AMEN: Adversarial Multi-Space Embedding Network for Text-based Person Re-identification *

Anonymous submission

Paper ID: 51

Abstract. Many of the existing methods manage to extract modality-invariant features from both modalities by learning a joint latent space, in which the visual/textual feature vector can be better aligned. However, though misaligned information can be removed when mapping features from two high-dimensional spaces into a common space, discriminative clues as well may be lost. To this end, merely embedding features into a joint latent space may not be sufficient to give satisfactory performance, and the utilization of both visual and textual high-level spaces deserves more in-depth exploration. In this paper, we proposed a novel Adversarial Multi-space Embedding Network (AMEN) to learn and match embeddings in multiple spaces. Following a encoder-decoder manner, the inter-modal reconstruction paradigm works in concert with the intra-modal reconstruction paradigm to properly embedding a feature into the opposite modality space while learning a strong common space. A consistency constraint is adapted to ensure that the learned visual and textual spaces are trained jointly and work consistently. To enhance both the common space learning and feature reconstruction, the adversarial mechanism is utilized. Our proposed AMEN is evaluated on CUHK-PEDES, which is currently only accessible dataset for text-base person re-identification task. Extensive experimental results demonstrate that AMEN outperforms previous methods and achieves the state-of-the-art performance.

Keywords: Multi-space Embedding · Text-based Person Re-identification · Cross-modal Retrieval · Adversarial Learning.

1 Introduction

Person re-identification aims to search for the corresponding pedestrian image according to a given query, which can be an image, a video, a set of attributes or a text description. Considering that in most of the scenes, text descriptions of a target person are much more accessible than any other type of queries, text-based person re-identification [11, 10, 17, 7, 16, 23, 13] has drawn remarkable attention. The task of text-based person re-identification involves handling multimedia data, which can be regarded as a subtask of cross-modal retrieval [25, 8,

* Student Paper

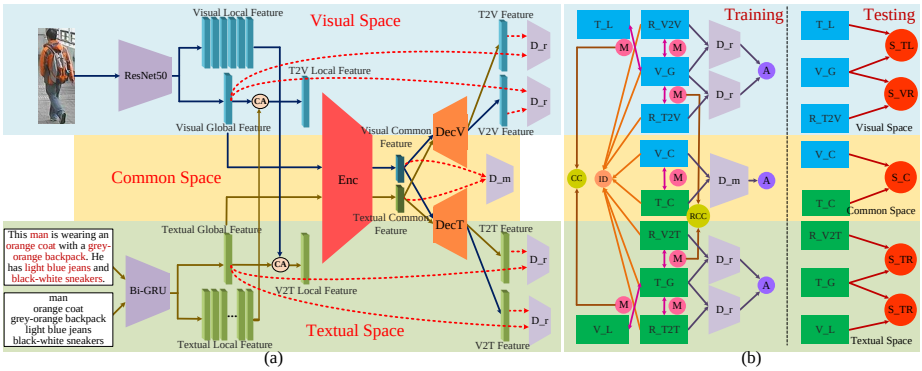

Fig. 1. The overall architecture of our proposed Adversarial Multi-space Embedding Network (AMEN) is illustrated in (a). Following an encoder-decoder manner, three spaces including a common space, a visual space and a textual space are learned to complement each other. Inter-modal (which forms $V2T$ and $T2V$ features) and intra-modal (which forms $V2V$ and $T2T$ features) reconstruction paradigms are utilized to realize the cross modal embedding while learning a stronger common space. Losses for training and similarities for testing AMEN are shown in (b). ID, M, A, CC and RCC denote the proposed ID loss, triplet ranking loss, adversarial loss, consistency constraint and reconstruction consistency constraint, respectively. S_C, S_{TL}, S_{VL}, S_{VR} and S_{TR} are the five corresponding cosine similarities employed to test AMEN.

