Cross-Modal Person Search: A Coarse-to-Fine Framework using Bi-directional Text-Image Matching

Xiaojing Yu[1]*, Tianlong Chen[1]*, Yang Yang[2], Michael Mugo[2], Zhangyang Wang[1]
[1]Texas A&M University, [2]Walmart Technology
{vicky_yu,wiwjp619,atlaswang}@tamu.edu
{yang.yang2,michael.mugo}@walmart.com

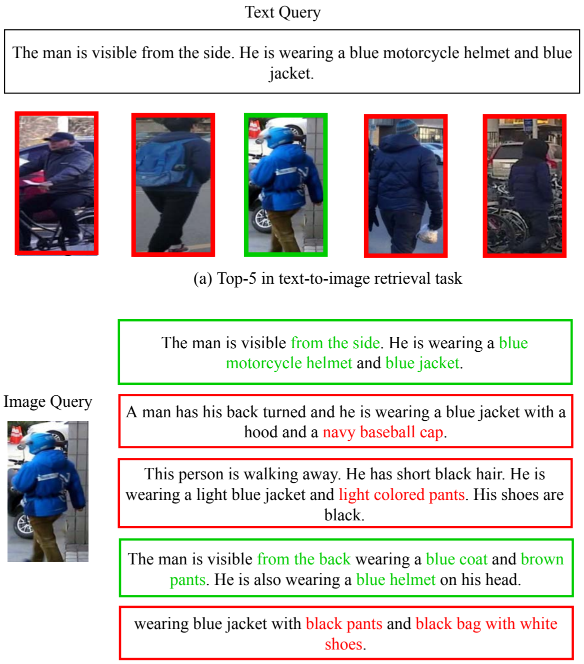
Abstract

*Searching person images from a gallery based on natural language descriptions remains to be a challenging and under-explored cross-modal retrieval problem. To improve the accuracy off an image-based retrieval task, e.g., person re-identification (Person Re-Id), re-ranking is known to be an effective post-processing tool. In this paper, we extend re-ranking from uni-modal retrieval to cross-modal retrieval for the first time, and develop a bi-directional coarse-to-fine framework (BCF) for cross-modal person search. Built on a recent state-of-the-art Person Re-Id model [5], BCF exploits first text-to-image and then image-to-text relevance, in a two-stage refinement fashion. BCF ranks competitively against a strong baseline[24] on the newly-introduced WIDER Person Search dataset [1], boosting validation set performance by 9.01%(top-1)/3.87%(mAP) for **val1** and 6.60%(top-1)/3.49%(mAP) for **val2** , respectively. With a high score, our solution ranks competitively in the ICCV 2019 WIDER Person Search by Language Challenge.*

1. Introduction

Searching person by natural language descriptions is an important application instance of cross-modal retrieval. Given a textual description of a specific person, its objective is to find images from gallery which best match the description. Based on existing methods for text-to-image retrieval, significant challenges remain to be addressed. For improvements, our work is motivated by multi-fold observations:

- **Cross-modality gives rise to (more) ambiguity:** In addition to the semantic ambiguities in either modality, it is typically unrealistic to assume one-to-one image-text mapping. For example, in the newly-introduced WIDER Person Search dataset [1], one person ID can have multiple gallery images, which means that for one textual description, there could be multiple "ground-truth" image-text matching pairs.
- **Bidirectional matching is more reliable:** Intuitively (and empirically observed by us), if an image and a text are found to match each other in both directions (i.e., text-to-image, and image-to-text), then they are more likely to make a correct match. Additionally, we observe the image-to-text retrieval to be often more accurate than the other way around, e.g., in Figure 1.



The man is visible from the side. He is wearing a blue motorcycle helmet and blue jacket.

(a) Top-5 in text-to-image retrieval task

The man is visible from the side. He is wearing a blue motorcycle helmet and blue jacket.

A man has his back turned and he is wearing a blue jacket with a hood and a navy baseball cap.

This person is walking away. He has short black hair. He is wearing a light blue jacket and light colored pants. His shoes are black.

