Exploring the Quality of GAN Generated Images for Person Re-Identification

Yiqi Jiang*
Weihua Chen*
yiqi.jyq@alibaba-inc.com
kugang.cwh@alibaba-inc.com
Alibaba Group
China

Xiuyu Sun†
Xiaoyu Shi
xiuyu.sxy@alibaba-inc.com
linyin.sxy@alibaba-inc.com
Alibaba Group
China

Fan Wang
Hao Li
fan.w@alibaba-inc.com
lihao.lh@alibaba-inc.com
Alibaba Group
China

ABSTRACT

Recently, GAN based method has demonstrated strong effectiveness in generating augmentation data for person re-identification (ReID), on account of its ability to bridge the gap between domains and enrich the data variety in feature space. However, most of the ReID works pick all the GAN generated data as additional training samples or evaluate the quality of GAN generation at the entire data set level, ignoring the image-level essential feature of data in ReID task. In this paper, we analyze the in-depth characteristics of ReID sample and solve the problem of "What makes a GAN-generated image good for ReID". Specifically, we propose to examine each data sample with id-consistency and diversity constraints by mapping image onto different spaces. With a metric-based sampling method, we demonstrate that not every GAN-generated data is beneficial for augmentation. Models trained with data filtered by our quality evaluation outperform those trained with the full augmentation set by a large margin. Extensive experiments show the effectiveness of our method on both supervised ReID task and unsupervised domain adaptation ReID task.

CCS CONCEPTS

• **Computing methodologies** → **Object identification**.

KEYWORDS

GAN, Person Re-Identification, Dataset, Augmentation, Sampling

ACM Reference Format:
Yiqi Jiang, Weihua Chen, Xiuyu Sun, Xiaoyu Shi, Fan Wang, and Hao Li. 2021. Exploring the Quality of GAN Generated Images for Person Re-Identification. In *Proceedings of the 29th ACM International Conference on Multimedia (MM '21), October 20–24, 2021, Virtual Event, China.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3474085.3475547

*Both authors contributed equally to this research.
†Corresponding author

MM '21, October 20–24, 2021, Virtual Event, China

ACM ISBN 978-1-4503-8651-7/21/10…$15.00
https://doi.org/10.1145/3474085.3475547

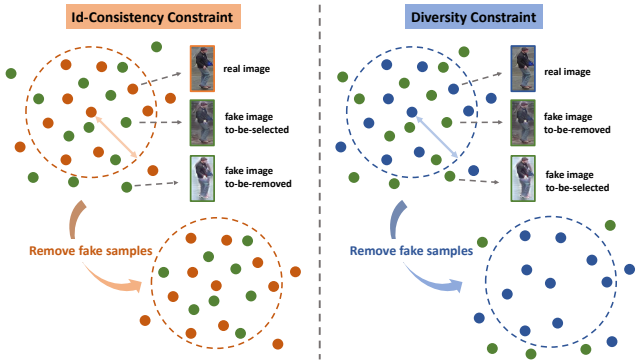


Figure 1: The illustration of id-consistency constraint (left) and diversity constraint (right). Id-consistency constraint favors the generated images which are close to the corresponding real ones in consistency space. Diversity constraint discards the generated images which are within a certain distance with the real image in diversity space.

1 INTRODUCTION

Given a query image, person re-identification (ReID) is the task of retrieving person images of the same identity from the gallery consisting of images collected across multiple cameras. Cameras often differ from each other regarding resolution, viewpoints, and illumination, which result in drastic changes in appearance and background of person images. Therefore, it has been crucial for ReID to learn representations that are robust against intra-class variations. Many methods have been proposed to learn a stable feature representation among different cameras, such as KISSME [18] and DNS [49]. Relying on the strong representation power of convolutional neural networks (CNNs), ReID methods [40, 52] achieve more powerful deep embeddings through deep learning. But it is still very challenging to learn a robust and discriminative representation to handle the appearance variance across domains (*e.g.* different cameras or different datasets).

Recently, more works exploit generative adversarial network (GAN) [11] to introduce additional augmented data into training set [55] so that ReID model can "see" more intra-class variations during training. For example, CamStyle [58] uses CycleGAN [60] to transfer images from one camera to the style of other cameras, and PTGAN [46] applies human pose conditioned GANs to generate pedestrian images of the same identity but with different poses.

Focusing on domain adaptive person ReID, SPGAN [6], CR-GAN [3] and PDA-Net [23] transfer labeled images from source domain into target domain to learn discriminative models on target domain. Although existing generative methods can synthesize plenty of visually pleasing data, there is no guarantee that all the generated images are beneficial to the final ReID training.

In this paper, We introduce two constraints when evaluating whether the generated images are favorable for ReID augmentation.

- **Id-consistency constraint**. The generated images should preserve identity information consistent with the real ones as much as possible.
- **Diversity constraint**. The generated images should diverse from the real images as much as possible to introduce more variations.

To better describe these constraints, we map the generated samples onto two different feature spaces, consistency space and diversity space. Intuitively, the consistency feature space represents identity-related appearance information in an image (*e.g.* clothing, hair, gender), which is the major information used for identifying a particular person. The diversity feature space contains all variations which are not shared within the images of the same identity (*e.g.* pose, illumination, camera views). As illustrated in Figure 1, an ideal generated image for augmentation, should be close to the corresponding real image in consistency space, and also should keep a certain distance from the real image in diversity space. An image that fails to satisfy these two constraints would be discarded because it might bring negative impact during training. Therefore, we design a metric-based sampling method to select a subset from the full set of augmentation data. Our experiments show that training with the filtered augmentations could achieve better ReID results compared with that trained with the full set of generated data.

Our contributions can be summarized as below:

1) We demonstrate that GAN generated images are not all beneficial for ReID training. Consistency and diversity constraints are presented to assess whether a generated image is suitable for ReID data augmentation.
2) The principles of consistency and diversity feature space are provided. Example projections for each feature space are presented, which we believe are great choice to be applied in practice.
3) Experiments show that, by mapping generated images to consistency and diversity feature spaces and filtering images with a simple sampling method, the augmentation set becomes more beneficial for ReID training, and the resulting ReID model achieves better performance compared with the model trained with full set of augmentation data.

