Text-Guided Visual Feature Refinement for Text-Based Person Search

Liying Gao, Kai Niu*, Zehong Ma, Bingliang Jiao, Tonghao Tan, Peng Wang*
School of Computer Science and Engineering
Northwestern Polytechnical University
Xi'an, PR China, 710072
gaoliying,npumzh,bingliang.jiao,tantonghao@mail.nwpu.edu.cn;kai.niu,peng.wang@nwpu.edu.cn

ABSTRACT

Text-based person search is a task to retrieve the corresponding person in a large-scale image database given a textual description, which has important value in various fields like video surveillance. In the inferring phase, language descriptions, serving as queries, guide to search the corresponding person images. Most existing methods apply cross-modal signals to guide feature refinement. However, they employ visual features from the gallery to refine textual features, which may cause high similarity between unmatched pairs. Besides, the similarity-based cross-modal attention could disturb the choice of interested areas for descriptions. In this paper, we analyze the deficiency of previous methods and carefully design a Text-guided Visual Feature Refinement network (TVFR), which utilizes text as reference to refine visual representations. Firstly, we divide each visual feature into several horizontal stripes for fine-grained refinement. After that, we employ a text-based filter generation module to generate description-customized filters, which are used to indicate the corresponding stripes mentioned in the textual input. Thereafter, we employ a text-guided visual feature refinement module to fuse part-level visual features adaptively for each description. In experiments, we validate our TVFR through extensive experiments on CUHK-PEDES, which is the only available dataset for text-based person search. To the best of our knowledge, the TVFR outperforms other state-of-the-art methods.

CCS CONCEPTS

• **Information systems** → **Image search**.

KEYWORDS

Text-Based Person Search, Cross-Modal Retrieval, Text-Guided Visual Feature Refinement

ACM Reference Format:
Liying Gao, Kai Niu*, Zehong Ma, Bingliang Jiao, Tonghao Tan, Peng Wang*. 2021. Text-Guided Visual Feature Refinement for Text-Based Person Search. In *Proceedings of the 2021 International Conference on Multimedia Retrieval (ICMR '21), August 21–24, 2021, Taipei, Taiwan.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3460426.3463652

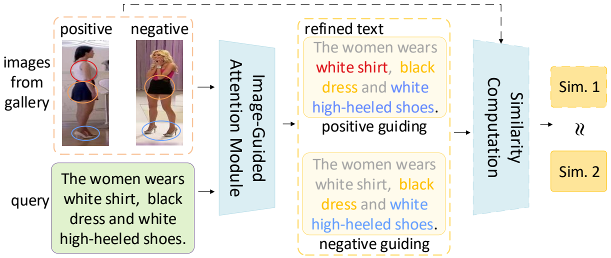
positive
negative
images from gallery
query
The women wears white shirt, black dress and white high-heeled shoes.
Image-Guided Attention Module
refined text
The women wears white shirt, black dress and white high-heeled shoes.
positive guiding
The women wears white shirt, black dress and white high-heeled shoes.
negative guiding
Similarity Computation
Sim. 1
≈
Sim. 2

Figure 1: The Sim.1 denotes the similarity of positive image with its guiding refined text, and Sim.2 denotes the similarity of the negative pair. The refined texts represent the focus of text under image-guided attention. The figure shows that the image-guided attention module leads the text to focus on the shared elements with the comparing image from gallery, and causes high similarity between the negative image and text query, which thus may result in a failure retrieval.

1 INTRODUCTION

Recently, person search has attracted growing attention due to its potential application in fields like intelligent surveillance. Person search is to retrieve the corresponding person in an image database given a query, such as a person image or a sentence. According to the type of queries, person search can be classified into two types: image-based person search [4, 6, 14, 16, 18, 21, 24, 29, 31, 33] and text-based person search [2, 12, 13, 32]. Among them, image-based person search requires at least an image of this specific person as the query, which is not always available in practical applications. For example, in the field of criminal investigation, sometimes the images of suspects are not given, while the descriptions of witness are available. The text-based person search proposed in [13] is to retrieve the corresponding person with provided pedestrian descriptions. Considering that pedestrian descriptions are more available in practical applications, we focus on text-based person search in this work.

The most challenging factor of current task is the huge domain gap between features from different modalities. Most existing methods, such as [9, 19], are devoted to eliminating the variation of visual and textual inputs by applying cross-modal signals to guide feature refinement. These methods have made a certain improvement, while we argue the feature refinement modules employed in

*Corresponding authors.

ICMR '21, August 21–24, 2021, Taipei, Taiwan
© 2021 Association for Computing Machinery.
ACM ISBN 978-1-4503-8463-6/21/08…$15.00
https://doi.org/10.1145/3460426.3463652

previous methods are not quite reasonable for this task. As shown in Figure 1, the widely used image-guided cross-modal attention mechanism may lead the final textual features to focus on the shared elements with the comparing image, *i.e.*, "black dress" and "white high-heeled shoes". However, the different elements between image and text, such as the "white shirt" in Figure 1, usually are ignored by the attention mechanism, while may truly contribute to distinguishing similar yet unmatched image-text pairs. It thus results in a high similarity between unmatched image-text pairs, which may deteriorate the retrieval performance.

We argue that there are two reasons for the problem above, namely, the unsuitable guiding direction and similarity-based cross-modal attention module. We deem that the current task is to retrieve the corresponding pedestrian given a description, while the image-guided cross-modal attention mechanism aims to search the related textual elements according to salient parts of a person image. However, these salient visual parts may not be mentioned in corresponding pedestrian descriptions, which may lead inaccurate matching results. Therefore, in this paper, we are devoted to exploiting an effective feature interaction model, which utilizes textual features to guide the refinement of visual features.

Apart from the guiding direction, the similarity-based cross-modal attention mechanism employed in previous algorithms is also unsuitable for the current task. This mechanism may lead the feature interaction model to focus on shared elements between query text and the comparing image and ignore the other discriminative cues, as exhibited in Figure 1. Intuitively, we could easily infer the visual areas ought to be focused on only according to the given pedestrian description. For instance, we will naturally pay attention to the foot parts of human body when we are searching "a woman wearing white shoes". We thus argue that it is reasonable to capture salient visual areas only depending on textual features without referring to the comparing visual instances, which may introduce extra noises.

