LANGUAGE PERSON SEARCH WITH MUTUALLY CONNECTED CLASSIFICATION LOSS

Yuyu Wang[1], *Chunjuan Bo*[2], *Dong Wang*[1*], *Shuang Wang*[3], *Yunwei Qi*[3], *Huchuan Lu*[1]

[1]School of Information and Communication Engineering, Dalian University of Technology
[2]College of Electromechanical Engineering, Dalian Minzu University, [3]Alibaba Group
Corresponding author: Dong Wang*, wdice@dlut.edu.cn

ABSTRACT

In this work, we develop an effective person search algorithm with natural language descriptions. The contributions of this work mainly include two aspects. First, we design a baseline language person search framework including three basic components: a deep CNN model to extract visual features, a bi-directional LSTM to encode language descriptions and the triplet loss to conduct cross-modal feature embedding. Second, we propose a novel mutually connected classification loss to fully exploit the identity-level information, which not only introduces the identification information into both image and language descriptions but also encourages the cross-modal classification probabilities of the same identity to be more similar. The experimental results on the CUHK-PEDES dataset demonstrate that our method achieves significantly better performance than other state-of-the-art algorithms.

Index Terms— Person search, language information, identity-level information, classification loss, mutual learning

1. INTRODUCTION

Language person search is a special person re-identification problem [1, 2, 3, 4], which is useful in some extreme surveillance cases (such as telephone alarm). This task poses more challenges due to the ambiguity of cross-modal matching.

Usually, language descriptions for person re-identification include two aspects: attribute [3, 5] and natural language [4, 6] descriptions. The attribute descriptions attempt to exploit a set of pre-defined semantic attributes that describe the basic appearance information of persons . Compared with attributes, natural language can precisely describe the details of person appearance and provide much textual information.

Language-image related works have drawn increasing attentions in recent years. For image captioning, Vinyal *et al.* [7] fed high-level image features from CNN into LSTM for sequence estimation. For VQA, answering questions about given images [8, 9, 10, 11, 12, 13] learn a dynamic parameter layer with hashing techniques, which adaptively adjusts image features based on different questions for accurate answer classification. In addition, visual semantic embedding methods [14, 15, 16, 17, 18] learn to embed both language and images into a common space for both image classification and retrieval. In [4], a language person search method is proposed to retrieval person images based on a given sentence, which studies the matching probability of texts and images features extracted from LSTM and CNN respectively.

Motivated by the above-mentioned discussions, this paper presents an effective language person search algorithm. The main contributions of this work are two folds. First, we construct baseline person search framework with language descriptions. This framework consists of three fundamental components: a deep CNN model to extract visual features, a bi-directional LSTM to encode language descriptions and the triplet loss to conduct cross-modal feature embedding. Second, we propose a novel mutually connected classification loss to promote the discriminative feature embedding, which provides a better way to exploit the identity-level information. The experimental results on the CUHK-PEDES dataset demonstrate that our method achieves significantly better performance than other state-of-the-art algorithms.

2. PROPOSED APPROACH

We first design a baseline framework for the language person search problem, shown in Figure 1. The architecture consists of three components: a bi-directional LSTM with attention (left) to extract text features, a deep convolutional neural network (CNN) to learn image features and a cross-modal learning module to promote the cross-modal feature embedding (middle). We note that the cross-modal learning module of the baseline framework merely includes the matching scheme with triplet loss. Then, we propose a novel mutually connected classification loss to fully exploit the identity-level information, within the red rectangle in Figure 1.

2.1. The baseline framework with triplet loss

Deep text bi-LSTM with attention. The left part of our framework is aimed to extract text features, mainly including bi-directional LSTM (bi-LSTM) and attention modules. The bi-LSTM extracts deep caption features and the attention network emphasizes the importance of each word in captions. Given a caption, we first split it into words and obtain the code $\mathbf{T} = [\mathbf{t}_1, \mathbf{t}_2, ..., \mathbf{t}_n]$ by encoding each raw word into vectors based on the word2vec dictionary (n is the number of words).