2 RELATED WORKS

2.1 Deep Learning-based ReID Methods

Recent methods mainly rely on CNNs for its strong representation power to learn metric spaces and discriminative feature representations to handle data variations. Zheng *et al.* [52] regards ReID problem as a classification problem, considering each person as a particular class. Wu *et al.* [47] and Zheng *et al.* [54] combine identification loss with verification loss to achieve better metric space. Cheng *et al.* [4], Hermans *et al.* [13], and Ristani *et al.* [33] apply triplet loss with hard sample mining and achieve greater improvement in performance. Several recent works [24, 39, 45] employ pedestrian attributes and multi-task learning to enrich feature learning with more supervisions. Alternatives employ pedestrian alignment and human part matching to leverage on the human structure prior. Li *et al.* [22] and Sun *et al.* [40] split input images or feature maps horizontally to take advantage of local spatial cues. Other work, like ResNet-IBN [30], improves feature representation by enhancing backbone with instance normalization [15] to better catch style migration.

2.2 GAN-based Augmentation Methods

CNN training might suffer from under-fitting if training set doesn't possess enough variation or there is cross-domain problem where we don't have any labeled data in target domain. With recent progress in generative adversarial networks (GANs) [11], generative models have become an appealing choice to solve these problem. Zheng *et al.* [55] use DC-GAN [32] to improve the discrimination ability of learned CNN embeddings. Qian *et al.* [31] and Wei *et al.* [46] enrich the training space by generating different pose images. Zhong *et al.* [58] propose CamStyle data augmentation approach which transfers images from one camera to the style of another camera. FD-GAN [10] generates a new person image of the same identity as the input with a given pose. Different from the works of Wei *et al.* [46] and Zhong *et al.* [58], it distills identity-related and pose-unrelated features from the input image, getting rid of pose-related information disturbing the ReID task.

2.3 Quantitative Measurements for GAN

Measuring the quality of GAN generated images has attracted extensive attention in various applications. One of the most common ways to evaluate GANs is the Inception Score [35]. It uses an Inceptionv3 network [41] pretrained on ImageNet to compute the logits of the generated samples. Similar to the Inception Score, Fréchet Inception Distance [14] also relies on Inception's evaluation to measure quality of generated samples. Different from evaluating GAN results on image level, FID takes the features from the Inception's penultimate layer and estimates Gaussian approximations from both real and generated images, measuring quality by computing distribution distance. GANtrain and GANtest [36] take generated data and real data as training sample to train a classification network and evaluate on the other set, respectively. Scores of generated images are given by comparing the performance of GANtrain and GANtest with a baseline network trained and evaluated both on real data. FID and GANtrain/GANtest evaluate GAN quality on set-level, which estimate whether the distribution of generated images is good or not, while the problem remains untouched that whether they are suitable to serve as augmentation data for training. In this paper, we focus on assessing the GAN generated image at image level for their potential benefits in training a ReID model.

3 THE PROPOSED METHOD

The pipeline of our proposed method is illustrated in Figure 2, Firstly, GAN models are applied to generate more images for augmentation. Then, E_c and E_d is provided as projectors to map both

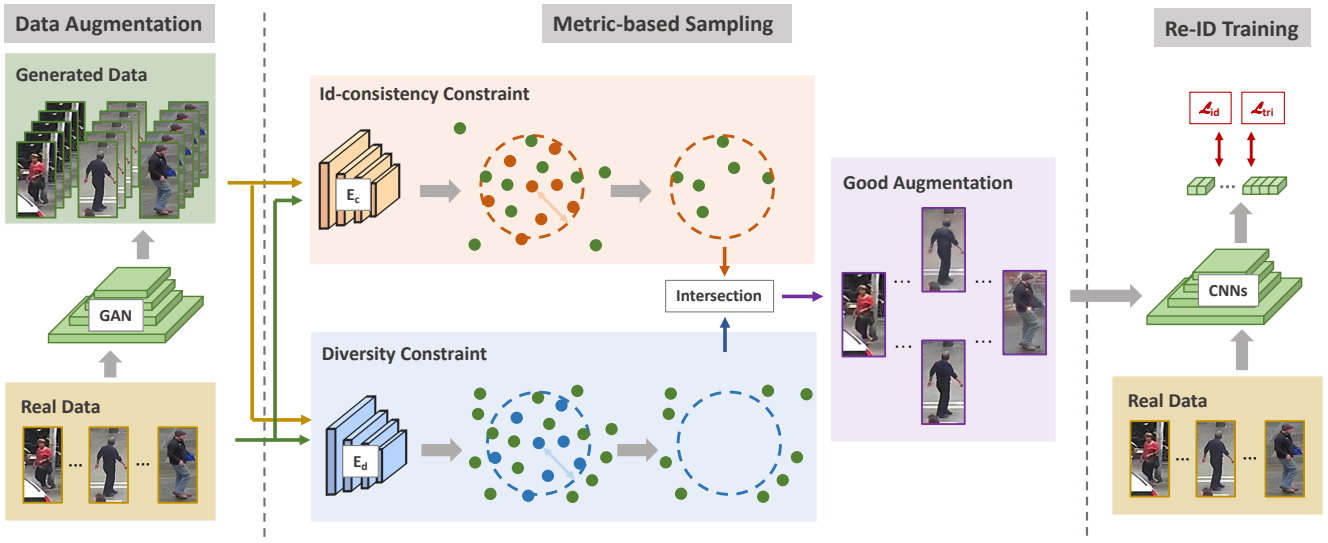


Figure 2: The pipeline of the proposed method. For each real image, we firstly use GAN to generate more images for augmentation. Then, Consistency encoder E_c and diversity encoder E_d are applied to both real data and generated data to extract consistency features and diversity features. Compared with the corresponding real image, generated images with small distance in consistency space and large distance in diversity space are selected, respectively. The intersection of images in consistency space and images in diversity space is applied to achieve good augmentations. Finally, good augmentations are combined with real data to train the ReID task.

real and generated images onto consistency and diversity feature space, respectively. Metric-based sampling method is further applied to select good augmentations. Finally, good augmentations are combined with real data to train the ReID task. Details of above modules are introduced as follows.

3.1 Projection Ecoders

We introduce two constraints, id-consistency constraint and diversity constraint, to assess the suitability of GAN generated image for augmentation purpose. In practice, the generated images are mapped onto two different feature space to estimate whether they satisfy these constraints. Two encoders are provided as feature space projectors and are denoted as E_c and E_d, respectively.

