ASPD-Net: Self-aligned Part Mask for Improving Text-based Person Re-identification with Adversarial Representation Learning

Aichun Zhu, Zijie Wang, Xili Wan, Xiaomei Zhu, Tian Wang, Yifeng Li, Gang Hua, Hichem Snoussi

Abstract—As the aim of text-based person re-identification is to retrieve images of the corresponding person from a large visual database according to a natural language description, it can be quite challenging to measure the cross-modal similarity between images and text descriptions due to the existing modality heterogeneity. Considering that the partial image regions are related to the given textual description, many researchers seek to utilize local features for more accurate matching. When it comes to visual local information extraction, most of the state-of-the-art methods adopt either a strict uniform strategy which can be too rough to catch local details properly, or pre-processing with external cues which may suffer from the deviations of the pre-trained model and the large computation consumption. In this paper, we proposed an Adversarial Self-aligned Part Detecting Network (ASPD-Net) model which extracts and combines fine-grained local/global visual and textual features. To address above issues, a novel Self-aligned Part Mask Detection Module is adopted to autonomously learn and extract human part information with more accuracy. Seeing the main model branches as a generator, a discriminator is utilized to determine whether the representation vector comes from the visual modality or the textual modality. Trained with the Adversarial Loss, ASPD-Net is able to learn more robust representations as long as it deceives the discriminator successfully. Experimental results demonstrate that the proposed ASPD-Net outperforms the previous methods and achieves the state-of-the-art performance on the CUHK-PEDES dataset.

Index Terms—Part Mask Detection, Text-based Person Re-identification, Adversarial Learning

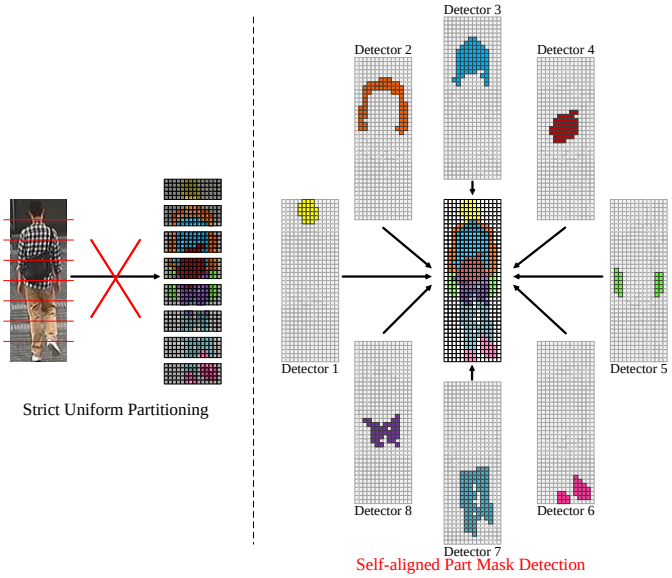


Fig. 1. Illustration of the effectiveness of the Strict Uniform Partitioning Strategy and the Self-aligned Part Mask Detection Strategy. It is obvious that the Strict Uniform Partitioning Strategy can be too rough to catch local information properly. By employing K self-aligned part mask detectors, ASPD-Net is able to autonomously learn and extract more detailed human part features.

I. Introduction

NOWADAYS, a large amount of cities are equipped with a great many surveillance cameras, by which large-scale videos are generated every second. These videos can be utilized to assist the police in criminal investigation and case handling. Considering that in many criminal scenes textual descriptions may be the only accessible information to search for one certain pedestrian, text-based person re-identification [1], [2], [3], [4] has drawn remarkable attention with the development of deep learning methods [5], [6], [3], [7], [8], [9], [10], which meets the urgent need for automatic person searching. Therefore, an efficient approach is desired to properly match a textual query with the video snapshot images of the corresponding pedestrian within a large visual database.

The major challenge of text-based person re-identification is to properly extract and match feature representations from both the visual and textual modalities. In addition, compared with the general cross-modal retrieval task, the text-based person re-identification task has its own particularities. Specifically, each image handled by the general cross-modal retrieval task commonly contains various categories of objects and information carried by the query textual descriptions is somehow rough or abstract. On the contrary, every image cared by the text-based person re-identification task contains just one certain pedestrian, and in the meantime the textual description queries offer much more details about the target person. The above mentioned particularity of text-

A. Zhu and G. Hua are with School of Information and Control Engineering, China University of Mining and Technology (email: aichun.zhu@njtech.edu.cn, ghua@cumt.edu.cn). A. Zhu, Z. Wang, X. Wan, X. Zhu and Y. Li are with School of Computer Science and Technology, Nanjing Tech University, China (email: aichun.zhu@njtech.edu.cn, zijiewang9928@gmail.com, xiliwan@njtech.edu.cn, njiczxm@njtech.edu.cn, lyffz4637@163.com). T. Wang is with School of Automation Science and Electrical Engineering, Beihang University, China (email: wangtian@buaa.edu.cn). H. Snoussi is with Institute Charles Delaunay-LM2S FRE CNRS 2019, University of Technology of Troyes, France (email: hichem.snoussi@utt.fr).

based person re-identification caused that many previous methods proposed on general cross-modal retrieval benchmarks (e.g. Flickr30K [11] ans MSCOCO [12]) generalize poorly on it, and thereby fine-grained cross-modal information ought to be taken into consideration for superior performance.

Many of the existing methods [1], [4], [13] employ multi-granular cues from both the visual and textual modalities to improve the searching accuracy. Considering the structural characteristics of the text data, phrases extracted from each sentence are utilized to obtain the textual fine-grained local information [1], [4], [13], which are commonly acquired with the Natural Language ToolKit (NLTK) by syntactic analysis, word segmentation and part-of-speech tagging. When it comes to the visual modal, some of the previous methods [4] introduce pre-processing with external cues (e.g. pose) to locate the local components, from which the visual local features are extracted. These approaches utilize an extra model to pre-process the input data. Due to the domain gap between data the extra model pre-trained on and data to be processed in the text-based person re-identification task, however, external cues generated by directly applying the pre-trained model without fine-tuning may suffer from great deviations. Unfortunately, as there is no annotation of body part in the dataset of text-based person re-identification, to fine-tune or re-train the proposed extra model seems impossible. Besides, introducing additional models is computationally costly as well. Following [5], some other approaches [1], [13] adopt a strict uniform strategy which horizontally crops the feature map into a fixed number of non-overlapping stripes for local visual feature extraction. Although free from the extra model and computational consumption problem, this strict uniform partitioning strategy still has its limitations. As person re-identification is commonly considered as the next high-level task after a pedestrian detection system, this strategy is based on the assumption that the bounding boxes provided by the previous detection model are perfect, which is tough for most current detection models, thereby introducing errors. Therefore, the proper extraction and utilization of visual local features deserve more in-depth exploration.

