UnrealPerson: An Adaptive Pipeline towards Costless Person Re-identification

Tianyu Zhang[1], Lingxi Xie[4], Longhui Wei[5], Zijie Zhuang[4], Yongfei Zhang[1,2,3*], Bo Li[1,2,3], Qi Tian [6]

[1]Beijing Key Laboratory of Digital Media, School of Computer Science and Engineering, Beihang University,

[2]State Key Laboratory of Virtual Reality Technology and Systems, Beihang University,

[3]Pengcheng Laboratory, [4]Tsinghua University, [5]University of Science and Technology of China, [6]Xidian University

{tianyu1949, 198808xc, weilh2568, jayzhuang42, wywqtian}@gmail.com,

{yfzhang, boli}@buaa.edu.cn

Abstract

The main difficulty of person re-identification (ReID) lies in collecting annotated data and transferring the model across different domains. This paper presents ***UnrealPerson****, a novel pipeline that makes full use of unreal image data to decrease the costs in both the training and deployment stages. Its fundamental part is a system that can generate synthesized images of high-quality and from controllable distributions. Instance-level annotation goes with the synthesized data and is almost free. We point out some details in image synthesis that largely impact the data quality. With 3,000 IDs and 120,000 instances, our method achieves a 38.5% rank-1 accuracy when being directly transferred to MSMT17. It almost doubles the former record using synthesized data and even surpasses previous direct transfer records using real data. This offers a good basis for unsupervised domain adaption, where our pre-trained model is easily plugged into the state-of-the-art algorithms towards higher accuracy. In addition, the data distribution can be flexibly adjusted to fit some corner ReID scenarios, which widens the application of our pipeline. We publish our data synthesis toolkit and synthesized data in* https://github.com/FlyHighest/UnrealPerson.

1. Introduction

Person re-identification aims to retrieve the same pedestrian (*i.e.*, an identity) from the images captured by a camera network. As a fundamental task of intelligent surveillance, ReID has attracted increasing attention in the computer vision community. Recently, with the emergence of large-scale ReID datasets [59, 60, 46], effective algorithms have been proposed and achieved satisfying performance in these benchmarks. However, there is still a significant gap in deploying the ReID algorithms to real-world scenarios, arguably because (i) the trained models are often vulnerable to domain changes, yet (ii) annotating identities in new scenarios requires exhausting human labors. We attribute such an application gap to the fact that the current pipeline is hindered by the data limitation and thus not optimized for generalizing across different scenarios.

To alleviate this problem and pave a new path for the community, we propose **UnrealPerson**, a new pipeline that makes full use of unreal (synthesized) image data towards a powerful ReID algorithm that easily and costlessly deploys to a wide range of scenarios. The key observation is that the synthesized pedestrian data sampled from a virtual world comes naturally with free and perfect annotations. From the perspective of the generalization ability, the synthesized data enjoys two-fold benefits. **First**, compared to the manually collected data from restricted real scenarios, the synthesized data from infinite virtual scenes is more diverse, making it less prone to domain-specific patterns. **Second**, the parameters of data synthesis can be freely adjusted to fit the domains in which collecting real data is difficult (*e.g.*, the low-illumination scenario). To fully utilize these characteristics, our entire pipeline consists of ***pre-training the model using abundant synthesized data and then fine-tuning it with off-the-shelf domain adaptation algorithms***. This pipeline has broad applications since it fits both the fully-supervised and unsupervised setting and transfers well to a few corner scenarios.

The quality and richness of our synthesized pedestrian data is the cornerstone of our UnrealPerson pipeline. To synthesize abundant and authentic samples, we first create a set of scenarios (*e.g.*, street, plaza, *etc.*) in the virtual 3D world with changeable environmental parameters (*e.g.*, illumination, lighting, *etc.*). Then, we place an arbitrary num-

This work was partially supported by the National Key R&D Program of China (Grant No.2020AAA0130200), the National Natural Science Foundation of China (No.61772054, 62072022), the NSFC Key Project (No.61632001) and the Fundamental Research Funds for the Central Universities. We thank Dr. Weichao Qiu for instructive discussions.

* Corresponding author

ber of pedestrians with configurable appearance (*e.g.*, gender, height, clothing, *etc.*) into the scenarios, and they move according to the pre-defined paths. Finally, the images are captured by the virtual cameras in the scenes. With this data synthesis system, UnrealPerson is flexible to assemble suitable training data and learning approaches to achieve the best performance for different ReID tasks. We verify its effectiveness through two groups of experiments. **First**, we verify the effectiveness of the synthesized data on improving the generalization ability. We train the ReID model only on the synthesized data and test it directly on conventional ReID benchmarks. Quantitatively, our vanilla baseline [62] achieves a rank-1 accuracy of 38.5% on the MSMT17 dataset, which almost doubles the previous synthesis-based record: 20.0% by RandPerson [45]. **Second**, we adapt our pipeline for specialized ReID scenarios, *e.g.*, all pedestrians are in black, or the illumination is very low. In these tough scenarios, UnrealPerson achieves competitive performance, even surpassing the models that are pre-trained in manually labeled datasets with real-world images. Our major contribution is summarized as follows.

- We propose a novel pipeline that largely reduces the deployment costs of ReID. For the first time, the model pre-trained purely on synthesized data outperforms that pre-trained on real, annotated data.
- We verify the usefulness of our pipeline in a wide range of downstream tasks, including direct transfer, supervised domain adaptation, and unsupervised domain adaptation settings.
- We provide a detailed analysis of the factors in data synthesis, which offers practical guidelines for reusing our toolkit for future research.

