Person Search via Background and Foreground Contrastive Learning

1st Qing Tang, 2nd Kang-Hyun Jo
Department of Electrical, Electronic and Computer Engineering
University of Ulsan
Ulsan, Korea
tangqing@islab.ulsan.ac.kr; acejo@ulsan.ac.kr

***Abstract*—The specific person search is the foundation of a wide range of applications in intelligent security and surveillance systems. Although detection and re-id have been widely studied, they are difficult to apply to practical applications directly. Therefore, this paper focuses on person search, which aims to solve person detection and person re-identification (re-id) jointly. The common practice is to append the standard detection loss and re-id branches parallelly on Faster RCNN. The traditional re-id utilized Online Instance Matching (OIM) to pull a sample closer to its identity class. However, the relationship among RoIs of an image has not been fully explored in previous methods. To address this issue, we propose Background and Foreground Contrastive Loss (BFCL) to further boost re-id performance. We consider that RoIs from one image have a high probability of containing similar patterns, which might disturb the re-id performance. Therefore, we proposed BFCL to strengthen the learning of distinguishing similar background and foreground by leveraging inter-RoIs pairwise similarity. In summary, our method jointly optimizes the regression loss, classification loss, re-id loss, and the proposed BFCL for achieving optimal performances in person search model. Experiments are performed on two large-scale person search datasets, CUHK-SYSU and PRW. Results show that the proposed BFCL consistently boosts the performance of the baseline framework SeqNet in two datasets. The improved results demonstrate the effectiveness of the proposed BFCL and the necessity of exploring the relationship among RoIs.**

***Index Terms*—Person search, contrastive learning**

I. Introduction

Intelligent security and surveillance systems have become an active research area in recent years because of the increasing demand for public safety, the widespread camera network in public places, the expensive human labor, and the growing practicability of computer vision in the industry. As the foundation of a wide range of applications in intelligent security and surveillance systems, person search has drawn considerable attention recently with the increasing demand for person re-identification (re-id) in real-world applications, such as the specific person searching, multi-object multi-camera tracking [1], and human activity analysis [2].

The person re-id aims to retrieve images containing the same identity. Although extensive person re-id methods [3]–[7] have been proposed, recent researchers [8]–[10] found out that there is still a big gap between the person re-id system setting and real-world application. The person re-id systems are trained using well-cropped images, however, person detectors might produce wrong-cropped images in practical applications. To close the gap, recent researches [8]–[13] tend to solve person detection and re-id jointly, namely person search.

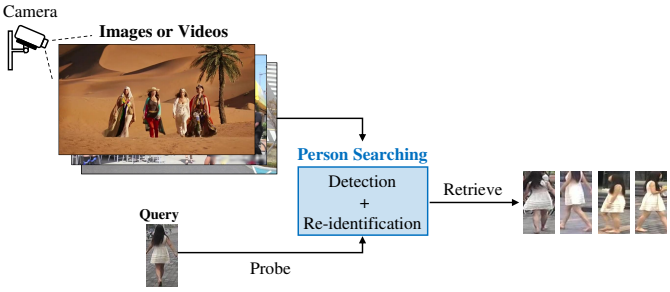


Fig. 1. Illustration of the person searching system.

An illustration of the person search system is shown in Fig. 1. The person search system aims to detect the specific person regions from realistic and uncropped images. Then, based on detected person regions, the system retrieves the specific person regions that contained the same identity as a query image by matching detected regions with query images.

The person search can be considered as an integrated task of person detection and person re-identification. The existing person search framework can be summarized into two categories: two-step framework and end-to-end framework. Two-step methods [14], [15] tackled the detection and re-id with two separate models. End-to-end methods [9]–[13] unified detection and re-id tasks in one model by attaching the original detection and re-id branches parallelly. In general, two-step methods yield better performance but they are time-consuming and heavy, and end-to-end methods are faster and simpler but they can not obtain satisfactory re-id results. It is because the inconsistent objectives [1], [14], [16], [17] between detection and re-id. More specifically, detection tends to produce similar features for person regions to distinguish them from backgrounds, but re-id tends to produce different features for person regions to further subdivide them into identities.