To this end, we proposed a Self-aligned Part Mask Module to autonomously learn human part information, which extracts the visual local features in a soft-attentional manner. In addition, an Adversarial Self-aligned Part Detecting Network (ASPD-Net) model is proposed to extract and combine rich visual/textual global and fine-grained local features. As shown in Fig. 1, comparing to the Strict Uniform Partitioning Strategy which can be too rough to catch local information properly, ASPD-Net is able to autonomously catch more detailed human part features by employing K Self-aligned Part Mask Detectors. Besides, considering that the main goal of Text-based Person Re-identification is to extract discriminative information from either the visual or the textual modality for the following similarity measuring step, it is reasonable that representations from both the visual and the textual modality should include sufficient general information about the targeted person, rather than information with respect to a single modality. In other words, ideally information contained either in the visual feature vector or in the textual feature vector is supposed to be the intersection of the visual and the textual modality. Seeing the network branches which generates the four local/global visual and textual representations as generators, we employ a discriminator to determine whether the representation vector comes from the visual modality or the textual modality. Intuitively, the ASPD-Net can extract much more discriminative feature vectors as long as it can deceive the discriminator successfully. Our proposed method is evaluated on the CUHK-PEDES dataset [3], which is a challenging dataset currently only available for text-based person re-identification. Experimental results present that the proposed ASPD-Net outperforms the previous methods and achieves the state-of-the-art performance on this dataset.

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

- A Self-aligned Part Mask Module is proposed to autonomously learn human part information, which extracts the visual local features in a soft-attentional manner while does not introduce extra pre-processing and computational consumption.
- An Adversarial Self-aligned Part Detecting Network (ASPD-Net) model is proposed to extract and combine visual/textual global and fine-grained local features.
- Considering the main model branches as a generator, a discriminator is utilized to determine whether the representation vector comes from the visual or textual modality, which enables the ASPD-Net to learn more robust modality-invariant representations.
- A comprehensive study is carried out to evaluate the proposed ASPD-Net model. Experimental results demonstrate that the proposed ASPD-Net significantly outperforms previous methods.

The rest of this paper is organized as follows. Section II illustrates related work for person re-identification and text-based person re-identification. Section III introduces the proposed model. The experiments and the comparison results are provided in Section IV. Finally, Section V gives the conclusion of this paper.

II. Related Works

A. Person Re-identification

Person re-identification has drawn increasing attention in both academical and industrial fields [5], [14], [6], [15], [16], [3], [17], [18], [7], [8], [9], and deep learning methods are in general playing a major role in current state-of-the-art works. Yi et al. [?] firstly proposed deep learning methods to match people with the same identification. Xia et al. [19] proposed the Second-order Non-local Attention (SONA) Module to learn local/non-local information

and relationships in a more end-to-end way. In order to strengthen the representation capability of the deep neural network, Hou et al. [20] proposed the Interaction-and-Aggregation (IA) Block, which consists of a Spatial Interaction-and-Aggregation (SIA) Module and a Channel Interaction-and-Aggregation (CIA) Module and can be inserted into deep CNNs at any depth.

B. Text-based Person Re-identification

Text-based person re-identification has been studied from various perspective [1], [2], [3], [4]. It is challenging to directly measure the affinity between images and descriptions, on account of the cross-modality heterogeneity. Li et al. [3] came up with the first work with deep learning methods in the text-based person re-identification task, which proposed a VGG-16 to extract global visual features. More importantly, they provided the CUHK Person Description Dataset (CUHK-PEDES), which currently is still the only accessible dataset for the text-based person re-identification task. Then in [24], Li et al. proposed an identity-aware two-stage framework for the textual-visual matching. Identity-aware representation is learned in Stage 1, while in Stage 2 salient image regions and latent semantic concepts are matched for the following textual-visual affinity estimation. Following this work, Chen et al. [2] propose an efficient patch-word matching model in order to capture the local similarity between image and text. More recently, many works attempt to fuse local and global cross-modal cues to further enhance the performance. Nikolaos Sarafianos et al. [?] propose a Text-Image Modality Adversarial Matching approach (TIMAM) to learn modality-invariant feature representation by means of adversarial and cross-modal matching objectives. Besides that, they employed the pre-trained BERT, a publicly-available language model, to better extract word embeddings. However, only some certain parts of an image and part of phrases of a textual description are discriminative enough to search for the corresponding person. In addition, just partial image regions are related to the given textual description. Considering these problems mentioned, many researchers seek to utilize local features for more accurate matching. Niu et al. [1] propose a Multi-granularity Image-text Alignments (MIA) model to extract fine-grained features by partitioning the feature map horizontally into multiple non-overlapping parts and then adopting 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 [4] employ pose information as inner-modal attention to provide soft partial image regions, which help to localize the discriminative 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.

III. Methodology

In this section, we introduce the Adversarial Self-aligned Part Detecting Network (ASPD-Net) in detail. First, the Self-aligned Part Mask Detection module is presented. Second, we describe the whole network architecture. The overall architecture of ASPD-Net is shown in Fig. 2.

A. Preliminary

B. Overall Architecture

C. Self-aligned Part Mask Detection

As discussed in the Introduction, roughly partitioning the image to extract local features may not be a good choice in some cases. On the contrary, the self-aligned part masks autonomously learn to align vision parts from the image.

The self-aligned part mask detection module consists of K self-aligned part mask detectors. Each detector takes in a 3-dimensional feature map $X \in \mathbb{R}^{h\times w\times c}$, where h, w and c denote height, width and channel number of the feature map, and then gives out a 2-dimensional part mask $M_i \in \mathbb{R}^{h\times w}, i \in \{1, 2, \cdots, K\}$. The detectors are implemented as a 1×1 conv layer followed by a Sigmoid layer:

$$M_i = Sigmoid(W_{detector} \times X + b_{detector}), \tag{1}$$

where $W_{detector}$ and $b_{detector}$ denote the 1×1 convolution layer.

Each mask is then duplicated to form a 3-dimensional mask $M_i^{duplicated} \in \mathbb{R}^{h\times w\times c}$. The self-aligned part feature maps $X_i^{aligned} \in \mathbb{R}^{h\times w\times c}, i \in \{1, 2, \cdots, K\}$ are obtained by a Hadamard product:

$$X_i^{aligned} = M_i^{duplicated} \times X. \tag{2}$$

D. ASPD-Net

As the task of text-based person re-identification involves both textual and visual modality, ASPD-Net respectively has two branches to extract textual and visual features separately. In order to provided more discriminative information, ASPD-Net extracts global information as well as local information from each modality, which finally includes a global visual representation vector, a local visual representation vector, a sentence representation vector and a phrase representation vector.