2. Related Work

2.1. ReID: Full Supervision and Direct Transfer

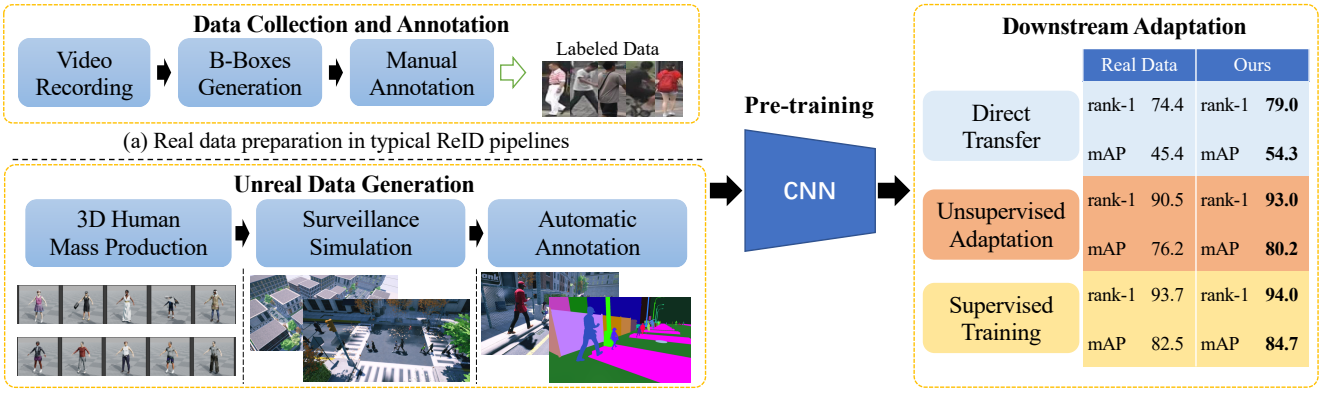
A typical ReID framework for supervised learning requires annotating identities, and hence deep networks can be optimized by learning to separate different persons. The methods can be roughly classified into two categories, *i.e.*, improving feature extraction and designing better objective functions. To make the feature more robust to pose variations, part-based methods [41, 35, 47, 58, 57] are proposed. Some methods target to enhance person-related feature extraction by eliminating the background via semantic parsing [42] or attention mechanism [37, 22]. Also, some works [21, 44, 34] contribute to more effective network architecture for person feature extraction. Moreover, powerful objective functions are introduced to ReID, including triplet loss [17, 53], contrastive loss [43], center loss [28, 48], circle loss [40], and so on. These methods achieve good performance on the same domain evaluation but report unsatisfactory results when directly transferred to other domains. Some methods aim to overcome overfitting on the training domain. For example, Liao *et al.* [23] conduct pairwise matching to find explainable local similar regions. Song *et al.* [38] follow meta-learning pipelines and extract domain-invariant features via sampled sub-domains. Zhuang *et al.* [62] propose to align the feature distribution of all cameras. However, the domain gap limits the performance of these methods.

2.2. Domain Adaptation for Person ReID

Domain adaptation on the target domain usually boosts the performance by shrinking the huge domain gap. Fine-tuning with annotated data can be regarded as a basic supervised domain adaptation method [15]. Further, Xiao *et al.* [49] propose to fine-tune the pre-trained model with domain-guided dropout to filter out useless neurons. On the other hand, unsupervised domain adaptation (UDA) attracts more attention because it requires cheaper unlabeled data of the target domain. The means of UDA include data augmentation [46, 11, 61, 27], distribution alignment [31, 19], predicting pseudo labels [20, 13, 3, 25, 52, 12], spatial-temporal consistency mining [29, 20], model ensemble [54, 56], and so on. The final performance of UDA also relies on the transferability of pre-training data. For example, pseudo label based methods rely on the pre-trained model to provide the initial labels, and the accuracy of pseudo labels directly influences the model convergence and the final performance. Therefore, the quality of pre-training data is the basis for domain adaptation.

2.3. Synthesized ReID Data

Recently, some researchers adopt data synthesis techniques in ReID tasks. SOMAset [2] is proposed to assist deep CNNs training. It contains only 50 persons. SyRI [1] has 100 persons under 140 different lighting conditions. These two datasets are rather small, and the diversity of backgrounds and human appearance is limited. PersonX [39] is a large-scale synthesized dataset, containing 1,266 persons of multiple viewpoints. This dataset aims to explore how viewpoints affect ReID systems. PersonX adopts ready-made human models from a 3D human dataset. RandPerson [45] is a recent dataset proposed to generalize current ReID models. It contains a maximum of 8,000 pedestrians of 19 cameras. By combining RandPerson and other real-world data as the training set, the ReID models achieve better supervised learning results. Although RandPerson is diverse and flexible for its human generation pipeline, it is still weaker than real-world datasets in terms of generalization ability. In what follows, we will discuss how to improve the quality of synthesized data. Our insights make the synthesized data surpass real-world datasets and achieve the best performance on multiple ReID tasks.



(a) Real data preparation in typical ReID pipelines

(b) Unreal data generation for our UnrealPerson pipeline

(c) Downstream adaptation on Market-1501

Figure 1. The UnrealPerson pipeline. The pre-training stage of our pipeline utilizes data synthesis, while conventional pipelines usually need annotated real-world data. We compare the best performance achieved by real data on the Market-1501 dataset. On all three tasks, our unreal data surpasses real data.

3. The UnrealPerson Pipeline

3.1. Problem: Person Re-identification

Given an annotated image dataset $\mathcal{S} = \{(\boldsymbol{I}_1, \boldsymbol{y}_1), (\boldsymbol{I}_2, \boldsymbol{y}_2), ..., (\boldsymbol{I}_N, \boldsymbol{y}_N)\}$, where each $\boldsymbol{I}_i$ represents the image, and $\boldsymbol{y}_i$ is the ground truth label of the identity, the goal of ReID is to learn a proper feature embedding function $\boldsymbol{f}(\theta; \boldsymbol{I}_i)$ that maps images into a feature space $\mathcal{X} = \{\boldsymbol{x}_i | \boldsymbol{x}_i = \boldsymbol{f}(\theta; \boldsymbol{I}_i), 1 \leq i \leq N\}$, where the distances of the same identity are smaller than those of different identities. A straight-forward way to achieve this is to minimize the identity prediction error on $\mathcal{S}$:

$$\min \mathbb{E}_{(\boldsymbol{I}_i, \boldsymbol{y}_i) \in \mathcal{S}} [\boldsymbol{y}_i - \boldsymbol{g}(\boldsymbol{f}(\boldsymbol{I}_i)], \tag{1}$$

where $\boldsymbol{g}$ is the classifier. In this formulation, the quality of the learnt mapping relies on the data distribution of $\mathcal{S}$. The data distribution can be disassembled into two parts, the identity appearance distribution and background distribution. Hence, we denote the data distribution of $\mathcal{S}$ as $\mathcal{D}^{\mathcal{S}}(\mathrm{FG}, \mathrm{BG})$, where FG and BG represent foreground and background, respectively.