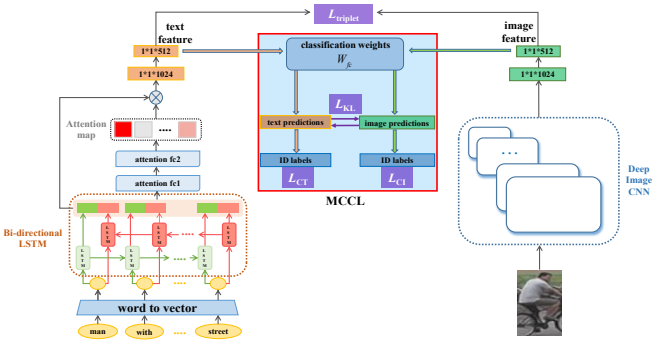


Fig. 1. The overall flowchart of the proposed approach. Better viewed in color and zoom in for details.

Next, we sequentially process the code $\mathbf{T}$ with a bidirectional LSTM to extract deep features. After forward and backward directions, we obtain two lists of hidden states,

$$\begin{aligned} \overrightarrow{\mathbf{h}}_t &= LSTM\left(\mathbf{t}_1, \overrightarrow{\mathbf{h}}_{t-1}\right) \\ \overleftarrow{\mathbf{h}}_t &= LSTM\left(\mathbf{t}_n, \overleftarrow{\mathbf{h}}_{t-1}\right) \end{aligned}, \tag{1}$$

where the subscript t is the temporal variable. $\overrightarrow{\mathbf{h}}$ and $\overleftarrow{\mathbf{h}}$ are the hidden states of two directions (we contact them to $\mathbf{h}$ in the following text, for simplicity). Let the hidden unit number for each unidirectional LSTM be u, thus, the initial word representations $\mathbf{h}_t$ from Bi-LSTM have size $2u$. We record all $\mathbf{h}_t$ to be a matrix $\mathbf{H} \in \mathbb{R}^{2u \times n}$ for initial text descriptions ($u = 512$ in our work).

Intuitively, different words tend to have different contributions for text descriptions but LSTM treats them equally. To address this issue, we introduce an attention module to process the initial matrix $\mathbf{H}$ with an attention map $\mathbf{A} \in \mathbb{R}^{r \times n}$,

$$\mathbf{A} = softmax\left(\mathbf{W}_{s2} \tanh\left(\mathbf{W}_{s1}\mathbf{H}^\top\right)\right). \tag{2}$$

This model contains two weight matrices, $\mathbf{W}_{s1} \in \mathbb{R}^{d_a \times 2u}$ and $\mathbf{W}_{s2} \in \mathbb{R}^{r \times d_a}$ ($d_a = 50$ and $r = 10$ in this paper).

Then, we multiply the attention matrix $\mathbf{A}$ with the bi-LSTM hidden states $\mathbf{H}$ and apply the max-pooling operation, obtaining a new feature $\mathbf{m} \in \mathbb{R}^{2u \times 1}$ as $\mathbf{m} = \max_r(\mathbf{AH})$ ($\max_r(.)$ is the maximum operation along the row direction). Finally, the feature $\mathbf{m}$ is further proceeded using one fully connected layer to obtain the final text feature $\mathbf{f}_T \in \mathbb{R}^{512 \times 1}$.

Deep image CNN. To extract visual image features, we employ the MobileNet [19] pre-trained on the ImageNet dataset as our basic model (see the right part in Figure 1). Given an input image of size 224×224, after the forward pass of the network, we extract the initial feature from the last pooling layer. After that, one fully connected layer is applied to obtain the final image feature $\mathbf{f}_I \in \mathbb{R}^{512 \times 1}$.