15, 14, 20, 9]. Instead of containing various categories of objects in an image, however, each image cared by text-based person re-identification task only contains one pedestrian while its corresponding text description offers more clues. This particularity of the text-based person re-identification task causes that many previous methods proposed on common cross-modal retrieval benchmarks (e.g. MSCOCO [12] and Flickr30k [18]) generalize poorly on it.

The main challenge of text-based person re-identification is to effectively extract and match feature vectors from both visual and textual modalities. Many of the existing methods [16, 13] manage to extract modality-invariant features from both modalities by learning a joint latent space, in which the visual/textual feature vector can be better aligned. Intuitively, bringing feature vectors into a specific space can be conducive to the following matching process, hence striving to learn a stronger joint latent space makes sense. However, though misaligned information can be removed when mapping features from two high-dimensional spaces into a common space, discriminative clues as well may be lost. To this end, merely embedding features into a joint latent space may not be sufficient to give satisfactory performance, and the utilization of both visual and textual high-level spaces deserves more in-depth exploration.

In this paper, we proposed a novel Adversarial Multi-space Embedding Network (AMEN) (shown in Fig.1), which follows an encoder-decoder manner to learn and match embeddings in multiple spaces, which includes a common space, a visual space and a textual space. AMEN first extracts global and fine-grained

local features from both modalities, and then the global features are mapped into a latent common space with a shared encoder. For the purpose of properly embedding one certain feature into the opposite modality space while learning a strong common space, AMEN contains two different ways of reconstruction, namely inter-modal reconstruction and intra-modal reconstruction, which play different roles. The inter-modal reconstruction paradigm aims to embed the common features encoded from one modality space into the opposite one, enabling the features to be matched in both high-dimensional spaces. In contrast, the intra-modal reconstruction paradigm reconstructs the common feature back to the original modality. By minimizing the differentiate between the original and reconstructed features, a stronger common space can be learned. To adequately exploit fine-grained clues, a cross-modal attention (CA) mechanism [17, 23] is utilized to match a local feature matrix from one modality space with the global feature in the other. In the mean time, when performing visual-to-textual (V2T) and textual-to-visual (T2V) embeddings simultaneously, it is crucial that the two high-level spaces are learned consistently and jointly. Thus, we introduce a consistency constraint into the training process of AMEN to avoid the situation where the visual and textual spaces develop and work independently, or even oppositely. Moreover, we utilize adversarial mechanism to enhance the performance of AMEN. A modality discriminator is used to determine whether a feature in the common space is encoded from the visual or textual modality. Meanwhile, a reconstruction discriminator is proposed to distinguish reconstructed features from original ones. As long as AMEN deceives the discriminators successfully, much more discriminative feature vectors can be extracted and generated.

The main contributions can be summarized as fourfold: (1) We proposed a novel Adversarial Multi-space Embedding Network (AMEN) to learn and match embeddings in multiply spaces. Following a encoder-decoder manner, the inter-modal reconstruction paradigm works in concert with the intra-modal one to properly embedding a feature into the opposite modality space while learning a strong common space. (2) We adapt a consistency constraint to ensure that the learned visual and textual spaces are trained jointly and work consistently. (3) We utilize adversarial mechanism to enhance both the common space learning and feature reconstruction. (4) We evaluate our proposed AMEN on CUHK-PEDES [11], which is currently only accessible dataset for the text-base person re-identification task. Extensive experimental results demonstrate that AMEN outperforms previous methods and achieves the state-of-the-art performance on CUHK-PEDES.

