Less is More: Learning from Synthetic Data with Fine-grained Attributes for Person Re-Identification

Suncheng Xiang\textsuperscript{1}, Guanjie You\textsuperscript{2}, Mengyuan Guan\textsuperscript{1}, Hao Chen\textsuperscript{1}, Binjie Yan\textsuperscript{1}, Ting Liu\textsuperscript{1}, Yuzhuo Fu\textsuperscript{1}
\textsuperscript{1}Shanghai Jiao Tong University, \textsuperscript{2}National University of Defense Technology
\{xiangsuncheng17, gemini.my, 958577057, yanbinjie, louisa.liu, yzfu\}@sjtu.edu.cn, ygjsxxz@163.com

Abstract

Person re-identification (re-ID) plays an important role in applications such as public security and video surveillance. Recently, learning from synthetic data, which benefits from the popularity of synthetic data engine, has attracted great attention from the public eyes. However, existing datasets are limited in quantity, diversity and reality, and cannot be efficiently used for re-ID problem. To address this challenge, we manually construct a large-scale person dataset named FineGPR with fine-grained attribute annotations. Moreover, aiming to fully exploit the potential of FineGPR and promote the efficient training from millions of synthetic data, we propose an attribute analysis pipeline called AOST, which dynamically learns attribute distribution in real domain, then eliminates the gap between synthetic and real-world data and thus is freely deployed to new scenarios. Experiments conducted on benchmarks demonstrate that FineGPR with AOST outperforms (or is on par with) existing real and synthetic datasets, which suggests its feasibility for re-ID task and proves the proverbial less-is-more principle. Our synthetic FineGPR dataset is publicly available at \url{https://github.com/JeremyXSC/FineGPR}.

1. Introduction

Given a query image, person re-identification aims to match images of the same person across non-overlapping camera views, which has attracted lots of interests and attention in both academia and industry. Encouraged by the remarkable success of deep learning networks \cite{16, 26} and the availability of re-ID datasets \cite{28, 46}, performance of person re-ID has been significantly boosted and made great progress. However, in practice, manually labelling a large diversity of training data is time-consuming and labor-intensive when directly deploying re-ID system to new scenarios. During intensive annotation, one needs to associate a pedestrian across different cameras, which is a difficult and laborious process as people might exhibit very different appearances in different cameras. In addition, there also has been an increasing concern over data safety and ethical issues, \textit{e.g.} DukeMTMC-reID \cite{28, 47} has been taken down due to privacy problem. Some European countries already passed privacy-protecting laws \cite{13} to prohibit the acquisition of personal data without authorization, which makes collection of large-scale datasets extremely difficult.

To address this challenge, several works \cite{1, 2, 32, 38} have proposed to employ off-the-shelf game engines to generate synthetic person images. For example, Barbosa \textit{et al.} \cite{2}...
Figure 2. System workflow of the AOST method, which is based on the dataset with fine-grained attribute annotations.

construct a SOMAset which contains 50 3D human models. SyRI [1] provides 100 virtual humans rendered with 140 HDR environment maps. Wang et al. [38] collect a Rand-Person dataset with 1,801,816 synthesized person images. Although these datasets provide considerable benefits of the data scale and enable some preliminary research in person re-ID, they are quite limited in both the attribute distribution and collected environment, e.g., SyRI does not have concept of cameras, and SOMAset is uniformly distributed along an environment with clothing variations. In essence, these synthetic datasets either focus on independent attribute, or require annotators to carefully simulate specific scenes in detail, few datasets consider fine-grained attribute annotations or high-quality image resolution, which limits their scalability and diversity in terms of synthesized person. Another challenge we could observe is that, previous methods mainly focus on achieving competitive performance with large-scale data at the sacrifice of expensive time costs and intensive human labors, while neglect to perform efficient training with a higher quality of attribute annotations from millions of synthetic data. Considering the fact that existing real-world datasets can be very different in terms of content and style, e.g., Market-1501 [46] consists mostly of summer scenes captured in campus, while light in the CUHK03 [21] covers a wide range of indoor scenes, directly using all synthetic dataset for training will undoubtedly produce negative effects for domain adaptation, which makes it infeasible in practical scenarios.

In order to alleviate the problems identified above and facilitate the study of re-ID community, we start from two perspectives, namely data and methodology. From the aspect of the data, we propose to collect data from synthetic world on the basis of GTA5 game engine, and manually construct a Fine-grained GTA Person Re-ID dataset called FineGPR, which provides accurate and configurable annotations, e.g., viewpoint, weather, diverse and informative illumination and background, as well as the various pedestrian attribute annotations at the identity level. Compared to existing person re-ID datasets, FineGPR is explicitly distinguished in richness, quality and diversity. It is worth noting that our data synthesis engine is still extendable to generate more data, which can be edited/extended not only for this study, but also for future research in re-ID community.

From the aspect of methodology, we introduce a novel Attribute Optimization and Style Transfer pipeline AOST to perform training on re-ID task in a data-efficient manner. AOST can dynamically select samples which approximates the attribute distribution in real domain. As illustrated in Fig. 2, the proposed AOST contains Stage-I (Attribute Optimization) and Stage-II (Style Transfer). Firstly, Stage-I is adopted to mine the attribute distribution of real domain, following by the Stage-II to reduce the intrinsic gap between synthetic and real domain. Finally, the transferred data are adopted for performing training on downstream vision task. This is the first time as far as we know, to greatly promote efficient training from millions of synthetic data on re-ID task, experiments across diverse datasets suggest that the “less-is-more” principle is highly effective in practice.