Cross-modal feature embedding. It requires to conduct cross-modal feature embedding for the language person search problem, since the text and image descriptions are from different domains. In this work, we first develop a baseline method with the triplet loss function. Given a quadric embedding features $\left(\mathbf{f}_T^+, \mathbf{f}_I^+, \mathbf{f}_T^-, \mathbf{f}_I^-\right)$ ($\left(\mathbf{f}_T^+, \mathbf{f}_I^+\right)$ are positive text and image features, and $\left(\mathbf{f}_T^-, \mathbf{f}_I^-\right)$ are negative ones). The triplet loss function (3) explicitly constructs the relationship between the image and text descriptions.

$$\begin{aligned} L_{triplet} = &\max\left[0, \alpha + D\left(\mathbf{f}_T^+, \mathbf{f}_I^-\right) - D\left(\mathbf{f}_T^+, \mathbf{f}_I^+\right)\right] + \\ &\max\left[0, \alpha + D\left(\mathbf{f}_I^+, \mathbf{f}_T^-\right) - D\left(\mathbf{f}_I^+, \mathbf{f}_T^+\right)\right] \end{aligned} \tag{3}$$

where $D(.,.)$ is the cosine distance to measure the similarity between two samples, α is a margin (simply set to 1 in this work). The former term is with respect to the image anchor; while the latter one corresponds to the text anchor. Thus, this triplet loss function encourages the matched image feature and text description to have a higher similarity score.

The combination of the above-mentioned text and image extractors merely with the triplet loss function serves as the baseline algorithm in this work.

3. MUTUALLY CONNECTED CLASSIFICATION LOSS

We note that although the triplet loss function explicitly considers the matching process between features from two modalities, it cannot fully exploit the feature distribution in each individual domain. The identity information has been shown to be effective in coping with the image-based person re-identification and language person search problems.

By introducing fully connected classification layers after the cross-modal embedded features, the prediction probabilities can be obtained as

$$\begin{aligned} P_T &= softmax\left(\mathbf{W}_{fc}^\top \mathbf{f}_T\right) \\ P_I &= softmax\left(\mathbf{W}_{fc}^\top \mathbf{f}_I\right) \end{aligned}, \tag{4}$$

where $\mathbf{W}_{fc}$ is the shared parameter of the fully connected layer for classification. The corresponding classification losses with the cross-entropy criterion can be obtained as

$$\begin{aligned} L_{CT} &= \mathbf{E}\left[-\log P_T\right] \\ L_{CI} &= \mathbf{E}\left[-\log P_I\right] \end{aligned}, \tag{5}$$

where P_T and P_I denote the predictions of text and image classifications, respectively. L_{CT} and L_{CI} stand for their corresponding classification losses.

Simply, the overall classification loss can be defined as a summation of L_{CT} and L_{CI} terms, i.e.,

$$L_C = L_{CT} + L_{CI}. \tag{6}$$

Besides the traditional classification loss (6), we introduce an additional mutually connected constraint between the predictions P_T and P_I as

$$L_{KL} = \frac{1}{|S^+|} \sum_{i \in S^+} \left[KL\left(P_T^i, P_I^i\right) + KL\left(P_I^i, P_T^i\right)\right] \tag{7}$$

where S^+ indicates the matched sample set with all samples from same identities and $|S^+|$ stands for the number of all matched samples. $KL(.,.)$ denotes the Kullback-Leibler (KL) distance. Thus, we obtain a novel mutually connected classification loss (MCCL) as

$$L_{MCC} = L_C + L_{KL}. \tag{8}$$

The final cross-modal feature embedding is conducted to minimize the combination of the triplet loss and MCC one,

$$L = L_{triplet} + L_{MCC}. \tag{9}$$

Our MCCL scheme facilitates a better cross-modal feature embedding due to the following two reasons. First, the classification weight $\mathbf{W}_{fc}$ is shared between two modalities, encouraging the learned text and image features within the same subspace. Second, the Kullback-Leibler divergence is adopted to match the predictions of text and image branches. This makes the matched text and image samples (with same identities) have similar classification probabilities and further encourages their embedded features to be similar.

