Engineering Applications of Artificial Intelligence 116 (2022) 105419

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai

ASPD-Net: Self-aligned part mask for improving text-based person re-identification with adversarial representation learning

Zijie Wang [a], Jingyi Xue [a], Xili Wan [a,*], Aichun Zhu [a,b], Yifeng Li [a], Xiaomei Zhu [a], Fangqiang Hu [a]

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

ARTICLE INFO

Keywords:
Part mask detection
Text-based person re-identification
Adversarial learning

ABSTRACT

Text-based person re-identification aims to retrieve images of the corresponding person from a large visual database according to a natural language description. 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 multi-granular visual and textual features. A novel Self-aligned Part Mask Module was presented to autonomously learn the information of human body parts, and obtain visual local features in a soft-attention manner by using K Self-aligned Part Mask Detectors. Regarding the main model branches as a generator, a discriminator is employed to determine whether the representation vector comes from the visual modality or the textual modality. With Adversarial Loss training, ASPD-Net can learn more robust representations, as long as it successfully tricks the discriminator. Experimental results demonstrate that the proposed ASPD-Net outperforms the previous methods and achieves the state-of-the-art performance on the CUHK-PEDES and RSTPReid datasets.

1. Introduction

Given a textual query, text-based person re-identification (Niu et al., 2020; Chen et al., 2018b; Li et al., 2017b; Jing et al., 2020) aims to search for the video snapshot images of the corresponding pedestrian within a large visual database. This task is attracting more and more attention considering that in many criminal scenes textual descriptions may be the only accessible information to search for one certain pedestrian. Nevertheless, there still exist many challenges to be solved.

The major challenge of text-based person re-identification task is how to accurately 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 unique characteristics. Specifically, each image processed by the general cross-modal retrieval task commonly contains various categories of objects, and information carried by the query textual descriptions is to some extent crude or abstract. In contrast, each image concerned by the text-based person re-identification task contains just one specific pedestrian, while the textual description queries provide much more local details about the target person. The above mentioned particularity of text-based person re-identification caused that many previous methods proposed on general cross-modal retrieval benchmarks (e.g. Flickr30K Plummer et al., 2015 and MSCOCO Lin et al., 2014) generalize poorly on this task, and thereby fine-grained cross-modal information ought to be taken into consideration for superior performance.

Many of the existing methods (Niu et al., 2020; Jing et al., 2020; Wang et al., 2020b) employ multi-granular (global/local) cues from both the visual and textual modalities to improve the searching accuracy. Multi-granular clues of visual and textual modalities are widely used in existing methods to improve the accuracy of their retrieval. Considering the structural characteristics of the text data, phrases extracted from each sentence are utilized to obtain the fine-grained local information of text (Niu et al., 2020; Jing et al., 2020; Wang et al., 2020b, 2022), which is usually obtained by parsing, word segmentation and part-of-speech tagging in Natural Language Toolkit (NLTK). When it comes to the visual modal, some of the previous methods (Jing et al., 2020) 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

[*] Corresponding author.
E-mail address: xiliwan@njtech.edu.cn (X. Wan).

https://doi.org/10.1016/j.engappai.2022.105419
Received 16 January 2022; Received in revised form 15 August 2022; Accepted 31 August 2022
Available online xxxx

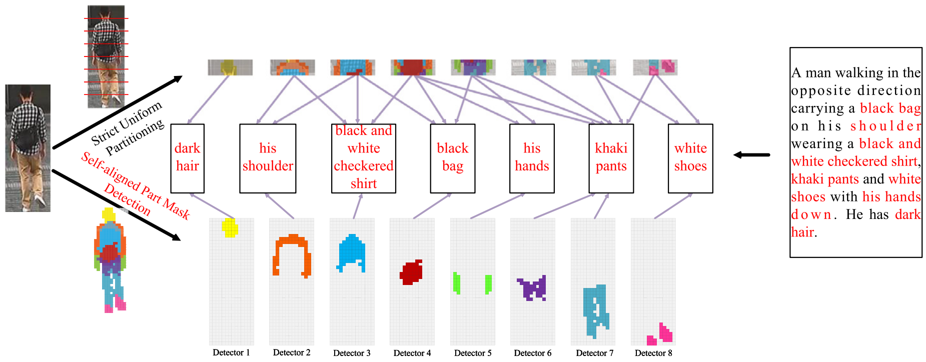

Strict Uniform Partitioning
Self-aligned Part Mask Detection
dark hair
his shoulder
black and white checkered shirt
black bag
his hands
khaki pants
white shoes
A man walking in the opposite direction carrying a black bag on his shoulder wearing a black and white checkered shirt, khaki pants and white shoes with his hands down. He has dark hair.
Detector 1
Detector 2
Detector 3
Detector 4
Detector 5
Detector 6
Detector 7
Detector 8

Fig. 1. Illustration of the effectiveness of the Strict Uniform Partitioning Strategy and the Self-aligned Part Mask Detection Strategy. With the Strict Uniform Partitioning Strategy, each stripe may contain multiple local parts while each of them are not complete, which can do harm to the performance. By employing K self-aligned part mask detectors, ASPD-Net is able to autonomously learn and extract more detailed human part features.

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 Sun et al. (2018), some other approaches (Niu et al., 2020; Wang et al., 2020b) 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. Since person re-identification is usually regarded as the next high-level mission of pedestrian detection task, this view is based on the assumption that the previous detection model are perfect, which is very difficult for most current detection models, thus 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, with the Strict Uniform Partitioning Strategy, each stripe may contain multiple local parts while each of them are not complete, which can do harm to the performance. By employing K self-aligned part mask detectors, ASPD-Net is able to autonomously learn and extract more detailed human part features. Therefore, each detected self-aligned part mask can capture more pure information about one certain local part, thereby greatly improving the retrieval accuracy. In addition, the major goal of the Text-based Person Re-identification task is to extract discriminative information from either the visual or the textual modality, which can be used for the subsequent similarity calculation step. Therefore, both visual and textual feature vector should include sufficient general information of the target person, instead of modality-specific information. 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 speaking, as long as ASPD-Net can successfully deceive the discriminator, it can extract more discriminative feature vectors. Our proposed method is evaluated on the CUHK-PEDES (Li et al., 2017b) and RSTPReid (Zhu et al., 2021a) datasets. Experimental results present that the proposed ASPD-Net outperforms the previous methods and achieves the state-of-the-art performance.

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 2 illustrates related work for person re-identification and text-based person re-identification. Section 3 introduces the proposed model. The experiments and the comparison results are provided in Section 4. Finally, Section 5 gives the conclusion of this paper.

