Multi-View Label Prediction for Unsupervised Learning Person Re-Identification

Qingze Yin, Guan’an Wang, Guodong Ding, Shaogang Gong, and Zhenmin Tang

Abstract—**Person re-identification (ReID) aims to match pedestrian images across disjoint cameras. Existing supervised ReID methods utilize deep networks and train them with identity-labeled images, which suffer from limited annotations. Recently, clustering-based unsupervised ReID attracts more and more attention. It first clusters unlabeled images and assigns cluster index to the pseudo-identity-labels, then trains a ReID model with the pseudo-identity-labels. However, considering the slight inter-class variations and significant intra-class variations, pseudo-identity-labels learned from clustering algorithms are usually noisy and coarse. To alleviate the problems above, besides clustering pseudo-identity-labels, we propose to learn pseudo-patch-labels, which brings two advantages: (1) Patch naturally alleviates the effect of backgrounds, occlusions, and carryings since they usually occupy small parts in images, thus overcome noisy labels. (2) It is plausible that patches from different pedestrians belong to the same pseudo-identity-label. For example, pedestrians have a high probability of wearing either the same shoes or pants but a low possibility of wearing both. The experiments demonstrate our proposed method achieves the best performance by a large margin on both image- and video-based datasets.**

Index Terms—**Unsupervised learning, multi-view learning, person re-identification, clustering.**

I. Introduction

PERSON re-identification (ReID) is an important computer vision task which deals with pedestrian image matching under non-overlapping camera installations. Recently, deep supervised ReID methods [1]–[3], [36], [43], [42] have made impressive progress. However, those supervised methods become somewhat limited in some real-world scenarios where data annotation is not available. This motivates the community to explore the task in an unsupervised setting.

Unsupervised ReID methods can be mainly grouped into four streams: conventional, domain-transfer-based, trajectory-based, and cluster-based. **Conventional** unsupervised ReID methods are built on top of hand-crafted features [16], [38], dictionary learning [4], and saliency analysis [14], [15]. Due to the lacking of semantic information, the methods often have inferior performance. **Domain-transfer-based** unsupervised ReID methods [17], [20], [26], [40] transfer knowledge from a labeled source domain to a target unlabeled domain. Yang *et al.* [35] aims to learn discriminative ‘individual’ patch features to solve the transfer gap on the image-level feature by pre-training on a very large dataset to learn the label knowledge. Those methods seriously rely on the small gap between source and target domains, which may not be very flexible in real scenarios. **Trajectory-based** unsupervised ReID methods utilize trajectories from tracking algorithms [22], [23], [29], [40], and suppose all images in a trajectory are the same pedestrians, then train a ReID model in a self-supervised way. Trajectory-based unsupervised ReID methods may suffer from inaccurate tracking and are not applicable in image-based ReID scenes.

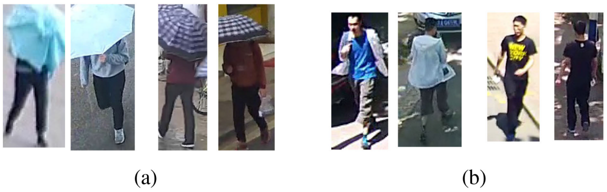


Fig. 1. Challenges in unsupervised person ReID with only pseudo-identity-labels. (a) Occlusions make distinct identities visually similar. (b) Large intra-identity variance is caused by viewpoint changes and illumination conditions.

Clustering-based unsupervised ReID [7], [10], [37] alternatively generate pseudo-labels with clustering and training a ReID model with them. This stream requires neither manual annotations nor detected trajectory, which is more flexible in real-world application meanwhile achieving decent accuracy, Ding *et al.* [7] proposed a dispersion-based clustering framework that progressively merges similar clusters based on a dispersion criterion and learns representations with cluster labels. However, only holistic image features are considered in this work.