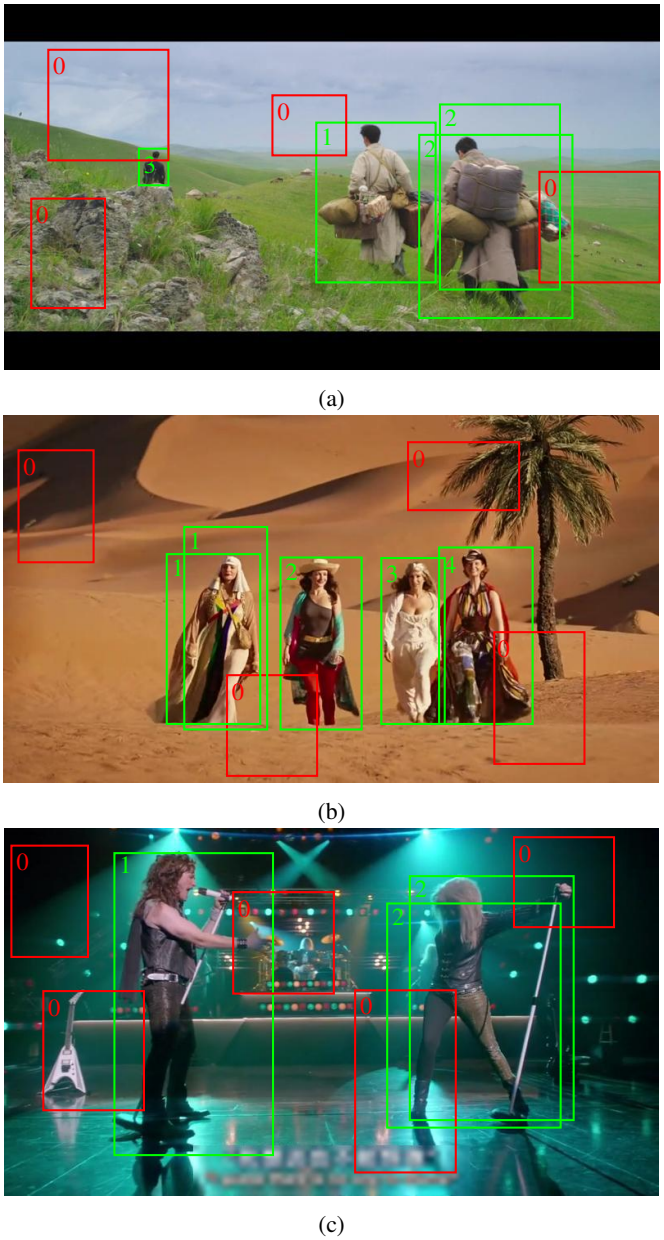
Chen et al. [14] first revealed the above goal conflict between the detection and re-ID. They argued that sharing features between the detection and re-ID tasks is not appropriate and therefore two-step methods yield better performance

than end-to-end methods. Chen et al. presented a Mask-Guided Two-Stream (MGTS) method to eliminate the conflict. Wang et al. [18] considered the consistency between detection and re-ID stages and introduced a Task-Consist Two-Stage (TCTS) framework. Recent end-to-end works also start to tackle the goal conflict between detection and re-id and further improve the person search performance. Chen et al. [13] proposed a Norm-Aware Embedding to decompose embedding into norm and angle for detection and re-id respectively. Li et al. [12] proposed a Sequential End-to-end Network (SeqNet), which employed an extra detection head to provide high-quality Region of Interests (RoIs) and embedding for re-id.

Although previous methods achieved good results, the relationship among RoIs of an image has not been explored. Intuitively, an input image has its own characteristic patterns and therefore RoIs from the image have high probabilities of containing similar patterns. Three examples are illustrated in Fig. 2. The RoIs in Fig. 2(a) mainly contain sky, grassland, or stone. RoIs in Fig. 2(b) contain desert. RoIs in Fig. 2(c) contain a stage with green light. Moreover, person regions (foregrounds) from the same image are more difficult to classify than person regions from different images because of similar patterns. To address this issue, we propose Background and Foreground Contrastive Loss (BFCL) to further boost re-id performance by leveraging inter-RoIs pairwise similarity. With the help of BFCL, the person search model is able to differentiate similar RoIs for re-id.