1) Textual Feature Extraction: As for textual representation vectors extraction, each word $w \in \mathbb{R}^W$ in the text description is first embedded into a vector $x \in \mathbb{R}^E$ following

$$x = W_e \times w, \tag{3}$$

where $W_e \in \mathbb{R}^{E\times W}$. Then, a shared bi-directional gated recurrent unit (Bi-GRU) is adopted to determine the dependencies between adjacent words for both sentences and phrases, which follows

$$\overrightarrow{h_t} = \overrightarrow{GRU}(x, \overrightarrow{h_{t-1}}), \tag{4}$$

$$\overleftarrow{h_t} = \overleftarrow{GRU}(x, \overleftarrow{h_{t-1}}), \tag{5}$$

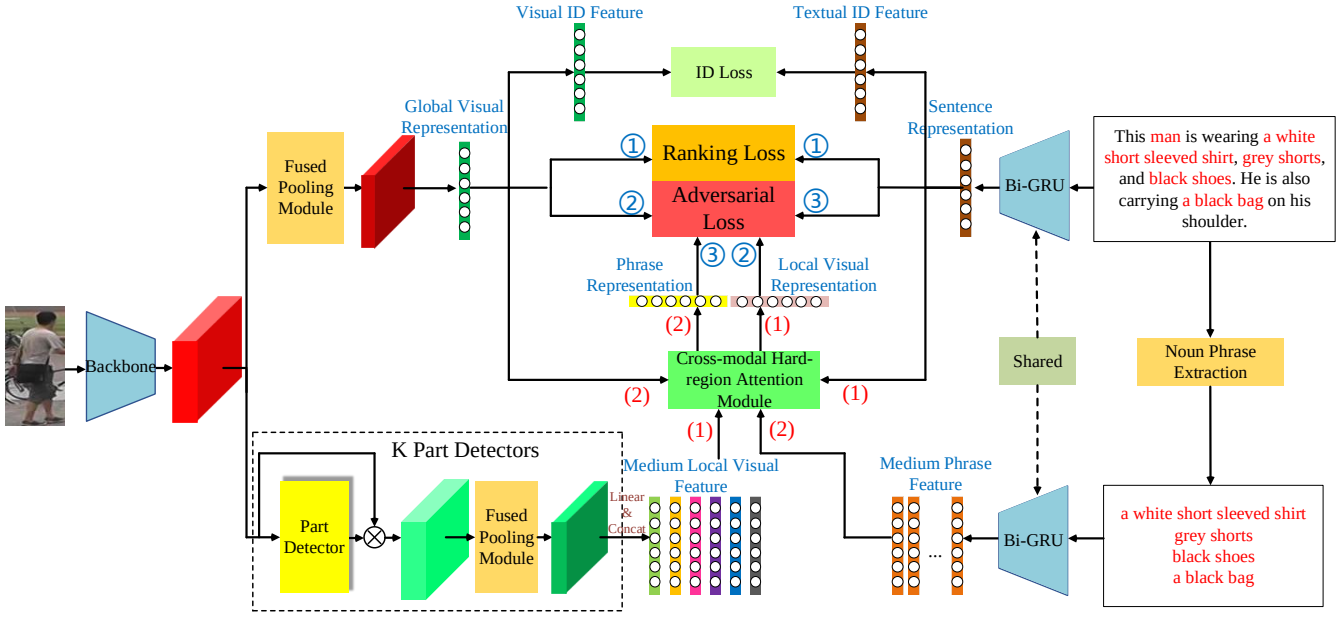


Fig. 2. The overall architecture of our proposed Adversarial Self-aligned Part Detecting Network (ASPD-Net). It extracts four representations of different granularities: global visual representation, local visual representation, sentence representation and phrase representation and matches visual and textual information via three different cross-modal combinations including global2sentence, local2sentence and global2phrase. ID Loss, Ranking Loss and Adversarial Loss are employed to train ASPD-Net. The sequence numbers (1) and (2) denote the corresponding vectors processed and generated by the Cross-modal Hard region Attention Module, while ①, ② and ③ denote the corresponding combinations when training ASPD-Net.

where $\overrightarrow{GRU}$ and $\overleftarrow{GRU}$ relatively denote the forward and backward GRUs, $t \in \{1, \cdots, n\}$, n is the number of words in the input sentence. After that, ASPD-Net concatenates the last hidden states of the forward and backward GRUs $\overleftarrow{h_n}$ and $\overrightarrow{h_n}$ to give the sentence or phrase feature:

$$e = concat(\overrightarrow{h_n}, \overleftarrow{h_n}), \tag{6}$$

where $e \in \mathbb{R}^P$, and e^k can be e^S or e^P which denotes sentence or phrase feature, respectively. To extract the sentence representation, the concatenated feature is passed through a batch normalization layer followed by a Fully-connected (FC) layer to obtain the sentence representation vector $T^S \in \mathbb{R}^P$. When it comes to phrases, each concatenated phrase feature $e_i^P \in \mathbb{R}^P, i \in \{1, \cdots, n\}$, is separately passed through a multi-layer perceptron to obtain a medium phrase representation column $M_i^P \in \mathbb{R}^P$. Then all of the medium phrase representation columns are concatenated to form the medium phrase feature matrix $M^P \in \mathbb{R}^{n \times P}$.

2) Visual Feature Extraction: To extract visual representations, the input image is first passed through a CNN backbone to obtain the shared medium feature map $\varphi(I) \in \mathbb{R}^{w \times h \times c}$, where h, w and c respectively denote height, width and channel number of the feature map. Then we handle the shared medium feature map separately to get the global and the local visual representations. For the global path, we adopt a Fused Pooling Module to downscale $\varphi(I)$ to $6 \times 1 \times c$:

$$\phi(I) = AvgPooling(\varphi(I)) + MaxPooling(\varphi(I)), \tag{7}$$

where the $\varphi(I)$ is separately passed through a maximum pooling layer and an average pooling layer and then added the output of the two pooling layers are added to the final output $\phi(I) \in \mathbb{R}^{6 \times 1 \times c}$. The $\phi(I)$ is flattened to a $(6 \times c)$-dimensional vector $\phi_{flattened}(I)$ and then passed through a group normalization layer followed by a Fully-connected (FC) layer, which gives out the global visual representation vector $V^G \in \mathbb{R}^P$.

For the local path, $\varphi(I)$ is first processed by the Self-aligned Part Mask Detector Module to obtain K self-aligned part feature maps $X_i^{aligned} \in \mathbb{R}^{h \times w \times c}, i \in \{1, 2, \cdots, K\}$. Then similar to the global path, these self-aligned part feature maps are separately passed through a Fused Pooling Module, a group normalization layer and 2 FC layers with a ReLU layer between them to form the medium local visual vectors $M_i^V \in \mathbb{R}^P, i \in \{1, 2, \cdots, K\}$. After that, we concatenate them to form the medium local visual feature matrix $M^V \in \mathbb{R}^{K \times P}$.