Consistency Feature Encoder E_c The consistency feature space represents identity related information, such as face, body shape, clothing, hair, gender, and other attributes which do not change across cameras and could help identify a particular person. Ideally, a ReID model could represent images in a feature space which is id-consistent in the domain of training data. The better the model being trained, the more compact the feature distribution would be for each identity cluster.

Furthermore, some researchers present representation disentangling methods [7, 10, 20, 53] which decompose feature into identity related and unrelated explicitly to enhance the representation power in id-related space. Inspired by this line of research, we build the encoder E_c based on the identity related branch in ISGAN [7].

Specifically, to disentangle id-related and unrelated features, two encoders E_R and E_U join the structure of GAN. Given a pair of images I_a and I_p of the same identity, E_R and E_U extract identity-related features $\phi_R(I_a)$ and $\phi_R(I_p)$, and identity-unrelated features, $\phi_U(I_a)$ and $\phi_U(I_p)$. After shuffling and regrouping ϕ_R and ϕ_U, the GAN reconstructs images and force the reconstructed image to recreate the original input when $I_a = I_p$, and to generate images with same identity when $I_a \neq I_p$. The shuffling loss and identity loss is defined as following:

$$L_S = \sum_{i,j \in a,p} \left\| I_i - G(\phi_R(I_j) \oplus \phi_U(I_i)) \right\|_1 \quad (1)$$

where $\oplus$ represents the concatenation of features, please refer to [7] for more details. With this constraint design, the encoder E_R extracts feature with a strong identity consistency, which makes it eligible to serve as our consistency space encoder E_c.

Diversity Feature Encoder E_d Diversity feature space contains all feature variations which are not shared within images of the same identity (*e.g.* poses, lighting conditions, resolutions, viewpoints). These features enrich the variance in ReID samples and could also bridge the domain gap between different scenarios or datasets.

Generally, there are many candidate models which could be regarded as an encoder or a projector to map images to diversity feature space, but to different extent. A random initialized model projects images to a random feature space, which could capture random representations of diversities in ReID. The id-unrelated encoder branch of disentangling model ISGAN, seems to be another good choice, but most of the constraints in ISGAN are focus on enhancing the id-related encoder E_R, leaving constraint on unrelated encoder E_U a Gaussian noise constraint.

ImageNet [34] is a large scale classification data set and contains a great deal of variations in illuminations, view point, *etc.* A well-trained classification model on ImageNet possesses powerful representation ability, which makes it able to build a rich diversity feature space. As a result, we choose a ResNet-50 [12] pretrained on ImageNet as one of our diversity space encoders.

3.2 Metric-based Sampling

As mentioned above, two encoders, E_c and E_d, are built to extract consistency features and diversity features, and map each image onto these two feature spaces simultaneously. In each space, we calculate the id center C_i of all real image features for every identity cluster i, denoted as C_c^i in consistency space and C_d^i in diversity space.

$$C_c^i = \frac{1}{n_i}\sum_{j=1}^{n_i} E_c(x_j), \qquad C_d^i = \frac{1}{n_i}\sum_{j=1}^{n_i} E_d(x_j), \tag{2}$$

where n_i denotes the number of images within identity i, and x_j represents the j_{th} image in identity i.

For each generated image with identity i, L2 distance between its feature and the corresponding id center is computed in the two feature spaces:

$$d_c^j = \left\|E_c(x_j) - C_c^i\right\|_2, \qquad d_d^j = \left\|E_d(x_j) - C_d^i\right\|_2, \tag{3}$$

A simple sampling rule is designed based on the distance to measure the effectiveness of features in each space. Specifically, to narrow down the domain gap in consistency feature space, we keep the generated images whose distance to C_c^i are less than a specified threshold T_c and regard these images as consistency-candidates. To enrich potential variations in diversity feature space, the generated images whose distances to C_d^i are greater than a specified threshold T_d are kept as diversity-candidates. A sample beneficial to ReID training should satisfy above requirements in both consistency and diversity space. Therefore, an intersection of both candidate sets is made to generate sampled augmentation data set.

$$S_{sampled} = S_{d_c<T_c} \cap S_{d_d>T_d} \tag{4}$$

where $S_{d_c<T_c}$ and $S_{d_d>T_d}$ denote consistency candidates and diversity candidates respectively.

Besides the criteria of filtering images one by one, their relationship should also be taken into consideration. Generated images which are close to each other in the diversity feature space do not bring in much diversity information; instead, too many duplicated training samples would increase training time and even cause imbalanced training which needs to be treated carefully.

To handle this problem, we employ Local Outlier Factor (LOF) [1] to monitor the density of each generated image in diversity feature space. If the image holds a high density, we will randomly drop it with a probability of α, which is set to 0.3 in our experiments by experience.

The final sampled augmentation can be obtained by

$$S_{final} = S_{d_c<T_c} \cap S_{d_d>T_d} \cap S_{lof} \tag{5}$$

where S_{lof} indicates the diversity candidate images filtered by LOF-based monitor in diversity feature space.

3.3 ReID Training

Given the new training set consists of all the real images and the sampled fake images, the training strategy should be carefully designed to exploit the effectiveness of these augmentations.

Firstly, it's important to control the balance between real and fake images during training. We use $B = P \times K$ images to form the mini-batch, where P and K denotes the number of different person-ids and the number of different images per person-id, respectively. In K, we set $K = M + N$, where M and N indicates the number of real and fake training samples. By setting the ratio of $\frac{M}{N}$, we can control the effect of real and fake images on training.

Secondly, most of works combine cross-entropy loss and triplet loss together to train ReID model. However, from experiments, we found the triplet loss on the fake data brings a negative effect to the model. A possible explanation is that there are still noises in the fake augmented data to some extent. The triplet loss is more sensitive to the noise than the classification loss, because a noise sample would affect all the related pairs in the triplet loss. As a result, we only use the cross-entropy loss for training on fake augmented images.

The loss function is described as:

$$L_{ReID} = \frac{1}{N}\sum_{i=1}^{N} L_R^i + \frac{1}{M}\sum_{i=1}^{M} L_F^i, \tag{6}$$

where $L_R^i = L_{id}(x_R^i) + L_{tri}(x_R^i)$ and $L_F^i = L_{id}(x_F^i)$. x_R^i is the real image and x_F^i is the generated image. L_{id} is cross-entropy loss function and L_{tri} is triplet loss function.