To solve the problems above, we construct an effective Text-guided Visual Feature Refinement network (TVFR), in which a Text-Based Filter Generation Module (TBFGM) and a Text-Guided Visual Feature Refinement Module (TVFRM) have been proposed to realize the single-directional guided feature refinement from textual features to visual features. The TBFGM is designed to generate customized filter for each description. The generated filter is able to capture description-specific interested areas (like "upper body"), which will not be disturbed when the visual instances to be compared changed. To flexibly capture the interested visual areas via the generated filter, we apply the horizontal partition method as in [25] to divide the visual feature into several stripes. Thereafter, we carefully design the TVFRM, which enables visual features to focus on the description-mentioned areas by adopting the generated filter. Then, we could achieve promising retrieval performance by measuring the similarity between the query feature and the visual features which emphasize corresponding stripes indicated in the description. Our proposed method is evaluated on the only available dataset for the task of text-based person search, CUHK-PEDES [13]. Experiments show that our TVFR outperforms other state-of-the-art methods.

The main contributions of our paper can be summarised as follows:

- We propose a Text-guided Visual Feature Refinement network (TVFR) to adaptively extract visual features under the guidance of textual features, which regards textual features as the core of the current task.
- In TVFR, we design a text-based filter generation module to dynamically produce customized filter for each language description, which could effectively assist the network to capture description-sensitive visual areas.
- We carefully design extensive experiments and ablation studies to substantiate the superiority of our TVFR. The proposed module can achieve significant improvements on CUHK-PEDES [13].

The rest of the paper is organized as follows: We briefly review related work of our paper in Section 2. In Section 3 the proposed TVFR is elaborated. The experimental results are reported and analysed in Section 4. And we conclude the paper in Section 5.

2 RELATED WORK

In this section, we briefly review several relevant researches in the aspects of text-guided attention mechanism in other vision-language tasks, part-based person re-identification and text-based person search. To clarify, we start with the description of text-guided attention mechanism in other vision-language tasks.

2.1 Text-Guided Attention Mechanism in Other Vision-Language Tasks

In vision-language tasks, quite a few algorithms explored the interaction between cross-modal features, especially text-guided attention for visual features, to promote the capability of designed frameworks. For instance, in Visual Question Answering (VQA) task, Shi *et al.* [23] proposed a question type guided attention module, which utilizes the information of question type to adaptively extract task-specific visual features. Besides, Yang *et al.* [28] and Xu *et al.* [8] applied question-guided spatial attention mechanism specifically for VQA task. As for the task of referring expression comprehension, Wang *et al.* [27] proposed a language-guided attention mechanism in order to learn a discriminative object feature that can adapt to the expression. Yu *et al.* [15] decomposed expressions into three modular components related to different aspects and employed language-based attention to learn the module weights, as well as word/phrase attention that captures the words focused on by each module. Liu *et al.* [17] employed a similar attention mechanism for cross-modal attention-guided erasing.

For the task of text-based person search, we deem that language descriptions play the crucial role, and thus we construct a text-based filter generation module which produces a customized filter for each textual input. Thereafter, we exploit a visual feature refinement module to generate description-specific visual features by employing the filters.

2.2 Part-Based Person Re-identification

The part-based alignment method has been proposed and regarded as one of the mainstream solutions for Person Re-identification (PRID).For example, Sun *et al.* [25] proposed a Part-based Convolutional Baseline (PCB) which partitions the holistic-level feature into several horizontal stripes. After that, local features corresponding to specific human parts are extracted in each horizontal stripe and they are employed for loss computation individually. In addition, Chen *et al.* [3] proposed a Salience-guided Cascaded Suppression Network (SCSN) model to mine diverse salient features belonging to different pre-defined human parts by suppressing the salient areas of the previous network layers. Besides, the MGN proposed by Wang *et al.*[26] uniformly partitions the images into several stripes, and varies the number of parts in different local branches to obtain local feature representations with multiple granularities.

Inspired by the existing methods, in this paper, we construct a visual feature partition module which extracts discriminative part-level visual features by performing the fine-grained horizontal partition. Thereafter, the visual feature partition module is assembled with the designed text-based filter generation module to adaptively generate description-specific visual features.

2.3 Text-Based Person Search

Due to its application in fields like intelligent surveillance and tracking, text-based person search has attracted explosive attention in recent years. The main challenge of this task is to extract discriminative feature representations of descriptions and images, as well as to design objective functions for the features from different modalities.

To solve the first problem, some algorithms struggled to explore a robust and reasonable feature extractor to obtain expressive feature representations belonging to each modality, so as to promote the model capability and performance. For example, Dual-Path [32] constructed an end-to-end dual-path convolutional network to learn the representations of image and text pairs.

For the second challenge, some previous methods researched metric learning of two modalities and performed effective matching loss on the joint space. Since directly deploying the ranking loss is hard for cross-modal features in network learning, Dual-Path [32] adopted instance loss to learn more discriminative feature representations and offers better weight initialization for the ranking loss. Zhang *et al.* [30] proposed a cross-modal projection matching (CMPM) loss and a cross-modal projection classification (CMPC) loss for computing the similarity of image-text pair data and learning more discriminative image-text embedding representations.

The algorithms mentioned above extract visual and textual features independently, which neglect to utilize the semantic connections between pedestrian images and textual descriptions in the feature extraction phase. In order to extract visual features and textual features with interaction, quite a few algorithms are devoted to making use of the inner-relevance between them. For instance, Li *et al.* [12] applied a latent co-attention mechanism concluding a special attention module and a latent semantic attention module to refine the matching results in the stage-2 training. Niu *et al.* [19] proposed a Multi-granularity Image-text Alignment (MIA) module to enhance the accuracy of identification by multi-grained feature alignment between visual representations and textual representations. In all three different granularity modules of MIA, textual feature and visual feature mutually and equally provide complementary information.

