Weakly Supervised Sketch Based Person Search

Lan Yan
State Key Laboratory for Management
and Control of Complex Systems,
Institute of Automation,
Chinese Academy of Sciences,
Beijing, China.
School of Artificial Intelligence,
University of Chinese Academy of Sciences,
Beijing, China.
yanlan2017@ia.ac.cn

Wenbo Zheng
School of Software Engineering,
Xi'an Jiaotong University,
Xi'an, China.
State Key Laboratory for Management
and Control of Complex Systems,
Institute of Automation,
Chinese Academy of Sciences,
Beijing, China.
zwb2017@stu.xjtu.edu.cn

Fei-Yue Wang
State Key Laboratory for Management
and Control of Complex Systems,
Institute of Automation,
Chinese Academy of Sciences,
Beijing, China.

Chao Gou*
School of Intelligent Systems Engineering,
Sun Yat-sen University,
Guangzhou, China.
gouchao@mail.sysu.edu.cn

ABSTRACT

Person search often requires a query photo of the target person. However, in many practical scenarios, there is no guarantee that such a photo is always available. In this paper, we define the problem of sketch based person search, which uses a sketch instead of a photo as the probe for retrieving. We tackle this problem in a weak supervision setting and propose a clustering and feature attention based weakly supervised learning framework, which contains two stages of pedestrian detection and sketch based person re-identification. Specially, we introduce multiple detectors, followed by fuzzy c-means clustering to achieve weakly supervised pedestrian detection. Moreover, we design an attention module to learn discriminative features in subsequent re-identification network. Extensive experiments show the superiority of our method.

CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence**; • **Information systems** → *Specialized information retrieval*; Data mining.

KEYWORDS

Sketch based person search, sketch based person re-identification, weakly supervised learning

ACM Reference Format:
Lan Yan, Wenbo Zheng, Fei-Yue Wang, and Chao Gou. 2021. Weakly Supervised Sketch Based Person Search. In *Proceedings of the 2021 International Conference on Multimedia Retrieval (ICMR '21), August 21–24, 2021, Taipei, Taiwan.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3460426.3463596

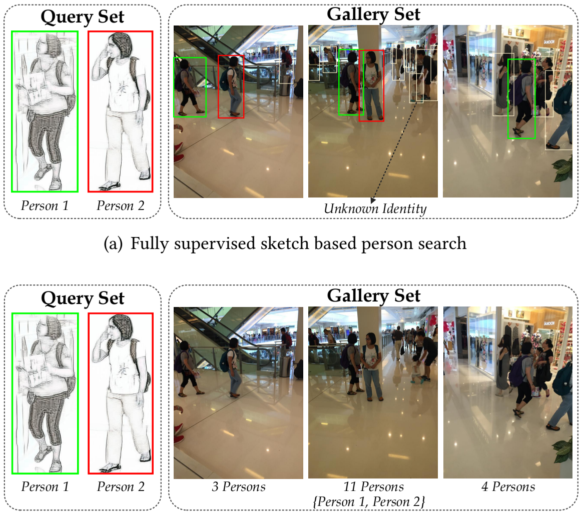
1 INTRODUCTION

Person search [29] aims at retrieving a target person from a gallery of unconstrained scene images. Although great progress has been made in recent years [3–8, 10, 12, 15, 20, 26, 27, 31, 37], a crucial issue that is often ignored is the availability of such a target photo. In practice, it is commonplace that useful suspect photographs cannot be easily acquired. This issue has long been cognizant by law enforcement, and has motivated studies on sketch based face recognition [33], which can match facial sketches drawn by professional artists based on the recollection of eyewitnesses to a facial photo database. In this work, we consider the case where the query photo is not available and only a sketch probe can be used as a substitute, and propose the **sketch based person search** problem for the first time. In sketch based person search, sketches (rather than photos) are used as query probes to search a target person in a whole scene photo gallery.

