Cross-Modal Knowledge Adaptation for Language-Based Person Search

Yucheng Chen, Rui Huang, Hong Chang, *Member, IEEE*, Chuanqi Tan, *Member, IEEE*, Tao Xue, and Bingpeng Ma

***Abstract*— In this paper, we present a method named Cross-Modal Knowledge Adaptation (CMKA) for language-based person search. We argue that the image and text information are not equally important in determining a person's identity. In other words, image carries image-specific information such as lighting condition and background, while text contains more modal agnostic information that is more beneficial to cross-modal matching. Based on this consideration, we propose CMKA to adapt the knowledge of image to the knowledge of text. Specially, text-to-image guidance is obtained at different levels: individuals, lists, and classes. By combining these levels of knowledge adaptation, the image-specific information is suppressed, and the common space of image and text is better constructed. We conduct experiments on the CUHK-PEDES dataset. The experimental results show that the proposed CMKA outperforms the state-of-the-art methods.**

***Index Terms*— Language-based person search, cross-modal knowledge adaptation, image-specific information.**

I. Introduction

LANGUAGE-BASED person search aims to retrieve specific person images from a large scale database according to a given textual description. It plays an important role in video surveillance due to its wide application in public security, information retrieval, and large-scale video analysis. Compared with images, textual descriptions of a query person are sometimes more accessible and contain comprehensive information. For example, it is difficult to obtain a photo of a suspect in advance, but there is a description from the victim. Thus, language-based person search has been attracting increasing research attention in recent years.

This man is wearing glasses, a maroon polo shirt, tan shorts, black sneakers and is shouldering a black back pack.

Fig. 1. Imbalanced information between image and text. The image-specific information such as lighting condition and background rarely appears in the text. When we compare the image and the text of a pedestrian, we only consider whether the common information of them is consistent.

The general way to perform language-based person search is to calculate the similarity between images and texts and rank the candidate images according to the similarity. However, the inconsistent representation of different modalities makes it difficult to directly measure the similarity between visual images and textual descriptions. To address the problem, some methods build a similarity learning network [1]–[3] to compute the similarity between image and text. Other methods project the features of different modalities into a shared space [4]–[7] to learn the common representation. One advantage of the common representation learning method is that the distance between features can be calculated directly in the shared space, so there is no need to repeatedly extract the features of the same image or text for each image-text pair as similarity learning methods do. Considering the great potential in learning discriminative features and the higher computational efficiency in the test stage, in this paper, we tackle language-based person search from the perspective of common representation learning.

Most of the common representation learning methods project image features and text features into a shared space in an equal manner. However, the information contained in an image and a text is not equal. Since text provides a description of the person in an image, it summarizes partial image information. In other words, images contain image-specific information that is rarely described by texts, such as lighting condition, image resolution, viewpoint, background, pedestrian subtle gestures, noise, and so on. Therefore, the two modalities actually carry unequal amount of information and the relationship between them is asymmetric.

Unequal amount of information of image and text will result in redundancy of image information and difficulty in aligning features across different modalities. As shown

Manuscript received September 17, 2020; revised February 15, 2021; accepted March 15, 2021. Date of publication March 31, 2021; date of current version April 7, 2021. This work was supported in part by the Natural Science Foundation of China (NSFC) under Grant 61876171 and Grant 61976203 and in part by the Open Project Fund from Shenzhen Institute of Artificial Intelligence and Robotics for Society under Grant AC01202005015 and 2019-INT006. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Guo-Jun Qi. *(Corresponding author: Bingpeng Ma.)*

Yucheng Chen and Hong Chang are with the Key Laboratory of Intelligent Information Processing, Chinese Academy of Sciences (CAS), Beijing 100190, China, also with the Institute of Computing Technology, Chinese Academy of Sciences (CAS), Beijing 100190, China, and also with the School of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: yucheng.chen@vipl.ict.ac.cn; changhong@ict.ac.cn).

Rui Huang is with Shenzhen Institute of Artificial Intelligence and Robotics for Society, The Chinese University of Hong Kong, Shenzhen 518172, China (e-mail: ruihuang@cuhk.edu.cn).

Chuanqi Tan and Tao Xue are with Tencent, Beijing 100193, China (e-mail: jamestan@tencent.com; emmaxue@tencent.com).

Bingpeng Ma is with the School of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: bpma@ucas.ac.cn).

Digital Object Identifier 10.1109/TIP.2021.3068825

in Fig.1, when humans compare the image and text descriptions of a pedestrian, they only consider whether the common information (e.g., glasses, maroon polo shirt) of the image and the text is consistent. The image-specific information such as background and lighting condition does not contribute to cross-modal matching, but instead leads to heterogeneity between different modalities and increases the difficulty of projecting the features of images and texts into a common space. The performance of retrieval will therefore be affected.

Another problem of the image-specific information is that it may be detrimental to the learning of image representation. For example, in Fig.2a, the key difference between these two different pedestrians with similar appearance is the patterns of the clothes. During the training stage, due to different lighting conditions, the model tends to distinguish the two images by lighting condition rather than the details of the clothes, which is more difficult to capture. At the testing stage, ignoring such details will result in the neglect of an important clue for cross-modal matching. Similarly, in Fig.2b, due to the misalignment caused by different viewpoints, it is difficult to attend to the attribute of whether the pedestrian holds a white bag in his right hand. In these two figures, texts which describe persons' characteristics are inherently capable of telling the key and subtle differences between two people. Therefore, text can be used to guide image features to enrich them with important person details while avoiding the interference of image-specific information.

In light of the above observation, in this paper, we propose a method named Cross-Modal Knowledge Adaptation (CMKA) to suppress image-specific information while getting benefit from text knowledge. In the proposed method, we adapt the knowledge of image to the knowledge of text at different levels. For the knowledge of individual data, we propose feature adaptation between the image and the text. For the structural knowledge of the entire data embedding space, we perform list-wise adaptation between modalities. For the knowledge of class relationship, we propose class probability adaptation. By combining these levels of knowledge adaptation, the amount of information from the image modality and the text modality is balanced, so that the features of different modalities can be projected more effectively into one common space. Moreover, the text feature that tells important details drives the image representation network to discover more textual-visual correspondences between the image and the text. Thus, an image feature that can better match the corresponding text is obtained.

To demonstrate the effectiveness of our method, we conduct extensive experiments on the CUHK-PEDES dataset. The results show that the proposed method improves the performance of language-based person search with a significant margin and achieves state-of-the-art performance.

The remainder of this paper is organized as follows: Section II reviews the relevant works. Section III introduces the proposed CMKA method. Extensive experimental results are presented and analyzed in Section IV, and finally the paper is concluded in Section V.

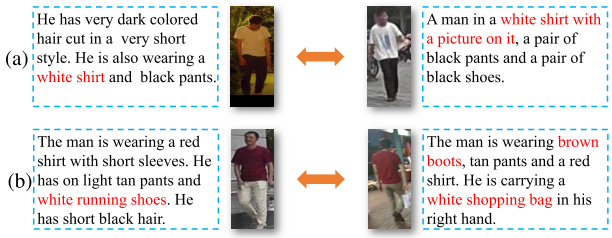


Fig. 2. Image-specific information hinders the model from learning discriminative details of the person. Texts are inherently capable of telling the key and subtle differences between two people.

II. Related Work

A. Language-Based Person Search

Language-based person search is a task of retrieving person images from image database according to a textual description, which is different from image or attribute-based person search [8]–[15] and language-based person segmentation [16]. Li *et al.* [1] collect the CUHK-PEDES dataset with detailed language description annotations and propose the task of language-based person search. The methods of language-based person search can be divided into two categories: similarity learning methods and common representation learning methods.

The similarity learning methods calculate the similarity score of image and text in the network. Li *et al.* [1] propose a recurrent neural network with gated neural attention mechanism (GNA-RNN) to learn affinities between sentences and person images for this task. In the network, the unit activation is used to represent the existence of certain appearance pattern in an image. The unit-level attention is used to relate each word with the units. The word-level gate represents the importance of each word. Li *et al.* [3] propose an identity-aware two-stage framework for the task. The stage-1 network compares the input feature with the dynamic buffer of the other modality. It filters out easy negative samples and provides initialization for the stage-2 network. In the stage-2 network, the spatial attention module is used to weight different image regions for each word. The latent semantic attention module is used to align different sentence structures with decoder LSTM. Chen *et al.* [2] propose an efficient patch-word matching model to accurately capture the local matching details between image and text. In the method, the affinity between an image and a word is calculated as the maximum patch-word affinity. They also propose an adaptive threshold mechanism to limit the affinity scores that exceed the threshold corresponding to the word, thereby reducing the sensitivity of the model to the affinity scores of certain image-word pairs. Chen *et al.* [17] build global and local associations for image-text pairs. In the global discriminative association, the binary cross-entropy loss pulls the positive image-text pairs closer and pushes the negative ones far away. In the local reconstructive association, they use the image feature attended by the text feature to reconstruct the text.