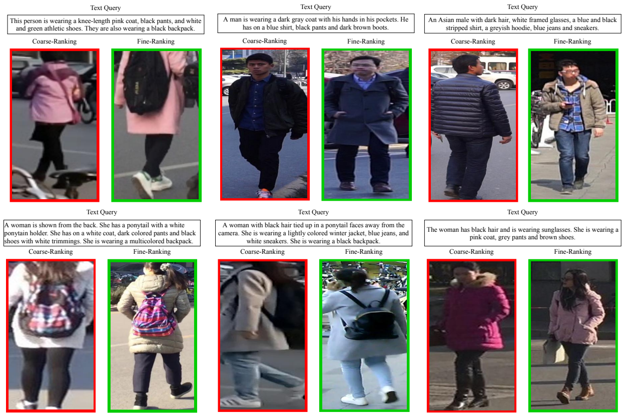
The man is visible from the back wearing a blue coat and brown pants. He is also wearing a blue helmet on his head.

wearing blue jacket with black pants and black bag with white shoes.

(b) Top-5 in image-to-text retrieval task

Figure 1. Top-5 results in text-to-image retrieval and image-to-text retrieval. The image/text with green border indicates a correct match. Using text and image embedding extracted with the same image-text matching model, image-to-text retrieval is consistently more accurate than text-to-image retrieval.

*Equal Contribution. Work is done in Walmart Technology.



This person is wearing a knee-length pink coat, black pants, and white and green athletic shoes. They are also wearing a black backpack.

A man is wearing a dark gray coat with his hands in his pockets. He has on a blue shirt, black pants and dark brown boots.

An Asian male with dark hair, white framed glasses, a blue and black stripped shirt, a greyish hoodie, blue jeans and sneakers.

A woman is shown from the back. She has a ponytail with a white ponytain holder. She has on a white coat, dark colored pants and black shoes with white trimmings. She is wearing a multicolored backpack.

A woman with black hair tied up in a ponytail faces away from the camera. She is wearing a lightly colored winter jacket, blue jeans, and white sneakers. She is wearing a black backpack.

The woman has black hair and is wearing sunglasses. She is wearing a pink coat, grey pants and brown shoes.

Figure 2. Visualization of person search by nature language. In each of the six examples, given a text query (top), the left image shows the top-1 image result in coarse-ranking stage while right image shows the top-1 image in the fine-ranking stage.

Driven by these observations, we propose a novel bi-directional framework, based on an image-text matching model [24] and a state-of-the-art Person Re-Id model [5]:

- Given a text query T_q, gallery images are first ranked through the image-text matching model M, which is termed as the **coarse-ranking** stage (text-to-image retrieval). A person Re-Id model [5] is then applied to each image in the coarse-ranking list L to find its top-k similar images (image-to-image retrieval).
- In the next **fine-ranking** stage (image-to-text retrieval), for each of the top-k images, query texts are ranked through the model M. The rank of T_q in the list is used to compute the final score of each image candidate in L.

The proposed coarse and fine ranking stages boost the performance of person search, as shown in Figure 2. For the first time, we integrate image-to-text, image-to-image (by Re-Id) and text-to-image retrieval, to refine the ranking from coarse to fine. We hence call our framework **bi-directional coarse-to-fine framework** (BCF).

2. Related work

Text-based Person Search: Attention mechanism has been a popular tool to capture relations between words and image regions [13, 12, 9, 4, 6]. Auxiliary tasks, such as classification and image caption, are also used to improve the image and text embeddings [24, 4, 8]. Because text-based person search is a specific application of image-text matching, most image-text matching methods can also be used in text-based person search [7, 11, 10, 22, 17, 15]. These methods use different models to learn image and text embeddings separately and then measure similarity of image and text embeddings by calculating cosine distance or inner product of both embeddings. The image embeddings and text embeddings can be used for text-to-image retrieval and image-to-text retrieval in similar ways.

An interesting observation is that, based on the same model e.g., [24], the accuracy of image-to-text retrieval is often higher than the accuracy of text-to-image retrieval. This disparity indicates that text-to-image retrieval is perhaps a more difficult and ambiguous task. Our proposed method exploits the disparity and generates coarse initial proposal of top image-text pairs by text-to-image retrieval, and image-to-text retrieval (combined with a novel scoring algorithm) is used to refine the initial proposal.