Sketch based person search has great practical value for law enforcement. Conceivably, when an outlaw is witnessed but not be photographed by a surveillance camera, legal agent or police officer can call on a professional artist to draw a sketch based on the witness's description. According to the sketch drawing, a sketch based person search system can automatically locate this outlaw by retrieving all the surveillance videos. This helps police and relevant law enforcement agencies quickly target those potential suspects. Once a suspect is caught on a surveillance camera, his movements and behaviors can be tracked and the witnesses around him can be successfully found, which saves a lot of manpower and material resources for the legal agent or police officer.

*Corresponding author.

ICMR '21, August 21–24, 2021, Taipei, Taiwan
© 2021 Association for Computing Machinery.
ACM ISBN 978-1-4503-8463-6/21/08…$15.00
https://doi.org/10.1145/3460426.3463596



(a) Fully supervised sketch based person search

(b) Weakly supervised sketch based person search

Figure 1: Illustration of two settings. Different colored bounding boxes represent different individuals.

Moreover, considering the tremendous labeled data requirement of fully supervised models as well as the high cost of data-labeling, in this work, we focus on the *weakly supervised sketch based person search*. Similar to the weakly supervised person search[30], the gallery set only provides the annotations regarding the presence of the identity in the set and the number of persons in each image, while the annotations suggesting in which image the identity appears and the pixel-level bounding box are not given. More specifically, as illustrated in Figure 1(b), the first picture in the gallery set is labeled with "3 Persons" implying that there are three pedestrians in this picture. The gallery set is annotated by a tag "{Person 1, Person 2}" suggesting that Person 1 and Person 2 are present in the set, while detail information about bounding box and which individual is Person 1 or Person 2 is not available. Thus, these labels are weak. Moreover, this weakly supervised setting is clearly inexact supervision [38], where the training process only provides coarse-grained labels. In such weak supervision setting, the labeling cost of sketch based person search can be sharply decreased and the scalability of models can be improved compared to the full supervision setting.

Under this setting, given a sketch query set of probe individuals, our goal is searching and locating these probe persons in the whole scene image gallery. To this end, we design a novel clustering and feature attention (CFA) based weakly supervised learning framework. Specifically, we use a two-step strategy, i.e., conducting pedestrian detection and sketch based person re-identification (Re-ID) separately. Multiple detectors are leveraged in pedestrian detection stage to offer richer bounding boxes. Then, the fuzzy c-means (FCM) clustering algorithm [1] is introduced to remove the false detection results. Subsequently, the bounding box photos are sent to a generator and translated to sketches. Considering the sparsity of sketches' pixel distribution, we design an attention module to attach more importance to discriminative features. Furthermore, we integrate it into the sketch based person Re-ID network to benefit the learning of discriminative features.

It is noteworthy that there is no publicly available sketch based person search dataset, but a sketch Re-ID dataset (PKUSketchRE-ID [22]) and two popular person search datasets (CUHK-SYSU [28] and PRW [36]) are usable. Therefore, we first train an image-to-image translation model MUNIT [13] on the PKUSketchRE-ID dataset and use the trained model to translate the images of the query set of CUHK-SYSU and PRW datasets into sketch. The obtained new datasets are named Sketch CUHK-SYSU and Sketch PRW. We conduct extensive experiments on these two new datasets and the results suggest the robustness and effectiveness of our method.

In summary, our main contributions are as follows:

1) To the best of our knowledge, this is the first attempt to both propose and tackle the problem of **sketch based person search** which has widely potential applications.

2) A novel weakly supervised learning framework termed as CFA with multiple detectors, fuzzy c-means, and attention mechanism is introduced to address the problem of sketch based person search.

3) Experimental results on the Sketch CUHK-SYSU and Sketch PRW datasets validate the effectiveness of the proposed method.

2 THE PROPOSED METHOD

In this section, our proposed framework is introduced. As evident in Figure 2, a panoramic image is first fed into K detector and the bounding boxes are obtained. Subsequently, the FCM clustering algorithm is used to cluster these bounding boxes, and the number of persons in the whole scene image (which is three in Figure 2) is adopted as the number of clusters. Once a bounding box has a low degree of membership for any class, it will be removed. After that, the rest bounding box photos are translated to sketches by a generator. Finally, these sketches are fed into the sketch based Re-ID network which can learn discriminative features under the guidance of two losses. Figure 2 also illustrates the testing procedure.