Thirdly, along the assumption above, the fake augmented data still contain noise to some extent. We involve Label Smoothing Regularization(LSR) [22] to alleviate the impact of noise. Specifically, we apply LSR on both the real images and the fake images to softly distributed their labels as follow:

$$q_{LSR}(c) = \begin{cases} 1 - \epsilon + \frac{\epsilon}{C} & c = y \\ \frac{\epsilon}{C} & c \neq y, \end{cases} \tag{7}$$

where ϵ is a small constant to encourage the model to be less confident on the training set. As fake images contain much more noise, in our study, ϵ_r is set to be 0.1 for real images and ϵ_f is set to be 0.3 for fake images, respectively.

4 EXPERIMENTS

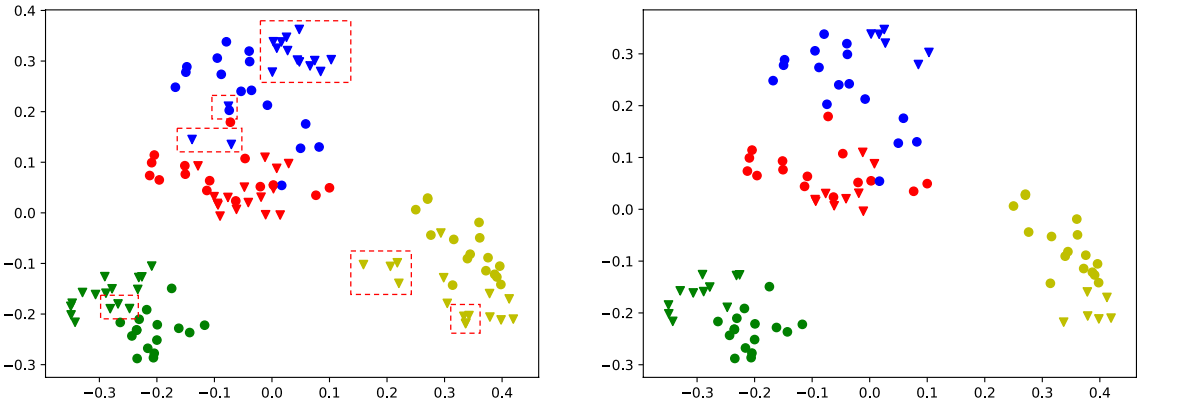
4.1 Experiment Settings

Datasets and Evaluation Settings We evaluate our method on three representative ReID datasets, Market-1501 [51], DukeMTMC-reID [55] and MSMT-17 [46]. Cumulative Matching Characteristic (CMC) at rank-1 and mAP [51] are used as the evaluation metrics.
GAN Model for Augmentation CamStyle [58] has proven its ability to generate person images across different camera style, but it has two weaknesses. First it is based on the CycleGAN [60], which requires to train a translation model for each camera pair. For a N-camera dataset, it would be too time-consuming to train $N(N-1)$ models. To solve this problem, we import StarGAN [5] into the CamStyle [58] which allows us to train multi-camera image-image translation with a single model.

Second, the original backbone of CamStyle model only contains several layers, which is hard to take full advantages of CamStyle.

Table 1: Comparisons with the baseline model and state-of-the-art supervised ReID methods on Market-1501, DukeMTMC-ReID and MSMT-17 datasets (%). The best and second best results are presented in bold and underline.

Method		Market-1501		DukeMTMC-reID		MSMT-17	
		Rank@1	mAP	Rank@1	mAP	Rank@1	mAP
MGN [44]	MM2018	95.7	86.9	88.7	78.4	76.9	52.1
ISGAN [7]	NIPS2019	95.2	87.1	<u>90.9</u>	79.5	-	-
DGNet [53]	CVPR2019	94.8	86.0	86.6	74.8	-	-
PGFA [29]	ICCV2019	91.2	76.8	82.6	65.5	-	-
OSNet[59]	ICCV2019	94.8	84.9	88.6	73.5	78.7	52.9
CBN [62]	ECCV2020	91.3	77.3	82.5	67.3	72.8	42.9
SAN [16]	AAAI2020	<u>96.1</u>	88.0	87.9	75.7	79.2	55.7
SCSN [2]	CVPR2020	95.7	88.5	**91.0**	79.2	**83.8**	53.5
HOReID [43]	CVPR2020	94.2	84.9	86.9	75.6	-	-
ISP [61]	ECCV2020	95.3	<u>88.6</u>	89.6	**80.0**	-	-
Baseline	-	95.1	87.8	89.6	79.1	81.1	<u>56.5</u>
OriginalCamStyle	-	94.9	85.5	86.6	75.6	80.6	53.7
ModifiedCamStyle	-	95.5	87.9	88.4	78.2	81.0	56.2
Ours	-	**96.2**	**89.2**	**91.0**	<u>79.9</u>	<u>81.9</u>	**57.2**



(a) Feature distribution of real samples and full augmentations (b) Feature distribution of real samples and good augmentations

Figure 3: t-SNE [27] visualization on Market-1501. Circle dots denote the real samples and triangles denote the generated samples. Different colors represent different identities. (a) The feature distribution of real samples and full augmentation. (b) The feature distribution of real samples and good augmentations. Full augmentations introduce a large number of noisy or duplicate samples (red boxes in (a)). By proper sampling based on the proposed consistency constraint and diversity constraint, most of bad samples are removed, as shown in (b).

So we modify the backbone carefully. Specifically, the modified generator consists of four down-sampling blocks, four intermediate blocks and four up-sampling blocks. Instance normalization (IN) [3] and adaptive instance normalization (AdaIN) [15] are used for down-sampling and up-sampling respectively. Discriminator is a multi-task discriminator [28], which contains L linear output branches, where L indicates the number of cameras. The discriminator contains six pre-activation residual blocks with leaky ReLU [26]. Finally, we use Adam optimizer [17] with $\beta_1 = 0.5$ and $\beta 2 = 0.999$. The learning rates for G and D are set to 10^{-4}. The results of the original CamStyle (*OriginalCamstyle*) and our modified CamStyle (*ModifiedCamStyle*) are shown in Table. 1. It is obvious that our modified CamStyle provides a much better performance than the original CamStyle. And the augmented images of the original CamStyle is even worse than the Baseline.

