Cross-Modal Cross-Domain Moment Alignment Network for Person Search

Ya Jing,[1,3] Wei Wang,[1,3]* Liang Wang,[1,2,3] Tieniu Tan[1,2,3]
[1]Center for Research on Intelligent Perception and Computing (CRIPAC),
National Laboratory of Pattern Recognition (NLPR)
[2]Center for Excellence in Brain Science and Intelligence Technology (CEBSIT),
Institute of Automation, Chinese Academy of Sciences (CASIA)
[3]University of Chinese Academy of Sciences (UCAS)
ya.jing@cripac.ia.ac.cn, {wangwei, wangliang, tnt}@nlpr.ia.ac.cn

Abstract

Text-based person search has drawn increasing attention due to its wide applications in video surveillance. However, most of the existing models depend heavily on paired image-text data, which is very expensive to acquire. Moreover, they always face huge performance drop when directly exploiting them to new domains. To overcome this problem, we make the first attempt to adapt the model to new target domains in the absence of pairwise labels, which combines the challenges from both cross-modal (text-based) person search and cross-domain person search. Specially, we propose a moment alignment network (MAN) to solve the cross-modal cross-domain person search task in this paper. The idea is to learn three effective moment alignments including domain alignment (DA), cross-modal alignment (CA) and exemplar alignment (EA), which together can learn domain-invariant and semantic aligned cross-modal representations to improve model generalization. Extensive experiments are conducted on CUHK Person Description dataset (CUHK-PEDES) and Richly Annotated Pedestrian dataset (RAP). Experimental results show that our proposed model achieves the state-of-the-art performances on five transfer tasks.

1. Introduction

Person search is a fundamental task in video surveillance and has gained great attention in recent years due to its wide applications. Existing methods of person search mainly focus on image-based person search also called person re-identification (Person Re-ID) [36], which aims to predict whether two images from different cameras belong to the same person. Great successes have been achieved with the development of deep neural networks [9].

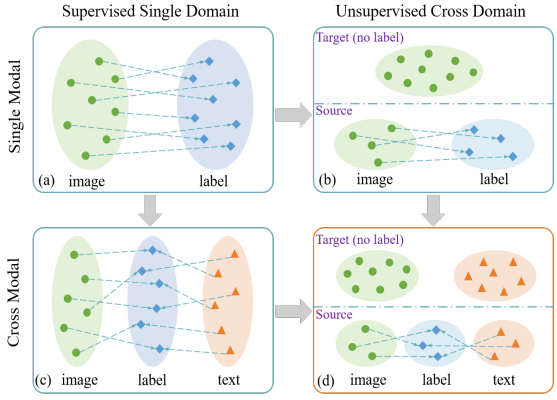

Figure 1. Our proposed (d) cross-modal cross-domain text-based person image search combines the challenges from both (b) cross-domain image-based person search and (c) cross-modal text-based person image search.

However, most of the person Re-ID models are trained in a supervised learning manner based on manually labeled pairwise datasets, which are expensive or impossible to collect in practice. This limits the applications where only unlabeled data is available. A feasible method is to transfer the learned models in labeled source domain to a new unlabeled target domain, which is called cross-domain person Re-ID [27] as shown in Fig. 1 (b). Nevertheless, due to the dataset bias, it is still difficult to generalize the model trained on labeled dataset to an unlabeled dataset. As a result, how to relieve the domain shift becomes increasingly important. To address this issue, several attempts have been proposed recently. They aim to learn a discriminative representation for target set with labeled source dataset and unlabeled target dataset. A common practice [27, 34] is to transform the source image into target domain style and thus utilizes the supervised learning.

Moreover, person Re-ID needs at least one image of

*Corresponding Author: Wei Wang

the queried person which in many cases is very difficult to obtain. Since textual descriptions are more accessible, text-based person search [15] can solve query person image missing problem, which is also a cross-modal task as shown in Fig. 1 (c). Text-based person search aims to retrieve the corresponding person images to a textual description from a large-scale person database, which faces the challenges of semantic gap between image and text. Similar to image-based person search, text-based person search still has the problem of lacking labeled pairwise data and needs domain adaptation from labeled source domain to unlabeled target domain. Furthermore, text-based person search also faces the domain shifts in images between different datasets on account of different collecting surveillance scenarios.

In this paper, we propose a new task of cross-modal cross-domain text-based person image search as shown in Fig. 1 (d), which combines the challenges from both cross-domain image-based person search and cross-modal text-based person image search. To the best of our knowledge, there is no such task proposed in previous work. Furthermore, we propose a moment alignment network (MAN) for this task, which contains three alignment modules to reduce the domain discrepancy in a complementary way: domain alignment (DA), cross-modal alignment (CA) and exemplar alignment (EA). Fig. 2 shows the architecture of MAN.

Specifically, we first learn four classifiers for source/target image/text classification. Considering the alignment and preserved semantic structure between the class-level mean of data distribution and classifier parameter introduced in [20], we take classifier parameter of a category as the mean of the data distribution of that category, also called class mean. Through measuring the similarity between sample representation in target domain and class means (classifier parameters) in source domain, we can compute pseudo labels for unlabeled target data, which are used as supervision information of the target classifiers. Particularly, we call the class means (classifier parameters) and their variance as class moments which will be used for alignments.

