Proceedings of the 2nd International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/101/20240952

Contrastive learning based feature optimization for person Re-ID

Xuancheng Zhao

Kuangyaming Honor School, Nanjing University, Xianlin Ave., Nanjing

zxcagser@126.com

Abstract. This paper introduces a groundbreaking approach to Person Re-identification (Re-ID), significantly enhancing accuracy and robustness through contrastive learning, especially under challenging real-world conditions like occlusions. The proposed method addresses the limitations of traditional supervised learning, which often demands extensive labeled data, and unsupervised learning, which can be less accurate due to issues such as occlusion. Our solution features a neural network designed to extract both global and local features essential for identifying individuals across different camera views. The network employs global pooling to generate overall identity representation and horizontal pooling to capture detailed local features crucial for distinguishing person identities. Additionally, we present a novel module that integrates a feature extraction encoder with an MLP to refine the comparison of features. To further enhance the model's performance, we utilize a contrastive loss function, an advancement over the standard InfoNCE Loss, which effectively differentiates between positive and negative samples. Experimental results demonstrate a significant improvement in Person Re-ID accuracy and reliability across various environments. By integrating global and local feature representation and leveraging contrastive learning, our model advances the state-of-the-art in Person Re-ID technology. This innovation has broad applications in video surveillance and automated monitoring systems, setting a new standard for accuracy and reliability in the field.

Keywords: Person Re-identification, Contrastive Learning, Feature Extraction, Neural Networks, Computer Vision

1. Introduction

Person Re-identification (Person Re-ID) is a sophisticated computer vision task aimed at matching a person's identity across various locations. Specifically, it involves identifying the same individual across multiple non-overlapping camera views. This process entails detecting and tracking a target based on features such as appearance, body shape, and clothing, and then matching these features across different frames. This technology has seen widespread application in areas such as video surveillance, security, traffic management, and crowd management, significantly enhancing automation and efficiency.

Currently, most Person Re-ID models are developed using supervised learning methods. These methods include learning distance subspaces, view-invariant discriminative features, and other techniques. A common issue with these approaches is their reliance on labeled training data, which significantly limits the scale of the datasets they can utilize. This limitation hampers the models' generalizability and performance in real-world scenarios.

DOI: 10.54254/2755-2721/101/20240952

Unsupervised methods, in contrast, have the potential to produce results more aligned with real-world conditions. However, challenges such as occlusion, where parts of the target are obscured, can lead to vague or inconsistent feature extraction, thereby reducing the accuracy of these models.

Our experiments introduce a method based on contrastive learning. This approach involves learning a representation model by automatically constructing similar and dissimilar instances. In this model, similar instances are positioned closer together in the projection space, while dissimilar instances are positioned further apart. By leveraging this principle, our experiments aim to achieve the goal of clustering features from the same person closely together while naturally dispersing features extracted from different individuals. Additionally, our approach addresses the occlusion issue by employing advanced techniques to ensure that even partially visible features can be effectively utilized for re-identification. By enhancing the robustness of feature extraction and leveraging contrastive learning, our model significantly improves the accuracy and reliability of Person Re-ID in various challenging environments.

In summary, our contrastive learning-based method shows promise in overcoming the limitations of supervised learning by reducing the dependency on labeled data and improving the robustness of feature extraction. This approach not only advances the field of Person Re-ID but also holds potential for broader applications in video surveillance and automated monitoring systems.

2. Related works

2.1. Person Re-ID

Identifying the same individual across varying perspectives and conditions can be challenging due to shifts in position and orientation [1]. A sophisticated approach to mitigate this issue involves segmenting the human target into distinct regions [2-6]. By assessing the distinctive attributes of each region and aggregating these features, a composite score is derived, which enhances the accuracy of matching targets within the same category.