Person Re-Identification and Re-Ranking: Image-based Person Re-Id has witnessed tremendous progress [20, 21, 14, 2]. Among many successful practices, re-ranking methods are widely used to refine the ranking list in Person Re-Id and other image-to-image retrieval tasks. Some re-ranking methods encode feature vectors to explore the similarities or differences of sub-feature vectors [2, 23]. Nguyen et al. [18] compute final score of gallery images using constraint information between people appearing simultaneously. The final score of a gallery image in [19] is jointly determined by the ranking list of the query image and query's k nearest neighbor (kNN) images. Jaccard distance is used to measure the distance of k-nearest neighbors in [3]. Zhong et al. [25] proposed k-reciprocal encoding to re-rank gallery images. However, existing methods only consider single modality, while we introduce a novel re-ranking framework for cross-modal person search for the first time.

3. Bi-directional coarse-to-fine framework

Figure 3 depicts the overall framework of BCF, consisting of two stages: coarse-ranking stage and fine-ranking stage. In the coarse-ranking stage, the feature embeddings of image and text are firstly extracted from an image-text matching model. By applying text-to-image retrieval, most irrelevant images are removed and we obtain potential image candidates for each query text. In the fine-ranking stage, we expand each candidate to a set of images by applying person Re-Id to add k nearest neighbours of the candidate, then we apply image-to-text retrieval for each image in image candidate sets to calculate the final score of image candidates and refine the ranking list.

Coarse-Ranking In the coarse ranking stage, we utilized the dual-path image-text embedding model [24] to learn the joint embeddings of image and text. This model takes gallery image and query text as input and embeds them into a common space via separate deep CNN models. For image embedding, the ImageNet pre-trained ResNet-50 is used as image CNN model to encode the input image of size 224 x

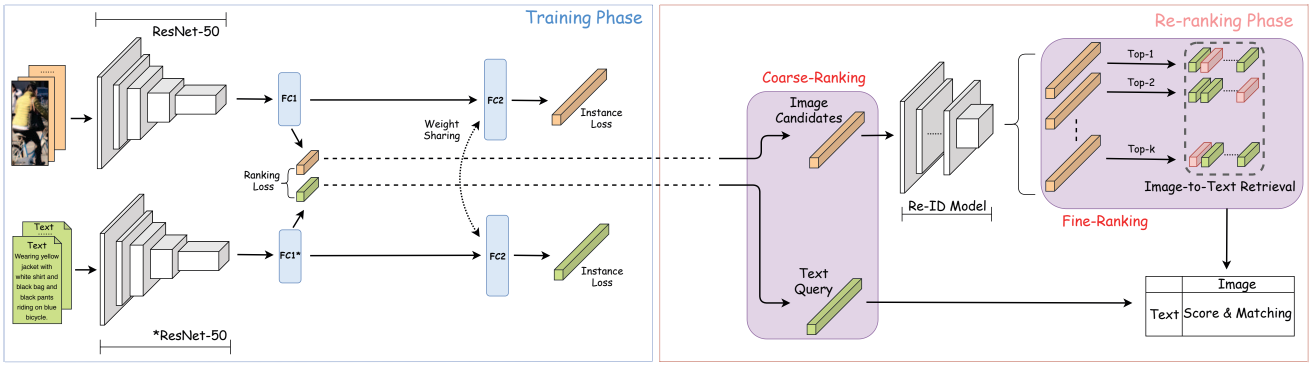


Figure 3. Framework of our proposed method. **In the training phase**, text CNN and image CNN learn the text and image embeddings respectively by instance loss and ranking loss [24], and share weights of last two fully-connected layer. **Re-ranking phase** includes 2 stages. In the **coarse-ranking stage**, the text embedding and image embedding are extracted to generate an initial ranking list. Then the initial ranking list of the text-to-image retrieval by query text is computed. In the **fine-ranking stage**, person Re-Id model takes the initial ranked images as input to retrieve the potential images of same persons and expand the initial images to a larger candidate set. Then image-to-text retrieval is applied for each image in the image set. The fine-ranking results are calculated by compute a novel score from the image-to-text retrieval results with the query text

224 into a 2048-dimension feature embedding. For text embedding, each word is encoded into a one-hot vector based on the word2vec dictionary in [16]. Hence, the sentence is converted into a matrix T belong to $\mathbb{R}^{n\times d}$, where n denotes the length of the sentence and d denotes the size of *word2vec* dictionary. Then we feed T into the text CNN model and generate a 2048-dimension feature embedding.