Our contributions can be summarized into three aspects:

- We open source the largest person dataset with fine-grained attribute annotations for the community without privacy and ethics concerns.

- Based on it, we propose a two-stage pipeline AOST to learn from fine-grained attributes, then eliminate style differences between synthetic and real domain for more efficient training.

- Extensive experiments conducted on benchmarks show that our FineGPR is promising and can achieve competitive performance with AOST in re-ID task.

2. Related Works

2.1. Person re-ID Methods

In the field of person re-ID, early works [23,45] either concentrate on hand-crafted feature or low-level semantic feature. Unfortunately, these methods always fail to produce competitive results because of their limited discriminative learning ability. Recently, benefited from the advances of deep neural networks, person re-ID performance in supervised learning has been significantly boosted to a new level [5,22,36], which learned robust feature extraction and reliable metric learning in an end-to-end manner. Typically, person re-ID model can be trained with the identification loss [41], contrastive loss [34,35] and triplet loss [17]. Recently, a strong baseline [25] for re-ID is employed to extract the discriminative feature, which has been proved to have great potential to learn a robust and discriminative model in person re-ID models. Besides, several literatures [9,40] focus on the image-level to allow different domains to have similar feature distributions, or adopt an adversarial domain adaptation approach to mitigate the distribution shift [11,33], which has attracted considerable attention from various fields in re-ID community.
### 2.2. Person re-ID Datasets

Being the foundation of more sophisticated re-ID techniques, the pursuit of better datasets never stops in the area of person re-ID. Early attempts could be traced back to VIPeR [15], ETHZ [30] and RAID [7]. More challenging datasets are proposed subsequently, including Market-1501 [46], CUHK03 [21], MSMT17 [39], etc. However, labelling such a large-scale real-world dataset is labor-intensive and time-consuming, sometimes there even exists security and privacy problems. Besides, all of these datasets only have limited attribute distribution and are lack of diversity. As the performance gain is gradually saturated on the above datasets, newly large-scale datasets are needed urgently to further boost re-ID performance. Recently, leveraging synthetic data is an effective idea to alleviate the reliance on large-scale real-world datasets. This strategy has been applied in various computer vision tasks, e.g., object detection [27], crowd counting [37] and semantic segmentation [6]. In the person re-ID community, many re-ID methods [1, 2, 32, 38] have proposed to take advantage of game engine to construct large-scale synthetic re-ID datasets, which can be used to pre-train or fine-tune CNN network. For example, Barbosa et al. [2] propose a synthetic dataset SOMAset created by photo-realistic human body generation software to enrich the diversity. Recently, Wang et al. [38] collect a virtual dataset RandPerson with 3D characters containing 1,801,816 synthetic images of 8,000 identities. However, these datasets are either in a small scale or lack of diversity, few of them provide rich attribute annotations, which cannot satisfy the need of attribute learning in person re-ID task. So new fine-grained annotated datasets are urgently needed.

### 3.1. Dataset Collection

Our FineGPR dataset is collected from a popular game engine called the Grand Theft Auto V (GTA5). Practically, we create a synthetic controllable world containing 2,028,600 synthesized person images of 1,150 identities. Images in this dataset generally contain different attributes in a large scope, e.g., Viewpoint, Weather, Illumination, Background and ID-level annotations, also including many hard samples with occlusion. It is worth noting that, all images are simultaneously captured by 36 non-overlapping cameras with a high resolution and image quality. In the process of image generation, each person walks along a schedule route, and cameras are set up and fixed at the chosen locations. As a controllable system, it can satisfy various data requirements in a fine-grained fashion.

### 3.2. Properties of FineGPR dataset

The goal of our FineGPR dataset is to introduce a new challenging benchmark with high-quality annotations and multiple attribute distribution to re-ID community. To the best of our knowledge, this is the first large-scale person dataset over 4 environment-level attributes and 13 ID-level attribute annotations.

**Identities.** According to the Table 1, FineGPR contains 1,150 hand-crafted identities including females and males, with resolution of 200 × 480. To ensure diversity, we cropped human region with different angles. As shown in Fig. 1 (second row), different person has different body shape, clothing, hairstyle, and the motion can be randomly set as walking, running, standing and so on. Particularly, the clothes of these characters include jeans, pants, shorts, skirts, T-shirts, dress shirts, etc., and some of these identities have a backpack, shoulder bag, and wear glasses or hat. In total, we manually annotate the FineGPR with 13 different pedestrian attributes at the identity level (e.g., wearing dress or not), the distribution diagram is demonstrated in Fig. 3.

**Viewpoint.** We construct the image exemplars under specified viewpoints. Those images are randomly sampled during normal walking, running, etc. Formally, a person image is sampled every 10° from 0° to 350°. (36 different
types of viewpoints in total). There are 49 images for each viewpoint of an identity in the entire FineGPR, so each person has 1,764 (49 × 36) images in total.

**Weather.** Currently, the proposed FineGPR has 7 different weather conditions, including Sunny, Clouds, Overcast, Foggy, Neutral, Blizzard and Snowlight. It is worth mentioning that the number of instances in each weather condition is the same, but not a natural heavy-tail distribution, which makes it adaptable to various real-world scenarios.