2.1 Detection and Clustering

Considering that training a detector from scratch under our weak supervision setting is impractical, we choose the existing preeminent detection network as our pedestrian detector. In addition, since the detection results of a single detector are not sufficient for the subsequent learning of an accurate sketch based Re-ID model, we apply multiple detectors to provide abundant detection results.

While multiple detectors give more bounding boxes, they also provide more incorrect candidate results. In addition, we cannot know the exact value of the wrong or correct bounding boxes, that is, they are fuzzy. Thus, a classic fuzzy clustering method, i.e., the FCM clustering algorithm, is adopted to acquire the membership degree of each bounding box sample to each cluster, and discard the samples with low membership degree. It can be formulated an objective function as:

$$J_q = \sum_{n=1}^{N} \sum_{m=1}^{M} (u_{nm})^q \left\|\mathbf{b}_n - \mathbf{c}_m\right\|^2, 1 \leq q < \infty \tag{1}$$

where M is the number of clusters, N is the number of all bounding box samples, $\mathbf{b}_n$ denotes the n-th bounding box sample, $\mathbf{c}_m$ is the center of the m-th cluster and u_{nm} is the membership of the n-th

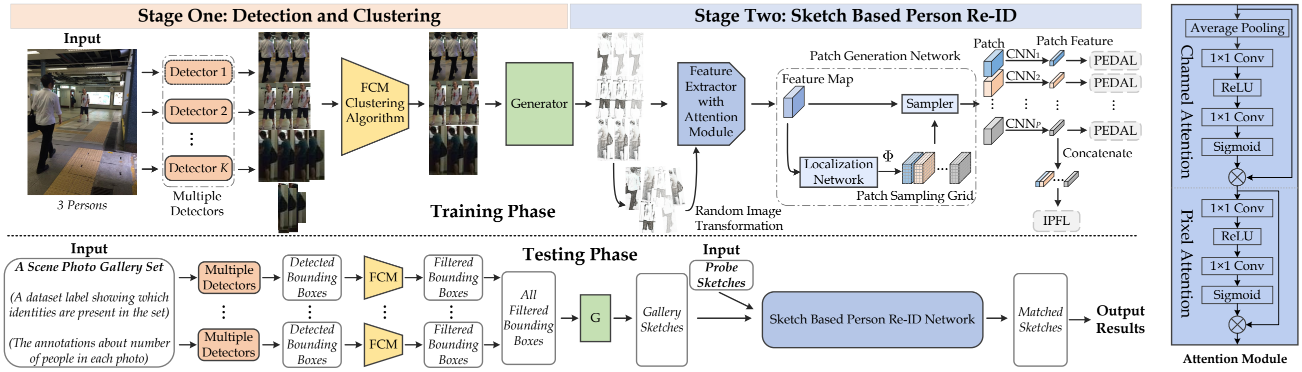


Figure 2: The proposed framework. IPEL [32] and PEDAL [32] are two loss functions. ⊗ means element-wise multiplication.

sample to the m-th cluster. q denotes a weighting exponent and in our experiments we set it as 2.

We can update and iterate the below two expressions to minimize above objective function:

$$\mathbf{c}_m = \frac{\sum_{n=1}^{N} (u_{nm})^q \mathbf{b}_n}{\sum_{n=1}^{N} (u_{nm})^q} \tag{2}$$

$$u_{nm} = \frac{1}{\sum_{j=1}^{M} \left(\frac{\|\mathbf{b}_n - \mathbf{c}_m\|}{\|\mathbf{b}_n - \mathbf{c}_j\|} \right)^{\frac{2}{q-1}}} \tag{3}$$

The iteration continues until $\left\| U^{(e+1)} - U^{(e)} \right\| < \varepsilon$ is satisfied. U is the membership matrix, e denotes the number of iterations. ε denotes iteration termination parameter and we set it as 1×10^{-6}.