In the above formulation, two drawbacks are revealed. **First**, a major difficulty lies in collecting and annotating the training dataset $\mathcal{S}$. A large-scale dataset often takes much time and labor to construct. For instance, MSMT17 is a large-scale ReID dataset, consisting of 4,101 identities captured from 15 cameras. Researchers collected 180 hours of high-resolution videos and three labelers annotated for two months. Such costs are unbearable in application scenarios. **Second**, the data distribution $\mathcal{D}^{\mathcal{S}}$ is easily interfered with by foreground and background changes. Thus, ReID data collected in one scene often fails to transfer well to other scenes. For example, a ReID model trained on Market-1501 reports 91.4% rank-1 accuracy when evaluating on Market-1501 testing set, where cameras are the same as the training set, but only obtains 25.7% rank-1 accuracy on MSMT17. This dramatic accuracy drop implies the huge domain gap and also reveals the weakness of the current ReID pipeline.

3.2. Towards a Generalized and Costless Pipeline

In the current ReID pipeline that involves data annotation from real scenes (as shown in Fig. 1(a)), the two drawbacks mentioned above are inevitable. Usually, the costs of annotating a lot of cross-camera pedestrians are unbearable in applications. To say the least, even if we do not consider the annotation costs, there is still a dilemma in real data preparation. On the one hand, to offer sufficient coverage of different scenarios, the dataset should contain a large amount of labeled data. On the other hand, a large amount of data may scatter data distribution, leading to an unsatisfying performance in some corner scenarios. We owe such a dilemma to the real data lacking flexibility and turn to generating unreal data for training stronger ReID systems.

Unlike real data, synthesized data enjoys the benefits of free annotation and flexible data distribution. Based on the toolkit of synthesizing data, we can easily pre-train a ReID model on an arbitrary distribution and, if needed, fine-tune it to adjust various downstream tasks. As shown in Fig. 1, our proposed pipeline involves three components: unreal data generation, model pre-training, and downstream adaptation, of which the unreal data is the fundamental part of the whole pipeline.

Our UnrealPerson pipeline liberates ReID systems from the manual annotation. In terms of time cost for data preparation, we are able to construct a labeled unreal dataset of 3,000 identities within 48 CPU-hours, in comparison to the real dataset MSMT17 of 4,101 identities, which took 180 person-days. This unreal dataset also surpasses MSMT17 on several downstream adaptation tasks, as shown in Fig. 1(c). Moreover, our pipeline is flexible

Datasets	#Identities	#Cameras	#BBoxes	Clothing	Accessories	Hard samples	Surveillance Simulation	Scalable	Rank-1 on MSMT17 (%)
SOMAset [2]	50	-	100,000	Real	No	No	No	No	3.1
SyRI [1]	100	-	56,000	Real	No	No	No	No	21.8
PersonX [39]	1,266	6	273,456	Real	No	No	No	Yes	22.2
RandPerson [45]	8,000	19	228,655	Generated+Real	No	No	Yes	Yes	20.0
Our synthesized data	3,000	34	120,000	Real	Various	Many	Yes	Yes	**38.5**

Table 1. Detailed comparisons of synthesized datasets. Note that SOMAset and SyRI do not have a camera network, so numbers of cameras are left blank. The rank-1 accuracy on MSMT17 is the direct transfer performance of ReID models trained on these synthesized datasets.

compared to the typical ReID pipeline because the data distribution of synthesized data is fully controllable. The flexibility of data synthesis empowers our pipeline to transfer well to many corner scenarios, like night-time ReID, indoor ReID, and so on, where annotated real data is hard to collect. With our pipeline, we can easily fit these scenarios using unreal data.

In this paper, we mainly focus on dataset preparation and show that with a better-annotated dataset, the demands for pre-training methods and fine-tuning algorithms will be lowered greatly. The details of our synthesized data will be discussed in the next subsection.

3.3. Data Synthesis: The Devil Lies in Details

We use synthesized data for our new ReID pipeline. A data synthesis system is developed to generate costless and flexible ReID image data. There are four major differences between our data synthesis system and others. **(i) More realistic.** In foregrounds, we use real clothing images on generated 3D humans that comply with biological structures; in backgrounds, we mimic real surveillance systems in high-quality virtual environments. Compared to PersonX or RandPerson that use low-poly assets in their generation systems, our assets are more realistic. **(ii) More details.** We add more details to 3D human models. A total of 248 types of clothes are used in our generated 3D humans. In addition, our 3D humans randomly carry accessories, including masks, glasses, hats, earphones, scarves, bags, and backpacks. These things are commonly seen in real scenes but hardly used in previous synthesis systems. **(iii) Hard samples.** Apart from increasing diversity, we also consider the difficulty of the training set. Adding more details inevitably makes humans easy to recognize. Therefore, we deliberately add pedestrians with similar appearances in our synthesized data as hard samples. Persons that look quite similar but differs in small discriminative regions play an important role in guiding ReID models to focus on local areas. **(iv) Scalability.** Different from SyRI and PersonX, our synthesis system supports arbitrary numbers of pedestrians and cameras because we have a 3D human production tool and a universal program that fits almost all virtual scenes in Unreal Engine 4. The summarized comparisons are shown in Tab. 1. We will validate the advantages mentioned above in Sec. 4.2. In the rest of this subsection, we briefly introduce the steps to generate the synthesized ReID data.

3D Human Mass Production. The 3D human models are produced in MakeHuman [9], an open-source program that generates realistic human models. We overwrite a community plugin, *massproduce* [8], to generate a large number of models in one click. To increase the diversity of human models, we use real-world clothing images of two datasets, Clothing-co-parsing [51] and DeepFashion [26], as the clothing texture images for the generated humans.