3) Local Representation Vectors Converting: In order to convert the two medium feature matrices M^V and M^P to the corresponding representation vectors, a cross-modal attention method is firstly employed to compute how firmly each column in a feature matrix relates to the global representation vector of another modality. Taking the local visual representation vector for example, the similarity between each local part column and the sentence representation T^S is calculated following:

$$\alpha_i^V = \frac{exp(cos(M_i^V, T^S))}{\sum_{j=1}^{6} exp(cos(M_j^V, T^S))}, \tag{8}$$

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

threshold-guided weighted summation is used to finally convert the medium local visual feature matrix M^V to the local visual representation vector $V^L \in \mathbb{R}^P$:

$$V^L = \sum_{\alpha_i^V > \frac{1}{6}} \alpha_i^V \cdot M_i^V. \tag{9}$$

Similarly, the phrase representation vector is given out by

$$\alpha_i^P = \frac{exp(cos(V^G, M_i^P))}{\sum_{j=1}^{n} exp(cos(V^G, M_j^P))}, \tag{10}$$

$$T^P = \sum_{\alpha_i^P > \frac{1}{n}} \alpha_i^P \cdot M_i^P, \tag{11}$$

4) Image-Text Matching: In the image-text matching phase of ASPD-Net, 3 cross-modal combinations of the 4 obtained representation vectors are included, namely, global-to-sentence matching (global2sentence, GS), local-to-sentence matching (local2sentence, LS) and global-to-phrase matching (global2phrase, GP).

For each way of matching, the similarity between the two proposed representation vectors is measured by a cosine similarity function:

$$Simi_{vt} = cos(V^v, T^t), \tag{12}$$

where v can be either G or S which denotes either the global visual representation vector or the local visual representation vector is utilized, while t can be either S or P which denotes either a sentence representation vector or a phrase representation vector is employed.

In the test stage, the 3 similarities are fused with a weighted summation:

$$Simi_{ASPDNet} = Simi_{GS} + \frac{1}{2}(Simi_{LS} + Simi_{GP}). \tag{13}$$

E. Loss Functions And Training Strategy

To train ASPD-Net, we adopt 3 different loss functions, including identification (ID) loss, triplet ranking loss and adversarial loss. The complete training process contains two stages.

1) Stage-1: First, parameters of the visual backbone are fixed, while the left parts of ASPD-Net are trained solely with the ID loss to cluster people into different bunches according to their identifications. Considering that global representations can provide more complete information for this clustering operation, only the two global representation vectors V^G and T^S are utilized here. The two proposed ID losses L_{id}^V and L_{id}^T for visual representation and textual representation are defined as

$$L_{id}^V = -log(softmax(W_{id} \times GN(V^G))), \tag{14}$$

$$L_{id}^T = -log(softmax(W_{id} \times GN(T^S))), \tag{15}$$

where $W_{id} \in \mathbb{R}^{Q \times P}$ is a shared transformation matrix which is carried out as a FC layer without bias and Q is the number of different people in the training set. GN denotes the group normalization layer. The 2 branches shared the transformation matrix W_{id} so that the visual and textual ID representations are mapped into the same feature space. We sum up L_{id}^V and L_{id}^T to give the general ID loss:

$$L_{id} = L_{id}^V + L_{id}^T. \tag{16}$$

Thus, the final loss function in Stage-I is

$$L_{stage1} = L^{id}. \tag{17}$$

2) Stage-2: Then all the parameters of ASPD-Net are fine-tuned together including ones in the visual backbone. Here we train ASPD-Net with all the 3 loss functions.

Considering that the main goal of Text-based Person Re-identification is to extract discriminative information from either the visual or the textual modality for the following similarity measuring step, it is reasonable that representations from both the visual and the textual modality should include sufficient general information about the targeted person, rather than information with respect to a single modality. In other words, ideally information contained either in the visual feature vector or in the textual feature vector is supposed to be the intersection of the visual and the textual modality.

Seeing the network branches described above which generates the four representations V^G, V^L, T^S and T^P as generators, we utilize a discriminator to determine whether the representation vector comes from the visual modality or the textual modality. The discriminator is implemented as two FC layers with a ReLU layer between them and a Sigmoid layer, which gives a scalar value to predict the modality where the input representation vector comes from. Intuitively, the ASPD-Net can extract much more discriminative feature vectors as long as it can deceive the discriminator successfully. An adversarial loss is adopted to optimize the discriminator:

$$\begin{aligned} L_{adversarial}^k = &- \mathop{\mathbb{E}}_{V_i \sim V}[logD(V_i)] \\ &- \mathop{\mathbb{E}}_{T_i \sim T}[1 - logD(T_i)], \end{aligned} \tag{18}$$

where $L_{adversarial}^k$ denotes $L_{adversarial}^{GS}$, $L_{adversarial}^{LS}$ or $L_{adversarial}^{GP}$. V can be V^G or V^L while T can be T^S or T^P respectively according to $L_{adversarial}^k$. ASPD-Net calculates the adversarial loss between 3 pairs of cross-modal representation vectors, namely, $V^G \sim T^S$, $V^L \sim T^S$ and $V^G \sim T^P$:

$$\begin{aligned} L_{adversarial} = L_{adversarial}^{GS} &+ L_{adversarial}^{LS} \\ &+ L_{adversarial}^{GP}. \end{aligned} \tag{19}$$

The triplet ranking loss is commonly adopted in either person re-identification or description-based person re-identification tasks, which aims to constrain the match pairs to be closer than the mismatched pairs in a mini-batch with a margin α. Following [21], we employ the sum of all pairs within each mini-batch when computing the

hinge-based triplet ranking loss instead of utilizing the furthest positive and closest negative sampled pairs:

$$L^{k}_{ranking} = \sum_{\widehat{T}} max\{\alpha - cos(V,T) + cos(V,\widehat{T})\} + \sum_{\widehat{V}} max\{\alpha - cos(V,T) + cos(\widehat{V},T)\}, \quad (20)$$

where $L^{k}_{ranking}$ denotes $L^{GS}_{ranking}$, $L^{LS}_{ranking}$ or $L^{GP}_{ranking}$. V can be V^G or V^L, while T can be T^S or T^P respectively according to $L^{k}_{ranking}$. (V,T) denotes the matched visual-textual pairs while $(V,\widehat{T})$ or $(\widehat{V},T)$ denotes the mismatched pairs and α is a margin. The general triplet ranking loss is calculated following:

$$L_{ranking} = L^{GS}_{ranking} + L^{LS}_{ranking} + L^{GP}_{ranking}. \quad (21)$$

The complete loss function in Stage-2 is

$$L_{stage2} = L_{id} + L_{ranking} + L_{adversarial}. \quad (22)$$

Intuitively, the identification loss mainly focuses on the ID category of a given person, which functions more like a loose constraint thereby failing to provide adequate accuracy for the fine-grained matching task. As the triplet ranking loss regards the description sentences annotated for a certain image as negative for any other images even with the same person ID, it is much stricter. Thus, the ID loss in Stage-1 can eliminate obvious mismatched pairs and as well provide an initialization for Stage-2. Then in Stage-2 the triplet ranking losses are employed to catch more fine-grained information and in this stage the ID losses are still reserved to function as a regularization for the model. With the help of the adversarial loss, ASPD-Net is capable of extracting much more discriminative representation vectors without being impaired by information from one single modality.