2. Related works

2.1. Person re-identification

Person re-identification has drawn increasing attention in both academical and industrial fields (Sun et al., 2018; Zhong et al., 2019; Song et al., 2018; Wang et al., 2021b; Zhang et al., 2021b; Li et al., 2021; Chen et al., 2021; Zhang et al., 2021a; Liu et al., 2018; Li et al., 2017b; Su et al., 2017; Cheng et al., 2016; Yuan et al., 2020; Lu et al., 2020; Zhu et al., 2021b, 2020; Daihong et al., 2021, 2022). With the development of deep learning, deep learning methods are in general playing a major role in current state-of-the-art works. The success of deep learning in image classification spreads to re-identification (re-ID) in 2014, when Yi et al. (2014) firstly proposed deep learning methods which employ a siamese neural network to determine if a pair of input images belong to the same ID. The reason for choosing the siamese 1 is probably that the number of training samples for each identity is limited (usually two). Xia et al. (2019) 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. (2019) 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. Yuan et al. (2020) propose a Gabor convolution module for deep neural networks based on Gabor function, which has a good texture representation ability and is effective when it is embedded in the low layers of a network. Taking advantage of the hinge function, they also design a new regularizer loss function to make the proposed Gabor Convolution module meaningful. To bridge the gap between theoretical research and practical application, Zhang et al. (2019) propose a large and real-scenario person re-identification dataset for night scenario named KnightReid. Image denoising networks combined with common used person re-identification networks can be adapted to this kind of problem. Li et al. (2018) first formulate the open-world group-based person re-identification problem.

2.2. Body part-aligned representations

To address the misalignment of body parts in person re-identification, more and more methods are devoted to extracting local body features of pedestrians and generally can be divided in to two categories. One category is to locate body parts by using existing human pose estimation algorithms. Wei et al. (2016) first introduce Convolutional Pose Machines (CPMs) to predict human body joints and generate body regions to extracts part features. Zhao et al. (2017b) proposed a novel network named Spindle Net which is the first time human body structure information is considered in a CNN framework to facilitate feature learning. Zheng et al. (2019) build PoseBox fusion (PBF) to reduce the impact of pose estimation errors and detail loss.

The other category is to locate human body parts or silent regions based on attention models, which can be considered as an unsupervised part detection method. Zhao et al. (2017a) learn several human part maps supervised only by triplet loss for person re-identification. Yang et al. (2019) design a novel multi-branch attention-driven network that simultaneously learns and fuses discriminative and complementary features from both global whole-body and local body-part images. Zhao et al. (2021) introduced a deep part-aware representation learning method which employed a localization branch to drive the model to focus on discriminative parts.

2.3. Text-based person re-identification

Text-based person re-identification has been studied from various perspective (Niu et al., 2020; Chen et al., 2018b; Li et al., 2017b; Jing et al., 2020; Wang et al., 2022). It is challenging to directly measure the affinity between images and descriptions, on account of the cross-modality heterogeneity. Li et al. (2017b) 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 Li et al. (2017a), 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. (2018b) 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. Sarafianos et al. (2019) 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. (2020) 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 (Jing et al., 2020) 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. An IMG-Net model is proposed by Wang et al. (2020b) to incorporate inner-modal self-attention and cross-modal hard-region attention with the fine-grained model for extracting the multi-granular semantic information. Liu et al. (2019) generate fine-grained structured representations from images and texts of pedestrians with an A-GANet model to exploit semantic scene graphs. Wang et al. (2020a) proposed a novel model called ViTAA which learns to disentangle the feature space of a person into sub-spaces corresponding to attributes using a light auxiliary attribute segmentation layer. A cross-modal momentum contrastive learning framework is introduced by Han et al. (0000) to enrich the training data for a given mini-batch. To effectively transfer the knowledge learned from large-scale generic image-text pairs, they proposed to perform cross-modal pre-training, but for the text modality, only word embedding is transferred. In order to facilitate the practical application, Chen et al. (0000) proposed a simple but effective framework for text-based person retrieval named TIPCB. In contrast to the existing local-matching methods, the structure of TIPCB does not need any additional models and complex evaluation strategies. Wang et al. (2022) propose a Serialized Updating and Matching (SUM) method for text-based person retrieval to bridge the heterogeneity gap between cross-modal data in a step-by-step manner. A novel model named SSAN is proposed by Ding et al. (0000) which does not split the textual description or perform cross-modal operations. SSAN explores the relatively well aligned body parts in images as supervision and utilizes the contextual cues in language descriptions to achieve this goal. A new approach CMAAM is introduced by Aggarwal et al. (2020) which learns an attribute-driven space along with a class-information driven space by introducing extra attribute annotation and prediction. Zheng et al. (2020a) propose a Gumbel attention module to alleviate the matching redundancy problem and a hierarchical adaptive matching model is employed to learn subtle feature representations from three different granularities. In this paper, by proposing a Self-aligned Part Mask Module, ASPD-Net is able 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.

3. Methodology

In this section, we introduce the Adversarial Self-aligned Part Detecting Network (ASPD-Net) (shown in Fig. 2) in detail.

3.1. Problem formulation

The aim of the proposed framework is to measure the similarity between a given textual description and a gallery pedestrian image (pedestrian images in the large scale person image database). Formally, let $\mathcal{D} = \{i_i, t_i\}_{i=1}^{N}$ be a training set containing N image-text pairs. Each pair consists of a person image captured by one certain surveillance camera and its corresponding textual description. The IDs of pedestrian in $\mathcal{D}$ are $Y = \{y_i\}_{i=1}^{Q}$. Given a textual description, the objective is to identify images of the most relevant pedestrian from a large scale person image gallery.

3.2. Overall architecture

In this work, we propose a novel Adversarial Self-aligned Part Detecting Network (ASPD-Net) model to deal with the text-based person re-identification task. As shown in Fig. 2, ASPD-Net has two branches to separately extract visual and textual features. In order to utilize overall and detailed information more properly for a high retrieval accuracy, both global and local features are extracted from the visual

