Deep Learning-based Person Search with Visual Attention Embedding

Liviu-Daniel Ştefan[1], Şeila Abdulamit[1], Mihai Dogariu[1], Mihai Gabriel Constantin[1], and Bogdan Ionescu[1]
[1] University Politehnica of Bucharest, Romania
Contact author e-mail: stefan.liviu.daniel@gmail.com

***Abstract*—In this work, we consider the problem of person search, which is a challenging task that requires both person detection and person re-identification run concurrently. In this context, we propose a person search approach based on deep neural networks that incorporates attention mechanisms to perform retrieval more robustly. Global and local features are extracted for person detection and person identification, respectively, boosted by attention layers that allow the extraction of discriminative feature representations, all in an end-to-end manner. We evaluate our approach on three challenging data sets and show that our proposed method improves the state-of-the-art networks.**

***Index Terms*—person search, person detection, person re-identification, visual attention layers.**

I. Introduction

Person search is a relatively new image retrieval task with increasing attention in the computer vision community [1], [2]. Given a query image, person search aims to localize the specific person from the given query in a gallery of images. It is an extension of the classical person re-identification classification problem [3], [4], which is based on two assumptions: (i) the spatial coordinates of each person from the gallery are given, and (ii) the spatial coordinates are perfectly aligned. However, in practice, these two assumptions do not hold. In this context, person search is more challenging, as it requires both person detection and re-identification, concurrently. Furthermore, aside from the errors introduced by the changes in illumination, camera viewpoint, background, and occlusions, that are generally considered influential problems for the person re-identification task, other factors need to be considered for the person search problem that deeply affects the retrieval quality. These include miss-alignments, miss-detections, false alarms or small changes in person's appearance and clothing attributes, across the query and the images collection. We show the differences between person search and person re-identification in Figure 1.

To date, only a few methods have been proposed to address person search [5]–[8]. Typical methods divide the problem into two categories: person detection and person re-identification, and tackle each task sequentially, via a two-stage strategy based on separate supervised algorithms. For the former, person detection is achieved by densely scanning the image in a sliding window fashion [9]–[11] or by using a proposal mechanism and leveraging CNNs to classify a sparsified set of proposals [12]–[14]. For the latest, different metric learning methods are employed to learn an embedding space that

Fig. 1: Demonstration of the search process for one gallery image for the person re-identification (a), and person search (b), problems. The red boxes indicate the wrong matched results while the green box represents the truly matched person.

cluster images corresponding to the query by minimizing the intra-person distance while maximizing the inter-person distance [15], [16], or different classification set-up [17]–[19].

In the last years, deep neural networks have become increasingly predominant choices for person localization and re-identification. One of the consolidated findings of state-of-the-art deep learning architectures [20]–[22] is that they are able to learn invariant deep embeddings, with superior representation capabilities for high-dimensional data, and jointly, with their classifying capabilities, dramatically outperforming conventional descriptors and classifiers.

In this article, we take advantage of the deep learning advancements, and propose an end-to-end person search framework that integrates multiple DNN architectures. Specifically, given the whole scene, and the query image, we leverage a person proposal network for person candidate detection to extract the proposal locations of the persons. Then, a re-identification network extracts discriminant visual representations between the query and each proposal to retrieve the results. It is an improvement of the work in [5] via the use of visual attention.

The remainder of the paper proceeds as follows. We first position our work in the literature, discussing related approaches and concepts, in Section 2. Then, we present our

proposed end-to-end person search system, in Section 3, by examining three types of CNNs used in constructing the framework. Experimental setup, and the analysis of the results are presented in Section 4 and Section 5, respectively. Lastly, Section 6 presents our conclusions and discusses future work.

II. Previous work

Considering that our person search framework is composed of two stages: (i) person detection, and (ii) person re-identification, we first review the advances in both fields. Afterward, we review existing works on person search, which is a recently proposed topic.

A. Person Detection

Person detection is canonical object detection, an extensively studied field over the past few decades, that became a testbed for the leading deep convolutional neural networks. In this context, the region-based CNN [12] and feature pyramid networks [23] can be considered milestones for object detection, due to the immense detection performance they achieved on this task, over the traditional methods such as the deformable part model (DPM) [24].