IV. Experiments

A. Experimental Setup

1) Dataset and Metrics: The CUHK-PEDES dataset is currently the only dataset for Text-Based Person Re-identification task. We follow the same data split approach as [3]. In detail, the training set contains 34054 images, 11003 persons and 68126 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 performance is evaluated by the top-k accuracy. 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 search. We report top-1, top-5, and top-10 accuracies for all experiments.

2) Implementation Details: In our experiments, we set dimensionality $P = 1024$. The word number W is 4984 after dropping the words that appears less than twice and the dimensionality E of embedded word vectors is set to 300. We choose the pre-trained VGG-16 and ResNet-50 as the visual CNN backbone. We obtain noun phrases of each sentence with the Natural Language ToolKit (NLTK) by syntactic analysis, word segmentation and part-of-speech tagging. The total number of noun phrases obtained from each sentence is kept flexible. A L2 regularization is utilized to make the self-align part masks focus more on diverse local parts.

In training, we initialize the weights of the visual CNN backbone pre-trained on the ImageNet classification task. An Adam optimizer is adopted to train the model with a batch size of 32. 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 CNN 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 own way down. After that, the initial learning rate for later epochs is defined as:

$$lr = 2 \times 10^{-4} \times (\frac{1}{10})^{epoch//10}, \quad (23)$$

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 18 epochs.

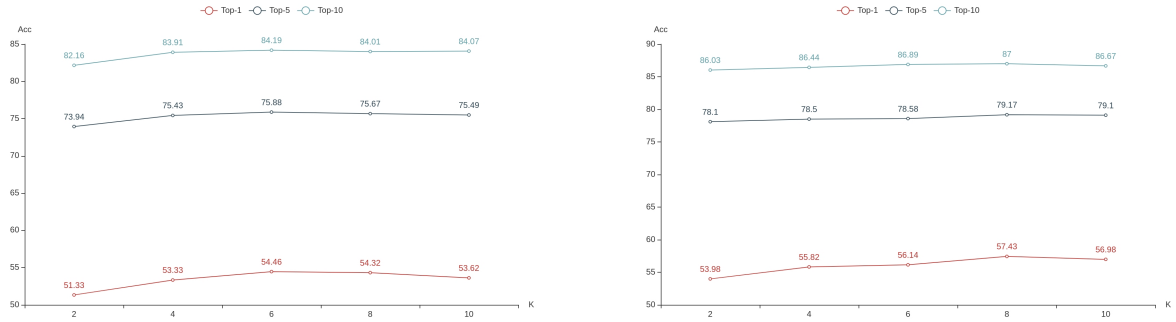
B. Ablation Analysis

To further investigate several components of ASPD-Net, we carry out plenty of ablation experiments. As shown in Table I, Table II and Table III, '-GS', '-LS' and '-GP' denote respectively the global2sentence, local2sentence or global2phrase combination is utilized while training and testing the model, while '-ALL' means all of them are employed. 'Tra-' and 'Tes-' denote the results of models trained with varied combinations and performances of a fully-trained ASPD-Net model tested with diverse granularities, respectively. 'AP', 'MP', 'FP' and 'AL' denote whether Average Pooling, Maximum Pooling, Fused Pooling and Adversarial Loss is utilized.

1) Combination of Granularities: Table I provides analysis on the effect of each granularity and the way they are combined. The results show that either training or testing with more than one single granularity brings performance gain, which indicates that the multi-granular cross-modal matching can provide more comprehensive information, hence leading to a more accurate retrieval. Specifically, as for the three combinations which combines two granularities, the one combines local2sentence and global2phrase outperforms the other two, 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 ASPD-Net model which employs both the coarse global and the fine-grained local information undoubtedly

TABLE I
Ablation analysis of granularity combination while training and testing

Method	Backbone	Tra-Top-1	Tra-Top-5	Tra-Top-10	Tes-Top-1	Tes-Top-5	Tes-Top-10
GS	VGG-16	47.21	70.31	78.98	49.97	73.94	82.62
LS	VGG-16	47.44	71.01	79.62	49.81	74.27	82.46
GP	VGG-16	42.87	67.79	77.33	45.00	69.88	78.95
GS-LS	VGG-16	51.74	74.93	83.14	53.09	75.44	83.59
GS-GP	VGG-16	50.96	74.11	81.99	51.69	74.76	83.01
LS-GP	VGG-16	53.82	75.83	83.57	52.99	75.88	83.85
ALL	VGG-16	54.46	75.88	84.19	54.46	75.92	84.19
GS	ResNet-50	51.01	73.01	81.82	53.09	75.44	83.59
LS	ResNet-50	51.14	71.28	79.95	51.69	74.76	83.01
GP	ResNet-50	46.14	70.58	79.89	47.79	73.62	83.62
GS-LS	ResNet-50	54.62	77.89	86.01	56.99	78.93	86.85
GS-GP	ResNet-50	53.09	77.39	84.77	55.43	77.88	85.83
LS-GP	ResNet-50	56.74	78.76	86.52	56.14	78.59	86.26
ALL	ResNet-50	57.43	79.17	87.00	57.43	79.17	87.00



(a) VGG-16 (b) ResNet-50

Fig. 3. Illustration of ablation analysis on the number K of self-aligned part masks with the VGG-16 backbone or the ResNet-50 backbone.

outperforms any other model utilizes part of the three granularities. What's more, the results show that the single '-LS' gives better performance than the single '-GS', while the single'-GS' is better than the single '-GP'. It is intuitive as the local2sentence matching takes more detailed information into consideration than the global2sentence matching, and the phrases for the global2phrase matching may be too short to offer sufficient features.