**Illumination.** Illumination is another critical factor that contributes to the success of generalizable re-ID, which consists of 7 different types of illumination, e.g. midnight (time period during 23:00–4:00 in 24 hours a day), dawn (4:00–6:00), forenoon (6:00–11:00), noon (11:00–13:00), afternoon (13:00–18:00), dusk (18:00–20:00) and night (20:00–23:00). Parameters like time setting can be modified manually for each illumination type. By editing the values of these terms, various kinds of illumination environments can be created.

**Background.** GTA5 has a very large environment map, including thousands of realistic urban areas and wild scenes. From now, 9 different scenes are selected to represent real-world scenarios with annotations, e.g. street, mall, school, park and mountain, etc., which are distributed evenly across all identities. The different scenes are shown in Fig. 1 (first row). More additional details related to our FineGPR can be found in the Supplementary Material.

**Ethical Considerations.** People-centric datasets pose some challenges of data privacy [10] and intersectional accuracy disparities [3]. To address such concerns, our dataset were created with careful attention to ethical questions, which we encountered throughout our work. Access to our dataset will be provided for research purposes only and with restrictions on redistribution. Furthermore, we are very cautious of annotation procedure of FineGPR dataset to avoid the social and ethical implications. As for re-ID system, governments and officials must establish strict regulations to control the usage of this technology since it mainly relies on (not all) surveillance data. Motivated by this, we do not consider the dataset for developing non-research systems unless further professional processing or augmentation.

### 4. Methodology of Proposed AOST

In this section, we design an effective training strategy AOST to directly select samples on the basis of synthetic FineGPR for initializing the re-ID backbone. And the overall framework is illustrated in Fig. 4, which includes two stages: Attribute Optimization and Style Transfer.

**Attribute Optimization.** Intuitively, since FineGPR is a large-attribute-range dataset, using entire FineGPR for training is time-consuming and low-efficient. To further exploit the potential of FineGPR and promote the training efficiency, we introduce a novel strategy to learn some representative attributes with prior target knowledge. Following the procedure in [12], we adopt a widely used backbone VGG-19 [31] pre-trained on ImageNet [8] to obtain the style distance $D_{\text{style}}$ and content distance $D_{\text{content}}$ respectively, which are formulated as:

$$ D_{\text{content}} = \frac{1}{2} \sum_{i,j} \left( F_{ij} - P_{ij} \right)^2 $$

$$ D_{\text{style}} = \sum_{l=0}^{L} w_l \frac{1}{4N_l^2M_l^2} \sum_{i,j} \left( G_{ij} - A_{ij} \right)^2 $$

where $F_{ij}$ and $P_{ij}$ denote representations extracted by the $i^{th}$ filter at position $j$ in layer $l$ of VGG-19. $w_l$ is a hyper-parameter which controls the importance of each layer to the style. $N_l$ represents the number of filters and $M_l$ is size of the feature map. $G_{ij}$ and $A_{ij}$ denote the Gram Matrix of real and synthetic images in layer $l$. Then total distance for attribute metric is represented as

$$ D_{\text{total}} = \alpha \cdot D_{\text{style}} + \beta \cdot D_{\text{content}} $$

where $\alpha$ and $\beta$ are two hyper-parameters which control the relative importance of style and content distance respectively. As depicted in Algorithm 1, a tree boosting system named XGBoost [4] model $\theta_0$ is trained with FineGPR.
Figure 4. The two-stage pipeline AOST to learn attribute distribution of target domain. Firstly, we learn attribute distribution of real domain on the basis of XGBoost & PSO learning system. Secondly, we perform style transfer to enhance the reality of optimal dataset. Finally, the transferred data are adopted for downstream re-ID task.

Figure 5. Some visual examples of collected MSCO dataset.

and \( D_{\text{total}} \). Based on the upgraded Xgboost model \( \theta^* \), we continuously adopt a wildly used Particle Swarm Optimization (PSO) \([19]\) method to search some optimized attributes. Typically, it selects a similar style or content distribution with respect to a real target dataset. The optimization framework can be shown in Fig. 4 (Stage I). To our knowledge this is the first demonstration to perform attribute optimization with large-scale synthetic dataset on person re-ID task. In essence, compared with existing methods for attribute optimization, such as reinforcement learning \([29]\) and attribute descent \([42]\), our method investigates and learns these attribute distributions only with few parameters to optimize, which makes it more flexible and adaptable.

**Style Transfer.** In the above attribute optimization stage, there exists serious domain gap or distribution shift between synthetic and real-world scenario. Generative Adversarial Networks (GAN) \([14]\) which have demonstrated impressive results on image-to-image translation seem to be a natural solution to this problem. However, existing methods are both inefficient and ineffective in practical application. Their inefficiency results from the fact that a new generator needs to be retrained when given a new real-world scenario. Meanwhile, these methods mainly employ low-resolution images to train a generator, and they are incapable of fully exploiting the potential of GAN, which is likely to limit the quality of generated images. To provide a remedy to this dilemma, we build a high-resolution dataset MSCO and crawled over 20K images with a size of nearly \( 200 \times 480 \), which mainly from COCO \([24]\) dataset and few from other real-world person datasets. Different locations are also considered to cover a large diversity. We believe that a unified dataset with high-resolution can provide more useful

