A Technical Report for ICCV 2021 VIPriors Re-identification Challenge

Cen Liu, Yunbo Peng, Yue Lin
NetEase Games AI Lab
{liucen01, gzpengyunbo, gzlinyue}@corp.netease.com

Abstract

Person re-identification has always been a hot and challenging task. This paper introduces our solution for the re-identification track in VIPriors Challenge 2021. In this challenge, the difficulty is how to train the model from scratch without any pretrained weight. In our method, we show use state-of-the-art data processing strategies, model designs, and post-processing ensemble methods, it is possible to overcome the difficulty of data shortage and obtain competitive results. (1) Both image augmentation strategy and novel pre-processing method for occluded images can help the model learn more discriminative features. (2) Several strong backbones and multiple loss functions are used to learn more representative features. (3) Post-processing techniques including re-ranking, automatic query expansion, ensemble learning, etc., significantly improve the final performance. The final score of our team (ALONG) is 96.5154% mAP, ranking first in the leaderboard.

1. Introduction

2021 VIPriors Re-identification Challenge is one of "2nd Visual Inductive Priors for Data-Efficient Deep Learning Workshop" in the ICCV2021 conference. It focuses on obtaining high mean Average Precision (mAP) on a dataset coming from short sequences of basketball games. This dataset contains 8,569 person images of 436 identities in the training set while testing data is composed of the 9,171 images of 468 identities.

Person re-identification (ReID) aims at recognizing pedestrians across non-overlapping camera views, which draws wide attention due to its wide applications in surveillance, tracking, smart retail, etc. [16, 15, 12, 7]. As deep learning prevails, the CNN based ReID methods progress rapidly and achieve impressive performance on benchmark datasets. However, this challenge restricts the use of external data and pre-trained weights, due to the small dataset, it is difficult to perform well with general training due to the small dataset..

We solve this problem by experimenting with various techniques. In our method, based on strong backbones, we use multiple loss functions, data augmentation strategies, and ensemble learning to improve the performance.

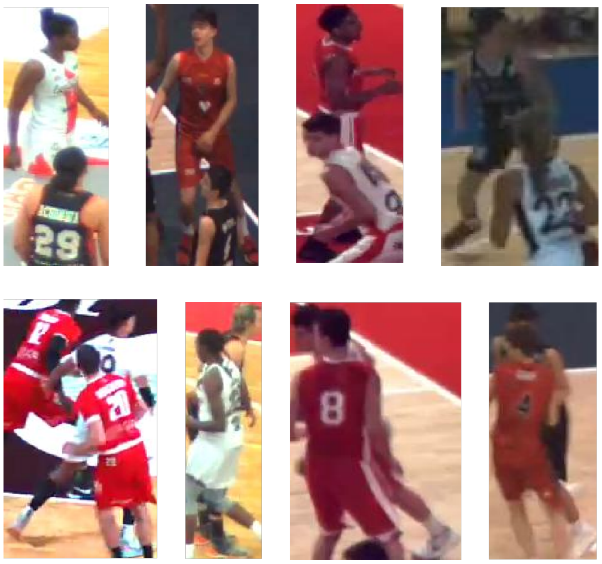

Figure 1. Illustration of same noise data samples with varying degrees of occlusion. The is partial occlusion is in the first row; the second row is completely occluded (the label of the images in the second row is the person occluded).

First, we focus on the data augmentation in data pre-processinging such as Random Erasing [18], Local Grayscale Transformation [2] and random flip horizontally, etc., to improve the diversity of person images. In addition, there are different degrees of occlusion in the raw data, which lead to the problem of noise in network training and result in hard samples. To address these challenges, we use hard sampling method to select difficult samples and clean noise samples. Then, we apply strong backbones, multiple loss functions and add some training tricks by this paper [8] to enable model training well with even a small amount of data. Finally, we apply automatic query expansion (AQE) [1] and re-rank [17], and post-processing ensemble techniques, etc. to compensate for the difficult problems through various models and multi-scale testing

