Stronger Baseline for Person Re-Identification

Fengliang Qi, Bo Yan, Leilei Cao, Hongbin Wang

Ant Group

Abstract

Person re-identification (re-ID) aims to identify the same person of interest across non-overlapping capturing cameras, which plays an important role in visual surveillance applications and computer vision research areas. Fitting a robust appearance-based representation extractor with limited collected training data is crucial for person re-ID due to the high expanse of annotating the identity of unlabeled data. In this work, we propose a **Stronger Baseline** for person re-ID, an enhancement version of the current prevailing method, namely, Strong Baseline, with tiny modifications but a faster convergence rate and higher recognition performance. With the aid the **Stronger Baseline**, we obtained the *third* place (i.e., 0.94 in mAP) in *2021 VIPriors Re-identification Challenge* without the auxiliary of ImageNet-based pre-train parameter initialization and any extra supplemental dataset.

Introduction

Person re-identification (re-ID) has become a core, and widely used technique in visual surveillance applications and computer vision research areas (Sun et al. 2019; Yan et al. 2019; Liu et al. 2019). It aims at locating and recognizing a person of interest across multiple non-overlapping cameras in various spots (Zhang et al. 2019; Subramaniam, Nambiar, and Mittal 2019; Fu et al. 2019).

The main challenge for the person-reID task is that compared with the conventional benchmark for classification task (e.g., ImageNet), the size of the available training dataset can be limited, making the risk of over-fitting increase, hence the deterioration of the final generalization performance. To alleviate this dilemma, most of the current work focus on two main technical routines: 1) Designing customized and lightweight neutral network structure (e.g., OSNet); 2) Introducing discriminative margin-based loss function (e.g., Triplet loss) widely used in metric learning and related research field (e.g., face verification).

In this work, we follow the second technical routine and propose the **Stronger Baseline** by modifying the Strong baseline proposed by (Luo et al. 2019) with tiny overhead but faster convergence rate and higher recognition performance as shown in Fig1. In contrast to the argument claimed in (Luo et al. 2019), we argue that the **BNNeck** is not the critical factor in alleviating the conflict between the triplet
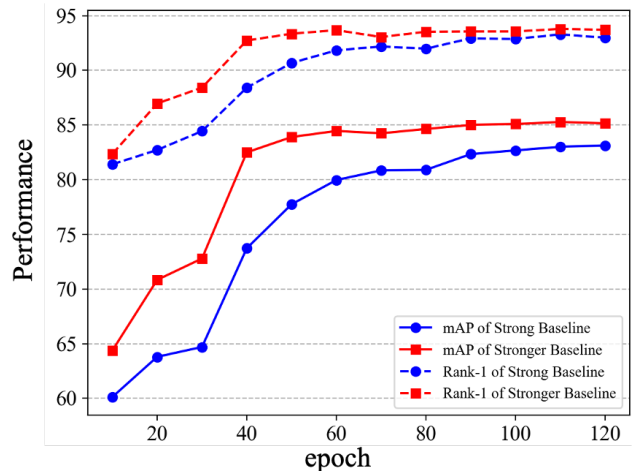

Figure 1: **Convergence rate and performance are both enhanced considerably on Market1501 via the proposed Stronger Baseline.**

loss and cross-entropy loss minimization process. Instead, we claim the BNNeck is simply a standardization procedure. When combined with a softmax classifier, whose output follows the multinomial distribution, one of the exponential family distribution, the loss landscape can be smoother. The gradient update direction can be more stable, making the optimization algorithm more possible to arrive at the optimal global solution.

Furthermore, we attribute the occurrence of optimization conflict between cross-entropy loss and triplet loss to two main aspects: 1) inconsistency in identifying the hard samplers during the hard mining process and 2) inconsistency in gradient update direction during the objective minimizing process in two metric space (i.e., Cosine Metric Space and Euclidean Metric Space). As illustrated in Fig.2(a), compared to the anchor sample f_a, the positive sample f_p is easy in Cosine Metric Space for it lies in the similar radial direction with anchor sample f_a; however, when it comes to the Euclidean Metric Space, the positive sample f_p should role as a hard sample for its significant Euclidean distance from anchor sample. The inconsistency in identifying the hard-

ness of the positive sample still holds for the negative sample (e.g., f_n in Fig.2(a)) in a similar principle. To alleviate this conflict, we use Batch-Normalization (BN) module to standardize the distribution of the samplers as shown in Fig.2(b) and make the identification of samples' hardness consistent in both Cosine and Euclidean metric space.