2 Related Works

2.1 Person re-identification

Person re-identification has drawn increasing attention in both academical and industrial fields. This technology addresses the problem of matching pedestrian images across disjoint cameras. The key challenges lie in the large intra-class and small inter-class variation caused by different views, poses, illuminations, and

occlusions. Existing methods can be grouped into handed-crafted descriptors, metric learning methods and deep learning methods, and deep learning methods generally plays a major role in current state-of-the-art works. Yi et al. [26] firstly proposed deep learning methods to match people with the same identification. Hou et al. [6] proposed an Interaction-and-Aggregation (IA) Block, which consists of a Spatial Interaction-and-Aggregation (SIA) Module and a Channel Interaction-and-Aggregation (CIA) Module to strengthen the representation capability of the deep neural network. Xia et al. [24] proposed the Second-order Non-local Attention (SONA) Module to learn local/non-local information in a more end-to-end way. Sun et al. [21] proposed a visibility-aware part model to significantly improve the learned representation and the achieving accuracy by considering a few parts of the Re-ID scenes combined with the self-supervising model of some feature observations to perceive the visibility of the region.

2.2 Text-based person re-identification

Text-based person re-identification searches for the corresponding pedestrian image according to a given text query. This task is first put forward by Li et al. [11] and they further take an LSTM to handle the input image and text. An efficient patch-word matching model [3] is proposed to capture the local similarity between image and text. Jing et al. [7] utilize pose information as soft attention to localize the discriminative regions. Niu et al. [17] propose a Multi-granularity Image-text Alignments (MIA) model exploit the combination of multiple granularities. Nikolaos et al. [16] propose a Text-Image Modality Adversarial Matching approach (TIMAM) to learn modality-invariant feature representation by means of adversarial and cross-modal matching objectives. Besides that, in order to better extract word embeddings, they employ the pre-trained publicly-available language model BERT. Wang et al. [23] proposed an IMG-Net model to incorporate inner-modal self-attention and cross-modal hard-region attention with the fine-grained model for extracting the multi-granular semantic information. Liu et al. [13] generate fine-grained structured representations from images and texts of pedestrians with an A-GANet model to exploit semantic scene graphs.

3 Methodology

In this section, we describe the proposed Adversarial Multi-space Embedding Network (AMEN) in detail (shown in Fig.1). First, we explain how the local and global visual/textual features are extracted in Section 3.1. Then we introduce the two reconstruction paradigms which are adopted to reconstruct features back into high-level spaces in Section 3.2. The proposed loss functions and training strategy are detailed in Section 3.3.

3.1 Feature extraction

A ResNet-50 [5] backbone pretrained on ImageNet is utilized to extract global/local visual features. Given an image I, the global feature $V_G \in \mathbb{R}^P$ is obtained by

down-scaling the output before the last pooling layer of ResNet-50 to a vector $\in \mathbb{R}^{1\times1\times2048}$ with an average pooling layer and then passing it through a group normalization (GN) layer followed by a fully-connected (FC) layer. The same output is first horizontally k-partitioned by pooling it to $k \times 1 \times 2048$, and then the local strips are separately passed through a GN and two FCs with a ReLU layer between them to form k P-dim vectors, which are finally concatenated to obtained the local visual feature matrix $M^V \in \mathbb{R}^{k\times P}$. To obtain global/local textual features, a whole sentence and the n phrases extracted from it are used as text materials. We employ a bi-directional GRU (bi-GRU) to handle the text materials, whose last hidden states of the forward and backward GRUs are concatenated to form both the processed global and local P-dim feature vectors. The P-dim vector got from a whole sentence is passed through a GN followed by an FC to give the global textual feature $T_G \in \mathbb{R}^P$ With each certain input phrase, the corresponding output P-dim vector is processed consecutively by a GN and two FCs with a ReLU layer between them and then concatenated with each other to form the local textual feature matrix $M^T \in \mathbb{R}^{n\times P}$.

