2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI) | 978-1-6654-0898-1/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICTAI52525.2021.00171

# Object Quality Guided Feature Fusion for Person Re-identification

Lei Zhang[1], Na Jiang[2], Qishuai Diao[1], Danyang Huang[1], Zhong Zhou[1*] and Wei Wu[1]
[1]*State Key Lab of Virtual Reality Technology and Systems, Beihang University, China*
[2]*Information Engineering College, Capital Normal University, China*
{zhangleilei0125, diaoqishuai, hdychi, zz, wuwei}@buaa.edu.cn, jiangna@cnu.edu.cn

***Abstract*—Person re-identification (Re-ID) is an essential task in computer vision, which aims to match a person of interest across multiple non-overlapping camera views. It is a fundamental challenging task because of the conflicts between large variations of samples and the limited scale of training sets. Data augmentation method based on generative adversarial network (GAN) is an efficient way to relieve this dilemma. However, existing methods do not consider how to keep identity information and filter the noise of the generated auxiliary samples during Re-ID training. In this paper, we propose object quality guided feature fusion network for person re-identification, which consists of a self-supervised object quality estimation module and a feature fusion module. Specifically, the former evaluates the quality of the auxiliary data to filter the noise and the disturbing features, while the later accomplishes the feature fusion based on object quality estimation in the collection-to-collection recognition manner to make full use of auxiliary data. Extensive performance analysis and experiments are conducted on two benchmark datasets (Market-1501 and DukeMTMC-reID) to show that our proposed approach outperforms or shows comparable results to the existing best performed methods.**

***Index Terms*—person re-identification, feature fusion, quality estimation, data augmentation**

## I. Introduction

Person re-identification (Re-ID) is an essential and demanding task in computer vision. Given a query person of interest, the fundamental Re-ID problem is to identify whether this person has appeared in another place at a distinct time across camera views [1], [2].

Person Re-ID is a fundamental challenging task due to conflicts between large variations of samples and the limited scale of training sets. And data augmentation [3]–[5] based on generative adversarial network (GAN) is an efficient way to relieve this dilemma. It considers the demand for large data volume and diversity in deep learning-based person Re-ID and generates auxiliary samples (pose transfer or style transfer) for training with GAN. However, GANs aim to generate images that look realistic. As shown in Fig. 1, it is easy to mix various appearance features and disturbing noise into the generated samples, which is destructive to the performance of person Re-ID.

Existing methods deal with this problem in two ways [6]–[8] of retaining identity information as much as possible during generative adversarial training and optimizing the model with label smooth regularization during Re-ID training. These two ways do not consider how to keep identity information and filter the noise of the generated samples during Re-ID training. And we argue that existing person re-identification methods based on data augmentation do not make full use of the auxiliary data.

(a) (b)

Fig. 1. Examples of the real images and the generated images (pose transfer images). (a) are the real images. (b) denote the generated images which contain noise and disturbing features.

In this paper, we consider that the key to make full use of the generated samples (pose transferred) lies in two ways: (1) fusing the features of the real images and the generated auxiliary data; (2) filtering the noise and the disturbing features, which would contain some information from pedestrians in different identities. In consequence, we propose object quality estimation guided feature fusion network for person re-identification. It consists of a self-supervised object quality estimation module and a feature fusion module. The former proposes to filter the noise and the disturbing features in the pose transferred images based on object quality estimation. And the later aims to fuse the features of the real images and the pose transferred images based on object quality estimation in collection-to-collection recognition manner.

Specifically, we propose a self-supervised object quality estimation module to evaluate the quality for filtering the noise and the disturbing features in the generated samples. It can assign different weights to the features of the auxiliary data. In detail, the dimension of high-quality features will be assigned

* Corresponding Author.