Surveillance Simulation. We simulate real surveillance scenes in Unreal Engine 4 (UE4) [18], a mature platform for high-quality video games and VR applications. The resources for UE4 are sufficient to generate various ReID datasets. We choose 4 scenes from UE4 marketplace, namely Scene 1, ..., Scene 4, among which three are outdoor city environments, and one is an indoor scene. For our virtual humans, we provide 4 walking animations and 2 idling animations. They are given pre-defined paths to walk along in the unreal scenes. Humans may occlude each other, just like in real-world scenes. The occlusion level can also be controlled. If needed, more serious occluded scenarios can be achieved by putting more obstacles or increasing pedestrians with different walking speeds. For cameras, we set virtual cameras in the unreal environments of different viewpoints and different distances to humans. Multiple views of pedestrians can be obtained. We also make skylight change during data collection.

Data Annotation. Data annotation is conducted automatically by our annotating scripts. We adopt UnrealCV [32, 33] to collect pixel-level instance segmentation annotations for every image the virtual cameras capture. Then, we crop every pedestrian in the images after filtering out small bounding boxes on the edge and discarding seriously occluded pedestrians. To simulate detection bounding boxes or manually cropped boxes, the bounding boxes are randomly enlarged by a factor of 0.1.

Summary. Our synthesized dataset sets up a better platform for downstream tasks, and even the pre-trained model itself achieves good results on many datasets. The UnrealPerson pipeline lowers the needs for annotated real data and boosts transferring performance across different domains, which will be presented in the next section. More technical details about data synthesis and visualizations of our synthesized data are shown in supplementary materials.

#IDs	#Cameras	Clothing Textures			Accessories	Hard Samples	Market		Duke		MSMT17	
		Random	Generated	Real			rank-1	mAP	rank-1	mAP	rank-1	mAP
800	6	✓					52.0	26.2	41.4	22.4	18.5	6.1
			✓				61.0	34.4	49.8	29.8	19.6	6.6
				✓			64.5	37.9	54.3	33.8	20.7	6.9
				✓	✓		65.3	38.1	57.0	36.3	21.6	7.4
				✓	✓	✓	65.2	38.8	56.7	36.6	21.9	7.6
800	16			✓	✓	✓	69.9	42.5	61.0	38.4	26.3	9.0
	22			✓	✓	✓	71.1	43.9	61.9	41.1	30.3	10.7
	28			✓	✓	✓	73.7	46.5	62.9	41.6	33.4	12.6
	34			✓	✓	✓	74.9	48.2	64.9	42.9	35.4	13.3
1,500	34			✓	✓	✓	75.7	50.7	67.3	46.6	36.3	13.9
2,000				✓	✓	✓	76.8	52.0	69.0	48.0	37.9	14.7
3,000				✓	✓	✓	**79.0**	**54.3**	**69.7**	**49.4**	**38.5**	**15.3**

Table 2. Direct transfer performance of Unreal to three real datasets. We control several parameters in our synthesized data, *i.e.*, clothing textures, accessories, hard samples and numbers of identities and cameras, towards better performance.

4. Experiments

4.1. Implementation Details

We adopt ResNet-50 [16], which is pre-trained on ImageNet [10], as our backbone for all experiments. We also replace all batch normalization layers with camera-based batch normalization (CBN) [62] layers in the network. An extra CBN layer is added after global average pooling on the last residual block of ResNet-50, followed by a linear layer as the classifier. To train with labeled data, the softmax cross-entropy loss is used. For unlabeled data, we use the joint visual and temporal consistency (JVTC) [20] framework for unsupervised domain adaptation (UDA). The images are resized to 256×128. For direct transfer experiments, we set batch size as 64; for UDA experiments, we sample 128 images from the source domain and 128 images from the target domain to form a mini-batch. In each mini-batch, for annotated data, we adopt a balanced sampling strategy proposed in [55]. For unlabeled data, we randomly sample images. We adopt the SGD optimizer for training, with a momentum of 0.9 and weight decay of 5×10^{-4}. The initial learning rate is 0.01 and decays to 0.001 after 40 epochs. The training stops at the $60th$ epoch. Besides, as synthesized datasets contain more images, we observe convergence with fewer iterations. Therefore, for synthesized datasets, in direct transfer experiments, we decay the learning rate at the $10th$ epoch and stop training at the $15th$ epoch; in UDA experiments, we only sample 300 mini-batches from the source domain in each epoch.

4.2. Direct Transfer Evaluation

The synthesized ReID data is the fundamental part of the UnrealPerson pipeline. Here, we adopt direct transfer performance on real datasets as the indicator to show the quality of synthesized data because direct transfer is the basis for all other tasks. Three real datasets, Market-1501 [59], DukeMTMC-reID [60], and MSMT17 [46], are used as the testing sets. For short, we refer to Market-1501 and DukeMTMC-reID as Market and Duke, respectively. For our synthesized data, we refer to as Unreal. We sample 40 images for each pedestrian to form our Unreal dataset. We first explore how to synthesize high-quality virtual humans and then scale up the number of cameras and identities. A summarized report is presented in Tab. 2.

(a) Generated textures (b) Real textures (c) Accessories

(d) Two groups of similar pedestrians.

Figure 2. Visualizations of our synthesized data. (a) Regular generated clothing textures. (b) Real textures. (c) 3D humans with accessories (handbag, backpack, and umbrella). (d) Pedestrians with similar appearance.

Clothing Textures. The appearance of clothes is a main part of the foreground and provides much discriminative information. In synthesized data, we randomly replace the clothing textures to enrich clothing appearance. We compare three different ways of enriching clothing textures. Random images are from universal image datasets, like ImageNet [10] or COCO [24]. Generated color textures (Fig. 2(a)) are proposed in RandPerson [45], with a few predefined patterns applied to generated color palettes. Real textures are cropped clothing image patches from clothing segmentation datasets. From the top 3 lines in Tab. 2, we can conclude that the textures of real clothes are more suitable for enriching 3D human models. Random images or

Figure 3. Top-1 results on MSMT17. The ReID models are trained using our unreal data with or without hard samples, *i.e.*, similar pedestrians. Images in the same column are interrelated. Red: false matches; Green: correct matches.

generated patterns may vary from real clothes largely and decay the reality of unreal data.

Accessories. The accessories of pedestrians are important clues to recognize their identities. Adding accessories to 3D humans makes the synthesized data more realistic. As shown in Fig. 2(c), our data synthesis system supports various accessories on human models. From Tab. 2, we can see that the synthesized datasets with accessories achieve better performance on all three testing sets.