ReID Baseline Model Our method can be applied to many ReID models, this paper uses reid-strongbaselne [25] as an example

Table 2: Performance evaluation with different encoders in Supervised ReID on Market-1501, DukeMTMC-ReID and MSMT-17. Top-left corner shows *Baseline* model trained with real images only, and *ModifiedCamStyle* trained with real images and the full set of CamStyle augmentation, performance shown in the format of Rank@1/mAP(# of images). '# of images' indicates the number of generated images used for augmentation.

Market-1501 Baseline: 95.1% / 87.8% (12936) ModifiedCamStyle: 95.5% / 87.9% (64680)		E_d			
		random-initilized	ISGAN-unrelated	ReID-Duke	ImageNet
		Rank@1 mAP (# of images)	Rank@1 mAP (# of images)	Rank@1 mAP (# of images)	Rank@1 mAP (# of images)
E_c	ISGAN-related	95.4% 88.2% (18900)	95.3% 88.5% (17733)	95.6% 88.6% (24842)	**96.2% 89.2%** (17835)
	ReID-Market	95.3% 87.9% (17988)	95.4% 88.1% (19848)	95.4% 88.2% (17273)	95.7% 88.7% (12771)

DukeMTMC-ReID Baseline: 89.6% / 79.1% (16522) ModifiedCamStyle: 88.4% / 78.2% (115654)		E_d			
		random-initilized	ISGAN-unrelated	ReID-Market	ImageNet
		Rank@1 mAP (# of images)	Rank@1 mAP (# of images)	Rank@1 mAP (# of images)	Rank@1 mAP (# of images)
E_c	ISGAN-related	89.0% 78.5% (29500)	89.3% 78.8% (28628)	89.8% 79.2% (35849)	**91.0% 79.9%** (23108)
	ReID-Duke	88.5% 77.9% (34208)	89.1% 78.2% (33767)	89.4% 78.9% (32384)	89.7% 79.0% (22439)

MSMT-17 Baseline: 81.1% / 56.5% (32621) ModifiedCamStyle: 81.0% / 56.2% (195726)		E_d			
		random-initilized	ISGAN-unrelated	ReID-Market	ImageNet
		Rank@1 mAP (# of images)	Rank@1 mAP (# of images)	Rank@1 mAP (# of images)	Rank@1 mAP (# of images)
E_c	ISGAN-related	81.3% 56.3% (44646)	81.3% 56.4% (45144)	81.5% 56.8% (51965)	**81.9% 57.2%** (45382)
	ReID-MSMT	81.0% 56.2% (44356)	80.9% 55.9% (45052)	81.1% 56.4% (49981)	81.7% 56.6% (43256)

baseline to verify the effectiveness of our approach. We train baseline model with 256x128 input images, which is processed by random cropping [19], random horizontal flipping [37] and random erasing [56]. We perform 60 epochs training on the model, using $B = P \times K = 6 \times 9$ as mini-batch, where $P = 6$ indicates the number of person-ids and $K = 9$ indicates the number of real images per person-id from training dataset.

4.2 Comparison with State of the arts

The comparison with state-of-the-arts is shown in Table.1. *Baseline* is our implementation based on reid-strongbaseline [25], where only the real data are used for supervised ReID training. The original CamStyle and the StarGAN-based CamStyle are employed to generate images based on the real data which are considered as full augmentation, and denoted as *OriginalCamStyle* and *ModifiedCamStyle* respectively. *Ours* shows the results of our method, which samples the generated images of StarGAN-based CamStyle with id-consistency and diversity constraints before adding into the training. With a state-of-the-art baseline model, the *ModifiedCamStyle* only brings a slight improvement on some datasets. Comparing *Ours* with *ModifiedCamStyle*, the performance is improved largely, which suggest that NOT every generated image benefits the ReID training. Eliminating the augmentations that do not satisfy our constraints could further improve the effectiveness of data augmentation.

Figure. 3 gives the t-SNE [27] visualization of feature distribution on Market-1501, circle dots denote real images and triangles denote generated images. Although full augmentations can fill in gaps between real images and extend class boundaries, it also brings in a large number of noisy or duplicate samples, as shown in Figure. 3(a), which have negative impact on ReID training. Figure. 3(b) shows the feature distribution after the generated samples being filtered based on proposed constraints. The removal of noise samples and duplicate samples help defining more reasonable distribution of training data, making it easier for ReID model to learn a robust yet discriminative model with better performance.

4.3 Ablation Study

In this section, we firstly explore different choices of encoders to build the consistency and diversity spaces. Then we conduct ablation studies to evaluate each component in our method.

Consistency Encoders Table.2 shows the results of different encoders on Market-1501, DukeMTMC-ReID and MSMT-17. Before comparison, on the left top, *Baseline* represents the result without any GAN augmentation and $ModifiedCamStyle$ represents using full set of CamStyle augmentation. The number of augmented fake images is listed in the brackets behind the performance. It is worth noting that MSMT-17 has 15 cameras and there will be 247380

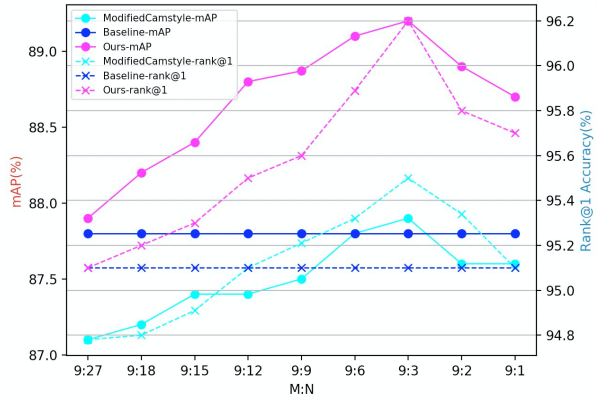


Figure 4: Evaluation with different ratio of real data and sampled augmentations ($M : N$) in a mini-batch on Market-1501.

Table 3: Performance Evaluation on Market-1501 using different sets. '# of images' indicates the number of generated images used for augmentation.

Method	# of Images	Rank@1	mAP
Baseline	-	95.1	87.8
ModifiedCamStyle	64680	95.5	87.9
Random	17835	94.9	87.6
Consistency	32340	95.4	88.3
Diversity	32340	95.7	88.9
Ignored	46845	94.2	87.1
Ours	20386	95.6	88.8
Ours w/ lof	17835	**96.2**	**89.2**

images, which is a large number. In order to save disk space and training time, we randomly select six cameras to do camstyle.

