Journal of Big Data

RESEARCH Open Access

Modified centroid triplet loss for person re-identification

Alaa Alnissany[1*] and Yazan Dayoub[2]

*Correspondence:
alaa.alnissany@hiast.edu.sy

[1] Department of Electronic and Mechanical Systems, Higher Institute for Applied Sciences and Technology, Damascus, Syria
[2] Department of Computer Science, HSE University, Moscow, Russia

Abstract

Person Re-identification (ReID) is the process of matching target individuals to their images within different images or videos captured from a variety of angles or cameras. This is a critical task for surveillance applications, in particular, these applications that operate in large environments such as malls and airports. Recent studies use data-driven approaches to tackle this problem. This work continues on this path by presenting a modification of a previously defined loss, the centroid triplet loss (CTL). The proposed loss, modified centroid triplet loss (MCTL), emphasizes more on the interclass distance. It is divided into two parts, one penalizes for interclass distance and second penalizes for intraclass distance. Mean Average Precision (mAP) was adopted to validate our approach, two datasets are also used for validation; Market-1501 and DukeMTMC. The results were calculated for first rank of identification and mAP. For dataset Market-1501 dataset, the results were 98.4% rank1, 98.63% mAP, and 96.8% rank1, 97.3% mAP on DukeMTMC dataset, the results outweighed those of existing studies in the domain.

Keywords: Person ReID, Triplet loss, Center loss, Inter class distance, Centroid triplet loss, DukeMTMC-ReID, Market-1501

Introduction

With the advent of technology and the increasing usage of cameras, the topic of re-identification has gained greater attention. It has many surveillance applications, such as detecting criminals, tracking individuals, and activity analysis [17, 30, 37]. The problem is summarized as the recognition of a target person through a variety of images captured at different times or with different cameras. This makes ReID complex problem due to its intraclass variations, and the fact that images can be taken in different conditions, different poses of the target person, or even different angles. Moreover, the target person may be partially visible due to being obscured by another item, such as another person.

While researchers used a variety of techniques to achieve this goal, the main concept is the same: to learn a person's discriminative features that aid in the matching process. Previously, this was accomplished manually by extracting and selecting features [1, 12, 39, 45]. However, with the advancement of deep learning model such as Convolution Neural Network (CNN), researchers moved their focus to deep learning techniques, in which models automatically learn pedestrian features. Yi et al in [36] proposed a Siamese

neural network that can learn the similarity between features directly from images, while Li et al in [15] presented a filter pairing neural network (FPNN) that handles the misalignment automatically. In comparison to prior methodologies, neural networks significantly improved the performance in this domain.

Recently, many deep learning methods have been proposed in this domain. Some works presented new loss functions [2, 11, 16, 38]. Others use an external source of information, for example pose estimation, to improve model performance [8, 19]. Generative Adversarial Networks (GANs) have also been used in ReID, they were mainly used to augment existing datasets with more images as they are captured from cameras [13, 33, 43]. Other works use re-ranking methods, which are post-processing techniques. For example, Zhong et al in [45] used a k-reciprocal encoding to encode an image using its k-reciprocal nearest neighbors, which is then used for re-ranking by the Jaccard distance.

Luo et al in [18] focused on developing a strong base model for ReID tasks. They used ResNet as features extractor and introduced novel neck structure named as batch normalization neck BNNeck with various training tricks to improve the model performance. Wieczorek et al in [35] proposed a modification to the retrieval method; they implemented retrieval using the centroid representations of identities, which saves memory and significantly reduces the amount of the required computation. Additionally, they introduced centroid triplet loss, which is a modification of triplet loss where negative and positive samples are replaced with clusters' centroids. They were able to achieve state-of-the-art results on different retrieval datasets.

In this paper, modified centroid triplet loss (MCTL) is proposed. The loss is broken into two terms. The first term is equivalent to the center loss and focuses on the intra-distance. The second term focuses on the inter-class to increase the margin between classes which makes it easier for the model to separate classes from each other. Using the described modification above, new results were achieved on both DukeMTMC-reID [22] and Market-1501 [41].

In market-1501 dataset for example, the gallery consists of 19,732 images, each $128 * 64 * 3$ pixels in size, this equates to 484 million data. As a result, ReID is a big data problem, and processing this volume of data using conventional approaches is challenging. While the use of centroids as the representations of identities resulted in a decrease in space and time, the necessity remains enormous and requires special attention.