Hard Samples. In real-world scenarios, some persons may look quite similar. They are *hard samples* for ReID algorithms and are important for efficient learning. As shown in Fig. 2(d), we also synthesize hard samples in unreal data. Specifically, we generate human models by group. The five human models in one group share similar appearances, but everyone is a little different from the others. Note that hard samples are better provided with a rather larger number of identities and cameras because in a small dataset, hard samples may heavily interfere learning from normal cases. As shown in Tab. 2, when we have 800 identities in 6 cameras, the hard samples slightly improve mAP on the three testing sets. From Fig. 4, with more cameras and more identities, adding hard samples significantly improves the rank-1 accuracy. We also show some query results on MSMT17 in Fig. 3. The ReID model trained with synthesized data without hard samples tends to ignore the obvious differences in backpacks and handbags. After adding hard samples, we empower ReID methods to focus on local regions.

Number of cameras. We prepare 4 scenes for the 3D human models to walk around, where 34 virtual cameras are deployed to capture images. We construct several synthesized datasets containing 800 humans with the number of cameras increasing from 6 to 34 and involving scenes from 1 to 4. The direct transfer evaluation results are shown in the middle part of Tab. 2. Virtual environments differ from each other in many aspects, such as illumination, styles of buildings and roads, and crowdedness. By recognizing persons

Training set	Market		Duke		MSMT	
	rank-1	mAP	rank-1	mAP	rank-1	mAP
Market	*91.4*	*76.8*	56.7	36.5	25.7	9.6
Duke	72.4	41.9	*82.1*	*67.5*	35.8	13.1
MSMT17	74.4	45.4	67.1	46.8	*72.5*	*42.4*
SyRI	48.5	22.6	38.9	18.2	21.8	5.7
PersonX	58.7	32.7	49.4	28.9	22.2	7.9
RandPerson	64.7	39.3	59.4	38.4	20.0	6.8
Unreal	**79.0**	**54.3**	**69.7**	**49.4**	**38.5**	**15.3**

Table 3. Direct transfer results with the CBN method [62]. The fully-supervised learning results are in *italics*.

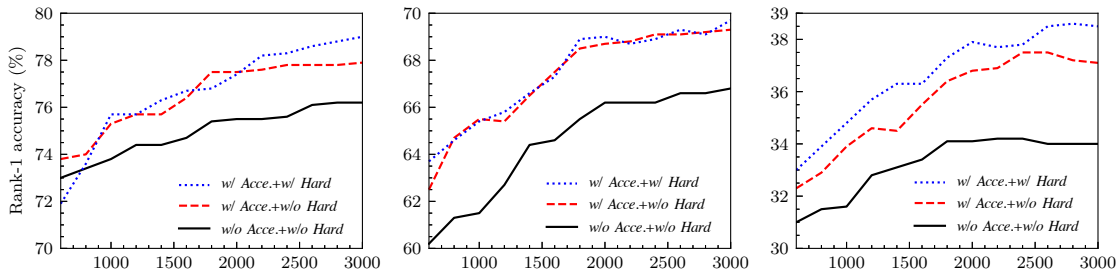
Training set	Market		Duke		MSMT	
	rank-1	mAP	rank-1	mAP	rank-1	mAP
Market	*94.6*	*86.3*	29.0	15.6	9.2	3.0
Duke	49.7	23.7	*86.5*	*76.5*	14.3	4.5
MSMT	56.2	29.8	51.8	34.3	*73.9*	*49.9*
SyRI	15.2	4.9	16.6	5.9	7.8	1.6
PersonX	27.6	11.0	15.9	7.1	4.1	1.2
RandPerson	53.4	27.9	44.0	26.4	14.0	4.8
Unreal	**64.0**	**37.2**	**58.0**	**37.5**	**26.8**	**9.9**

Table 4. Direct transfer results (the BN variant of Tab. 3). The baseline is BoT [28], and the fully-supervised results are in *italics*.

across virtual environments, ReID models are more robust to person-unrelated variations of different scenes.

Number of pedestrians. Apart from the results shown in the bottom part of Tab. 2, we also conduct fine-grained experiments to explore how many pedestrians are suitable for training models. The accessories and hard samples are also validated in these experiments. From Fig. 4, we can see that generally, more pedestrians lead to better performance. The results also demonstrate the effectiveness of our proposed additional components for foreground synthesis, *i.e.*, accessories and hard samples. Note that we achieve the best direct transfer performance using 3,000 identities, much fewer than RandPerson. We also observe that, when adding more pedestrians (larger than 3,000), the performance is hardly promoted. This issue will be discussed in Sec. 5.

Summary. In the above analysis, we explore the key factors of how to improve synthesized ReID datasets. In foreground synthesis, we generate 3D human models with accessories to introduce more diversity and produce similar humans as hard samples. In background synthesis, we validate the importance of cross-scene pedestrians. Based on these conclusions, we take a representative synthesized dataset that exploits all the advantages we find. This dataset contains 120,000 images of 3,000 pedestrians collected from 34 cameras deployed in 4 different virtual scenes. For convenience, we refer to this dataset as Unreal. The direct transfer performance of Unreal and other datasets is compared in Tab. 3 and Tab. 4. On both the CBN baseline [62] and BN baseline [28], our Unreal dataset surpasses other datasets in direct transfer evaluation.



(a) Evaluation on Market (b) Evaluation on Duke (c) Evaluation on MSMT17

Figure 4. Results of direct transfer evaluation on Market, Duke and MSMT17, with the number of pedestrians in synthesized datasets increasing from 600 to 3,000. The number of cameras in all the experiments shown in this figure is 34.