Algorithm 1 The Proposed AOST Method

Input: Labeled synthetic data \( L \); Unlabeled real target data \( U \);

- Initialized VGG model \( \phi \); Xgboost model \( \theta_0 \);

- Two hyper-parameters \( \alpha \) and \( \beta \); Iteration rounds \( n \);

Output: Best re-ID model \( f (w, x_i) \)

1: Initialize: \( m = 1 \), \( \text{iter} = 1 \);

2: \( \triangleright \) Attribute Optimization ***

3: Extract \( D_{\text{style}} \& D_{\text{content}} \) between \( L \) and \( U \) with model \( \phi \);

4: \( D_{\text{total}} \leftarrow \alpha \ast D_{\text{style}} + \beta \ast D_{\text{content}} \);

5: while \( m \leq \|U\| \) do

6: Optimized model \( \theta^* \leftarrow \text{train} \ \theta_0 \) with \( L \) and \( D_{\text{total}} \);

7: Optimized attributes \( V^* \leftarrow \text{update} \ \theta^* \) with PSO;

8: Update the sample size: \( m \leftarrow m + 1 \);

9: \( \textbf{end while} \)

10: Generate a new dataset \( L^* \) according to \( V^* \);

11: \( \triangleright \) Style Transfer ***

12: Performing style transfer with GAN on \( L^* \);

13: if \( \text{iter} \leq n \) then

14: Initializing re-ID model with \( L^* \) by softmax loss ;

15: \( \text{iter} \leftarrow \text{iter} + 1 \);

16: \( \textbf{end if} \)
and discriminative information during translation. Some visual examples of collected MSCO dataset are illustrated in Fig. 5. By doing so, we only need to train one generator and translate the synthetic images into photo-realistic style at testing phase. The details can be seen in Fig. 4 (Stage II). To verify the priority of MSCO, we adopt several state-of-the-art methods for style-level domain adaptation, e.g., CycleGAN [49], PTGAN [39] and SPGAN [9].

5. Experiments

5.1. Datasets and Evaluation

Market-1501 [46] contains 32,668 labeled images of 1,501 identities captured from campus in Tsinghua University. Each identity is captured by at most 6 cameras. The training set contains 12,936 images from 751 identities and the test set contains 19,732 images from 750 identities.

MSMT17 [39] has 126,441 labeled images belonging to 4,101 identities and contains 32,621 training images from 1,041 identities. For the testing set, 11,659 bounding boxes are used as query images and other 82,161 bounding boxes are used as gallery images.

CUHK03 [21] contains 14,097 images of 1,467 identities. Following the CUHK03-NP protocol [48], it is divided into 7,365 images of 767 identities as the training set and the remaining 6,732 images of 700 identities as the testing set. We adopt mean Average Precision (mAP) and Cumulative Matching Characteristics (CMC) at rank-1 and rank-5 for evaluation on re-ID task.

5.2. Experiment Settings

We mainly use the newly-built FineGPR to conduct the experiments. For attribute optimization, we empirically set $w_1 = 0.2$ in Eq. 2, and $\alpha = 0.9, \beta = 1$ in Eq. 3. It is worth mentioning that our re-ID baseline system is built only with commonly used softmax cross-entropy loss [44] on vanilla ResNet-50 [16] with no bells and whistles. Following the practice in [25], person images are resized to $256 \times 128$, then a random horizontal flipping with 0.5 probability is used for data augmentation. The batch size of training samples is set as 128. Adam method [20] is adopted for optimization. The initial learning rate is set to $3.5 \times 10^{-4}$ for the backbone network. Then, these learning rates are decayed to $3.5 \times 10^{-5}$ and $3.5 \times 10^{-6}$ at 40th epoch and 70th epoch respectively, and the training stops after 120 epochs.

5.3. Comparison with the State-of-the-arts

To evaluate the superiority of our synthetic dataset, we perform training on FineGPR and testing on each individual real dataset. The evaluation results are reported in Table 2. Surprisingly, when initializing with whole FineGPR dataset, we can achieve a rank-1 accuracy of 50.5%, 12.5% and 8.7% when tested on Market-1501, MSMT17 and CUHK03 respectively. Although there is a slight inferiority of performance when compared with RandPerson [38], our FineGPR selected by AOST with fine-grained attributes can lead a significant improvement by +0.7% and +0.8% in rank-1 accuracy on Market-1501 and CUHK03 dataset respectively. When compared with real-world datasets, FineGPR also outperforms these benchmarks by an impressively large margin in terms of rank-1 accuracy, leading +0.3% and +13.9% improvement on Market-1501 compared with MSMT17 and CUHK03 separately. However, initializing with whole FineGPR dataset is time-consuming and low-efficient, this motivates the investigation of data selecting techniques that can potentially address this problem.