Next, we compute three alignments. First, to directly alleviate the domain discrepancy, we propose a domain alignment by minimizing the distance of the class moments between source and target domains. Based on the domain alignment strategy, the learned representations are constrained to be invariant across domains. Second, due to the semantic gap between person image and textual description, we propose a cross-modal alignment by minimizing the distance of the class moments between image and text in target domain. Note that the cross-modal alignment in source domain is implemented by minimizing the ranking loss function [3]. Third, considering that the representations in target domain learned from pseudo labels may be not optimally discriminative, we propose an exemplar alignment to further enhance the clustering characteristics of representations in target domain by maximizing the probability of a target exemplar belonging to its class. By modelling these three alignments jointly, we can learn domain-invariant and semantic aligned cross-modal representations.

Our proposed method is evaluated on CUHK Person Description dataset (CUHK-PEDES) [15] and Richly Annotated Pedestrian dataset (RAP) [12], and achieves the state-of-the-art results on five transfer tasks.

In summary, the main contributions are three-fold:

(1) We make the first attempt to conduct domain adaptive text-based person search, which is a challenging cross-modal cross-domain task. (2) We propose a novel cross-modal cross-domain moment alignment network, where domain alignment, cross-modal alignment, and exemplar alignment are jointly modeled to reduce the domain discrepancy and semantic gap in a complementary way. (3) The experimental results indicate that our MAN achieves the best performances. And extensive ablation studies demonstrate the effectiveness of each component in the MAN for domain adaptive text-based person search.

2. Related Work

In this section, we introduce the related works, including text-based person search, cross-domain person Re-ID and unsupervised domain adaptation.

2.1. Text-Based Person Search

Li et al. [15] propose the task of text-based person search and further employ a CNN (Convolutional Neural Network) [10]-LSTM (Long Short-Term Memory) [7] network with gated neural attention for this task. To exploit the person identification, Li et al. [14] propose an identity-aware two-stage network. PWM+ATH [1] utilizes a patch-word matching model to exploit the local similarity. Different from the above methods which all aim to learn the correspondence between image and text by attention mechanism, Dual Path [33] employs an identification loss for instance-level image-text matching. CMPM+CMPC [31] devises a cross-modal projection matching (CMPM) loss and a cross-modal projection classification (CMPC) loss for learning discriminative image-text embeddings. TIMAM [21] employs an adversarial discriminator that aims to identify whether the input is originating from the visual or textual modality to learn discriminative modality-invariant embeddings. In contrast to them, we learn the unsupervised domain adaptive text-based person search, which does not need paired image-text data in target domain.

2.2. Cross-Domain Person Re-ID

Cross-domain person Re-ID is proposed due to the expensive or impossible identity labeling, which aims to learn a generalized retrieval model by labeled source dataset and

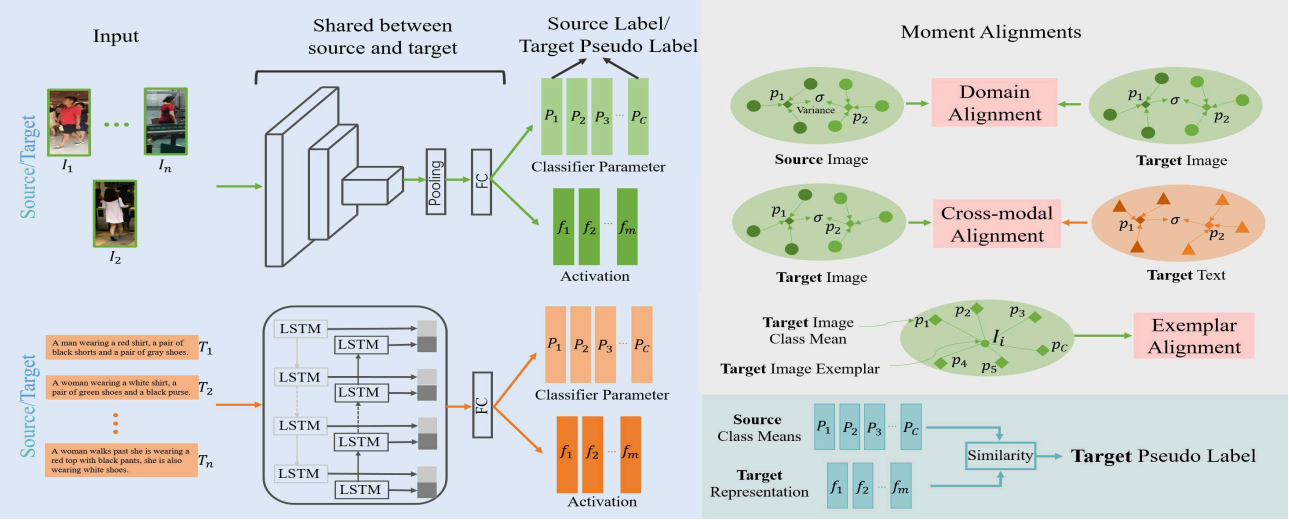

Figure 2. The architecture of our proposed cross-modal cross-domain moment alignment network (MAN). A visual CNN and a bidirectional LSTM are utilized to encode the image and text, respectively. The classifier parameter is utilized to represent the class-level mean of the data distribution, i.e., class mean. Note that the class means and their variance are called as class moments. Subsequently, three moment alignments including domain alignment, cross-modal alignment, and exemplar alignment are proposed to relieve the domain shift. Moreover, we calculate a pseudo label for each target sample to guide the learning of target classifiers. Best viewed in color.