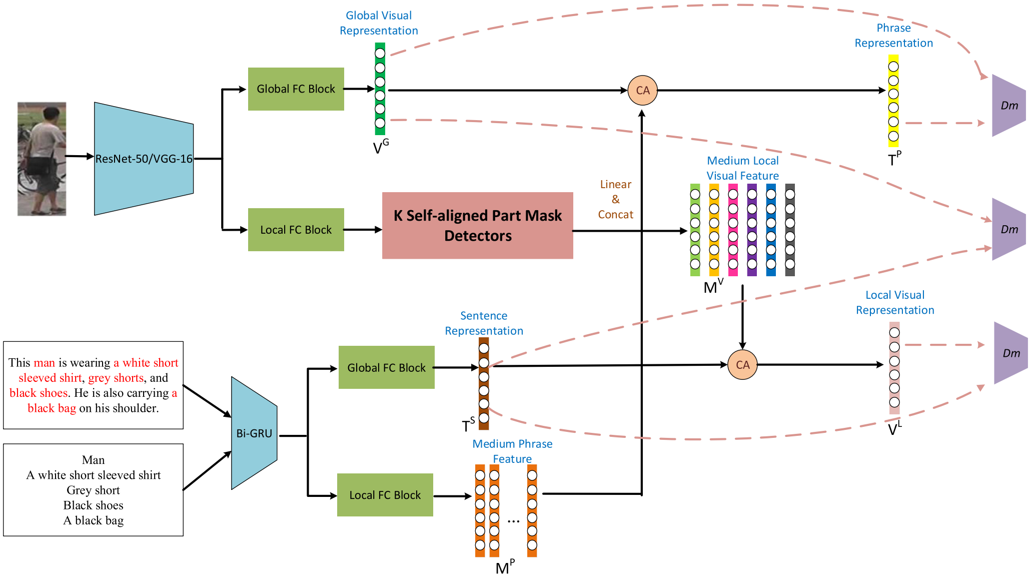

Global Visual Representation
Phrase Representation
ResNet-50/VGG-16
Global FC Block
V^G
CA
T^P
Dm
Local FC Block
K Self-aligned Part Mask Detectors
Linear & Concat
Medium Local Visual Feature
M^V
Dm
Sentence Representation
Local Visual Representation
This man is wearing a white short sleeved shirt, grey shorts, and black shoes. He is also carrying a black bag on his shoulder.
Bi-GRU
Global FC Block
T^S
CA
V^L
Dm
Man
A white short sleeved shirt
Grey short
Black shoes
A black bag
Medium Phrase Feature
Local FC Block
...
M^P

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. With the proposed Self-aligned Part Mask Module, human part information can be learned autonomously, which enables ASPD-Net to extract the visual local features in a soft-attentional manner. D_m denotes the discriminator for modality while CA denotes the proposed Cross-modal Attention mechanism.

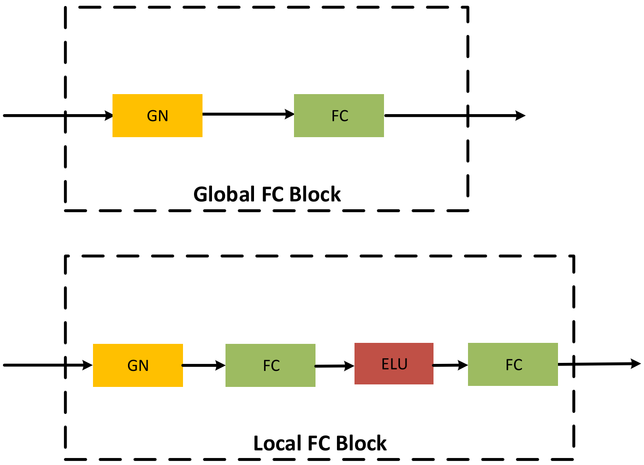


Fig. 3. Detailed structures of the Global and Local FC modules.

and textual modalities. Given an input person image, a P-dimensional (P-dim) global visual representation vector $V^G \in \mathbb{R}^P$ and K P-dim intermediate local visual feature vectors $M^V \in \mathbb{R}^{P \times K}$ are mainly extracted via a VGG-16 (Simonyan and Zisserman, 0000) or ResNet-50 (He et al., 2016) backbone and a global fully connected (FC) block (detailed in Fig. 3). In particular, with a proposed novel Self-aligned Part Mask Module to learn human part information autonomously, the visual local features can be obtained in a soft-attentional manner by employing K Self-aligned Part Mask Detectors. In terms of textual feature extraction, a whole sentence and the n phrases extracted from it are taken as global and local textual materials, which are processed by a shared bi-directional gated recurrent unit (bi-GRU). The last hidden states of the forward and backward GRUs are concatenated and then respectively passed a FC block to give a P-dim sentence representation vector $T^S \in \mathbb{R}^P$ and m P-dim intermediate phrase feature vectors $M^P \in \mathbb{R}^{P \times m}$. Then, the two intermediate local feature matrices are converted into two P-dim local representation vectors, that is, the local visual representation vector $V^L \in \mathbb{R}^P$ and the phrase representation vector $T^P \in \mathbb{R}^P$, by using cross-modal attention method.

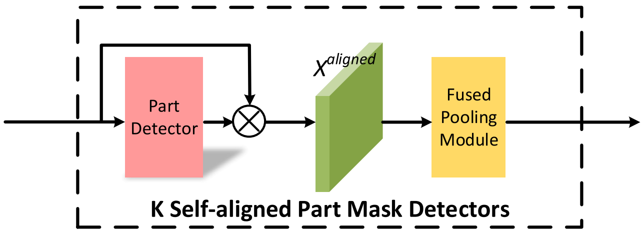


Fig. 4. Detail of the self-aligned part mask module.

3.3. Self-aligned part mask detection

As discussed in the Introduction, it may not be a good choice to roughly divide the picture to extract local features in some cases, which may result in the loss of some information. 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 (illustrated in Fig. 4). 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 $M_i \in \mathbb{R}^{h \times w}, i \in \{1, 2, \ldots, K\}$. The detectors are implemented as a 1 × 1 conv layer followed by a Sigmoid layer:

$$M_i = Sigmoid(Conv_{1\times1}(X)). \tag{1}$$

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 $X_i^{aligned} \in \mathbb{R}^{h \times w \times c}, i \in \{1, 2, \ldots, K\}$ are obtained by a Hadamard product:

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

3.4. ASPD-Net

Since the job of text-based person re-identification has two modalities, namely, visual modality and textual modality, ASPD-Net also contains two branches to extract visual and textual features respectively. 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.

3.4.1. 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 $W_e \in \mathbb{R}^{E \times W}$ denotes the embedding layer. 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

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

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

where $\overrightarrow{GRU}$ and $\overleftarrow{GRU}$ relatively denote the forward and backward GRUs and h_t stands for the hidden state of GRU in the tth step. $t \in \{1, \dots, n\}$, 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:

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

where $e^k \in \mathbb{R}^P$, 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 $e_i^P \in \mathbb{R}^P, i \in \{1, \dots, m\}$, 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 $M^P \in \mathbb{R}^{P \times n}$.

3.4.2. 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 $K \times 1 \times c$:

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

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 $\phi(I) \in \mathbb{R}^{K \times 1 \times c}$. The $\phi(I)$ is flattened to a $(K \times c)$-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 $X_i^{aligned} \in \mathbb{R}^{h \times w \times c}, i \in \{1, 2, \dots, K\}$. 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 ELU layer between them to form the medium local visual vectors $M_i^V \in \mathbb{R}^P, i \in \{1, 2, \dots, K\}$. After that, we concatenate them to form the medium local visual feature matrix $M^V \in \mathbb{R}^{K \times P}$.

3.4.3. Local representation vectors converting

In order to convert the two medium feature matrices M^V and M^P to the corresponding representation vectors, we first adopted the cross-modal attention method to calculate how much each column in one feature matrix is related to the representation vector of the other 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:

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

where α_i^V represents the relation between the ith 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 $V^L \in \mathbb{R}^P$:

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

Similarly, the phrase representation vector T^P is given out by

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

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

3.5. Loss functions and training strategy

To train ASPD-Net, 3 different loss functions is adopted, including identification (ID) loss, triplet ranking loss and adversarial loss (shown in Fig. 5). The complete training process contains two stages.