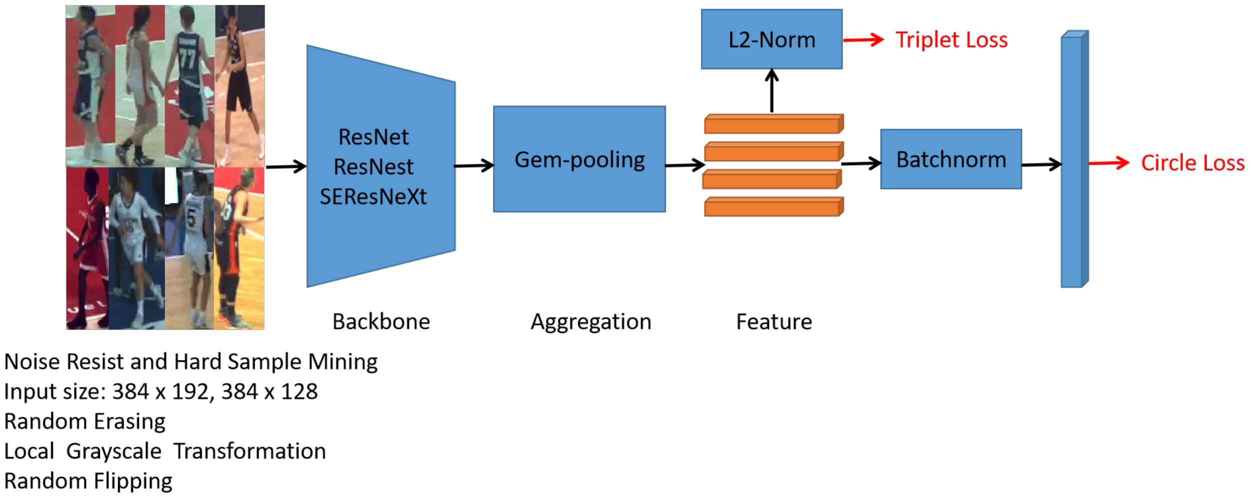

Figure 2. Illustration of our proposed framework.

to produce good performance. Through the model applies with these techniques, our approach can significantly improve the ReID performance and achieve a competitive result on the test set. According to the rules of the competition, we do not use any external image/video data or pretrained weights. The implementation details of our solution are described in section 2 and section 3.

2. Our Method

In this section, our method for this challenge is introduced in detail. The critical parts of our method include data pre-processing, model architecture, loss function and post-processing techniques, which are elaborated in detail as follows.

2.1. Pre-processing

Due to the limited quantity and quality of original data, effective data pre-processing technology can improve the recognition performance of the model.

Noise Resist and Hard Sample Mining: Most of the person re-identification datasets contain a small portion of noisy labels, which requires the learning algorithm to resist certain amount of noises. Through our observation of this dataset, we find that there is also a lot of labeling noise. Therefore, we choose the online difficult sample mining algorithm proposed by Shrivastava et al. [10] to mine the “hard sample (noise)” data according to the loss of the training set in the training stage. The noise data we filtered out is shown in Figure 1. We can see that this part of the data has different degrees of person occlusion.

Our initial approach is to delete this part of “noise” data, but we find that if all occlusion data are deleted through experiments, the results of the test set fall rather than rise. We find the main reason is that there is occlusion data in the test set, but the occlusion is not serious (not complete occlusion). As shown in figure 1, the partial occlusion is in the first row; the second row is completely occluded (the label of the images in the second row is occluded by person). Therefore, inspired by [10], we divide samples with different degrees of occlusion into partial occlusion and full occlusion, according to the training loss and the selected threshold. Our approach is to delete completely occluded samples as noise data. The data with slight occlusion is not deleted, but as the “hard samples” in training stage, and several data augmentations are used for the “hard samples”, increasing the proportion of hard samples in the original training set. In this way, we make the network focus on the optimization of this occlusion sample. The results show that our method of hard sample mining is effective to improve the performance of our network.

Data Augmentation: Data augmentation can effectively prevent overfitting. To solve the limitation of training data, we adopt some data augmentation strategies, such as Random Erasing [18], Local Grayscale Transformation [2].

Random Erasing [18]: In person ReID, persons in the images are sometimes occluded by other objects. To address the occlusion problem and improve the generalization ability of ReID models, Zhong et al. [18] proposed a new data augmentation approach named as Random Erasing Augmentation.

Local Grayscale Transformation(LGT) [2]: Through the observation of the original dataset, we find that most of the personnel are players of the same team, so their dresses and appearances are very similar. Howerver, they have the different spatial structures. To address this problem, this method can be used as an effective data augmentation by introducing grayscale information, which proposed by Gong [2]. With this strategy, our model can achieve significant improvements in ReID task.