Pre-training Dataset	Fine-tuning Dataset	Market		Duke		MSMT	
		rank-1	mAP	rank-1	mAP	rank-1	mAP
ImageNet	Market	*91.4*	*76.8*	56.7	36.5	25.7	9.6
MSMT17		*93.7*	*82.5*	68.5	50.1	47.3	21.4
Unreal		***94.0***	***84.7***	72.8	54.7	35.6	14.8
ImageNet	Duke	72.4	41.9	*82.1*	*67.5*	35.8	13.1
MSMT17		76.5	47.6	*85.7*	*72.1*	48.5	20.6
Unreal		82.5	57.9	***86.8***	***74.2***	40.5	16.3
ImageNet	MSMT17	74.4	45.4	67.1	46.8	*72.5*	*42.4*
Duke+Market		80.0	53.6	70.5	52.6	*73.7*	*44.7*
Unreal		80.1	53.8	71.5	52.8	***74.5***	***46.0***

Table 5. Results of supervised fine-tuning on the pre-trained models. The fully-supervised learning results are in *italics*. Direct transfer performance is also shown in this table conveniently.

Source Domain	Methods	Market		Duke		MSMT17	
		rank-1	mAP	rank-1	mAP	rank-1	mAP
Market	JVTC	-	-	76.5	59.6	46.1	20.4
Duke		89.0	73.1	-	-	52.5	23.5
MSMT17		89.9	74.5	79.0	63.5	-	-
Unreal		**90.8**	**78.3**	**81.2**	**66.1**	**53.7**	**25.0**
Market	JVTC+	-	-	84.6	68.8	63.7	30.1
Duke		89.3	74.6	-	-	66.8	32.5
MSMT17		90.5	76.2	85.2	72.1	-	-
Unreal		**93.0**	**80.2**	**88.3**	**75.2**	**68.2**	**34.8**

Table 6. Unsupervised domain adaptation performance on three benchmark datasets.

4.3. Domain Adaptation

Supervised Fine-tuning. When abundant labeled data in the target domain is accessible, domain adaptation can be simply done by supervised fine-tuning on the pre-trained model. From Tab. 5, we see that the results of fully-supervised learning are promoted by using Unreal as the pre-training data. For example, when training and testing on Market with ImageNet pre-trained model, the rank-1 accuracy is 91.4%, while our Unreal pre-trained model reaches 94.0%. Our Unreal dataset also surpasses other real-world ReID datasets, showing its universality and transferability.

Unsupervised Domain Adaptation. Unsupervised domain adaptation (UDA) is a popular direction to leverage unlabeled data from the target domain. The training sets include labeled data of the source domain and unlabeled data from the target domain. In UDA experiments, we implement JVTC [20] as the off-the-shelf algorithm in the network with CBN [62] layers. JVTC+ denotes using joint similarity of visual and spatial-temporal features in the testing stage. The results are shown in Tab. 6. With the assistance of our unreal data, the UDA performance is largely boosted. On Duke, JVTC+ further promotes the rank-1 accuracy to 88.3%, not only setting up a state-of-the-art record but also surpassing the fully-supervised learning results shown in Tab. 5. Note that these results are obtained without any manual annotation. It demonstrates the application values of our UnrealPerson pipeline.

4.4. Task-specific Adaptation for Corner Scenarios

Our UnrealPerson pipeline enjoys the benefits of flexible data synthesis. The distribution of synthesized data can be adaptively adjusted by modifying parameters of synthesis, and the process can be done easily in our UnrealPerson pipeline. This advantage makes our pipeline more suitable for corner scenarios of ReID, where labeled data is hard to obtain. We present three examples, *i.e.*, indoor ReID, low illumination ReID and black ReID. Previously little attention has been paid to these scenarios for the lack of abundant training data. For evaluation on indoor ReID and low illumination ReID tasks, we use GRID [7] and LIPS [30] dataset, respectively. GRID is a dataset collected in an underground station, where the cameras are located at high angles of view. LIPS is constructed with two night cameras, and thus the illumination is extremely low. We take 250 persons in GRID and 50 persons in LIPS for testing. Black ReID problem was first defined in [50]. In some scenes, most pedestrians wear similar clothes,*e.g.*, many wear dark clothes in winter. We prepare two datasets, Market-black and Duke-black, according to the annotations provided by the Black-reID dataset [50]. All persons in Duke-black and Market-black wear black clothes. Duke-black contains 1,216 images of 145 persons, among which 436 images are used as queries and the rest as gallery images. Market-black contains 41 persons. It has 174 images as queries and 251 images as the gallery.

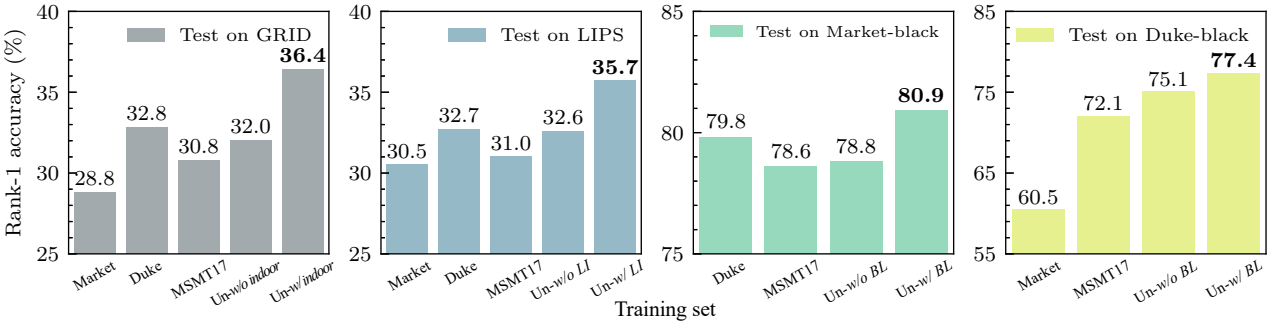


Figure 5. Direct transfer performance on various corner scenarios. *Un* denotes the *unreal* data are used for training.

In Fig. 5, we show the direct transfer performance on these datasets. The compared training sets include Market, Duke, and MSMT17. For the indoor dataset GRID, we use the Unreal dataset with 6 extra indoor cameras, achieving 36.4% rank-1 accuracy. On the low illumination dataset, we adjust our virtual scenes to night time, and train the network with low illumination (LI) unreal data, surpassing real data by 3.0%. For black ReID, we apply dark clothing textures to 3D humans to construct our Unreal-*w/ BL*. This dataset gets significant promotions on the two black datasets.

5. Open Problems

Our research leaves a few open problems for the community. We summarize them as follows, and hence suggest the community pays more attention to this new pipeline that is potentially the future trend of ReID.