The Region-based Convolutional Neural Networks (R-CNNs) have become very successful due to their low cost that resulted trough sharing convolutions across proposals. It is based on a bottom-up grouping and saliency cues selective search method that reduces the cost and the searching space in object detection. The next iteration, Fast R-CNN [25], further improved the results using very deep networks, by performing a single convolution operation per image, instead of performing for each region proposal. Finally, in order to have an almost cost-free region proposals prediction, Faster R-CNN was proposed in [26] consisting of two models, namely an RPN and a Fast R-CNN model, where the first generates object proposals which are fed into the Fast R-CNN, and the latest refine the proposals with the goal of increasing their quality, leading to an overall increased object detection accuracy.

Other object detectors, such as Feature Pyramid Network (FPN) [27] allow for the detection of objects at different scales. In FPN, the network builds a multi-scale feature pyramid, with each level of the pyramid being able to detect objects at different scales.

B. Person Re-identification

Person re-identification addresses the problem of searching specific persons across spatially non-overlapping cameras, by estimating visual similarities between different probe-gallery pairs. Various data sets [2]–[4], [28], [29] have been proposed to support the research of the re-identification problem. Existing person re-identification methods focus either on manually or automatically designing discriminative features [16], [17], [30], [31] for representing person images, or designing a learning distance metrics [32]–[36] for measuring similarity between person images. These include novel deep architectures [15], [16] that combines both of the practices.

For instance, Song et al. [37] proposes a domain generalizable person re-identification model that maps the image with its identity classifier, bypassing the need of updating the model for the target domain. Wang et al. [38], proposed a deep network that tackles the misalignments and color differences across camera problems. The authors in [39] proposed a weakly-supervised framework that matches a person with untrimmed video data. Wu et al. [40] developed a Multi-teacher adaptive similarity distillation framework that uses lightweight models to reduce the testing computation. The framework uses multiple teacher-single student settings and proposes an adaptive knowledge aggregator to measure the teachers contributions, achieving performances comparable to state-of-the-art unsupervised and semi-supervised Re-ID methods.

C. Person Search

Person search is a relatively new challenge that consists of both pedestrian detection and person re-identification to run concurrently, in a coherent system. In this context, Xu et al. [41] firstly introduced the problem of person search in images using human body appearance, in contrast to the existing works on people detection and person re-identification. Xiao et al. [5] further proposed the first end-to-end person search framework based on body and parts regions in order to fully understand the pedestrian representations. Munjal et al. [7] showed the benefits of end-to-end optimization by proposing QEEPS, a query-guided person search for online instance matching via a query-guided Siamese squeeze-and-excitation network.

In this paper, we present an end-to-end trainable deep neural network for person search. The contribution beyond state of the art can be summarized with the following: (i) we integrate attention mechanisms in the stem CNN to train the network to attend to representative parts in pedestrain patches. This allow the network to focus on discriminative regions, e.g., face and body parts, or on different accessories such as glasses, bags, etc.; (ii) we perform spatial transformations to increase the robustness of the system to spatial variances.

III. Proposed method

In this section, we provide a detailed description of the proposed person search network. In addition, we study, test, and integrate attention layers in the architecture, to improve the predictive power of the algorithm. The methodology for performing person search in an unified way, consists of three parts, namely: (i) *global features*, represented by low-level features extracted from the whole input image, (ii) *region proposals* for transforming low-level features into pedestrian proposals, and (iii) *local features* represented by discriminative features corresponding to the identities in the image.