As for the second inconsistency in gradient update direction, the gradient for the positive sample f_p in Euclidean Metric Space is parallel to the link between f_a and f_p. However, in Cosine Metric Space, the gradient update direction is parallel to the tangent direction, confusing the final comprehensive update direction and deteriorating the optimization process. To mitigate this confusion, we use the L2 Normalization operation for the feature before triplet loss estimation, that is to say, triplet loss optimizing the distance in the same cosine metric space with cross-entropy loss as illustrated in Fig.2(c).

Based on the analysis for alleviating the conflict in optimizing the cross-entropy loss and triplet loss, we propose our **Stronger Baseline** with tiny modification on the Strong Baseline proposed in (Luo et al. 2019). As illustrated in Fig.3(b), the feature f_t extracted from the backbone network (e.g., ResNet) is processed by subsequent Batch-Normalization module to generate the feature f_i for cross-entropy loss calculation in the training stage and similarity evaluation in inference stage, which is the same as the pipeline of Strong baseline in Fig.3(a). Different from Strong Baseline, the **Stronger Baseline** calculate the triplet loss on the L2 normalization version of feature f_i instead of feature f_t. We discard the center loss for its limited influence on the final performance(Luo et al. 2019).

In a nutshell, the main contribution of this work can be summarized in the following folds:

1. We conclude two kinds of inconsistency while simultaneously optimizing the Cosine and Euclidean metric space, inducing the conflict in minimizing the cross-entropy and triplet loss.

2. We propose the **Stronger Baseline** based on analysis of how to alleviate the inconsistency during optimization by ameliorating the Strong Baseline with limited overhead.

3. We verify the superiority of our proposed method on the Market1501 and SynergySports benchmark.

Related Works

Person Re-identification

Gheissari et al. (Gheissari, Sebastian, and Hartley 2006) first defined the person re-ID as a specific computer vision task. Most of the traditional methods focused on feature extraction via handcraft design before the emergence of deep learning(Kviatkovsky, Adam, and Rivlin 2012; Farenzena et al. 2010; Gray and Tao 2008; Matsukawa et al. 2016; Cheng et al. 2011). With the rise of deep learning in recent years, convolutional neural networks (CNN) based feature representation methods have become the mainstream for image-based person re-ID (Li, Zhu, and Gong 2018; Yan et al. 2019; Sun et al. 2018; Xiao et al. 2016). According to Zheng et al. (Zheng, Yang, and Hauptmann 2016), most of the pre-existing image-based works concentrated on

discriminative learning and metric learning. Discriminative learning (Li, Zhu, and Gong 2018; Chang, Hospedales, and Xiang 2018; Guo and Cheung 2018; Liu et al. 2017) aims at getting representative features for identity classification. While metric learning (Sun et al. 2017; Zheng et al. 2017) learns to project extracted features from different cameras and views into a common feature representation subspace.

Distance Metric Learning

The goals of the training and testing are slightly different in person re-ID tasks. The training process mainly focuses on classification or metric learning, while the testing process is a retrieval problem. Most existing supervised person re-ID methods apply identification loss for identity classification (e.g., cross-entropy loss)(Sun et al. 2017) and verification loss for metric learning (e.g., triplet loss)(Hermans, Beyer, and Leibe 2017). Triplet loss (Schroff, Kalenichenko, and Philbin 2015) is designed initially for face recognition problem and has been regarded as a commonly used method in the retrieval related tasks, especially in person re-ID. In the triplets, an anchor, a positive sample, and a negative sample are included. Since triplet loss is calculated by two randomly sampled person identification, it is difficult to ensure that the distance between the anchor and positive samples are smaller than the distance between the anchor and negative examples in the whole training dataset(Luo et al. 2019). Quadruplet loss(Chen et al. 2017) is an improved version of triplet loss, which contains two different negative samples to learn a larger inter-class distance and a smaller intra-class distance compared with the triplet loss(Jiang et al. 2020).

Our Method

Holistic Pipeline

In our experiments, the Stronger Baseline is applied as our holistic pipeline to realize the representation extracting and loss objective calculation. We use ResNet50 as a backbone for controlled experiments and *OSNet* **without ImageNet pretrain** for final result submission. The influence of the Backbone is empirically limited to the yielded performance, and the gain of the performance mostly owes to the proposed **Stronger Baseline**. Both the BNNeck and FC layers are initialized through Kaiming initialization proposed in(He et al. 2015). Each image in a training batch is prepossessed and resized to 256×128 in pixels. In each batch, we sample 8 identities, each with 4 images.