5.4. Ablation Study

Important Parameters. In Eq. 3, both $\alpha$ and $\beta$ controls the relative importance of the style and content distance respectively between real and synthetic samples. Since $D_{\text{total}}$ is a linear combination between $D_{\text{style}}$ and $D_{\text{content}}$, we can smoothly regulate the emphasis or adjust the trade-off between the content or the style. As depicted in Fig. 6 (a),

| Training set | Reference | Synthetic data | Rank-1 | Rank-5 | mAP | Rank-1 | Rank-5 | mAP | Rank-1 | Rank-5 | mAP |
|--------------|-----------|----------------|--------|--------|-----|--------|--------|-----|--------|--------|-----|
| Market-1501  | ICCV 15   | ×              | 92.7   | 97.9   |     | 6.0   | 11.2   | 1.9 | 9.3    | 12.4   | 6.2 |
| MSMT17 [39]  | CVPR 18   | ×              | 50.2   | 67.7   | 25.7 | 25.7  | 59.6   | 5.4 | 9.9    | 20.4   | 10.7|
| CUHK03 [21]  | CVPR 14   | ×              | 36.6   | 53.9   | 16.6 | 4.6   | 10.1   | 1.3 | 4.3    | 62.9   | 41.5|
| SOMAset* [2] | CVIU 18   | ✓              | 4.5    | 1.3    |     | 1.4   | 0.3    |     | 0.4    | 0.4    |     |
| SyRi* [1]    | ECCV 18   | ✓              | 29.0   | 10.8   |     | 16.4  | 4.4    |     | 4.1    | 3.5    |     |
| Unreal† [43] | CVPR 21   | ✓              | 37.4   | 55.2   | 15.9 | 3.9   | 7.4    | 1.3 | 4.3    | 10.0   | 4.7 |
| PersonX* [32]| CVPR 19   | ✓              | 44.0   | 20.4   |     | 11.7  | 3.6    |     | 7.4    | 6.2    |     |
| RandPerson*†[58]| MM 20 | ✓              | 55.6   | 28.8   | 20.1 | 20.1  | 6.3    |     | 13.4   | 10.8   |     |

Table 2. Performance comparison with existing Real and Synthetic datasets on Market-1501, MSMT17 and CUHK03, respectively. Red indicates the best and Blue the second best. * means results are reported by RandPerson [38]. † represents results reproduced with Unreal v2.1 on our baseline. Underline denotes supervised learning. ↑ means performing selecting with our AOST method.
we observe that when $\alpha/\beta$ is small, the performances is not optimal because the style representation is way too limited to a very small portion, and thus our AOST could only mine the discriminative information in terms of content representation of the re-ID data. The $\alpha/\beta$ should also not be set too large, otherwise the performances drops dramatically since the model mine too many samples in style representation. Specially, $\alpha/\beta = 0.9$ yields the best accuracy.

**Evaluation of Attribute Importance.** Based on the end-to-end AOST system, we evaluate the impacts of different attributes in a fine-grained manner by the XGBoost in terms of gain [4] (feature importance score). As illustrated in Fig. 6 (b-d), it can be easily observed that **Identity** accounts for the largest proportion no matter which real dataset is employed for testing, followed by **Viewpoint** attribute. Typically, this conclusion is in accordance with our prior knowledge on generalizable re-ID problem, that is, using more IDs as training samples is always beneficial to the re-ID system, and viewpoint also plays a key role in recognizing the clothes appearance. More details about the importance related to ID-level attributes is provided in the Supplementary. We hope these analysis about attribute importance will provide useful insights for dataset building and future practical usage to the community.

**Effectiveness of Attribute Optimization.** We proceed study on dependency by testing whether the Attribute Optimization (AO) matters. According to Table 3, our attribute optimization strategy FineGPR+AO (w/o transfer) can lead a significant improvement in rank-1 of +5.6%, +4.5% and +2.6% on Market-1501, MSMT17 and CUHK03 respectively when compared with random sampling (FineGPR+R). We suspect this is due to samples selected by attribute optimization strategy are much closer to real target domain, and the learned attribute distribution has a higher quality, which have a direct impact on downstream re-ID task. Meanwhile, fast training is our second main advantage since the scale of training set can be largely decreased by attribute optimization, e.g., it costs nearly 20 GPU-days$^1$ when pre-training on entire FineGPR with 2,028,600 images. Fortunately, training time will be considerably reduced by $15 \times (20 \text{ vs. } 1.3 \text{ GPU-days})$ by our proposed AOST without performance degradation, which leads a more efficient deployment to real-world scenarios. Surprisingly, even with fewer samples for training, our approach still yields its competitiveness when compared with existing datasets, e.g. 56.3% vs. 55.6% in rank-1 on Market in Table 2, proving the proverbial less-is-more principle.

To go even further, we also adopt AOST on synthetic datasets to prove the priority of FineGPR. Unfortunately, due to lack of fine-grained attribute annotations, these datasets (e.g. PersonX, Unreal and RandPerson) cannot satisfy the need for AOST in re-ID. We instead randomly sample 124,200 images from these datasets individually, and then perform style transfer. As illustrated in Table 3, it can Figs. 6(a-c).

---

**Table 3.** Controlled experiments by different regulations of our proposed AOST method on Market-1501, MSMT17 and CUHK03, respectively. “R” indicates random sampling, “AO” represents Attribute Optimization. “ST” means Style Transfer.