To adequately utilize the fine-grained local information, we adopt a cross-modal attention (CA) mechanism to covert each local feature matrix to the V2T local feature V_L or T2V local feature T_L according to the global feature in the opposite modality:

$$\alpha_i^X = \frac{exp(S(M_i^X, Y_G))}{\sum_{j=1}^{q} exp(S(M_j^X, Y_G))}, \tag{1}$$

$$X_L = \sum_{\alpha_i^X > \frac{1}{q}} \alpha_i^X \cdot M_i^X, \tag{2}$$

where (X, Y) can be (V, T) or (T, V) which denotes the corresponding pair of visual and textual features and q can be k or n denoting the number of local strips or phrases. S$(\cdot, \cdot)$ is the cosine similarity. The visual and textual global features are mapped into a latent common space with a shared encoder, which is implemented as a two-layered perceptron to encode the features from visual and textual spaces to B-dim common features V_C and T_C, respectively.

3.2 Feature reconstruction

AMEN proposes two different reconstruction paradigms, namely inter-modal and intra-modal reconstruction, aiming to properly embed one certain feature into the opposite modality space while learning a strong common space. To do this, we employ a visual decoder and a textual decoder which decode the common features into the visual and textual spaces respectively. The inter-modal reconstruction paradigm embeds the common features encoded from one modality space into the opposite space, which gives the P-dim V2T feature R_{V2T} and T2V feature R_{T2V}, enabling the features to be matched in both high-dimensional spaces. In contrast, the intra-modal reconstruction paradigm reconstructs the

common feature back into the original modality space to get V2V feature R_{V2V} and T2T feature R_{T2T}. By minimizing the differentiate between the original and reconstructed features (which will be detailed in Section 3.3), a stronger common space can be learned.

3.3 Loss functions and training strategy

Losses for training and similarities for testing AMEN are illustrated in Fig.1(b). The complete training process includes 3 stages.

Stage-1 We first fix the parameters of the visual backbone and train the left parts of AMEN with the identification (ID) loss

$$L_{id}(X) = -log(softmax(W_{id} \times GN(X)) \tag{3}$$

to cluster person images into groups according to their identification, where $W_{id} \in \mathbb{R}^{Q \times P}$ is a shared transformation matrix implemented as a FC layer without bias and Q is the number of different people in the training set.

As global features can provide more complete information for clustering, only V_G and T_G are utilized here:

$$L_{ID} = L_{id}(V_G) + L_{id}(T_G). \tag{4}$$

And the entire loss in Stage-1 is

$$L_{Stage1} = L_{ID}. \tag{5}$$

Stage-2 In this stage all the parameters of AMEN without the two decoders are fine-tuned together. Besides ID loss, three more loss functions are adopted here.

First, a triplet ranking loss

$$L_{rk}(X_1, X_2) = \sum_{\widehat{X_2}} max\{\alpha - S(X_1, X_2) + S(X_1, \widehat{X_2}), 0\} + \sum_{\widehat{X_1}} max\{\alpha - S(X_1, X_2) + S(\widehat{X_1}, X_2), 0\} \tag{6}$$

is utilized to more accurately constrain the matched pairs to be closer than the mismatched pairs with a margin α, where $(X_1, \widehat{X_2})$ or $(\widehat{X_1}, X_2)$ denotes a mismatched pair. Instead of using the furthest positive and closest negative sampled pairs, we adopt the sum of all pairs within each mini-batch when computing the loss following [4]. The proposed matching loss in Stage-2 is

$$L_M = L_{rk}(V_C, T_C) + L_{rk}(V_G, T_L) + L_{rk}(V_L, T_G). \tag{7}$$

When transforming a local feature matrix into the opposite modality via the cross-modal attention mechanism, a consistency constraint loss

$$L_{CC} = MSE(S(V_G, T_L), S(V_L, T_G)) \tag{8}$$

is proposed to ensure that the two high-level spaces are learned consistently and jointly, where MSE denotes the Mean Square Error.