978-1-6654-0898-1/21/$31.00 ©2021 IEEE
DOI 10.1109/ICTAI52525.2021.00171

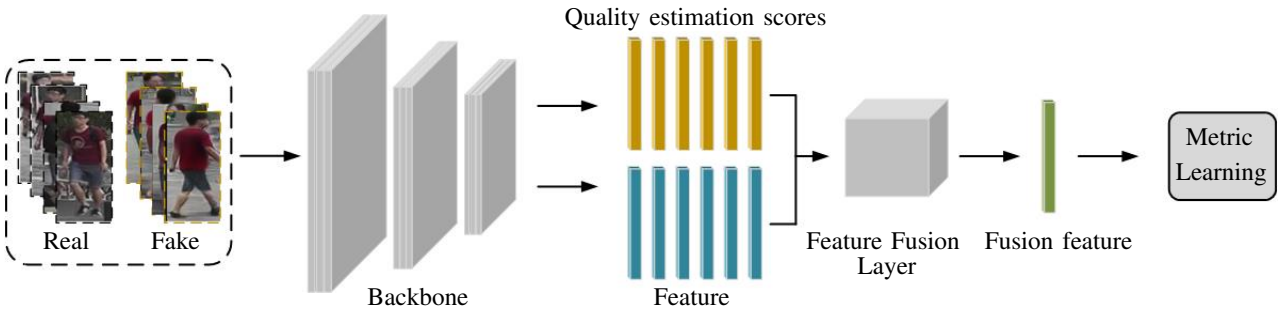



Fig. 2. Outline of our proposed feature fusion network based on object quality estimation. We consider to extract features of the same person in different poses for fusion and introduce a self-supervised object quality estimation to assign different weights to the features of the auxiliary data, which can filter the noise and the disturbing features.

higher weights than those low-quality features. In addition, we propose a feature fusion network based on object quality estimation in collection-to-collection recognition manner to make full use of the auxiliary data. It aims at fusing the features of images in different poses to alleviate the problem of misalignment. And we introduce a novel data sampling strategy to accomplish the collection-to-collection recognition. In summary, our contributions can be summarized into three aspects as follows:

(1) we propose a self-supervised object quality estimation module to assign different weights to the features of the auxiliary data. The module can filter the noise and the disturbing features for driving feature fusion;

(2) we propose a feature fusion network based on object quality estimation in collection-to-collection recognition manner to fuse the features of the real images and the generated auxiliary data (pose transferred images);

(3) Extensive performance analysis and experiments are conducted on two benchmark datasets to show that our proposed approach outperforms or shows comparable results to the existing best performed methods.

## II. METHODOLOGY

In this section, we propose our object quality guided feature fusion for person re-identification. We first introduce our self-supervised object quality estimation module. And then we briefly describe the proposed feature fusion network based on object quality estimation in collection-to-collection recognition manner.

### *A. Self-Supervised Object Quality Estimation*

In this paper, our approach aims to make full use of the generated auxiliary in the manner of feature fusion. The quality of the pose transferred images is uneven, there are lots of pool-quality samples. The noise in the low-quality images is the key factor that prevents us from improving the feature discrimination. The feature discrimination ability after fusion is insufficient. Therefore, we propose a self-supervised object quality estimation module to describe the importance of features for filtering the noise.

During training, existing methods usually utilize the classical Euclidean distance and cosine distance to measure the likelihood that two images belong to the same person. Euclidean distance calculates the sum of the squares of the difference between the features of each dimension and calculates the square root of the arithmetic. Cosine distance computes the cosine of the angle between two feature vectors in a special space. However, they do not consider the quality of the features. In this paper, during feature fusion, we consider to obtain the quality score of each image. Therefore, during distance metric, we desire that the dimensions of high-quality features obtain higher weight. In addition, during person search, images which contain lots of noise may lead to false matches. Specifically, the distance person images in different identities which are blurry may be less. We consider to add penalties for poor quality images. Based on the above discussion, in this paper we propose a distance measure function driven by quality scores, which is expressed as:

$$D = \sum_{k=1}^{K} \left( \frac{\left(f_k^i - f_k^j\right)}{\left(q_k^i\right)^2 + \left(q_k^j\right)^2} + log\left( \left(q_k^i\right)^2 + \left(q_k^j\right)^2 \right) + c \right) \quad (1)$$

$K$ denotes the dimension of the feature, which is usually 1024 or 2048. $c$ is a constant value. $q_k^i$ and $q_k^j$ are quality scores. And we use the reciprocal of the quality scores as the weight. The higher the quality score in the dimension, the lower the $q_k^i(q_k^j)$. At this moment, the confidence is superior. In other words, this dimension has a higher weight. In addition, to alleviate the mismatching problem, we add a penalty term in our loss function. If the quality of two images to be measured are low, this term results in a large penalty value to increase the distance.

