving description-based person re-identification by multi-granularity image-text alignments, IEEE Transactions on Image Processing (TIP) 29 (2020) 5542–5556.

[20] D. Chen, H. Li, X. Liu, Y. Shen, J. Shao, Z. Yuan, X. Wang, Improving deep visual representation for person re-identification by global and local image-

language association, in: The European Conference on Computer Vision (ECCV), 2018, pp. 54–70.

[21] S. Aggarwal, V. B. RADHAKRISHNAN, A. Chakraborty, Text-based person search via attribute-aided matching, in: The IEEE Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 2617–2625.

[22] Y. Kim, D. Kim, A cnn-based 3d human pose estimation based on projection of depth and ridge data, Pattern Recognition (PR) 106 (2020) 107462.

[23] L. Tian, P. Wang, G. Liang, C. Shen, An adversarial human pose estimation network injected with graph structure, Pattern Recognition (PR) 115 (2021) 107863.

[24] J. Fu, J. Liu, Y. Li, Y. Bao, W. Yan, Z. Fang, H. Lu, Contextual deconvolution network for semantic segmentation, Pattern Recognition (PR) 101 (2020) 107152.

[25] X. Wang, S. Zheng, R. Yang, B. Luo, J. Tang, Pedestrian attribute recognition: A survey, arXiv preprint arXiv:1901.07.

[26] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805.

[27] L. Zheng, Y. Yang, A. G. Hauptmann, Person re-identification: Past, present and future, arXiv preprint arXiv:1610.02984.

[28] G. Zhang, T. Jiang, J. Yang, J. Xu, Y. Zheng, Cross-view kernel collaborative representation classification for person re-identification, Multimedia Tools and Applications (2021) 1–19.

[29] G. Wang, Y. Yuan, X. Chen, J. Li, X. Zhou, Learning discriminative features with multiple granularities for person re-identification, in: The ACM international conference on Multimedia (ACMMM), 2018, pp. 274–282.

[30] L. Zhao, X. Li, Y. Zhuang, J. Wang, Deeply-learned part-aligned representations for person re-identification, in: The IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3219–3228.

[31] G. Song, B. Leng, Y. Liu, C. Hetang, S. Cai, Region-based quality estimation network for large-scale person re-identification, in: The AAAI Conference on Artificial Intelligence (AAAI), Vol. 32, 2018.

[32] H. Zhao, M. Tian, S. Sun, J. Shao, J. Yan, S. Yi, X. Wang, X. Tang, Spindle net: Person re-identification with human body region guided feature decomposition and fusion, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1077–1085.

[33] M. M. Kalayeh, E. Basaran, M. G¨okmen, M. E. Kamasak, M. Shah, Human semantic parsing for person re-identification, in: the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1062–1071.

[34] X. Zhu, X.-Y. Jing, X. You, X. Zhang, T. Zhang, Video-based person re-identification by simultaneously learning intra-video and inter-video distance metrics, IEEE Transactions on Image Processing (TIP) 27 (11) (2018) 5683–5695.

[35] J. Meng, A. Wu, W.-S. Zheng, Deep asymmetric video-based person re-identification, Pattern Recognition (PR) 93 (2019) 430–441.

[36] G. Zhang, Y. Chen, Y. Dai, Y. Zheng, Y. Wu, Reference-aided part-aligned feature disentangling for video person re-identification, arXiv preprint arXiv:2103.11319.

[37] G. Wang, S. Yang, H. Liu, Z. Wang, Y. Yang, S. Wang, G. Yu, E. Zhou, J. Sun, High-order information matters: Learning relation and topology for occluded person re-identification, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 6449–6458.

[38] M. Ye, X. Lan, P. C. Yuen, Robust anchor embedding for unsupervised video person re-identification in the wild, in: The European Conference on Computer Vision (ECCV), 2018, pp. 170–186.

[39] Y. Lin, X. Dong, L. Zheng, Y. Yan, Y. Yang, A bottom-up clustering approach to unsupervised person re-identification, in: The AAAI Conference on Artificial Intelligence (AAAI), Vol. 33, 2019, pp. 8738–8745.

[40] L. Song, C. Wang, L. Zhang, B. Du, Q. Zhang, C. Huang, X. Wang, Unsupervised domain adaptive re-identification: Theory and practice, Pattern Recognition (PR) 102 (2020) 107173.

[41] J. Jiao, W.-S. Zheng, A. Wu, X. Zhu, S. Gong, Deep low-resolution person re-identification, in: The AAAI Conference on Artificial Intelligence (AAAI), Vol. 32, 2018.

[42] Z. Cheng, Q. Dong, S. Gong, X. Zhu, Inter-task association critic for cross-resolution person re-identification, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 2605–2615.

[43] K. Han, Y. Huang, C. Song, L. Wang, T. Tan, Adaptive super-resolution for person re-identification with low-resolution images, Pattern Recognition (PR) 114 (2021) 107682.

[44] Y. Lu, Y. Wu, B. Liu, T. Zhang, B. Li, Q. Chu, N. Yu, Cross-modality person re-identification with shared-specific feature transfer, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 13379–13389.

[45] S. Choi, S. Lee, Y. Kim, T. Kim, C. Kim, Hi-cmd: hierarchical cross-modality disentanglement for visible-infrared person re-identification, in: the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10257–10266.

[46] Y. Hao, J. Li, N. Wang, X. Gao, Modality adversarial neural network for visible-thermal person re-identification, Pattern Recognition (PR) 107 (2020) 107533.

[47] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778

[48] X. Jiang, Y. Gong, X. Guo, Q. Yang, F. Huang, W.-S. Zheng, F. Zheng, X. Sun, Rethinking temporal fusion for video-based person re-identification on semantic and time aspect, in: The AAAI Conference on Artificial Intelligence (AAAI), Vol. 34, 2020, pp. 11133–11140.

[49] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: The International Conference on Neural Information Processing Systems (NeurIPS), 2017, pp. 6000–6010.

[50] H. T. Le, C. Cerisara, A. Denis, Do convolutional networks need to be deep for text classification?, arXiv preprint arXiv:1707.04108.

[51] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: a large-scale hierarchical image database, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2009, pp. 248–255.

[52] L. Van der Maaten, G. Hinton, Visualizing data using t-sne, Journal of Machine Learning Research (JMLR) 9 (11).