2) Number of Self-aligned Part Masks: Ablation experiments are conducted to search for the optimal number K of self-aligned part masks, whose results are recorded in Table III. As can be observed from the illustrations of ablation analysis results in Fig. 3(a) and Fig. 3(b), initially with increase of K, performance of ASPD-Net keeps improving. Then after reaching a peak, the performance begins to turn worse as K continues to go larger. It is conceivable that with more self-aligned part mask to autonomously learn to align vision parts from the image, ASPD-Net can catch more detailed information. Nevertheless, in spite of the L2 regularization, when K becomes too large, some of the masks may still focus on similar local parts, which can do little benefit to the performance. What's more, too many parameters can be introduced as the number of mask branches goes too large. Some examples of self-aligned part masks learned by ASPD-Net for some images from the testing set are shown in Fig. 4. As can be seen, our self-learned detector can focus on most key components to help improve the accuracy, which are self-learned with similarity information in an end-to-end manner instead of relying on labeling information. As for the 2 images of people holding a backpack in the first row, for example, the backpack part is detected by the 5th mask of the left image and the 4th mask of the right image, respectively. Other key parts contributing to the matching process like head, limbs, foot, etc. are detected properly as well.

3) Effectiveness of Adversarial Loss: As is shown in Table II, with the assistance of the adversarial loss, the top-1 accuracy of ASPD-Net from 53.33% to 54.46% and from 56.67% to 57.43% with VGG-16 and ResNet-50 as visual CNN backbone respectively. The results prove that being able to deceive the discriminator successfully, ASPD-Net can extract much more discriminative representation vectors.

4) Fused Pooling Module: Table II also shows the ablation analysis results of the combination of pooling methods. As is shown in the table, while both utilizing one single pooling layer in the model, model with maximum pooling method mildly outperforms the one with the average pooling method, which is reasonable as the

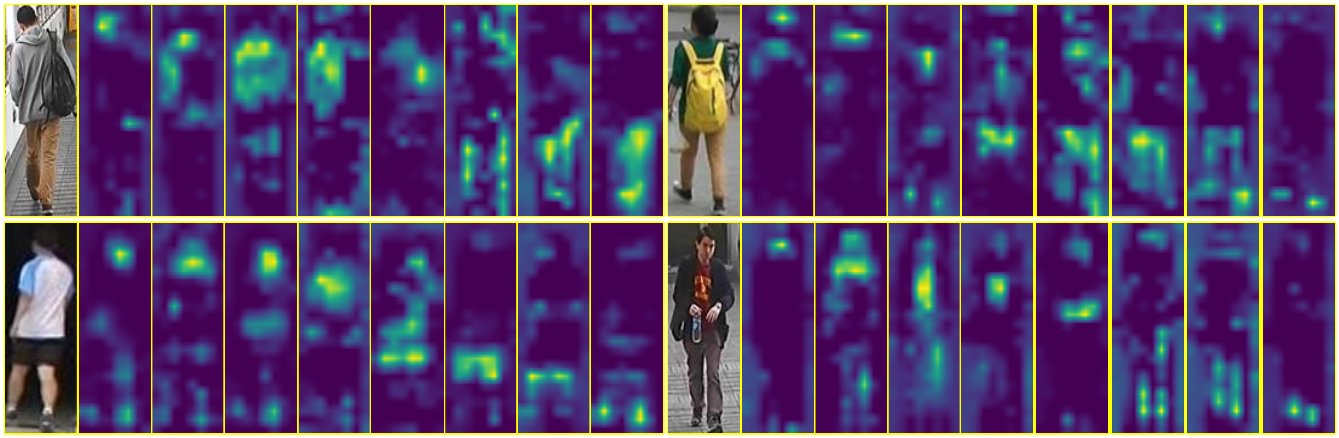


Fig. 4. Examples of self-aligned part masks learned by ASPD-Net for some images from the testing set. Our self-learned detector can focus on most key components to help improve the accuracy, which are self-learned with similarity information in an end-to-end manner instead of relying on labeling information. As for the 2 images of people holding a backpack in the first row, for example, the backpack part is detected by the 5th mask of the left image and the 4th mask of the right image, respectively. Other key parts contributing to the matching process like head, limbs, foot, etc. are detected properly as well.

TABLE II
Comparison of key components

AP	MP	FP	AL	Backbone	Top-1	Top-5	Top-10
✓	×	×	✓	ResNet-50	56.54	78.82	86.87
×	✓	×	✓	ResNet-50	56.86	78.98	86.91
×	×	✓	✓	ResNet-50	57.43	79.17	87.00
×	×	✓	×	VGG-16	53.33	75.59	84.11
×	×	✓	✓	VGG-16	54.46	75.88	84.19
×	×	✓	×	ResNet-50	56.67	78.84	86.89
×	×	✓	✓	ResNet-50	57.43	79.17	87.00

TABLE III
Ablation analysis of the number K of Self-aligned Part Masks

K	Backbone	Top-1	Top-5	Top-10
2	VGG-16	51.33	73.94	82.16
4	VGG-16	53.33	75.43	83.91
6	VGG-16	54.46	75.88	84.19
8	VGG-16	54.32	75.67	84.01
10	VGG-16	53.62	75.49	84.07
2	ResNet-50	53.98	78.10	86.03
4	ResNet-50	55.82	78.50	86.44
6	ResNet-50	56.14	78.58	86.89
8	ResNet-50	57.43	79.17	87.00
10	ResNet-50	56.98	79.10	86.67

average pooling method average pooling layer is able to take contextual information into consideration while maximum pooling method cannot. However, as the maximum pooling method is capable of catching the most salient signals in the feature map, it can help as well in case signals surrounded a salient signal are relatively weak, where the average pooling method may blur the discriminative signals. Therefore, after fusing the two methods together, ASPD-Net performs best with contextual information and the most salient signals complementing each other.

TABLE IV
Comparison with other state-of-the-art methods

Method	Backbone	Top-1	Top-5	Top-10
CNN-RNN [22]	VGG-16	8.07	-	32.47
Neural Talk [23]	VGG-16	13.66	-	41.72
GNA-RNN [3]	VGG-16	19.05	-	53.64
IATV [24]	VGG-16	25.94	-	60.48
PWM-ATH [25]	VGG-16	27.14	49.45	61.02
Dual Path [?]	VGG-16	32.15	54.42	64.30
GALM [4]	VGG-16	47.82	69.83	78.31
MIA [?]	VGG-16	48.00	70.70	79.30
ASPD-Net(ours)	VGG-16	54.46	75.88	84.19
Dual Path [?]	ResNet-50	44.40	66.26	75.07
GLA [25]	ResNet-50	43.58	66.93	76.26
MIA [?]	ResNet-50	53.10	75.00	82.90
GALM [4]	ResNet-50	54.12	75.45	82.97
TIMAM [26]	ResNet-101	54.51	77.56	84.78
ASPD-Net(ours)	ResNet-50	57.43	79.17	87.00

Finally, we display some examples of top-5 text-based person re-identification results by full ASPD-Net model with $k = 8$, full ASPD-Net with $k = 4$ and ASPD-Net with $k = 8$ while without Adversarial Loss. All the three ASPD-Net proposed here utilize ResNet-50 as visual backbone. Images of the target person are marked by green rectangles in Fig. 5.