unlabeled target dataset. To reduce the domain shift between datasets, several methods are proposed. TJ-AIDL [26] learns an attribute-semantic and identity-discriminative feature space transferrable to unseen domains. PTGAN [27] and HHL [34] both utilize Generative Adversarial Networks (GAN) [4] to transfer the source domain images into target domain style. Then the model is trained on these translated images and suitable for the target domain. ECN [35] investigates into the intra-domain invariance of the target domain. Another line of works [13, 30] focus on learning pseudo identity labels in target domain. Yu et al. [30] propose to learn a soft multilabel for each unlabeled person by the comparison with a set of known reference persons from an auxiliary dataset. Different from them, we study the domain adaptation in text-based person search, which needs to mine not only the cross-domain alignment but also the cross-modal alignment. In addition, we are probably the first to study the domain adaptive text-based person search.

2.3. Unsupervised Domain Adaptation

Domain adaptive person search is closely related to the unsupervised domain adaptation (UDA) [18, 22, 24, 19]. ADDA [24] designs a domain classifier based on adversarial learning objectives to reduce the domain gap. TPN [19] utilizes Transferrable Prototypical Networks for adaptation such that the prototypes for each class in source and target domains are close in the embedding space and the score distributions predicted by prototypes separately on source and target data are similar. However, most of UDA methods assume that the classes are shared between source and target domains, which cannot be applied to person search. The completely different identities between source and target domains make domain adaptive person search more challenging than UDA.

3. Cross-Modal Cross-Domain Moment Alignment Network

In this section, we first explain the proposed cross-modal cross-domain moment alignment network in detail. Afterwards, we introduce the learning procedure of the model.

3.1. Notations and Definitions

In domain adaptive text-based person search, we are given a labeled source dataset $X^s = \{I^s = \{I_i^s, y_i^s\}_{i=1}^{N_I^s}, T^s = \{T_j^s, y_j^s\}_{j=1}^{N_T^s}\}$, where N_I^s and N_T^s are the numbers of images I^s and texts T^s, respectively. Each sample is provided with an identity y. Additionally, we are also provided with an unlabeled target dataset $X^t = \{I^t = \{I_i^t\}_{i=1}^{N_I^t}, T^t = \{T_j^t\}_{j=1}^{N_T^t}\}$, where identity annotation is not available. It is worth noting that the target persons are completely non-overlapping with the source persons due to different surveillance scenarios. Based on this, our goal is to learn a deep neural network which can retrieve the corresponding person images by query text in target domain.

3.2. Class Moments

Given source and target data, a visual CNN and a bidirectional Long Short-Term Memory network (bi-LSTM) are utilized to encode the images and texts, respectively.

Note that the feature extraction networks are shared between source and target domains. Following it, four classifiers are learned for source/target image/text classification.

Due to the shift in data distributions across different domains, the performance drops a lot when directly exploiting the model trained on source dataset to target dataset. To measure the data distribution, an optional approach in UDA is to compute the prototypes [19] based on the samples in a mini-batch at each iteration, which results in the fact that the maximum number of classes the model can support is less than the batch size. Therefore, this approach cannot be applied to the cross-domain text-based person search proposed in this paper. Generally, there are more classes in person search than those in previous UDA tasks, which results in the fact that the samples in a mini-batch cannot contain all the classes. The categorical information in each mini-batch is insufficient and may has huge bias between the estimated distribution and the true distribution. Considering that there exists an alignment and preserved semantic structure between the class-level mean of data distribution and the classifier parameter, we propose to utilize the classifier parameter learned on source and target datasets to approximately represent the mean of data distribution in each class (also called class mean), which can eliminate the mini-batch constraints and obtain up-to-date class representations.

$$\mu^s_{I_k} = p^s_{I_k}, \mu^s_{T_k} = p^s_{T_k}, \tag{1}$$

$$\mu^t_{I_k} = p^t_{I_k}, \mu^t_{T_k} = p^t_{T_k}, \tag{2}$$

where $\mu^s_{I_k}$ and $p^s_{I_k}$ denote the class mean and classifier parameter of the k-th class in visual source domain, respectively. Based on class means, we can calculate the variances between them:

$$\sigma^s_I = \frac{1}{C}\sum_{k=1}^{C}(\mu^s_{I_k} - \frac{\sum_{k=1}^{C}\mu^s_{I_k}}{C})^2, \tag{3}$$

$$\sigma^t_I = \frac{1}{C}\sum_{k=1}^{C}(\mu^t_{I_k} - \frac{\sum_{k=1}^{C}\mu^t_{I_k}}{C})^2, \tag{4}$$

$$\sigma^t_T = \frac{1}{C}\sum_{k=1}^{C}(\mu^t_{T_k} - \frac{\sum_{k=1}^{C}\mu^t_{T_k}}{C})^2, \tag{5}$$

where C is the number of classes in source domain. The σ^s_T is not utilized because there is no domain shift between source texts and target texts. In addition, the cross-modal alignment in source domain is implemented by minimizing the ranking loss. Particularly, the class means (classifier parameters) and their variances are called class moments used in alignment. The mean between class means is not utilized since we find that it cannot improve the performance.