Both instance loss and ranking loss are used in [24], we follow the two-stage training procedure to fine-tune our model. After extracting the embeddings of image and text, we calculate the cosine similarity of each image-text pair and generate the ranking lists for both image-to-text retrieval and text-to-image retrieval.

Fine-Ranking In person search dataset, one person Id may have multiple gallery images. Therefore, instead of only using images from text-to-image retrieval directly, we first expand the original image candidates of the query text by a Person Re-Id model.

For each candidate in the top-k ranked text-to-image retrieval result, we use a state-of-the-art person Re-Id model ABD-Net [5] to retrieve its k nearest neighbors (kNN) and add them into the original image candidate set. The expansion is justified by two reasons: First, image-based Person Re-Id achieved very good performance (e.g. 95.60% top-1 accuracy on Market-1501 [5]). Second, by using kNN, the proposed method can effectively reduce the noise introduced due to the occlusions, variance of illumination conditions, and changes of viewpoints [25].

Let T and I represent the entire text set and image set respectively. Given a query text $T_i \in T$, we denote the rank of a gallery image $I_j \in I$ in the ranking list as $R_1(T_i, I_j)$. Given a gallery image I_i, we denote the rank of query text $T_j \in T$ in the ranking list as $R_2(I_i, T_j)$. Given a gallery image I_i, we denote the person Re-Id rank of another gallery image I_j as $R_3(I_i, I_j)$.

A query text's k-nearest neighbors in text-to-image retrieval task are the top-k gallery images. A gallery image's k-nearest neighbors in image-to-text retrieval task are the top-k text queries. It is very likely that the best matching gallery image will be included in the query text's k-nearest neighbors while the query text will be included in the best matching gallery image's k-nearest neighbors. Based on this observation, a non-scoring ranking method to refine the top-1 image I_{T_i} of query text T_i can be simply defined as follows:

$$I_{T_i} = \arg\max_{I_j}(R_1(T_i, I_j)|R_1(T_i, I_j) \leq k_0 \cap R_2(I_j, T_i) \leq k_0) \tag{1}$$

where k_0 is the number of nearest neighbors of text-to-image retrieval result. In this simple ranking method (1), if multiple gallery images have the same query text in its k-nearest neighbor set as I_j, the final rank of image I_j is determined by its position in $R_1(T_i, I_j)$ without considering actual rank of query text in $R_2(I_j, T_i)$. However, we notice that if the query text is 1-nearest neighbor of an image in image-to-text retrieval, this image should have higher possibility to be selected. Therefore, we proposed a scoring-based fine-ranking method which both considers the actual rank of text-to-image retrieval and image-to-text retrieval results. The score of a gallery image I_j is calculated as follows,

$$S(I_j) = \begin{cases} 0, R_1(T_i, I_j) > k_0 \cup R_2(I_j, T_i) > k_0 \\ +\infty, R_1(T_i, I_j) = 1 \cap R_2(I_j, T_i) \leq k_1 \\ 2k_0 - R_1(T_i, I_j) - R_2(I_j, T_i) + S_p(I_j), o.w. \end{cases} \tag{2}$$

where k_1 is the number of nearest neighbors of image-to-text retrieval result, and k_1 is much smaller than k_0 empirically. $S_p(I_j)$is a scoring term computed from expanded image candidates by person Re-Id model.

Since the rank of image may be affected by the illuminations, occlusions and viewpoints[25], both text-to-image retrieval and image-to-text retrieval of same person may not have a high rank as expected. To reduce the influence of these variances, we utilize a person Re-Id model to find potential images of the same person. For each image candidate, we collect k_2 nearest neighbour images of the candidate to form an image set. Then, the image-to-text retrieval result of each image in the set can be used to refine the rank of the candidate. We denote the k_2-nearest neighbor set of a gallery image I_j from ABD-Net[5] as $P_j = \{p \in I | R_2(p, T_i) \leq k_0 \cap R_3(I_j, p) \leq k_2\}$. The final score of I_j in k_2 nearest neighbors is determined as follows

$$S_p(I_j) = \sum_{p \in P_j} (k_0 - R_2(p, T_i)) \tag{3}$$

where k_2 is the number of nearest neighbors of image-to-text retrieval result. After we calculate the final score of original k-nearest neighbor images in coarse-ranking stage, the final rank list can be generated by sorting scores in descending order.

Our proposed BCF only requires the ranking results of image-text matching model and person Re-Id model (image-to-image retrieval). It can be used to refine the result of other image-text matching model to improve final accuracy in an efficient way.