Given the final membership matrix U^*, we can obtain the maximum membership degree u_{nm^*} of bounding box sample $\mathbf{b}_n$. The bounding box $\mathbf{b}_n$ will be removed, if u_{nm^*} is lower than a given threshold. The detection results of each image are clustered separately, and then the bounding boxes with membership degree lower than a certain threshold are discarded. The remainder are used to the sketch based person Re-ID network.

2.2 Sketch Based Person Re-ID

Since the probe sketches may be drawn by different artistes, that is, these sketches have a variety of artistic styles. We empirically and experimentally choose to pre-train MUNIT [13], rather than CycleGAN [39], UNIT [19] or NICE-GAN [6], on the PKUSketchRE-ID dataset which has five artistic styles. Subsequently, we employ the pre-trained MUNIT to translate the cropped images to sketches. Then, we aim to extract discriminative features from these sketches.

Similar with PAUL [32], we develop an unsupervised sketch based Re-ID network which mainly consists of a feature extractor and a patch generation network, as shown in Figure 2. Moreover, different from PAUL, we introduce attention mechanism and design an attention module including channel and pixel attention integrated into the feature extractor to learn discriminative features better.

Feature Attention. Our feature attention mechanism consists of channel attention which offers different weights to different channel features and pixel attention which attaches greater importance to discriminative features. As shown in Figure 2, firstly, the channel-wise global spatial information is obtained by global average pooling. Then, the feature passes through two convolution layers with kernel size 1×1, ReLU as well as sigmoid activation layer. The output is acquired by conducting element-wise multiplying between the weights of channel and input feature.

As for pixel attention, similar to the channel attention, the outputs of channel attention are fed into two convolutional layers with kernel size 1×1, ReLU and sigmoid activation layer. Finally, we use element-wise multiplication to acquire the output feature.

Loss Function. The patch-based discriminative feature learning loss (PEDAL) [32] is introduced to pull similar patches together and push the dissimilar patches away. Let $\mathbf{x}_i^p$ denote the p-th patch feature of i-th cropped sketch. During training, a memory bank W^p is maintained, where $W^p = \{\mathbf{w}_l^p\}_{l=1}^{L}$ and L represents the number of sketches sample:

$$\mathbf{w}_{l,t}^p = \begin{cases} (1-r) \times \mathbf{w}_{l,t-1}^p + r \times \mathbf{x}_{l,t}^p, & t > 0, \\ \mathbf{x}_{l,t}^p, & t = 0, \end{cases} \tag{4}$$

where t is the training epoch, $\mathbf{x}_{l,t}^p$ is the latest patch feature and $r \in [0,1]$ denotes the update rate. Subsequently, given $\{\mathbf{w}_l^p\}_{l=1}^{L}$, we can obtain the set of k nearest patches of $\mathbf{x}_i^p$, i.e., $\mathcal{K}_i^p$, by computing the l_2 distance between $\mathbf{w}_l^p$ and $\mathbf{x}_i^p$. Thus, PEDAL is defined by:

$$\mathcal{L}_{\text{PEDAL}}^p = -\log \frac{\sum_{\mathbf{w}_l^p \in \mathcal{K}_i^p} e^{-\frac{s}{2}\left\|\mathbf{x}_i^p - \mathbf{w}_l^p\right\|_2^2}}{\sum_{l=1, l \neq i}^{L} e^{-\frac{s}{2}\left\|\mathbf{x}_i^p - \mathbf{w}_l^p\right\|_2^2}} \tag{5}$$

where s is a scaling factor. Moreover, the patch feature learning loss (IPFL) [32], which harnesses all patch features of a sketch to provide sketch-level guidance, can be expressed as:

$$\mathcal{L}_{\text{IPFL}} = \max\{\|\mathbf{x}_i - \mathbf{p}_i\|_2 - \|\mathbf{x}_i - \mathbf{n}_i\|_2 + \eta, 0\} \tag{6}$$

where $\mathbf{p}_i$ and $\mathbf{n}_i$ represent the feature of a proxy positive sample and the hardest negative sample for $\mathbf{x}_i$, respectively. η denotes a margin of the loss.