For unlabeled target samples, we utilize the self-labeling [11], where the pseudo label is assigned for target sample by class means in source domain:

$$y^t_{I_k} = \frac{\exp(\cos(f(I^t), \mu^s_{I_k}))}{\sum_{r=1}^{C}\exp(\cos(f(I^t), \mu^s_{I_r}))}, \tag{6}$$

$$y^t_{T_k} = \frac{\exp(\cos(f(T^t), \mu^s_{T_k}))}{\sum_{r=1}^{C}\exp(\cos(f(T^t), \mu^s_{T_r}))}, \tag{7}$$

where $y^t_{I_k}$ is the k-th entry of pseudo label for target image I^t, $\cos$ means the cosine similarity, and f indicates the activation of the final fully-connected layer (FC in Fig. 2). Note that the learned pseudo label is a soft label since different classes across domains, and is further used as supervision information of the target classifiers.

3.3. Domain Alignment

In domain adaptive text-based person search, due to the fact that the vocabularies of different datasets differ little from each other, the domain shift is mainly caused by different person appearance distributions across domains. To eliminate the image distribution variations, we propose a domain alignment network by relieving the divergence of domain statistics. The basic idea is that the moments calculated on different domains are the same if the distributions of source and target domains are identical. Different from modelling the global distribution of domain representations where crucial class-level information may be lost, we enforce the class-level domain alignment to ensure the samples with similar visual features being mapped nearby in the feature space. In other words, a fine-grained alignment of domain distributions is performed on class level.

Based on the analysis above, we propose a cross-domain consistent loss:

$$L_{CDC} = \sum_{k=1}^{C} d(\mu^s_{I_k}, \mu^t_{I_k}) + \gamma_1 d(\sigma^s_I, \sigma^t_I), \tag{8}$$

where $d(.,.)$ indicates the distance between two class moments and γ is a hyperparameter to control the importance of each distance. We use the squared Euclidean distance in our experiments:

$$d(x_1, x_2) = \| x_1 - x_2 \|^2 . \tag{9}$$

The justification is that the cluster mean yields optimal cluster representatives when a Bregman divergence (e.g. squared Euclidean distance and Mahalanobis distance) is used [23].

Through explicitly restricting the distance between domain statistics, the feature extractor is guided to learn domain-invariant representations.

3.4. Cross-Modal Alignment

The domain alignment only restricts the cross-domain visual statistics but ignores the latent semantic alignmen-

t between image and text, which is the key to this cross-modal matching task. In source domain, the pairwise ranking loss [3] is utilized to ensure the cross-modal alignment. Due to the lack of labels, the ranking loss cannot be employed in target domain. Like domain alignment, we adopt the class-level cross-modal moment alignment module to ensure the target samples with similar semantic being closer in the feature space. Based on the class moments, the cross-modal consistent loss is defined as follows:

$$L_{CMC} = \sum_{k=1}^{C} d(\mu_{I_k}^t, \mu_{T_k}^t) + \gamma_2 d(\sigma_I^t, \sigma_T^t). \quad (10)$$

Thus the learned representations can be independent from their data modalities.

3.5. Exemplar Alignment

In target domain, due to the lack of ground truth labels, the learned representations may be not optimally discriminative. Considering the fact that a target exemplar belongs to one of the target classes, we propose an exemplar alignment module by enforcing each exemplar to be close to its nearest class means to enhance the clustering characteristics of representations. We first calculate the cosine similarities between $f(I^t)$ and target class means. Then the probability that I^t belongs to class k is calculated as follows:

$$q(k|I^t) = \frac{\exp(\cos(f(I^t), \mu_{I_k}^t))}{\sum_{r=1}^{C} \exp(\cos(f(I^t), \mu_{I_r}^t))}. \quad (11)$$

The exemplar consistent loss is defined as:

$$L_{EC} = -\log(\max q(k|I^t)|_{k=1}^{C}), \quad (12)$$

which aims to maximize the probability of the class that target exemplar belongs to. In this way, we can further improve the robustness of our model.

3.6. Model training and testing

For labeled source dataset, ranking loss with the hardest negative samples [3] is utilized to ensure the cross-modal alignment by making positive pair being closer than the hardest negative pair:

$$\begin{aligned} L_r(I^s, T^s) = & \, max(\alpha - S(I^s, T^s) + S(I^s, T_{\widehat{h}}^s), 0) \\ & + max(\alpha - S(I^s, T^s) + S(I_{\widehat{h}}^s, T^s), 0), \end{aligned} \quad (13)$$

where $T_{\widehat{h}}^s$ is the hardest text sample in a mini-batch for the source image I^s, S denotes the similarity score between image and text, i.e., cosine score, and α is a margin.

Besides ranking loss, the identification loss is also adopted for the identity-level matching. The image and text identification losses L_{I^s} and L_{T^s} are defined as follows:

$$L_{I^s} = -\frac{1}{n_I^s} \sum_{i=1}^{n_I^s} y_i^s log(softmax(W_{idi} g(I_i^s))), \quad (14)$$

$$L_{T^s} = -\frac{1}{n_T^s} \sum_{i=1}^{n_T^s} y_i^s log(softmax(W_{idt} g(T_i^s))), \quad (15)$$

where n_I^s is the number of source images in a training batch, g denotes the activation of classifier, and W_{idi} is the transformation matrix to categorize the visual representations.