Background and related work

In [20], Ming et al. grouped deep learning-based person ReID methods into four categories. These categories are: 1. depth metric learning, 2. local feature learning, 3. generative adversarial learning, and 4. sequences feature learning. The last category is for video-based ReID, so it will not be discussed in this paper.

Deep metric learning

In deep metric learning, different cost functions are introduced to help the network learn more robust features, and increase its generation ability in the presence of large environment variations. For instance, center loss was first introduced in [34]. It penalizes the intra-class distance and seeks compact clusters representations. Triplet loss

helps the network to output close features for images of the same identities while driving away images' features of different classes [14, 23, 31]. Another loss is classification loss, which helps the network to classify images, instead of measuring their similarities [42]. For example, although center loss reduces the intra-class distance, it doesn't consider inter-class distance, thus a trivial solution would be to output zero features for all images which achieve zero loss. It won't be difficult for the network to find this trivial solution. Therefore, recent studies use combinations of them [18, 35].

Feature learning

Feature learning can be subdivided into global and local features learning. This is based on how the network represents the image, with either a single global feature vector for the whole image or with many vectors taken from different parts of the image. In works based on global features (this work included), the network converts the whole image into one single embedding vector. This is done through a base network which is commonly an ImageNet pre-trained network, in this case, ResNet50. However, although this approach is more computation friendly, it is difficult for this method to capture the detailed information about the pedestrian [20].

On the other hand, methodologies for extracting local features are various. They can also be classified into subcategories [20]. Sun et al. [27] proposed a Part-based Convolutional Baseline (PCB). They divided the image into parts and aligned them using content consistency inside each part, rather than relying on external sources such as the pose estimator. Many subsequent works, such as multiple granularity network (MGN) [29] and spatial-channel parallelism network (SCPNet) [4] were inspired by PCB. Other researchers accessed local information via the attention mechanism. Sharma et al. [24] presented the Locally Aware Transformer (LA-Transformer), which employs a PCB-like strategy to combine the globally enhanced local tokens while keeping their ordering in correspondence with the image dimension. Other models make use of external data. For example, Song et al. [26] introduced a mask-guided contrastive attention model (MGCAM). They used a binary mask to learn features separately from the body and background regions. He et al. used both pose-guided and mask-guided approach to build a saliency heat-map to aid in the learning of stronger discriminative features [10].

Generative adversarial learning (GANs)

GANs were first introduced in 2014 by Goodfellow et al. [7]. They have been established as a prominent player in the domain of deep learning. Regarding ReID, GANs are used in data generation to increase existing datasets. For instance, Zheng et al. [43] used DCGAN to create unlabeled samples that were then combined with annotated real images to train the model semi-supervised. Zhao et al. [40] proposed an adversarial network for Hard Triplet Generation (HTG). GANs are also used to transfer styles between datasets or even cameras. For example, Zhong et al. [46] introduced camera style(CamStyle) to solve the problem of camera style disparities within the same dataset. Additionally, GANs are used to learn resolution invariant features of pedestrian images. For instance, Chen et al. [3] proposed Resolution Adaptation and re-Identification Network (RAIN) that can learn resolution-invariant representations for re-ID in

an end-to-end manner. Wang et al. [32] used a cascaded super-resolution GAN (CSR-GAN) to tackle the resolution mismatch problem.

All the studies in the domain of ReID are concerned with improving the results depending on ways that are not related to optimizer loss function. Even though updating the loss function to a new one might be complex, this will effectively improve the ReID problem and which was the target; to work on new ways of improvement such as loss function.

Objective of the paper

In general, loss function measures the difference between predicted output and actual output label. This difference is also known as training loss. It is a crucial metric that guides the optimization algorithm to adjust model weights and biases to minimize the training loss. The choice of the loss function depends on the nature of the problem to be solved (in this article, problem is ReID). For instance, regression problems require different loss functions, such as mean squared error, mean absolute error, or Huber loss, to estimate how well the model predicts a continuous value. In contrast, classification problems demand cross-entropy loss, hinge loss, or focal loss to calculate the performance of the model in predicting classes. Thus, selecting an appropriate loss function can ensure that the model trains effectively and converges to a robust solution for the given problem. In summary, the loss function is crucial for accurate training and optimization of deep learning models.

This study aims to improve the global feature baseline by introducing a loss function that takes into account both inter- and intra-class distances, which yields encouraging experimental results. Additionally, the introduced loss is applicable to problems involving general classification. The global feature approach is used because it is more efficient in terms of computation and storage than alternative methods.