These existing algorithms extract visual features and textual features interdependently. However, in the bidirectional attention mechanisms, visual features play the guiding role in image-guided attention module, which cause textual query features to be adjusted by visual features of the gallery, which is not reasonable for this task. Considering the problem, we argue that it is the language descriptions that should play the crucial role in the task of text-based person search, instead of being guided by visual features. Therefore, we innovatively propose the TVFR, which extracts description-specific visual features dynamically under the guidance of filters generated by textual features.

3 APPROACH

In this section, we illustrate our proposed framework. First, we give an overall explanation of it. Following that, we elaborate on the details of the Text-Based Filter Generation Module and Text-Guided Visual Feature Refinement Module in order, which are the main innovations in our paper. Finally, we discuss the difference between similarity-based attention mechanism and our TBFGM.

3.1 Overall Framework

In this subsection, we briefly illustrate the Text-guided Visual Feature Refinement network (TVFR). As shown in Figure 2, our framework consists of three major components which are responsible for feature extraction, feature interaction and cross-modal features matching, respectively.

Feature Extraction Component. Firstly, we would like to introduce the feature extraction component which is constructed with a visual representation extraction branch and a textual representation extraction branch. The textual branch, *i.e.*, the upper branch in Figure 2, is designed to process textual descriptions and generate global textual features. For each textual input, in the first step, each word is represented with its word index in the pre-defined vocabulary, and then is embedded into a 512-dimension feature vector by look-up embedding. In order to refine the textual features, we employ a bidirectional LSTM (Bi-LSTM) [22], which is able to capture the embedding feature of each word in a sentence combined with its inner-relevance and integrate the content of the textual description. In this work, the feature dimensions of hidden states and output vectors are both set to 512. After that, the hidden states of forward and backward directions in Bi-LSTM are concatenated into a 1024-d text representation. In the sequel, the global textual features are extracted by calculating the mean of the hidden states of all the words in the textual description, so as to capture the holistic-level content of textual inputs. The visual branch, *i.e.*, the lower branch in Figure 2, is responsible for extracting visual features, which adopts MobileNet [7] pre-trained on ImageNet [5] as the backbone module. And then, we perform the fine-grained horizontal partition on the final convolutional visual feature V_{base} to extract part-level ones $(V^1_{part}, V^2_{part}, ..., V^k_{part})$. The size of these part-level features is designed as $(C, H/k, W)$, where C, k and (H, W) represent the channel dimension, the number of horizontal stripes and the feature map size respectively. In this paper, C, H, W and k are set to 1024, 12, 4 and 6.

Feature Interaction Component. One of the most challenging difficulties of text-based person search is cross-modal feature matching. Obviously, it is hard to tackle this issue by extracting features belonging to different modalities independently without

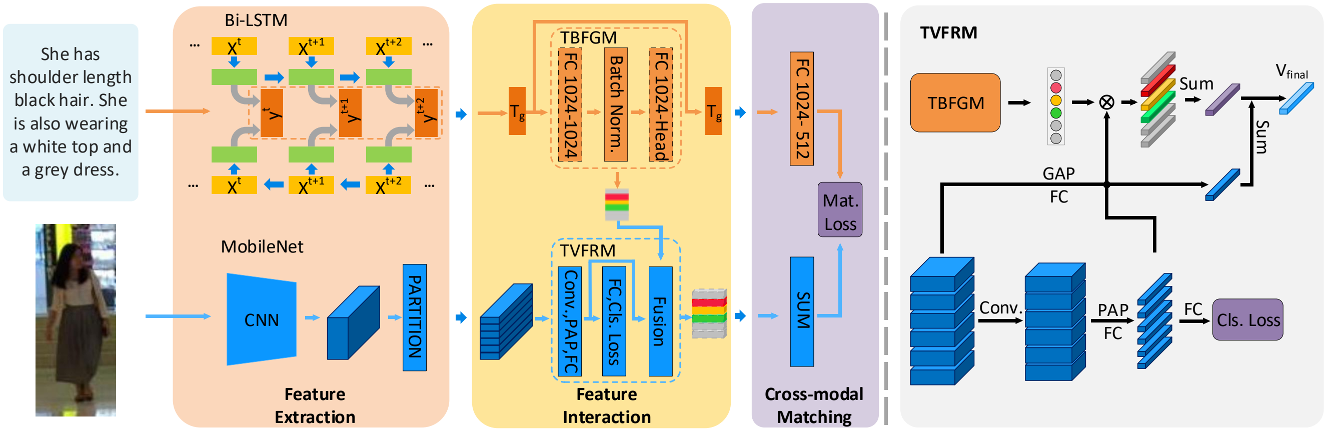

Figure 2: Overview of the proposed TVFR framework. a) Feature Extraction Component: Bi-LSTM and MobileNet are used for textual and visual feature extraction. b) Feature Interaction Component: The text-specific filter is generated by the Text-Based Filter Generation Module (TBFGM) to measure the salience of different human parts. In Text-Guided Visual Feature Refinement Module (TVFRM), part-level visual features are weighted-summed with the text-guided filter as weight. And the details of TVFRM are shown in the right part. PAP denotes part-level average pooling. c) Cross-modal Matching Component: The features from two modalities are mapped into a joint embedding space for cross-modal matching loss computation.

information transaction. To solve this problem, we design a feature interaction component. To be specific, the component is composed of a text-based filter generation module and a text-guided visual feature refinement module, which regard textual features as the controller signal to adaptively extract description-specific visual features. The details about these two modules are introduced in the following subsections.