4. EXPERIMENTS AND DISCUSSIONS

4.1. Dataset and implementation details

To our knowledge, there is only one publicly person search dataset with natural language descriptions, i.e., CUHK-PEDES [6]. The CUHK-PEDES dataset is constructed by selecting images from many different person re-identification datasets and adding the corresponding language annotations. It contains $40,206$ images of $13,003$ identities and $80,440$ textual descriptions (each image is described by about 2 sentences). Based on [6], this dataset is divided into three parts: (1) $11,003$ training individuals with $34,054$ images and $68,126$ captions; (2) $1,000$ validation persons with $3,078$ images and $6,158$ sentences; and (3) $1,000$ test identities with $3,074$ images and $6,156$ captions. Some visual illustrations are presented in Figure 2 (see the text queries and the retrieval images with green rectangles).

The proposed framework and loss functions are all implemented in the TensorFlow platform with a NVIDIA GEFORCE GTX 1080Ti GPU. To be specific, we use the MobileNet model [19] to obtain the visual features and employ the Bi-LSTM structure to extract the textual features. In addition, we adopt the Adam optimizer [20] to conduct the optimization process with $lr = 0.0002$. The batch size is 64. We adopt Recall@K (R@K for short) for evaluation.

4.2. Results on the CUHK-PEDES dataset

In Table 1, we evaluate the proposed methods in comparison with existing algorithms on the CUHK-PEDES dataset. These methods include deeper LSTM Q+norm [21], iBOWIMG [22], NeuralTalk [23], Word CNN-RNN [17], GNA-RNN [6], GMM+HGLMM [24], and Latent Co-attention [4]. We can see that our final model (Baseline+MCCL) achieves the best performance (50.58% of R@1 and 79.06% of R@10), outperforming the second best one (Latent Co-attention) by a large margin. The underlying reasons are three aspects: (1) a more effective baseline method is designed for the language person search problem; (2) the usage of the classification loss facilitates obtaining more discriminative feature embedding since it adds the identification supervision into the traditional triplet loss; (3) the proposed MCCL provides a better way to exploit the identification information.

Table 1. Comparison of person search results (R@K(%)) on the CUHK-PEDES dataset.

Method	R@1	R@10
deeper LSTM Q+norm [21]	17.19	57.82
iBOWIMG [22]	8.00	30.56
NeuralTalk [23]	13.66	41.72
Word CNN-RNN [17]	10.48	36.66
GNA-RNN [6]	19.05	53.64
GMM+HGLMM [24]	15.03	42.27
Latent Co-attention [4]	25.94	60.48
Baseline (Triplet)	45.55	75.50
Baseline+CL	48.21	78.27
Baseline+MCCL	50.58	79.06

4.3. Visual comparisons and ablation analysis

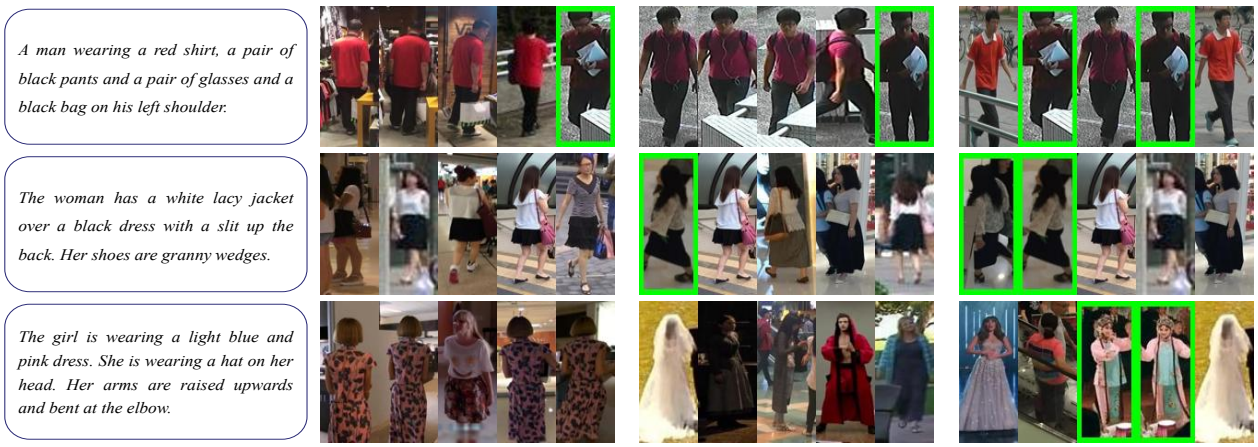
Figure 2 shows visual results of different algorithms (Baseline, Baseline+CL and Baseline+MCCL). Figure 2 (a) provides text queries, and (b-d) illustrates the related retrieval results. We can see that our MCCL method works well compared with other related ones. To better understand the MCCL, we provide some representative examples for test feature distributions learned by different methods (shown in Figure 3). We can observe that the MCCL method facilitates obtaining a much better embedding of text and image features.