The common representation learning methods aim to project features of different modalities into a common space. Zheng *et al.* [4] view each image / text group as a class

and propose the instance loss to improve the intra-modal discriminability of the model. They build a dual-path CNN to conduct end-to-end training on both image and text branches. Different from the above methods of building similarity learning networks, they project the features of different modalities into a common space and then calculate the similarity. To learn discriminative image-text embeddings, Zhang and Lu [18] propose a cross-modal projection matching loss and a cross-modal projection classification loss. The cross-modal projection matching loss minimizes the KL divergence from the true matching probability to the estimated matching probability of an image-text pair. The cross-modal projection classification loss replaces the original single modal feature with cross-modal projection to calculate the norm-softmax. Wang *et al.* [19] propose a mutually connected classification loss to effectively utilize the identity-level information. They also use an attention module to capture the different contributions of words in the text representation network. Jing *et al.* [20] propose a pose-guided multi-granularity attention network to learn multi-granularity image-text relevance. In the coarse alignment network, a hard attention module selects the image regions related to the text to calculate the similarity score. In the fine-grained alignment network, the human body part and noun phrase are aligned through the guidance of pose information.

However, these methods ignore the information imbalance between the image and the text, and include image-specific information in the feature extraction process. In this paper, we suppress image-specific information through knowledge adaptation. We also extract more robust image features under the supervision of texts, which are inherently capable of telling key and subtle details of people.

B. Knowledge Distillation

Knowledge distillation is a way of transferring knowledge from the teacher model to a student model to improve the performance of student networks. The idea can be traced back to Breiman and Shang's work [21]. In recent years, it has been proven effective by many works [22]–[29]. Hinton *et al.* [22] exploit this method in the image classification task and name it knowledge distillation. Romero *et al.* [30] propose a framework to compress wide and deep networks into thin and deeper ones, by introducing intermediate-level hints from the teacher hidden layers to guide the training process of the student. Zagoruyko and Komodakis [24] present several ways of transferring attention from one network to another. Sau and Balasubramanian [31] propose a simple methodology to include a noise-based regularizer while training the student from the teacher, which provides a healthy improvement in the performance of the student network. Yim *et al.* [32] define the distilled knowledge to be transferred in terms of flow between layers, which is calculated by computing the inner product between features from two layers. Park *et al.* [33] introduce a approach, dubbed relational knowledge distillation (RKD), that transfers mutual relations of data examples.

Recently, Knowledge distillation is also designed to adapt to other tasks such as object detection, semantic segmentation, and natural language processing. Chen *et al.* [34] use knowledge distillation to improve object detection performance. They propose to handle the regression component through a teacher bounded loss. They also help the student to learn from the intermediate teacher distributions by using adaptation layers. Liu *et al.* [35] propose the pair-wise distillation and the holistic distillation for dense prediction tasks. The pair-wise distillation constrains the pair-wise similarities of the outputs of the student network and the teacher network to be consistent. The holistic distillation transfers holistic knowledge by using adversarial learning. For natural language processing tasks, Jiao *et al.* [36] propose to distill the knowledge of BERT, which has good performance but is computationally intensive. A two-stage learning framework is proposed to perform Transformer-based distillation at the pre-training and fine-tuning stages.

Note that in most of these works, knowledge distillation aims to improve the performance of a single student network, which has a smaller size than the teacher network. The teacher network and the student network are of the same type and have the same input. However, the motivation and framework of this paper are different. In this paper, knowledge adaptation aims to improve the matching ability of two networks by balancing the information of them. And the image network and the text network are of different types and have inputs from different modalities.

C. Cross-Modality Re-Identification

Language-based person search is related to other cross-modality re-identification tasks such as RGB-infrared(IR) re-identification. Several methods are proposed for the task. Ye *et al.* [37] propose a two-stage framework. In the first stage, the identity loss and the contrastive loss are used for feature learning. In the second stage, the modality-specific and modality-shared metrics are optimized for metric learning. Wu *et al.* [38] introduce a focal modality-aware similarity-preserving loss to keep the cross-modality similarity consistent with the same-modality similarity. They also design a modality-gated node as a universal structure that can represent both modality-specific and shared nodes to construct a structure-learnable feature extractor. Wang *et al.* [39] jointly model pixel alignment and feature alignment. By playing a min-max game among the pixel generator, the feature generator and the joint discriminator, the identity-consistent features are learned, and the cross-modality and intra-modality variation is reduced.

Most of these methods treat two modalities in the same way. However, in language-based person search, the relationship between the image modality and the text modality is asymmetric. Since text is annotated based on image, text information can be regarded as a subset of image information. Therefore, unlike the above methods, we perform knowledge adaptation in an asymmetric manner. We only adapt image knowledge to text knowledge to suppress image-specific information.

III. Proposed Method

In this section, we present the proposed approach in detail. First, we introduce the overall framework of the method.

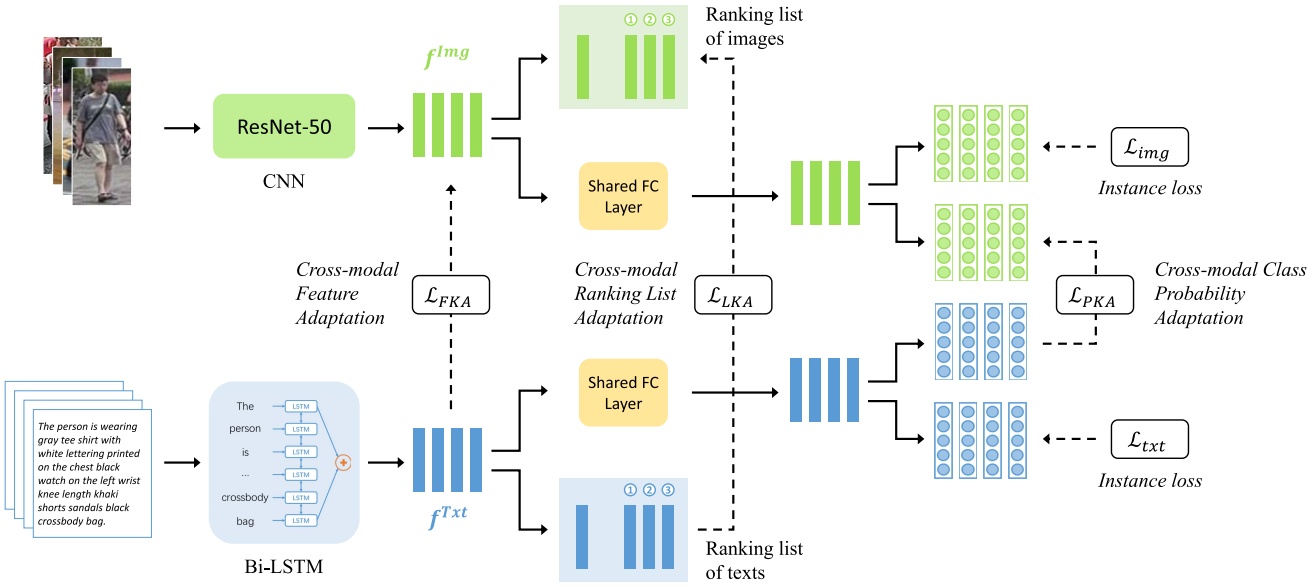


Fig. 3. The overall architecture of the proposed CMKA method. An image representation network is used to extract image features. A text representation network is used to extract text features. The proposed cross-modal knowledge adaptation is performed at three different levels: individuals, lists, and classes. The instance losses $\mathcal{L}_{img}$ and $\mathcal{L}_{txt}$ are used to ensure the discriminability of the image representation network and the text representation network.

Then we describe the network architecture of the image representation network and the text representation network. Next, we illustrate the proposed CMKA method. Finally, the objective function is presented.

The overall framework of the method is shown in Fig.3. As shown in the figure, an image representation network is used to extract image features. A text representation network is used to extract text features. After the shared fully connected layer projects the features of the two modalities into one common space, instance loss is adopted to ensure the discriminability of the image and text networks through the supervision of person identity. Cross-modal knowledge adaptation is performed at three levels so that the information extraction of the image representation network is guided by the text representation network.