Hence, the total loss function for sketch based person Re-ID network can be formulated as:

$$\mathcal{L} = \mathcal{L}_{\text{IPFL}} + \lambda \frac{1}{P} \sum_{P}^{p=1} \mathcal{L}_{\text{PEDAL}}^p \tag{7}$$

where λ controls the impact of the PEDAL.

3 EXPERIMENTS

In this section, we quantitative evaluate our method on both Sketch CUHK-SYSU and Sketch PRW datasets, whose sketch query sets are translated from the query images of CUHK-SYSU [28] and PRW [36] by a pre-trained MUNIT model, respectively. We follow OIM [28] and choose the mean Average Precision (mAP) and top-1 matching rate metric as performance indicators.

Experimental Setting. We select three off-the shelf detectors, including Faster-RCNN [23], Cascade R-CNN [2] and RetinaNet [17]. They all adopt ResNet-101 [11] with FPN as backbone network and are pre-trained on MS COCO [18] and then frozen. As shown in Figure 3, we experimentally and empirically set the threshold of membership to 0.8. We apply ImageNet-pretrained ResNet-50 as the backbone of feature extractor. We remove the last fully connected layer and insert an attention module after each residual block, along with set the stride of last residual block as 1. We use the MUNIT model trained on the PKUSketchRE-ID [22] dataset to transform DukeMTMC-reID [24] dataset into sketches, and leverage these sketches to pre-train our sketch Re-ID network. For the loss function, similar to [32], we set update rate r as 0.1, the scaling factor s as 5, λ and η are set to 2. During the sketch Re-ID network training, the sketches are resized to 384×128. SGD [25] optimizer is adopted with the momentum of 0.9. We initialize the learning rate at 0.0001 and decay it by 0.1 every 40 epochs. The Re-ID network is trained for 50 epochs with a batch size of 48.

Performance Comparison. Consider that there is no other weakly supervised sketch based person search method, we compare our approach with **fully supervised** person search approaches including OIM [28] and other methods combining different pedestrian detectors (LDCF [21], R-CNN [9]) and person descriptors (LOMO [16], DSIFT [34], BoW [35]) along with distance metric (XQDA [16], KISSME [14]). Not only are these methods trained in a fully supervised manner, but the query sets used in test are also photos. The comparison results are listed in Table 1 and Table 2.

From Table 1, as one weakly supervised method, our method can achieve a top-1 matching rate of 78.4% which is comparable with photo based fully supervised method of OIM and significantly outperforms other methods. Besides, from Table 2, we can obtain a similar conclusion. These consistently indicate the effectiveness and robustness of our method.

Ablation Study. To demonstrate the effectiveness of key components in our method, we consider different model configurations and perform ablation experiments. As shown in Figure 4, the methods with multiple detectors outperform the methods which only use a single detector by a large margin. As we can see in the top four lines of Figure 4, both FCM clustering algorithm and our attention module are effective to improve the model performance. Moreover, it is obvious that the performance keeps improving by adding each key component incrementally. This indicates that all these components are reasonable and effective, and combing them together can realize the maximum gain.

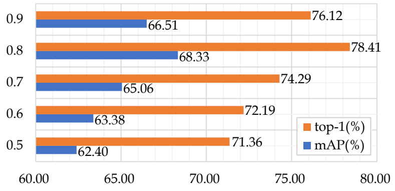


Figure 3: The impact of membership threshold.

Table 1: Performance comparison on the CUHK-SYSU dataset with a gallery size of 100.

Method	Type	mAP(%)	top-1(%)
CNN + DSIFT + Euclidean	Photo & Fully	34.5	39.4
CNN + DSIFT + KISSME		47.8	53.6
CNN + BoW + Cosine		56.9	62.3
OIM		**75.5**	**78.7**
Ours	Sktech & Weakly	*68.3*	*78.4*

Table 2: Performance comparison on PRW dataset.