Then the total source loss is defined as:

$$L_S = L_r + \beta(L_I + L_T), \quad (16)$$

where β aims to control the relative importance of each loss function.

By combining the losses defined above, the final objective of our MAN is formulated as:

$$L = L_S + \lambda_1 L_{CDC} + \lambda_2 L_{CMC} + \lambda_3 L_{EC}. \quad (17)$$

Based on this objective, we can obtain two deep embedding networks (CNN and bi-LSTM), where the learned cross-modal representations are domain-invariant and semantic aligned.

The training procedure is split into two stages. At the first stage, only the L_S is adopted and the model is trained in source domain. At the second stage, we utilize the final objective L to train our model, where both source data and target data are used. This procedure not only ensures the accuracy in source domain, but also transfers the knowledge learned in source domain to target domain.

During testing, we calculate the visual and textual representations via CNN and bi-LSTM, respectively. Then the similarity scores between them are ranked to retrieve the corresponding person images based on the query text.

4. Experiments

We conduct extensive evaluations of MAN on CUHK-PEDES dataset [15] and RAP dataset [12]. Performance comparisons with the state-of-the-art methods as well as the ablation studies are presented.

The CUHK-PEDES dataset [15] is currently the only dataset for text-based person search, where the images are collected from five different existing person re-identification datasets, CUHK03 [17], Market-1501 [32], SSM [28], VIPER [5], and CUHK01 [16]. Since these five datasets are collected from different surveillance scenarios, there are domain shifts between them.

To further verify the effectiveness of our MAN, we conduct the experiments on RAP dataset [12], which is collected from 25 cameras in an indoor shopping mall. For distant cameras, due to the large changes of viewpoints, backgrounds, cloth appearances and lighting conditions, there

Figure 3. Example images from CUHK-PEDES and RAP datasets. Best viewed by zooming in.

exists distribution bias between the images captured by different cameras. We choose two sets of images (RAP-1 and RAP-2) to perform domain/view adaptation, where the camera viewpoints of these two image sets are very different. The domain/view gap between these two sets has been proved by the performance gap between supervised learning (SL) in target domain and transfer learning from labeled source domain to unlabeled target domain (SO) in Table 2. Note that the two sets collected from different cameras have different pedestrian identities in our setting to perform domain adaptive person search.

4.1. Datasets and Metrics

CUHK-PEDES. The CUHK-PEDES dataset contains 40,206 images and 80,440 textual descriptions of 13,003 identities. To learn the domain adaptation in text-based person search, we regard CUHK-PEDES as five independent datasets, where each dataset is considered as a domain. Note that we choose SSM (S) as the source dataset and consider four transfer tasks S→C03 (CUHK03), S→M (Market-1501), S→V (VIPER) and S→C01 (CUHK01).

RAP. We choose two sets of images from different cameras to perform domain adaptation, where images from 10 cameras (i.e., CAM31, CAM30, CAM29, CAM28, CAM27, CAM25, CAM22, CAM21, CAM20, and CAM19) are regarded as the source dataset (RAP-1) and images from other 5 cameras (i.e., CAM01, CAM06, CAM09, CAM10 and CAM11) are regarded as the target dataset (RAP-2). As a result, there are 12,985 and 3,084 images in RAP-1 and RAP-2, respectively. To perform text-based person search, we choose 104 attributes out of 152 except age, customer, employee, viewpoint, occlusion and position. Moreover, we concatenate all selected attributes about an image into a sentence to describe this image.

Several example images are shown in Fig. 3 to illustrate the domain gap.

Metrics. We choose top-1, top-5 and top-10 accuracies to evaluate the performance of text-based person search. Specifically, given a query text, all test images are ranked by the similarities with the text. If the corresponding images are within the top-k images, we regard it as a successful search.

4.2. Implementation Details

We resize the input image to 384 $\times$ 128 and utilize ResNet-50 [6] to encode it. For the textual representation, we build a vocabulary by collecting all the words in sentences. Then the word is embedded to a 300-dimensional vector and entered into a 1024-dimensional bi-LSTM. In addition, the dimension of feature space is set to 1024.

At the first training stage, we first fix ResNet-50 and train the other parts of the model with learning rate lr = $1e^{-3}$. Then we train the whole model with learning rate lr = $2e^{-4}$. At the second training stage, we directly train the whole model with learning rate $2e^{-4}$. The Adam optimizer [8] is employed for optimization. The hyperparameters γ, β and λ are empirically set to 1. Moreover, the batch size for each domain and margin are set to 128 and 0.2, respectively.

Specially, for the S→C03 and RAP-1→RAP-2 transfer tasks, we train the model at the first training stage for 400, and 300 epochs, respectively. At the second training stage, we train the model for 80 and 60 epochs, respectively. We take about 7 hours to train the model.