For the backbone of our architecture, we have implemented and tested three popular deep neural networks (DNNs) from the literature, namely GoogleNet [20], ResNet50 [21], and DenseNet101 [42]. For reproducibility purposes, we have selected the best performer network as the base of the framework

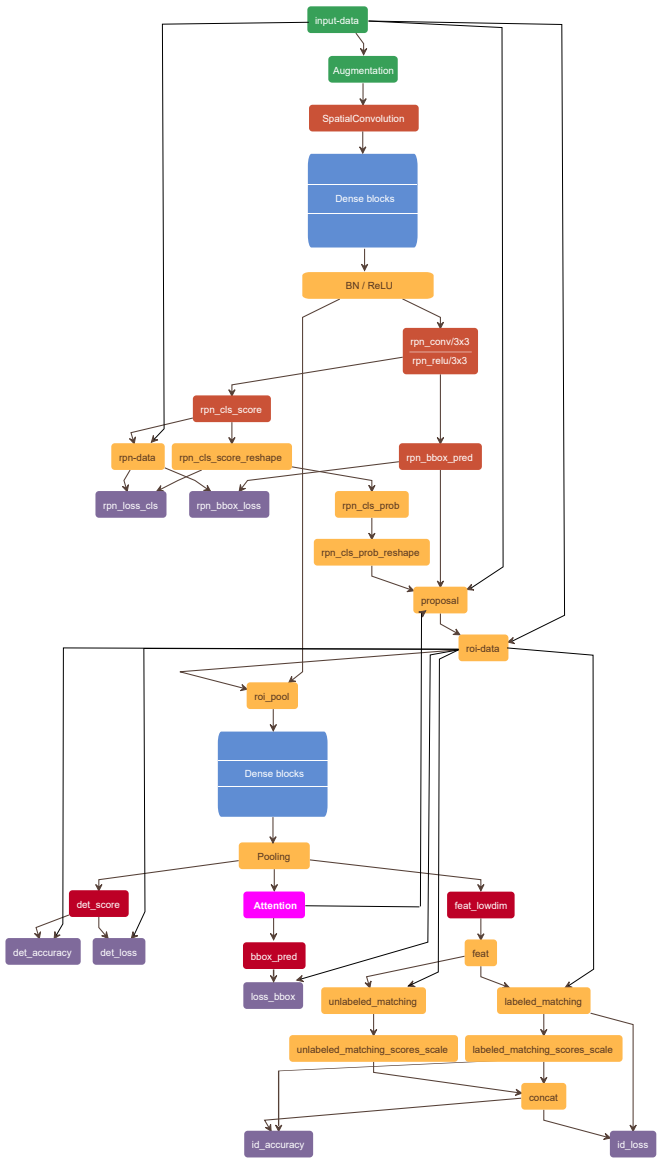

Fig. 2: Illustration of the unified person search architecture. The framework extracts low-level features via a set of 3 dense blocks with 6, 12, and 24 dense layers respectively, that are then transformed into pedestrian proposals and finally fed to an identification network based on a dense block with 16 layers which extracts down-sampled representative features of the identity.

in Figure 2. The same topology for connecting the models was used for the two others base DNNs architectures.

Global features. Global features represent a generalization of a whole image, revealing contour and shape representations of the objects in the scene, allowing us to discern between different objects, and background. Given as input an RGB scene image of height and width multiples of 32, we first perform an initial convolution and max-pooling with 7×7 and 3×3 kernel sizes, followed by a stem CNN composed of 3 dense blocks with 6, 12, and 24 dense layers respectively, with a growth rate of 32. Each dense block is an iterative concatenation of previous feature maps so that each layer has direct access to the gradients from the loss function and the original input signal. The stem CNN will produce 512 channels of features maps, which have 1/16 resolutions of the input image. Then, spatial transformations are applied to the generated features to increase the robustness of the system to spatial variances.

Region proposals. Region proposals represent candidate spatial coordinates of the objects in the scene. We follow the methodology described in [5] to transform feature maps in spatial coordinates by building a pedestrian proposal network to detect person candidates. We achieve this by training a SoftMax classifier to discern whether the features, mapped to a set of 9 anchors of sizes 128×128, 256×256, 512×512 with height/width ratios of 1:1, 1:2 and 2:1 respectively, represent a person or not. In addition, we use a linear regressor to further refine the respective locations using the protocol described in [43]. Therefore, the multi-task loss function accommodates the losses of classification and bounding box regression:

$$L = L_{cls} + L_{reg} \tag{1}$$

The loss function for an image is described by the following equation:

$$\begin{aligned} L\left(\{p_i\},\{t_i\}\right) = \frac{1}{N_{cls}} \sum_i L_{cls}\left(p_i, p_i'\right) + \\ \frac{\lambda}{N_{reg}} \sum_i p_i' \times L_1^{smooth}\left(t_i - t_i'\right) \end{aligned} \tag{2}$$

where pi represents the predicted probability of anchor i to be an object, p_i' represents the ground truth label of the anchor i, t_i represents the predicted spatial coordinates of the object i, t_i' represents the ground truth spatial coordinates for the object i, N_{cls} and N_{reg} represent the normalization term set to a mini-batch and to the number of anchor locations (∼2,400), respectively, and λ represents a weighted parameter for L_{cls} and L_{reg}, set to be 10. L_1^{smooth} is the smooth L_1 loss.

The log loss function over two classes L_{cls}, predicting a sample being a target object or background is computed as following:

$$L_{cls}\left(p_i, p_i'\right) = -p_i' \times log_{p_i} - \left(1 - p_i'\right) \times log(1 - p_i') \tag{3}$$

Local features. To discern between identities, we further extract local features from pedestrian image patches generated by the region proposal network. In this context, the predictions represented by the spatial coordinates, are fed into an RoI-Pooling layer by pooling a $1,024 \times 14 \times 14$ region for each proposal and then, we pass the predicted proposal to an identification network. The identification network is composed of a dense block with 16 dense layers, with a growth rate of 32. These are then followed by a global average pooling layer which outputs a 2,048 dimensional features vector representing the individual features of the identity.

To further improve the precision of the architecture, we include an attention mechanism before the identification network. The integrated attention mechanism parameters are

learned throughout the end-to-end training, helping the networks to focus on key elements of the input. In this regard, we have selected a soft (stochastic) attention setting, where the mask of values is constrained to be between 0 or 1, therefore, ignoring non-discriminant features, and training the neural network to attend to specific parts of the input. Finally, the identification is supervised using the OIM loss [5].

IV. Experimental Setup

We evaluate our proposed method on three large-scale end-to-end person detection and re-identification benchmarks, namely: Person Re-identification in the Wild [2], CUHK03 [4], and CUHK-SYSU [5].

A. Data sets

Person Re-identification in the Wild (PRW) contains 34,304 images of 932 identities (avg. 36.8 per identity), and 8,806 images of distractors, captured by six cameras.

CUHK03 contains 14,097 images of 1,467 person recorded with six surveillance cameras with each identity captured by two disjoint camera views (avg. 4.8 images in each view).

CUHK-SYSU contains 18,184 images of 8,432 identities (avg. 11.4 images per identity) recorded from two data sources, namely, from point-and-shoot cameras capturing street snaps around an urban city, and from movies with pedestrians.

B. Implementation Details

Our framework is implemented with the pytorch [44] framework and integrates the py-faster-rcnn repository [26], using the parameters values proposed by the authors. In the implementation process, we study and test a set of good practices for training DNNs. Specifically, we implement a set of data augmentation techniques for images, i.e., contrast changes, hue/saturation, affine transformations, perspective transformations changes, blurring, gaussian noise, dropout of regions, cropping/padding, and perform transfer learning, by using the ImageNet-pretrained models for parameters initialization, with the goal of improving the predictive power of the proposed algorithm.

The framework takes input images resized to have at least 900 $\times$ 1,500 pixels either on the short side or the long side. Then, the detection image patches are re-scaled to 256 $\times$ 128. In the training stage, the learning rate is initialized to 0.001, and dropped to 0.0001 after 25 epochs, until the model converges at 50 epochs. The optimization algorithm is the stochastic gradient descent (SGD) with the Softmax loss and Smooth L1 loss for detection, whereas for the identification, we use the OIM [5] loss function with a circular size of 2,048. In both training and testing, a detected bounding box is considered correct if the Intersection over Union (IoU) score with the ground truth bounding box is bigger than 0.75.

