Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 3, No. 1, 2023

Unsupervise contrastive learning person re-identification Method based on Transfomer

Duiyu Chen, Jing Chen and Chengxu He

School of Physics and Optoelectronic Engineering, Guangdong University of Technology, Guangzhou 510006, China

Abstract: Unsupervised learning person re-identification is a meaningful and challenging person retrieving problem resulting from its application on secure surveillance and missing of label. To handle the lack of person identities, CNN-based method has been proposed to use self-contrastive learning based on a memory dictionary. However, CNN's partial bias is not well addressed in the field of ReID. To overcome these limitations, we propose a transformer-based framework called (TransCC). Specifically, we take the ViT transformer encoder trained on ImageNet as the feature extraction network, which makes up for some disadvantage of the CNN model. In addition, to overcome the issue of cluster inconsistency in instance-based memory method, unique cluster vectors are used to represent clusters in Characteristic space stored in the memory dictionary. By updating the cluster centriods and calculating contrastive loss, the model can be optimized and learn person-identifying traits. The experiment results outperform that of most methods, which demonstrate the effectiveness of our approach on USL ReID.

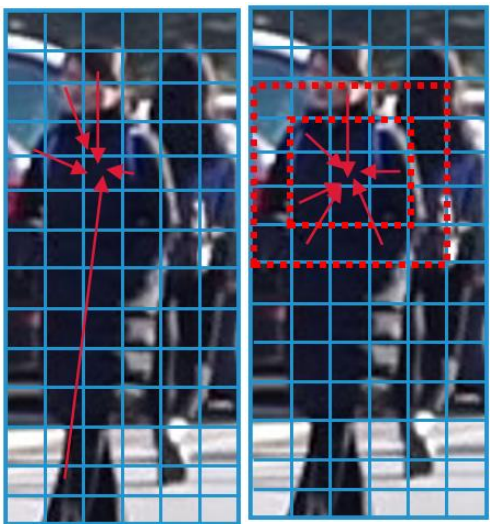
Keywords: Unsupervised learning; Person re-identification; Cluster-wise contrastive learning; Transformer-based.

1. Introduction

Widely regarded as a sub-problem of image retrieval, person re-identification (ReID) aims to judge whether a particular pedestrian is present in an image or video sequence. It has caught the eye of many researchers as a key component of intelligent surveillance and pedestrian tracking. In real life applications, ReID still faces multiple challenges, such as target occlusion, the deviation of perspective from different cameras and difference in appearance due to clothing or hair. Strictly speaking, each person is treated as a class, which leads to supervised learning of person re-identification requiring countless labels. In practice it is more suitable for open world tasks. Because it is time-consuming and expensive to annoate pedestrians' images across cameras, more research tends to focus on unsupervised person re-identification [1,2].

Without the abundant person annotation images, unsupervised person ReID (USL ReID) excels at learning Pedestrian re-recognition excels at learning discriminative features from unlabeled images by training the model, similar to contrast learning or self-training approach [3,4,5]. For a long time,CNN-based methods have been dominated USL ReID feature exaction. By reviewing convolution structure, we find that two fatal cissues for person identification. Firstly, exploiting the rich structural patterns in a global scope is crucial for object ReID[6]. The induction bias of CNN believes that information has spatial locality, which makes the CNN-based method have the problem of limited receptive field (see Figure 1.a). Secondly, fine-grained information is also important to ReID. However, the spatial resolution of image features is reduced by down-sampling operation, such as pooling, compromising discrimination ability to identify similar looking objects. Luo et.al [7] propose a trick to change the last stride of down-sampling from 2 to 1. Higher spatial resolution brings significant improvement. However, detail information is still lost during down-sampling.