| Training set ↓ | Bboxes | Time (GPU-days) ↓ | Rank-1 | Rank-5 | mAP | Rank-1 | Rank-5 | mAP | Rank-1 | Rank-5 | mAP |
|----------------|--------|-------------------|--------|--------|-----|--------|--------|-----|--------|--------|-----|
| FineGPR        | 2,028,600 | 20               | 50.5   | 67.7   | 24.6| 12.5   | 18.5   | 3.9 | 8.7    | 18.2   | 8.4 |
| FineGPR+R      | 124,200  | 1.3              | 39.9   | 57.1   | 18.3| 4.6    | 9.7    | 1.4 | 5.9    | 15.4   | 5.4 |
| FineGPR+AO     | 124,200  | 1.3              | 45.5   | 63.2   | 23.8| 9.1    | 14.7   | 3.1 | 8.5    | 16.9   | 8.3 |
| PersonX+R+ST   | 124,200  | 1.3              | 28.7   | 45.9   | 11.8| 7.1    | 13.4   | 2.1 | 3.1    | 7.4    | 3.1 |
| Unreal+R+ST    | 124,200  | 1.3              | 42.8   | 59.3   | 18.4| 11.3   | 20.5   | 3.5 | 5.4    | 12.6   | 5.1 |
| RandPerson+R+ST| 124,200  | 1.3              | 51.4   | 68.3   | 25.0| 15.8   | 26.7   | 5.0 | 8.4    | 18.1   | 7.5 |
| FineGPR+AO+ST  | 124,200  | 1.3              | 56.3   | 70.4   | 29.2| 19.7   | 27.4   | 6.1 | 14.2   | 20.6   | 11.2 |

---

---

1 All timings use one Nvidia Tesla P100 GPU on a server equipped with an Intel Xeon E5-2690 V4 CPU.
be easily observed that FineGPR+AOST can perform significantly greater than PersonX+R+ST, Unreal+R+ST and RandPerson+R+ST separately, which successfully proves the applicability of our proposed dataset and approach.

**Effectiveness of Style Transfer.** Synthetic data engine can generate images and annotations at lower labor costs. However, there exists obvious domain gap between synthetic and real-world scenario, which hinders the further improvement in performance on downstream task. Note that for training efficiency, we instead consider an easier but practical strategy, that is, employing off-the-shelf style transfer model to generate photo-realistic images for further effective training. As shown in Table 3, w/o style transfer by AOST, the rank-1 accuracy drops sharply from 56.3% to 45.5% and the mAP drops from 29.2% to 23.8%. *This confirms that mitigating domain gap between synthetic and real dataset is a crucial ingredient to make the performance to an excellent level.* For simplicity, the SPGAN is used as the style transfer model in the following experiments.

### 5.5. Qualitative and Quantitative Results

**Qualitative Evaluations.** Fig. 7 presents a visual comparison of different transfer methods trained on low-resolution Market and high-resolution MSCO respectively. It can be easily noticed that GANs create artifacts and coarse results (indicated by yellow box) when trained on low-resolution images, which still remains problematic. In comparison, our method with MSCO can successfully address the artifacts and produce most visually pleasant results (indicated by red box) in an even better fashion, which is implicitly beneficial to the downstream re-ID mission.

**Quantitative Evaluations.** Our qualitative observations above are confirmed by the quantitative evaluations. To be more specific, we adopt Fréchet Inception Distance (FID) [18] to measure the distribution difference between synthetic and real photos. Generally, FID measures how close the distribution of generated images is to the real. As shown in Fig. 8, by adding new regulation terms, *e.g.* attribute optimization or style transfer, the FID score gradually decreases no matter which dataset is employed for evaluation, suggesting the learned attribute distributions are more and more similar to the real images. Even prior to this point, according to Fig. 9, training a generator with low-resolution images always produce low-quality images (indicated by higher FID score) no matter which style transfer model is employed. Still, SPGAN and PTGAN rank the best, while SPGAN shows a slight quantitative advantage. In all, the introduction of high-resolution MSCO dataset can always improve the adaptability to style changes and mitigate previously mentioned domain gap effectively, even in much more complex scenarios.

### 6. Conclusion

In this work, we take the first step to construct the largest person dataset FineGPR with fine-grained attribute labels and high-quality annotations. On top of FineGPR, we introduce an attribute analysis methodology called AOST to learn important attribute distribution, which enjoys the benefits of small-scale dataset for more efficient training. Continuously, style transfer is adopted to further mitigate domain gap between synthetic and real photos. With this, we proved, for the first time, that a model trained on limited synthetic data can yield a competitive performance in generalizable re-ID task. Extensive experiments also demonstrate the superiority of FineGPR and effectiveness of AOST. We hope our dataset and method will shed light into potential tasks for the community to move forward.
References

[1] Slawomir Bak, Peter Carr, and Jean-Francois Lalonde. Domain adaptation through synthesis for unsupervised person re-identification. In Proceedings of the European Conference on Computer Vision (ECCV), pages 189–205, 2018. 1, 2, 3, 6

[2] Igor Barroso Barbosa, Marco Cristani, Barbara Caputo, Alexander Rognhaugen, and Theoharis Theoharis. Looking beyond appearances: Synthetic training data for deep cnns in re-identification. Computer Vision and Image Understanding, 167:50–62, 2018. 1, 3, 6

[3] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Conference on fairness, accountability and transparency, pages 77–91. PMLR, 2018. 4

[4] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, pages 785–794, 2016. 4, 7

[5] Weihua Chen, Xiaotang Chen, Jianguo Zhang, and Kaiqi Huang. Beyond triplet loss: a deep quadruplet network for person re-identification. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 403–412, 2017. 2

[6] Yuhua Chen, Wen Li, Xiaoran Chen, and Luc Van Gool. Learning semantic segmentation from synthetic data: A geometrically guided input-output adaptation approach. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 1841–1850, 2019. 3