We can still observe some drawbacks of the existing clustering-based unsupervised methods as below: (1) **Noisy labels**. As in Fig. 1 a, due to the small inter-class variation (such as backgrounds, occlusions, and carryings), the pseudo-labels are easily polluted by some local similarity and lead to noisy labels. (2) **Coarse labels**. As in Fig. 1 b, due to the large intra-class variation (such as poses, views, illuminations), images from a pedestrian often have more than one pseudo-labels, lead to coarse labels and confusing model training. Considering the

Manuscript received April 13, 2021; revised June 8, 2021; accepted June 11, 2021. Date of publication June 23, 2021; date of current version July 20, 2021. This work was supported by the program of China Scholarships Council under Grant 202006840102. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Ferdous Sohel. *(Corresponding author: Qingze Yin)*

Qingze Yin, Guodong Ding, and Zhenmin Tang are with the Nanjing University of Science and Technology, Nanjing 210094, China (e-mail: qingzeyin@njust.edu.cn; GUODONG.DING@NJUST.EDU.CN; Tzm.cs@njust.edu.cn).

Guan’an Wang is with the Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: guan.wang0706@gmail.com).

Shaogang Gong is with the Queen Mary University of London, London E1 4NS, U.K. (e-mail: s.gong@qmul.ac.uk).

Digital Object Identifier 10.1109/LSP.2021.3090258

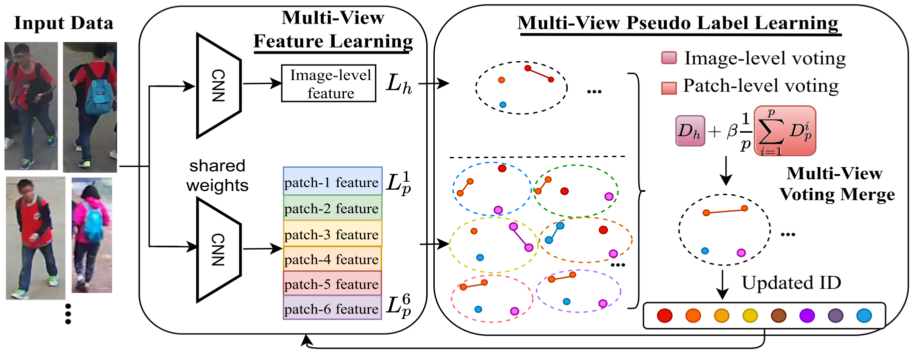


Fig. 2. Our proposed Multi-View Learning framework. CNN model and pseudo-labels are trained with multi-view features in an alternating manner. The multi-view voting mechanism helps to form more robust clusters. *i.e*, the circles of different colours represent the corresponding patch-clusters. Their merging results will be used in the final merging stage (multi-view voting merge) with the image-cluster merging result.

observations above, we propose a novel multi-view unsupervised ReID method, which both learns pseudo-patch-labels and pseudo-identity-labels. Our proposed multi-view unsupervised ReID enjoys two merits: (1) Patches naturally alleviate the effect of backgrounds, occlusions, and carryings since they usually occupy small parts in images. Besides, it is more reasonable that patches from different pedestrians belong to the same pseudo-patch-label. For example, pedestrians dress in the same shoes. (2) Patches focus more on texture information, alleviating the effect of intra-class variations. Specifically, MV-ReID includes a multi-view feature learning module and a multi-view pseudo-label learning. The former learns holistic and patch features from given images with convolutional neural networks (CNN). The latter separately learns pseudo-labels for holistic and every patch with a clustering algorithm. Then holistic and patch features are trained by their corresponding pseudo-labels. They are iteratively performed until the model converges.

We list our contributions as (1) We propose a novel multi-view unsupervised ReID (MV-ReID), which learns, aside from pseudo-identity-labels, noise-tolerant, and complementary pseudo-patch-labels to cumulatively vote for the identity tag. Such design enables a more stable learning process. (2) MV-ReID proposes to describe a person identity by a multi-view feature set. The descriptor set comprises the holistic and patch-level representations, which contain more discriminative cues than other single feature descriptors. (3) Extensive experiments on both image- and video-based ReID datasets demonstrate the effectiveness of our proposed MV-ReID.