C. Evaluation

For each data set, we adopted the original evaluation protocol that the data set provides, namely the mean average precision (mAP). Finally, the results are obtained in a single-query setting, without re-ranking.

TABLE I: Effectiveness of the pedestrian detection module expressed in AP (IoU $>$ 0.75).

Method	PRW	CUHK03	CUHK-SYSU
GoogleNet	0.755	0.818	0.764
ResNet50	0.794	0.845	0.776
DenseNet121	0.847	0.862	0.792

TABLE II: Effectiveness of the person search framework expressed in mAP (IoU $>$ 0.75).

Method	PRW	CUHK03	CUHK-SYSU
GoogleNet	0.263	0.668	0.685
ResNet50	0.327	0.714	0.753
DenseNet121	0.335	0.703	0.778
$GoogleNet_{att}$	**0.281**	**0.694**	**0.692**
$ResNet50_{att}$	**0.347**	**0.721**	**0.775**
$DenseNet121_{att}$	**0.358**	**0.717**	**0.783**

V. Results and Discussion

We compare our proposed person search framework (with or without using attention mechanisms) using three consecrated deep networks, namely GoogleNet, ResNet50, and DenseNet121 as the backbone of the framework, to evidentiate their utility in person search. The results are summarized in Table II. Comparing the baseline models, DenseNet121 outperforms the two other networks on two data sets out of three, namely on PRW and CUHK-SYSU, with an mAP score of 0.335 and 0.778, respectively. The best performer on the CUHK03 data set is ResNet50 with an mAP score of 0.714. Furthermore, we observe that the trend is valid also for the variants with attention mechanisms, with DenseNet121ATT achieving an mAP score of 0.358 and 0.783 on the PRW and CUHK-SYSU data set, and ResNet50, with an mAP of 0.721 on the CUHK03.

Analyzing the results with respect to attention mechanism for the best performers on each data set, we can observe that $DenseNet121_{att}$ achieved a boost in performance from 0.335 to 0.358 and from 0.778 to 0.783 on the PRW and CUHK-SYSU data sets, respectively. On the CUHK03 data set, $ResNet50_{att}$ increased from 0.714 to 0.721. These results indicate the advantage of our proposed framework as the attention mechanisms constantly outperform the baseline variants on all three data sets.

When building a person search system, an important question needs to be addressed, namely: *How does the detector performance affect the overall performance?* In this context, we analyze the detection precision impact on person search. Intuitively, better detector results advocate a higher overall accuracy. Table I presents the results achieved by the GoogleNet, ResNet50 and DenseNet121 networks. We can observe that the detection accuracy is consistent with the person search performance evaluated using the IoU $>$ 0.75 criterion, with DenseNet121 achieving the best results with an AP score of 0.847, 0.862 and 0.792 on the PRW, CUHK03 and CUHK-SYSU data sets, respectively.

VI. Conclusions

In this paper, we have presented a unified framework for person search based on deep neural networks. In this context, we have tested three popular DNNs namely, GoogleNet, ResNet50, and DenseNet121, as the backbone for the entire framework, enhanced with attention layers to furthe improve the predictive capabilities of the proposed approach. In the training phase, we have also included a set of good practices for training deep networks targeting image processing and transfer learning. Extensive experiments show that the attention mechanism consistently improves the overall performance of the system, achieving a mAP score of 0.358, 0.721, and 0.783 on the PRW, CUHK03, and CUHK-SYSU data sets, respectively. In future work, we will extend the attention mechanisms in the detection stage, to better cope with occlusions and perform retrieval more robustly.

References

[1] C. C. Loy, D. Lin, W. Ouyang, Y. Xiong, S. Yang, Q. Huang, D. Zhou, W. Xia, Q. Li, P. Luo *et al.*, "Wider face and pedestrian challenge 2018: Methods and results," *arXiv preprint arXiv:1902.06854*, 2019.

[2] L. Zheng, H. Zhang, S. Sun, M. Chandraker, Y. Yang, and Q. Tian, "Person re-identification in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1367–1376.

[3] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1116–1124.

[4] 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*, 2014, pp. 152–159.