3.5.1. Stage-1

First, we fix the parameters of the visual backbone, 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

$$L_{id}^V = -log(softmax(W_{id} \times GN(V^G))), \tag{12}$$

$$L_{id}^T = -log(softmax(W_{id} \times GN(T^S))), \tag{13}$$

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:

$$L_{id} = L_{id}^V + L_{id}^T. \tag{14}$$

Thus, the final loss function in Stage-I is

$$L_{stage1} = L_{id}. \tag{15}$$

3.5.2. 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 subsequent similarity measuring step, it is reasonable that representations from both the visual and the

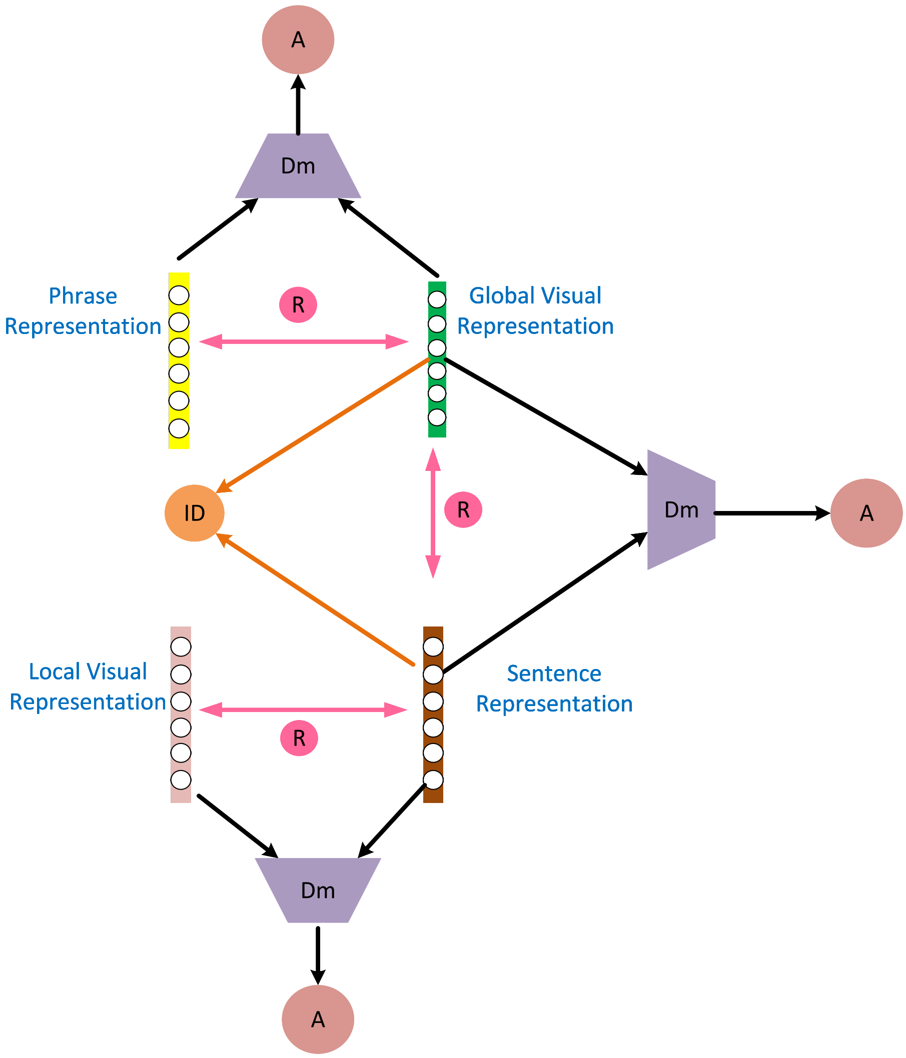


Fig. 5. Illustration of the loss functions utilized to train ASPD-Net on CUHK-PEDES. ID Loss, Ranking Loss and Adversarial Loss are employed to train ASPD-Net, which are denoted as ID, R and A in circles, respectively. D_m denotes the discriminator for modality.

textual modality should include sufficient general information about the targeted person, rather than information with respect to a single modality. In other words, ideally, we think that 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 ELU 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:

$$L^k_{adversarial} = -\mathop{\mathbb{E}}_{V_i \sim V}[log D(V_i)] - \mathop{\mathbb{E}}_{T_i \sim T}[1 - log D(T_i)], \tag{16}$$

where $L^k_{adversarial}$ denotes $L^{GS}_{adversarial}$, $L^{LS}_{adversarial}$ or $L^{GP}_{adversarial}$. V can be V^G or V^L while T can be T^S or T^P respectively according to $L^k_{adversarial}$. 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 $V^G \sim T^P$:

$$L_{adversarial} = L^{GS}_{adversarial} + L^{LS}_{adversarial} + L^{GP}_{adversarial}. \tag{17}$$

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 Faghri et al. (2018), 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:

$$L^k_{ranking} = \sum_{\hat{T}} max\{\alpha - cos(V,T) + cos(V,\hat{T}), 0\} + \sum_{\hat{V}} max\{\alpha - cos(V,T) + cos(\hat{V},T), 0\}, \tag{18}$$

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 $(V,\hat{T})$ or $(\hat{V},T)$ denotes the mismatched pairs and α is a margin. The general triplet ranking loss is calculated following:

$$L_{ranking} = L^{GS}_{ranking} + L^{LS}_{ranking} + L^{GP}_{ranking}. \tag{19}$$

The complete loss function in Stage-2 is

$$L_{stage2} = L_{id} + L_{ranking} + L_{adversarial}. \tag{20}$$

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.

3.5.3. Image-text matching for inference

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:

$$Simi_{vt} = cos(V^v, T^t), \tag{21}$$

where v can be either G or L 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:

$$Simi_{ASPDNet} = Simi_{GS} + \frac{1}{2}(Simi_{LS} + Simi_{GP}). \tag{22}$$

4. Experiments

4.1. Experimental setup

4.1.1. Dataset and metrics

Our approach is evaluated on two challenging Text-based Person Retrieval datasets, including CUHK-PEDES (Li et al., 2017b) and RSTPReid (Zhu et al., 2021a).

CUHK-PEDES. Following the official data split approach (Li et al., 2017b), the training set of CUHK-PEDES contains 34 054 images, 11 003 persons and 68 126 textual descriptions. The validation set contains 3078 images, 1000 persons and 6158 textual descriptions while the test set has 3074 images, 1000 persons and 6156 descriptions.

Rstpreid. The RSTPReid dataset contains 20 505 images of 4101 persons. Each person has 5 corresponding images taken by different cameras and each image is annotated with 2 textual descriptions. For data division, 3701, 200 and 200 identities are utilized for training, validation and test, respectively.