The rest of the paper is organized as follows: Section II elaborates on the proposed framework and algorithm. Then, experiments, evaluations, and discussions are presented in Section III. Finally, Section IV concludes the paper.

II. The Proposed Method

We present our novel Multi-View unsupervised ReID (MV-ReID) approach in Fig. 2.

A. *Multi-View Feature Learning*

We first utilize a CNN to learn feature maps from given batch images. The learned feature maps are then post-processed to be holistic features vector via a global average pooling (GAP), and a group of the patch features p by horizontally stripping and GAP. The generated holistic and patch feature vectors own the same dimension (in channel), and are trained with a non-parametric cross-entropy loss, similar as [7]–[10]:

$$\ell = -log(p(y|x,V)), p(y|x,V) = \frac{exp(V_y^T v/\tau)}{\sum_{j=1}^{C} exp(V_j^T v/\tau)} \qquad (1)$$

where $p(y|x,V)$ is the probability of x belonging to the y-th cluster. Here, x could be a image x or the i-th patch x^i and so does y for both image and p patch cluster. $v = \frac{\phi(\theta;x)}{\|\phi(\theta;x)\|}$ is l_2 normalized image/patch feature, V is a look up table which stores the centroid features of that cluster and updated on-the-fly. C is the number of clusters at present, τ is a temperature parameter that controls the softness of probability distribution over clusters.

The overall multi-view loss is calculated as a combination of all $1+p$ losses by a weighting term α:

$$L_{multi-view} = \ell_h + \alpha\frac{1}{p}\sum_{i=1}^{p}\ell_p^i, \qquad (2)$$

p is the number of local views, ℓ_h represents the loss of holistic feature, and ℓ_p^i represents the loss for i-th patch.

With the above loss formulation, our model learns discriminative holistic-view and local-view features. The resulting features collectively form a multi-view feature set considered as the final person descriptor for testing written as:

$$f = \{f_h, f_p^i\}, \quad i = 1, \ldots, p \qquad (3)$$

B. *Multi-View Pseudo-Labels Learning*

It is nontrivial to design an alternative supervision signal to train CNN models when labels are not provided. Inspired by [30], we can obtain pseudo-labels by utilizing agglomerative clustering. Initially, each image is considered as a singleton cluster. As clustering progresses, pairs of clusters will merge according to a given similarity criterion. The idea is based on the observation that images of the same identity need to be visually similar and should be close [7], [10]. However, existing works only use the image-level feature to calculate similarities, easily introducing errors due to the significant variations in person images. Thus, it is critical to find a robust criterion.

This multi-view pseudo-label learning module considers the multi-view distance of samples from two different level clusters to calculate the dissimilarity. In this work, we consider seven views; each view is in the form of feature representation. As shown in Fig. 2, for each sample, we can obtain one *unique view* of holistic appearance and six different patches providing a complementary set of different *multi-views* of the person image. We calculate the distances between each corresponding patch of image pairs and aggregate the average of all patch distances with a balancing parameter (β) to influence the contributions of local features in each pair of image matching (Re-ID). Moreover, these local feature distances are assisted by a global feature distance to estimate whether the image pairs belong to the same cluster when voting. Then we rank all calculated final distances of pairs images and merge the top-ranked similar pairs with the smallest distance. The fact that both images and patches are considered, our concept of *multi-view* is also very different from the conventional idea of *multi-patch* in both model design and rational. The multi-view voting criterion is formulated as below:

$$D(A, B) = \min_{f \in A, g \in B} d(f_h, g_h) + \beta \frac{1}{p} \sum_{i=1}^{p} d(f_p^i, g_p^i) \qquad (4)$$

β is a hyper-parameter to balance the impact of holistic and patch distance similarities, f and g denote the samples from two different clusters A and B, respectively.

The overall training strategy is the iterative update of feature learning and pseudo-label learning. Algorithm 1 provides the detailed training process.