To learn modality-invariant features in the common space, we employ a modality discriminator D_m to predict which modality the input feature comes from. An adversarial loss

$$L_{ADM} = \mathop{\mathbb{E}}_{V_C^i \sim V_C}[logD_m(V_C^i)] + \mathop{\mathbb{E}}_{T_C^i \sim T_C}[1 - logD_m(T_C^i)] \tag{9}$$

is adopted to optimize D_m and the loss for training AMEN is

$$L_{AM} = -L_{ADM}. \tag{10}$$

After being able to successfully deceive the D_m, much more discriminative modality-invariant features can be extracted by AMEN. The complete loss in this stage is

$$L_{Stage2} = L_{Stage1} + L_M + L_{CC} + L_{AM}. \tag{11}$$

Stage-3 Then the two reconstruction paradigms are added to train AMEN. Instead of being superficially look-alike with the target feature, the reconstructed feature ought to be discriminative for a proper matching. Thus, rather than utilizing the traditional Euclidean Distance to guide the reconstruction, the reconstruction matching loss

$$L_{RM} = L_{rk}(R_{V2V}, V_G) + L_{rk}(R_{T2T}, T_G) + L_{rk}(R_{T2V}, V_G) + L_{rk}(R_{V2T}, T_G) \tag{12}$$

is adopted. Besides, a reconstruction ID loss

$$L_{RID} = L_{id}(R_{V2V}) + L_{id}(R_{T2T}) + L_{id}(R_{T2V}) + L_{id}(R_{V2T}) \tag{13}$$

is used to ensure that the reconstructed feature can be correctly related to the corresponding person. While performing V2T and T2V reconstruction simultaneously, we also utilize a consistency constraint loss

$$L_{RCC} = MSE(S(R_{T2V}, V_G), S(R_{V2T}, T_G)) \tag{14}$$

to avoid the situation where visual and textual spaces develop and work independently, or even oppositely. A reconstruction discriminator D_r is employed to distinct a reconstructed feature with the original one, for which the loss is

$$L_{DAR} = \sum_{(X,Y)\in\Omega} \mathop{\mathbb{E}}_{X^i \sim X}[logD(X^i)] + \mathop{\mathbb{E}}_{Y^i \sim Y}[1 - logD(Y^i)], \tag{15}$$

where $\Omega = \{(V_G, R_{V2V}), (T_G, R_{T2T}), (V_G, R_{T2V}), (T_G, R_{V2T})\}$. The reconstruction adversarial loss for AMEN is

$$L_{AR} = \sum_{Y\in\Upsilon} \mathop{\mathbb{E}}_{Y^i \sim Y}[logD(Y^i)] \tag{16}$$

where $\Upsilon = \{R_{V2V}, R_{T2T}, R_{T2V}, R_{V2T}\}$. The final loss for training AMEN is

$$L_{Stage3} = L_{Stage2} + L_{RID} + L_{RM} + L_{RCC} + L_{AR}. \quad (17)$$

For testing, the combined similarity is defined as

$$S_{AMEN} = S(V_C, T_C) + \frac{1}{2}(S(V_G, T_L) + S(V_L, T_G)) + \frac{1}{2}(S(V_G, R_{T2V}) + S(R_{V2T}, T_G)). \quad (18)$$

4 Experiments

4.1 Experimental setup

Dataset and metrics At present CUHK-PEDES [11] is the only dataset for text-based person re-identification task. Follow the official data split approach, the training set contains 34054 images, 11003 persons and 68126 textual descriptions. The validation set contains 3078 images, 1000 persons and 6158 textual descriptions while the testing set has 3074 images, 1000 persons and 6156 descriptions. Almost every image has two descriptions, and each sentence is generally no shorter than 23 words. After dropping words that appear less than twice, the word number is 4984. The performance is evaluated by the top-k accuracy. Given a query description, all test images are ranked by their similarities with this sentence. If any image of the corresponding person is contained in the top-k images, we call this a successful search. We report the top-1, top-5, and top-10 accuracy for all experiments.

