

Spatio-Temporal Enhanced Unsupervised Person Re-Identification with 3C-Based Clustering and Transformer Features

Ziming Zhu

*Qian Xuesen Academy, Xi'an Jiaotong University, Xi'an, China
3652169852@stu.xjtu.edu.cn*

Abstract. Recently, unsupervised person re-identification still suffers from the absence of reliable annotations. Clustering algorithms are highly dependent on noise, and some clustering-based approaches always lack stability due to unreliable pseudo-labels. In previous work, it is clearly shown that simply using visual similarity cannot resolve the above problem. This work developed a unified approach by fusing the spatiotemporal model, 3c-based clustering approaches, and transformer-based feature extractions. The paper adopted a DSTA backbone network to obtain better global representation, and it no longer treated the temporal information as auxiliary information, instead consolidating temporal information directly to estimate similarity, it is guided by Harmonic Difference Clustering in the clustering phase. This paper also adds a camera-aware weighting strategy to learn contrast. Experiments on the Market-1501 dataset demonstrate that the proposed framework in this paper can already make reasonable pseudo-labels. The results suggest that combining temporal priors with clustering mechanisms leads to more structured feature distributions.

Keywords: Unsupervised reid, Spatio-temporal modeling, Clustering algorithm

1. Introduction

Person re-identification has long been formulated as a supervised problem with identity labels to guide learning features and matching [1, 2]. Although it provides a simple optimization problem, this supervised learning problem requires a large-scale annotated dataset, which is expensive to label and impractical in real-world surveillance settings. To get over this obstacle, unsupervised person re-identification has caught much attention. Clustering-based methods are a dominating strategy for unsupervised person re-identification, where pseudo labels are initialized by clustering feature representations and then refined in the learning process [3, 4]. This approach is effective but highly dependent on feature quality, and suffers from noisy pseudo labels. Previous works showed that dependence on visual similarity is not sufficient for robust matching. In practical applications, the pedestrian movements conform some spatial and temporal constraints across the camera network. Some works have attempted to utilize spatio-temporal information to improve the matching robustness [5, 6].

Meanwhile, recent advances on clustering strategies suggest that structure information within clusters also provide supervisory cues other than pairwise similarity. Methods such as cluster consistency learning emphasize the role of global structure to enhance the quality of pseudo-labels [7]. Transformer-based methods have also shown strong ability in modeling global dependencies, and have been successfully applied to person reidentification [8, 9] tasks. Inspired by these insights, this paper proposes a unified framework that fuses Transformer-based feature representation, spatiotemporal features and structure-aware clustering, with the aim of enhancing the quality of representation and stabilizing the pseudo-label.

2. Methodology

2.1. Problem formulation

The goal of unsupervised person re-identification is to match pedestrian identities across non-overlapping camera views without manual annotations. Compared with the supervised setting, deleting identity labels means more uncertainty introduced to feature learning and similarity estimation. This side effect poses two difficulties: feature representations become poorly discriminative as the optimisation objective is not supervised. At the same time, clustering-based pseudo-labelling is necessarily noisy, the error brought into feature representations by early iterations is potentially amplified during training process.

A straightforward strategy is to optimize feature extractors, and expect clustering quality to improve at the same time. However, these strategies ignore an important nature characteristic of pedestrian data, which is the intrinsic spatial and temporal structure of pedestrian appearance. In real world surveillance, pedestrian appearances are not independent events; they are shaped by the limitation of cameras. The transition of a pedestrian from one camera to another has implicit temporal and spatial consistency and provides information beyond visual similarity. Motivated by this observation, the paper presents a unified framework for feature representation, spatio-temporal modeling and clustering optimization to combine into a coupled learning scheme. Instead of treating these different components as independent, the proposed method emphasizes mutual influences among them, such that structure information can help representation learning and vice versa, to achieve more reliable pseudo-label generation.

2.2. Framework overview

Overall framework is progressive optimization, in which the model is learned in more refined steps with more structured constraints. In the first stage, this paper adopts a Transformer-based backbone to encode global features of the samples. The features represent a reasonably smooth embedding space but still suffer from ambiguities caused by poor supervision. To avoid ambiguity, spatio-temporal information is incorporated into the framework to reshape similarity relations between the samples. Specifically, image similarity is no longer the only criterion for matching; it is regulated by temporal consistency and the camera transition pattern. Fusion of information can lead to the suppression of spurious matches and induce more reliable similarity estimation.