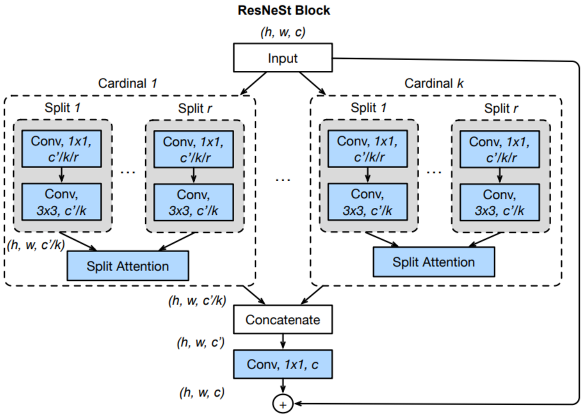

Figure 3. Split-Attention Block of ResNeSt.

Besides the above two specific data augmentation methods, some regular augment methods are applied, such as random affine transformation, pixel padding, random flipping, etc.

Imbalance Identities Resist: Even if the data provider claims that in the training set, sequence is composed by 20 frames in the training set. But through statistics, we find that there are less than 20 images in IDs such as ID31 and ID164. We use the balanced ID data augmentation on the training set, copy few images up to 5 times, and several data augmentations are used for the copied images.

2.2. Model Architecture

Baseline model is important for the final ranking, we use a CNN-based baseline [8] shown in Figure 2. And besides ResNet [4], we also use ResNeSt [14] and SE-ResNeXt [13] as our backbones. These models are introduced as follows.

ResNeSt: The key part of ResNeSt is Split-Attention block. Split-Attention block is a computational unit consisting of feature-map group and split attention operations. Figure 3 depicts an overview of the Split-Attention Block.

SE-ResNeXt: SE-ResNeXt is constructed by repeating a building block that aggregates a set of transformations with the same topology. SE block [6] adaptively recalibrates channel-wise feature responses by explicitly modelling interdependencies between channels. SE-ResNeXt has good performance in multiple tasks and is widely used.

Other Techniques: Various techniques are also applied. First, the aggregation layer aims to aggregate feature maps generated by the backbone into a global feature, and we use the Generalized-Mean (GeM) pooling [9] instead of GAP proposed in [8]. Second, dropout is used for preventing overfitting. Finally, we remove the last spatial downsampling operation in the backbone network to increase the size of the feature map. For convenience, we denote the last spatial down-sampling operation in the backbone network as last stride.

2.3. Loss function

Categorical cross-entropy(CE) loss after softmax and triplet loss [5] are widely used in person ReID task. But for VIPriors ReID dataset, we find that circle loss [11] performs better than CE loss. The total loss is proposed as follows for better performance.

$$L = \alpha L_{Triplet} + \beta L_{circle} \tag{1}$$

Triplet Loss: The triplet loss is computed as:

$$L_{Triplet} = [d_p - d_n + \alpha]_+ \tag{2}$$

where d_p and d_n are feature distances of positive pair and negative pair. α is the margin of triplet loss, and $[z]_+$ equals to $max(z, 0)$. In this paper, α is set to 0.4. However, triplet loss only considers the difference between d_p and d_n and ignores the absolute values of them.

Circle Loss: The circle loss benefits deep feature learning with high flexibility in optimization and more definite convergence target. It has a unified formula for two elemental learning approaches, i.e., learning with class-level labels and learning with pair-wise labels, more suitable for person ReID task. The derivation process of circle loss is not described here in detail, which can refer to [11].

It is worth mentioning that in the Circle Loss, the relaxation factor m is a very important hyper-parameter. We find that different m may have different results in training, but this difference is very conducive to the ensemble of models. This point will be explained in detail in section 3.3.

2.4. Post-processing

Post-processing can significantly improve ReID performance in the inference stage. In this section, we will introduce several post-processing methods in this paper.

Augmentation Test: For each test image, we flip these two images horizontally and additionally extract two features. We get two features totally and then average them to obtain the final ReID feature.

Re-Ranking: We adopt a widely-used Re-ranking (RK) method [17] to update the final result. We notice that Jaccard distance is more suitable than Euclidean distance due to many people dressing in the same team uniform. We set $\lambda = 0.1$ in this paper.

Query expansion: We also use QE [1] to improve the performance of the retrieval system. But compared to reranking, the improvement of this technique is not significant.

3. Experiments

3.1. Implementation Details

All of our experiments are conducted in 8 NVIDIA A100 GPUs. In the training stage, we train the model for 180 epochs with warming up [3] with initial learning rate of 0.0001 in first 10 epochs, increasing the learning rate from 0.0001 to 0.005 and dropout with probability of 0.2, weight decay of 0.0001 and label smooth are used for learning. In early stage, we trained model on training set for methods attempt and verification. And in the final stage, we trained models on both training set and validation set, and no external images or pre-trained weights are used. In addition, in order to obtain a better model, the model is trained several times by varying the random seed, and the results are combined together and treated as a whole, improving model mAP by 0.8%.