Recently years have witnessed the widespread success of transfromer on computer vision. Vision Transformer (ViT) [8] and distillation Transformer (DeiT) [9] shine in the field of image classification, and Deformable Transformer (DETR) [10] leads the new research trend in object detection. Through the multi-head self-attention mechanism, transformer adaptively aggregate more perceptual attention from global features. Transformer exploit self-attention on the whole image at the patch level, thus representation obtained have long-range dependencies and lack of local induction bias like CNN (see Figure 1.b). In conclusion, the comprehensive attributes of the transformer are better than that of CNN, which cater to the need of capturing the global context information for person re-identification. In this paper, wo introduce pure transformer to replace CNN as the framework. Due to lack of a prior induction, the model do not converge well when trained on insufficient data. Then a ImageNet-pretrained model can work at the small datasets.



(a) (b)

Figure 1. The sight of CNN and transformer. (a) CNN aggregates relational features from pixels locally and focus on small regions. (b) Transformer explore long-range dependencies, thereby calculating attention mechanisms globally.

State-of-the-art unsupervised re-ID methods train a neural network with a memory-based dictionary without true label.

Sample representation vectors stored in the dictionary are assigned with pseudo labels to calculate contrastive loss. However, the varying cluster size makes the dictionary inconsistent after each iteration of cluster update. Previous work mostly stores and update instance features in the memory dictionary, and then calculate the contrastive loss at the instance level. Inspired by cluster contrast [11], we propose a cluster-wise memory-based contrastive learning method. Specifically, we use unique mean vectors to represent clusters, and make a memory dictionary with clusters centriods. Then a cluster-based contrastive loss is used to narrow distance between query and cluster.

The main contributions of our work are summarizing as three-fold: 1) We introduce Transformer structure into USL ReID to take advantage of its long-range dependencies to learn global representation of images. 2) we propose a cluster instance dictionary to storage sample assigned by pseudo label. 3) wo verify that our proposed unsupervised approach can achieve good performance on two common ReID datasets Market1501, MSMT17.

2. Relation work

2.1. Unsupervised person re-identification

Unsupervised ReID can be summarized into two categories in terms of solving ideas. One is unsupervised domain adaptive ReID (UDA ReID) [12,13,14], which transfer knowledge learning from source domain with true label to unlabeled target domain. Specifically, they gained a pre-trained model by supervised learning on th source domain, and fine-tune parameters of the framework with target domain data. IDM [15] puts forward a intermediate domain to bridge the gap between target and source domain, which can improve model's discriminability on the unlabeled target domain The second category is pure unsupervised learning person re-identification (USL ReID), which train model to learn discriminative representation of pedestrians from unlabeled image data [16,17]. Fu et.al [18] use the self-similarity grouping method to learn fine-grained features from the whole image and different part which divide an image into two stripes horizontally. Ge et.al [16] proposed self-paced contrastive learning with hybrid memory, and exploited reliable clusters to update the instance in the memory bank. However, those methods still do not break through the limitations of CNN, and the learned representation will be affected by successive pixels. In this paper,we propose a strong baseline that introduces the pure transformer to the application of USL ReID.

2.2. Transformer in vision

Transformer model is proposed to solve sequential data in terms of natural language processing (NLP). ViT [8] processes a picture as a bunch of one-dimensional signals like sequence data which called patch, and introduce transformer into the field of computer vision. He et.al [19] are the first to conduct research in the application of Transformer on object ReID. Besides, side information embedding encodes camera or viewpoint into embedding representations, similar to the position embedding. TransReID achieves a remarkable improvement over state-of-the-art CNN-based methods in supervised person re-identification. Zhong et.al [20] takes advantage of both CNN and Transformer and propose a Hierarchical Aggregation Transformer (HAT) for image-based person Re-ID with high performance. HAT uses ResNet-50 to extract features on the list, while Transformer integrating multi-scale features from a global view is used after each stage of ResNet-50. Because lack of the inductive bias, ViT training may result in catastrophic failure or accuracy degradation, Luo et.al [21] investigate self-supervised learning (SSL) methods with Vision Transformer which is pre-trained on unlabeled ReID dataset LUPerson and find it significantly surpasses ImageNet pre-trained models on ReID downstream tasks.