The results of different consistency encoders are listed in different rows. Intuitively, a state-of-the-art ReID model itself should be powerful enough to extract id-consistency information. Therefore, besides using the ISGAN id-related model to serve as E_c, we also train a state-of-the-art ReID model[25] as an alternative choice to build the consistency space. ISGAN id-related model and the state-of-the-art ReID model are denoted as $ISGAN-related$ and $ReID-X$ respectively, where X is the dataset name.

Taking Market-1501 as an example, comparing the top row and bottom row of Table 2, it can be observed that the consistency space built through ISGAN id-related branch outperforms the space built through the ReID model, even though the ReID model has already been carefully trained. It is expected that the generated images should preserve as much identity information as possible. ISGAN employs additional supervision and a shuffle strategy to enhance the ability of identity consistent representation. This conclusion is also consistent on DukeMTMC-ReID and MSMT-17.

Diversity Encoders Four different choices of the diversity feature space encoders are explored in columns in Table.2: (1) a model with random initialized parameters, denoted as *random-initialized*;

Table 4: Performance Evaluation on Market-1501 using different threshold. T_c and T_d are the thresholds for consistency space and diversity space respectively. '# of images' indicates the number of generated images used for augmentation.

T_c	T_d	# of Images	Rank@1	mAP
Baseline		-	95.1	87.8
mean	mean	16996	95.6	88.9
mean	median	17990	95.8	89.1
median	mean	15162	95.6	88.8
median	median	17835	**96.2**	**89.2**

Table 5: Performance Evaluation on Market-1501 using different loss functions. ID: Cross-Entropy Loss, Tri: Triplet Loss.

Training data	L_R	L_F	Rank@1	mAP
Real	ID+Tri	None	95.1	87.8
Real+Fake	ID+Tri	ID	**96.2**	**89.2**
Real+Fake	ID+Tri	Tri	93.2	86.1
Real+Fake	ID+Tri	ID+Tri	94.3	86.8

(2) a model from ISGAN id-unrelated branch, denoted as *ISGAN-unrelated*; (3) a well-trained ReID model on dataset in different domain, denoted as *ReID-X* with X representing the domain different from consistent space; (4) the proposed model pretrained on ImageNet, denoted as *ImageNet*. We can see that *ImageNet* outperforms all other choices of diversity encoders. As we argued in diversity constraint, the generated image should contain as much diversity changes as possible. The ImageNet pretrained model generates a feature space with rich texture, illumination and other representations. *random-initialized* and *ISGAN-unrelated* produce the lowest performances, because the space built by these two models is either a random space or a Gaussian noise space, which captures little information about image diversity. The *ReID-X* encoder outperforms *random-initialized* but is inferior to *ImageNet*. This is because the ReID diversity encoder is trained on the data from another domain, which could bring some variance. But it still contains a certain degree of the consistency, which leads to some conflicts with the consistency space during sampling. For instance, some "best augmentation sample" could be eliminated in diversity space since they are with the same identity and with insufficient distance from the id-center in diversity space. Consequently, the sampling result of ReID diversity encoder is not comparable with the result of *ImageNet* encoder.

Sampling Method Analysis Table. 3 shows the effectiveness of the proposed sampling method. $ModifiedCamStyle$ denotes using full fake images for augmentation. *Random* denotes randomly selecting the same number of fake images as our proposed method. $Consistency$ and $Diversity$ denotes using only id-consistency constraint and diversity constraint, respectively. *Ignore* denotes selecting samples that are ignored by our proposed method. We can see that *Random* and *Ignore* has the lower performance while $Consistency$ and $Diversity$ consistently gets higher performance

Table 6: Comparisons with the baseline model and state-of-the-art UDA ReID methods between Market-1501 and DukeMTMC-ReID datasets (%). The best and second-best results are presented in bold and underline.

Method		Market-1501 → DukeMTMC-reID				DukeMTMC-reID → Market-1501			
		mAP	Rank@1	Rank@5	Rank@10	mAP	Rank@1	Rank@5	Rank@10
PT-GAN [46]	CVPR2018	-	27.4	-	50.7	-	38.6	-	66.1
SPGAN-LMP [6]	CVPR2018	26.4	46.9	62.6	68.5	26.9	58.1	76.0	82.7
ECN [57]	CVPR2019	40.4	63.3	75.8	80.4	43.0	75.1	87.6	81.6
Theory [38]	PR2020	49.0	68.4	80.1	83.5	53.7	75.8	89.5	93.2
SSG [8]	CVPR2019	53.4	73.0	80.6	83.2	58.3	80.0	90.0	92.4
AD-Cluster [48]	CVPR2020	54.1	72.6	82.5	85.5	68.3	86.7	94.4	96.5
MMCL [42]	CVPR2020	51.4	72.4	82.9	85.0	60.4	84.4	92.8	95.0
JVTC [21]	ECCV2020	56.2	75.0	85.1	88.2	61.1	83.8	93.0	95.2
NMRT [50]	ECCV2020	62.2	77.8	86.9	89.5	71.7	87.8	94.6	96.5
MMT [9]	ICLR2020	**68.7**	**81.8**	**91.2**	**93.4**	74.5	91.1	**96.5**	**98.2**
Baseline	-	56.1	74.1	84.1	87.3	62.8	84.7	93.4	95.6
ModifiedCamStyle	-	58.7	76.0	85.7	89.5	74.9	90.8	95.6	97.3
Ours	-	64.4	79.4	88.6	91.0	**78.6**	**91.3**	96.3	97.7

than *Baseline*. *Ours* uses both id-consistency and diversity constraint to do sampling and gets much higher performance. Considering that generated images which hold a high density do not bring in much diversity information and need to be dropped, *Ours* $w/\ lof$ uses LOF monitor to further improve the quality of augmentation data set, which also partly handles the data imbalance problem and achieves the best performance. The best performance of *Ours* $w/\ lof$ shows the effectiveness of the monitor and the monitor can be regarded as a beneficial complement to our proposed constraints.

Threshold Analysis Threshold T_c and T_d in Equ.5 are important parameters in our sampling process. We explore the median value and the mean value of the distances from the center of all pictures in each category, as our thresholds. The results are shown in Table 4. The performance of ReID model varies with the setting of threshold, but they all consistently higher than baseline, which shows the effectiveness of consistency constraints and diversity constraints. Meanwhile, using median as threshold get highest performance, and it is used as our default setting through all the experiments.