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

Compared Methods. To verify the merit of our MAN, we compare with the following representative approaches in various experimental settings: (1) The labeled target data is utilized to train the model, e.g., Supervised Learning (SL). (2) Only the labeled source data is utilized to train the model, e.g., Source Only (SO), CMPM+CMPC [31], Attribute query (AQ) and Adv-attReID [29]. CMPM+CMPC is a traditional text-based person search method with high performance, which utilizes a cross-modal projection matching (CMPM) loss and a cross-modal projection classification (CMPC) loss to learn discriminative image-text representations. Attribute query (AQ) retrieves the corresponding person images based on the attributes of the person. Adv-attReID is a traditional attributed-based person search method with high performance, which imposes an attribute-guided attention mechanism for images and a semantic consistent adversary strategy for attributes. (3) The labeled source data and unlabeled target data are both utilized to train the model, e.g., SPGAN [2], ADDA [24] and ECN [35]. SPGAN translates the labeled images from source domain to target domain in an unsupervised manner, and then trains Re-ID model with the translated images by supervised method. ADDA designs a domain classifier based on adversarial learning objectives. ECN investigates into the intra-domain invariance of the target domain and proposes three types of invariance, i.e., exemplar-invariance, camera-invariance and neighborhood-invariance.

It is noteworthy that since we are probably the first to conduct domain adaptive text-based person search, there is no method reporting the performance on CUHK-PEDES and RAP. As a result, we perform the experiments based

Table 1. Methods comparison when tested on C03, M, V and C01. Supervised Learning (SL): Model trained with labeled target data utilizing L_S. Source Only (SO): Model trained with only labeled source data utilizing L_S. Top-1, top-5 and top-10 accuracies (%) are reported. The best performance is **bold**.

	S→C03			S→M			S→V			S→C01		
Method	Top-1	Top-5	Top-10	Top-1	Top-5	Top-10	Top-1	Top-5	Top-10	Top-1	Top-5	Top-10
SL	54.7	79.4	87.1	68.5	90.1	95.2	67.8	91.8	96.4	59.3	81.5	88.4
SO	41.5	65.8	78.1	62.7	84.5	91.2	59.1	83.2	93.4	44.3	69.5	80.8
CMPM+CMPC[31]	42.3	69.2	79.9	63.4	85.1	92.2	57.8	84.7	92.4	44.8	70.9	81.7
SPGAN[2]	44.7	72.5	82.6	63.3	85.3	92.4	60.7	85.7	93.3	45.3	71.2	83.1
ADDA[24]	45.1	72.8	82.5	63.9	85.7	92.3	61.4	86.0	93.2	45.7	71.6	83.0
ECN[35]	45.8	73.2	82.8	64.3	86.1	93.4	62.5	86.4	93.7	46.6	72.1	83.2
MAN(ours)	**48.5**	**74.8**	**84.3**	**65.1**	**87.4**	**94.6**	**64.2**	**87.2**	**94.3**	**48.2**	**73.2**	**83.6**

Table 2. Performances when tested on RAP-1→RAP-2 and RAP-2→RAP-1 transfer tasks.

	RAP-1→RAP-2			RAP-2→RAP-1		
Method	Top-1	Top-5	Top-10	Top-1	Top-5	Top-10
SL	46.5	73.2	81.5	33.4	58.7	69.2
SO	42.3	68.9	78.4	19.1	37.6	47.0

Table 3. Methods comparison when tested on RAP dataset (RAP-1→RAP-2).

Method	Top-1	Top-5	Top-10
SL	46.5	73.2	81.5
SO	42.3	68.9	78.4
AQ	32.1	60.6	70.3
Adv-attReID [29]	25.3	47.2	56.9
SPGAN[2]	42.6	68.4	79.0
ADDA[24]	43.3	69.4	79.4
ECN[35]	43.5	69.1	79.6
MAN(ours)	**44.2**	**69.8**	**80.1**

on the above methods and modify them to fit our task.

Results on the CUHK-PEDES dataset. We compare our proposed MAN with the state-of-the-art methods on four transfer tasks from CUHK-PEDES dataset. Table 1 shows the results, where S→C03 means that S is the source dataset and C03 is the target dataset. Note that Source Only and CMPM+CMPC [31] are two methods without domain adaptation. Compared with the strong baseline (CMPM+CMPC) which achieves high performances in text-based person search, our MAN still improves the experimental results. Particularly, the performances of Source Only can be regarded as a lower bound. Incorporating the domain adaptation can obviously outperform the Source Only, which demonstrates the advantage of relieving the domain discrepancy. Moreover, our proposed MAN achieves the best performances of all domain adaptation methods on every transfer task, which proves the effectiveness of our MAN. Specifically, compared with the methods (SPGAN and ADDA) which both utilize Generative Adversarial Networks to eliminate the domain shift, our MAN significantly outperforms them. The improved performances illustrate the superiorities of class-level moment alignment network in restricting the cross-domain statistics, where the features with similar semantic are constrained being mapped nearby and thus are learned more domain-invariant. Furthermore, compared with ECN which exploits intra-domain alignment, our MAN also achieves better performances, which basically indicates the advantage of jointly modelling the inter-domain, intra-domain and inter-modal moment alignments in cross-modal cross-domain person search. Note that although our MAN achieves the best performances compared with existing domain adaptation methods, which is inferior to the supervised cross-modal matching methods (Supervised Learning). This suggests that there is still room for improvement in the methods of domain adaptation.

Results on the RAP dataset. To prove the domain gap between RAP-1 and RAP-2, we perform extensive experiments as shown in Table 2. The performance gap between SL (Supervised Learning) and SO (Source Only) proves that there is indeed domain shift between the two datasets.