C. Discussion

Our proposed MV-ReID is related to two types of existing methods. PCB [11] learns patch features in the supervised setting. [7], [10] learn holistic features and cluster to produce pseudo labels with different criteria. Different from them, our proposed methods form a multi-view voting mechanism to learn a robust pseudo label at both the patch-level and the holistic-level. These multi-granularity features vote for a more robust pseudo label, supervising the feature learning module to generate better representations in our alternating update scheme. Our proposed method also outperforms these works by a large margin (See Section III-D).

III. Experiments

A. Datasets

We evaluate the proposed MV-ReID on three datasets. **Market-1501** [6] is an image-based dataset of 1501 identities captured by 6 out-door cameras, of which the training set composes 12936 images of 751 identities, and the testing set composes 19732 images of 750 identities. **DukeMTMC-reID** [12] contains 36411 images of 1404 identities captured by 8 cameras. Its training set composes 16522 images of 702 identities, and its query set has 2228 images of the other 702 identities, 17661 for gallery images. **DukeMTMC-VideoReID** [25] is derived from DukeMTMC dataset which is a large-scale video-based dataset contains 2196 tracklets of 702 identities for training, and 2636 tracklets of other 702 identities for testing. Same as most existing work, rank-k, an indicator of retrieval accuracy, and mean average precision (mAP) for precision are reported in our results.

Algorithm 1: Multi-View Unsupervised Learning.

```
Require: training data $X = \{x_i\}_{i=1}^N$, merging rate
  $m \in (0, 1)$, parameter $\alpha$, $\beta$, CNN model $\phi(\cdot; \theta)$, number
  of partitions $p$
Ensure: Optimised CNN model $\hat{\phi}(\cdot; \hat{\theta})$
Initialize: cluster labels $Y = \{y_i = i\}_{i=1}^N$, cluster
  number $C = N$, number of iterations $T$
while $t < T$ do
  Multi-view Feature Learning (Section II-A):
  train $\phi(\cdot; \theta^t)$ with $X$ and $Y$ with loss in 2 for a given
    number of epochs;
  Multi-view Pseudo-Label Learning
    (Section II-B):
  extract both holistic and local feature sets
    $\{F_h, F_p\} = \phi(X; \theta^t)$ on the entire training set $X$;
  calculate the distance of $f_h$ and each patch-level
    feature $f_p^i$ between all pairs across all clusters with
    Eq.4.
  merge $m$ of top-ranked cluster pairs to form new
    clusters;
  update $Y^t$ with updated clustering results;
  evaluate ReID performance with person descriptor $f$
    in 3 on validation data and denote as $P^t$;
  if $P^t > \hat{P}$ then
    $\hat{P} = P^t$, $\hat{\phi}(\cdot; \hat{\theta}) = \phi(\cdot; \theta^t)$
  end if
end while
```

B. Implementation Details

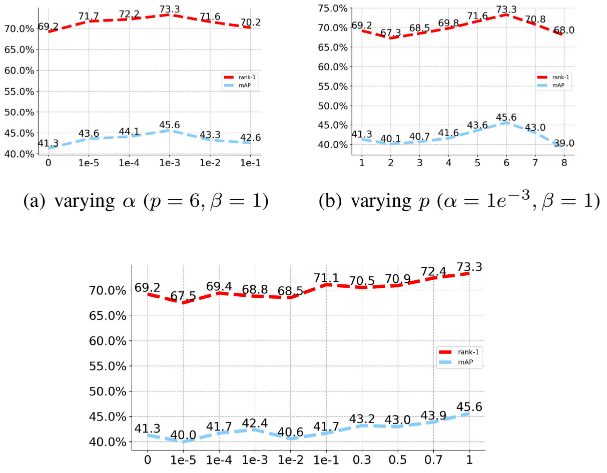
For a fair comparison, the pre-trained ResNet50 [34] is adopted as our CNN backbone. The learning comprises of 20 iterations of alternating updates. The feature learning module is trained 20 epochs for the first iteration and two epochs each after, with a batch size of 16. We optimize the model by Stochastic Gradient Descent (SGD) with momentum set to 0.9; the initial learning rate is 0.1 and decayed to 0.01 after 15 iterations. The merging percentage for each update iteration is set to 0.05 and τ is set to 0.1. Our choice of the number of local views p is 6. α is set to 0.001 and β is 1 in Market-1501 and DukeMTMC-VideoReID, α is 0.01 in DukeMTMC-reID.