Implementation Details The feature dimension P is set to 1024 while $B = 512$. The phrases of each sentence are obtained with the Natural Language ToolKit (NLTK) by syntactic analysis, word segmentation and part-of-speech tagging. The total number of noun phrases n obtained from each sentence is kept flexible with an upper bound 26. Number of local strips k is set to 6. An Adam optimizer is used to train AMEN with a batch size of 32. The margin α of ranking losses is set to 0.2. In training stage-1, AMEN is trained with a learning rate of 1×10^{-3} for 10 epochs with the ResNet-50 backbone fixed. In stage-2, the learning rate is initialized as 2×10^{-4} to optimize all parameters of AMEN including the visual backbone except the two decoders for 3 epochs. Then in stage-3 the two reconstruction paradigms are added and we train the complete AMEN for extra 25 epochs. The learning rate is down-scaled by $\frac{1}{10}$ every 10 epochs.

4.2 Ablation analysis

Extensive ablation experiments are carried out to further investigate several key components of AMEN (shown in Table 1). ✓ and × denote whether the corresponding component or similarity is used. Comparing the results in line 2, 3

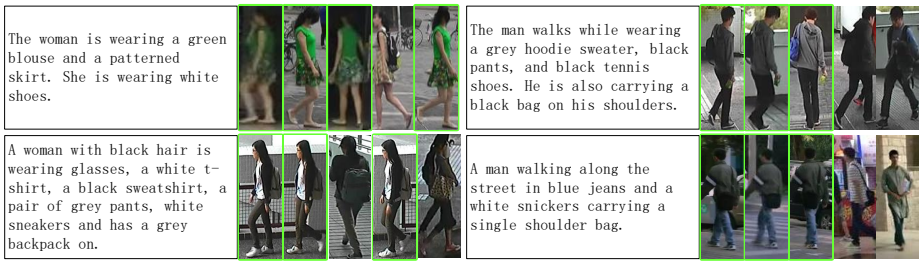

Fig. 2. Examples of top-5 text-based person re-identification results by AMEN. Images of the target pedestrian are marked by green rectangles.

Table 1. Ablation analysis of AMEN

No.	V2V	T2T	T2V	V2T	CC	RCC	D_m	D_r	S_C	S_{TL}	S_{VL}	S_{VR}	S_{TR}	Top-1	Top-5	Top-10
1	×	×	×	×	×	×	×	×	✓	✓	✓	×	×	53.08	76.54	84.86
2	✓	✓	×	×	×	×	✓	✓	✓	✓	✓	×	×	54.89	77.48	85.20
3	×	×	✓	✓	×	×	✓	✓	✓	✓	✓	✓	✓	54.93	77.83	84.93
4	✓	✓	✓	✓	×	×	✓	✓	✓	✓	✓	✓	✓	55.38	77.94	85.27
5	✓	✓	✓	✓	✓	×	✓	✓	✓	✓	✓	✓	✓	55.81	78.12	85.44
6	✓	✓	✓	✓	×	✓	✓	✓	✓	✓	✓	✓	✓	55.78	78.16	85.59
7	✓	✓	✓	✓	✓	✓	×	×	✓	✓	✓	✓	✓	54.98	77.81	85.01
8	✓	✓	✓	✓	✓	✓	✓	×	✓	✓	✓	✓	✓	55.99	78.27	85.48
9	✓	✓	✓	✓	✓	✓	×	✓	✓	✓	✓	✓	✓	56.15	78.31	85.43
10	✓	✓	✓	✓	✓	✓	✓	✓	✓	×	×	×	×	47.82	72.35	81.64
11	✓	✓	✓	✓	✓	✓	✓	✓	×	✓	×	×	×	48.99	73.65	82.16
12	✓	✓	✓	✓	✓	✓	✓	✓	×	×	✓	×	×	50.84	75.41	83.76
13	✓	✓	✓	✓	✓	✓	✓	✓	×	×	×	✓	×	50.26	74.11	83.07
14	✓	✓	✓	✓	✓	✓	✓	✓	×	×	×	×	✓	50.23	74.46	83.53
15	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	×	×	55.49	78.07	85.46
16	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	**57.16**	**78.64**	**86.22**