The contributions could be summarized as four-fold. (1) Although CQ is widely used in previous works [9], [12], [13], we found out CQ cannot yield consistent gains in two datasets. (2) The proposed BFCL helps the model learn more discriminative features of the foreground. (3) The proposed BFCL can exploit the unlabeled identities without building a storage-consuming memory bank as [9]. (4) Our proposed BFCL shows exceptionally strong performances.

The remainder of this paper is organized as follows. Section II describes the proposed method. In section III the dataset, the implementation details, and extensive experimental results are reported and analyzed. Finally, Section IV concludes this paper.



(a) (b) (c)

Fig. 2. Examples of backgrounds (red boxes) and foregrounds (green boxes) RoIs in (a)-(c) three different input images. The number at the top left corner represents the identity i. $i = 0$ indicates background, and $i > 0$ indicates foreground. Different i means different identity.

II. Proposed Methods

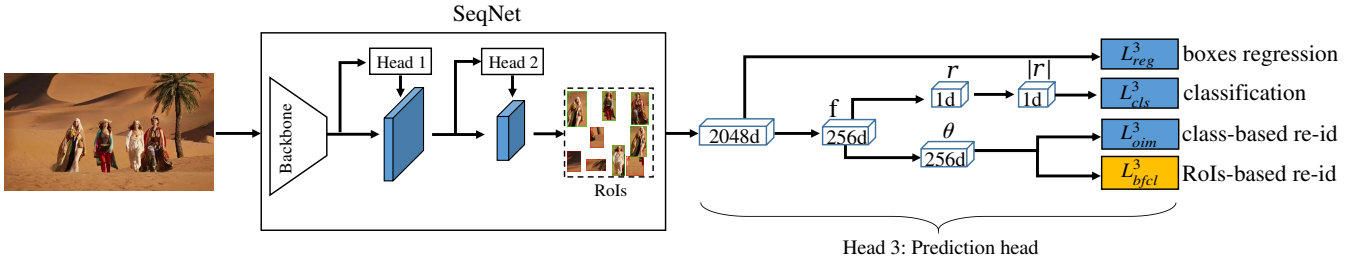
In this section, we revisit the end-to-end person search network SeqNet [12], which is the baseline framework of our work. Then, we introduce the overview of person search architecture. At last, he proposed Background and Foreground Contrastive Loss (BFCL) in detail.

A. Architecture Overview

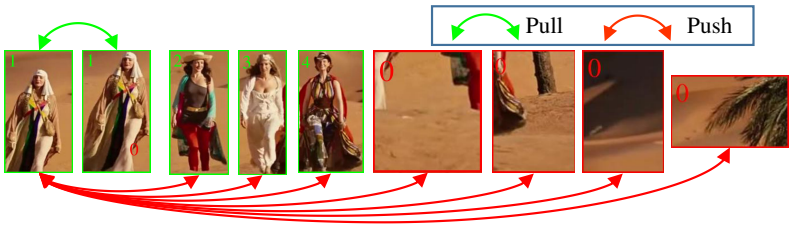
SeqNet is the baseline framework of our work. The architecture of the SeqNet with our proposed BFCL is illustrated in Fig. 3(a). SeqNet extracts the $2048d$ features using Faster RCNN [19], which contains a backbone network ResNet50 [20], a Region Proposal Network (RPN).

During training step, there are four loss in SeqNet, i.e., L^1_{reg}, L^1_{cls}, L^2_{reg}, L^2_{cls}. Superscripts ^{1} and ^{2} indicate the first and second head of SeqNet, and subscripts $_{reg}$ and $_{cls}$ indicate the regression and classification loss, respectively.