[7] Abir Das, Anirban Chakraborty, and Amit K Roy-Chowdhury. Consistent re-identification in a camera network. In European conference on computer vision, pages 330–345. Springer, 2014. 3

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition, pages 248–255. Ieee, 2009. 4

[9] Weijian Deng, Liang Zheng, Qixiang Ye, Guoliang Kang, Yi Yang, and Jianbin Jiang. 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, pages 994–1003, 2018. 2, 6

[10] Matteo Fabbri, Guillem Brasó, Gianluca Maugeri, Orcun Cetintas, Riccardo Gasparini, Aljosa Osep, Simone Calderara, Laura Leal-Taixe, and Rita Cucchiara. Motsynth: How can synthetic data help pedestrian detection and tracking? In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 10849–10859, 2021. 4

[11] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In International conference on machine learning, pages 1180–1189. PMLR, 2015. 2

[12] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 2414–2423, 2016. 4

[13] Michelle Goddard. The eu general data protection regulation (gdpr): European regulation that has a global impact. International Journal of Market Research, 59(6):703–705, 2017. 1

[14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. Advances in neural information processing systems, 27, 2014. 5

[15] Douglas Gray, Shane Brennan, and Hai Tao. Evaluating appearance models for recognition, reacquisition, and tracking. In Proc. IEEE international workshop on performance evaluation for tracking and surveillance (PETS), volume 3, pages 1–7. Citeseer, 2007. 3

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016. 1, 6

[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] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems, 30, 2017. 8

[19] James Kennedy and Russell Eberhart. Particle swarm optimization. In Proceedings of ICNN'95-international conference on neural networks, volume 4, pages 1942–1948. IEEE, 1995. 5

[20] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014. 6

[21] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 152–159, 2014. 2, 3, 6

[22] Wei Li, Xiatian Zhu, and Shaogang Gong. Harmonious attention network for person re-identification. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 2285–2294, 2018. 2

[23] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z Li. Person re-identification by local maximal occurrence representation and metric learning. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 2197–2206, 2015. 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 European conference on computer vision, pages 740–755. Springer, 2014. 5

[25] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pages 0–0, 2019. 2, 6

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

[27] Bojan Pepik, Michael Stark, Peter Gehler, and Bernt Schiele. Teaching 3d geometry to deformable part models. In 2012 IEEE conference on computer vision and pattern recognition, pages 3362–3369. IEEE, 2012.

[28] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In European conference on computer vision, pages 17–35. Springer, 2016.

[29] Nataniel Ruiz, Samuel Schulter, and Mannmohan Chandraker. Learning to simulate. arXiv preprint arXiv:1810.02513, 2018.

[30] William Robson Schwartz and Larry S Davis. Learning discriminative appearance-based models using partial least squares. In 2009 XXII Brazilian symposium on computer graphics and image processing, pages 322–329. IEEE, 2009.

[31] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.

[32] Xiaoxiao Sun and Liang Zheng. Dissecting person re-identification from the viewpoint of viewpoint. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 608–617, 2019.

[33] Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In CVPR 2011, pages 1521–1528. IEEE, 2011.

[34] Rahul Rama Varior, Mrinal Haloi, and Gang Wang. Gated siamese convolutional neural network architecture for human re-identification. In European conference on computer vision, pages 791–808. Springer, 2016.

[35] Rahul Rama Varior, Bing Shuai, Jiwen Lu, Dong Xu, and Gang Wang. A siamese long short-term memory architecture for human re-identification. In European conference on computer vision, pages 135–153. Springer, 2016.

[36] Faqiang Wang, Wangmeng Zuo, Liang Lin, David Zhang, and Lei Zhang. Joint learning of single-image and cross-image representations for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1288–1296, 2016.

[37] Qi Wang, Junyu Gao, Wei Lin, and Yuan Yuan. Learning from synthetic data for crowd counting in the wild. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 8198–8207, 2019.

[38] Yanan Wang, Shengcai Liao, and Ling Shao. Surpassing real-world source training data: Random 3d characters for generalizable person re-identification. In Proceedings of the 28th ACM International Conference on Multimedia, pages 3422–3430, 2020.

[39] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 79–88, 2018.

[40] Suncheng Xiang, Yuzhuo Fu, Guanjie You, and Ting Liu. Unsupervised domain adaptation through synthesis for person re-identification. In 2020 IEEE International Conference on Multimedia and Expo (ICME), pages 1–6. IEEE, 2020.

[41] Tong Xiao, Shuang Li, Bochao Wang, Liang Lin, and Xiaogang Wang. Joint detection and identification feature learning for person search. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3415–3424, 2017.

[42] Yue Yao, Liang Zheng, Xiaodong Yang, Milind Naphade, and Tom Gedeon. Simulating content consistent vehicle datasets with attribute descent. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020. Proceedings, Part VI, pages 775–791. Springer, 2020.

[43] Tianyu Zhang, Lingxi Xie, Longhui Wei, Zijie Zhuang, Yongfei Zhang, Bo Li, and Qi Tian. Unrealperson: An adaptive pipeline towards costless person re-identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 11506–11515, 2021.

[44] Zhilu Zhang and Mert R Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. In 32nd Conference on Neural Information Processing Systems (NeurIPS), 2018.

[45] Rui Zhao, Wanli Ouyang, and Xiaogang Wang. Learning mid-level filters for person re-identification. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 144–151, 2014.