A. Image and Text Representation Network

Given a collection of N tuples, denoted $\Psi = (I_i, T_i, c_i^{gt})_{i=1}^{N}$, each tuple contains an image sample I_i, a text sample T_i, and an identity label of the image sample and the text sample c_i^{gt}. We construct an image representation network to extract the feature of I_i and a text representation network to extract the feature of T_i.

For the image I_i, we use ResNet-50 [40] with the final fully-connected layer removed as image representation network to extract the visual feature. To make the dimension of the image feature consistent with the dimension of the text feature, we perform a 1×1 convolution before the average pooling layer to convert the number of channels to 1,024. Finally, the network generates a 1,024-dimensional feature vector f_i^{Img} as follows,

$$f_i^{Img} = F^{Img}(I_i), \tag{1}$$

where $F^{Img}(\cdot)$ represents the forward pass process of the image representation network.

For a sentence T_i with the same identity as the image I_i, we encode its t-th word into a length-K one-hot vector $v_i(t)$, where K is the vocabulary size. Then we transform the one-hot vector into a 300-dimensional word embedding:

$$x_i(t) = W_e v_i(t), \tag{2}$$

where W_e is a $300 \times K$-dim embedding matrix. Since bi-directional LSTM (bi-LSTM) is useful for capturing the bidirectional long-term dependencies between words, We adopt bi-LSTM to extract text features. The hidden states of forward and backward directions are formulated as

$$\overrightarrow{h_i(t)} = \overrightarrow{LSTM}(x_i(t), \overrightarrow{h_i(t-1)}), \tag{3}$$

$$\overleftarrow{h_i(t)} = \overleftarrow{LSTM}(x_i(t), \overleftarrow{h_i(t+1)}), \tag{4}$$

where $\overrightarrow{h_i(t-1)}$ is the previous hidden state of forward direction. $\overleftarrow{h_i(t+1)}$ is the next hidden state of backward direction. Then the hidden states of forward and backward directions are concatenated, and a max-pooling strategy is used to obtain the 1,024-dimensional text representation f_i^{Txt}, which can be formalized as

$$f_i^{Txt} = \max_t(concat(\overrightarrow{h_i(t)}, \overleftarrow{h_i(t)})) \tag{5}$$

B. Cross-Modal Feature Adaptation

At the individual level, we propose cross-modal feature adaptation to avoid the imbalanced relationship between different modalities. Intuitively, a text is a description based on the person in its corresponding image, which means that most of its information is derived from the image. However, some of the information in the image is hardly mentioned by natural language descriptions. For example,

when people describe pedestrians in images, they usually pay attention to their appearance, and rarely describe lighting condition, background, or very specific pose of pedestrians. We refer to these image information that rarely appears in texts as image-specific information. Due to the existence of image-specific information, the information of the image representation and the text representation is unbalanced, which increases the gap between the two modalities.

To help the image representation network suppress the image-specific information, we enforce the image representation to fit the text representation. The objective function of cross-modal feature knowledge adaptation (FKA) can be formulated as minimizing the squared difference between the image feature and the corresponding text feature:

$$\mathcal{L}_{FKA} = \left\| f_i^{Img} - f_i^{Txt} \right\|^2 \tag{6}$$

Since we only adapt the image feature to the text feature, $\mathcal{L}_{FKA}$ is not back-propagated through the text representation network during model training.

C. Cross-Modal Ranking List Adaptation

In addition to the knowledge of an individual data point, we also adapt the high-order knowledge of the ranking list. For different modalities, the ranking lists of candidates may be different for a given query. In the image modality, the ranking takes image-specific information into account. The candidate image with background or illumination similar to the query image may have high rank inappropriately. In contrast, the text is naturally close to high-level semantics. It directly summarizes the appearance characteristics of a person. Therefore, the candidates are mainly ranked based on people's appearance information in the text modality, which is more reasonable. For this consideration, we adapt the list-wise knowledge of image modality to the list-wise knowledge of text modality.

Inspired by the learning-to-rank technique [25], [41], we perform permutation probability adaptation between different modalities. Given the i-th sample in the batch as query and the remaining samples as candidate samples, a permutation of the candidate samples is denoted by $\pi_i = \langle \pi_i(1), \pi_i(2), \ldots, \pi_i(N-1) \rangle$, which means that the $\pi_i(j)$-th sample in the batch is ranked j-th. Then the probability of the permutation in the image modality is defined as

$$P_i^{Img}(\pi_i) = \prod_{j=1}^{N-1} \frac{exp(S(f_i^{Img}, f_{\pi_i(j)}^{Img})}{\sum_{k=j}^{N-1} exp(S(f_i^{Img}, f_{\pi_i(k)}^{Img}))}, \tag{7}$$

where $S(f_a, f_b)$ is the similarity score between two features. We use the similarity score function defined based on Euclidean distance:

$$S(f_a, f_b) = -\alpha \left\| f_a - f_b \right\|^{\beta}, \tag{8}$$

where α and β are two tunable parameters. Similarly, the probability of permutation π_i in the text modality can be obtained as

$$P_i^{Txt}(\pi_i) = \prod_{j=1}^{N-1} \frac{exp(S(f_i^{Txt}, f_{\pi_i(j)}^{Txt})}{\sum_{k=j}^{N-1} exp(S(f_i^{Txt}, f_{\pi_i(k)}^{Txt}))} \tag{9}$$

Then the permutation which has the maximum probability in text modality is

$$\pi_i^{\star} = \arg\max_{\pi_i} P_i^{Txt}(\pi_i) \tag{10}$$

A nice property of the permutation probability defined in Eq.9 is that $\pi_i^{\star}$ is the sequence sorted in descending order of similarity score [42], which makes the calculation easier.

To adapt the permutation probability of image modality to the permutation probability of text modality, we maximize the probability of $\pi_i^{\star}$ in the image modality. The objective function of cross-modal list-wise knowledge adaptation (LKA) is formulated as

$$\mathcal{L}_{LKA} = -log(P_i^{Img}(\pi_i^{\star})) \tag{11}$$

Like $\mathcal{L}_{FKA}$, in our experiments, $\mathcal{L}_{LKA}$ is not back-propagated through the text representation network during model training unless specified.

The cross-modal ranking list adaptation naturally fits the objective of the retrieval task. It adapts rich structural knowledge of the entire data embedding space, which reflects the relationship of multiple samples.

D. Cross-Modal Class Probability Adaptation

The class probability distribution of a sample contains dark knowledge of the relationship of classes, which is essential to the generalization ability of the model. For example, the classes with high predicted probability for a sample are likely to be close to each other. Compared with image modality, the inter-class relationship of text modality can reflect the semantic relationship of the classes better, because there is no interference from the image-specific information. Based on this consideration, we further adapt knowledge at the level of predicted class probability to constrain the image network to generalize in the same way as the text network.

Inspired by knowledge distillation of Hinton *et al.* [22], we adapt the predicted class probability of the image to the class probability produced by the text model. The image feature and the text feature pass through the final fully connected layer of shared parameters:

$$z_i^{Img} = W^{share} f_i^{Img} \tag{12}$$

$$z_i^{Txt} = W^{share} f_i^{Txt}, \tag{13}$$

where z_i^{Img} and z_i^{Txt} are d-dimensional output logits of the two modalities. d is the number of classes. W^{share} is the parameter of the final fully connected layer. It is shared by the two modalities to encourage the image feature and the text feature to be projected into one common space.

Then, we convert the logits of each class, $z_i^{Img}(c)$ and $z_i^{Txt}(c)$, into the class probability. $z_i^{Img}(c)$ and $z_i^{Txt}(c)$ are compared with other logits by softmax with temperature:

$$q_i^{Img}(c) = \frac{exp(z_i^{Img}(c)/\tau)}{\sum_{c'} exp(z_i^{Img}(c')/\tau)} \tag{14}$$

$$q_i^{Txt}(c) = \frac{exp(z_i^{Txt}(c)/\tau)}{\sum_{c'} exp(z_i^{Txt}(c')/\tau)}, \tag{15}$$

where $q_i^{Img}(c)$ and $q_i^{Txt}(c)$ are class probability of the two modalities. τ is the temperature parameter that has an influence on the transformation performance of softmax function. The higher value τ has, the softer class probability it produces.

In order to suppress the image-specific information, We use Kullback–Leibler divergence, a commonly used measure of the difference between two probability distributions, to match the predicted class probabilities of the image and the corresponding text. The objective function of cross-modal class probability knowledge adaptation (PKA) is formulated as

$$\mathcal{L}_{PKA} = KL(q_i^{Txt}, q_i^{Img}), \tag{16}$$

where $KL(\cdot)$ is the Kullback-Leibler divergence between two probabilities. Similarly, $\mathcal{L}_{PKA}$ is not back-propagated through the text representation network during model training unless specified.