Proposed method

Modified centroid triplet loss

Triplet loss aims to increase the similarity between images' features that belong to the same class while decreasing similarity to features of images from different classes. It compares a reference image (referred to as the anchor) to a positive image (an image belonging to the same class) and a negative image (an image from a different class). The distance between the anchor and positive image is minimized because they are from the same class with high similarity while maximizing the distance between the anchor image and the negative image. The loss function is described by the following formula:

$$L_{triplet} = \left\lfloor d\big(f(A), f(P)\big) - d\big(f(A), f(N)\big) + m \right\rfloor_{+} \tag{1}$$

Where d is a distance function, commonly euclidean distance, f learned embedding function by the network, A anchor image, P positive example, N negative example, m margin parameter, $\lfloor x \rfloor_{+} = \max(x, 0)$.

However, because the triplet loss is particularly susceptible to the sample triplets (A, P, N), they must be carefully chosen. One technique is hard mining [11, 21, 25], in which a triplet is composed of an anchor, the farthest positive image from the anchor, and the nearest negative image. This can be expressed mathematically as follows:

$$L_{hm} = \lfloor \max\left(d\left(f(A), f(P)\right)\right) - \min\left(d\left(f(A), f(N)\right)\right) + m \rfloor_+ \tag{2}$$

Chen et al. [2] proposed that instead of a single negative example, they used two negative examples and refer to their loss as quadruplet loss. The triplet is changed into a quadruplet consisting of (A, P, N_1, N_2), with $(A, P, N1)$ defined as before and N_2 being the second closest negative example to the anchor. Mathematically:

$$\begin{aligned} L_{quadruplet} = &\lfloor \max\left(d\left(f(A), f(P)\right)\right) - \min\left(d\left(f(A), f(N_1)\right)\right) + m_1 \rfloor_+ \\ &+ \lfloor \max\left(d\left(f(A), f(P)\right)\right) - \min\left(d\left(f(A), f(N_2)\right)\right) + m_2 \rfloor_+ \end{aligned} \tag{3}$$

Where m_2 a margin constant satisfies $m_2 < m_1$, N_1, N_2 are chosen from different classes. This method contributes to increasing the inter-class distance by pushing negative samples further away.

Another adjustment was proposed in [35], where Wieczorek et al. changed the triplet to become (A, c_p, c_n) an anchor image A, the centroid of its class c_p, and c_n the centroid of a negative class. They called it centroid triplet loss, Mathematically:

$$L_{ctl} = \lfloor d\left(f(A), c_P\right) - \min\left(d\left(f(A), c_N\right)\right) + m \rfloor_+ \tag{4}$$

This work presents a modified version of the CTL loss to emphasize more on the inter-class distance. Our approach is to define a loss function that is subdivided into two distinct terms. The first term is a scaled center loss that attempts to get features as close to their center as possible, while the second term attempts to push features away from the negative centroids behind a defined margin. This can be written as follows:

$$L_{mctl} = \omega_1 \left(d\left(f(A), c_p\right)\right)^2 + \omega_2 \lfloor m - \min\left(d\left(f(A), c_N\right)\right) \rfloor_+^{\,2} \tag{5}$$

Where ω_1, ω_2 are hyper-parameters. Choosing $\omega_1 < \omega_2$, penalizes the inter-class more heavily. To facilitate hyperparameter selection, w_2 is set such that $w_2 = (1 - w_1)$.

The main difference between the two loss functions is in how they weight the importance of the inter-class distances. By using a squared distance term in L_{mctl}, the difference between the feature representation and the positive class centroid is amplified, making it more important to minimize this distance. Additionally, by giving higher weights to inter-class distances, the model is encouraged to place a larger emphasis on separating the individuals based on their unique characteristics, which can help to reduce the misclassification errors and improve the overall performance of the model.

ResNet50

ResNet50[1] is a variant of Residual Networks (ResNet) [9] with a depth of 50 layers. The architecture is shown in Fig. 1. It consists of a convolution layer, followed by a max-pooling layer, and four residual blocks with an average pooling layer at the end. A residual block contains a skip connection between its inputs and outputs. Each residual block contains a repeated three chained Conv layers, as shown in Fig. 2. In this study, the last

[1] Only the encoder part of ResNet50 is shown.