Cross-modal Matching Component. In order to promote the capacity of our TVFR and ensure convergence, we adopt the cross-modal matching loss in the last part of our TVFR. Since the features to be matched are extracted from different modalities, we firstly map them into a joint embedding space. In practice, we map the global textual feature T_g into the joint embedding space by a mapping function, a fully connected layer here, to produce a 512-dimension textual feature vector. For visual features, we also map the part-level features separately to the joint embedding space by a series of fully connected layers, which do not share parameters. In addition, the global visual feature generated from the backbone network has also been employed as a supplement for part-level visual features. After that, the refined part-level visual features after the feature interaction module and the global visual feature are summed to produce the final visual feature, *i.e.*, V_{final} in Figure 2. After obtaining these embedded cross-modal features, we employ them for similarity and loss computation. In the training phase, We utilize the cross-modal projection matching (CMPM) loss function and cross-modal projection classification (CMPC) loss function proposed in [30] to maximize the similarity of features from matched image-text pairs and the variance between unmatched pairs, namely, the Mat. Loss in Figure 2. Besides, in order to enhance the part-level visual features, we employ part-level identity classification loss on visual stripes as in [25]. The overall loss function is calculated by

$$\mathcal{L} = \mathcal{L}_{cmpm} + \mathcal{L}_{cmpc} + \lambda \cdot \mathcal{L}_{part_{CLS}} \tag{1}$$

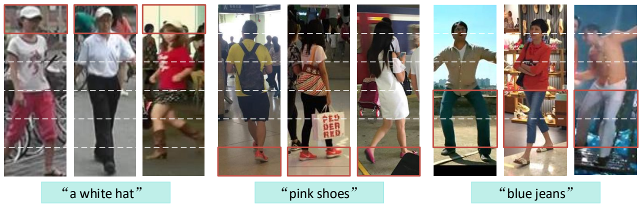

Figure 3: The figure exhibits several groups of pedestrian instances with shared visual attributes, *i.e.*, a white hat, pink shoes and blue jeans. It can be found that the same visual attribute belonging to different pedestrian images is usually located in the fixed stripe.

where λ is a hyper-parameter that controls the weight of part-level identity classification loss, *i.e.*, the Cls. Loss in Figure 2. In the test phase, cosine distance is employed as similarity measurement to construct the similarity ranking list.

3.2 Text-Based Filter Generation Module

In this subsection, we would like to elaborate on the details of the major part in the feature interaction component of TVFR, which is termed as Text-Based Filter Generation Module (TBFGM). As we mentioned in Section 2.3, most existing relevant algorithms neglect to follow the task setting and design inappropriate feature interaction modules. Instead, we argue that language descriptions play the crucial role in the task of text-based person search, and it is reasonable to utilize textual features to guide the visual feature extraction.

To supply this gap, we are going to exploit a module which could adaptively adjust visual features under the guidance of each textual input. Generally, the description on a specific pedestrian always focuses on several salient objects, such as "a white hat", "pink shoes"

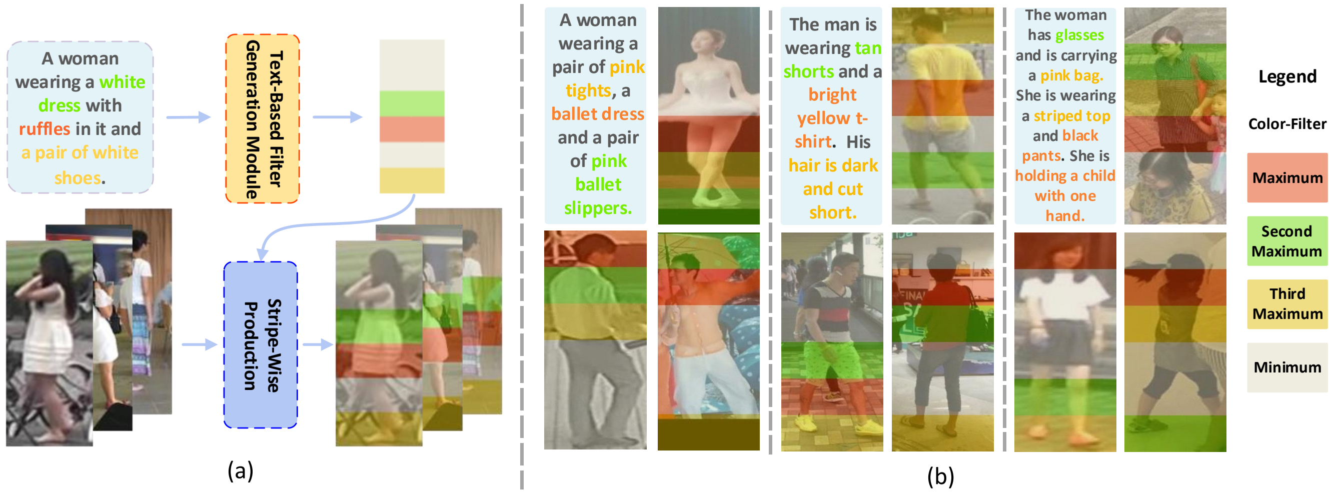
A woman wearing a white dress with ruffles in it and a pair of white shoes.
Text-Based Filter Generation Module
Stripe-Wise Production
(a)
A woman wearing a pair of pink tights, a ballet dress and a pair of pink ballet slippers.
The man is wearing tan shorts and a bright yellow t-shirt. His hair is dark and cut short.
The woman has glasses and is carrying a pink bag. She is wearing a striped top and black pants. She is holding a child with one hand.
Legend
Color-Filter
Maximum
Second Maximum
Third Maximum
Minimum
(b)

Figure 4: (a): The illustration of Text-Based Filter Generation Module (TBFGM). Given a textual description, the TBFGM generates customized filter, which could capture the stripes of visual objects mentioned in the description. (b): Several visualization examples of text-guided customized filters (the first row) and similarity-based attention filters (the second row). The filters generated by TBFGM are decided only by textual description, while the similarity-based attention filters are also influenced by image contents, which sometimes misunderstand the semantic information and mislead the interested areas. As shown in the first example, the similarity-based attention filters focus on the pink shirt and pink umbrella.

and "blue jeans". By observing and analyzing the dataset, we come to the point of view that the same visual attributes belonging to different pedestrian images are always located in the fixed stripes, as shown in Figure 3. Therefore, it is reasonable to infer that we could capture the possible locations of visual attributes mentioned in descriptions by analyzing textual inputs.

Under this observation, we propose the TBFGM, which generates customized filters by analyzing textual inputs, so as to capture salient visual areas. The illustration of TBFGM is shown in Figure 4 (a), and it can be found that the TBFGM adaptively generates a customized filter for the textual input to capture the positions of visual objects, which is not influenced by the contents of images.