Table 2. Comparison of different person re-identification methods (R@K(%) and MAP) on the Market-1501 dataset. Baseline* stands for the method presented in [25].

Method	R@1	R@5	MAP
PUL [26]	44.70	59.1	20.1
WARCA [27]	45.16	68.12	–
Histogram [28]	59.47	80.73	–
Bilinear-CNN [29]	66.36	85.01	41.17
P2S [30]	70.72	–	44.27
Spindle Net [31]	76.90	91.50	–
k-reciprocal [32]	77.11	–	63.63
DML [33]	87.73	–	68.83
MTMCT [34]	89.46	–	75.67
HA-CNN [35]	91.20	–	75.70
Baseline*	82.72	95.29	66.61
Baseline*+CL	87.89	95.34	71.51
Baseline*+MCCL	88.21	95.37	71.90

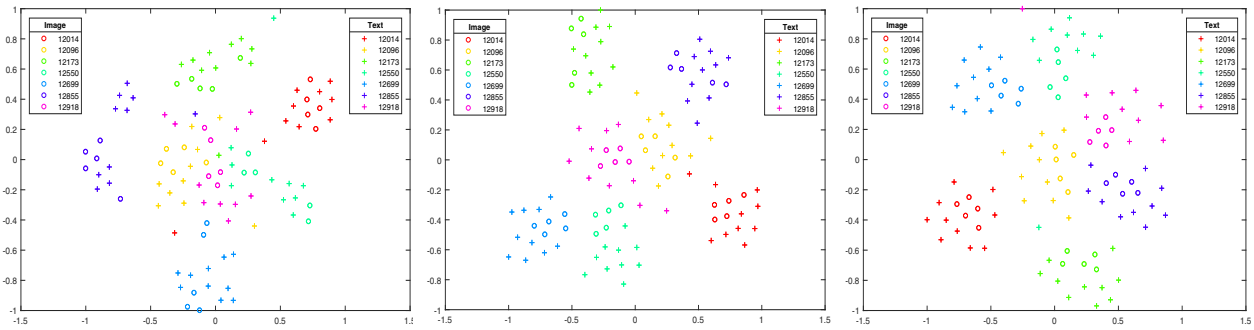
4.4. Generalization ability of our MCCL scheme

We note that the proposed mutually connected classification loss is not only useful for the language person search prob-



(a) Text queries (b) Baseline (c) Baseline+CL (d) Baseline+MCCL

Fig. 2. Examples of top-5 language person search results by different algorithms, including (b) Baseline, (c) Baseline+CL and (d) Baseline+MCCL. Corresponding images are marked by green bounding boxes.



(a) Baseline (b) Baseline+CL (c) Baseline+MCCL

Fig. 3. Comparison of feature distribution learned with different algorithms (better viewed in color).

lem, but also works for improving the matching model (triplet loss) with classification information. Here, we demonstrate the generalization ability of our MCCL scheme for image-based person re-identification.

To be specific, we choose a baseline method with triplet loss presented in [25] (denoted as Baseline* in Table 2). We adopt the commonly used Market-1501 [36] dataset to compare the proposed method with other competing ones. The Market-1501 dataset is widely used for person re-identification. The bounding boxes are selected by cross-joint overlap from person detectors and manually annotated ones. It contains $32,668$ images of $1,501$ identities captured from six non-overlapping camera views, with 12936 images of 751 identities for training and $19,732$ images of 750 identities for testing. The compared algorithms include PUL [26], WARCA [27], Histogram [28], Bilinear-CNN [29], P2S [30], Spindle Net [31], k-reciprocal [32], DML [33], MTMCT [34], and HA-CNN [35]. From Table 2, we can see that our MCCL scheme also improves the baseline method by a large margin. After this improvement, the re-identification performance is comparable with recent top-ranked algorithms (notice that our algorithm does not use complicated feature extraction schemes or part-based models).