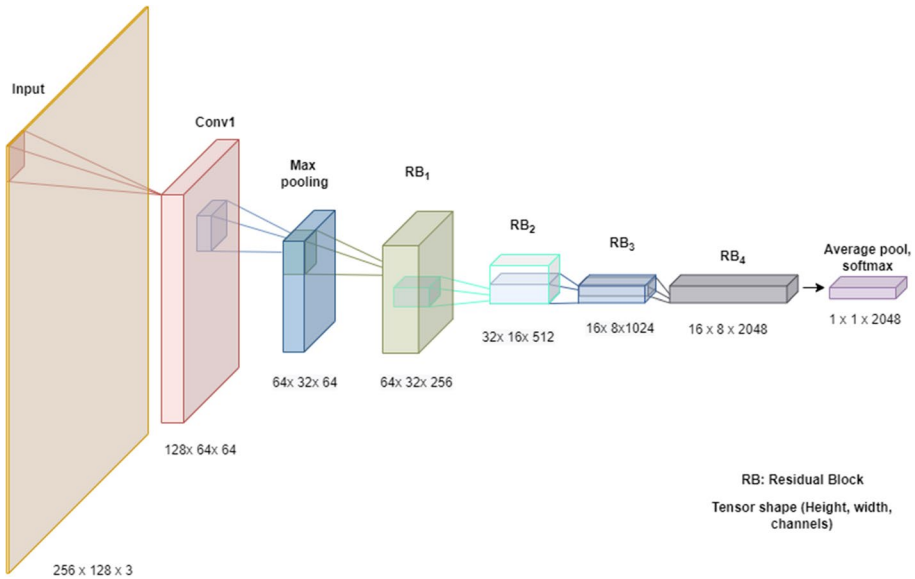

Fig. 1 ResNet architecture

Conv1:	**kernel (7×7), stride (2), channels (64)**
Max pooling:	kernel (3×3), stride(2)
Residual block1 : 3×	kernel (1×1), channels (64) kernel (3×3), channels (64) kernel (1×1), channels (256)
Residual block2 : 4×	kernel (1×1), channels (128) kernel (3×3), channels (128) kernel (1×1), channels (512)
Residual block2 : 4×	kernel (1×1), channels (256) kernel (3×3), channels (256) kernel (1×1), channels (1024)
Residual block2 : 4×	kernel (1×1), channels (512) kernel (3×3), channels (512) kernel (1×1), channels (2048)

Fig. 2 Resnet Detailed architecture

stride is modified to become 1 instead of 2 as proposed in [18], which helps in increasing the spatial dimensions of the feature map before converting it to a vector.

Network architecture

The network's architecture is shown in Fig. 2. It is based on [35]. The model is composed of a CNN feature extractor (encoder) and a classifier. The encoder, in this work ResNet50, takes a batch of images, made of C distinct classes with N instances per class; consequently, the batch size is equal to $C * N$. Images are converted into feature map and represented by embedding vectors of dimension d, where $d = 2048$ (number of channels in the last conv in resnet50) in this work. These features are used to calculate three losses: triplet loss, center loss, and modified centroid triplet loss (Fig. 3).

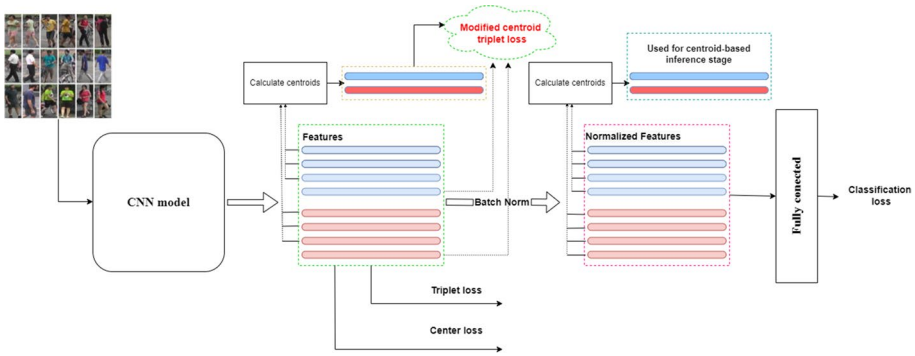

Fig. 3 Model architecture used in [35]

To calculate the classification loss, a classifier is used. The classifier plays an auxiliary role in helping the encoder learn more discriminative features. It consists of a 1D-batch normalization layer, which feeds its output to a linear layer augmented with a softmax activation function(BNNeck [18]). Categorical cross-entropy is used to calculate the classification loss.