4. Experiments

4.1. Dataset and Implementation

Dataset We evaluate our method in WIDER Person Search dataset [1], which is a large-scale image-language person search dataset with identity-level annotations. The images are collected from several Person Re-Id datasets, such as Market1501, CUHK01, CUHK03 and MSMT17. Each image has more than 2 text descriptions. Its training set includes 37132 images, and its two validation sets include 3074 images and 999 images respectively. The images in the first validation dataset are collected from the same Person Re-Id dataset as training set. The images in the second validation dataset are collected from MSMT17 dataset.

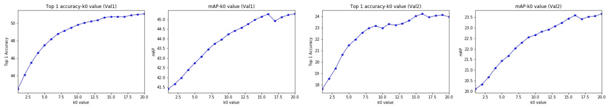


Figure 4. Top 1 accuracy and mAP using different parameter k_0.

Implementation When extracting image embedding, images are resized to 224x224 pixels by randomly cropping images. For hyper-parameters in the framework, i.e., k_0, k_1, k_2, we set k_0=20 (text-to-image), k_1=2 (image-to-text), k_2=1 (person Re-Id) for val1 set and k_0=16 (text-to-image), k_1=1 (image-to-text), k_2=1 (person Re-Id) for val2 set. The dual-path image-text matching model is pre-trained in CUHK-PEDES dataset. The Person Re-Id model is pre-trained on MSMT17 dataset and then is used to extract the ranking list of images in the training set and two validation sets.

4.2. Evaluation

The model is evaluated on both validation sets. We compare it against a strong baseline model, i.e., the dual-path image-text matching model in [24]. We evaluate the simple non-scoring fine-ranking method (NSF) in Eq.(1) and the scoring-based fine-ranking (SF) method in Eq.(2) with/without term $S_p(I_j)$ (i.e., using person Re-Id or not) to analyze the effectiveness of each method. We use mAP and top-1 accuracy to evaluate these methods.

We compared BCF against the baseline on WIDER Person Search dataset: see Table 1. Using NSF improves baseline on both validation sets by 4.95%(top-1)/ 0.51%(mAP) on val1 set and 3.15%(top-1)/ 1.16 %(mAP) on val2. Using SF gains additional improvements on both metrics by 3.30%(top-1)/ 2.70%(mAP) on val1 set and 2.45%(top-1)/ 1.57%(mAP) on val2 set, against with NSF. By combining ABD-Net and SF, the performance is further boosted by 0.31%(top-1)/ 0.66%(mAP) on val1 and 1.00%(top-1)/ 0.76%(mAP) on val2, which validates the effectiveness of utilizing person Re-Id result to expand image candidate set for fine-ranking.

Table 1. Ablation Study of Proposed Framework

Method	val1		val2	
	top1 (%)	mAP (%)	top1 (%)	mAP (%)
Baseline	42.50	41.41	17.62	20.10
Baseline + NSF.	47.45	41.92	20.77	21.26
Baseline + SF.	50.75	44.62	23.22	22.83
BCF (last + ABD)	51.06	45.28	24.22	23.59

4.3. Parameter analysis

We analyze the selection of hyper-parameters k_0, k_1, k_2 in our proposed framework. Figure 4,5,6 show the top-1

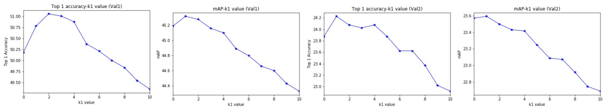


Figure 5. Top 1 accuracy and mAP using different parameter k_1.

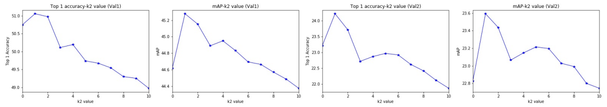


Figure 6. Top 1 accuracy and mAP using different parameter k_2.

accuracy of BCF with different hyper-parameters. The accuracy keeps improving on val1 set when k_0 increases from 1 to 20. When k_0=16, top-1 accuracy and mAP on val2 increase to the highest value. When k_1=2 and 1, val1 and val2 achieve highest accuracy respectively. So we choose k_0=20, k_1=2 for val1 set and k_0=16, k_1=1 for val2 set comparison. When k_2=1 (1-nearest neighbor), both validation sets achieve peak value on mAP. This indicates each person id might have 2 gallery images on average. So we choose k_2=1 for comparison.