In practice, we simply employ a multilayer perceptron (MLP) module to generate a customized filter for each textual description in our algorithms, as illustrated in the TBFGM of Figure 2. To be more specific, the designed TBFGM can be formulated as follows,

$$\pi(x) = \delta(\omega_2(\sigma(\mathcal{N}(\omega_1 x + \beta_1)) + \beta_2) \qquad (2)$$

where $\pi(\cdot)$ denotes the MLP module, $\sigma(\cdot)$ and $\delta(\cdot)$ represent the ReLU and Sigmoid activation function respectively, $\mathcal{N}$ indicates batch normalization, β and ω are trainable parameters of fully connected layers in the three-step transformation function mapping $\mathbb{R}^{C_o} \mapsto \mathbb{R}^{C_o} \mapsto \mathbb{R}^{Head}$ with C_o and $Head$ as the input channel dimension and the number of horizontal stripes. In this paper, we set C_o and $Head$ to 1024 and 6 respectively.

In the first row of Figure 4 (b), we give visualization of the generated text-guided feature filters. It can be found that filters generated by the TBFGM accurately capture the positions of visual objects mentioned in descriptions. The comparison with similarity-based attention mechanism will be discussed in Section 3.4.

3.3 Text-Guided Visual Feature Refinement Module

As we mentioned above, text-guided filters generated by the TBFGM are capable of capturing salient areas in visual features. It is therefore reasonable to exploit an adaptive visual feature fusion algorithm which aims to extract description-specific visual features by employing customized filters. We thereby propose the Text-Guided Visual Feature Refinement Module (TVFRM), which correlates the text-guided filters with visual features. Following the existing works [19, 25], we apply the horizontal partition on the convolutional visual features to extract part-level features, consistent with the numbers of generated filter. Then, simply, we apply a stripe-wise production, namely, weighted sum on part-level visual representations with the filter as weight. We argue that this architecture is a simple and powerful one, with increased capacity thanks to its description-adaptive scheme, yet without an excessive increase in the number of model parameters or the amount of computation.

In more detail, as shown in the right part in Figure 2, the TVFRM is composed of two branches in terms of the visual feature extraction, *i.e.*, the part-level branch and the global-level branch. For part-level visual feature extraction, we insert a series of parameter-unshared 1×1 convolution layers after each partitioned feature in order to individually refine features belonging to specific stripes. Then we extract features of each part by applying a simple part-level average pooling (PAP). After that, a series of fully connected layers are employed as the mapping functions after pooled part-level visual features to embed the visual features into the joint embedding space and compress them with the same channel dimension as textual features, 512. Noting that, all these fully connected layers do not sharing parameters, so as to further increase the diversity among these part-level visual features. Besides, to make them more expressive and discriminative, we insert an identity classification loss on each part-level visual feature, as in PCB [25]. After that, they are refined by utilizing the text-guided filters generated by TBFGM to obtain the aggregated part-level visual features.

Simultaneously, for the global-level visual feature extraction, the global visual convolutional feature after the CNN backbone is compressed to a 512-dimension vector by global average pooling (GAP) and a fully connected layer, in order to be embeded into the joint embedding space and match the dimension of textual features. Then we add the global visual feature to the refined part-level visual features to provide complementary information from a holistic-level view. The final visual feature is calculated by

$$V_{final} = V_{global} + \alpha^i \cdot V_{part}^i \tag{3}$$

where V_{global} denotes the global visual representation, V_{part}^i represents the part-level visual feature of the i-th horizontal stripe and α^i is the text-guided weight score for the i-th visual stripe.

3.4 Difference between Similarity-based Attention Mechanism and Our TBFGM

As mentioned earlier, most existing methods [12, 19, 20] treat visual features and textual features equally, while we argue that textual features play the crucial role in the text-based person search task. To be specific, we adopt the partition algorithm on the visual feature extraction module as most existing algorithms [3, 25, 26] to extract expressive visual features, and employ TBFGM (introduced in Section 3.2) to produce a customized filter for each textual input. Thereafter, we aggregate partitioned visual features together by the generated filters. Meanwhile, the existing similarity-based attention module in [2] is also appropriate for guiding the part-level visual feature aggregation. In more detail, they adopt the similar partition algorithm as that in our framework and generate weight scores to aggregate partitioned visual features by measuring the similarity between textual features with part-level visual features.

By comparison, we deem that the major divergence between our proposed TBFGM and the similarity-based attention mechanism is the way to generate filters. That is, our TBFGM employs textual features to single-directionally guide visual features by generating filters which are not influenced by visual features, while the other one measures the similarity between textual features and visual features to capture the salient human parts. Although the similarity-based attention mechanism has been widely used, we argue that this strategy is not definitely fit for text-guided person search. Our key interpretation is that textual features ought to dominate the retrieval process and decide the interested image areas in the task whose query instances are all from the textual modality, while the weight of image parts produced by the similarity-based attention module could be disturbed by salient yet irrelevant visual features. As shown in the second row of Figure 4 (b), compared with our TBFGM, the attention-based algorithm leads to several inaccurate interested areas due to the similarity between non-corresponding human-part features and local textual features in the description, which will deteriorate the performance of the retrieval module. Consequently, it may not be the optimal solution for visual feature refinement in the text-based person search task.

Table 1: Comparison with the state-of-the-art methods on the CUHK-PEDES [13] dataset. R@1, R@5 and R@10 accuracy are reported. The best results are bold.

Methods	R@1	R@5	R@10	Total
Prior Works				
GNA-RNN(2017) [13]	19.05	-	53.64	-
PWM-ATH(2018) [2]	27.14	49.45	61.02	137.61
IATVM(2017) [12]	25.94	-	60.48	-
GLA(2018) [1]	43.58	66.93	76.26	186.77
Dual-Path(2017) [32]	44.40	66.26	75.07	185.73
MIA(2020) [19]	53.10	75.00	82.90	211.00
Baseline				
CMPM+CMPC (2018) [30]	49.37	-	79.27	-
Proposed				
TVFR(ours)	**53.87**	**75.25**	**83.47**	**212.59**

4 EXPERIMENTS

In this section, we first introduce the experimental setup including dataset, evaluation metrics and implementation details. Then we compare our TVFR with several other state-of-the-art algorithms to show the superiority of our framework. Thereafter, we give the ablation studies in the aspects of filter generation algorithm, comparison with the similarity-based attention mechanism, hyper-parameter settings, and so on. Finally, we visualize and analyze several examples, including successful searches and failure cases.