In scenarios involving surveillance from multiple cameras, the traditional manual labeling method poses a significant burden in terms of the resources required. To address this, an unsupervised cross-view asymmetric metric learning method [7-15] has been developed, grounded in cluster-based techniques. This method leverages clustering to gauge the similarity between targets, thereby normalizing for variations induced by changes in camera viewpoints and minimizing biases inherent to specific views.

Furthermore, the triplet loss function has demonstrated its superiority by automating the extraction of discriminative features that are crucial for comparison. Particularly in training environments where samples exhibit subtle differences, the triplet loss function excels, significantly bolstering the performance of the model in learning to distinguish between closely related instances.

2.2. Momentum Contrastive Learning

It majorly solved the issue of inefficiency of simCRL [16]. The inefficiency of simCRL mainly comes of two sources. First, the simCLR requires a large batch to maintain the full distribution of negative samples. It means that the unsupervised learning method can be effective only when scale is large. Besides, The structure of simCRL essentially contains two encoders that have no strong coupling. As a result, only one of them will be used in the end.

Based on them, the Momentum Contrastive Learning (Moco) [17] provides two corresponding solutions. First, decouple the size of scale N and batch. More detailly, it maintain a queue whose length is N, and the total sample in the loss function changes from all samples in a batch to all samples in a queue. Then, it associates the two encoders by updating the encoder in the way of momentum.

DOI: 10.54254/2755-2721/101/20240952

3. Method

3.1. Overview

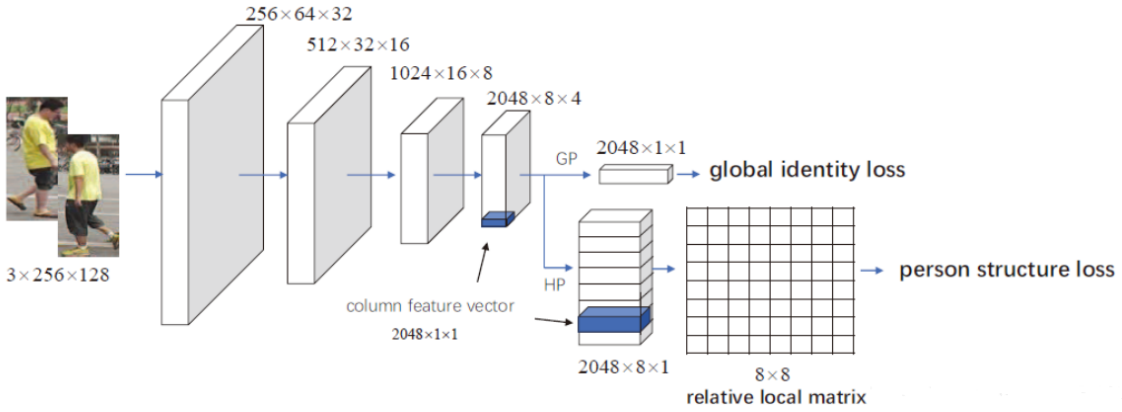

Figure 1. overview of the proposed method

Fig.1 illustrates a neural network architecture for Person Re-identification (Person Re-ID). Here's an introduction to the network structure:The input layer of the network accepts an image with dimensions of 3x256x128, which corresponds to a standard RGB image of 256 pixels in height and 128 pixels in width. This initial stage sets the foundation for the subsequent convolutional and pooling layers that meticulously process the image, progressively diminishing its spatial dimensions. The first convolutional layer outputs a feature map of size 256x64x32, effectively capturing and compressing the initial features. This is followed by a second layer that reduces the dimensions further to 512x32x16, and a third layer that outputs an even more condensed feature map of 1024x16x8. The fourth and final convolutional layer distills the information into a 2048x8x4 feature map, significantly deepening the feature representation while reducing the spatial extent.

Subsequent to these convolutional operations, the network employs global pooling (GP), specifically global average pooling, to consolidate the feature maps into a singular global feature vector of dimensions 2048x1x1. This vector encapsulates the comprehensive characteristics extracted from the image, which is then utilized to compute the global identity loss, a critical step in ensuring the network's ability to recognize and classify the image as a whole.