Table 3 shows the comparison results on RAP-1 → RAP-2 transfer task. According to the performance gap between SL (Supervised Learning) and SO (Source Only), RAP-1 → RAP-2 is a more balanced transfer task than transfer tasks in CUHK-PEDES dataset. Based on this, our model still outperforms the existing domain adaptation methods but with less improvement compared to the results on CUHK-PEDES dataset. This further demonstrates that the proposed MAN can learn domain-invariant representations. In addition, we report the results of attribute query method (AQ), which is set similar to SO except the query type. The reduced performances mean that the relationship between attributes is significant to the learning of textual representation. Moreover, our MAN also achieves better performances compared with traditional attributed-based person search method Adv-attReID [29].

4.4. Ablation Study

To systematically investigate the effectiveness of each component in MAN, we perform a set of ablation studies

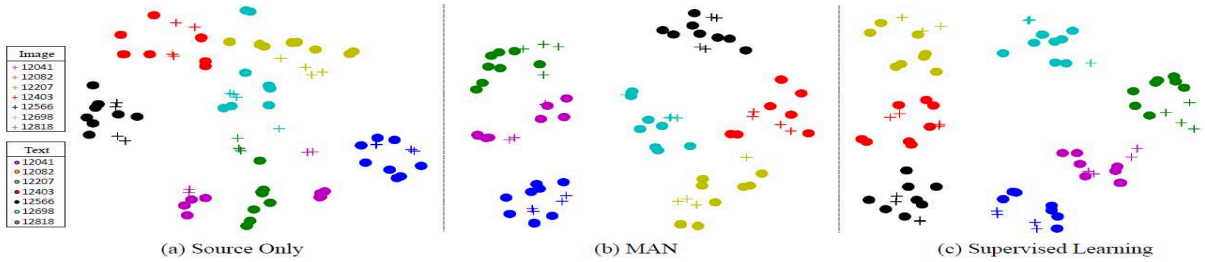

Figure 4. The t-SNE visualization of image and text features learned by Source Only, MAN and Supervised Learning on S→C03 transfer task.

on S→C03 transfer task. The results are illustrated in Table 4. We first investigate the importance of domain alignment (DA) by performing experiments on MAN (wo DA), where DA is completely missing. It can be seen that the top-1 accuracy declines 2.2%, which proves the effectiveness of domain alignment in benefitting the domain-invariant representations learning. When comparing MAN with MAN (wo CA) where the cross-modal alignment is missing, we can observe that MAN significantly outperforms the MAN (wo CA) in top-1 accuracy (48.5% / 46.8%). The improved performances indicate that cross-modal alignment can help our model learn more discriminative cross-modal representations by modelling the constraints between image and text, and thus benefits the performances. To investigate the importance of exemplar alignment, we perform experiments on MAN (wo EA). The results indicate that it is appropriate to alleviate domain discrepancy by enhancing the clustering characteristics of representations in target domain. In addition, we report the results of discarding the σ in moment alignment. The performance is inferior to the MAN by 0.5%, which illustrates the superiorities of σ in relieving the divergence of data distributions.

In summary, we can observe that all the components are designed reasonably and the performance degrades when any of these components is removed.

4.5. Visual Results

To figure out whether our model can learn the cross-modal features in target domain well, we utilize t-SNE [25] to visualize the feature distribution learned by Source Only, MAN and Supervised Learning on S→C03 transfer task (randomly select 7 classes). As shown in Fig. 4 (c), we can see that the learned image-text features are distributed along radial spokes, where the corresponding visual and textual features lie in the same direction due to the usage of cosine similarity between image and text. From the comparison between Fig. 4 (a) and Fig. 4 (b), we can clearly observe that the features learned by Source Only can not be discriminated very well, where some features from different classes are mixed up in the feature space. In contrast, the features learned by our MAN are more discriminative and dispersed for different classes. This illustrates that our model can enlarge the inter-class dispersion in target domain by three moment alignments.

Table 4. Ablation analysis of different components in the proposed MAN on S→C03 transfer task. The wo means without.

Method	Top-1	Top-5	Top-10
SO	41.5	65.8	78.1
MAN(wo DA)	46.3	73.2	82.7
MAN(wo CA)	46.8	73.4	83.0
MAN(wo EA)	47.3	73.8	83.4
MAN(wo σ)	48.0	74.1	83.7
MAN	**48.5**	**74.8**	**84.3**

5. Conclusion

In this work, we propose a novel moment alignment network for domain adaptive text-based person search. To the best of our knowledge, this is the first attempt to investigate this problem. To achieve the goal, we propose three effective moment alignments including domain alignment, cross-modal alignment, and exemplar alignment in target domain. These three alignment mechanisms are complementary to each other to learn domain-invariant and semantic aligned cross-modal representations. We perform extensive experiments on five transfer tasks from CUHK-PEDES and RAP datasets and demonstrate the effectiveness of our model by significant performance improvements.

6. Acknowledgments

This work is jointly supported by National Key Research and Development Program of China (2016YFB1001000), National Natural Science Foundation of China (61420106015, 61572504, 61721004), Shandong Provincial Key Research and Development Program (Major Scientific and Technological Innovation Project) (NO.2019JZZY010119), and Science and Technology Project of SGCC Research on feature recognition and prediction of typical ice and wind disaster for transmission lines based on small sample machine learning method.