and 4, it can be concluded that the inter-modal reconstruction paradigm works in concert with the intra-modal one can properly embedding a feature into the opposite modality space while learning a strong common space. Via the consistency constraints, as shown from line 4, 5, 6 and 16, AMEN can learn the two high-level spaces consistently and jointly and improve in performance. As can be seen from line 7 and 16, the top-1 accuracy increases by 1.56% with the assistance of the tow proposed adversarial mechanisms, which indicates that being able to deceive the discriminator successfully, more discriminative representation vectors can be extracted and generated.

Table 2. Comparison with other state-of-the-art methods

Method	Top-1	Top-5	Top-10
CNN-RNN [19]	8.07	-	32.47
Neural Talk [22]	13.66	-	41.72
GNA-RNN [11]	19.05	-	53.64
IATV [10]	25.94	-	60.48
PWM-ATH [3]	27.14	49.45	61.02
Dual Path [27]	44.40	66.26	75.07
GLA [2]	43.58	66.93	76.26
MIA [17]	53.10	75.00	82.90
A-GANet [13]	53.14	74.03	81.95
GALM [7]	54.12	75.45	82.97
TIMAM [16]	54.51	77.56	84.78
IMG-Net [23]	56.48	76.89	85.01
CMAAM [1]	56.68	77.18	84.86
AMEN(ours)	**57.16**	**78.64**	**86.22**

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

The proposed AMEN is compared against 13 previous SOTA methods, including CNN-RNN [19], Neural Talk [22], GNA-RNN [11], IATV [10], PWM-ATH [3], Dual Path [27], GLA [2], MIA [17], A-GANet [13], GALM [7], TIMAM [16], IMG-Net [23] and CMAAM [1]. As shown in Table 2, AMEN achieves the best performance under top-1, top-5 and top-10 accuracy metrics in the text-based person retrieval task on CUHK-PEDES, which approves the effectiveness of our proposed method. Compared with the best competitor MIA, the AMEN model significantly outperforms it by 3.58% under top-1 metric, indicating the effectiveness of the cross-modal attention. Pose-guided joint global and attentive local matching network (GALM) utilizes pose information to help localize the discriminative regions. With ResNet-50 as visual backbone, AMEN surpasses GALM by about 3% without suffering from the deviations of the pose estimation and the large computation consumption,which proves the effectiveness of two different reconstruction paradigms, namely inter-modal and intra-modal reconstruction.

Some examples of top-5 text-based person re-identification results by AMEN are displayed in Fig.2. Images of the target pedestrian are marked by green rectangles.

5 Conclusion

In this paper, we proposed a novel Adversarial Multi-space Embedding Network (AMEN) to learn and match embeddings in multiply spaces. Following a encoder-decoder manner, the inter-modal reconstruction paradigm works in concert with the intra-modal one to properly embedding a feature into the opposite modality space while learning a strong common space. A consistency constraint

is adapted to ensure that the learned visual and textual spaces are trained jointly and work consistently. To enhance both the common space learning and feature reconstruction, the adversarial mechanism is utilized. We evaluated our proposed AMEN on the CUHK-PEDES dataset, which is currently only accessible dataset for text-base person re-identification task. Extensive experimental results demonstrate that AMEN outperforms previous methods and achieves the state-of-the-art performance.