- The **quality** of synthesized data. It is easy to recognize the synthesized images from the real images, because the realness of 3D human models, the richness of facial expressions and contexts, the illumination changes, *etc.*, are still far from perfect. We guess that there is a saturation point for synthesized data (beyond it, continuing improving reality brings marginal gains), but we are not sure when it will be reached and whether the domain transfer methods can relieve the burden of image synthesis. Within a short period, we believe that continuing mimicking the real-world data property can bring us non-trivial benefits.
- The **quantity** of synthesized identities. Currently, our algorithm reaches a plateau in direct transfer at 3,000 pedestrians. Surprisingly, this number is even smaller than MSMT17, the real-world ReID dataset. There is no doubt about the potential of introducing more data, but there are problems to solve, including the quality issue (described above) and the data distribution issue (*e.g.*, the number of identities to be placed in one scenario, the function that samples the details for each identity, *etc.*). Clues may be found by analyzing some meta-information (*e.g.*, the distribution of identity similarity) and compare it to real-world datasets.
- The **efficiency** of learning from infinite data. Prior works [36, 4] have shown that active learning or hard example mining are potentially more efficient strategies when there is an infinite amount of data. We hope to validate these techniques in our pipeline and thus decrease the complexity, in particular when the data quantity becomes much larger.
- The **relationship** with other problems. The flexibility of data synthesis allows us to augment the scope of ReID, or investigate the relationship between ReID and other vision problems. To name a few, (i) one can generalize ReID from image-based to video-based, where our pipeline enjoys a larger advantage over the public benchmarks; (ii) one can generate high-quality segmentation mask for other objects in the virtual scenarios, allowing the researchers to consider the contexts for more accurate ReID; (iii) one can also study the self-supervised learning methods [14, 5, 6], which often require more data and are believed stronger in domain transfer.

6. Conclusions

This paper presents **UnrealPerson**, a novel pipeline for person re-identification (ReID). It aims to relieve the burden of costly data annotation and alleviate the difficulty of domain transfer. Synthesized data plays an important role in this research. We reveal that there is still much room of improvement by synthesizing data from more virtual scenarios/cameras as with richer details. With our pre-trained ReID model, the direct transfer accuracy to MSMT17, the largest publicly available dataset, is almost doubled compared to the previous best pipeline that uses synthesized training data. UnrealPerson enjoys stronger transferability to real-world ReID datasets because (i) the pre-trained model is specialized in and better at processing ReID data, and (ii) the synthesized environment can be flexibly adjusted to the corner scenarios in which collecting real-world data is difficult.

References

[1] Slawomir Bak, Peter Carr, and Jean-Francois Lalonde. Domain adaptation through synthesis for unsupervised person re-identification. In *ECCV*. 2, 4

[2] Igor Barros Barbosa, Marco Cristani, Barbara Caputo, Aleksander Rognhaugen, and Theoharis Theoharis. Looking beyond appearances: Synthetic training data for deep cnns in re-identification. *Computer Vision and Image Understanding*, 2017. 2, 4

[3] Guangyi Chen, Yuhao Lu, Jiwen Lu, and Jie Zhou. Deep credible metric learning for unsupervised domain adaptation person re-identification. In *ECCV*, 2020. 2

[4] Qi Chen, Weichao Qiu, Yi Zhang, Lingxi Xie, and Alan Yuille. Sampleahead: Online classifier-sampler communication for learning from synthesized data. 2018. 8

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020. 8

[6] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. Big self-supervised models are strong semi-supervised learners. In *NeurIPS*, 2020. 8

[7] Chen Change Loy, T. Xiang, and S. Gong. Multi-camera activity correlation analysis. In *CVPR*, 2009. 7

[8] Makehuman community. Massproduce plugin for makehuman, 2017. https://github.com/makehumancommunity/community-plugins-massproduce. 4

[9] Makehuman community. Makehuman, 2020. http://www.makehumancommunity.org. 4

[10] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*, 2009. 5

[11] Weijian Deng, Liang Zheng, Guoliang Kang, Yi Yang, Qixiang Ye, and Jianbin Jiao. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person reidentification. In *CVPR*, 2018. 2

[12] Yixiao Ge, Dapeng Chen, and Hongsheng Li. Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification. In *ICLR*, 2020. 2

[13] Yixiao Ge, Feng Zhu, Dapeng Chen, Rui Zhao, and Hongsheng Li. Self-paced contrastive learning with hybrid memory for domain adaptive object re-id. In *NeurIPS*, 2020. 2

[14] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020. 8

[15] Kaiming He, Ross Girshick, and Piotr Dollar. Rethinking imagenet pre-training. In *ICCV*, 2019. 2

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 5

[17] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017. 2

[18] Epic Games Incorporated. Unreal engine, 2020. https://www.unrealengine.com. 4

[19] Xin Jin, Cuiling Lan, Wenjun Zeng, and Zhibo Chen. Global distance-distributions separation for unsupervised person re-identification. In *ECCV*, 2020. 2

[20] Jianing Li and Shiliang Zhang. Joint visual and temporal consistency for unsupervised domain adaptive person re-identification. In *ECCV*, 2020. 2, 5, 7

[21] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*, 2014. 2

[22] Wei Li, Xiatian Zhu, and Shaogang Gong. Harmonious attention network for person re-identification. In *CVPR*, 2018. 2

[23] Shengcai Liao and Ling Shao. Interpretable and Generalizable Person Re-Identification with Query-Adaptive Convolution and Temporal Lifting. In *European Conference on Computer Vision (ECCV)*, 2020. 2

[24] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *ECCV*, 2014. 5

[25] Yutian Lin, Xuanyi Dong, Liang Zheng, Yan Yan, and Yi Yang. A bottom-up clustering approach to unsupervised person re-identification. In *AAAI*, 2019. 2

[26] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *CVPR*, 2016. 4

[27] Chuanchen Luo, Chunfeng Song, and Zhaoxiang Zhang. Generalizing person re-identification by camera-aware invariance learning and cross-domain mixup. In *ECCV*, 2020. 2

[28] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In *CVPR Workshops*, 2019. 2, 6