3. Methodology

In this section, wo build a Transformer-based cluster contrast (TransCC) framework for unsupervised person re-identification. Figure 2 shows the overall framework we proposed.Section 3.1 describes Transformer architecture as the feature encoder of TransCC. Then a cluster dictionary is designed to storage cluster representation for contrastive learning in Section 3.2. Finally, we explain the details of cluster contrast including its initialization and update, as well as the optimization of the whole pipeline.

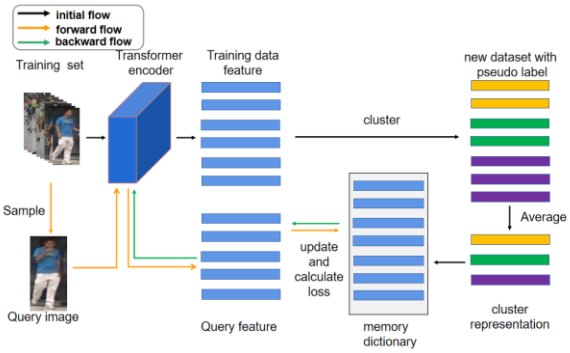


Figure.2 Illustration of the overall framework

3.1. Feature exaction-based Transformer

In USL ReID, we are given an unlabeled dataset $X = \{x_1, x_2, \cdots, x_N\}$ denoted the training set with N instances. We denote backbone network as $g = f_\theta(x)$, through which the input X is mapped to the hidden representation $G = \{g_1, g_2, \cdots, g_N\}$ as shown in Figure.3. Given an image $x \in \{H \times W \times C\}$,where h,w,c denote its height,width and number of channels, a convolution layer split it into fixed-sized patches. CNN-based can increase the receptive field as the network deepens, while ViT split images into non-overlapping patches, losing local continuity around the patches. To solve such problem, we use slide convolution kernel to split into overlapping patches. The fixed-size of patch and sliding stride of kernel denote K nd S respectively, then the area covered by two adjacent patches is $(K - S) \times K$ and a image will be split into N patches. As S becomes smaller, the image will be split into more patches, which will improve the performance and increase the calculation cost.

$$N = N_H \times N_W = \left\lfloor \frac{H + S - K}{S} \right\rfloor \times \left\lfloor \frac{W + S - K}{S} \right\rfloor \quad (1)$$

Following patches, the linear projection operation will map each patch to D dimension, generally 768. A extra learnable embedding token called cls is assigned to each image and denoted as its global feature. Then an image xi will be coded into input sequences, denoted as $z_i = \{cls, p_i^1, p_i^2, p_i^3, \cdots, p_i^N\}$, which are then fed to the transformer layers. Spatial information is attached by adding learnable position embedding to z_i. p denotes the patch of a image.

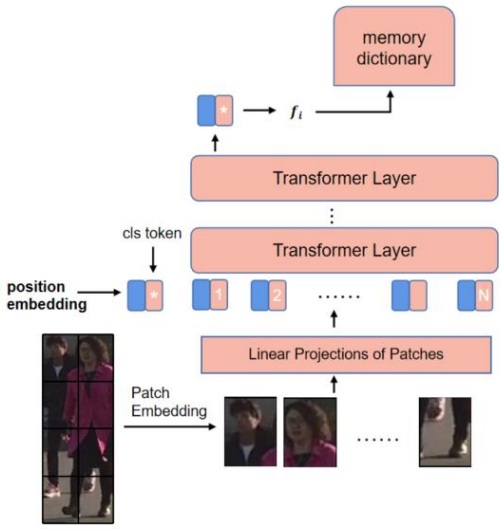


Figure.3 Overview of Transformer-based feature encoder

Transformer encoder is made up a stack of B transformer layers. The Transformer layer consists of multi-head self-attention block (MSA) and MLP block. Each layer starts with an LN layer normalisation, and the residual links are added to the output. Finally, we use the encoder's cls token (z0) as the global feature of the picture.