4.4. Visualization of Query Results

We show some representative person search results in Figure 7. The retrieved images in coarse-ranking stage failed in associating color information with specific objects description (e.g. orange bag), while the retrieved images in fine-ranking stage managed to correctly locate the top-1 best match. The BCF is able to map phrase to fine-grained image attributes, such as backpack etc.

5. Conclusion

We proposed a novel bi-directional coarse-to-fine framework (BCF) for cross-modal person search. The experiments demonstrate its effectiveness in the person search by natural language task. Note that different image-text matching models and ReID models can be plugged into the BCF framework without hassle. Our future work includes applying this framework on other cross-modal retrieval task and designing more robust scoring rules.

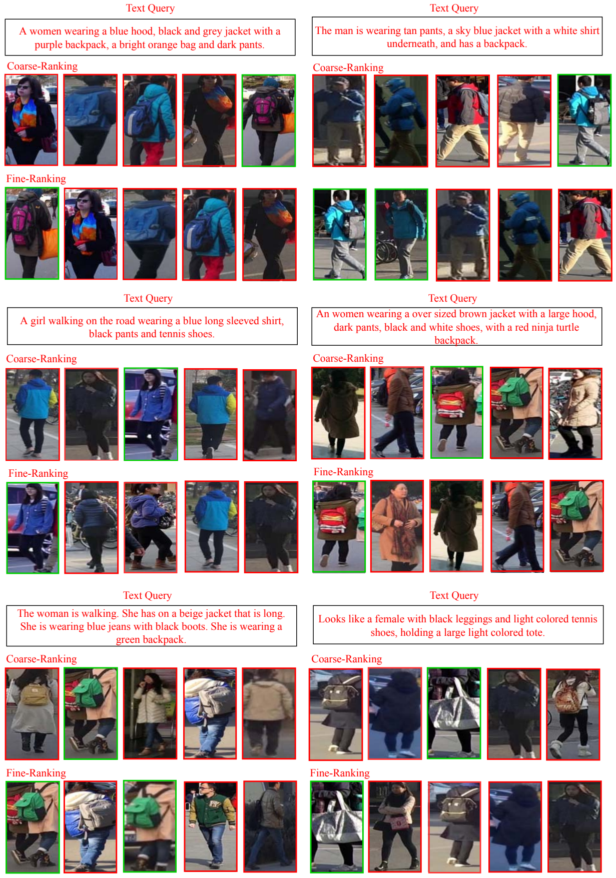

Text Query
A women wearing a blue hood, black and grey jacket with a purple backpack, a bright orange bag and dark pants.
Coarse-Ranking
Fine-Ranking
Text Query
The man is wearing tan pants, a sky blue jacket with a white shirt underneath, and has a backpack.
Coarse-Ranking
Fine-Ranking
Text Query
A girl walking on the road wearing a blue long sleeved shirt, black pants and tennis shoes.
Coarse-Ranking
Fine-Ranking
Text Query
An women wearing a over sized brown jacket with a large hood, dark pants, black and white shoes, with a red ninja turtle backpack.
Coarse-Ranking
Fine-Ranking
Text Query
The woman is walking. She has on a beige jacket that is long. She is wearing blue jeans with black boots. She is wearing a green backpack.
Coarse-Ranking
Fine-Ranking
Text Query
Looks like a female with black leggings and light colored tennis shoes, holding a large light colored tote.
Coarse-Ranking
Fine-Ranking

Figure 7. Visualization of top-5 images in coarse-ranking stage and fine-ranking stage. As shown in the examples, our method benefits from the text retrieval result and successfully map text phrase to fine-grained image attributes, such as backpack etc.