Experiments

Two datasets were used: Market-1501 and DukeMTMC-reID. The model was trained and tested in Colab environment on GPU Tesla V100-SXM2–16GB.

Market-1501 was presented for the first time in 2015. The dataset is composed of images collected in front of a supermarket at Tsinghua University for 1501 distinct pedestrians, which is how it got its name. The dataset contains 32668 images of a 128 * 64 shape captured by six cameras.

DukeMTMC-reID is a subset of a bigger dataset DukeMTMC. Data is collected at Duke University from eight cameras. In total, the dataset contains 36411 images which are divided as follows: 16, 522 of 702 individuals for train, 2228 of an additional 702 for query, and 17, 661 for search gallery.

Implementation details

The implementation is based on.[2]. The same training procedure in [35] was followed with one sole modification, the replacement of the CTL loss with MCTL of $w_1 = 0.1$, $w_2 = 0.9$.

During training, data is augmented with: random horizontal flip, random erasing, and random cropping. To optimize the model's parameters, two optimizers were used: Adam and Stochastic gradient descent SGD. Adam optimizer was used to train the model for 120 epochs, with a multi-step scheduler that reduces its learning rate LR by a factor of 10 at epoch 40th, 70th epoch starting with an LR of 0.00035. SGD was used to optimize the center loss separately with an LR of 0.5. The total loss is a weighted sum of the four

[2] https://github.com/mikwieczorek/centroids-reid Github official repository for [35].

Table 1 Person re-identification results on market-1501 dataset

Market-1501				
Model	**mAP**	**rank1**	**rank5**	**rank10**
Unsupervised Pre-training [5]	96.21	97.03		
RGT & RGPR [6]	95.6	96.9		
LA-Transformer [24]	98.27	98.27		
st-ReID [28]	95.5	98.0	98.9	99.1
Ctl model [35]	98.3	98.0	98.6	99.5
Ours-Mctl	**98.63**	**98.4**	**99.1**	**99.6**

* The numbers reported in the paper are different from this reproted here, for more info, please visit their https://github.com/DengpanFu/LUPerson

Table 2 Person re-identification results on DukeMTMC-reID dataset

DukeMTMC-reID				
Model	**mAP**	**rank1**	**rank5**	**rank10**
Unsupervised Pre-training	92.77	93.99		
st-ReID	92.7	94.5		
Viewpoint-Aware Loss [44]	91.8	93.9	96.5	
Ctl model	96.1	95.6	96.2	97.9
Ours-Mctl	**97.3**	**96.8**	**98.4**	**98.9**

losses, where all losses are weighted with 1, except center loss is weighted by $5e-4$. The centroids are calculated from raw unnormalized features. Each sample from a class is treated as a query and the rest samples are used to calculate the centroids by averaging their features. Re-sampling is omitted because it adds noises to centroids vectors. Additionally, label smoothing is used in cross-entropy loss. During inference, the centroids are calculated first. Each class is represented with its centroid, which is the mean of all samples belonging to this class. Then query images are matched to centroids instead of gallery images.

Results and discussion

In Tables, 1, 2 model results are presented alongside those of other state-of-the-art models, highlighting the best results in bold. MCTL loss increased the model's performance greatly across all criteria. the results ensured that the proposed approach outperformed research approaches with the same tested datasets.

CMC for both datasets is shown in Fig. 4. Zooming in Fig. 4 to enlarge the interval [96, 100]% and this is shown in Fig. 5; this illustrates that the proposed approach is effective and the proposed MCTL can be adopted in problems of researches. As shown in Fig. 5, the model reached its maximum matching rate of 99.9% on market-1501 at rank-42, while it reached its maximum of 99.5% at rank-32 on DukeMTMC-reID.