4.1 Dataset and Evaluation Metrics

We use the CUHK-PEDES [13] dataset to verify our method, which is the only large-scale dataset for the task of text-based person search. This dataset contains 40, 206 pedestrian images of 13, 003 identities, with each image described by two textual descriptions on average. The dataset is split into 11, 003 training identities with 68, 126 image-text pairs, 1000 validation persons with 6, 158 image-text pairs and 1000 test individuals with 6, 156 image-text pairs. We follow the train-test split in [13]. To make a good comparison with the previous methods and analyze for ablation experiments, we adopt Recall@K (K=1, 5, 10) [10] for retrieval evaluation.

4.2 Implementation Details

All the models are implemented in TensorFlow with a NVIDIA GEFORCE GTX 2080 Ti GPU. In the visual branch, we employ MobileNet [7] pre-trained on ImageNet [5] as backbone. All the images are resized to (384, 128). For each input image, we normalize it to limit its values into the range of $[0, 1]$, and transform them to $[-1, 1]$ by simple computation before sending them to the backbone module. In the training phase, we horizontally flip each image with the probability of 0.5 for data augmentation. In the textual branch, we employ Bi-LSTM to extract textual features, where the feature size of the hidden states is set to 512. And the mini-batch size is set to 16. We adopt Adam optimizer[11] whose learning rate is 2×10^{-4}. In our experiments, all models are trained for 60 epochs. The dimension of visual features and textual features in the joint embedding space is set to 512. During the test phase, we employ global textual features and fused visual features for distance computation by utilizing cosine distance for similarity measuring.

4.3 Comparison with the State-of-the-art Methods

Table 1 shows the comparison results of our TVFR and other state-of-the-art methods on the CUHK-PEDES dataset. Preliminarily, it can be seen that our TVFR outperforms all the other approaches regarding R@1, R@5 and R@10 accuracy. Noting that, our baseline is competitive with some of the previous methods, which is partly because the loss function is carefully designed for this task, although the architecture is a simple two-branch model and the visual backbone is pre-trained MobileNet instead of deeper networks like ResNet-50.

The best competitor, MIA proposed by Niu *et al.*[19], is a multi-granularity image-text alignment method, consisting of three different granularities, *i.e.*, global-global alignment, global-local alignment and local-local alignment. Compared with MIA, which employs three different alignment modules to achieve multi-grained matching, our TVFR achieves a slightly better performance by capturing the essential relationships between visual and textual features in the current task, where textual features play the dominant role in the test phase.

Dual-Path [32] employs instance loss in addition to the cross-modal matching loss, and CMPM+CMPC [30] proposes effective cross-modal projection loss functions. However, both of these two methods extract visual features and textual features independently while neglecting to exploit the interaction between them. PWM-ATH [2] and GNA-RNN [13] employ fine-grained patch-word matching and word-image matching respectively, but lack the global alignment filtering the uninvolved information.

4.4 Ablation Experiments

During the experiments, we conduct ablation studies from different aspects on CUHK-PEDES dataset to analyze the rationality and effectiveness of the proposed components in TVFR. The results are shown in Table 2 and Table 3.

Baseline Module. In this paper, we follow the existing work [30] to construct a simple yet effective baseline module. The experimental results of the baseline module are shown in the first row of Table 2. Following this, we demonstrate the effectiveness of the components in our framework by gradually adding them to the naive baseline module and comparing the performance improvements.

Effects of Partition Algorithm in Visual Branch. Following the previous work [25] in image-based person search, we apply the horizontal partition on the convolutional visual features to extract part-level features. Thereafter, we fuse them together by calculating the mean of part-level visual features or concatenating them together and then add the global visual feature to them as we mentioned in Section 3.1. Then the fused visual features are employed for similarity computation with global textual features. In Table 2, Avg. in row 2 refers to the average of part-level visual features, and Cat. represents concatenating the part-level visual features. As can be found, part-based methods (row 2 and 3) can bring a significant improvement of about 3 points of R@1 and more than 2 points of R@5 and R@10, compared with the baseline method. It can be concluded that part-level visual features can supply useful complementary information to the global visual features.

Table 2: Ablation study.

	Method	R@1	R@5	R@10	Total
1	baseline	48.86	71.32	79.97	200.15
2	Avg.	51.58	73.93	82.07	207.58
3	Cat.	52.05	73.90	82.24	208.18
4	Rand.(uniform)	51.73	74.44	82.52	208.67
5	Rand.(norm)	52.32	74.45	82.65	209.42
6	w/o V_{global}	52.08	73.79	82.42	208.29
7	Attn. (cosine)	49.95	71.88	80.33	202.16
8	TVFR (ours)	**53.87**	**75.25**	**83.47**	**212.59**

Effects of Proposed TBFGM. To verify the effectiveness of the proposed TBFGM, we design a series of comparison experiments, as shown in row 2, 4 and 5 of Table 2. The Avg. in row 2 can be regarded as a method of applying static filters, in which all the elements are $1/k$, where k denotes the number of horizontal stripes. In addition to the static filters, we also compare our TBFGM with the dynamically random-sampled filter generation algorithms. To be specific, we randomly sample filters from a uniform distribution (row 4) in range of $[0, 1)$, and a truncated normal distribution (row 5) from 0 to 1, whose mean is 0.5, and standard deviation is 0.25. By using TBFGM (row 8), which generates a customized filter for each textual input, our TVFR outperforms the other three compared filter generation methods by 1.84 points in R@1 on average. Therefore, we argue that the filters generated by TBFGM can effectively capture the interested human-parts for each description and contribute to text-based person search.