C. Comparison with Other State-of-the-art Methods

The comparisons with other state-of-the-art methods are shown in Table IV. It can be seen that our ASPD-Net model achieves the best performance under top-1, top-5 and top-10 metrics. PWM-ATH proposes an efficient patch-word matching model to capture the local similarity between image and text, but ignores the global-local relations. With VGG-16 backbone, ASPD-Net outperforms PWM-ATH by over 27% under top-1 metric, which validates the significance of the local2sentence and

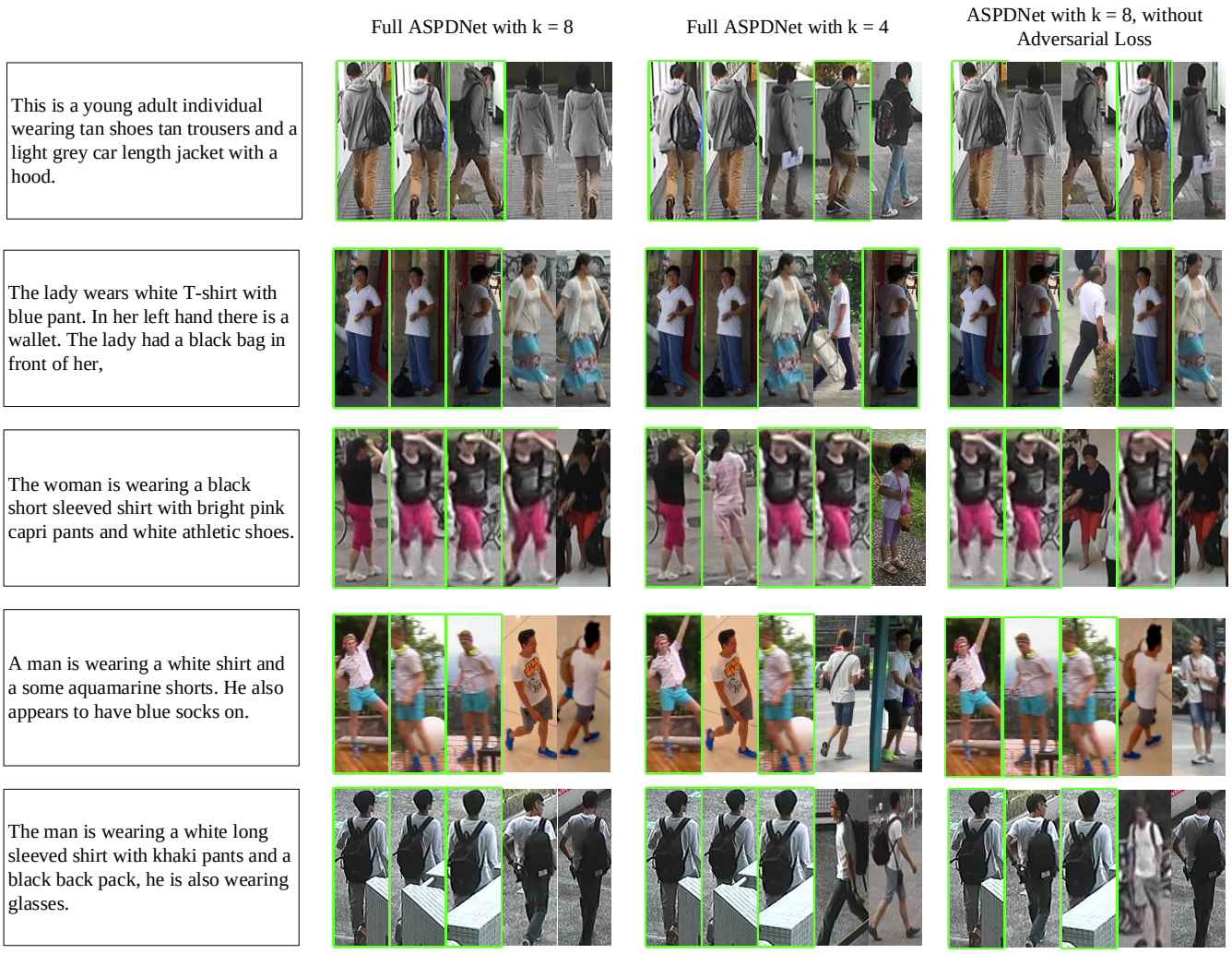


Fig. 5. Examples of top-5 text-based person re-identification results by full ASPD-Net model with $k = 8$, full ASPD-Net with $k = 4$ and ASPD-Net with $k = 8$ while without Adversarial Loss. All the three ASPD-Net proposed here utilize ResNet-50 as visual backbone. Images of the target person are marked by green rectangles.

global2phrase granularities in our method. Compared with the best competitor MIA using VGG-16 as visual backbone, ASPD-Net significantly outperforms it by 6.46% under top-1 metric, indicating the superiorities of the self-align part mask detector module and the adversarial loss. Even with the VGG-16 backbone, ASPD-Net outperforms GALM, the best competitor with ResNet-50 backbone, by 0.34% under top-1 metric, which proves the effectiveness of the self-align part mask detector module without utilizing extra pre-trained cues(e.g., pose). With ResNet-50 backbone, ASPD-Net even outperforms TIMAM, the currently best competitor with ResNet-101 backbone by 2.93% under top-1 metric.

D. Apply ASPD-Net To Other Text-to-image Retrieval Tasks

Considering that description-based person re-identification can be considered as a subdomain of text-to-image retrieval, it's reasonable that ASPD-Net can be applied to other tasks besides people image. Thus we train ASPD-Net on Flickr30K and test it with text-to-image retrieval task. As shown in Table V, ASPD-Net

TABLE V
Performance of ASPD-Net trained on Flickr30K dataset

Method	Backbone	Top-1	Top-5	Top-10
m-CNN[27]	VGG-19	26.2	56.3	69.6
DSPE[28]	VGG-19	29.7	60.1	72.1
RRF-Net[29]	ResNet-152	35.4	68.3	79.9
CMPM+CMPC[?]	ResNet-152	37.3	65.7	75.5
DANs[?]	ResNet-152	39.4	69.2	79.1
NAR[30]	ResNet-152	39.4	68.8	79.9
VSE++[21]	ResNet-152	39.6	70.1	79.5
SCO[31]	ResNet-152	41.1	70.5	80.1
GXN[32]	ResNet-152	41.5	-	80.1
TIMAM[26]	ResNet-152	42.6	71.6	81.9
SCAN[33]	Faster R-CNN	48.6	77.7	85.2
BFAN[34]	Faster R-CNN	50.8	78.4	-
PFAN[35]	Faster R-CNN	50.4	78.7	86.1
ASPD-Net(ours)	ResNet-50	49.8	78.7	85.3

with ResNet-50 backbone has better performance than all previous work with ResNet-152 and close performance with ones with Faster R-CNN, which indicates our part detectors can detect key component in multi-object tasks as well.