Data Ratio Analysis The ratio of $\frac{M}{N}$ is critical for ReID model training, where M and N indicate the number of real and fake samples in mini-batch. Figure. 4 shows results on different ratio. The x-axis shows the different M and N used for training. Our proposed method with different $\frac{M}{N}$ consistently improve over baseline, while *ModifiedCamStyle* achieves lower performance than baseline when using more fake data than real data. We argue that *ModifiedCamStyle* contains a large number of fake images that do not meet id-consistency and diversity constraints, and are not suitable for ReID training. *Ours* filters out those fake images and achieve the best performance with the data ratio $M : N = 9 : 3$.

Loss Function Analysis Table. 5 gives results of using different loss functions on sampled fake images. Recently works have proved that using ID loss and triplet loss together to train ReID model is beneficial. However, ReID performance drops a lot when we apply either two losses or only triplet loss on sampled augmentations. A possible explanation is that there still exists noise in fake images even after our sampling. The triplet loss is more sensitive to these noises than the classification loss, resulting in worse result.

4.4 Experiments on UDA ReID Task

Besides the work on the Supervised ReID task, we also conduct experiments on Unsupervised Domain Adaptive(UDA) ReID to further show the effectiveness of our method. The comparison is shown in Table.6. *Baseline* is our implementation based on [38], where only the real data in target domain are used for UDA ReID training. StarGAN-based CamStyle is employed to generate images based on the target-domain real data, which is considered as full augmentation and denoted as *ModifiedCamStyle*. *Ours* shows the results of our method, which samples the generated images of StarGAN-based CamStyle with the proposed two constraints before adding into the training.

The *ModifiedCamStyle* increases the *Baseline* performance of UDA ReID without bells and whistles. Comparing *Ours* with *ModifiedCamStyle*, the performance is further improved, which suggest that NOT every generated image benefits the ReID training. Eliminating the augmentations that do not satisfy our constraints could further improve the effectiveness of data augmentation. More interesting, we can find that a low baseline (*i.e.* 84.7%/62.8%) can be improved to a state-of-the-art result (*i.e.* 91.3%/78.6%) with the help of our method.

5 CONCLUSION

In this paper, we claim that not every GAN generated sample is beneficial for ReID Training. A good augmentation should satisfy both identity consistency constraint and diversity constraint. We propose to project person image into consistency and diversity feature spaces and select good augmentations through a simple distance metric sampling method. Experiments on several benchmarks demonstrate that the constraints on both consistency and diversity are necessary for GAN based augmentation. We believe that the consistency and diversity spaces could be further explored with other well-designed encoders. Besides, a more complicated sampling method which exploits the interaction between samples or performs a combinatorial optimization, could yield better sub set of augmentation. We leave these as our future work.

REFERENCES

[1] Markus M Breunig, Hans-Peter Kriegel, Raymond T Ng, and Jörg Sander. 2000. LOF: identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*. 93–104.

[2] Xuesong Chen, Canmiao Fu, Yong Zhao, Feng Zheng, Jingkuan Song, Rongrong Ji, and Yi Yang. 2020. Salience-Guided Cascaded Suppression Network for Person Re-Identification. In *CVPR*.

[3] Yanbei Chen, Xiatian Zhu, and Shaogang Gong. 2019. Instance-guided context rendering for cross-domain person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*. 232–242.

[4] De Cheng, Yihong Gong, Sanping Zhou, Jinjun Wang, and Nanning Zheng. 2016. Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In *Proceedings of the iEEE conference on computer vision and pattern recognition*. 1335–1344.

[5] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. 2018. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 8789–8797.

[6] Weijian Deng, Liang Zheng, Qixiang Ye, Guoliang Kang, Yi Yang, and Jianbin Jiao. 2018. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 994–1003.

[7] Chanho Eom and Bumsub Ham. 2019. Learning disentangled representation for robust person re-identification. In *Advances in Neural Information Processing Systems*. 5297–5308.

[8] Yang Fu, Yunchao Wei, Guanshuo Wang, Yuqian Zhou, Honghui Shi, and Thomas S Huang. 2019. Self-similarity grouping: A simple unsupervised cross domain adaptation approach for person re-identification. In *ICCV*. 6112–6121.

[9] Yixiao Ge, Dapeng Chen, and Hongsheng Li. 2020. Mutual Mean-Teaching: Pseudo Label Refinery for Unsupervised Domain Adaptation on Person Re-identification. In *ICLR*.

[10] Yixiao Ge, Zhuowan Li, Haiyu Zhao, Guojun Yin, Shuai Yi, Xiaogang Wang, et al. 2018. Fd-gan: Pose-guided feature distilling gan for robust person re-identification. In *Advances in neural information processing systems*. 1222–1233.

[11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in neural information processing systems*. 2672–2680.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. (2016).

[13] Alexander Hermans, Lucas Beyer, and Bastian Leibe. 2017. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737* (2017).

[14] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in neural information processing systems*. 6626–6637.

[15] Xun Huang and Serge Belongie. 2017. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*. 1501–1510.

[16] Xin Jin, Cuiling Lan, Wenjun Zeng, Guoqiang Wei, and Zhibo Chen. 2020. Semantics-aligned representation learning for person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 11173–11180.

[17] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[18] Martin Koestinger, Martin Hirzer, Paul Wohlhart, Peter M Roth, and Horst Bischof. 2012. Large scale metric learning from equivalence constraints. In *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2288–2295.

[19] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2017. Imagenet classification with deep convolutional neural networks. *Commun. ACM* 60, 6 (2017), 84–90.

[20] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. 2018. Diverse image-to-image translation via disentangled representations. In *Proceedings of the European conference on computer vision (ECCV)*. 35–51.

[21] Jianing Li and Shiliang Zhang. 2020. Joint Visual and Temporal Consistency for Unsupervised Domain Adaptive Person Re-Identification. In *ECCV*.

[22] Wei Li, Xiatian Zhu, and Shaogang Gong. 2017. Person re-identification by deep joint learning of multi-loss classification. *arXiv preprint arXiv:1705.04724* (2017).

[23] Yu-Jhe Li, Ci-Siang Lin, Yan-Bo Lin, and Yu-Chiang Frank Wang. 2019. Cross-dataset person re-identification via unsupervised pose disentanglement and adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 7919–7929.