After feature fusion, we desire that the fused features have good discrimination. Therefore, based on the proposed sampling strategy above, we utilize the classical triplet loss to optimize the model. FaceNet propose the triplet loss which has been proven that it is critical for fast convergence. It engages in that a pedestrian image of a specific identity should be closer to all other person images from the same identity than it is to any person image from any other identity. Specifically, given an image $x_i^a$(anchor), choosing the hardest positive pedestrian image $x_i^p$ such that argmax $\|f(x_i^a) - f(x_i^p)\|_2^2$ and the hardest negative image $x_i^n$ such that argmin $\|f(x_i^a) - f(x_i^n)\|_2^2$ similarly, where $f(x)$ denotes the feature of image $x$. To avoid model collapse, FaceNet replaces

the hardest negative images with semi-hard samples such that $\|f(x_i^a) - f(x_i^p)\|_2^2 < \|f(x_i^a) - f(x_i^n)\|_2^2$ .

Triple loss function takes into account not only the absolute distance between positive (negative) samples and anchor, but also the relative distance between positive and negative samples. When training with it, the network can learn the commonalities between samples in the same category, as well as the differences between different identities. Therefore, during feature fusion, the network tends to retain the features in higher discrimination, while ignoring the features unrelated to classification. As the model converges, samples with higher quality tend to be given higher weights. In this way, a self-supervised object quality estimation is achieved without manual annotation and introducing other information into the training process.

### *B. Feature Fusion Network based on Collection-to-Collection Recognition*

Annotating large-scale datasets is effective but prohibitively expensive. Some works based on data augmentation have been proposed, which aim at generating auxiliary data in different poses to increase the diversity of images. The key factors that help us to improve the performance are the manner of the generated image usage. In this paper, we propose a feature fusion network based on collection-to-collection recognition to make full use of the auxiliary data generated (pose transferred). It aims at fusing the features of images in different poses to alleviate the problem of misalignment.

Pose variations commonly produce severe misalignment between pedestrian images. The results of distance cannot accurately express the similarity of the images in quite different poses. To deal with this issue, we propose a novel feature fusion network based on collection-to-collection recognition to use the features of the real images and the generated auxiliary data diversity.

Misalignment will produce an inaccurate similarity measurement. Therefore, we extract features of the same person in different poses for fusion to accomplish distance metric based on collection-to-collection recognition. Examples of two kinds of similarity measurement are shown in Fig. 3. The upper is the original similarity measurement, which extracts feature of each image separately to calculate the Euclidean distance or cosine distance. It cannot describe the similarity of pedestrians accurately due to pose variations. The bottom denotes the distance metric based on collection-to-collection recognition. The images in the blue border are real and the ones in the red border are the pose transferred images (auxiliary data). We extract the features of the real images and the auxiliary data in different poses respectively for fusing to improve the discrimination of the features and then calculate the similarity.

Based on the above discussion, we propose a novel feature fusion network to alleviate the misalignment problem. As shown in Fig. 1, the quality of the pose transferred images is uneven. Therefore, we design a feature fusion network, which aims to estimate the quality of each fake image and fuses the features of different images in the same identity. The overall framework is shown in Fig. 2. We consider that the pose transferred images contain noise. It makes the importance of each dimension feature different. To alleviate the problem, we propose a quality estimation module to describe the confidence for assigning different weights to the features of the auxiliary data to filter the noise and the disturbing features. The features after processing are then fused with the features of real images to improve the discrimination and alleviate the misalignment problem.

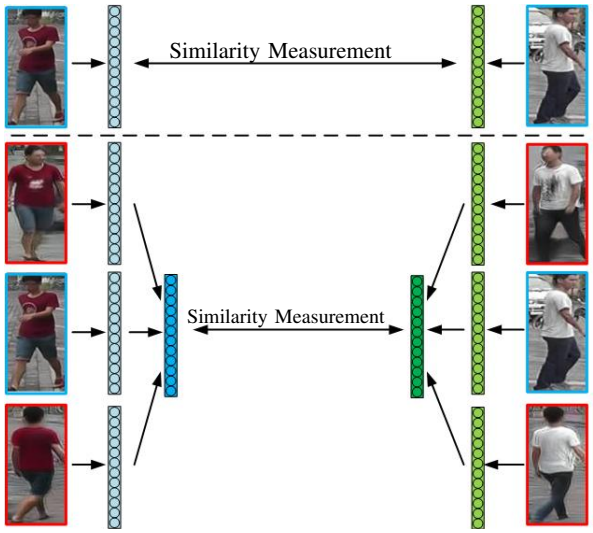



Fig. 3. Comparison of two kinds of similarity measurement. The upper is the original similarity measurement and the bottom denotes the distance metric based on collection-to-collection recognition.