Method	Type	mAP(%)	top-1(%)
LDCF + LOMO + XQDA	Photo & Fully	11.0	31.1
LDCF + IDEdet		18.3	44.6
LDCF + IDEdet + CWS		18.3	45.5
OIM		**21.3**	**49.9**
Ours	Sktech & Weakly	*19.2*	*49.7*

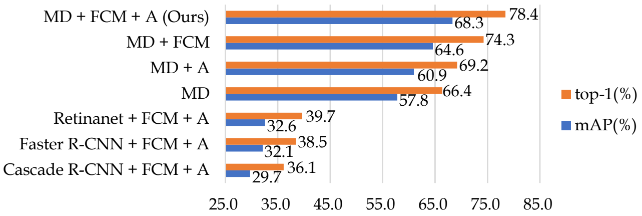


Figure 4: Evaluation results for different model configurations on Sketch CUHK-SYSU dataset with gallery size of 100. "MD" stands for multiple detectors. "FCM" stands for FCM clustering algorithm. "A" denotes attention module.

4 CONCLUSION

In this paper, we try to address the problem of **sketch based person search** in a weakly supervised setting. Specially, we design a new clustering and feature attention (CFA) based weakly supervised learning framework and introduce attention mechanism to improve the performance. Additionally, multiple detectors and FCM clustering algorithm are employed to provide great bounding box results and boost the performance of our model. Experimental results validate the effectiveness and robustness of our approach.

ACKNOWLEDGMENTS

This work is supported in part by National Key R&D Program of China (2020YFB1600400), in part by Key Research and Development Program of Guangzhou (202007050002) and in part by Shenzhen Science and Technology Program (RCBS20200714114920272).

REFERENCES

[1] James C Bezdek, Robert Ehrlich, and William Full. 1984. FCM: The fuzzy c-means clustering algorithm. *Computers & Geosciences* 10, 2-3 (1984), 191–203.

[2] Zhaowei Cai and Nuno Vasconcelos. 2018. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*. 6154–6162.

[3] Xiaojun Chang, Po-Yao Huang, Yi-Dong Shen, Xiaodan Liang, Yi Yang, and Alexander G Hauptmann. 2018. RCAA: Relational context-aware agents for person search. In *European conference on computer vision(ECCV)*. 84–100.

[4] Di Chen, Shanshan Zhang, Wanli Ouyang, Jian Yang, and Bernt Schiele. 2020. Hierarchical Online Instance Matching for Person Search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 10518–10525.

[5] Di Chen, Shanshan Zhang, Wanli Ouyang, Jian Yang, and Ying Tai. 2018. Person search via a mask-guided two-stream cnn model. In *European conference on computer vision(ECCV)*. 734–750.

[6] Runfa Chen, Wenbing Huang, Binghui Huang, Fuchun Sun, and Bin Fang. 2020. Reusing Discriminators for Encoding: Towards Unsupervised Image-to-Image Translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*. 8168–8177.

[7] Wenkai Dong, Zhaoxiang Zhang, Chunfeng Song, and Tieniu Tan. 2020. Bi-Directional Interaction Network for Person Search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*. 2839–2848.

[8] Wenkai Dong, Zhaoxiang Zhang, Chunfeng Song, and Tieniu Tan. 2020. Instance Guided Proposal Network for Person Search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*. 2585–2594.

[9] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*. 580–587.

[10] Chuchu Han, Jiacheng Ye, Yunshan Zhong, Xin Tan, Chi Zhang, Changxin Gao, and Nong Sang. 2019. Re-ID Driven Localization Refinement for Person Search. In *Proceedings of the IEEE international conference on computer vision(ICCV)*. 9814–9823.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*. 770–778.

[12] Zhenwei He and Lei Zhang. 2018. End-to-end detection and re-identification integrated net for person search. In *Asian Conference on Computer Vision*. Springer, 349–364.

[13] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. 2018. Multimodal unsupervised image-to-image translation. In *European conference on computer vision(ECCV)*. 172–189.

[14] Martin Koestinger, Martin Hirzer, Paul Wohlhart, Peter M Roth, and Horst Bischof. 2012. Large scale metric learning from equivalence constraints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*. 2288–2295.