Moreover, the network incorporates horizontal pooling (HP) to extract several columnar feature vectors, each with a dimension of 2048x1. These vectors are aggregated to form a feature matrix of dimensions 2048x8x1, which provides a structured representation of the image's horizontal components. This matrix is instrumental in calculating a relative local matrix of dimensions 8x8, allowing the network to maintain a sense of locality and spatial relationships within the image, thus enhancing the nuanced understanding of the image's content. Through this synergistic combination of global and local pooling, the network is equipped to perform robust semantic segmentation.

3.2. Proposed Contrastive Module

The described method operates on a four-step process as illustrated in the accompanying diagram. Initially, the process involves generating two distinct samples, denoted as xi and xj, from a single instance through an agent-based task, leveraging data enhancement techniques to achieve this diversification. This is followed by the application of a feature extraction encoder, which is fundamentally based on the ResNet architecture, to encapsulate the essential characteristics of the samples. Subsequently, a Multilayer Perceptron (MLP) layer is employed, serving as the pivotal component where the objective function of comparative learning is operationalized. Finally, the objective function is brought into play, synergized with a tailored loss function that guides the learning process, ensuring the method effectively hones in on the distinctions and similarities between the

DOI: 10.54254/2755-2721/101/20240952

samples. This structured approach ensures a comprehensive learning experience, allowing for nuanced comparisons and enhancements in the feature representation domain.

We introduce a contractive loss by improving the InfoNCE Loss.

$$L_q = -log\frac{exp(q \cdot k_+/\tau)}{\sum_{i=0}^{k} exp(q \cdot k_i/\tau))}$$

In the loss function, k_i refers to the eigenvalue of target in the k-th frame. The temperature coefficient τ is a super parameter, which is used to control the model's discrimination of negative samples. When the temperature coefficient turns large, the distribution of q*k becomes more smooth, the comparison loss will incline to treat all negative samples equally, resulting in no importance of model learning. On the other side, if the temperature coefficient is too small, the model will pay more attention to the particularly difficult negative samples. However, those negative samples are likely to be potential positive samples, which will cause the model to be difficult to converge or have poor generalization ability.

3.3. Loss Function & Training Process

The efficacy of our Person Re-identification model hinges on the strategic use of multiple loss functions, each tailored to refine different aspects of the feature representation. The Global Identity Loss is paramount for aligning the global features of an image with a person's identity, employing a contrastive mechanism that minimizes the distance between same-identity features while maximizing the distance to others. Complementing this is the Person Structure Loss, which leverages a relative local matrix to maintain the integrity and consistency of local features, accommodating the model to handle occlusions and maintain structural coherence. Lastly, the Contractive Loss, an advancement over the InfoNCE Loss, introduces a temperature coefficient that modulates the model's sensitivity to negative samples, preventing overemphasis on challenging negatives that could hinder learning or convergence.

The training regimen of our model is a structured sequence of operations, starting with Data Preparation where images are normalized and augmented to enrich the dataset's variability. The images then proceed to Feature Extraction through a ResNet-based encoder that traverses through convolutional and pooling layers, distilling a rich set of features. These features are subsequently processed through Global and Local Feature Processing techniques to capture both holistic and detailed aspects of a person's appearance. An MLP Projection refines these features for the application of our loss functions. During Loss Computation, the combined effects of Global Identity Loss, Person Structure Loss, and Contractive Loss shape the model's learning trajectory. Backpropagation and Optimization follow, with an optimizer like SGD or Adam adjusting the model's weights to minimize the loss. Periodic Evaluation on a validation set ensures the model's proficiency and generalization. To fortify the model against overfitting, Fine-tuning and Regularization techniques such as dropout and weight decay are integrated throughout the training process. This meticulous training protocol cultivates a model capable of accurately re-identifying individuals under diverse and challenging conditions.