Note that we do not directly use the class probability of instance loss(see Sec.III-E) for knowledge adaptation, where the temperature can be regarded as 1. This is because the less extreme probability is more informative for knowledge adaptation. It implies the relative probabilities of incorrect answers, which tell a lot about how the text model tends to generalize.

In the three levels of knowledge adaptation, we adapt the knowledge of image to the knowledge of text. Another option for performing knowledge adaptation is to adapt the knowledge of text to the knowledge of image. The knowledge adaptation of this direction is not appropriate. Since text is generated from image, text information can be regarded as a subset of image information. Therefore, text does not have much modal-specific information to be eliminated. We also conduct knowledge adaptation experiments in this direction and both directions. Both are worse than just adapting the knowledge of image to the knowledge of text, which is discussed in Sec.IV-B.

E. Objective Function

In order to ensure the intra-modal discriminability of the image representation network and the text representation network, we also adopt the instance loss [4] to learn to classify the features of each modality. It can be formulated as

$$o_i^{Img} = softmax(W^{share} f_i^{Img}), \tag{17}$$

$$\mathcal{L}_{img} = -log(o_i^{Img}(c_i^{gt})), \tag{18}$$

$$o_i^{Txt} = softmax(W^{share} f_i^{Txt}), \tag{19}$$

$$\mathcal{L}_{txt} = -log(o_i^{Txt}(c_i^{gt})), \tag{20}$$

where $o_i(c_i^{gt})$ is the predicted possibility of the right class c_i^{gt}.

To summarize, the final objective function is expressed as the weighted sum of the identification loss and the proposed losses:

$$\mathcal{L} = \lambda_0(\mathcal{L}_{img} + \mathcal{L}_{txt}) + \lambda_1\mathcal{L}_{FKA} + \lambda_2\mathcal{L}_{LKA} + \lambda_3\mathcal{L}_{PKA}, \tag{21}$$

where λ_0, λ_1, λ_2, and λ_3 are the weights to balance the effects of different losses.

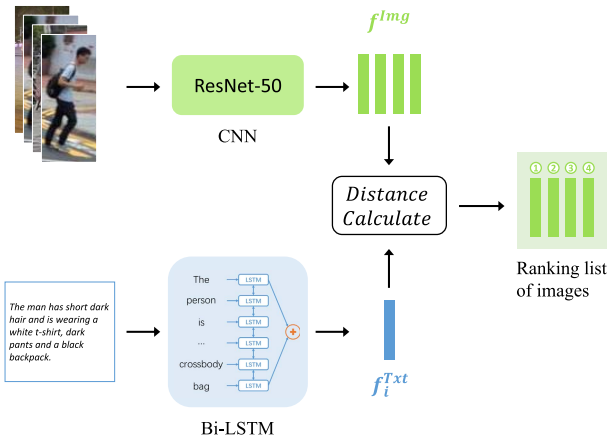


Fig. 4. The test network of the proposed CMKA method. It consists of an image representation network and a text representation network.

F. Model Testing

As shown in Fig.4, during testing, we extract the 1,024-dimensional text feature f_i^{Txt} by text representation network and the 1,024-dimensional image feature f_j^{Img} by image representation network. We use the cosine distance to evaluate the similarity between the query sentence and each candidate image. The final retrieval result is based on the similarity ranking.

IV. Experiments

We evaluate the proposed approach following the standard protocol on the language-based person search benchmark. We compare the proposed CMKA to state-of-the-art methods. Extensive experiments demonstrate that CMKA achieves superior performance over the state-of-the-art methods. Moreover, we investigate the effectiveness of each component of CMKA.

A. Experimental Settings

1) Dataset: To our knowledge, there is only one public dataset for language-based person search, i.e., CUHK-PEDES [1]. The CUHK-PEDES dataset contains 40,206 images of 13,003 person identities. Each image is described by two sentences. The training set consists of 11,003 persons, 34,054 images, and 68,108 sentence descriptions. The validation set and test set contain 3,078 and 3,074 images, respectively, and both of them have 1,000 persons.

2) Evaluation Metric: The Recall@K (K = 1, 5, 10) is chosen to evaluate the performance following [1]. Recall@K (or R@K) indicates the percentage of the queries where at least one ground-truth is retrieved among the top-K results.

3) Implementation Details: For the text representation network, we set the hidden dimension of bi-LSTM as 512. The vocabulary size is 7,163. We use the word2vec initialization for the embedding matrix W_e. For the image representation network, we follow some of the settings in [4]. We initialize the weights of image representation network with ResNet-50 pre-trained on ImageNet [43]. While training, the images are resized to 224 × 224 pixels which are randomly cropped

TABLE I
ABLATION STUDIES ON DIFFERENT COMPONENTS OF THE PROPOSED FRAMEWORK ON THE CUHK-PEDES DATASET

Method	Losses					R@1	R@5	R@10
	$\mathcal{L}_{img}$	$\mathcal{L}_{txt}$	$\mathcal{L}_{FKA}$	$\mathcal{L}_{LKA}$	$\mathcal{L}_{PKA}$			
Baseline	✓	✓				46.91	68.13	77.00
FKA	✓	✓	✓			49.64	70.66	79.40
FKA+LKA	✓	✓	✓	✓		51.88	72.69	80.00
FKA+LKA+PKA (CMKA)	✓	✓	✓	✓	✓	**54.69**	**73.65**	**81.86**

from images whose shorter size is 256. We also perform simple data augmentation such as horizontal flipping. Dropout is applied to the image representation network and the text representation network, and the dropout rate is 0.8. We set the max text length to 56. The batch size is 32. We set both α and β in Eq.8 to be 3. We set temperature τ to be 4. To better train our model, we split the training procedure into two stages. In the first stage, we fix the parameters of the pre-trained image representation network and use only the instance loss ($\lambda_0 = 1$, $\lambda_1 = 0$, $\lambda_2 = 0$, $\lambda_3 = 0$) to train the remaining part for 50 epochs. The SGD optimizer is employed for optimization with a learning rate of 0.001. In the second stage, we use the overall loss as Eq.21 ($\lambda_0 = 1$, $\lambda_1 = 1$, $\lambda_2 = 0.1$, $\lambda_3 = 10$) and fine-tune the entire network. The model is optimized with the Adam optimizer. We start training with a learning rate of 0.0001 for 40 epochs. Then we train with a learning rate of 0.00001. We stop training when the loss converges. We also conduct the horizontal flipping when testing and use the average features (no flip and flip) as the image feature.

B. Ablation Studies

1) Effectiveness of Each Component: In Table I, we evaluate the effectiveness of different components of the proposed method on the CUHK-PEDES dataset. We implement a baseline and several variants of the model. The loss constraints adopted by each method are also illustrated in the table. *Baseline* only uses instance loss. *FKA* additionally adapts knowledge at the feature level. *FKA + LKA* uses the feature adaptation and the list-wise adaptation. *FKA + LKA + PKA* combines *FKA*, *LKA*, and *PKA* to adapt knowledge at three levels.

For R@1, R@5, and R@10, *FKA* improves 2.73%, 2.53%, and 2.4% over *Baseline*. The result indicates that feature adaptation can promote the balance of information between modalities. Textual-visual correspondences between the image and the text are more effectively learned. *LKA* further increases the performance from 49.64%, 70.66%, and 79.40% to 51.88%, 72.69%, and 80%, respectively. When *FKA*, *LKA*, and *PKA* are combined, the performance is once again enhanced to 54.69%, 73.65%, and 81.86%, and achieves the best. Compared with *Baseline*, the improvement reaches 7.78%, 5.52%, and 4.86% on the three indicators. This result demonstrates that *FKA*, *LKA*, and *PKA* are complementary to each other. Each component of our method makes contribution for better matching the image and the text.

2) Adapting Text Knowledge to Image Knowledge: Another option for knowledge adaptation is to adapt the information of the text to the information of the image, which means that during model training, the loss is not back-propagated through the image representation network, but through the text representation network. The result is shown by *CMKA (Image-to-Text)* in Table II. It can be seen that Image-to-Text CMKA deteriorates the performance of *Baseline*. The R@1, R@5, and R@10 of Image-to-Text CMKA are 2.08%, 1.87%, and 1.68% lower than *Baseline*, respectively. The table also shows the result of CMKA in both directions. The R@1, R@5, and R@10 of *CMKA (Both directions)* are 50.28%, 70.05%, and 77.49%, respectively. The result is better than *Baseline* but worse than *CMKA (Text-to-Image)*.