In addition, to accomplish the collection-to-collection recognition in feature fusion network, we design a data sampling strategy correspondingly based on triplet loss. During training, each batch contains 64 images from 4 identities, where every identity represents a pedestrian without duplication.

### *C. Training Strategy based on Data Augmentation*

In this paper, we introduce a feature fusion network based on collection-to-collection recognition. We aim to fuse the features in each collection which are from the same identity. And images in different poses can provide the diversity of appearance information. Therefore, we select person from different poses in each collection to alleviate the problem of large disparity in appearance feature caused by pose change. We divide the dataset into three subsets of the front, back and side. For a pedestrian with lack of pose information, we use generate adversarial network to supplement it for data augmentation. In the beginning, the performance of the model is poor. Therefore, the images of each collection are all real in the early training stage. As the model converges, we increase the proportion of the auxiliary data gradually. In the beginning of training, we just utilize the real images. As the model converges, we increase the number of the pose transferred samples gradually. Besides, we maintain the diversity of pose variations in each collection as rich as possible.

TABLE I
EFFECTIVENESS OF THE PROPOSED APPROACH

| Methods | Market-1501 | | DukeMTMC-reID | |
|---|---|---|---|---|
| | Rank-1 | mAP | Rank-1 | mAP |
| Baseline [10] | 94.1 | 85.7 | 86.2 | 75.9 |
| +Auxiliary Data | 94.3 | 86.0 | 87.1 | 76.0 |
| +Feature Fusion | 94.9 | 86.5 | 87.8 | 76.5 |
| +Quality Estimation | 95.8 | 88.5 | 90.4 | 78.5 |

## III. EVALUATION AND ANALYSIS

We evaluate our proposed approach following standard experimental protocols on Market-1501 [2] and DukeMTMC-reID [9]. We provide comparisons to the state-of-the-art methods and perform ablation studies to verify the effectiveness.

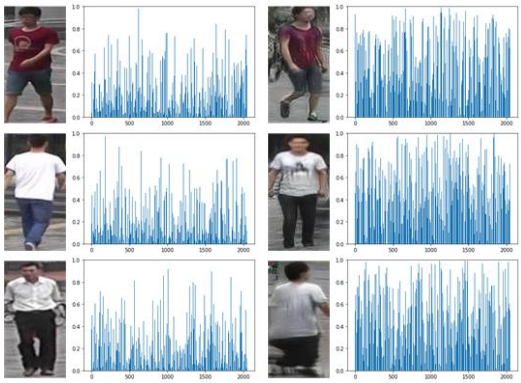
### A. Dataset Descriptions and Implementation Details

Market-1501 [2] contains 32668 labeled images of 1501 identities from 6 camera views. DukeMTMC-reID [9] contains 36411 labeled images of 1404 identities, which captured from 8 camera views. We implement our approach using Pytorch. Recently, a strong baseline is developed, it has achieved state-of-the-art performance on several public datasets. Therefore, we adopt the strong baseline [10] without center loss as our base network. We train all models utilizing a stochastic gradient descent method. The batch size is 64. The learning rate is initialized as 0.00035 and we spend 10 epochs increasing it to 0.0035 linearly. Then, the learning rate is decreased by 0.1 at the 30th epoch and 60th epoch, respectively. There are 150 epochs to train totally. Besides, we train with label smoothing regularization and triplet loss. The epsilon is set to 0.1 and 0.5 for real and fake images respectively. The input images are resized to 256×128 pixels and to pad the resized image 10 pixels utilizing zero values. Simultaneously, horizontal flip each image with 50% probability.

### B. Analysis of Contribution Effectiveness

In this section, we first implement a baseline which is trained using auxiliary data directly without any strategy, i.e., combining the origin images with auxiliary data directly to obtain an extended dataset for training. Then we train the baseline with combinations of feature fusion and object quality estimation to make full use of the auxiliary data. The experimental results are shown in TABLE I.