3.2. Results

We show the ablation study of above different strategies in Table 1. The baseline with ResNet50 trains on training set and evaluates on the validation set with the input size of 256×128. The baseline with ResNet50 backbone network can achieves 93.4708% on validation set. Besides, it can be find that the strategies we proposed improve the performance by almost 18% in terms of mAP, which shows the effectiveness of these techniques.

Table 1. The ablation study of different strategies on validation set. With all methods, ResNet50 achieves 93.4708% mAP.

Method	mAP(%)
Baseline(ResNet50)	76.2044
+ Noise Resist	78.0190
+ Hard Sample Mining	80.9477
+ Data Augmentation(Random Erasing and LGT)	81.4552
+ Gem Pooling	83.8275
+ Circle Loss	84.9631
+ Augmentation Test	85.2631
+ Re-rank	93.0423
+ AQE	93.1172
+ Large Input Size(384×128)	93.4708

3.3. Ensembling

Experimental evidence shows that the ensemble method is usually much more accurate than a single model. In our method, the ensemble method is the addition of the distance matrix of all prediction. For a better performance, we have ensembled predictions of above methods in total 24 models including ResNet101, ResNet152, ResNet200, ResNeSt-101, ResNeSt-152, ResNeSt-200, SE-ResNeXt101, SE-ResNeXt152, SE-ResNeXt200, with different input sizes and different relaxation factor m of circle loss.

Tips: We trained all models on both training set and validation set, for each backbone, we use two different sizes (384×128 and 384×192) as input. Besides, different relaxation factor m of circle loss will get different results, and we find that fusing the results has a certain improvement. Specifically, we set $m = 0.3$ and $m = 0.4$. Therefore, we fuse all the results as the final submission. The 96.5154% in mAP is the final ensemble results, ranking first in the learderboard.

4. Conclusion

In our method, three strong network architectures were taken as the backbones. The usage of multiple data preprocessing and post-processing strategies improves the performance of the models. Besides, multiple testing methods and ensemble strategies improve the generalization and robustness of the models and prevent overfitting. Finally, we win the 1st place in VIPriors re-identification competition.

References

[1] O. Chum, J. Philbin, J. Sivic, M. Isard, and A. Zisserman. Total recall: Automatic query expansion with a generative feature model for object retrieval. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8. IEEE, 2007.

[2] Y. Gong. A general multi-modal data learning method for person re-identification. 2021.

[3] P. Goyal, P. Dollár, R. Girshick, P. Noordhuis, L. Wesolowski, A. Kyrola, A. Tulloch, Y. Jia, and K. He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017.

[4] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[5] A. Hermans, L. Beyer, and B. Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017.

[6] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[7] C. Liu, L.-J. Guo, and R. Zhang. Hlfnet: High-low frequency network for person re-identification. *IEEE Signal Processing Letters*, 2021.

[8] 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*, pages 0–0, 2019.

[9] F. Radenović, G. Tolias, and O. Chum. Fine-tuning cnn image retrieval with no human annotation. *IEEE transactions on pattern analysis and machine intelligence*, 41(7):1655–1668, 2018.

[10] A. Shrivastava, A. Gupta, and R. Girshick. Training region-based object detectors with online hard example mining. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 761–769, 2016.

[11] Y. Sun, C. Cheng, Y. Zhang, C. Zhang, L. Zheng, Z. Wang, and Y. Wei. Circle loss: A unified perspective of pair similarity optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6398–6407, 2020.

[12] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *Proceedings of the European conference on computer vision (ECCV)*, pages 480–496, 2018.

[13] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017.

[14] H. Zhang, C. Wu, Z. Zhang, Y. Zhu, H. Lin, Z. Zhang, Y. Sun, T. He, J. Mueller, R. Manmatha, et al. Resnest: Split-attention networks. *arXiv preprint arXiv:2004.08955*, 2020.

[15] L. Zhao, X. Li, Y. Zhuang, and J. Wang. Deeply-learned part-aligned representations for person re-identification. In *Proceedings of the IEEE international conference on computer vision*, pages 3219–3228, 2017.

[16] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision*, pages 1116–1124, 2015.

[17] Z. Zhong, L. Zheng, D. Cao, and S. Li. Re-ranking person re-identification with k-reciprocal encoding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1318–1327, 2017.

[18] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang. Random erasing data augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13001–13008, 2020.