TABLE II
COMPARISON OF THE RESULTS OF DIFFERENT KNOWLEDGE ADAPTATION DIRECTIONS ON THE CUHK-PEDES DATASET

Method	R@1	R@5	R@10
Baseline	46.91	68.13	77.00
CMKA (Image-to-Text)	44.83	66.26	75.32
CMKA (Text-to-Image)	**54.69**	**73.65**	**81.86**
CMKA (Both directions)	50.28	70.05	77.49

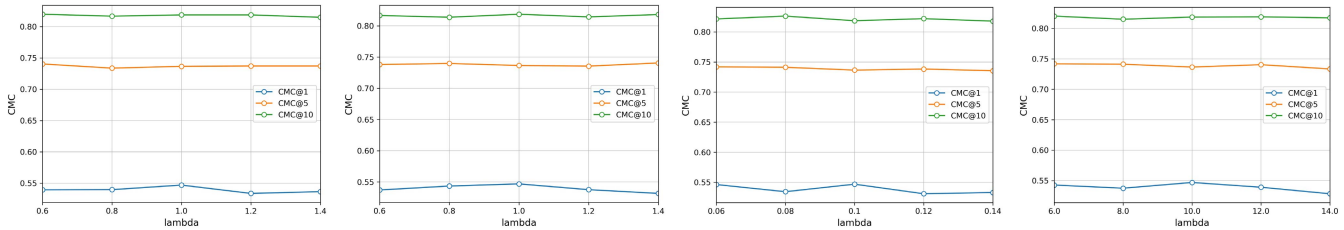
TABLE III
COMPARISON OF RESULTS USING DIFFERENT TRAINING WAYS ON THE CUHK-PEDES DATASET

Method	R@1	R@5	R@10
Baseline	46.91	68.13	77.00
CMKA-f	50.57	70.19	79.00
CMKA	**54.69**	**73.65**	**81.86**

These results suggest that adapting text knowledge to image knowledge is not appropriate for the language-based person search task. On the one hand, in language-based person search, the information of the texts is mainly derived from images. Texts do not have much modal-specific information to be eliminated. On the other hand, it is not practical to enforce the text representation to fit the image representation, which inherently has extra information.

3) Influence of Hyper-Parameters: We evaluate the effect of λ_0, λ_1, λ_2, and λ_3 on our method. When investigating the impact of a hyper-parameter, we fix other hyper-parameters. The results are shown in Fig.5. The results show that the performance of our method remains stable with the change of λ_0, λ_1, λ_2, and λ_3.

4) Knowledge Adaptation With Text Representation Network Fixed: In our approach, the text representation network as the guider is being trained while the image knowledge is adapted to the text knowledge. We also evaluate the way in which knowledge adaptation and text representation network



(a) λ_0 (b) λ_1 (c) λ_2 (d) λ_3

Fig. 5. Evaluation of the proposed method with different values of hyper-parameter λ_0, λ_1, λ_2, and λ_3 on the CUHK-PEDES dataset.

TABLE IV
COMPARISON OF RESULTS USING FKA AND OTHER ALIGNMENT METHODS ON THE CUHK-PEDES DATASET

Method	R@1	R@5	R@10
CMKA-FKA	**54.69**	**73.65**	**81.86**
CMKA-MMD	52.76	73.03	81.17
CMKA-Deep CORAL	52.97	72.63	80.73
CMKA-adversarial learning	53.44	73.39	81.22

training are conducted separately. Specifically, we use the *Baseline* model as the pre-trained model. We adapt the image knowledge to the text knowledge and fix the text representation network. The results are shown in Table III. In the table, *CMKA-f* is the method of knowledge adaptation with the text representation network fixed. It can be seen that *CMKA-f* outperforms *Baseline*. However, its performance is worse than *CMKA*. The result shows that knowledge adaptation can in turn help the learning of text representation. These two processes promote each other.

5) Replacing FKA With Other Alignment Methods: Compared with existing alignment methods [44]–[52], FKA adopts squared distance loss and only back-propagates the gradient through the image representation network to adapt knowledge in a single direction. We compare the results of using FKA or other methods such as MMD [44], Deep CORAL [45], and adversarial learning [46] at the individual level. The results are shown in Table IV. The methods in the table use different losses at the individual level and the same losses at the other two levels. It can be seen that replacing FKA with MMD, Deep CORAL, or adversarial learning does not improve performance. The results show that FKA is not only easy to implement, but also effective.

C. Results of I2T, I2I and T2T Re-ID

In addition to Text-to-Image re-ID (i.e., Language-based person search), we also evaluate the results of Image-to-Text (I2T), Image-to-Image (I2I) and Text-to-Text (T2T) re-ID in Table V. All evaluations use the same image features and text features as T2I re-ID. In the I2I and T2T re-ID settings, since a sample can be easily matched by itself, for each query, we remove the same sample from the gallery as this query. Some interesting findings can be found in the results.

For I2T re-ID, CMKA increases R@1, R@5, and R@10 from 61.68%, 84.48%, and 90.40% to 63.47%, 88.03%, and 93.56%, respectively. This result once again demonstrates the effectiveness of our method, which reduces the gap between different modalities to facilitate cross-modal matching.

Interestingly, for I2I re-ID, we find that the performance of CMKA is worse than the method of not using CMKA. This is because knowledge adaptation eliminates some of the image-specific information that is useful for distinguishing images of different classes. In I2I re-ID task, due to a person's limited scope of motion, the images of the same person may be in similar lighting conditions and background. And the images of different people may be in different lighting conditions and background. The image-specific information such as lighting condition and background can indeed be used as factors in determining the identity of an image in I2I re-ID. And it is not necessary to consider whether information from an image is also owned by other modalities. However, these image-specific information does not contribute to cross-modal re-ID because only information shared by different modalities can be used as a reference for comparing image and text. The result reflects the key difference between the I2I re-ID task and the T2I re-ID task and shows that CMKA can effectively suppress the image-specific information.

Another finding from the table is that although the proposed knowledge adaptation is only applied to image modality, the T2T re-ID performance is also improved. This is because CMKA reduces the interference of image-specific information and makes the model capture more textual-visual correspondences, which in turn improves the generalization ability of the text representation network.

D. Comparison to State-of-the-Arts

In Table VI, the proposed approach is compared with several state-of-the-art language-based person search methods (including deeper LSTM Q+norm I [53], iBOWIMG [54], NeuralTalk [55], Word CNN-RNN [56], GNA-RNN [1], GMM+HGLMM [57], IATV [3], PWM-ATH [2], DPCE [4], GLIA [17], CMPM+CMPC [18], MCCL [19], TIMAM [58], and PMA [20]) on the CUHK-PEDES dataset. To compare fairly, for TIMAM, we show the results without BERT. We use the same protocol to evaluate these methods.

Among existing methods, PWM-ATH, GLIA, and PMA model the correspondences between noun-phrases and image regions. GNA-RNN, IATV, PWM-ATH, GLIA, MCCL, and PMA introduce attention mechanisms to enhance the learning of image representation network or text representation

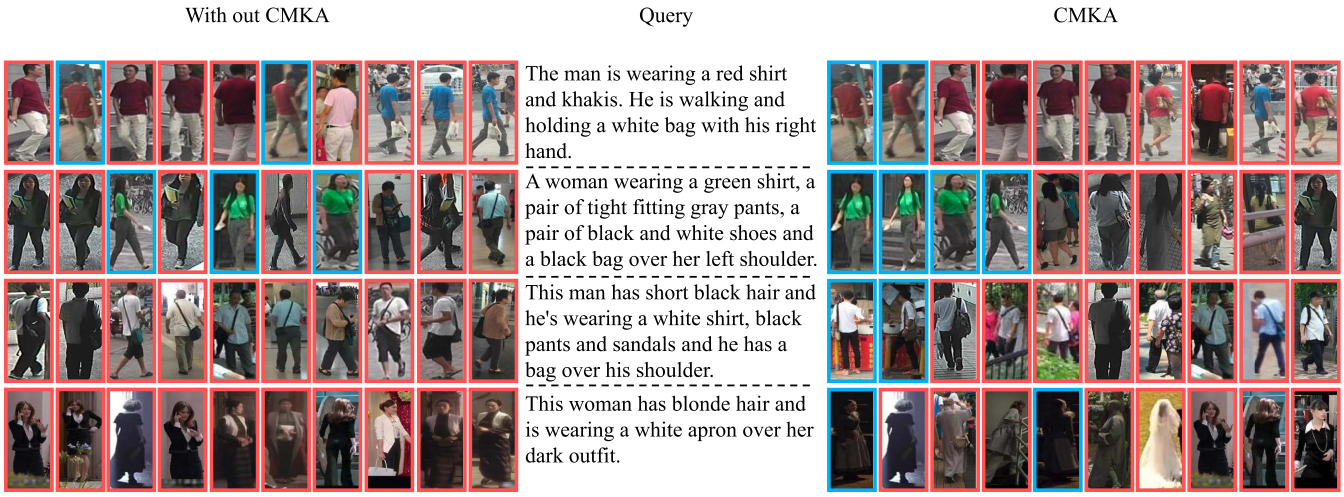


Fig. 6. Retrieval results on the CUHK-PEDES dataset. Positive images are marked by blue rectangles. Negative images are marked by red rectangles.