5. CONCLUSIONS

Language person search requires a robust cross-modal feature embedding to promote the retrieval process.This work fist designs a cross-modal matching framework with the triplet loss function. In our framework, a deep CNN model is used to extract image features and a bi-LSTM with the attention scheme is adopted to encode the text information. Second, a novel mutually connected classification loss (MCCL) is proposed to effectively exploit the identity-level information. The experimental results on the CUHK-PEDES dataset show that our final model improves the baseline model and other competing methods by a large margin. In addition, the experimental results on the Market-1501 dataset also demonstrate the good generalization ability of the proposed MCCL scheme.

Acknowledgement. This paper was supported in part by the Natural Science Foundation of China #61751212, #61872056, #61806037, #61725202, and in part by the Fundamental Research Funds for the Central Universities under Grant #DUT18JC30. This work was also supported by Alibaba Group through Alibaba Innovative Research (AIR) program.

6. REFERENCES

[1] Zimo Liu, Dong Wang, and Huchuan Lu, "Stepwise metric promotion for unsupervised video person re-identification," in *ICCV*, 2017, pp. 2448–2457.

[2] Ju Dai, Pingping Zhang, Dong Wang, Huchuan Lu, and Hongyu Wang, "Video person re-identification by temporal residual learning," *IEEE Transactions on Image Processing*, vol. 28, no. 3, pp. 1366–1377, 2019.

[3] Yutian Lin, Liang Zheng, Zhedong Zheng, Yu Wu, and Yi Yang, "Improving person re-identification by attribute and identity learning," *CoRR*, vol. abs/1703.07220, 2017.

[4] Shuang Li, Tong Xiao, Hongsheng Li, Wei Yang, and Xiaogang Wang, "Identity-aware textual-visual matching with latent co-attention," in *ICCV*, 2017, pp. 1908–1917.

[5] Jingya Wang, Xiatian Zhu, Shaogang Gong, and Wei Li, "Transferable joint attribute-identity deep learning for unsupervised person re-identification," in *CVPR*, 2018, pp. 2275–2284.

[6] Shuang Li, Tong Xiao, Hongsheng Li, Bolei Zhou, Dayu Yue, and Xiaogang Wang, "Person search with natural language description," in *CVPR*, 2017, pp. 5187–5196.

[7] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan, "Show and tell: A neural image caption generator," in *CVPR*, 2015, pp. 3156–3164.

[8] Mengye Ren, Ryan Kiros, and Richard S. Zemel, "Exploring models and data for image question answering," in *NIPS*, 2015, pp. 2953–2961.

[9] Mateusz Malinowski, Marcus Rohrbach, and Mario Fritz, "Ask your neurons: A neural-based approach to answering questions about images," in *CVPR*, 2015, pp. 1–9.

[10] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach, "Multimodal compact bilinear pooling for visual question answering and visual grounding," in *EMNLP*, 2016, pp. 457–468.

[11] Hyeonwoo Noh, Paul Hongsuck Seo, and Bohyung Han, "Image question answering using convolutional neural network with dynamic parameter prediction," in *CVPR*, 2016, pp. 30–38.

[12] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alexander J. Smola, "Stacked attention networks for image question answering," in *CVPR*, 2016, pp. 21–29.

[13] Kuniaki Saito, Andrew Shin, Yoshitaka Ushiku, and Tatsuya Harada, "DualNet: Domain-invariant network for visual question answering," in *ICME*, 2017, pp. 829–834.

[14] Andrea Frome, Gregory S. Corrado, Jonathon Shlens, Samy Bengio, Jeffrey Dean, Marc'Aurelio Ranzato, and Tomas Mikolov, "DeViSE: A deep visual-semantic embedding model," in *NIPS*, 2013, pp. 2121–2129.