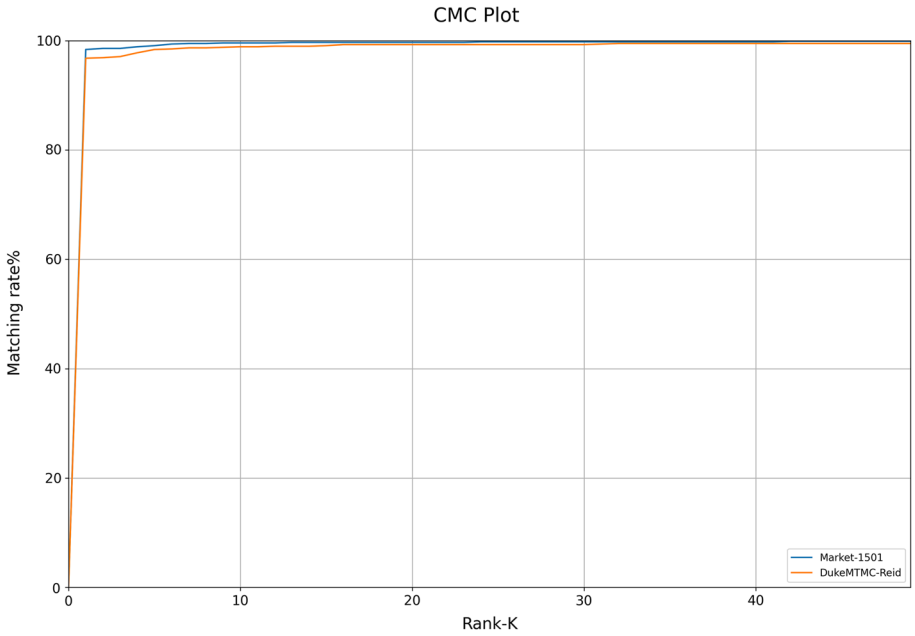

Fig. 4 CMC plots for results on Market and Duke datasets

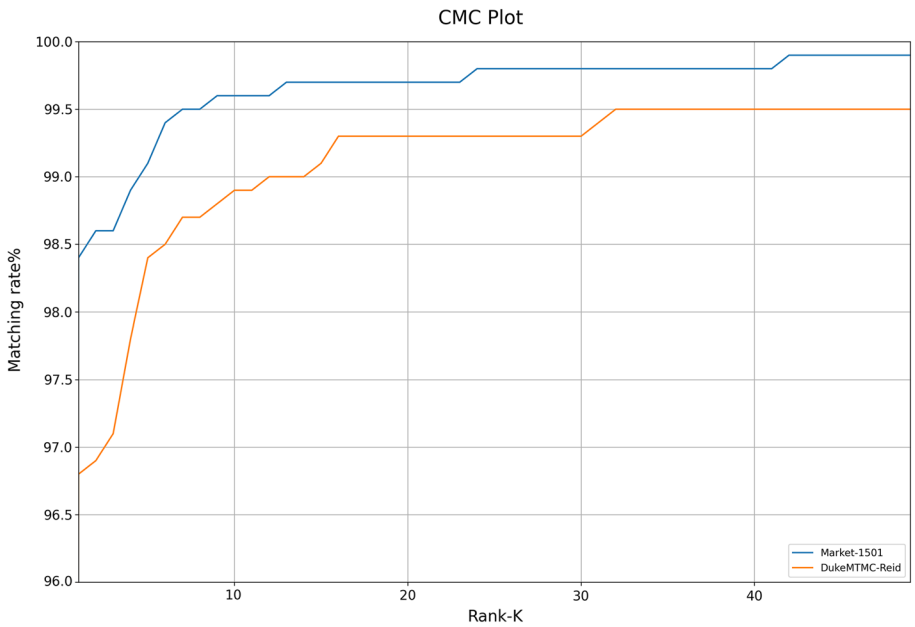

Fig. 5 Zoomed plot for CMC results on Market and Duke datasets

Conclusion

Modified centroid triplet loss for Reid model was presented in this article to improve performance. The developed method was validated on two datasets: the Market-1501 and DukeMTMC-ReID. The results show outperformance of the adopted approach compared to the already existing one. Despite this enhancement, additional research should be conducted. For example: experiments with various base models, investigation

of an euclidean classification loss function and investigation of the model's generality in cross-domain tasks.

Abbreviations

Ctl	Centroid triplet loss
Mctl	Modified triplet loss
FPNN	filter pairing neural network
GAN	Generative adversarial network
DCGAN	Deep convolutional generative adversarial network
RAIN	Adaptation and re-Identification Network
MGCAM	Mask-guided contrastive attention model
PCB	Part-based Convolutional Baseline
MGN	Multiple Granularity Network
SCPNet	Spatial-channel parallelism network
HTG	Hard triplet generation
CSR-GAN	Cascaded super-resolution generative adversarial network

Acknowledgements
The authors would like to thank Mr. Ammar Asaad for his helpful advice and comments throughout the development of this work.

Author contributions
Both authors contributed equally to the work.

Funding
The authors declare that they have no funding.