TABLE V
RESULTS OF T2I, I2T, I2I, AND T2T RE-ID ON THE CUHK-PEDES DATASET

Method	Text-to-Image			Image-to-Text			Image-to-Image			Text-to-Text		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
Baseline	46.91	68.13	77.00	61.68	84.48	90.40	**89.92**	**95.93**	**97.17**	38.26	60.92	69.92
CMKA	**54.69**	**73.65**	**81.86**	**63.47**	**88.03**	**93.56**	89.20	95.54	97.07	**40.09**	**61.91**	**70.63**

TABLE VI
COMPARISON OF LANGUAGE-BASED PERSON SEARCH RESULTS (R@K(%)) ON THE CUHK-PEDES DATASET

Method	R@1	R@5	R@10
deeper LSTM Q+norm I [53]	17.19	-	57.82
iBOWIMG [54]	8.00	-	30.56
NeuralTalk [55]	13.66	-	41.72
Word CNN-RNN [56]	10.48	-	36.66
GNA-RNN [1]	19.05	-	53.64
GMM+HGLMM [57]	15.03	-	42.27
IATV [3]	25.94	-	60.48
PWM-ATH [2]	27.14	49.45	61.02
DPCE [4]	44.40	66.26	75.07
GLIA [17]	43.58	66.93	76.26
CMPM+CMPC [18]	49.37	-	79.27
MCCL [19]	50.58	-	79.06
TIMAM [58]	51.30	-	**82.40**
PMA [20]	53.81	73.54	81.23
CMKA(ours)	**54.69**	**73.65**	81.86

network. CMKA uses only the global features of image and text and does not use any attention mechanism, which is simpler and more efficient. Moreover, our approach outperforms previous works and achieves state-of-the-art results. Specifically, CMKA is 0.88%, 0.11%, and 0.63% better than the previous best performing method PMA in terms of R@1, R@5, and R@10. Note that PMA requires training a pose estimation model on an additional dataset that annotates keypoints of human body. Our approach does not require any annotation of human key points or external dataset. The results validate the effectiveness of the proposed method. By suppressing the image-specific information, cross-modal knowledge adaptation can reduce the imbalance of information between modalities, and help the model capture more textual-visual correspondences. As a result, the performance of language-based person search is improved.

E. Results on the Flickr30K Dataset

To make our method more convincing and generic, we also conduct experiments on Flickr30K [65], which is one of the image captioning datasets. In Table VII, the proposed approach is compared with other methods including RRF-Net [60], CMPM+CMPC [18], DPCE [4], DAN [59], NAR [61], VSE++ [62], SCO [63], GXN [64], and TIMAM [58]. Similar to most methods, on this dataset, we use ResNet-152 for image representation network.

In image-to-text matching, CMKA achieves the best results on R@5 and R@10, which are 82.9% and 90.0%, respectively. In text-to-image matching, CMKA outperforms existing methods by a large margin in all metrics. Compared with the second best method TIMAM-BERT, CMKA does not introduce BERT, and is still 2.4%, 1.8%, and 0.8% better than it on R@1, R@5, and R@10. The results again demonstrate the advantage of our method, which balances information between different modalities through cross-modal knowledge adaptation.

F. Qualitative Results

We conduct a qualitative evaluation of the proposed method. Fig.6 shows the results with and without CMKA. The images on the left are search results without CMKA, and the images on the right are search results with CMKA. Positive images are marked by blue rectangles. Negative images are marked by red rectangles.

The first two examples show that CMKA can help the model to discover detailed textual-visual correspondences.

TABLE VII
MATCHING RESULTS (R@K(%)) ON THE FLICKR30K DATASET

Method	Image Backbone	Image-to-Text			Text-to-Image			Sum
		R@1	R@5	R@10	R@1	R@5	R@10	
DAN [59]	VGG-19	41.4	73.5	82.5	31.8	61.7	72.5	363.4
RRF-Net [60]	ResNet-152	47.6	77.4	87.1	35.4	68.3	79.9	395.7
CMPM+CMPC [18]	ResNet-152	49.6	76.8	86.1	37.3	65.7	75.5	391.0
TIMAM [58]	ResNet-152	50.1	-	-	37.9	-	-	-
DPCE [4]	ResNet-152	55.6	81.9	89.5	39.1	69.2	80.9	416.2
DAN [59]	ResNet-152	55.0	81.8	89.0	39.4	69.2	79.1	413.5
NAR [61]	ResNet-152	55.1	80.3	89.6	39.4	68.8	79.9	413.1
VSE++ [62]	ResNet-152	52.9	80.5	87.2	39.6	70.1	79.5	409.8
SCO [63]	ResNet-152	55.5	82.0	89.3	41.1	70.5	80.1	418.5
GXN [64]	ResNet-152	**56.8**	-	89.6	41.5	-	80.1	-
TIMAM-BERT [58]	ResNet-152	53.1	78.8	87.6	42.6	71.6	81.9	415.6
CMKA(ours)	ResNet-152	55.7	**82.9**	**90.0**	**45.0**	**73.4**	**82.7**	**429.7**

Specifically, in the description of the first line, "red shirt and khaki" can be easily observed from an image. Therefore, most of the search results are for people wearing red shirt and khaki pants, whether CMKA is used or not. However, some non-corresponding people also wear red shirt and khaki pants. "a white bag with his right hand" in the text is the key feature that helps to accurately find the corresponding person from these very similar people. Compared with the method without using CMKA, the method using CMKA can put the corresponding images at the top through this detail. This is because CMKA reduces the interference of image-specific information, enabling the model to capture more detailed textual-visual correspondences.

Similarly, most of the search results in the second row match the description of "green shirt", "gray pants" and "black and white shoes". CMKA can further place the corresponding images in the top position by "black bag over her left shoulder", which is what the method without CMKA fails to do.

The example in the third row shows that CMKA can reduce the negative effect of background information. In the example, the method without CMKA almost misses all the corresponding images. This may be because the corresponding images contain an uncommon red sign in the background. The image-specific information of background increases the gap between image modality and text modality. After CMKA suppressing the image-specific information, the amount of information in the image and text is balanced, and the corresponding image and text are matched.

The example in the fourth row shows that CMKA can reduce the negative effect of lighting condition information. In the example, the method without CMKA almost misses all the corresponding images. This may be caused by too much image-specific information on the low light condition in these corresponding images. By using CMKA, the image-specific information of lighting condition is suppressed, and the corresponding image and text are matched by the modal agnostic information.

G. Visualization

1) Visualization of Feature Map: To better understand how CMKA helps image network learn discriminative features, we visualize the activation map of the image feature before the average pooling layer in Fig.7. Following [66], we calculate the activation map by summarizing the absolute-valued feature maps along the channel dimension. As shown in the first example, both the features extracted by the baseline method and the features extracted by CMKA have a high response on the human body. However, the baseline model also focuses on the pink object in the background, which is redundant for the matching process. CMKA can focus on the white bag

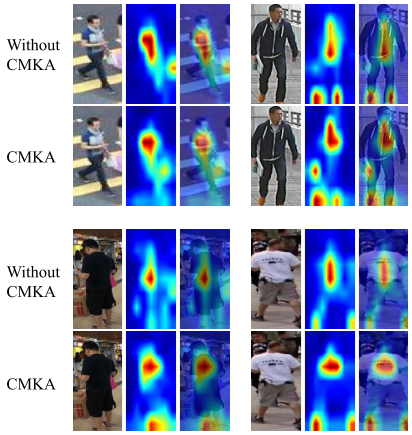


Fig. 7. Activation maps of image features.