$$z_l' = MSA(LN(z_{l-1})) + z_{l-1} \tag{2}$$

$$z_l = MLP(LN(z_l')) + z_l' \tag{3}$$

3.2. Cluster contrast

When robustness features are learnt,we focus on a contrast learning approach for unsupervised learning due to the absence of ground-truth. After exacting feature vectors, we apply clustering algorithms to assign same pseudo label for instances that cluster similarly, for example, DBSCAN or InfoMap. Such pseudo labels are refined as adjacent consistency, that two images are assigned with the same label if they share similar neighbors. As the network is updated, the data distribution changes. Therefore, the pseudo labels are regenerated and more reliable at the next epoch. In training, we discard outlier points from the clustering process and assign pseudo-labels y_i to the clustered data, which denotes the person identity assigned to an image x_i by clustering algorithms.

Memory initialization-based cluster Assume that the number of clusters per epoch is K (K is changeable), we have an updatable dataset $\{(x_i, y_i)|y_i \in K)\}$ with pseudo label. Then we storage K cluster distribution representations in memory dictionary. The memory dictionary is initialized by the mean vector of instance features in each cluster,that is

$$\phi_K = \frac{1}{N_K} \sum_{x_i \in C_K} f_\vartheta(x_i) \tag{4}$$

where C_K and N_K denote the K-th cluster and its number of instances. Thus ϕ_K can represent the feature distribution of center of cluster K.

Memory update During training, P person identities and a fixed number of X instance of each identities from the new dataset are sampled in a mini-batch. It is worth noting that all cluster representation are stored in the memory dictionary in the form of cluster centroids. Previous methods of memory-based contrast learning store query into dictionary and update feature vector both at instance level by Eq (5).

$$f_i \leftarrow mf_i + (1-m)q \tag{5}$$

Where f_i denotes instance feature stored in the dictionary, q denotes the query feature, and the size is usually batch-size. And m is a momentum updatig factor. Updating cluster dictionary for Instance feature can result cluster inconsistency. Specifically, the sizes of clusters are distributed inconsistently. In the training iteration, a small number of query feature vectors, relative to a large cluster, will be updated, while in smaller clusters all instance features will be updated. Then an instance-wise contrastive loss is used to measure the distance between query and cluster feature, as follow.$\tau$is a temperature hyper-parameter. Then f denotes the feature vector in the memory dictionary and that one with subscript "+" is those instance feature that share the identity with query.

$$L_q = -log \frac{exp(q \cdot f_+/\tau)}{\int_{i=0}^{N_K} exp(q \cdot f_i/\tau)} \tag{6}$$

Through investigation for instance-wise instance, we find two issue: (1) Only the features of the query are updated each iteration, which is insignificant compared to the whole dataset. That is inefficient and inconsistent oscillatory distribution of mini-batches. (2) Update the clustering imbalance. Fewer updates to small clusters result in performance degradation for instances with small samples. To solve those problems, we propose our cluster contrastive learning method that the vectors that we store in the dictionary and use to calculate contrastive losses are no longer instance features but cluster representations. In training iteration, The formula of cluster update is shown as follows:

$$\phi_K \leftarrow m\phi_K + (1-m)q \tag{7}$$

Then we calculate our cluster-wise contrastive loss, defined by InfoNCE loss as follows:

$$L_q = -log \frac{exp(q \cdot \phi_+/\tau)}{\int_{i=0}^{K} exp(q \cdot \phi_K/\tau)} \tag{8}$$

By Eq. (7) and Eq. (8), our proposed algorithm uses unique vectors to represent each cluster identity and remains distinct throughout the updating process,avoiding the problem of unbalanced cluster during instance update process.