Data Augmentation

We utilize the following data augmentation for enhancing the generalization performance: *Random Horizontal Flip* with probability 0.5, *Random Erasing* with probability 0.5, *Color Jitter* with probability 0.5, and *AutoAugmentation*(Cubuk et al. 2018). We are inconclusive in which augmentation strategy is beneficial to the final performance at most. We solely adopted the first two augmentation strategies while conducting the controlled experiment.
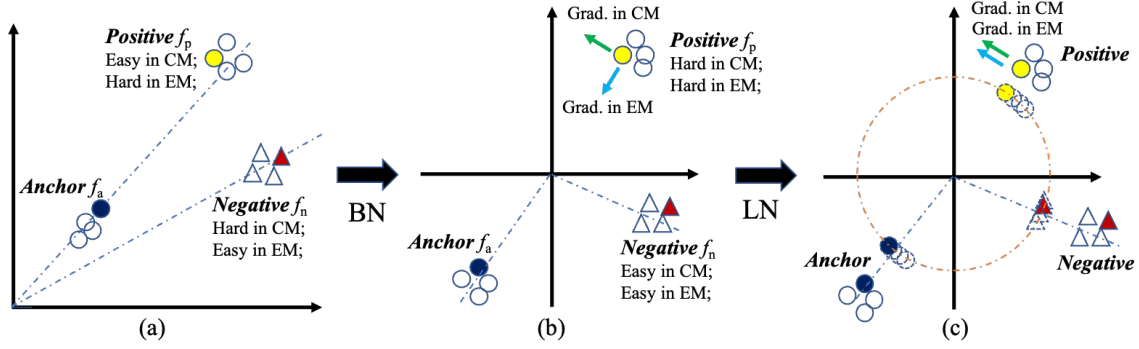
Figure 2: **Schematic illustration of the optimization conflict in Cosine and Euclidean metric space.** (a) The hardness of samples is identified inconsistently in Cosine and Euclidean metric space. (b) BN module can alleviate the first inconsistency in identifying the hardness of samples. (c) LN operation can alleviate the second inconsistency in gradient updating direction. Shape refers to the identity of the samples, and filling color refers to the role in metric pairs (e.g., anchor), Grad., BN, LN, CM, and EM refer to the Gradient, Batch-Normalization, L2 Normalization, Cosine Metric, and Euclidean Metric, respectively.
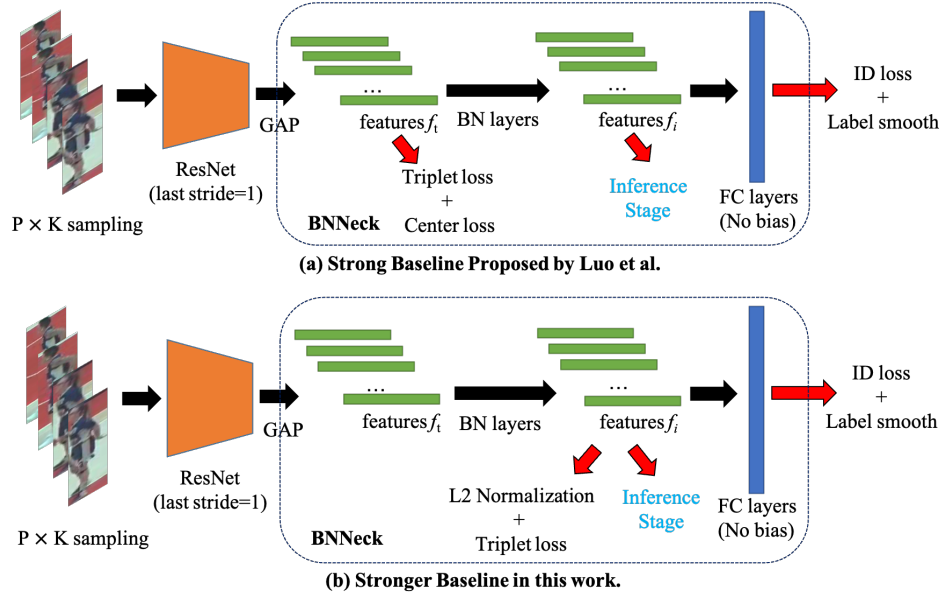

Figure 3: **The overview of the pipeline of Stronger Baseline**. (a) The primary pipeline of Strong Baseline. (b) The enhanced version of Strong Baseline, namely, Stronger Baseline.

Post-processing

Person re-ID can also be regarded as a retrieval problem. While presenting our final results, we utilize two common post-processing strategies, i.e., Query Expansion and ReRank, whose hyperparameters follow the default setting in FastReID.