In TABLE I, training with the extended dataset directly, the model achieves a gain of 0.2% and 0.9% for rank-1 on Market-1501 and DukeMTMC-reID respectively. And our proposed object quality guided feature fusion network can achieve higher gains. Specifically, on Market-1501 dataset, our feature fusion network achieves a gain of 0.8% and 0.8% for rank-1 and mAP respectively. And with object quality estimation, we achieve a gain of 0.9% and 2.0% for rank-1 and mAP further. Similarly, on DukeMTMC-reID dataset, the proposed feature fusion network achieves a gain of 1.6% and 0.6% respectively. And with our self-supervised object quality estimation, the gain is 2.6% and 2.0% for rank-1 and mAP respectively. Experimental results demonstrate the effectiveness of our method.



(a) (b)

Fig. 4. Comparison of the object quality scores. (a) and (b) denote the quality scores of the real images and the auxiliary data respectively.

To further verify the effectiveness of the object quality estimation module, we select two sets of images which are from the real dataset and the auxiliary data respectively. The image quality in the left set is better, and the images in the right set are the pose transferred images, which contain lots of noise and disturbing features. Then we describe and compare the quality scores of the two sets, the results are shown in Fig. 4. In Fig. 4, (a) and (b) describe quality scores of the real images and the auxiliary images respectively. Intuitively, images in (a) can provide more information than those in (b). From the comparison results, the quality scores of real images are lower than those of auxiliary images. The comparison results in the figure are basically in line with expectations.

In the early stage of training, we just utilize the real images to obtain an initial model. And with the model converge, we add the auxiliary images in the training procedure gradually. Specially, for some epochs interval, one real image is reduced and one generated image is added in each collection. And the difference in the size of the interval also affects the results of the experiment. We analyze the influence of different intervals on the model performance. The results are shown in TABLE II. We choose 1, 3, 5, 10, 15 and 20 epochs as interval. When the interval is 15 epochs, our approach achieves the best performance, which the rank-1 accuracy arrives at 96.0% for Market-1501 and 91.9% for DukeMTMC-reID.

### C. Performance Comparison on Public Datasets

In this section, we report the comparison of our approach with the state-of-the-arts in recent years on Market-1501 and DukeMTMC-reID. The results are shown in TABLE III.

On each dataset, we acquire competitive results compared with the state-of-the-art approaches. Specifically, we achieve a gain of 1.7% and 4.2% in Rank-1 for Market-1501 and DukeMTMC-reID respectively. In addition, we achieve a gain of 2.8% and 3.6% in mAP respectively. Our final rank-1 accuracy arrives at 95.8% for Market-1501 and 90.4%

TABLE II
THE INFLUENCE OF DIFFERENT INTERVALS WHEN ADD THE AUXILIARY IMAGES ON THE MODEL PERFORMANCE DURING PERSON RE-IDENTIFICATION TRAINING

| Interval (Epochs) | Market-1501 | | DukeMTMC-reID | |
|---|---|---|---|---|
| | Rank-1 | mAP | Rank-1 | mAP |
| 1 | 94.5 | 86.0 | 87.8 | 76.0 |
| 3 | 94.7 | 86.0 | 87.8 | 76.1 |
| 5 | 94.8 | 86.2 | 88.1 | 76.3 |
| 10 | 95.3 | 87.3 | 88.5 | 77.0 |
| **15** | **95.8** | **88.5** | **90.4** | **78.5** |
| 20 | 95.5 | 88.2 | 90.1 | 78.0 |

TABLE III
COMPARISON WITH STATE-OF-THE-ART ON MARKET-1501 AND DUKEMTMC-REID. BEST AND SECOND BEST RESULTS ARE COLORED WITH RED AND BLUE RESPECTIVELY.

| Methods | Market-1501 | | DukeMTMC-reID | |
|---|---|---|---|---|
| | Rank-1 | mAP | Rank-1 | mAP |
| BoW+kissme [2] | 44.4 | 20.8 | 25.1 | 12.1 |
| IDE [11] | 72.5 | 46.0 | 65.2 | 45.0 |
| LSRO [9] | 84.0 | 66.1 | 67.7 | 47.1 |
| PT [12] | 87.7 | 68.9 | 78.5 | 56.9 |
| PN-GAN [6] | 89.4 | 72.6 | 73.6 | 53.2 |
| Camstyle [7] | 89.5 | 71.6 | 78.3 | 57.6 |
| FD-GAN [8] | 90.5 | 77.7 | 80.0 | 64.5 |
| VPM [13] | 93.0 | 80.8 | 83.6 | 72.6 |
| CtF [14] | 93.7 | 84.9 | 87.6 | 74.8 |
| FSAM [15] | 94.6 | 85.6 | 86.4 | 75.7 |
| DG-net [5] | 94.8 | 86.0 | 86.6 | 74.8 |
| GPS [16] | 95.2 | 87.8 | 88.2 | 78.7 |
| SCSN [17] | 95.7 | 88.5 | 90.1 | 79.0 |
| Base1ine [10] | 94.1 | 85.7 | 86.2 | 75.9 |
| **Ours** | 95.8 | 88.5 | 90.4 | 78.5 |