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.

4.1.2. 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. The input images are resized to 384 × 128 × 3. The random cropping, random horizontal flipping, random erasing, label smoothing regularization strategies are employed for data augmentation. 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.

Table 1
Ablation analysis of granularity combination on CUHK-PEDES. Top-1, top-5 and top-10 accuracies are reported. The best results are marked bold.

Method	Backbone	Top-1	Top-5	Top-10
GS	VGG-16	48.71	71.23	79.27
LS	VGG-16	48.99	71.94	79.97
GP	VGG-16	44.37	68.71	77.65
GS-LS	VGG-16	53.25	75.81	83.39
GS-GP	VGG-16	52.48	75.00	82.31
LS-GP	VGG-16	55.32	76.34	83.91
ALL	VGG-16	**55.44**	**76.53**	**84.21**
GS	ResNet-50	52.97	74.34	82.23
LS	ResNet-50	53.16	72.56	80.26
GP	ResNet-50	48.13	71.88	80.19
GS-LS	ResNet-50	56.63	79.22	86.49
GS-GP	ResNet-50	55.09	78.86	85.17
LS-GP	ResNet-50	58.75	79.24	86.93
ALL	ResNet-50	**59.32**	**80.11**	**87.41**

Table 2
Ablation analysis of the number K of Self-aligned Part Masks on CUHK-PEDES. Top-1, top-5 and top-10 accuracies are reported. The best results are marked bold.

K	Backbone	Top-1	Top-5	Top-10
2	VGG-16	52.83	74.83	82.48
4	VGG-16	54.82	76.34	83.91
6	VGG-16	**55.44**	**76.53**	**84.21**
8	VGG-16	55.39	76.49	84.14
10	VGG-16	55.12	76.38	84.32
2	ResNet-50	55.96	79.11	86.43
4	ResNet-50	57.81	79.56	86.86
6	ResNet-50	58.15	79.63	87.34
8	ResNet-50	**59.32**	**80.11**	**87.41**
10	ResNet-50	59.01	80.08	87.17

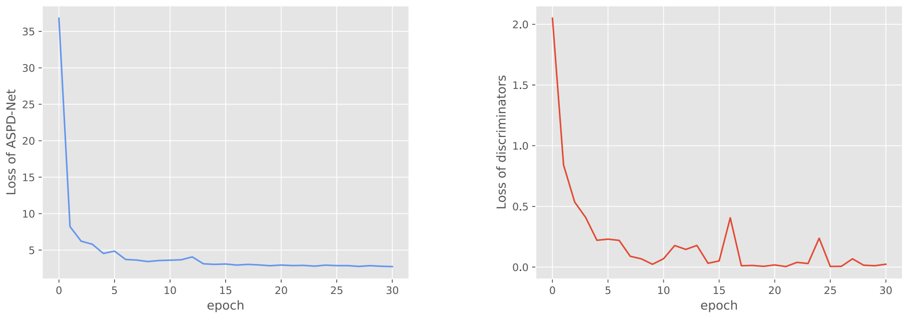
In training, we initialize the weights of the visual CNN backbone pre-trained on the ImageNet classification task. An Adam optimizer is adopted to train the model with a batch size of 32. The margin α of ranking losses is set to 0.2. In training stage-1, we start the iteration with a learning rate of 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:

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

where lr means the learning rate and $\lfloor \cdot \rfloor$ denotes the operation of taking the integer part. We totally train the stage-2 for 30 epochs. The overall training procedure of ASPD-Net takes about three days on NVIDIA 1080Ti. We illustrate the training process of our proposed ASPD-Net in Fig. 6.

4.2. Ablation analysis

To further investigate several components of ASPD-Net, we carry out plenty of ablation experiments. As shown in Table 1, Tables 4 and 2, 'GS', 'LS' or 'GP' denote respectively the global2sentence, local2sentence or global2phrase combination is utilized, while '-ALL' means all of them are employed. 'AP', 'MP', 'FP' and 'AL' denote whether Average Pooling, Maximum Pooling, Fused Pooling and Adversarial Loss is utilized.



(a) Loss of ASPD

(b) Loss of the discriminators

Fig. 6. Illustration of the training process of our proposed ASPD-Net on CUHK-PEDES. (a) The variation curve of the loss for training ASPD-Net. (b) The variation curve of the loss for training the discriminators.

Table 3
Comparison between ASPD-Net and a baseline utilizing the traditional Strict Uniform Partitioning Strategy on CUHK-PEDES. Top-1, top-5 and top-10 accuracies are reported. The best results are marked bold.

Method	Backbone	Top-1	Top-5	Top-10
Baseline	VGG-16	51.46	76.07	83.34
ASPD-Net	VGG-16	**55.44**	**76.53**	**84.21**
Baseline	ResNet-50	54.32	77.13	85.24
ASPD-Net	ResNet-50	**59.32**	**80.11**	**87.41**

Table 4
Comparison of key components on CUHK-PEDES. Top-1, top-5 and top-10 accuracies are reported. The best results are marked bold.

AP	MP	FP	AL	Backbone	Top-1	Top-5	Top-10
✓	×	×	✓	ResNet-50	58.54	79.81	87.27
×	✓	×	✓	ResNet-50	58.85	80.02	87.33
×	×	✓	✓	ResNet-50	**59.32**	**80.11**	**87.41**
×	×	✓	×	VGG-16	54.84	76.49	**84.42**
×	×	✓	✓	VGG-16	**55.44**	**76.53**	84.21
×	×	✓	×	ResNet-50	58.56	79.81	87.12
×	×	✓	✓	ResNet-50	**59.32**	**80.11**	**87.41**

4.2.1. Combination of granularities

Table 1 provides analysis on the effect of each granularity and the way they are combined. The results show that utilizing 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 outperforms any other model utilizes part of the three granularities. What is 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.

4.2.2. 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 2. As can be observed from the data, 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 is 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 test set are shown in Fig. 7. As can be seen, our self-learned detector of a full-trained ASPD-Net 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.

4.2.3. Comparison between self-aligned part mask module and strict uniform partitioning strategy

As shown in Table 3, comparing with a baseline adopting the traditional Strict Uniform Partitioning Strategy when handling local visual information, which is proposed following IMG-Net (Wang et al., 2020b) without the Inner-Modal Self-Attention Module and is also utilized in MIA (Niu et al., 2020), ASPD-Net outperforms the baseline by 3.98%, 0.46%, 0.87% and 5.00%, 2.98% and 2.17% under the top-1, top-5 and top-10 accuracies with VGG-16 and ResNet-50 as the visual backbone, respectively. The increase in model performance proves that our proposed method is able to autonomously catch more detailed human part features by employing K Self-aligned Part Mask Detectors than roughly partitioning the feature maps uniformly.