Availability of data and materials
The datasets used during the current study are available at these links: Market1501 dataset: https://drive.google.com/file/d/0B8-rUzbwVRk0c054eEozWG9COHM/view?resourcekey=0-8nyl7K9_x37HlQm34MmrYQ DukeMTMC-reID: https://drive.google.com/file/d/1jjE85dRCMOgRtvJ5RQV9-Afs-2_5dY3O/view.

Declarations

Ethics approval and consent to participate
The authors Ethics approval and consent to participate.

Consent for publication
The authors consent for publication.

Comprting interests
The authors declare that they have no competing interests.

Received: 29 March 2022 Accepted: 8 May 2023
Published online: 22 May 2023

References

1. An Le, Chen Xiaojing, Liu Shuang, Lei Yinjie, Yang Songfan. Integrating appearance features and soft biometrics for person re-identification. Multimed Tools Appl. 2017;76(9):12117–31.
2. Chen W, Chen X, Zhang J, and Huang K. Beyond triplet loss: a deep quadruplet network for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2017. p. 403-412.
3. Yun-Chun C, Yu-Jhe L, Xiaofei D, Yu-Chiang FW. Learning resolution-invariant deep representations for person re-identification. Proc AAAI Conf Artif Int. 2019;33:8215–22.
4. Fan X, Luo H, Zhang X, He L, Zhang C, and Jiang W. Scpnet: Spatial-channel parallelism network for joint holistic and partial person re-identification. In Asian conference on computer vision, 2018. pages 19–34. Springer,
5. Fu D, Chen D, Bao J, Yang H, Yuan L, Zhang L, Li H, Chen D. Unsupervised pre-training for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, 2021.
6. Gong Y. A general multi-modal data learning method for person re-identification. 2021. arXiv preprint arXiv:2101.08533.
7. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. Generative adversarial nets. Advances in neural information processing systems, 2014. 27.
8. Guo J, Yuan Y, Huang L, Zhang C, Yao JG, Kai Han K. Beyond human parts: Dual part-aligned representations for person re-identification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision,2019. pages 3642–3651.
9. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016. pages 770–778.