C. Parameter Study

To analyze the effects of balancing factors α, β, and view number p have on ReID task, we perform a parameter study on the Market-1501 dataset. Fig. 3 shows the comparisons for diverse settings of hyper-parameters α, β, and p. Note that when α and β are equal to 0, the model would be the same as DBC [7], where only holistic image features are considered.

TABLE I
EXPERIMENTAL RESULTS ON IMAGE-BASED REID DATASET: MARKET-1501, DUKEMTMC-REID AND VIDEO-BASED REID DATASET: DUKEMTMC-VIDEOREID. "ONEEX" DENOTES SINGLE-EXAMPLE ANNOTATION, WHERE EACH PERSON ID IN THE DATASET ONLY HAS ONE LABELED SAMPLE. "TRANSFER" USES FULL ANNOTATIONS FROM AN EXTERNAL DATASET

Methods	Labels	Market-1501				DukeMTMC-reID				DukeMTMC-VideoReID			
		1	5	10	mAP	1	5	10	mAP	1	5	10	mAP
CCLF[37]	Transfer	46.1	62.4	68.8	22.2	31.2	47.7	54.1	17.2	-	-	-	-
PAUL[35]	Transfer	68.5	82.4	87.4	40.1	72.0	82.7	86.0	53.2	-	-	-	-
ENC[33]	Transfer	75.1	87.6	91.6	43.0	63.3	75.8	80.4	40.4	-	-	-	-
SSG[32]	Transfer	80.0	90.0	92.4	58.3	73.0	80.6	83.1	53.4	-	-	-	-
DGM+IDE[22]	OneEx	-	-	-	-	-	-	-	-	42.3	57.9	69.3	33.6
Stepwise[29]	OneEx	-	-	-	-	-	-	-	-	56.2	70.3	79.2	46.7
EUG[25]	OneEx	49.8	66.4	72.7	22.5	45.2	59.2	63.4	24.5	72.7	84.1	-	63.2
OIM[24]	None	38.0	58.0	66.3	14.0	24.5	38.8	46.0	11.3	51.1	70.5	76.2	43.8
BUC[10]	None	66.2	79.6	84.5	38.3	47.4	62.6	68.4	27.5	69.2	81.1	85.8	61.9
DBC[7]	None	69.2	83.0	87.8	41.3	51.5	64.6	70.1	30.0	75.2	87.0	90.2	66.1
MV-ReID (Ours)	None	**73.3**	**85.3**	**89.1**	**45.6**	**54.5**	**67.5**	**72.1**	**31.7**	**80.9**	**92.0**	**95.3**	**73.8**



(a) varying α ($p = 6, \beta = 1$)

(b) varying p ($\alpha = 1e^{-3}, \beta = 1$)

(c) varying β ($p = 6, \alpha = 1e^{-3}$)

Fig. 3. Sensitivity study of balancing parameter α, β, and number of view p on Market-1501 dataset.

Loss term α. Fig. 3 a outlines the comparisons for various α with p=6 and β=1 partitions. As we can see, as α varies from 0 to 0.1, the rank-1 accuracy(73.3%) fluctuates and the highest value is at α=0.001. The same trend can be observed on mAP(45.6%).

Similarity term β. We fix α=1e^{-3}, p=6, and report in Fig. 3 c, the performance undulates and achieves the best result at β=1. Another noteworthy fact is that despite the choice of α and β (except 0), the model has consistently outperformed DBC [7] (69.2% rank-1 accuracy and 41.3% mAP), highlighting the importance of multi-view features.

Number of views p. We next fix α=1e^{-3}, β=1 to study how the number of views p affects performance. As Fig. 3 b shows, when p=6 the model achieves the best performance.

D. Comparison With State-of-The-Arts

We summarize and compare our approach with the state-of-the-art approaches on both image-based and video-based datasets in Table. I.