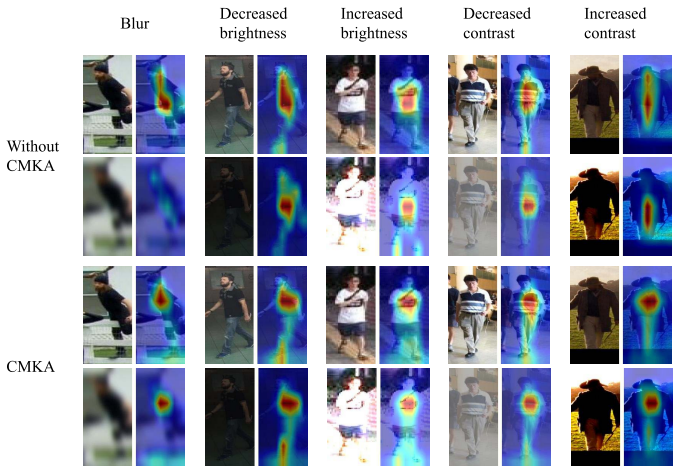
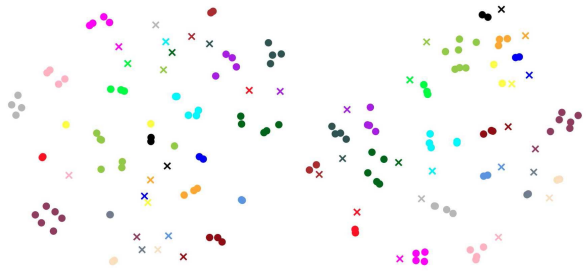


Fig. 8. Activation maps of images with different image-specific information.



(a) Without CMKA (b) CMKA

Fig. 9. Visualization of the feature distribution with and without CMKA. Different colors indicate different IDs. The circle represents the image feature. The cross represents the text feature.

in the human hand which is critical to the matching process instead of the background object. In other examples, CMKA can focus on the white book in the human hand, the sandals, and the pattern on T-shirt, respectively. In contrast, the baseline method ignores these important details. The results show that the proposed method can help the image network avoid the interference of image-specific information and discover discriminative information.

2) Stability of Feature Map to Changes in Image-Specific Information: In Fig.8, we compare the activation maps of the baseline method and CMKA by adjusting the image-specific information of the input image, such as blur, brightness, and contrast. As shown in the figure, after the image is blurred, the baseline method fails to capture important local regions. However, CMKA keeps focusing on the discriminative region of the image. When the brightness or contrast of the image changes, the high response region in the activation map of the baseline method is easily affected by the image variation. In contrast, the activation area of CMKA remains consistent. The results show that the image feature learned without CMKA contains more image-specific information and is therefore susceptible to image changes. CMKA can suppress the image-specific information and extract features with semantic information shared by modalities. As a result, the activation map is stable to changes in the image-specific information of the input image, such as blur, brightness, and contrast.

3) Visualization of Feature Distribution: In Fig.9, the feature distribution with and without CMKA is shown by t-SNE [67]. Features are sampled from the test set of the CUHK-PEDES dataset. In Fig.9a, the features of the same modality are well aggregated by identity. However, due to the unbalanced knowledge between the image modality and the text modality, the image features and text feature of the same identity (such as the red circles and red cross) are far apart. In Fig.9b, since the image-specific information is suppressed by cross-modal knowledge adaptation, the information gap between different modalities is narrowed. As a result, the image features and text feature of the same identity (such as the red circles and red cross) are pulled closer. The comparison demonstrates that cross-modal knowledge adaptation can effectively balance the amount of information of different modalities and better construct the common space of image and text.

V. Conclusion

In this paper, we propose a cross-modal knowledge adaptation method for language-based person search. To suppress the image-specific information, we adapt the knowledge of image to the knowledge of text at different levels. As a result, the information between modalities is balanced and more textual-visual correspondences are learned. Experimental results have shown that the proposed method outperforms state-of-the-art methods.

References

[1] S. Li, T. Xiao, H. Li, B. Zhou, D. Yue, and X. Wang, "Person search with natural language description," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1970–1979.

[2] T. Chen, C. Xu, and J. Luo, "Improving text-based person search by spatial matching and adaptive threshold," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2018, pp. 1879–1887.

[3] S. Li, T. Xiao, H. Li, W. Yang, and X. Wang, "Identity-aware textual-visual matching with latent co-attention," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1890–1899.

[4] Z. Zheng, L. Zheng, M. Garrett, Y. Yang, M. Xu, and Y.-D. Shen, "Dual-path convolutional image-text embeddings with instance loss," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 16, no. 2, pp. 1–23, Jun. 2020.

[5] E. Yang, C. Deng, C. Li, W. Liu, J. Li, and D. Tao, "Shared predictive cross-modal deep quantization," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 11, pp. 5292–5303, Nov. 2018.

[6] C. Deng, X. Xu, H. Wang, M. Yang, and D. Tao, "Progressive cross-modal semantic network for zero-shot sketch-based image retrieval," *IEEE Trans. Image Process.*, vol. 29, pp. 8892–8902, 2020.

[7] D. Xie, C. Deng, C. Li, X. Liu, and D. Tao, "Multi-task consistency-preserving adversarial hashing for cross-modal retrieval," *IEEE Trans. Image Process.*, vol. 29, pp. 3626–3637, 2020.

[8] S. Bai, X. Bai, and Q. Tian, "Scalable person re-identification on supervised smoothed manifold," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2530–2539.

[9] R. Yu, Z. Dou, S. Bai, Z. Zhang, Y. Xu, and X. Bai, "Hard-aware point-to-set deep metric for person re-identification," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 188–204.

[10] L. Zheng, Y. Huang, H. Lu, and Y. Yang, "Pose-invariant embedding for deep person re-identification," *IEEE Trans. Image Process.*, vol. 28, no. 9, pp. 4500–4509, Sep. 2019.

[11] Z. Zhong, L. Zheng, Z. Zheng, S. Li, and Y. Yang, "CamStyle: A novel data augmentation method for person re-identification," *IEEE Trans. Image Process.*, vol. 28, no. 3, pp. 1176–1190, Mar. 2019.

[12] Z. Zheng, L. Zheng, and Y. Yang, "Pedestrian alignment network for large-scale person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 10, pp. 3037–3045, Oct. 2019.

[13] L. Wu, R. Hong, Y. Wang, and M. Wang, "Cross-entropy adversarial view adaptation for person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 7, pp. 2081–2092, Jul. 2020.

[14] D. A. Vaquero, R. S. Feris, D. Tran, L. Brown, A. Hampapur, and M. Turk, "Attribute-based people search in surveillance environments," in *Proc. Workshop Appl. Comput. Vis. (WACV)*, Dec. 2009, pp. 1–8.

[15] C. Su, S. Zhang, J. Xing, W. Gao, and Q. Tian, "Deep attributes driven multi-camera person re-identification," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 475–491.

[16] H. Wang, C. Deng, J. Yan, and D. Tao, "Asymmetric cross-guided attention network for actor and action video segmentation from natural language query," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3939–3948.

[17] D. Chen *et al.*, "Improving deep visual representation for person re-identification by global and local image-language association," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 54–70.

[18] Y. Zhang and H. Lu, "Deep cross-modal projection learning for image-text matching," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 686–701.

[19] Y. Wang, C. Bo, D. Wang, S. Wang, Y. Qi, and H. Lu, "Language person search with mutually connected classification loss," in *Proc. ICASSP-IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 2057–2061.

[20] Y. Jing, C. Si, J. Wang, W. Wang, L. Wang, and T. Tan, "Pose-guided multi-granularity attention network for text-based person search," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 11189–11196.

[21] L. Breiman and N. Shang, "Born again trees," Dept. Statist., Univ. California, Berkeley, Berkeley, CA, USA, Tech. Rep., 1996.

[22] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," in *Proc. Adv. Neural Inf. Process. Syst. Workshop*, 2015, pp. 1–9.

[23] G. Urban *et al.*, "Do deep convolutional nets really need to be deep and convolutional?" in *Proc. Int. Conf. Learn. Represent.*, 2017, pp. 1–13.

[24] S. Zagoruyko and N. Komodakis, "Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer," in *Proc. Int. Conf. Learn. Represent.*, 2017.

[25] Y. Chen, N. Wang, and Z. Zhang, "DarkRank: Accelerating deep metric learning via cross sample similarities transfer," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 1–8.

[26] X. Gu, B. Ma, H. Chang, S. Shan, and X. Chen, "Temporal knowledge propagation for image-to-video person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9647–9656.

[27] Q. Li, S. Jin, and J. Yan, "Mimicking very efficient network for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6356–6364.

[28] T. He, C. Shen, Z. Tian, D. Gong, C. Sun, and Y. Yan, "Knowledge adaptation for efficient semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 578–587.

[29] Y. Liu, K. Chen, C. Liu, Z. Qin, Z. Luo, and J. Wang, "Structured knowledge distillation for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2604–2613.

[30] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, "FitNets: Hints for thin deep nets," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–12.