For an input image x, 128 numbers of proposals are selected then aligned into $1024 \times 14 \times 14$ RoIs by RoIAlign. The res5 in ResNet50 extracted these RoIs into $2048d$ features to calculate the box regression loss. Following the previous work NAE [13], $256d$ features f is extracted from $2048d$ by fully connection to perform classification and re-id loss. To overcome the goal conflict between the classification and re-



(a) Our proposed person searching network



(b) Illustration of learning direction of RoIs-based re-id.

Fig. 3. The architecture of the proposed person searching framework. The component in yellow is newly proposed by us. Our proposed Backgrounds and Foregrounds Contrasting Loss L^3_{bfcl} aims to push an RoI far away from other RoIs with different i and backgrounds.

id, NAE [13] decomposing f into norm r and angle θ in the polar corrdinate system as follows:

$$f = r \cdot \theta \tag{1}$$

where norm r is $1d$ value and angle θ is a $256d$ unit vector. To represent classification confidence using $r \in [0, +\infty)$, NAE normalize it to $|r| \in [0, 1]$. Four losses, i.e., regression loss L^3_{reg}, classification loss L^3_{cls}, online instance matching L^3_{oim} for re-id, and the proposed BFCL L^3_{bfcl} are used in the third head. The total learning objective function is then formulated as,

$$\begin{aligned} L = \lambda_1 L^1_{reg} + \lambda_2 L^1_{cls} + \lambda_3 L^2_{reg} + \lambda_4 L^2_{cls} + \\ \lambda_5 L^3_{reg} + \lambda_6 L^3_{cls} + \lambda_7 L^3_{oim} + \lambda_8 L^3_{bfcl} \end{aligned} \tag{2}$$

Following the SeqNet, $\lambda_1 = 10$, and the others are set to 1.

B. Background and Foreground Contrastive Loss (BFCL) for re-id

1) Traditional re-id loss: The Online Instance Matching (OIM) loss [9] is widely used in traditional person search methods. For a dataset with N numbers of identity classes, a $N \times 256$ sized look-up table $\mathcal{M}$ is built and maintained to store features for N classes. $\mathcal{M}$ is updated in every training iteration by θ as follows,

$$\mathcal{M}^t[i] = \alpha \mathcal{M}^{t-1}[i] + (1 - \alpha)\theta \tag{3}$$

where the superscript ^{t} denotes the t-th training iteration. $\alpha \in [0, 1]$, is the updating rate. i indicates the identity class of θ. During the whole training process. Then, the similarity s of the current input image and N identity classes can be computed using $\mathcal{M}$. OIM loss aims to pull θ close to its identity class i and push θ far away from other identity classes. In other words, when the similarity between θ and $\mathcal{M}[i]$ is high and the similarities between θ and other features in $\mathcal{M}$ is low, the OIM loss is small. OIM loss is defined as follows,

$$L^3_{oim} = -log \frac{\exp(< \theta, \mathcal{M}^+ >)/\tau}{\sum \exp(< \theta, \mathcal{M} >)/\tau} \tag{4}$$

where $<,>$ denotes cosine similarity of features, restricted between $[1, 1]$. $\mathcal{M}^+ = \mathcal{M}[i]$ is the i-th class that θ belongs to. τ is a temperature hyper-parameter. $< \theta, \mathcal{M} >$ is a N-dimensional vector, indicates the similarity between θ and N identity classes.

2) The proposed BFCL loss: OIM loss can be considered as class-based learning by pulling a feature closer to its corresponding class feature but OIM loss did not consider the relationship among RoIs of an image. We found that RoIs from the same image contains similar background pattern, as illustrated in Fig. 2. For example, Fig. 2(a) mainly contains grassland, Fig. 2(b) contains grassland desert, and Fig. 2(c) contains a stage with green light. The similar patterns lead the RoIs from the same image to be more difficult to classify. Therefore, we explore the relationship among RoIs additionally to help the model learn discriminative identity features from the foreground. The learning directions of the proposed BFCL loss are illustrated in Fig. 3(b).