[5] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang, "Joint detection and identification feature learning for person search," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3415–3424.

[6] 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*, 2017, pp. 1970–1979.

[7] B. Munjal, S. Amin, F. Tombari, and F. Galasso, "Query-guided end-to-end person search," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 811–820.

[8] H. Liu, J. Feng, Z. Jie, K. Jayashree, B. Zhao, M. Qi, J. Jiang, and S. Yan, "Neural person search machines," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 493–501.

[9] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[10] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," *arXiv preprint arXiv:1312.6229*, 2013.

[11] S. Zhang, R. Benenson, B. Schiele *et al.*, "Filtered channel features for pedestrian detection." in *CVPR*, vol. 1, no. 2, 2015, p. 4.

[12] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

[13] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[14] J. Hosang, R. Benenson, and B. Schiele, "Learning non-maximum suppression," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4507–4515.

[15] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," *arXiv preprint arXiv:1703.07737*, 2017.

[16] W. Chen, X. Chen, J. Zhang, and K. Huang, "Beyond triplet loss: a deep quadruplet network for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 403–412.

[17] C. Su, J. Li, S. Zhang, J. Xing, W. Gao, and Q. Tian, "Pose-driven deep convolutional model for person re-identification," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 3960–3969.

[18] Z. Zhong, L. Zheng, Z. Luo, S. Li, and Y. Yang, "Invariance matters: Exemplar memory for domain adaptive person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 598–607.

[19] C.-P. Tay, S. Roy, and K.-H. Yap, "Aanet: Attribute attention network for person re-identifications," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7134–7143.

[20] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[22] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[23] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.

[24] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 9, pp. 1627–1645, 2009.

[25] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1440–1448.

[26] R. G. J. S. Shaoqing Ren, Kaiming He, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems*, 2015, pp. 91–99.

[27] R. G. K. H. B. H. S. B. Tsung-Yi Lin, Piotr Dollár, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.

[28] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in *European Conference on Computer Vision*. Springer, 2016, pp. 17–35.

[29] S. Bak, P. Carr, and J.-F. Lalonde, "Domain adaptation through synthesis for unsupervised person re-identification," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 189–205.

[30] D. Chen, D. Xu, H. Li, N. Sebe, and X. Wang, "Group consistent similarity learning via deep crf for person re-identification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8649–8658.

[31] D. Li, X. Chen, Z. Zhang, and K. Huang, "Learning deep context-aware features over body and latent parts for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 384–393.

[32] L. Zhang, T. Xiang, and S. Gong, "Learning a discriminative null space for person re-identification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1239–1248.

[33] M. F. T Ali and S. Chaudhuri, "Maximum margin metric learning over discriminative nullspace for person re-identification," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 122–138.

[34] Y. Zhai, X. Guo, Y. Lu, and H. Li, "In defense of the classification loss for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 0–0.

[35] S. Paisitkriangkrai, C. Shen, and A. Van Den Hengel, "Learning to rank in person re-identification with metric ensembles," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1846–1855.

[36] D. Chen, Z. Yuan, G. Hua, N. Zheng, and J. Wang, "Similarity learning on an explicit polynomial kernel feature map for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1565–1573.

[37] Y.-Z. S. T. X. T. M. H. Jifei Song, Yongxin Yang, "Generalizable person re-identification by domain-invariant mapping network," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 719–728.

[38] F. W. G. W. Yicheng Wang, Zhenzhong Chen, "Person re-identification with cascaded pairwise convolutions," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 1470–1478.

[39] W.-S. Z. Jingke Meng, Sheng Wu, "Weakly supervised person re-identification," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 760–769.

[40] X. G. J.-H. L. Ancong Wu, Wei-Shi Zheng, "Distilled person re-identification: Towards a more scalable system," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 1187–1196.

[41] Y. Xu, B. Ma, R. Huang, and L. Lin, "Person search in a scene by jointly modeling people commonness and person uniqueness," in *Proceedings of the 22nd ACM international conference on Multimedia*, 2014, pp. 937–940.

[42] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[43] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[44] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," 2017.