[15] Wu Liu, Tao Mei, Yongdong Zhang, Cherry Che, and Jiebo Luo, "Multi-task deep visual-semantic embedding for video thumbnail selection," in *CVPR*, 2015, pp. 3707–3715.

[16] Mengye Ren, Ryan Kiros, and Richard S. Zemel, "Image question answering: A visual semantic embedding model and a new dataset," *CoRR*, vol. abs/1505.02074, 2015.

[17] Scott E. Reed, Zeynep Akata, Honglak Lee, and Bernt Schiele, "Learning deep representations of fine-grained visual descriptions," in *CVPR*, 2016, pp. 49–58.

[18] Andrej Karpathy and Li Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," *TPAMI*, vol. 39, no. 4, pp. 664–676, 2017.

[19] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam, "MobileNets: Efficient convolutional neural networks for mobile vision applications," *CoRR*, vol. abs/1704.04861, 2017.

[20] Diederik P. Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014.

[21] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh, "VQA: visual question answering," in *ICCV*, 2015, pp. 2425–2433.

[22] Bolei Zhou, Yuandong Tian, Sainbayar Sukhbaatar, Arthur Szlam, and Rob Fergus, "Simple baseline for visual question answering," *CoRR*, vol. abs/1512.02167, 2015.

[23] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan, "Show and tell: Lessons learned from the 2015 MSCOCO image captioning challenge," *TPAMI*, vol. 39, no. 4, pp. 652–663, 2017.

[24] Benjamin Klein, Guy Lev, Gil Sadeh, and Lior Wolf, "Associating neural word embeddings with deep image representations using fisher vectors," in *CVPR*, 2015, pp. 4437–4446.

[25] Alexander Hermans, Lucas Beyer, and Bastian Leibe, "In defense of the triplet loss for person re-identification," *CoRR*, vol. abs/1703.07737, 2017.

[26] Hehe Fan, Liang Zheng, Chenggang Yan, and Yi Yang, "Unsupervised person re-identification: Clustering and fine-tuning," *TOMCCAP*, vol. 14, no. 4, pp. 83:1–83:18, 2018.

[27] Cijo Jose and François Fleuret, "Scalable metric learning via weighted approximate rank component analysis," in *ECCV*, 2016, pp. 875–890.

[28] Evgeniya Ustinova and Victor S. Lempitsky, "Learning deep embeddings with histogram loss," in *NIPS*, 2016, pp. 4170–4178.

[29] Evgeniya Ustinova, Yaroslav Ganin, and Victor S. Lempitsky, "Multiregion bilinear convolutional neural networks for person re-identification," *CoRR*, vol. abs/1512.05300, 2015.

[30] Sanping Zhou, Jinjun Wang, Jiayun Wang, Yihong Gong, and Nanning Zheng, "Point to set similarity based deep feature learning for person re-identification," in *CVPR*, 2017, pp. 5028–5037.

[31] Haiyu Zhao, Maoqing Tian, Shuyang Sun, Jing Shao, Junjie Yan, Shuai Yi, Xiaogang Wang, and Xiaoou Tang, "Spindle net: Person re-identification with human body region guided feature decomposition and fusion," in *CVPR*, 2017, pp. 907–915.

[32] Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li, "Re-ranking person re-identification with k-reciprocal encoding," in *CVPR*, 2017, pp. 3652–3661.

[33] Ying Zhang, Tao Xiang, Timothy M. Hospedales, and Huchuan Lu, "Deep mutual learning," in *CVPR*, 2018, pp. 4320–4328.

[34] Ergys Ristani and Carlo Tomasi, "Features for multi-target multi-camera tracking and re-identification," in *CVPR*, 2018, pp. 6036–6046.

[35] Wei Li, Xiatian Zhu, and Shaogang Gong, "Harmonious attention network for person re-identification," in *CVPR*, 2018, pp. 2285–2294.

[36] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian, "Scalable person re-identification: A benchmark," in *ICCV*, 2015, pp. 1116–1124.