[24] Yutian Lin, Liang Zheng, Zhedong Zheng, Yu Wu, Zhilan Hu, Chenggang Yan, and Yi Yang. 2019. Improving person re-identification by attribute and identity learning. *Pattern Recognition* 95 (2019), 151–161.

[25] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. 2019. 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*. 0–0.

[26] Andrew L Maas, Awni Y Hannun, and Andrew Y Ng. 2013. Rectifier nonlinearities improve neural network acoustic models. In *Proc. icml*, Vol. 30. Citeseer, 3.

[27] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, Nov (2008), 2579–2605.

[28] Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. 2018. Which training methods for GANs do actually converge?. In *International conference on machine learning*. PMLR, 3481–3490.

[29] Jiaxu Miao, Yu Wu, Ping Liu, Yuhang Ding, and Yi Yang. 2019. Pose-guided feature alignment for occluded person re-identification. In *ICCV*. 542–551.

[30] Xingang Pan, Ping Luo, Jianping Shi, and Xiaoou Tang. 2018. Two at once: Enhancing learning and generalization capacities via ibn-net. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 464–479.

[31] Xuelin Qian, Yanwei Fu, Tao Xiang, Wenxuan Wang, Jie Qiu, Yang Wu, Yu-Gang Jiang, and Xiangyang Xue. 2018. Pose-normalized image generation for person re-identification. In *Proceedings of the European conference on computer vision (ECCV)*. 650–667.

[32] Alec Radford, Luke Metz, and Soumith Chintala. 2015. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434* (2015).

[33] Ergys Ristani and Carlo Tomasi. 2018. Features for multi-target multi-camera tracking and re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 6036–6046.

[34] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, and Michael Bernstein. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision* 115, 3 (2015), 211–252.

[35] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. 2016. Improved techniques for training gans. In *Advances in neural information processing systems*. 2234–2242.

[36] Konstantin Shmelkov, Cordelia Schmid, and Karteek Alahari. 2018. How good is my GAN?. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 213–229.

[37] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[38] Liangchen Song, Cheng Wang, Lefei Zhang, Bo Du, Qian Zhang, Chang Huang, and Xinggang Wang. 2020. Unsupervised domain adaptive re-identification: Theory and practice. *PR* 102 (2020), 107173.

[39] Chi Su, Shiliang Zhang, Junliang Xing, Wen Gao, and Qi Tian. 2016. Deep attributes driven multi-camera person re-identification. In *European conference on computer vision*. Springer, 475–491.

[40] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. 2018. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *Proceedings of the European Conference on Computer Vision (ECCV)*. 480–496.

[41] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1–9.

[42] Dongkai Wang and Shiliang Zhang. 2020. Unsupervised Person Re-identification via Multi-label Classification. In *CVPR*. 10981–10990.

[43] Guan'an Wang, Shuo Yang, Huanyu Liu, Zhicheng Wang, Yang Yang, Shuliang Wang, Gang Yu, Erjin Zhou, and Jian Sun. 2020. High-order information matters: Learning relation and topology for occluded person re-identification. In *CVPR*. 6449–6458.

[44] Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi Zhou. 2018. Learning discriminative features with multiple granularities for person re-identification. In *ACMMM*. 274–282.

[45] Jingya Wang, Xiatian Zhu, Shaogang Gong, and Wei Li. 2018. Transferable joint attribute-identity deep learning for unsupervised person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2275–2284.

[46] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. 2018. Person transfer gan to bridge domain gap for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 79–88.

[47] Lin Wu, Yang Wang, Junbin Gao, and Xue Li. 2018. Where-and-when to look: Deep siamese attention networks for video-based person re-identification. *IEEE Transactions on Multimedia* 21, 6 (2018), 1412–1424.

[48] Yunpeng Zhai, Shijian Lu, Qixiang Ye, Xuebo Shan, Jie Chen, Rongrong Ji, and Yonghong Tian. 2020. Ad-cluster: Augmented discriminative clustering for domain adaptive person re-identification. In *CVPR*. 9021–9030.

[49] Li Zhang, Tao Xiang, and Shaogang Gong. 2016. Learning a discriminative null space for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1239–1248.

[50] Fang Zhao, Shengcai Liao, Guo-Sen Xie, Jian Zhao, Kaihao Zhang, and Ling Shao. 2020. Unsupervised domain adaptation with noise resistible mutual-training for person re-identification. In *ECCV*.

[51] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. 2015. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE*

international conference on computer vision. 1116–1124.

[52] Liang Zheng, Yi Yang, and Alexander G Hauptmann. 2016. Person re-identification: Past, present and future. *arXiv preprint arXiv:1610.02984* (2016).

[53] Zhedong Zheng, Xiaodong Yang, Zhiding Yu, Liang Zheng, Yi Yang, and Jan Kautz. 2019. Joint discriminative and generative learning for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2138–2147.

[54] Zhedong Zheng, Liang Zheng, and Yi Yang. 2017. A discriminatively learned cnn embedding for person reidentification. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 14, 1 (2017), 1–20.

[55] Zhedong Zheng, Liang Zheng, and Yi Yang. 2017. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *Proceedings of the IEEE International Conference on Computer Vision*. 3754–3762.

[56] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. 2020. Random erasing data augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 13001–13008.

[57] Zhun Zhong, Liang Zheng, Zhiming Luo, Shaozi Li, and Yi Yang. 2019. Invariance matters: Exemplar memory for domain adaptive person re-identification. In *CVPR*. 598–607.

[58] Zhun Zhong, Liang Zheng, Zhedong Zheng, Shaozi Li, and Yi Yang. 2018. Camera style adaptation for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5157–5166.

[59] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. 2019. Omni-scale feature learning for person re-identification. In *ICCV*. 3702–3712.

[60] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*. 2223–2232.

[61] Kuan Zhu, Haiyun Guo, Zhiwei Liu, Ming Tang, and Jinqiao Wang. 2020. Identity-Guided Human Semantic Parsing for Person Re-Identification. *ECCV* (2020).

[62] Zijie Zhuang, Longhui Wei, Lingxi Xie, Tianyu Zhang, Hengheng Zhang, Haozhe Wu, Haizhou Ai, and Qi Tian. 2020. Rethinking the distribution gap of person re-identification with camera-based batch normalization. In *ECCV*. Springer, 140–157.