4. Experiments

4.1. Datasets and Evalution Metrics

To verify the performance of our proposed methods, we conduct experiments over benchmark datasets, i.e., Market1501, MSMT17. Tabel.1 describes the details of two datasets in our evaluation experiment, which are listed according to their essential information (number of person identities, imanges, train sets, query sets and gallery sets, image size and annoation method). In our experiments, we use the common evaluate metrics: Rank 1/5/10 of Cumulative Matching Characteristic (CMC) and mean average precision (mAP) and no post-processing tricks like re-rank are adapted in test.

4.2. Implementation Details

To facilitate the conduct of experiment,we fix the size of all images to 256 ×128 and use random horizontal flip, random crop and random erasing as data augmentation

technologies for training. At the beginning of each epoch, we use Transformer as backbone encoder of the feature extractor to obtain the feature of the whole dataset, assign pseudo label by cluster algorithms al as well as initialize the memory dictionary with the cluster representation. In a mini-batch, 64 images of 4-person identities are sampled for training (16 images for per person). We use Adam optimizer to update our TransCC model with weight decay 5e-4. The initial learning rate is set to 3.5e-4, then is reduced by 10 every 20 epochs in a total of 50 epochs.

Our method is implemented on the Pytorch, and four Nvidia RTX3090 GPUs are used for training and only one GPU is used for testing.

Table 1. Properties of the three benchmark datasets used in our experiment

Datasets	ID	Images	Train	Query	Gallery	Crop szie	Annotation method
Market1501	1201	32217	12936	19732	3368	128*64	Semi-automated (DPM)
MSMT17	4101	126441	32621	82161	11658	Variable	Faster R-CNN

4.3. Comparison experiment

we compare TransCC with unsupervised person re-identification method based on CNN. The experimental results are shown in Table2. The reslut of our proposed method boost performance by 3.6% and 2.8% in mAP and rank-1 accuracy compared with SPCL on Market1501, while 14% and 20.7% improvement on MSMT17. Compared with SSG, our method surpasss 24.3% and 13% in term of mAP on Market1501 and MSMT17 respectively. This is because the TransCC method proposed in paper exploit Transformer with overlapping-patch as extraction network which can help model to learn more global discriminative features to build a long-distance dependence, as well as adjacent features around patches.

Table.2 Performance comparison between USL ReID

Method	Market1501		MSMT17	
	mAP	R1	mAP	R1
MMCL	45.5	80.3	11.2	35.4
HCT	56.4	80.0	-	-
SSG	58.3	80.0	16.6	37.6
SPCL	76.2	90.2	19.1	43.3
TransCC(ours)	82.6	93.1	33.1	63.3

We also conduct experiments with different cluster algorithm, i.e. DBSCAN and Infomap on Market1501. The cluster proess is also important to unsupervised learnig, since the relible pseudo label can improve model'performence significally. The result of experiment is shown in Table.3.We can see that no matter which algorithm, our method can achieve good performance.

Table.3 Ablation study with cluster algorithm

Method	Market1501			
	mAP	R1	R5	R10
Ours/DBSCAN	82.6	93.1	96.7	97.9
Ours/Infomap	82.0	92.9	97.3	98.0

5. Conclusion

In this paper,we investigate a transformer-based framework for unsupervised learinng person re-identification task. A Pure transformer with overlapp-patch is used to extract discriminative feature, while a cluster-wise memory dictionary stored the whole cluster centriods is used to implement contrastibe loss. We believe that Transoformer has more potential to devoelop on ReID tasks. Experiments show that Transformer pipeline with cluster contrast surpassing mostly USL ReID methods.

References

[1] M. Zabłocki, K. Gosciewska, D. Frejlichowski, and R. Hofman, "Intelligent video surveillance systems for public spaces–a survey," Journal of Theoretical and Applied Computer Science, vol. 8(2014) no. 4, pp. 13–27.