References

[1] Tianlang Chen, Chenliang Xu, and Jiebo Luo. Improving text-based person search by spatial matching and adaptive threshold. In *WACV*, 2018.

[2] Weijian Deng, Liang Zheng, Qixiang Ye, Guoliang Kang, Yi Yang, and Jianbin Jiao. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In *CVPR*, 2018.

[3] Fartash Faghri, David J. Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. In *BMVC*, 2018.

[4] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014.

[5] Douglas Gray, Shane Brennan, and Hai Tao. Evaluating appearance models for recognition, reacquisition, and tracking. In *IEEE International Workshop on Performance Evaluation for Tracking and Surveillance*, pages 1–7, 2007.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[7] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 1997.

[8] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *Computer Science*, 2014.

[9] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NeurIPS*, 2012.

[10] Yann LeCun, Léon Bottou, Yoshua Bengio, Patrick Haffner, et al. Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*, 1998.

[11] Dong-Hyun Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on Challenges in Representation Learning, ICML*, 2013.

[12] Dangwei Li, Zhang Zhang, Xiaotang Chen, and Kaiqi Huang. A richly annotated pedestrian dataset for person retrieval in real surveillance scenarios. *IEEE transactions on image processing*, pages 1575–1590, 2018.

[13] Minxian Li, Xiatian Zhu, and Shaogang Gong. Unsupervised person re-identification by deep learning tracklet association. In *ECCV*, 2018.

[14] Shuang Li, Tong Xiao, Hongsheng Li, Wei Yang, and Xiaogang Wang. Identity-aware textual-visual matching with latent co-attention. In *ICCV*, 2017.

[15] Shuang Li, Tong Xiao, Hongsheng Li, Bolei Zhou, Dayu Yue, and Xiaogang Wang. Person search with natural language description. In *CVPR*, 2017.

[16] Wei Li, Rui Zhao, and Xiaogang Wang. Human reidentification with transferred metric learning. In *ACCV*, pages 31–44, 2012.

[17] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*, pages 152–159, 2014.

[18] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I Jordan. Learning transferable features with deep adaptation networks. In *ICML*, 2015.

[19] Yingwei Pan, Ting Yao, Yehao Li, Yu Wang, Chong-Wah Ngo, and Tao Mei. Transferrable prototypical networks for unsupervised domain adaptation. In *CVPR*, 2019.

[20] Siyuan Qiao, Chenxi Liu, Wei Shen, and Alan L Yuille. Few-shot image recognition by predicting parameters from activations. In *CVPR*, 2018.

[21] Nikolaos Sarafianos, Xiang Xu, and Ioannis A Kakadiaris. Adversarial representation learning for text-to-image matching. In *ICCV*, 2019.

[22] Rui Shu, Hung H Bui, Hirokazu Narui, and Stefano Ermon. A dirt-t approach to unsupervised domain adaptation. In *ICLR*, 2018.

[23] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *NeurIPS*, 2017.

[24] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *CVPR*, 2017.

[25] Laurens Van Der Maaten. Accelerating t-sne using tree-based algorithms. *The Journal of Machine Learning Research*, pages 3221–3245, 2014.

[26] Jingya Wang, Xiatian Zhu, Shaogang Gong, and Wei Li. Transferable joint attribute-identity deep learning for unsupervised person re-identification. In *CVPR*, 2018.

[27] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *CVPR*, 2018.

[28] Tong Xiao, Shuang Li, Bochao Wang, Liang Lin, and Xiaogang Wang. End-to-end deep learning for person search. *arXiv preprint arXiv:1604.01850*, 2016.

[29] Zhou Yin, Wei-Shi Zheng, Ancong Wu, Hong-Xing Yu, Hai Wan, Xiaowei Guo, Feiyue Huang, and Jianhuang Lai. Adversarial attribute-image person re-identification. In *IJCAI*, 2018.

[30] Hong-Xing Yu, Wei-Shi Zheng, Ancong Wu, Xiaowei Guo, Shaogang Gong, and Jian-Huang Lai. Unsupervised person re-identification by soft multilabel learning. In *CVPR*, 2019.

[31] Ying Zhang and Huchuan Lu. Deep cross-modal projection learning for image-text matching. In *ECCV*, 2018.

[32] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jiahao Bu, and Qi Tian. Person re-identification meets image search. *arXiv preprint arXiv:1502.02171*, 2015.

[33] Zhedong Zheng, Liang Zheng, Michael Garrett, Yi Yang, and Yi-Dong Shen. Dual-path convolutional image-text embedding. In *arXiv preprint arXiv:1711.05535*, 2017.

[34] Zhun Zhong, Liang Zheng, Shaozi Li, and Yi Yang. Generalizing a person retrieval model hetero-and homogeneously. In *ECCV*, 2018.

[35] Zhun Zhong, Liang Zheng, Zhiming Luo, Shaozi Li, and Yi Yang. Invariance matters: Exemplar memory for domain adaptive person re-identification. In *CVPR*, 2019.

[36] Qin Zhou, Heng Fan, Shibao Zheng, Hang Su, Xinzhe Li, Shuang Wu, and Haibin Ling. Graph correspondence transfer for person re-identification. In *AAAI*, 2018.