10. He L, Liu W. Guided saliency feature learning for person re-identification in crowded scenes. In European Conference on Computer Vision,2020. pages 357–373. Springer.
11. Hermans A, Beyer L, Leibe B. In defense of the triplet loss for person re-identification. 2017. arXiv preprint arXiv:1703.07737.
12. Hai-Miao Hu, Fang Wen, Zeng Guodong, Zihao Hu, Li Bo. A person re-identification algorithm based on pyramid color topology feature. Multimed Tools Appl. 2017;76(24):26633–46.
13. Huang Y, Zha ZJ, Fu X, Hong R, Li L. Real-world person re-identification via degradation invariance learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,2020. p. 14084–14094.
14. Li H, Wu G, Zheng WS. Combined depth space based architecture search for person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021. p. 6729–6738.
15. Li W, Zhao R, Xiao T, Wang X. Deepreid: Deep filter pairing neural network for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, 2014. p. 152–159.
16. Li W, Zhu X, Gong S. Person re-identification by deep joint learning of multi-loss classification.2017. arXiv preprint arXiv:1705.04724.
17. Loy CC, Xiang T, Gong S. Multi-camera activity correlation analysis. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009. pages 1988–1995. IEEE.
18. Luo Hao, Jiang Wei, Youzhi Gu, Liu Fuxu, Liao Xingyu, Lai Shenqi, Jianyang Gu. A strong baseline and batch normalization neck for deep person re-identification. IEEE Trans Multimed. 2019;22(10):2597–609.
19. Miao J, Wu Y, Liu P, Ding Y, Yang Y. Pose-guided feature alignment for occluded person re-identification. In: Proceedings of the IEEE/CVF international conference on computer vision,2019. p. 542–551.
20. Ming Z, Zhu M, Wang X, Zhu J, Cheng J, Gao Chengrui, Yang Yong, Wei Xiaoyong. Deep learning-based person re-identification methods: a survey and outlook of recent works. Image Vis Comput. 2022;119: 104394.
21. Mishchuk A, Mishkin D, Radenovic F, Matas J. Working hard to know your neighbor's margins: Local descriptor learning loss. 2017. arXiv preprint arXiv:1705.10872.
22. Ristani E, Solera F, Zou R, Cucchiara R, Tomasi C. Performance measures and a data set for multi-target, multi-camera tracking. In: European conference on computer vision,2016. p. 17–35. Springer.
23. Ristani E, Tomasi C. Features for multi-target multi-camera tracking and re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018. p. 6036–6046.
24. Sharma Charu, Kapil Siddhant R, Chapman David. Person re-identification with a locally aware transformer. arXiv Preprint. 2021. https://doi.org/10.48550/arXiv.2106.03720.
25. Shi H, Yang Y, Zhu X, Liao S, Lei Z, Zheng W, Li SZ. Embedding deep metric for person re-identification: a study against large variations. In: European conference on computer vision, 2016. pages 732–748. Springer.
26. Song C, Huang Y, Ouyang W, Wang L. Mask-guided contrastive attention model for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018. p. 1179–1188.
27. Sun Y, Zheng L, Yang Y, Tian Q, Wang S. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In: Proceedings of the European conference on computer vision (ECCV),2018. p. 480–496.
28. Wang Guangcong, Lai Jianhuang, Huang Peigen, Xie Xiaohua. Spatial-temporal person re-identification. Proc AAAI Conf Artif Intell. 2019;33:8933–40.
29. Wang G, Yuan Y, Chen X, Li J, Zhou L. Learning discriminative features with multiple granularities for person re-identification. In: Proceedings of the 26th ACM international conference on Multimedia,2018. p. 274–282.
30. Wang Xiaogang. Intelligent multi-camera video surveillance: a review. Pattern Recogn Lett. 2013;34(1):3–19.
31. Wang Y, Chen Z, Wu F, Wang G. Person re-identification with cascaded pairwise convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018.p. 1470–1478.
32. Wang Z, Ye M, Yang F, Bai X, Shin'ichi Satoh. Cascaded sr-gan for scale-adaptive low resolution person re-identification. In: IJCAI, 2018. 1, p. 4.
33. Wei L, Zhang S, Gao W, Tian Q. Person transfer gan to bridge domain gap for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition,2018. p. 79–88.
34. Wen Y, Zhang K, Li Z, Qiao Y. A discriminative feature learning approach for deep face recognition. In: European conference on computer vision,2016. p. 499–515. Springer.
35. Wieczorek M, Rychalska B, Dabrowski J. On the unreasonable effectiveness of centroids in image retrieval. arXiv. 2021. https://doi.org/10.48550/arXiv.2104.13643.
36. Yi D, Lei Z, Liao S, Li S. Deep metric learning for person re-identification. In: 2014 22nd International Conference on Pattern Recognition, 2014. p. 34–39. IEEE.
37. Yu S-I, Yi Yang, Hauptmann A. Harry potter's marauder's map: Localizing and tracking multiple persons-of-interest by nonnegative discretization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,2013. p. 3714–3720.
38. Ye Yuan, Chen W, Yang Y, Wang Z. In defense of the triplet loss again: Learning robust person re-identification with fast approximated triplet loss and label distillation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops,2020. p.354–355.
39. Zhao R, Ouyang W, Wang X. Person re-identification by salience matching. In: Proceedings of the IEEE international conference on computer vision, 2013. p. 2528–2535.
40. Zhao Y, Jin Z, Qi G-J, Lu H, Hua H-S. An adversarial approach to hard triplet generation. In: Proceedings of the European conference on computer vision (ECCV),2018. p. 501–517.
41. Zheng L, Shen L, Tian L, Wang S, Wang J, Qi Tian. Scalable person re-identification: A benchmark. In: Proceedings of the IEEE international conference on computer vision, 2015. p. 1116–1124.
42. Zheng L, Zhang H, Sun S, Chandraker M, Yi Yang, Qi Tian. Person re-identification in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,2017. p.1367–1376.
43. Zheng Z, Zheng L, Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In: Proceedings of the IEEE international conference on computer vision,2017. p. 3754–3762.

44. Zhihui Z, Xinyang J, Feng Z, Xiaowei G, Feiyue H, Weishi Z, Xing S. Viewpoint-aware loss with angular regularization for person re-identification. In Proceedings of the 33rd AAAI Conference on Artificial Intelligence (AAAI-19), Honolulu, HI, USA, 2019; 27.
45. Zhong Z, Zheng L, Cao D, Li S. Re-ranking person re-identification with k-reciprocal encoding. In: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017; p. 1318–1327.
46. Zhong Z, Zheng L, Zheng Z, Li S, Yi Yang. Camera style adaptation for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2018. p. 5157–5166.

Publisher's Note