Optimization Strategy

For controlled trials on Market1501, Adam(Kingma and Ba 2014) optimizer is selected with an initial learning rate of 3.5×10^{-4}. The commonly adopted warm-up strategy(Fan et al. 2019) is applied to bootstrap the network for better performance. In practice, the network is optimized for 120 epochs. We spend 10 epochs linearly increasing the learning rate from 3.5×10^{-6} to 3.5×10^{-4}, and it then decays at the 30th and 55th epoch. We discard the warm-up stage for final submission and use SGD optimizer instead due to the exclusion of pretrain. The network is optimized for 350 epochs, and the init learning rate of 0.065 decays at the 150th, 225th and 300th epoch by 0.1.

Experiments

We conduct two main experiments in this work, i.e., Controlled Experiment on the public benchmark (i.e., Market1501) and Submitted Experiment on SynergySports. The former aims to justify the superiority of the proposed Stronger Baseline, and the latter aims to obtain a better result with extra tricks (e.g., Complex Data Augmentation and Post-Processing Strategy).

Datasets and Evaluation Protocol

Market-1501 dataset is featured by 1,501 IDs, 19,732 gallery images and 12,936 training images captured by 6 cameras. Market-1501 are produced by the DPM detector. The Cumulative Matching Characteristics (CMC) curve is used for performance evaluation, which encodes the possibility that the query person is found within the top n ranks in the rank list. We also employ the mean Average Precision (mAP), which considers the retrieval process's precision and recall. The evaluation toolbox provided by the Market-1501 authors is used.

SynergySports generates from short sequences of basketball games, and each sequence is composed of 20 frames. For the validation and test sets, the query images are persons taken at the first frame, while the gallery images are identities taken from the 2nd to the last frame. There are 436 IDs, 8569 images in training split, and 50 IDs, 960 images in validation split.

Empirical Results

As illustrated in Fig1, the Stronger Baseline can enhance the convergence rate and performance simultaneously. The detailed results are listed in Table1. The mAP of Strong Baseline is improved by 2.03, which is a considerable enhancement considering the tiny modification and no extra introduced overhead. We further evaluate our method on the validation set of SynergySports; the results are listed in Table2, the mAP achieves the predictable improvement by 1.89.

As for the Submitted Experiment, we switch to the OSNet, a lightweight backone customized for person re-ID, and introduce the complex data augmentation and postprocessing strategy for better generalization performance. The results on the validation and test set are listed in Table3.

Protocol	Strong Baseline	Stronger Baseline
mAP	83.13	85.16 (+2.03)
Rank-1	92.99	93.71 (+0.72)

Table 1: The mAP and rank-1 comparison result of Strong Baseline and Stronger Baseline on Market1501. We use ResNet50 as backbone and ignore the complex data augmentation (i.e., Color Jitter and AutoAug) and Post-Processing here.

Protocol	Strong Baseline	Stronger Baseline
mAP	92.74	94.63 (+1.89)

Table 2: The mAP comparison result of Strong Baseline and Stronger Baseline on SynergySports (Validation Set). We use ResNet50 as backbone and ignore the complex data augmentation (i.e., Color Jitter and AutoAug) and Post-Processing here.

Protocol	Validation Set	Test Set
mAP	95.17	94.19

Table 3: The mAP result on SynergySports. We use OSNet1x0 as backbone and introduce the complex data augmentation (i.e., Color Jitter and AutoAug) and Post-Processing here.

Conclusions

In this paper, we concluded two kinds of inconsistency while optimizing the Cosine and Euclidean metric space simultaneously, i.e., hardness identification inconsistency and gradient update inconsistency, causing the conflict between minimizing the cross-entropy loss and triplet loss. To alleviate the inconsistency and ameliorate the optimization process, we proposed the Stronger Baseline with tiny modifications on the Strong Baseline but a faster convergence rate and higher evaluation performance. With the aid of Stronger Baseline, we obtain a third place in the 2021 VIPriors Re-identification Challenge without the auxiliary of ImageNet-based pre-train parameter initialization and any extra supplemental dataset.

References

Chang, X.; Hospedales, T. M.; and Xiang, T. 2018. Multi-level factorisation net for person re-identification. In *CVPR*, 2109–2118.

Chen, W.; Chen, X.; Zhang, J.; and Huang, K. 2017. Beyond Triplet Loss: A Deep Quadruplet Network for Person Re-Identification. In *CVPR*.

Cheng, D. S.; Cristani, M.; Stoppa, M.; Bazzani, L.; and Murino, V. 2011. Custom pictorial structures for re-identification. In *BMVC*, volume 1, 6.