For one RoI feature θ in an input image, the BFCL is computed as follows,

$$L^3_{bfcl} = \frac{1}{|\theta^+|} \sum_1^{|\theta^+|} -log \frac{\exp(< \theta, \theta^+ >)/\tau_c}{\sum \exp(< \theta, \mathcal{U} >)/\tau_c} \tag{5}$$

where $\mathcal{U} = \{\theta_1, ..., \theta_{128}\}$ indicates the collection of all IoU features in one input image. θ^+ indicates the positive RoIs in $\mathcal{U}$ that have the same class with θ. Intuitively, the above L^3_{bfcl}

encourages the θ to approach its positive RoIs, and leave its negative RoIs.

III. Experiments

The experiments are performed on two widely used datasets, CUHK-SYSU [9] and PRW [21].

A. Datasets

1) CUHK-SYSU: CUHK-SYSU [9] contains two data sources to diversify the scenes. The first one contains street snaps in an urban city, which are shot by hand-held cameras. The second data source is collected from movie snapshots, which contain person images with large variations of viewpoints, lighting, and background conditions. StreetSnap images and MovieTV screenshots. The datasets contains $18,184$ uncropped images, $96,143$ person bounding boxes with $8,432$ labeled identities in total. The training set has $11,206$ images, $55,272$ persons with $5,532$ different identities. The test set has $6,978$ images, $40,871$ persons with $2,900$ different identities.

2) PRW: PRW [21] are collected at Tsinghua university for about 10 hours with 6 cameras. The PRW aims to simulate real-world situations where pedestrians appear or disappear in different cameras. The datasets contains $11,816$ uncropped images, $43,110$ person bounding boxes with 484 labeled identities in total. Both CUHK-SYSU and PRW contain unlabeled identities. For example. In our paper, unlabeled identities are used in regression and classification losses but not used in re-id and our proposed BFCL.

3) Evaluation Metrics: Same with the re-id task, two evaluation metrics are used to measure model performance. The first is Mean Average Precision (mAP) (%). Another one is the Cumulative Matching Characteristic (CMC) curve. The CMC (%) of Top-1 is reported, which represents the probability of top-1 ranked gallery samples containing the query identity. Following the previous works, the detection evaluation metrics are not used to evaluate the performance of the person search model.

4) Implementation Details: Faster R-CNN [19] is adopt as the backbone network, in which ResNet-50 [20] pre-trained on ImageNet is used. The SeqNet [12] is the baseline framework of our method. The backbone network contains the res1, res2, res3, and res4 blocks of ResNet-50, and the output features of res4 are used for the first prediction head. The output features of res5 are used for the second prediction head.

The input images are resized to 900×1500. The batch size is 5. The network is trained by the Stochastic Gradient Descent (SGD) with a learning rate of 0.003 which is warmed up during the first epoch and decreased by 10 at the 16-th epoch. The model is trained for 20 epochs in CUHK-SYSU and 18 epochs in PRW. The circular queue size of OIM is not used here because the circular queue did not enhance model performance consistently in two datasets, the experimental results are reported in Table I. The updating rate α in Eq.(3) and τ in Eq.(4) are set to 0.5 and $1/3$, respectively.

The experiments are performed on one NVIDIA Tesla V100 GPU with 32 GB of memory. The total training time is around 24 hours on CUHK-SYSU, and 17 hours on PRW.

B. Ablation Study

Ablation studies are performed to demonstrate the effectiveness of the proposed BFCL and analyze the effectiveness of different temperature values τ_c.

1) Effectiveness of Circular Queue (CQ): We implement the analysis of Circular Queue (CQ) [9] on SeqNet-base model [12]. The results are reported in Table I. Our re-implementation of the SeqNet model without CQ is notated as "SeqNet-base" in Table I. Xiao et al. [9] proposed CQ to store the features of unlabeled identities situation. They demonstrate the effective use of CQ in their framework. However, we found out CQ did not enhance SeqNet performance consistently in two datasets, as reported in "SeqNet-base" and "SeqNet-base + CQ" in Table I. Adding CQ to SeqNet-base yields a gain of $+0.2$ in mAP and $+0.5$ in Top-1 in CUHK-SYSU but decreases -0.5 in mAP and -0.2 in Top-1 in PRW. Therefore, we consider that the CQ is not a robust method and therefore not used in our paper.