References

1. Aggarwal, S., Radhakrishnan, V.B., Chakraborty, A.: Text-based person search via attribute-aided matching. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 2617–2625 (2020)
2. Chen, D., Li, H., Liu, X., Shen, Y., Shao, J., Yuan, Z., Wang, X.: Improving deep visual representation for person re-identification by global and local image-language association. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 54–70 (2018)
3. Chen, T., Xu, C., Luo, J.: Improving text-based person search by spatial matching and adaptive threshold. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 1879–1887 (2018)
4. Faghri, F., Fleet, D.J., Kiros, J.R., Fidler, S.: Vse++: Improving visual-semantic embeddings with hard negatives. arXiv preprint arXiv:1707.05612 (2017)
5. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
6. Hou, R., Ma, B., Chang, H., Gu, X., Shan, S., Chen, X.: Interaction-and-aggregation network for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 9317–9326 (2019)
7. Jing, Y., Si, C., Wang, J., Wang, W., Wang, L., Tan, T.: Pose-guided multi-granularity attention network for text-based person search. arXiv preprint arXiv:1809.08440 (2018)
8. Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3128–3137 (2015)
9. Lee, K.H., Chen, X., Hua, G., Hu, H., He, X.: Stacked cross attention for image-text matching. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 201–216 (2018)
10. Li, S., Xiao, T., Li, H., Yang, W., Wang, X.: Identity-aware textual-visual matching with latent co-attention. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1890–1899 (2017)
11. Li, S., Xiao, T., Li, H., Zhou, B., Yue, D., Wang, X.: Person search with natural language description. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1970–1979 (2017)
12. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
13. Liu, J., Zha, Z.J., Hong, R., Wang, M., Zhang, Y.: Deep adversarial graph attention convolution network for text-based person search. In: Proceedings of the 27th ACM International Conference on Multimedia. pp. 665–673 (2019)

14. Liu, Y., Guo, Y., Bakker, E.M., Lew, M.S.: Learning a recurrent residual fusion network for multimodal matching. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4107–4116 (2017)
15. Nam, H., Ha, J.W., Kim, J.: Dual attention networks for multimodal reasoning and matching. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 299–307 (2017)
16. Nikolaos Sarafianos, Xiang Xu, I.A.K.: Adversarial representation learning for text-to-image matching. In: ICCV. pp. 5813–5823 (2019)
17. Niu, K., Huang, Y., Ouyang, W., Wang, L.: Improving description-based person re-identification by multi-granularity image-text alignments. IEEE Transactions on Image Processing **29**, 5542–5556 (2020)
18. Plummer, B.A., Wang, L., Cervantes, C.M., Caicedo, J.C., Hockenmaier, J., Lazebnik, S.: Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In: Proceedings of the IEEE international conference on computer vision. pp. 2641–2649 (2015)
19. Reed, S., Akata, Z., Lee, H., Schiele, B.: Learning deep representations of fine-grained visual descriptions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 49–58 (2016)
20. Sun, C., Song, X., Feng, F., Zhao, W.X., Zhang, H., Nie, L.: Supervised hierarchical cross-modal hashing. In: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 725–734 (2019)
21. Sun, Y., Xu, Q., Li, Y., Zhang, C., Li, Y., Wang, S., Sun, J.: Perceive where to focus: Learning visibility-aware part-level features for partial person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 393–402 (2019)
22. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3156–3164 (2015)
23. Wang, Z., Zhu, A., Zheng, Z., Jin, J., Xue, Z., Hua, G.: Img-net: inner-cross-modal attentional multigranular network for description-based person re-identification. Journal of Electronic Imaging **29**(4), 043028 (2020)
24. Xia, B.N., Gong, Y., Zhang, Y., Poellabauer, C.: Second-order non-local attention networks for person re-identification. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3760–3769 (2019)
25. Yan, F., Mikolajczyk, K.: Deep correlation for matching images and text. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3441–3450 (2015)
26. Yi, D., Lei, Z., Liao, S., Li, S.Z.: Deep metric learning for person re-identification. In: 2014 22nd International Conference on Pattern Recognition. pp. 34–39. IEEE (2014)
27. Zheng, Z., Zheng, L., Garrett, M., Yang, Y., Xu, M., Shen, Y.D.: Dual-path convolutional image-text embeddings with instance loss. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) **16**(2), 1–23 (2020)