4.2.4. Effectiveness of adversarial mechanism

As is shown in Table 4, with the assistance of the adversarial loss, the top-1 accuracy of ASPD-Net increases from 54.84% to 55.44% and from 58.56% to 59.32% with VGG-16 and ResNet-50 as visual CNN backbone respectively. The results prove the effectiveness of the proposed Adversarial Mechanism. Moreover, we display some examples of self-aligned part masks learned by ASPD-Net for some person images with/without the adversarial mechanism in Fig. 8. It is obvious that with the aid of the proposed adversarial mechanism, ASPD-Net is enabled to catch much more accurate and well-defined part masks, which is reasonable. The adversarial mechanism aims to extract modality-invariant features from both visual and textual modalities. The modality discriminator is proposed to decide which modality the input feature vector comes from and the aim of training ASPD-Net is to successfully deceive the discriminator. To achieve this goal, ASPD-Net is supposed to focus more on information about the identification

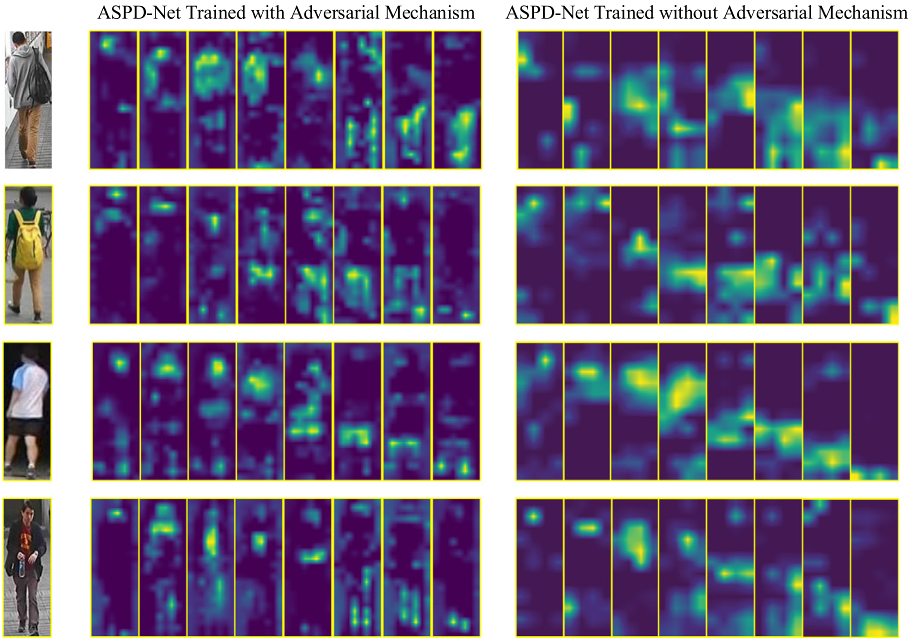
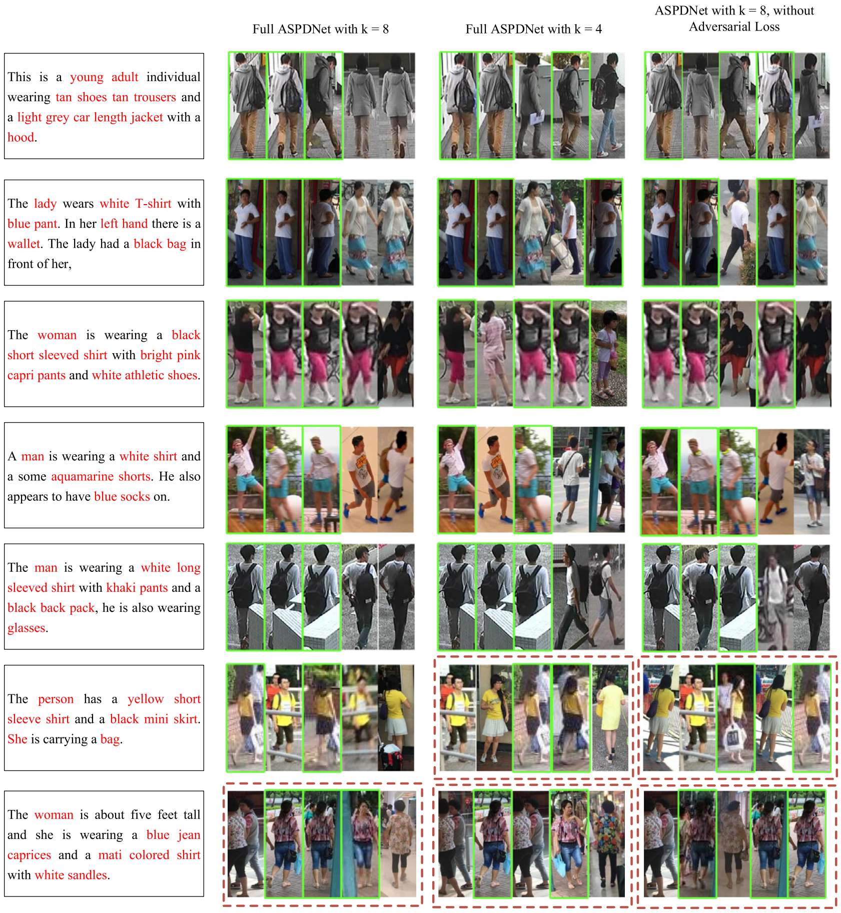

ASPD-Net Trained with Adversarial Mechanism
ASPD-Net Trained without Adversarial Mechanism

Fig. 7. Examples of self-aligned part masks learned by ASPD-Net for some images with/without the adversarial mechanism on CUHK-PEDES. 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 first person and the 4th mask of the second image, respectively. Other key parts contributing to the matching process like head, limbs, foot, etc. are detected properly as well. Moreover, it is obvious that with the aid of the proposed adversarial mechanism, ASPD-Net is enabled to catch much more accurate and well-defined part masks.



This is a young adult individual wearing tan shoes tan trousers and a light grey car length jacket with a hood.

The lady wears white T-shirt with blue pant. In her left hand there is a wallet. The lady had a black bag in front of her,

The woman is wearing a black short sleeved shirt with bright pink capri pants and white athletic shoes.

A man is wearing a white shirt and a some aquamarine shorts. He also appears to have blue socks on.

The man is wearing a white long sleeved shirt with khaki pants and a black back pack, he is also wearing glasses.

The person has a yellow short sleeve shirt and a black mini skirt. She is carrying a bag.

The woman is about five feet tall and she is wearing a blue jean caprices and a mati colored shirt with white sandles.

Fig. 8. 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 on CUHK-PEDES. All the three ASPD-Net proposed here utilize ResNet-50 as visual backbone. Images of the target person are marked by green rectangles and the failed cases in which the top-1 images does not belong to the target person is marked by dotted dark red rectangles.

Table 5
Comparison with other state-of-the-art methods on CUHK-PEDES. Top-1, top-5 and top-10 accuracies are reported. The best results are marked bold.