[29] Jianming Lv, Weihang Chen, Qing Li, and Can Yang. Unsupervised cross-dataset person re-identification by transfer learning of spatial-temporal patterns. In *CVPR*, 2018. 2

[30] Fei Ma, Xiaoke Zhu, Xinyu Zhang, Liang Yang, Mei Zuo, and Xiao-Yuan Jing. Low illumination person re-identification. *Multimedia Tools and Applications*, 78(1):337–362, 2019. 7

[31] Djebril Mekhazni, Amran Bhuiyan, George Ekladious, and Eric Granger. Unsupervised domain adaptation in the dissimilarity space for person re-identification. In *ECCV*, 2020. 2

[32] Weichao Qiu and Alan Yuille. Unrealcv: Connecting computer vision to unreal engine. In *ECCV*, 2016. 4

[33] Weichao Qiu, Fangwei Zhong, Yi Zhang, Siyuan Qiao, Zihao Xiao, Tae Soo Kim, and Yizhou Wang. Unrealcv: Virtual worlds for computer vision. In *ACMMM*, 2017. 4

[34] Ruijie Quan, Xuanyi Dong, Yu Wu, Linchao Zhu, and Yi Yang. Auto-reid: Searching for a part-aware convnet for person re-identification. In *ICCV*, 2019. 2

[35] M Saquib Sarfraz, Arne Schumann, Andreas Eberle, and Rainer Stiefelhagen. A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking. In *CVPR*, 2018. 2

[36] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *ICLR*, 2018. 8

[37] Chunfeng Song, Yan Huang, Wanli Ouyang, and Liang Wang. Mask-guided contrastive attention model for person re-identification. In *CVPR*, 2018. 2

[38] Jifei Song, Yongxin Yang, Yi-Zhe Song, Tao Xiang, and Timothy M Hospedales. Generalizable person re-identification by domain-invariant mapping network. In *CVPR*, 2019. 2

[39] Xiaoxiao Sun and Liang Zheng. Dissecting person re-identification from the viewpoint of viewpoint. In *CVPR*. 2, 4

[40] Yifan Sun, Changmao Cheng, Yuhan Zhang, Chi Zhang, Liang Zheng, Zhongdao Wang, and Yichen Wei. Circle loss: A unified perspective of pair similarity optimization. In *CVPR*, 2020. 2

[41] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *ECCV*, 2018. 2

[42] Maoqing Tian, Shuai Yi, Hongsheng Li, Shihua Li, Xuesen Zhang, Jianping Shi, Junjie Yan, and Xiaogang Wang. Eliminating background-bias for robust person re-identification. In *CVPR*, June 2018. 2

[43] Rahul Rama Varior, Mrinal Haloi, and Gang Wang. Gated siamese convolutional neural network architecture for human re-identification. In *ECCV*, 2016. 2

[44] Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi Zhou. Learning discriminative features with multiple granularities for person re-identification. In *ACMMM*, 2018. 2

[45] Yanan Wang, Shengcai Liao, and Ling Shao. Surpassing real-world source training data: Random 3d characters for generalizable person re-identification. In *ACMMM*. 2, 4, 5

[46] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *CVPR*, 2018. 1, 2, 5

[47] Longhui Wei, Shiliang Zhang, Hantao Yao, Wen Gao, and Qi Tian. Glad: global-local-alignment descriptor for pedestrian retrieval. In *ACMMM*, 2017. 2

[48] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *ECCV*. Springer, 2016. 2

[49] Tong Xiao, Hongsheng Li, Wanli Ouyang, and Xiaogang Wang. Learning deep feature representations with domain guided dropout for person re-identification. In *CVPR*, 2016. 2

[50] Boqiang Xu, Lingxiao He, Xingyu Liao, Wu Liu, Zhenan Sun, and Tao Mei. Black re-id: A head-shoulder descriptor for the challenging problem of person re-identification. In *ACMMM*, 2020. 7

[51] Wei Yang, Ping Luo, and Liang Lin. Clothing co-parsing by joint image segmentation and labeling. In *CVPR*, 2013. 4

[52] Hong-Xing Yu, Wei-Shi Zheng, Ancong Wu, Xiaowei Guo, Shaogang Gong, and Jian-Huang Lai. Unsupervised person re-identification by soft multilabel learning. In *CVPR*, 2019. 2

[53] Rui Yu, Zhiyong Dou, Song Bai, Zhaoxiang Zhang, Yongchao Xu, and Xiang Bai. Hard-aware point-to-set deep metric for person re-identification. In *ECCV*, September 2018. 2

[54] Yunpeng Zhai, Qixiang Ye, Shijian Lu, Mengxi Jia, Rongrong Ji, and Yonghong Tian. Multiple expert brainstorming for domain adaptive person re-identification. In *ECCV*, 2020. 2

[55] Tianyu Zhang, Lingxi Xie, Longhui Wei, Yongfei Zhang, Bo Li, and Qi Tian. Single camera training for person re-identification. In *AAAI*, 2020. 5

[56] Fang Zhao, Shengcai Liao, Guo-Sen Xie, Jian Zhao, Kaihao Zhang, and Ling Shao. Unsupervised domain adaptation with noise resistible mutual-training for person re-identification. In *ECCV*, 2020. 2

[57] Haiyu Zhao, Maoqing Tian, Shuyang Sun, Jing Shao, Junjie Yan, Shuai Yi, Xiaogang Wang, and Xiaoou Tang. Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In *CVPR*, 2017. 2

[58] Liming Zhao, Xi Li, Yueting Zhuang, and Jingdong Wang. Deeply-learned part-aligned representations for person re-identification. In *ICCV*, 2017. 2

[59] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *ICCV*, 2015. 1, 5

[60] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *ICCV*, 2017. 1, 5

[61] Zhun Zhong, Liang Zheng, Zhiming Luo, Shaozi Li, and Yi Yang. Invariance matters: Exemplar memory for domain adaptive person re-identification. In *CVPR*, 2019. 2

[62] Zijie Zhuang, Longhui Wei, Lingxi Xie, Tianyu Zhang, Hengheng Zhang, Haozhe Wu, Haizhou Ai, and Qi Tian. Rethinking the distribution gap of person re-identification with camera-based batch normalization. In *ECCV*. 2, 5, 6, 7