for DukeMTMC-reID. Experimental results show that the proposed feature fusion network can make full use of the auxiliary data generated with pose transferred model to obtain discriminative features, and our introduced self-supervised object quality estimation module can filter the noise.

## IV. CONCLUSIONS

In this paper, we propose object quality guided feature fusion for person re-identification. We first introduce a self-supervised object quality estimation module to estimate the quality of the auxiliary images. It can assign higher weight for the effective feature to filter the noise and the disturbing features generated with pose transferred model. Based on it we propose a novel feature fusion network. We propose a novel data sampling strategy to accomplish feature fusion in a collection-to-collection recognition manner for making full use of the auxiliary images. We demonstrate that our approach can effectively make full use of the auxiliary data significantly outperforms the existing state-of-the-art methods on two widely used public datasets (Market-1501 and DukeMTMC-reID).

## ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China under Grant No. 61872024, the National Key R&D Program of China under Grant No. 2018YFB2100603 and the Strategic Consulting Research Project of Henan Research Institute of China Engineering Science and Technology Development Strategy under Grant No. 2021HENZDA03.

## REFERENCES

[1] BG. Apurva and S. Shishir K. A survey of approaches and trends in person re-identification. *Image and vision computing*, 32(4):270–286, 2014.

[2] 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.

[3] G. Han, C. Yang, J. Liu, N. Sun, and X. Li. Person re-identification based on pose-guided generative adversarial network. In *Proceedings of the ACM Turing Celebration Conference-China*, pages 191–195, 2020.

[4] A. Khatun, S. Denman, S. Sridharan, and C. Fookes. Semantic consistency and identity mapping multi-component generative adversarial network for person re-identification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2267–2276, 2020.

[5] Z. Zheng, X. Yang, Z. Yu, L. Zheng, Y. Yang, and J. Kautz. Joint discriminative and generative learning for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2138–2147, 2019.

[6] X. Qian, Y. Fu, T. Xiang, W. Wang, J. Qiu, Y. Wu, YG. Jiang, and X. Xue. Pose-normalized image generation for person re-identification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 650–667, 2018.

[7] Z. Zhong, L. Zheng, Z. Zheng, S. Li, and Y Yang. Camera style adaptation for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5157–5166, 2018.

[8] Y. Ge, Z. Li, H. Zhao, G. Yin, S. Yi, X. Wang, and H. Li. Fd-gan: Pose-guided feature distilling gan for robust person re-identification. In *Advances in Neural Information Processing Systems*, pages 1222–1233, 2018.

[9] Z. Zheng, L. Zheng, and Y. Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3754–3762, 2017.

[10] 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 Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[11] L. Zheng, Y. Yang, and AG. Hauptmann. Person re-identification: Past, present and future. *arXiv preprint arXiv:1610.02984*, 2016.

[12] J. Liu, B. Ni, Y. Yan, P. Zhou, S. Cheng, and J. Hu. Pose transferrable person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4099–4108, 2018.

[13] Y. Sun, Q. Xu, Y. Li, C. Zhang, Y. Li, S. Wang, and J. Sun. Perceive where to focus: Learning visibility-aware part-level features for partial person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 393–402, 2019.

[14] G. Wang, S. Gong, J. Cheng, and Z. Hou. Faster person re-identification. In *European Conference on Computer Vision*, pages 275–292. Springer, 2020.

[15] P. Hong, T. Wu, A. Wu, X. Han, and W. Zheng. Fine-grained shape-appearance mutual learning for cloth-changing person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10513–10522, 2021.

[16] B. X. Nguyen, B. D. Nguyen, T. Do, E. Tjiputra, Q. D. Tran, and A. Nguyen. Graph-based person signature for person re-identifications. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3492–3501, 2021.

[17] X. Chen, C. Fu, Y. Zhao, F. Zheng, J. Song, R. Ji, and Y. Yang. Salience-guided cascaded suppression network for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3300–3310, 2020.