Method	Backbone	Top-1	Top-5	Top-10
CNN-RNN (Reed et al., 2016)	VGG-16	8.07	–	32.47
Neural Talk (Vinyals et al., 2015)	VGG-16	13.66	–	41.72
GNA-RNN (Li et al., 2017b)	VGG-16	19.05	–	53.64
IATV (Li et al., 2017a)	VGG-16	25.94	–	60.48
PWM-ATH (Chen et al., 2018a)	VGG-16	27.14	49.45	61.02
Dual Path (Zheng et al., 2020b)	VGG-16	32.15	54.42	64.30
GALM (Jing et al., 2020)	VGG-16	47.82	69.83	78.31
MIA (Niu et al., 2020)	VGG-16	48.00	70.70	79.30
IMG-Net (Wang et al., 2020b)	VGG-16	54.32	75.93	84.21
ASPD-Net(ours)	VGG-16	**55.44**	**76.53**	**84.21**
Dual Path (Zheng et al., 2020b)	ResNet-50	44.40	66.26	75.07
GLA (Chen et al., 2018a)	ResNet-50	43.58	66.93	76.26
MIA (Niu et al., 2020)	ResNet-50	53.10	75.00	82.90
GALM (Jing et al., 2020)	ResNet-50	54.12	75.45	82.97
TIMAM (Sarafianos et al., 2019)	ResNet-101	54.51	77.56	84.78
IMG-Net (Wang et al., 2020b)	ResNet-50	56.48	76.89	85.01
CMAAM (Aggarwal et al., 2020)	ResNet-50	56.68	77.18	84.86
HGAN (Zheng et al., 2020a)	ResNet-50	59.00	79.49	86.6
ASPD-Net(ours)	ResNet-50	**59.32**	**80.11**	**87.41**

Table 6
Comparison with other state-of-the-art methods on RSTPReid. Top-1, top-5 and top-10 accuracies are reported. The best results are marked bold.

Method	Backbone	Top-1	Top-5	Top-10
IMG-Net (Wang et al., 2020b)	ResNet-50	37.60	61.15	73.55
AMEN (Wang et al., 2021a)	ResNet-50	38.45	62.40	73.80
DSSL (Zhu et al., 2021a)	ResNet-50	39.05	62.60	73.95
ASPD-Net(ours)	ResNet-50	**39.90**	**63.15**	**74.40**

of a certain person and drop modality-specific interference information during the feature extraction process. Therefore, it is natural that ASPD-Net can generate more detailed self-align part masks to catch more effective information to benefit the following feature matching process. Except the visualized local part masks, under the constraints formed by the proposed adversarial mechanism, all multi-granular visual and textual features extracted by ASPD-Net can be much more modality-invariant and discriminative as long as ASPD-Net is able to deceive the discriminator successfully.

4.2.5. *Fused pooling module*

Table 4 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 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.

4.2.6. *Analysis of retrieval results*

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. 8. The failed cases in which the top-1 images does not belong to the target person is marked by dotted dark red rectangles. As can be seen in the figure, images of the target person in the candidate list given by models trained with different k or trained with/without Adversarial Loss may appear in varied orders. This is a reasonable phenomenon due to the non-convex optimization nature of the training of deep learning models. Though given in varied orders, target person images always can be retrieved properly as long as the key information is caught by the models. Besides, person in the top mismatched images (including in some of the top-1 mismatched cases) also look quite similar (especially w.r.t color information) to the target one. It seems necessary to dig deeper into the semantic information and find ways to better catch discriminative details like textual information, structural information, etc., which remains for our future work.

4.3. *Comparison with other state-of-the-art methods*

The comparisons with other state-of-the-art methods on CUHK-PEDES and RSTPReid are respectively shown in Tables 5 and 6. 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 28.30% under top-1 metric, which validates the significance of the local2sentence and global2phrase granularities in our method. Compared with MIA using VGG-16 as visual backbone, ASPD-Net significantly outperforms it by 7.44% under top-1 metric, indicating the superiority of the self-align part mask detector module and the adversarial loss. Besides, ASPD-Net outperforms IMG-Net, which is the best competitor with VGG-16 backbone, with 1.12% under top-1 metric. With the VGG-16 backbone, ASPD-Net even outperforms TIMAM, the currently best competitor with ResNet-101 backbone by 0.93% under top-1 metric. With ResNet-50 backbone, ASPD-Net achieves the best retrieval performance compared with all of the previous methods.

5. 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 and RSTPReid datasets and the results indicate that the proposed ASPD-Net improves the performance with a large margin. Besides, as can be observed from some of the retrieval result examples, person in the top mismatched images (including in some of the top-1 mismatched cases) also look quite similar (especially w.r.t color information) to the target one. It seems necessary to dig deeper into the semantic information and find ways to better catch discriminative details like textual information, structural information, etc., which remains for our future work.

CRediT authorship contribution statement

Zijie Wang: Software, Writing – original draft. **Jingyi Xue:** Software, Validation. **Xili Wan:** Conceptualization, Methodology. **Aichun Zhu:** Writing – review & editing. **Yifeng Li:** Supervision, Resources. **Xiaomei Zhu:** Data curation, Visualization. **Fangqiang Hu:** Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data that has been used is confidential.

Acknowledgments

This work is partially supported by the National Natural Science Foundation of China (Grant No. 62101245), China Postdoctoral Science Foundation (Grant No. 2019M661999), Natural Science Research of Jiangsu Higher Education Institutions of China (19KJB520009) and Future Network Scientific Research Fund Project, China (Grant No. FNSRFP-2021-YB-21).

References

Aggarwal, S., Radhakrishnan, V.B., Chakraborty, A., 2020. Text-based person search via attribute-aided matching. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 2617–2625.

Chen, D., Li, H., Liu, X., Shen, Y., Shao, J., Yuan, Z., Wang, X., 2018a. 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), pp. 54–70.

Chen, F., Wang, N., Tang, J., Liang, D., 2021. A negative transfer approach to person re-identification via domain augmentation. Inform. Sci. 549, 1–12.

Chen, T., Xu, C., Luo, J., 2018b. Improving text-based person search by spatial matching and adaptive threshold. In: 2018 IEEE Winter Conference on Applications of Computer Vision. WACV, pp. 1879–1887.

Chen, Y., Zhang, G., Lu, Y., Wang, Z., Zheng, Y., 0000. Tipcb: A simple but effective part-based convolutional baseline for text-based person search, Neurocomputing.

Cheng, D., Gong, Y., Zhou, S., Wang, J., Zheng, N., 2016. 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, pp. 1335–1344.

Daihong, J., Lei, D., Jin, P., et al., 2021. Facial expression recognition based on attention mechanism. Sci. Program..

Daihong, J., Sai, Z., Lei, D., Yueming, D., 2022. Multi-scale generative adversarial network for image super-resolution. Soft Comput. 26 (8), 3631–3641.