References

[1] Wider person search by language dataset. http://wider-challenge.org/2019.html. 1, 4

[2] E. Ahmed, M. Jones, and T. K. Marks. An improved deep learning architecture for person re-identification. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 3908–3916, 2015. 2

[3] S. Bai and X. Bai. Sparse contextual activation for efficient visual re-ranking. IEEE Transactions on Image Processing, 25(3):1056–1069, 2016. 2

[4] D. Chen, H. Li, X. Liu, Y. Shen, J. Shao, Z. Yuan, and X. Wang. 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), pages 54–70, 2018. 2

[5] T. Chen, S. Ding, J. Xie, Y. Yuan, W. Chen, Y. Yang, Z. Ren, and Z. Wang. Abd-net: Attentive but diverse person re-identification. In The IEEE International Conference on Computer Vision (ICCV), Oct 2019. 1, 2, 3, 4

[6] T. Chen, C. Xu, and J. Luo. Improving text-based person search by spatial matching and adaptive threshold. In 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), pages 1879–1887. IEEE, 2018. 2

[7] Y. Huang, W. Wang, and L. Wang. Instance-aware image and sentence matching with selective multimodal lstm. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2310–2318, 2017. 2

[8] Y. Huang, Q. Wu, C. Song, and L. Wang. Learning semantic concepts and order for image and sentence matching. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 6163–6171, 2018. 2

[9] Z. Ji, S. Li, and Y. Pang. Fusion-attention network for person search with free-form natural language. Pattern Recognition Letters, 116:205–211, 2018. 2

[10] Z. Ji, H. Wang, J. Han, and Y. Pang. Saliency-guided attention network for image-sentence matching. arXiv preprint arXiv:1904.09471, 2019. 2

[11] K.-H. Lee, X. Chen, G. Hua, H. Hu, and X. He. Stacked cros sattention for image-text matching. In Proceedings of the European Conference on Computer Vision (ECCV), pages 201–216, 2018. 2

[12] S. Li, T. Xiao, H. Li, W. Yang, and X. Wang. Identity-aware textual-visual matching with latent co-attention. In Proceedings of the IEEE International Conference on Computer Vision, pages 1890–1899, 2017. 2

[13] S. Li, T. Xiao, H. Li, B. Zhou, D. Yue, and X. Wang. Person search with natural language description. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1970–1979, 2017. 2

[14] W. Li, R. Zhao, T. Xiao, and X. Wang. Deepreid: Deep filter pairing neural network for person re-identification. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 152–159, 2014. 2

[15] J. Marin, A. Biswas, F. Ofli, N. Hynes, A. Salvador, Y. Aytar, I. Weber, and A. Torralba. Recipe1m+: A dataset for learning cross-modal embeddings for cooking recipes and food images. IEEE transactions on pattern analysis and machine intelligence, 2019. 2

[16] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781, 2013. 3

[17] H. Nam, J.-W. Ha, and J. Kim. Dual attention networks for multimodal reasoning and matching. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 299–307, 2017. 2

[18] V.-H. Nguyen, T. D. Ngo, K. M. Nguyen, D. A. Duong, K. Nguyen, and D.-D. Le. Re-ranking for person re-identification. In 2013 International Conference on Soft Computing and Pattern Recognition (SoCPaR), pages 304–308. IEEE, 2013. 2

[19] X. Shen, Z. Lin, J. Brandt, S. Avidan, and Y. Wu. Object retrieval and localization with spatially-constrained similarity measure and k-nn re-ranking. In 2012 IEEE Conference on Computer Vision and Pattern Recognition, pages 3013–3020. IEEE, 2012. 2

[20] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In Proceedings of the European Conference on Computer Vision (ECCV), pages 480–496, 2018. 2

[21] G. Wang, Y. Yuan, X. Chen, J. Li, and X. Zhou. Learning discriminative features with multiple granularities for person re-identification. In 2018 ACM Multimedia Conference on Multimedia Conference, pages 274–282. ACM, 2018. 2

[22] J. Yu, Y. Lu, Z. Qin, W. Zhang, Y. Liu, J. Tan, and L. Guo. Modeling text with graph convolutional network for cross-modal information retrieval. In Pacific Rim Conference on Multimedia, pages 223–234. Springer, 2018. 2

[23] R. Yu, Z. Zhou, S. Bai, and X. Bai. Divide and fuse: A re-ranking approach for person re-identification. arXiv preprint arXiv:1708.04169, 2017. 2

[24] Z. Zheng, L. Zheng, M. Garrett, Y. Yang, and Y.-D. Shen. Dual-path convolutional image-text embedding with instance loss. arXiv preprint arXiv:1711.05535, 2017. 1, 2, 3, 4

[25] Z. Zhong, L. Zheng, D. Cao, and S. Li. Re-ranking person re-identification with k-reciprocal encoding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1318–1327, 2017. 2, 3, 4