Add Global Visual Feature to Part-level Visual Features. As in the aforementioned Section 3.3, we add the global visual feature to part-level visual features, so as to supply the final visual feature with the holistic-level information. In order to verify the additional global feature is effective for pedestrian identifying, we compare the experimental results of our TVFR with the module which does not employ the additional global feature. As shown in the row 6 and 8 of Table 2, the additional global visual feature could bring 1.81% R@1 improvements, which verifies the effectiveness of the global visual feature to improve the performance.

Comparison with Similarity-based Attention Mechanism. In the Section 3.4, we have illustrated the reason why we do not employ the similarity-based attention algorithm to guide the aggregation of partitioned visual features in detail. Here, we give the experiential comparison of TBFGM with the similarity-based attention module. The attention module employed in the compared version receives the global textual feature and partitioned visual features as input, embeds them into a joint space, and calculates the cosine similarity between the textual feature and part-level visual features. Thereafter, the part-level visual features are fused by weighed-sum with the similarity scores as weight. As exhibited in row 7 and 8 of Table 2, our TVFR outperforms the compared version with the similarity-based attention mechanism by more than 3 points in R@1, which proves the suitability of our TBFGM for the text-based person search task.

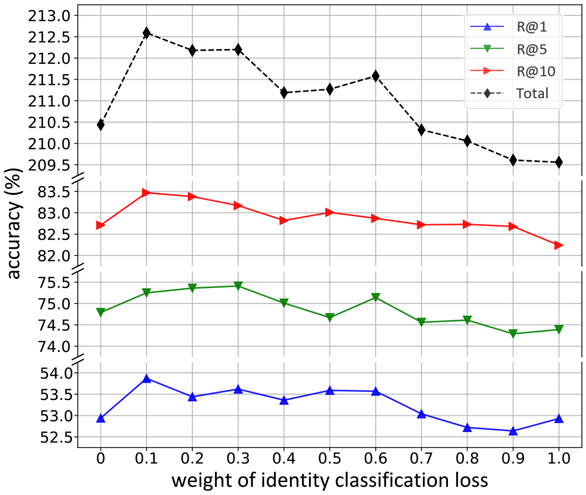

Figure 5: The line chart of results of different weights on identity classification loss for part-level visual features.

Table 3: Ablation study: the number of horizontal stripes.

Parts	R@1	R@5	R@10	Total
1	48.86	71.32	79.97	200.15
2	52.25	74.70	82.54	209.47
4	52.52	74.84	82.75	210.11
6	**52.94**	74.79	**82.71**	**210.44**
12	51.80	73.88	82.04	207.72

The Number of Horizontal Stripes. Table 3 shows the impact of the number of horizontal stripes on the retrieval results. Intuitively, it is equal to directly adopting the global feature if we set k to 1. By increasing the number of stripes, finer-grained visual features are captured and the performance is improved. However, the overall performance degrades if we blindly increase the number of horizontal stripes, as exhibited in row 5. A reasonable interpretation could be that if we employ the dense partition strategy, the stripes can hardly capture a rounded human-part. As a result, we achieve the best performance when we set k to 6.

Weight of Part-level Identity Classification Loss. In order to find the balance point between the global cross-modal matching losses, namely, CMPM and CMPC losses, and the local part-level identity classification loss, we adjust λ in Equation 3.1 from 0.0 to 1.0. The experimental results are exhibited in Figure 5, in which we can find that the performance of our module is promoted a lot when λ increases from 0 to 0.1. In a holistic view, the performance gradually decreases when λ increases from 0.1 to 1.0, which suggests that our module could achieve best performance when we set λ to 0.1.

4.5 Cases Analysis

We exhibit three top-10 retrieval results of our proposed TVFR in Figure 6. From the left colorful stripes, it can be found that the filters generated by TBFGM can capture the interested visual areas mentioned in descriptions. In the first two lines, our TVFR achieves accurate prediction results for the reason that the filters generated by TBFGM assist to focus on interested image areas and exclude mismatched person images. However, in the last line, the TVFR fails to recognize the corresponding pedestrian even though the TBFGM produces reasonable filter for the text, which we deem is caused by the inaccurate description, like "a multicolored design", and inadequate recognition ability of the framework.

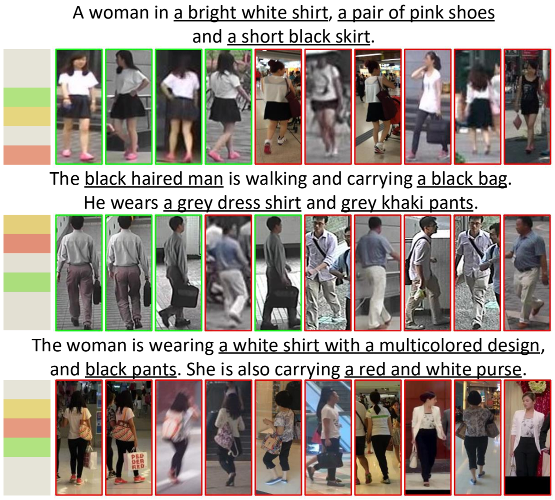
A woman in a bright white shirt, a pair of pink shoes
and a short black skirt.
The black haired man is walking and carrying a black bag.
He wears a grey dress shirt and grey khaki pants.
The woman is wearing a white shirt with a multicolored design,
and black pants. She is also carrying a red and white purse.

Figure 6: Examples of text-based person search by the proposed model on CUHK-PEDES [13]. Images of the corresponding pedestrian are marked with green rectangles.

5 CONCLUSION

In this paper, we analyze the problems of the similarity-based image-guided feature refinement modules in the previous works and propose to utilize the textual feature to generate filters and guide the visual feature refinement. Based on this, we introduce TVFR, a novel text-guided visual feature refinement framework that has two carefully designed sub-networks, namely, Text-Based Filter Generation Module (TBFGM) and Text-Guided Visual Feature Refinement Module (TVFRM). The TBFGM is constructed to generate a customized filter for each input description and the TVFRM is able to capture the interested human parts with the assistance of the generated filter and further adaptively refine part-level visual features for each textual input. The capability of our model is sightly promoted by employing textual features and their custom-made visual features for similarity computation. Our TVFR achieves state-of-the-art performance on the CUHK-PEDES dataset. We also demonstrate through extensive ablation studies that our proposed TBFGM and TVFRM are effective for text-based person search.