[2] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. H. Hoi, "Deep learning for person re-identification: A survey and outlook," IEEE Trans. Pattern Anal. Mach. Intell.vol. 44(2022) pp. 2872–2893.

[3] D. Wang and S. Zhang, "Unsupervised person re-identification via multilabel classification," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 10 981–10 990.

[4] Z. Dai, G. Wang, W. Yuan, S. Zhu, and P. Tan, "Cluster contrast for unsupervised person re-identification," in Proceedings of the Asian Conference on Computer Vision, 2022, pp. 1142–1160.

[5] Yang, H.-X. Yu, A. Wu, and W.-S. Zheng, "Patch-based discriminative feature learning for unsupervised person re-identification," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 3633–3642.

[6] Z. Zhang, C. Lan, W. Zeng, X. Jin, and Z. Chen, "Relation-aware global attention for person re-identification," 2020, pp. 3186–3195.

[7] H. Luo, Y. Gu, X. Liao, S. Lai, and W. Jiang, "Bag of tricks and a strong baseline for deep person re-identification," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, 2019.

[8] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly et al., "An image is worth 16x16 words: Transformers for image recognition at scale," arXiv preprint arXiv:2010.11929, 2020.

[9] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jegou, "Training data-efficient image transformers & distillation through attention," in International conference on machine learning. PMLR, 2021, pp. 10347–10357.

[10] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable detr: Deformable transformers for end-to-end object detection," arXiv preprint arXiv:2010.04159, 2020.

[11] Z. Dai, G. Wang, W. Yuan, S. Zhu, and P. Tan, "Cluster contrast for unsupervised person re-identification," in Proceedings of the Asian Conference on Computer Vision, 2022, pp. 1142–1160.

[12] Z. Zhong, L. Zheng, Z. Luo, S. Li, and Y. Yang, "Learning to adapt invariance in memory for person re-identification," IEEE Trans. Pattern Anal. Mach. Intell., vol. 43, no. 8, pp. 2723–2738, 2020

[13] K. Zheng, W. Liu, L. He, T. Mei, J. Luo, and Z. J. Zha, "Group-aware label transfer for domain adaptive person re-identification," pp. 5306– 5315, 2021

[14] Z. Zhong, L. Zheng, Z. Luo, S. Li, and Y. Yang, "Invariance matters: Exemplar memory for domain adaptive person re-identification," 2019, pp. 598–607.

[15] Y. Dai, J. Liu, Y. Sun, Z. Tong, C. Zhang, and L.-Y. Duan, "IDM: An intermediate domain module for domain adaptive person reid," 2021, pp. 11 844–11 854.

[16] Y. Ge, F. Zhu, D. Chen, R. Zhao et al., "Self-paced contrastive learning with hybrid memory for domain adaptive object reid," Advances in Neural Information Processing Systems, vol. 33, pp. 11 309–11 321, 2020.

[17] H. Chen, B. Lagadec, and F. Bremond, "ICE: Inter-instance contrastive encoding for unsupervised person re-identification," 2021, pp. 14960– 14969.

[18] Y. Fu, Y. Wei, G. Wang, Y. Zhou, H. Shi, and T. S. Huang, "Selfsimilarity grouping: A simple unsupervised cross domain adaptation approach for person re-identification," 2019, pp. 6112–6121.

[19] S. He, H. Luo, P. Wang, F. Wang, H. Li, and W. Jiang, "Transreid: Transformer-based object re-identification," in Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 15 013–15 022.

[20] G. Zhang, P. Zhang, J. Qi, and H. Lu, "Hat: Hierarchical aggregation transformers for person re-identification," in Proceedings of the 29th ACM International Conference on Multimedia, 2021, pp. 516–525.

[21] H. Luo, P. Wang, Y. Xu, F. Ding, Y. Zhou, F. Wang, H. Li, and R. Jin, "Self-supervised pre-training for transformer-based person reidentification," arXiv preprint arXiv:2111.12084, 2021.