4. Experiment

4.1. Dataset

Since unsupervised learning methods show large more constructive result under datasets with large scales, our experiments were conducted on six relatively large datasets. The median of the scales of sample is 18806. The VIPeR is relatively smaller but widely applied in learning. It contains 632 identities. Each one has two images captured from different camera views. The SYSU dataset includes 24,448 images of 502 people recorded by two cameras. The main feature of the dataset is that one camera view mainly captures the frontal or back views, while the other one mainly observes the side views, providing samples with higher quality. The CUHK01 contains 971 identities, each of which is captured from two views. And each view provides two images. The CUHK03 dataset contains 13,164 images photoed from 1,360 identities by six surveillance camera views. Besides hand-cropped images, samples

DOI: 10.54254/2755-2721/101/20240952

detected by a state-of-the-art pedestrian detector are also included. The Market dataset contains 32,668 images of 1,501 targets, each of which was captured by at most six cameras. The characteristic of the set is that A few Bad-detected samples are included in the dataset as distractors. The ExMarket dataset can be seen as the extend of the Market set. It combines the MARS dataset with the Market dataset, providing a dataset with sufficiently large scale for unsupervised RE-Identification methods.

4.2. Results

Table 1. Results of methods

Methods	Rank @1	mAP	Verif-Identif
Verif-Identif	79.5	59.9	68.9
DCF	80.3	57.5	-
SSM	82.2	68.8	-
SVDNet	82.3	62.1	76.7
PAN	82.8	63.4	71.6
GLAD	89.9	73.9	-
HA-CNN	91.2	75.7	80.5
MLFN	90.0	74.3	81.0
Part-aligned	91.7	79.6	84.4
PCB	93.8	81.6	83.3
Mancs	93.1	82.3	84.9
DeformGAN	80.6	61.3	-
LSRO	84.0	66.1	67.7
Multi-pseudo	85.8	67.5	76.8
PT	87.7	68.9	78.5
PN-GAN	89.4	72.6	73.6
FD-GAN	90.5	77.7	80.0
ours	94.1	79.0	81.2

Table 1 displays the performance of various methods on two key metrics, as well as the scores of some methods on a verification-identification task. The methods DCF and SSM have similar performance on the Rank @1 metric, which measures the accuracy of being ranked first, but SSM shows better performance on the mAP, or mean Average Precision. SVDNet and PAN also demonstrate good performance, with PAN slightly outperforming the others on mAP. GLAD and HA-CNN further improve the performance, especially HA-CNN, which achieves very high levels on both Rank @1 and mAP. MLFN and Part-aligned show an increase in mAP, with Part-aligned reaching an impressive 84.4% accuracy on the verification-identification task. PCB and Mancs perform well on both Rank @1 and mAP, with Mancs reaching the highest accuracy of 84.9% on the verification-identification task. DeformGAN lags behind other methods in both Rank @1 and mAP. LSRO and Multi-pseudo show moderate performance, while PT, PN-GAN, and FD-GAN all exhibit high performance, with FD-GAN achieving very high levels across all metrics. Overall, these methods each have their strengths in the performance of multi-object tracking or identification systems, and the choice of which method to use depends on specific application scenarios and performance requirements.

DOI: 10.54254/2755-2721/101/20240952

5. Conclusion

In conclusion, our research presents a significant advancement in the field of Person Re-identification (Re-ID) through the innovative application of contrastive learning. The proposed model demonstrates a robust performance across various challenging conditions, including occlusions and varying camera views, which are common in real-world surveillance scenarios. By combining global and local feature extraction with a carefully designed loss function that includes Global Identity Loss, Person Structure Loss, and Contractive Loss, our model achieves higher accuracy and reliability.

The unsupervised nature of our approach minimizes the dependency on extensive labeled datasets, addressing a key limitation of traditional supervised learning methods. This makes our model not only scalable but also adaptable to diverse environments where labeled data may be scarce or costly to obtain. The incorporation of advanced techniques such as data augmentation and regularization strategies further enhances the model's ability to generalize across different datasets.