Finally, the paper adopts a structure-aware clustering method to produce pseudo labels, which are optimized again by camera-aware contrastive learning. The stage not only promotes compactness of the clusters but also generalizes well for different camera views. The whole iteration framework can be viewed as a closed loop. Representation learning, similarity modelling, and clustering can feed each other mutually.

2.3. Transformer-based feature representation

To capture global contextual information, the paper use Transformer-based architecture as the backbone network. As compared with CNN, which bases on a local receptive field, Transformer performs long-range dependency calculation via the self-attention mechanism [10], which is particularly applicable to person re-identification, where the subtle appearance differences and global structural information are more important for identity discrimination. Given an input image x , it first splits into a sequence of patches and embedded to a latent space. The Transformer encoder then transforms the sequence into a feature representation f . To stabilize the similarity computation, the paper also normalizes the feature vector to be $\hat{f} = f/|f|$. The normalization makes the effects of the feature magnitude remove, and enables similarity to come just from the angular distance.

Although the Transformer-based features lead to a better representation, at the same time they lead to a finer granularity of the embedding space. Hence, samples that belong to the same identity can be grouped in several different sub-clusters, so clustering becomes more complicated. Hence additional constraints need to be enforced to regularize the feature space.

2.4. Spatio-temporal consistency modeling

In practical surveillance systems, each pedestrian sample is associated with metadata such as camera ID and timestamp. Even in the absence of identity labels, this information remains accessible and can be leveraged to enhance similarity estimation.

For two samples i and j , the paper first computes their visual similarity using cosine similarity [11]:

$$s_v(i, j) = \hat{f}_i \cdot \hat{f}_j \quad (1)$$

While this measure captures appearance similarity, it does not account for temporal feasibility. For instance, two visually similar images captured at incompatible time intervals are unlikely to correspond to the same identity.

To incorporate temporal information, the paper defines a temporal consistency function based on a Gaussian kernel [6]:

$$s_t(i, j) = \exp\left(-\frac{(t_i - t_j)^2}{2\sigma^2}\right) \quad (2)$$

where σ controls the sensitivity to time differences. This formulation assigns higher similarity to samples with closer timestamps, reflecting the continuity of pedestrian movement.

Instead of directly combining s_v and s_t through linear operations, the paper employs a nonlinear fusion strategy inspired by spatial-temporal modeling approaches. The fused similarity is defined as:

$$s_{ij} = \phi(s_v(i, j), s_t(i, j)) \quad (3)$$

where $\phi(\cdot)$ denotes a logistic transformation that balances the contributions of visual and temporal information. This design allows the model to dynamically adjust the importance of each modality, thereby improving robustness under challenging conditions such as occlusion or illumination changes.

2.5. Structure-aware clustering

Clustering plays a crucial role for unsupervised person re-identification and provides pseudo labels for further training. However, existing clustering methods [3, 12] only consider pairwise distances between samples and are very sensitive to noise. To make the clustering more robust, the paper presents a structure-aware clustering approach considering both the samples-to-samples structure and the sample-to-cluster structure.

For a sample i and a cluster C_k , the assignment score is defined as:

$$d(i, C_k) = \frac{1}{|C_k|} \sum_{j \in C_k} \frac{1}{|f_i - f_j| + \epsilon} \quad (3)$$

It quantifies the consistency between a sample and the whole cluster instead of using a single nearest neighbor. By coaggregating information from cluster members, the method eliminates the effect of outliers and ensures a more reliable assignment. Moreover, samples that do not have strong affinity to any cluster are ignored for pseudo-label generation. This purposeful selection filters out the low-confidence samples, avoiding the negative influence on the overall clustering quality, which is very important for the stable training dynamics.

2.6. Camera-aware contrastive learning

After giving pseudo labels, the paper trains the model with a contrastive learning objective [11, 13]. However, the standard contrastive learning uses all samples equally and ignores the diversity between camera views. In multi-camera settings, identities may change much in appearance due to appearance differences among cameras, viewpoints, illuminations and occlusions.