Ding, Z., Ding, C., Shao, Z., Tao, D., 0000. Semantically self-aligned network for text-to-image part-aware person re-identification, arXiv preprint arXiv:2107.12666.

Faghri, F., Fleet, D.J., Kiros, J.R., Fidler, S., 2018. Vse++: Improving visual-semantic embeddings with hard negatives. In: Proceedings of the British Machine Vision Conference (BMVC).

Han, X., He, S., Zhang, L., Xiang, T., 0000. Text-based person search with limited data. arXiv preprint arXiv:2110.10807.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778.

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

Jing, Y., Si, C., Wang, J., Wang, W., Wang, L., Tan, T. and, 2020. Pose-guided multi-granularity attention network for text-based person search. In: Proceedings of the AAAI Conference on Artificial Intelligence, 34, pp. 11189–11196.

Li, H., Pang, J., Tao, D., Yu, Z., 2021. Cross adversarial consistency self-prediction learning for unsupervised domain adaptation person re-identification. Inform. Sci. 559, 46–60.

Li, X., Wu, A., Zheng, W.-S., 2018. Adversarial open-world person re-identification. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 280–296.

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

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

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2014. Microsoft coco: Common objects in context. In: European Conference on Computer Vision. Springer, pp. 740–755.

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

Liu, J., Zha, Z.-J., Hong, R., Wang, M., Zhang, Y., 2019. Deep adversarial graph attention convolution network for text-based person search. In: Proceedings of the 27th ACM International Conference on Multimedia, pp. 665–673.

Lu, Y., Wu, Y., Liu, B., Zhang, T., Li, B., Chu, Q., Yu, N., 2020. Cross-modality person re-identification with shared-specific feature transfer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13379–13389.

Niu, K., Huang, Y., Ouyang, W., Wang, L., 2020. Improving description-based person re-identification by multi-granularity image-text alignments. IEEE Trans. Image Process. 29, 5542–5556.

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

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

Sarafianos, N., Xu, X., Kakadiaris, I.A., 2019. Adversarial representation learning for text-to-image matching. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 5814–5824.

Simonyan, K., Zisserman, A., 0000. Very deep convolutional networks for large-scale image recognition, CoRR abs/1409.1556.

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

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

Sun, Y., Zheng, L., Yang, Y., Tian, Q., Wang, S., 2018. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In: Proceedings of the European conference on computer vision (ECCV), pp. 480–496.

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

Wang, Z., Fang, Z., Wang, J., Yang, Y., 2020a. Vitaa: Visual-textual attributes alignment in person search by natural language. In: European Conference on Computer Vision. Springer, pp. 402–420.

Wang, Z., Xue, J., Zhu, A., Li, Y., Zhang, M., Zhong, C., 2021a. Amen: Adversarial multi-space embedding network for text-based person re-identification. In: Chinese Conference on Pattern Recognition and Computer Vision. PRCV, Springer, pp. 462–473.

Wang, J., Yuan, L., Xu, H., Xie, G., Wen, X., 2021b. Channel-exchanged feature representations for person re-identification. Inform. Sci. 562, 370–384.

Wang, Z., Zhu, A., Xue, J., Jiang, D., Liu, C., Li, Y., Hu, F., 2022. Sum: Serialized updating and matching for text-based person retrieval. Knowl.-Based Syst. 248, 108891.

Wang, Z., Zhu, A., Zheng, Z., Jin, J., Xue, Z., Hua, G., 2020b. Img-net: inner-cross-modal attentional multigranular network for description-based person re-identification. J. Electron. Imaging 29 (4), 043028.

Wei, S.-E., Ramakrishna, V., Kanade, T., Sheikh, Y., 2016. Convolutional pose machines. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pp. 4724–4732.

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

Yang, F., Yan, K., Lu, S., Jia, H., Xie, X., Gao, W., 2019. Attention driven person re-identification. Pattern Recognit. 86, 143–155.

Yi, D., Lei, Z., Liao, S., Li, S.Z., 2014. Deep metric learning for person re-identification. In: 2014 22nd International Conference on Pattern Recognition. IEEE, pp. 34–39.

Yuan, Y., Zhang, J., Wang, Q., 2020. Deep gabor convolution network for person re-identification. Neurocomputing 378, 387–398.

Zhang, Y., Ma, B., Feng, Y., Li, M., 2021a. Pmt-net: Progressive multi-task network for one-shot person re-identification. Inform. Sci. 568, 133–146.

Zhang, G., Yang, J., Zheng, Y., Wang, Y., Wu, Y., Chen, S., 2021b. Hybrid-attention guided network with multiple resolution features for person re-identification. Inform. Sci. 578, 525–538.

Zhang, J., Yuan, Y., Wang, Q., 2019. Night person re-identification and a benchmark. IEEE Access 7, 95496–95504.

Zhao, L., Li, X., Zhuang, Y., Wang, J., 2017. Deeply-learned part-aligned representations for person re-identification. In: Proceedings of the IEEE international conference on computer vision, pp. 3219–3228.

Zhao, Y., Shen, C., Yu, X., Chen, H., Gao, Y., Xiong, S., 2021. Learning deep part-aware embedding for person retrieval. Pattern Recognit. 116, 107938.

Zhao, H., Tian, M., Sun, S., Shao, J., Yan, J., Yi, S., Wang, X., Tang, X., 2017. Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1077–1085.

Zheng, L., Huang, Y., Lu, H., Yang, Y., 2019. Pose-invariant embedding for deep person re-identification. IEEE Trans. Image Process. 28 (9), 4500–4509.

Zheng, K., Liu, W., Liu, J., Zha, Z.-J., Mei, T., 2020a. Hierarchical gumbel attention network for text-based person search. In: Proceedings of the 28th ACM International Conference on Multimedia, pp. 3441–3449.

Zheng, Z., Zheng, L., Garrett, M., Yang, Y., Xu, M., Shen, Y.-D., 2020b. Dual-path convolutional image-text embeddings with instance loss. ACM Trans. Multimedia Comput., Commun., Appl. (TOMM) 16 (2), 1–23.

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

Zhu, A., Wang, Z., Li, Y., Wan, X., Jin, J., Wang, T., Hu, F., Hua, G., 2021. Dssl: Deep surroundings-person separation learning for text-based person retrieval. In: Proceedings of the 29th ACM International Conference on Multimedia, pp. 209–217..

Zhu, A., Wu, Q., Cui, R., Wang, T., Hang, W., Hua, G., Snoussi, H., 2020. Exploring a rich spatial–temporal dependent relational model for skeleton-based action recognition by bidirectional lstm-cnn. Neurocomputing 414, 90–100.

Zhu, A., Zheng, Z., Huang, Y., Wang, T., Jin, J., Hu, F., Hua, G., Snoussi, H., 2021. Cacrowdgan: Cascaded attentional generative adversarial network for crowd counting, IEEE Transactions on Intelligent Transportation Systems.