Cubuk, E. D.; Zoph, B.; Mané, D.; Vasudevan, V.; and Le, Q. V. 2018. AutoAugment: Learning Augmentation Policies from Data. *CoRR* abs/1805.09501. URL http://arxiv.org/abs/1805.09501.

Fan, X.; Jiang, W.; Luo, H.; and Fei, M. 2019. Spher-ereid: Deep hypersphere manifold embedding for person re-identification. *Journal of Visual Communication and Image Representation* 60: 51–58.

Farenzena, M.; Bazzani, L.; Perina, A.; Murino, V.; and Cristani, M. 2010. Person re-identification by symmetry-driven accumulation of local features. In *CVPR*, 2360–2367.

Fu, Y.; Wang, X.; Wei, Y.; and Huang, T. 2019. STA: Spatial-temporal attention for large-scale video-based person re-identification. In *AAAI*, volume 33, 8287–8294.

Gheissari, N.; Sebastian, T. B.; and Hartley, R. 2006. Person reidentification using spatiotemporal appearance. In *CVPR*, 1528–1535.

Gray, D.; and Tao, H. 2008. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *ECCV*, 262–275.

Guo, Y.; and Cheung, N.-M. 2018. Efficient and deep person re-identification using multi-level similarity. In *CVPR*, 2335–2344.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *ICCV*, 1026–1034.

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

Jiang, M.; Leng, B.; Song, G.; and Meng, Z. 2020. Weighted triple-sequence loss for video-based person re-identification. *Neurocomputing* 381: 314–321.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv:1412.6980* .

Kviatkovsky, I.; Adam, A.; and Rivlin, E. 2012. Color invariants for person reidentification. *IEEE Transactions on pattern analysis and machine intelligence* 35(7): 1622–1634.

Li, W.; Zhu, X.; and Gong, S. 2018. Harmonious attention network for person re-identification. In *CVPR*, 2285–2294.

Liu, C.-T.; Wu, C.-W.; Wang, Y.-C. F.; and Chien, S.-Y. 2019. Spatially and Temporally Efficient Non-local Attention Network for Video-based Person Re-Identification. In *arXiv:1908.01683*.

Liu, H.; Feng, J.; Qi, M.; Jiang, J.; and Yan, S. 2017. End-to-end comparative attention networks for person re-identification. *IEEE Transactions on Image Processing* 26(7): 3492–3506.

Luo, H.; Gu, Y.; Liao, X.; Lai, S.; and Jiang, W. 2019. Bag of Tricks and a Strong Baseline for Deep Person Re-Identification. In *CVPRW*.

Matsukawa, T.; Okabe, T.; Suzuki, E.; and Sato, Y. 2016. Hierarchical gaussian descriptor for person re-identification. In *CVPR*, 1363–1372.

Schroff, F.; Kalenichenko, D.; and Philbin, J. 2015. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, 815–823.

Subramaniam, A.; Nambiar, A.; and Mittal, A. 2019. Co-segmentation Inspired Attention Networks for Video-based Person Re- identification. In *ICCV*, 562–572.

Sun, H.; Chen, Z.; Yan, S.; and Xu, L. 2019. MVP Matching: A Maximum-Value Perfect Matching for Mining Hard Samples, With Application to Person Re-Identification. In *ICCV*, 6737–6747.

Sun, Y.; Zheng, L.; Deng, W.; and Wang, S. 2017. Svdnet for pedestrian retrieval. In *ICCV*, 3800–3808.

Sun, Y.; Zheng, L.; Yang, Y.; Tian, Q.; and Wang, S. 2018. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *ECCV*, 480–496.

Xiao, T.; Li, H.; Ouyang, W.; and Wang, X. 2016. Learning deep feature representations with domain guided dropout for person re-identification. In *CVPR*, 1249–1258.

Yan, S.; Xu, J.; Liu, Y.; and Xu, L. 2019. Hornet: a hierarchical offshoot recurrent network for improving person re-ID via image captioning. *arXiv:1908.04915* .

Zhang, R.; Li, J.; Sun, H.; Ge, Y.; Luo, P.; Wang, X.; and Lin, L. 2019. Scan: Self-and-collaborative attention network for video person re-identification. *IEEE Transactions on Image Processing* 28(10): 4870–4882.

Zheng, L.; Yang, Y.; and Hauptmann, A. G. 2016. Person re-identification: Past, present and future. *arXiv:1610.02984* .

Zheng, L.; Zhang, H.; Sun, S.; Chandraker, M.; Yang, Y.; and Tian, Q. 2017. Person re-identification in the wild. In *CVPR*, 1367–1376.