Image-based ReID. As we can see in Table I, 'Transfer' ReID methods, compared with others, can perform very competitive results on these two datasets. SSG [32] achieved 80% and 73% rank-1 scores on Market-1501 and DukeMTMC-reID, respectively. While 'OneEx' approach EUG [25] achieved 49.8% and 45.2% rank-1 accuracy on Market-1501 and DukeMTMC-reID, respectively. 'None' setting is more challenging than 'Transfer,' as it does not require any form of annotation. Fully unsupervised approach DBC [7] achieved 69.2%/41.3% (rank-1/mAP) and 51.5%/30.0% scores on two datasets. Compared with theirs, our MV-ReID model managed to achieve the best performance by a large margin of 4.1%(rank-1) and 4.3%(mAP) on Market-1501 and 3.0%(rank-1) and 1.7%(mAP) on DukeMTMC-reID. This has shown the effectiveness of the multi-view learning of features.

Video-based ReID. Results of different methods on video-based dataset DukeMTMC-VideoReID can also be found in Table I. Notably, DBC [7] which is strictly unsupervised achieves 75.2%(rank-1) and 66.1%(mAP) surpassing 'OneEx' approach EUG (72.7% and 63.2%). This big improvement shows that the clustering-based method has the potential to recognize person identities. By incorporating multi-view learning, our MV-ReID model further achieves the best performance at 80.9%(rank-1) and 73.8%(mAP), which is 5.7% and 7.7% higher than DBC [7] on rank-1 and mAP. The results have demonstrated the effectiveness of our proposed MV-ReID approach in this unsupervised setting.

IV. CONCLUSIONS

In this paper, we propose a Multi-View learning approach that works cyclically for the task of unsupervised person ReID. It can generate discriminative features and reciprocally produce more accurate pseudo labels. The approach exploits holistic and local-view features synchronously. Local-view features are reinforced within-patch consistency, which encodes finer-grained information, particularly for person ReID. Extensive studies and comparative evaluations have demonstrated the effectiveness of MV-ReID. In future work, as the unsupervised clustering is sensitive to the camera domain gap, we are interested in exploring how to incorporate cross-camera discriminative information to further improve the accuracy of our multi-view learning framework.

REFERENCES

[1] C. Tian, M. Zeng and Z. Wu, “Person re-identification based on spatiogram descriptor and collaborative representation,” in *IEEE Signal Process. Lett.*, vol. 22, no. 10, pp. 1595–1599, Oct. 2015.

[2] A. V. Subramanyam, V. Gupta, and R. Ahuja, “Robust discriminative subspace learning for person reidentification,” in *IEEE Signal Process. Lett.*, vol. 26, no. 1, pp. 154–158, Jan. 2019.

[3] S. Zhang, L. Zhang, W. Wang, and X. Wu, “AsNet: Asymmetrical network for learning rich features in person re-identification,” *IEEE Signal Process. Lett.*, vol. 27, pp. 850–854, 2020.

[4] E. Kodirov, T. Xiang, and S. Gong, “Dictionary learning with iterative laplacian regularisation for unsupervised person re-identification,” in *Proc. BMVC*, 2015, pp. 44.1–44.12.

[5] G. Ding, S. Khan, Z. Tang, J. Zhang, and F. Porikli “Towards better validity: Dispersion based clustering for unsupervised person re-identification,” 2019, *arXiv:1906.01308*.

[6] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, “Scalable person re-identification: A benchmark,” in *Proc. Proceed. IEEE Int. Conf. Comput. Vision (ICCV)*, Santiago, Chile, 2015, pp. 1116–1124.

[7] G. Ding, S. Khan, and Z. Tang, “Dispersion based clustering for unsupervised person re-identification,” in *Proc. BMVC*, 2019.

[8] G. Ding, S. Zhang, S. Khan, Z. Tang, J. Zhang, and F. Porikli, “Feature affinity-based pseudo labeling for semi-supervised person re-identification,” *IEEE Trans. Multimedia*, vol. 21, no. 11, pp. 2891–2902, Nov. 2019.