V. Conclusion

In this work, we address the problems in the field of the text-based person re-identification and design an Adversarial Self-aligned Part Detecting Network (ASPD-Net) model to extract and combine fine-grained local/global visual and textual features. Specifically, the Self-aligned Part Mask Module is employed to address the within-part consistency broken problem. With the aid of the Adversarial Loss, ASPD-Net can extract much more discriminative feature vectors as long as it can deceive the discriminator successfully. Furthermore, we evaluate our approach on the CUHK-PEDES dataset and the results indicate that the proposed ASPD-Net improves the performance with a large margin.

Acknowledgment

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

References

[1] K. Niu, Y. Huang, W. Ouyang, and L. Wang, "Improving description-based person re-identification by multi-granularity image-text alignments," IEEE Transactions on Image Processing, vol. 29, pp. 5542–5556, 2020.

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

[3] S. Li, T. Xiao, H. Li, B. Zhou, D. Yue, and X. Wang, "Person search with natural language description," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1970–1979.

[4] Y. Jing, C. Si, J. Wang, W. Wang, L. Wang, and T. Tan, "Pose-guided multi-granularity attention network for text-based person search," arXiv preprint arXiv:1809.08440, 2018.

[5] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, "Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)," ECCV, 2018.

[6] C. Song, Y. Huang, W. Ouyang, and L. Wang, "Mask-guided contrastive attention model for person re-identification," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1179–1188.

[7] Y.-J. Cho and K.-J. Yoon, "Pamm: Pose-aware multi-shot matching for improving person re-identification," vol. 27, no. 8. IEEE, 2018, pp. 3739–3752.

[8] H. Yao, S. Zhang, R. Hong, Y. Zhang, C. Xu, and Q. Tian, "Deep representation learning with part loss for person re-identification," IEEE Transactions on Image Processing, vol. 28, no. 6, pp. 2860–2871, 2019.

[9] J. Dai, P. Zhang, D. Wang, H. Lu, and H. Wang, "Video person re-identification by temporal residual learning," IEEE Transactions on Image Processing, vol. 28, no. 3, pp. 1366–1377, 2018.

[10] X. Zhang, T. Huang, Y. Tian, and W. Gao, "Background-modeling-based adaptive prediction for surveillance video coding," IEEE Transactions on Image Processing, vol. 23, no. 2, pp. 769–784, 2013.

[11] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik, "Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models," in Proceedings of the IEEE international conference on computer vision, 2015, pp. 2641–2649.

[12] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in European conference on computer vision. Springer, 2014, pp. 740–755.

[13] Z. Wang, A. Zhu, Z. Zheng, J. Jin, Z. Xue, and G. Hua, "Img-net: inner-cross-modal attentional multigranular network for description-based person re-identification," Journal of Electronic Imaging, vol. 29, no. 4, p. 043028, 2020.

[14] Z. Zhong, L. Zheng, Z. Luo, S. Li, and Y. Yang, "Invariance matters: Exemplar memory for domain adaptive person re-identification," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 598–607.

[15] Y. Sun, Q. Xu, Y. Li, C. Zhang, Y. Li, S. Wang, and J. Sun, "Perceive where to focus: Learning visibility-aware part-level features for partial person re-identification," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 393–402.

[16] J. Liu, B. Ni, Y. Yan, P. Zhou, S. Cheng, and J. Hu, "Pose transferrable person re-identification," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 4099–4108.

[17] C. Su, J. Li, S. Zhang, J. Xing, W. Gao, and Q. Tian, "Pose-driven deep convolutional model for person re-identification," in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 3960–3969.

[18] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng, "Person re-identification by multi-channel parts-based cnn with improved triplet loss function," in Proceedings of the iEEE conference on computer vision and pattern recognition, 2016, pp. 1335–1344.

[19] B. N. Xia, Y. Gong, Y. Zhang, and C. Poellabauer, "Second-order non-local attention networks for person re-identification," in Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 3760–3769.

[20] R. Hou, B. Ma, H. Chang, X. Gu, S. Shan, and X. Chen, "Interaction-and-aggregation network for person re-identification," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 9317–9326.

[21] F. Faghri, D. J. Fleet, J. R. Kiros, and S. Fidler, "Vse++: Improving visual-semantic embeddings with hard negatives," arXiv preprint arXiv:1707.05612, 2017.

[22] S. Reed, Z. Akata, H. Lee, and B. Schiele, "Learning deep representations of fine-grained visual descriptions," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 49–58.

[23] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3156–3164.

[24] S. Li, T. Xiao, H. Li, W. Yang, and X. Wang, "Identity-aware textual-visual matching with latent co-attention," in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 1890–1899.

[25] D. Chen, H. Li, X. Liu, Y. Shen, J. Shao, Z. Yuan, and X. Wang, "Improving deep visual representation for person re-identification by global and local image-language association," in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 54–70.

[26] I. A. K. Nikolaos Sarafianos, Xiang Xu, "Adversarial representation learning for text-to-image matching," in ICCV, 2019, pp. 5813–5823.

[27] L. S. Lin Ma, Zhengdong Lu and H. Li, "Multimodal convolutional neural networks for matching image and sentence," 2015.

[28] J. H. Liwei Wang, Yin Li and S. Lazebnik, "Learning two-branch neural networks for image-text matching tasks," in IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, pp. 394–407.

[29] E. M. B. Y. Liu, Y. Guo and M. S. Lew, "Learning a recurrent residual fusion network for multimodal matching," 2017.

[30] W. Z. Chunxiao Liu, Zhendong Mao and B. Wang, "A neighbor-aware approach for image-text matching," 2019.

[31] C. S. Yan Huang, Qi Wu and L. Wang, "Learning semantic concepts and order for image and sentence matching," 2018.

[32] S. R. J. L. N. Jiuxiang Gu, Jianfei Cai and G. Wang, "Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models," 2018.
[33] K.-H. Lee, X. Chen, G. Hua, H. Hu, and X. He, "Stacked cross attention for image-text matching," in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 201–216.
[34] A.-A. L. T. Z. B. W. Chunxiao Liu, Zhendong Mao and Y. Zhang, "Focus your attention: A bidirectional focal attention network for image-text matching," 2019.
[35] X. Q. L. M. J. L. B. L. Yaxiong Wang, Hao Yang and X. Fan, "Position focused attention network for image-text matching," 2019.