2) Effectiveness of Proposed BFCL: We implement the analysis of our proposed BFCL on the SeqNet-base model [12]. The results are reported in Table I. It is clear that adding BFCL to the SeqNet-base yields consistent gain in two datasets. Specifically, Adding BFCL to the SeqNet-base yields a gain of $+0.4$ in mAP and $+0.5$ in Top-1 in CUHK-SYSU but decreases $+1.5$ in mAP and $+0.8$ in Top-1 in PRW.

3) Comparison with Different Temperature τ_c in Eq.(5): Almost all contrastive learning-based methods [12], [22], [23] used the temperature value and have similar effects. [24] demonstrated that the contrastive loss is a hardness-aware loss function, and the temperature value τ_c controls the strength of penalties on hard negative samples. Small τ_c tends to pay more attention to the hard negative samples. Large τ_c tends to pay less attention to the hard negative samples, in other words, less sensitive to the hard negative samples. Our person search framework performance in two datasets with different temperatures τ_c are reported in Table II. For CUHK-SYSU, when $\tau_c = 0.03$, our framework achieves the best results 94.0% in mAP and 94.6% in Top-1. For PRW, when $\tau_c = 0.05$, our framework achieves the best results 48.7% in mAP and 84.4% in Top-1. Following the theory in [24], the different optimal value of τ_c in CUHK-SYSU and PRW demonstrates that paying more attention to the hard negative samples helps model performance in CUHK-SYSU. On the other hand, paying less attention to the hard negative samples helps model performance in PRW.

C. Comparison with the state-of-the-art Methods

We compare our method against state-of-the-art person search models in CUHK-SYSU and PRW in Table III. As the baseline of recent person search works [11]–[13], OIM [9] is the first paper that proposed Online Instance Matching (OIM) loss function to end-to-end train the re-id with detection jointly. NAE [13] notice the end-to-end training strategy enhance the goal conflict between detection and re-id, therefore NAE [13] decompose feature f into norm r and angle θ. Based on two head methods NAE, SeqNet [12] added

Methods	CUHK-SYSU		PRW	
	mAP	Top-1	mAP	Top-1
SeqNet-base	93.6	94.1	47.2	83.6
SeqNet-base + CQ	93.8 (+0.2)	94.6 (+0.5)	46.7 (-0.5)	83.4 (-0.2)
SeqNet-base + BFCL (proposed)	**94.0** (+0.4)	**94.6** (+0.5)	**48.7** (+1.5)	**84.4** (+0.8)

TABLE I. Ablation experiments on **CQ**: Circular Queue [9] and **BFCL**: our proposed Background and Foreground Contrastive Loss.

τ_c in Eq.(5)	CUHK-SYSU		PRW	
	mAP	Top-1	mAP	Top-1
0.03	**94.0**	**94.6**	48.5	84.1
0.05	93.6	94.2	**48.7**	**84.4**
0.10	92.7	93.5	47.9	84.1

TABLE II. Performance of our framework with different values of τ_c in Eq.(5).

Method	reference	CUHK-SYSU		PRW	
		mAP	Top-1	mAP	Top-1
OIM [9]	CVPR17	75.5	78.7	21.3	49.9
NAE [13]	CVPR20	91.5	92.4	43.3	80.9
AlignPS [11]	CVPR21	93.1	93.4	45.9	81.9
SeqNet [12]	AAAI21	93.8	**94.6**	46.7	83.4
BFCL	Ours	**94.0**	**94.6**	**48.7**	**84.4**

TABLE III. Comparison with state-of-the-art methods on two person searching datasets. The top result is highlighted in bold.

one prediction head to improve the detection accuracy for providing high-quality RoIs. Based on SeqNet, our proposed BFCL further boosts the person search performance in both CUHK-SYSU and PRW datasets, especially in PRW. The consistent improvements demonstrate the necessity of mining relationships among RoIs of an image and the effectiveness of our proposed BFCL.