[9] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, “A discriminative feature learning approach for deep face recognition,” in *Proc. Eur. Conf. Comput. Vision (ECCV)*, 2016, pp 499–515.

[10] Y. Lin, X. Dong, L. Zheng, Y. Yan, and Y. Yang, “A bottom-up clustering approach to unsupervised person re-identification,” in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 8738–8745.

[11] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, “Beyond part models: Person retrieval with refined part pooling,” in *Proc. Eur. Conf. Comput. Vision (ECCV)*, 2018, pp. 480–496.

[12] Z. Zheng, L. Zheng, and Y. Yang, “Unlabeled samples generated by GAN improve the person re-identification baseline in vitro,” in *Proc. IEEE Int. Conf. Comput. Vision (ICCV)*, 2017, pp. 3754–3762.

[13] L. Zheng *et al.*, “MARS: A. video benchmark for large-scale person re-identification,” in *Proc. Eur. Conf. Comput. Vision (ECCV)*, 2016, pp 868–884.

[14] R. Zhao, W. Ouyang, and X. Wang, “Unsupervised salience learning for person re-identification,” in *Proc. IEEE Conf. Comput. Vision Pattern Recognit. (CVPR)*, Portland, OR, USA, 2013, pp. 3586–3593.

[15] Y. Xie, H. Yu, X. Gong, Z. Dong, and Y. Gao, “Learning visual-spatial saliency for multiple-shot person re-identification,” *IEEE Signal Process. Lett.*, vol. 22, no. 11, pp. 1854–1858, Nov. 2015.

[16] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, “Person re-identification by local maximal occurrence representation and metric learning,” in *Proc. IEEE Conf. Comput. Vision Pattern Recognit. (CVPR)*, Boston, MA, USA, 2015, pp. 2197–2206.

[17] J. Wu, S. Liao, Z. lei, X. Wang, Y. Yang, and S. Z. Li, “Clustering and dynamic sampling based unsupervised domain adaptation for person re-identification,” in *Proc. 2019 IEEE Int. Conf. Multimedia Expo (ICME)*, Shanghai, China, 2019, pp. 886–891.

[18] Z. Li, J. Tang, and T. Mei, “Deep collaborative embedding for social image understanding,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 9, pp. 2070–2083, Sep. 2019.

[19] Z. Zheng, L. Zheng, and Y. Yang, “A discriminatively learned CNN embedding for person re-identification,” *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 14, no. 1, pp. 1–20, 2017.

[20] H. Fan, L. Zheng, C. Yan, and Y. Yang, “Unsupervised person re-identification: Clustering and fine-tuning,” *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 14, no. 4, pp. 1–18. 2018.

[21] Z. Zhong, L. Zheng, Z. Zheng, S. Li, and Y. Yang, “CamStyle: A. novel data augmentation method for person re-identification,” *IEEE Trans. Image Process.*, vol. 28, no. 3, pp. 1176–1190, Mar. 2019.

[22] M. Ye, A. J. Ma, L. Zheng, J. Li, and P. C. Yuen, “Dynamic label graph matching for unsupervised video re-identification,” in *Proc. IEEE Int. Conf. Comput. Vision (ICCV)*, Venice, Italy, 2017, pp. 5142–5150.

[23] M. Li, X. Zhu and S. Gong, “Unsupervised person re-identification by deep learning tracklet association,” in *Proc. Eur. Conf. Comput. Vision (ECCV)*, vol. 11208, 2018, pp. 772–788.

[24] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang, “Joint detection and identification feature learning for person search,” in *Proc. IEEE Conf. Comput. Vision Pattern Recognit. (CVPR)*, Honolulu, HI, USA, 2017, pp. 3415–3424.

[25] Y. Wu, Y. Lin, X. Dong, Y. Yan, W. Ouyang, and Y. Yang, “Exploit the unknown gradually: One-shot video-based person re-identification by stepwise learning,” in *Proc. IEEE Conf. Comput. Vision Pattern Recognit. (CVPR)*, Salt Lake City, UT, USA, 2018, pp. 5177–5186.