To address this issue, the paper introduces a camera-aware weighting scheme. For each cluster, the paper computes the entropy of its camera distribution:

$$H_k = - \sum_c p_c \log p_c \quad (4)$$

where p_c denotes the proportion of samples captured by camera c . Clusters with higher entropy correspond to identities observed across multiple cameras and thus contain richer variation.

Based on this observation, the paper assigns higher weights to samples belonging to high-entropy clusters. The modified contrastive loss is formulated as:

$$L = -w_i \log \frac{\exp(f_i \cdot m_{y_i} / \tau)}{\sum_k \exp(f_i \cdot m_k / \tau)} \quad (5)$$

where w_i is derived from cluster entropy and τ is a temperature parameter. This weighting strategy encourages the model to focus on more challenging samples, thereby improving cross-camera generalization.

2.7. Optimization strategy

Training operates iteratively. In the beginning, features are not so good and clustering results are not stable. Spatio-temporal modelling gives the initial guidance for coarse clustering to obtain better clusters. As training goes, features become better and the pseudo labels are more reliable, and provide a stronger supervision. An interesting observation here is that the optimization does not strictly follow a monotonic convergence. In fact, the loss stabilizes somewhere and reaches a

balance between exploration and refinement. Removing any component breaks the balance. Without spatio-temporal modelling, clustering results are noisy; without structure-aware clustering, quality of pseudo label drops; without camera-aware weighting, models tend to fit easy samples too easily.

3. Experiments

3.1. Experimental setup

Experiments on the Market1501 dataset are carried out, which is a typical dataset for testing the person re-identification model for the multi-camera environment [1]. The paper does not directly scale it to the whole Market1501 dataset. Instead, a trial ground setting with 1 000 images randomly selected in the dataset is adopted first. The proposed framework is trained in fully unsupervised way. No identity labels are used in training. Pseudo labels are learned by clustering and iteratively refined. The pipeline of training contains three stages: pretraining, spatio-temporal fusion and target domain learning.

Features are extracted by a Transformer-based backbone. The feature dimension is set to 768. Clustering feature is followed by structure-aware strategy and the number of clusters is set to 50 at initializing. The model is trained using contrastive loss with temperature scaling and weighting for contrastive learning in training using the entropy based weighting strategy to pay more attention to challenging samples.

3.2. Training dynamics analysis

Another interesting behavior of the model during training can be explained by the following findings. At the start of training, the loss value of the model is relatively high (57.78), which is natural since it lacks reliable supervision. However, a sharp decrease was observed after the first epoch (the loss becomes approximately 13.6), and the rapid convergence suggests that the model quickly acquires the coarse structural information in the feature space. Later, as training goes on, the loss remains in a narrow range (roughly 13.6 to 14.9). In contrast to the supervised learning case, where the loss usually decreases monotonically, the observed oscillation of loss reflects the uncertainty in pseudo-label supervision. It does not imply instability in learning, but instead reflects a good balance between feature refining and cluster re-assignment.

3.3. Clustering quality evaluation

Also, to better understand the clustering behaviour, the paper analyses both the number of clusters and the proportion of valid pseudo labels. The valid label number is about 850, and the number of cluster is 50. A high valid label ratio 85% means that most samples are classified into clusters and are assigned with high confidence. The relative high number of clusters also means that the space is divided into relatively finer regions. Such behaviour is consistent with Transformer representations, which are detailed feature distributions. However, an increased granularity of clusters may also lead to the problem of over-segmentation, i.e., samples that are assigned to the same identity but corresponding to different clusters. Such trade-off is a balance between discrimination and generalization.

3.4. Comparison with baseline methods

So as to measure the contribution of each component, the paper compares the proposed method with simple baselines. It compares the baselines in terms of clustering and training stability instead of common ranking metrics without using ground-truth labels.