[15] Xu Lan, Xiatian Zhu, and Shaogang Gong. 2018. Person search by multi-scale matching. In *European conference on computer vision(ECCV)*. 536–552.

[16] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z Li. 2015. Person re-identification by local maximal occurrence representation and metric learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*. 2197–2206.

[17] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision(ICCV)*. 2980–2988.

[18] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision(ECCV)*. Springer, 740–755.

[19] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. 2017. Unsupervised image-to-image translation networks. In *Advances in neural information processing systems*. 700–708.

[20] Bharti Munjal, Sikandar Amin, Federico Tombari, and Fabio Galasso. 2019. Query-Guided End-To-End Person Search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*. 811–820.

[21] Woonhyun Nam, Piotr Dollár, and Joon Hee Han. 2014. Local decorrelation for improved pedestrian detection. In *Advances in neural information processing systems*. 424–432.

[22] Lu Pang, Yaowei Wang, Yi-Zhe Song, Tiejun Huang, and Yonghong Tian. 2018. Cross-domain adversarial feature learning for sketch re-identification. In *Proceedings of the 26th ACM international conference on Multimedia*. 609–617.

[23] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*. 91–99.

[24] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. 2016. Performance measures and a data set for multi-target, multi-camera tracking. In *European conference on computer vision(ECCV)*. 17–35.

[25] Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. 2013. On the importance of initialization and momentum in deep learning. In *International conference on machine learning*. PMLR, 1139–1147.

[26] Cheng Wang, Bingpeng Ma, Hong Chang, Shiguang Shan, and Xilin Chen. 2020. TCTS: A Task-Consistent Two-Stage Framework for Person Search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*.

[27] Jimin Xiao, Yanchun Xie, Tammam Tillo, Kaizhu Huang, Yunchao Wei, and Jiashi Feng. 2019. IAN: the individual aggregation network for person search. *Pattern Recognition* 87 (2019), 332–340.

[28] Tong Xiao, Shuang Li, Bochao Wang, Liang Lin, and Xiaogang Wang. 2017. Joint detection and identification feature learning for person search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*. 3415–3424.

[29] Yuanlu Xu, Bingpeng Ma, Rui Huang, and Liang Lin. 2014. Person search in a scene by jointly modeling people commonness and person uniqueness. In *Proceedings of the 22nd ACM international conference on Multimedia*. 937–940.

[30] Lan Yan, Wenbo Zheng, Fei-Yue Wang, and Chao Gou. 2020. Weakly Supervised Person Search. In *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE, 188–196.

[31] Yichao Yan, Qiang Zhang, Bingbing Ni, Wendong Zhang, Minghao Xu, and Xiaokang Yang. 2019. Learning Context Graph for Person Search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*. 2158–2167.

[32] Qize Yang, Hong-Xing Yu, Ancong Wu, and Wei-Shi Zheng. 2019. Patch-Based Discriminative Feature Learning for Unsupervised Person Re-Identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*. 3633–3642.

[33] Wei Zhang, Xiaogang Wang, and Xiaoou Tang. 2011. Coupled information-theoretic encoding for face photo-sketch recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*. 513–520.

[34] Rui Zhao, Wanli Ouyang, and Xiaogang Wang. 2013. Unsupervised salience learning for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*. 3586–3593.

[35] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. 2015. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision(ICCV)*. 1116–1124.

[36] Liang Zheng, Hengheng Zhang, Shaoyan Sun, Manmohan Chandraker, Yi Yang, and Qi Tian. 2017. Person re-identification in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*. 1367–1376.

[37] Yingji Zhong, Xiaoyu Wang, and Shiliang Zhang. 2020. Robust partial matching for person search in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*. 6827–6835.

[38] Zhi-Hua Zhou. 2017. A brief introduction to weakly supervised learning. *National Science Review* 5, 1 (2017), 44–53.

[39] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networkss. In *Proceedings of the IEEE international conference on computer vision(ICCV)*. 8168–8177.