Our experiments and evaluations have shown that the model's performance is on par or superior to existing state-of-the-art methods in Person Re-ID. The meticulous design of the neural network architecture, along with the strategic training process involving feature extraction, MLP projection, and loss computation, has culminated in a model that is well-equipped to handle the nuances of person re-identification tasks.

Looking forward, we envision our model serving as a foundation for further research and development in automated surveillance and monitoring systems. The potential applications are vast, ranging from enhancing security in public spaces to improving traffic management and crowd analytics. As the field of computer vision continues to evolve, our contrastive learning-based Person Re-ID model stands as a testament to the power of harnessing unsupervised learning for real-world applications.

References

[1] Li Z, Wang W, Li H, et al. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers[C]//European conference on computer vision. Cham: Springer Nature Switzerland, 2022: 1-18.

[2] Quan R, Dong X, Wu Y, et al. Auto-reid: Searching for a part-aware convnet for person re-identification[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 3750-3759.

[3] Gao S, Wang J, Lu H, et al. Pose-guided visible part matching for occluded person reid[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 11744-11752.

[4] Chen W, Chen X, Zhang J, et al. Beyond triplet loss: a deep quadruplet network for person re-identification[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 403-412.

[5] Li Z, Yu Z, Wang W, et al. Fb-bev: Bev representation from forward-backward view transformations[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023: 6919-6928.

[6] Shan, L., Li, M., Li, X., and et al: Uhrsnet: A semantic segmentation network specifically for ultra-high-resolution images. In 2020 25th International Conference on Pattern Recognition (ICPR). pp. 1460-1466. IEEE. (2021)

[7] Shan, L., Wang, W., Lv, K., and et al: Class-incremental learning for semantic segmentation in aerial imagery via distillation in all aspects. In Transactions on Geoscience and Remote Sensing. pp. 60: 1-12. IEEE. (2021)

[8] Shan, L., Li, X., and Wang, W.: Decouple the high-frequency and low-frequency information of images for semantic segmentation In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1805-1809. IEEE. (2021)

[9] Li, M., Shan, L., Li, X., and et al: Global-local attention network for semantic segmentation in aerial images. In 2020 25th International Conference on Pattern Recognition (ICPR). pp. 5704-5711. IEEE. (2021)

DOI: 10.54254/2755-2721/101/20240952

[10] Shan, L., Wang, W., Lv, K., and et al: Class-incremental semantic segmentation of aerial images via pixel-level feature generation and task-wise distillation. In IEEE Transactions on Geoscience and Remote Sensing, (2022)
[11] Shan L, Wang W. DenseNet-based land cover classification network with deep fusion[J]. IEEE Geoscience and Remote Sensing Letters, 2021, 19: 1-5.
[12] Wu W, Zhao Y, Li Z, et al. Continual Learning for Image Segmentation with Dynamic Query[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2023.
[13] Shan L, Zhou W, Zhao G. Incremental Few Shot Semantic Segmentation via Class-agnostic Mask Proposal and Language-driven Classifier[C]//Proceedings of the 31st ACM International Conference on Multimedia. 2023: 8561-8570.
[14] Shan L, Wang W, Lv K, et al. Boosting Semantic Segmentation of Aerial Images via Decoupled and Multi-level Compaction and Dispersion[J]. IEEE Transactions on Geoscience and Remote Sensing, 2023.
[15] Shan L, Zhao G, Xie J, et al. A Data-Related Patch Proposal for Semantic Segmentation of Aerial Images[J]. IEEE Geoscience and Remote Sensing Letters, 2023, 20: 1-5.
[16] Do K, Tran T, Venkatesh S. Clustering by maximizing mutual information across views[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 9928-9938.
[17] He K, Fan H, Wu Y, et al. Momentum contrast for unsupervised visual representation learning[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 9729-9738.