ACKNOWLEDGMENTS

This work is supported by the National Key R&D Program of China (No. 2020AAA0106900), the National Natural Science Foundation of China (No.U19B2037), and the Open Projects Program of National Laboratory of Pattern Recognition (No. 202100028).

REFERENCES

[1] Dapeng Chen, Hongsheng Li, Xihui Liu, Yantao Shen, Jing Shao, Zejian Yuan, and Xiaogang Wang. 2018. 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)*. 54–70.

[2] Tianlang Chen, Chenliang Xu, and Jiebo Luo. 2018. Improving text-based person search by spatial matching and adaptive threshold. In *2018 IEEE Winter Conference on Applications of Computer Vision*. 1879–1887.

[3] Xuesong Chen, Canmiao Fu, Yong Zhao, Feng Zheng, Jingkuan Song, Rongrong Ji, and Yi Yang. 2020. Salience-guided cascaded suppression network for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3300–3310.

[4] YingCong Chen, WeiShi Zheng, and Jianhuang Lai. 2015. Mirror Representation for Modeling View-Specific Transform in Person Re-Identification. In *Proceedings of the 24th International Conference on Artificial Intelligence*. 3402–3408.

[5] Jia Deng, Wei Dong, Richard Socher, LiJia Li, Kai Li, and Li FeiFei. 2009. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 248–255.

[6] Yixiao Ge, Zhuowan Li, Haiyu Zhao, Guojun Yin, Shuai Yi, Xiaogang Wang, and Hongsheng Li. 2018. FD-GAN: Pose-guided feature distilling GAN for robust person re-identification. *arXiv preprint arXiv:1810.02936* (2018).

[7] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. 2017. MobileNets: efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861* (2017).

[8] Xu Huijuan and Saenko Kate. 2016. Ask, Attend and Answer: Exploring Question-Guided Spatial Attention for Visual Question Answering. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 451–466.

[9] Ya Jing, Chenyang Si, Junbo Wang, Wei Wang, Liang Wang, and Tieniu Tan. 2020. Pose-Guided Multi-Granularity Attention Network for Text-Based Person Search. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 11189–11196.

[10] Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3128–3137.

[11] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[12] Shuang Li, Tong Xiao, Hongsheng Li, Wei Yang, and Xiaogang Wang. 2017. Identity-aware textual-visual matching with latent co-attention. In *Proceedings of the IEEE International Conference on Computer Vision*. 1890–1899.

[13] Shuang Li, Tong Xiao, Hongsheng Li, Bolei Zhou, Dayu Yue, and Xiaogang Wang. 2017. Person search with natural language description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1970–1979.

[14] Zhang Li, Tao Xiang, and Shaogang Gong. 2016. Learning a Discriminative Null Space for Person Re-Identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1239—-1248.

[15] Yu Licheng, Lin Zhe, Shen Xiaohui, Yang Jimei, Lu Xin, Bansal Mohit, and Berg Tamara L. 2018. MAttNet: Modular Attention Network for Referring Expression Comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1307–1315.

[16] Jinxian Liu, Bingbing Ni, Yichao Yan, Peng Zhou, Shuo Cheng, and Jianguo Hu. 2018. Pose transferrable person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4099–4108.

[17] Xihui Liu, Zihao Wang, Jing Shao, Xiaogang Wang, and Hongsheng Li. 2019. Improving Referring Expression Grounding With Cross-Modal Attention-Guided Erasing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1950–1959.

[18] Martinel, Niki, Abir Das, Christian Micheloni, and Amit K. Roy-Chowdhury. 2016. Temporal Model Adaptation for Person Re-Identification. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 858—-877.

[19] Kai Niu, Yan Huang, Wanli Ouyang, and Liang Wang. 2020. Improving description-based person re-identification by multi-granularity image-text alignments. *IEEE Transactions on Image Processing* 29 (2020), 5542–5556.

[20] Nikolaos Sarafianos, Xiang Xu, and Ioannis A Kakadiaris. 2019. Adversarial representation learning for text-to-image matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 5814–5824.

[21] M Saquib Sarfraz, Arne Schumann, Andreas Eberle, and Rainer Stiefelhagen. 2018. A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 420–429.

[22] Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing* 45, 11 (1997), 2673–2681.

[23] Yang Shi, Tommaso Furlanello, Sheng Zha, and Animashree Anandkumar. 2018. Question type guided attention in visual question answering. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 151–166.

[24] Yumin Suh, Jingdong Wang, Siyu Tang, Tao Mei, and Kyoung Mu Lee. 2018. Part-aligned bilinear representations for person re-identification. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 402–419.

[25] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. 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)*. 480–496.

[26] Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi Zhou. 2018. Learning discriminative features with multiple granularities for person re-identification. In *Proceedings of the 26th ACM international conference on Multimedia*. 274–282.

[27] Peng Wang, Qi Wu, Jiewei Cao, Chunhua Shen, Lianli Gao, and Anton van den Hengel. 2019. Neighbourhood watch: Referring expression comprehension via language-guided graph attention networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1960–1968.

[28] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. 2016. Stacked attention networks for image question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 21–29.

[29] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven C.H. Hoi. 2021. Deep Learning for Person Re-identification: A Survey and Outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021), 1–1.

[30] Ying Zhang and Huchuan Lu. 2018. Deep cross-modal projection learning for image-text matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 686–701.

[31] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. 2015. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE International Conference on Computer Vision*. 1116–1124.

[32] Zhedong Zheng, Liang Zheng, Michael Garrett, Yi Yang, and Yi-Dong Shen. 2017. Dual-path convolutional image-text embedding with instance loss. *arXiv preprint arXiv:1711.05535* (2017).

[33] Qin Zhou, Heng Fan, Shibao Zheng, Hang Su, Xinzhe Li, Shuang Wu, and Haibin Ling. 2018. Graph correspondence transfer for person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*.