IV. Conclusions

This paper introduces an end-to-end person search model in this paper. To strengthen the re-id capability of the model, we propose a Background and Foreground Contrastive Loss (BFCL) which can leverage similarity relationships among RoIs to learn to distinguish similar backgrounds and foregrounds. Moreover, we demonstrate that the widely used CQ can not consistently enhance our model's performance in two datasets. In the future, we wish to focus on two aspects to enforce the practicability of person search in real applications: (1) Integrating our methods into a lightweight detection network, and (2) Integrating the proposed algorithm in high-level video surveillance tasks.

Acknowledgment

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the government (MSIT). (No.2020R1A2C200897212)

References

[1] C. Liang, Z. Zhang, Y. Lu, X. Zhou, B. Li, X. Ye, and J. Zou, "Rethinking the competition between detection and reid in multi-object tracking," *ArXiv*, vol. abs/2010.12138, 2020.

[2] L. Zheng, Y. Yang, and A. Hauptmann, "Person re-identification: Past, present and future," *ArXiv*, vol. abs/1610.02984, 2016.

[3] K. Zhou, Y. Yang, A. Cavallaro, and T. Xiang, "Omni-scale feature learning for person re-identification," *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3701–3711, 2019.

[4] Q. Tang and K.-H. Jo, "Unsupervised person re-identification via nearest neighbor collaborative training strategy," in *2021 IEEE International Conference on Image Processing (ICIP)*, 2021, pp. 1139–1143.

[5] Q. Tang, G. Cao, and K.-H. Jo, "Fully unsupervised person re-identification via multiple pseudo labels joint training," *IEEE Access*, vol. 9, pp. 165 120–165 131, 2021.

[6] D. Wang and S. Zhang, "Unsupervised person re-identification via multi-label classification," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10 978–10 987, 2020.

[7] Z. Zhong, L. Zheng, Z. Luo, S. Li, and Y. Yang, "Invariance matters: Exemplar memory for domain adaptive person re-identification," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 598–607, 2019.

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

[9] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang, "Joint detection and identification feature learning for person search," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3376–3385, 2017.

[10] Y. Zhang, X. Li, and Z. Zhang, "Efficient person search via expert-guided knowledge distillation," *IEEE Transactions on Cybernetics*, vol. 51, pp. 5093–5104, 2021.

[11] Y. Yan, J. Li, J. Qin, S. Liao, and X. Yang, "Efficient person search: An anchor-free approach," *ArXiv*, vol. abs/2109.00211, 2021.

[12] Z. Li and D. Miao, "Sequential end-to-end network for efficient person search," in *AAAI*, 2021.

[13] D. Chen, S. Zhang, J. Yang, and B. Schiele, "Norm-aware embedding for efficient person search," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12 612–12 621, 2020.

[14] D. Chen, S. Zhang, W. Ouyang, J. Yang, and Y. Tai, "Person search via a mask-guided two-stream cnn model," in *ECCV*, 2018.

[15] C. Han, J. Ye, Y. Zhong, X. Tan, C. Zhang, C. Gao, and N. Sang, "Re-id driven localization refinement for person search," *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9813–9822, 2019.

[16] B.-J. Han, K. Ko, and J.-Y. Sim, "End-to-end trainable trident person search network using adaptive gradient propagation," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 905–913, 2021.

[17] X. Zhang, X. Wang, J. Bian, C. Shen, and M. You, "Diverse knowledge distillation for end-to-end person search," in *AAAI*, 2021.

[18] C. Wang, B. Ma, H. Chang, S. Shan, and X. Chen, "Tcts: A task-consistent two-stage framework for person search," in *Proceedings of*

the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2020.

[19] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, pp. 1137–1149, 2015.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

[21] L. Zheng, H. Zhang, S. Sun, M. Chandraker, Y. Yang, and Q. Tian, "Person re-identification in the wild," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3346–3355, 2017.

[22] H. Chen, B. Lagadec, and F. Bremond, "Ice: Inter-instance contrastive encoding for unsupervised person re-identification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 14 960–14 969.

[23] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," *arXiv preprint arXiv:1911.05722*, 2019.

[24] F. Wang and H. Liu, "Understanding the behaviour of contrastive loss," *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2495–2504, 2021.