[26] P. Peng *et al.*, “Unsupervised cross-dataset transfer learning for person re-identification,” in *Proc. IEEE Conf. Comput. Vision Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, 2016, pp. 1306–1315.

[27] Y. Chen, X. Zhu, and S. Gong, “Deep association learning for unsupervised video person re-identification,” in *Proc. BMVC*, 2018, p. 48.

[28] M. Ye, X. Lan, and P. C. Yuen, “Robust anchor embedding for unsupervised video person re-identification in the wild,” in *Proc. Eur. Conf. Comput. Vision (ECCV)*, 2018, pp. 170–186.

[29] Z. Liu, D. Wang and H. Lu, “Stepwise metric promotion for unsupervised video person re-identification,” in *Proc. IEEE Int. Conf. Comput. Vision (ICCV)*, Venice, Italy, 2017, pp. 2429–2438.

[30] M. Bautista, A. Sanakoyeu, E. Tikhoncheva, and B. Ommer, “CliqueCNN: Deep unsupervised exemplar learning,” in *Proc. NeurIPS*, 2016, pp.3853–3861.

[31] Y. Chen, X. Zhu, and S. Gong, “Instance-guided context rendering for cross-domain person re-identification,” in *Proc. IEEE/CVF Int. Conf. Comput. Vision (ICCV)*, Seoul, South Korea, 2019, pp. 232–242.

[32] Y. Fu *et al.*, “Self-similarity grouping: A. simple unsupervised cross domain adaptation approach for person re-identification,” in *Proc. IEEE/CVF Int. Conf. Comput. Vision (ICCV)*, Seoul, South Korea , 2019, pp. 6112–6121.

[33] Z. Zhong, L. Zheng, Z. Luo, S. Li, and Y. Yang, “Invariance matters: Exemplar memory for domain adaptive person re-identification,” in *Proc. IEEE/CVF Conf. Comput. Vision Pattern Recognit. (CVPR)*, 2019, pp. 598–607.

[34] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE Conf. Comput. Vision Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, 2016, pp. 770–778.

[35] Q. Yang, H. Yu, A. Wu, and W. Zheng, “Patch-based discriminative feature learning for unsupervised person re-identification,” in *Proc. IEEE/CVF Conf. Comput. Vision Pattern Recognit. (CVPR)*, Long Beach, CA, USA, 2019, pp. 3633–3642.

[36] Y. Zhao, Y. Li, and S. Wang, “Open-world person re-identification with deep hash feature embedding,” *IEEE Signal Process. Lett.*, vol. 26, no. 12, pp. 1758–1762, Dec. 2019.

[37] J. Lu, Y. He, T. Liu, and X. Chen, “Centralized and clustered features for person re-identification,” *IEEE Signal Process. Lett.*, vol. 26, no. 6, pp. 933–937, Jun. 2019.

[38] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, “Person re-identification by symmetry-driven accumulation of local features,” in *Proc. IEEE Comput. Soc. Conf. Comput. Vision Pattern Recognit. (CVPR)*, San Francisco, CA, USA, 2010, pp. 2360–2367.

[39] J. Wang, X. Zhu, S. Gong, and W. Li, “Transferable joint attribute-identity deep learning for unsupervised person re-identification,” in *Proc. IEEE Conf. Comput. Vision Pattern Recognit. (CVPR)*, 2018, pp. 2275–2284.

[40] M. Li, X. Zhu, and S. Gong, “Unsupervised Person Re-Identification by Deep Learning Tracklet Association,” in *Proc. Eur. Conf. Comput. Vision (ECCV)*, 2018, pp. 737–753.

[41] G. Wang, Y. Yang, J. Cheng, J. Wang, and Z. Hou, “Color-sensitive person re-identification,” in *Proc. IJCAI*, 2019, pp. 933–939,

[42] G. Wang *et al.*, “High-order information matters: Learning relation and topology for occluded person re-identification,” in *Proc. IEEE/CVF Conf. Comput. Vision Pattern Recognit. (CVPR)*, 2020, pp. 6449–6458.