Table 1. Comparison with baseline variants

Method	Backbone	ST Modeling	Clusters
Baseline	ResNet50	No	28
Transformer	DSTA	No	33
ST fusion	DSTA	Yes	38
Proposed	DSTA	Yes	50

As shown in table 1, replacing the CNN backbone with a Transformer will increase the number of clusters immediately, which means an improvement of feature sensitivity. Adding spatio-temporal fusion refines the similarity links further so that clustering becomes more stable. The full model that includes entropy-aware contrastive learning creates the most distinct partition. This phenomenon of progress clearly indicates that each component contributes in its own way and their combination leads to richer representation feature space.

3.5. Ablation study

To further investigate the contribution of each module, the paper performs ablation experiments by selectively removing components from the full model. The results are shown in Table 2.

Table 2. Ablation study

Configuration	Valid Labels	Clusters
Full Model	850	50
w/o ST Fusion	790	34
w/o Entropy Weight	820	41
w/o Structure Clustering	760	30

Removing spatio-temporal fusion greatly reduces clustering quality, in terms of both number of clusters and valid label number. The observation implies that temporal constraints can benefit refining of similarity relationship. 3) Similarly, removing the spatial and temporal fusion weighting greatly reduces cluster diversity, indicating that the model still biases itself towards easy samples. Reducing structure-aware clustering produces fewer clusters and lower valid label ratios, indicating lower robustness. Both conclusions indicate that each component in the method is indispensable, not duplicate. Instead, they complement each other to shape feature space.

3.6. Discussion

These experimental results also give several key insights. Raising the feature representation capacity alone is not enough. Transformer-based models create more detailed embeddings but, at the same time, create finer-grained clusters which can lead to over-segmentation. Spatio-temporal modeling alleviates this by incorporating consistency constraints. It does not only refine similarity scores but

restructure feature space in a meaningful way. Finally, entropy-based weighting encourages the optimisation towards more difficult samples. This improves discrimination, but at the same time also increases sensitivity. The observed balance between cluster count and stability shows that the model operates in a regime in which exploration and refinement are both beneficial. Overall, the proposed framework has stable training behaviour and effective clustering in a truly fully unsupervised regime.

4. Conclusion

By combining Transformer features, HDC clustering and entropy-aware for temporal-spate enhanced unsupervised person re-identification, this paper has proposed a person re-identification framework. Temporal constraints and cluster structure complement each other well. Visual features alone are too limited. Context plays an important role in the design of representation space. Future work could be adaptive clustering, and more efficient temporal modelling. As is, the proposed framework reveals already that meaningful structure emerges under unsupervised settings.

References

- [1] Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., & Tian, Q. (2015). Scalable person re-identification: A benchmark. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 1116–1124.
- [2] Sun, Y., Zheng, L., Deng, W., & Wang, S. (2018). Beyond part models: Person retrieval with refined part pooling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(11), 2805–2818. <https://doi.org/10.1109/TPAMI.2018.2796534>
- [3] Ge, Y., Chen, D., Zhu, F., Zhao, R., & Li, H. (2020). Self-paced contrastive learning with hybrid memory for domain adaptive object re-ID. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3730–3739.
- [4] Dai, Z., Wang, G., Yuan, W., Zhu, S., & Tan, P. (2021). Cluster contrast for unsupervised person re-identification. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, 1621–1630.
- [5] Lv, J., Chen, W., Li, Q., & Yang, C. (2019). Learning invariant representation for person re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(7), 1740–1753.
- [6] Chen, Y., Zhu, X., & Gong, S. (2017). Person re-identification by deep learning multi-scale representations. In *IEEE International Conference on Computer Vision Workshops (ICCVW)*, 2590–2600.
- [7] Zhong, Z., Zheng, L., Kang, G., Li, S., & Yang, Y. (2020). Random erasing data augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 13001–13008.
- [8] He, S., Luo, H., Wang, P., Wang, F., Li, H., & Jiang, W. (2021). TransReID: Transformer-based object re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 15013–15022.
- [9] Li, Y., He, J., Zhang, T., & Zhang, X. (2022). Transformer-based person re-identification: A survey. *arXiv preprint arXiv: 2206.04347*. <https://arxiv.org/abs/2206.04347>
- [10] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*.
- [11] He, K., Fan, H., Wu, Y., Xie, S., & Girshick, R. (2020). Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9729–9738.
- [12] Ester, M., Kriegl, H. P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 226–231.
- [13] Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning (ICML)*, 1597–1607.