[31] B. B. Sau and V. N. Balasubramanian, "Deep model compression: Distilling knowledge from noisy teachers," 2016, *arXiv:1610.09650*. [Online]. Available: https://arxiv.org/abs/1610.09650

[32] J. Yim, D. Joo, J. Bae, and J. Kim, "A gift from knowledge distillation: Fast optimization, network minimization and transfer learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4133–4141.

[33] W. Park, D. Kim, Y. Lu, and M. Cho, "Relational knowledge distillation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3967–3976.

[34] G. Chen, W. Choi, X. Yu, T. Han, and M. Chandraker, "Learning efficient object detection models with knowledge distillation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 742–751.

[35] Y. Liu, C. Shu, J. Wang, and C. Shen, "Structured knowledge distillation for dense prediction," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Jun. 12, 2020, doi: 10.1109/TPAMI.2020.3001940.

[36] X. Jiao *et al.*, "TinyBERT: Distilling bert for natural language understanding," 2019, *arXiv:1909.10351*. [Online]. Available: https://arxiv.org/abs/1909.10351

[37] M. Ye, X. Lan, J. Li, and P. Yuen, "Hierarchical discriminative learning for visible thermal person re-identification," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 1–8.

[38] A. Wu, W.-S. Zheng, S. Gong, and J. Lai, "RGB-IR person re-identification by cross-modality similarity preservation," *Int. J. Comput. Vis.*, vol. 128, no. 6, pp. 1765–1785, Jun. 2020.

[39] G. Wang, T. Zhang, J. Cheng, S. Liu, Y. Yang, and Z. Hou, "RGB-infrared cross-modality person re-identification via joint pixel and feature alignment," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3623–3632.

[40] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[41] F. Xia, T.-Y. Liu, J. Wang, W. Zhang, and H. Li, "Listwise approach to learning to rank: Theory and algorithm," in *Proc. 25th Int. Conf. Mach. Learn. (ICML)*, 2008, pp. 1192–1199.

[42] Z. Cao, T. Qin, T.-Y. Liu, M.-F. Tsai, and H. Li, "Learning to rank: From pairwise approach to listwise approach," in *Proc. 24th Int. Conf. Mach. Learn. (ICML)*, 2007, pp. 129–136.

[43] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[44] E. Tzeng, J. Hoffman, N. Zhang, K. Saenko, and T. Darrell, "Deep domain confusion: Maximizing for domain invariance," 2014, *arXiv:1412.3474*. [Online]. Available: http://arxiv.org/abs/1412.3474

[45] B. Sun and K. Saenko, "Deep coral: Correlation alignment for deep domain adaptation," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 443–450.

[46] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 1180–1189.

[47] G.-J. Qi, X.-S. Hua, and H.-J. Zhang, "Learning semantic distance from community-tagged media collection," in *Proc. 17th ACM Int. Conf. Multimedia (MM)*, 2009, pp. 243–252.

[48] J. Wang, Z. Zhao, J. Zhou, H. Wang, B. Cui, and G. Qi, "Recommending Flickr groups with social topic model," *Inf. Retr.*, vol. 15, nos. 3–4, pp. 278–295, 2012.

[49] G.-J. Qi, W. Liu, C. Aggarwal, and T. Huang, "Joint intermodal and intramodal label transfers for extremely rare or unseen classes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 7, pp. 1360–1373, Jul. 2017.

[50] G.-J. Qi, C. C. Aggarwal, and T. S. Huang, "Online community detection in social sensing," in *Proc. 6th ACM Int. Conf. Web Search Data Mining (WSDM)*, 2013, pp. 617–626.

[51] Y. Zhao, Z. Jin, G.-J. Qi, H. Lu, and X.-S. Hua, "An adversarial approach to hard triplet generation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 501–517.

[52] G.-J. Qi, L. Zhang, H. Hu, M. Edraki, J. Wang, and X.-S. Hua, "Global versus localized generative adversarial nets," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1517–1525.

[53] S. Antol *et al.*, "VQA: Visual question answering," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2425–2433.

[54] B. Zhou, Y. Tian, S. Sukhbaatar, A. Szlam, and R. Fergus, "Simple baseline for visual question answering," 2015, *arXiv:1512.02167*. [Online]. Available: https://arxiv.org/abs/1512.02167

[55] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3156–3164.

[56] S. Reed, Z. Akata, H. Lee, and B. Schiele, "Learning deep representations of fine-grained visual descriptions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 49–58.

[57] B. Klein, G. Lev, G. Sadeh, and L. Wolf, "Associating neural word embeddings with deep image representations using Fisher vectors," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4437–4446.

[58] N. Sarafianos, X. Xu, and I. Kakadiaris, "Adversarial representation learning for text-to-image matching," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5814–5824.

[59] H. Nam, J.-W. Ha, and J. Kim, "Dual attention networks for multimodal reasoning and matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 299–307.

[60] Y. Liu, Y. Guo, E. M. Bakker, and M. S. Lew, "Learning a recurrent residual fusion network for multimodal matching," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4107–4116.

[61] C. Liu, Z. Mao, W. Zang, and B. Wang, "A neighbor-aware approach for image-text matching," in *Proc. ICASSP-IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 3970–3974.

[62] F. Faghri, D. J. Fleet, J. Ryan Kiros, and S. Fidler, "VSE++: Improving visual-semantic embeddings with hard negatives," 2017, *arXiv:1707.05612*. [Online]. Available: http://arxiv.org/abs/1707.05612

[63] Y. Huang, Q. Wu, C. Song, and L. Wang, "Learning semantic concepts and order for image and sentence matching," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6163–6171.

[64] J. Gu, J. Cai, S. Joty, L. Niu, and G. Wang, "Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7181–7189.

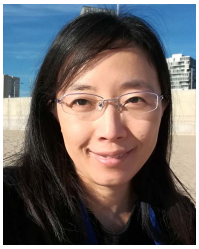
[65] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik, "Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2641–2649.

[66] K. Zhou, Y. Yang, A. Cavallaro, and T. Xiang, "Omni-scale feature learning for person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3702–3712.

[67] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.

Yucheng Chen received the bachelor's degree from the University of Electronic Science and Technology of China, Chengdu, China, in 2016. He is currently pursuing the M.S. degree in computer science with the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China. His research interests include computer vision and machine learning, especially person re-identification and vehicle re-identification based on deep neural networks.

Rui Huang received the B.Sc. degree from Peking University in 1999, the M.Eng. degree from the Chinese Academy of Sciences in 2002, and the Ph.D. degree from Rutgers University in 2008. He was a Post-Doctoral Researcher with Rutgers University, before he joined the faculty at the Huazhong University of Science and Technology, China, in 2010. From 2012 to 2016, he was a Researcher at NEC Laboratories China. He is currently an Associate Professor with The Chinese University of Hong Kong, Shenzhen. He has been involved in various research topics, including subspace analysis, deformable models, probabilistic graphical models, and their applications in computer vision, pattern recognition, and medical image analysis. He has authored more than 70 articles in related areas and has been the principal investigator of various research grants. His current research interests include computer vision and machine learning methods for surveillance and robots.



Hong Chang (Member, IEEE) received the bachelor's degree from the Hebei University of Technology, Tianjin, China, in 1998, the M.S. degree from Tianjin University, Tianjin, in 2001, and the Ph.D. degree from The Hong Kong University of Science and Technology, Hong Kong, in 2006, all in computer science. She was the Research Scientist of Xerox Research Centre Europe. She is currently a Professor with the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China. Her main research interests include algorithms and models in machine learning, and their applications in pattern recognition, computer vision, and data mining.

Chuanqi Tan (Member, IEEE) received the Ph.D. degree in computer sciences and technology from Tsinghua University in July 2019. He joined Tencent as a Researcher in 2019. Prior to this, he has worked in jike.com in 2012 and baidu.com in 2014. His research interests include computer vision, deep learning, transfer learning, brain–computer interface, and the related research topics.

Tao Xue received the B.S. degree from Northwestern Polytechnical University, Xi'an, China, in 2005, and the M.S. degree from Beihang University, Beijing, China, in 2009. Her research interests include computer vision, pattern recognition, and machine learning. She specially focuses on object detection, image segmentation, and image retrieval.

Bingpeng Ma received the B.S. degree in mechanics and the M.S. degree in mathematics from the Huazhong University of Science and Technology, in 1998 and 2003, respectively, and the Ph.D. degree in computer science from the Institute of Computing Technology, Chinese Academy of Sciences, China, in 2009. He was a Post-Doctoral Researcher with the University of Caen, France, from 2011 to 2012. He joined the School of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing, in March 2013, where he is currently an Associate Professor. His research interests include computer vision, pattern recognition, and machine learning. He especially focuses on person re-identification, face recognition, and the related research topics.