[46] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jindong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In Proceedings of the IEEE international conference on computer vision, pages 1116–1124, 2015.

[47] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In Proceedings of the IEEE international conference on computer vision, pages 3754–3762, 2017.

[48] Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. Re-ranking person re-identification with k-reciprocal encoding. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1318–1327, 2017.

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

This supplementary material accompanies our main manuscript “Less is More: Learning from Synthetic Data with Fine-grained Attributes for Person Re-Identification”, including more visual results related to our FineGPR dataset in terms of viewpoint, weather, illumination and background distribution, as well as ID-level attributes, and more analysis experiment results of our proposed AOST, which related to the Section 3, Section 4 and Section 5 in the main manuscript respectively. In addition, FineGPR is not limited to the specific research field of person re-ID, it can also be extended to broader research areas, including other human related task such as segmentation and pose estimation. We hope this dataset could open a door to related research of future tasks.

1. More Demonstrations of FineGPR Dataset

In this section, we will give more visual examples on the distribution of viewpoint, weather, illumination and background attributes of our proposed FineGPR dataset in the environment level, as well as multiple pedestrian attributes at the identity level, which will help us better understand the advantages of FineGPR. It is worth mentioned that all attributes at the environment level are distributed evenly across all identities in FineGPR, which is not heavily biased, but a uniform distribution, e.g. viewpoint, weather, illumination and background. So every identity has 36 different viewpoints, 7 different types of weather, 7 different types of illumination and 9 different scenes.

**Viewpoint.** We present the image exemplars under 36 different types of viewpoints. To be more specific, a person image is sampled at an interval of 10°. For example, a person has 9 different angles in total.

**Weather.** For a deeper understanding of our dataset, we present a brief demonstration of FineGPR which consists of 7 different types of weather, e.g. sunny, clouds, overcast, foggy, neutral, blizzard and snowlight, some visual examples are shown in Fig. 2. These attributes are thoroughly studied in the following experiments.

**Illumination.** As illustrated in Fig. 2, we present some exemplars in different illuminations on top of FineGPR, which consists of 7 different types of illumination, e.g., midnight (23:00–4:00), dawn (4:00–6:00), forenoon (6:00–11:00), noon (11:00–13:00), afternoon (13:00–18:00), dusk (18:00–20:00) and night (20:00–23:00). Intuitively, we classify these illumination distributions into three categories: Weak Light: midnight, night; Middle Light: dawn, dusk; Strong Light: forenoon, noon,
Weathers distribution

Illuminations distribution

Figure 2. The exemplars of different weather distribution (left) and illumination distribution (right) from the proposed FineGPR dataset.

Figure 3. The statistical information of proposed FineGPR on illumination (strong light, middle light and weak light) and background (urban area and wild scene) distribution.

afternoon. All these attributes are carefully designed and provide the new exploration space for fine-grained attribute analysis in re-ID task.

Locations in GTA5 World We have marked the position of each location of our 9 backgrounds in GTA5 world, as shown in Fig. 16. In particular, considering the fact that real-world scenarios or existing benchmark datasets in re-ID task are mainly located in the urban scene, so our locations are mainly concentrated in the urban area: street, mall, school, underground parking; few of them are wild scene: park, beach and mountain. The population distribution diagram of our FineGPR dataset is shown in Fig. 3. Those considerations result in complex backgrounds and scene variations, also make FineGPR more appealing and challenging, etc. In the future, we will continue to enrich our FineGPR datasets by selecting more representative locations and adding more human identities.  

Identity-level Attributes. Inspired by the work [2], our synthetic dataset FineGPR also provides pedestrian attribute label at the identity level, the category name for all images in the FineGPR dataset is illustrated in Table 4. The attributes are carefully selected and manually annotated according to the characteristics of the FineGPR, which can be used to further boost the performance of generalizable re-ID with attention mechanism. To be more specific, we have labeled 13 different attribute annotations at the identity level: Gender (male, female); hair length (long, short); Age (teenager, adult, older); Wear glass (yes, no); Wear short sleeve (yes, no); length of top clothes (yes, no); Wear dress (yes, no); Wear boot (yes, no); Wear hat (yes, no); Carry bag (yes, no); Color of shoes (dark, light); Color of upper-body clothes (black, white, red, purple, gray, yellow, blue, green, blown); Color of lower-body clothes (black, white, red, purple, gray, yellow, blue, green, blown). Some representative samples with attribute label at the identity level are also demonstrated in Fig. 4. Intuitively, there exists a potential direction that our FineGPR can also be applied in other surveillance task (e.g. pedestrian attribute recognition) for fine-grained analysis.

2. More Experiment Results

2.1. Analysis of ID-level Attribute Importance

Based on the end-to-end AOST system, we evaluate the impacts of different attributes in a fine-grained manner by the boosting system XGBoost in terms of gain1 (feature importance score). As illustrated in Fig. 5, it can be easily observed that Colors of upper-body clothes and Colors of lower-body clothes account for the largest proportion when testing in real-world domain. This observation is consistent with our intuition because the color of upper-body /lower-body clothes contain more detailed and discriminative information, a model trained with discriminative cloth appear-

---

1https://xgboost.readthedocs.io/en/latest/tutorials/model.html
Wear short sleeve  Wear dress  Wear hat  Carry bag

Figure 4. Some visual exemplars with ID-level pedestrian attributes in the proposed FineGPR dataset, such as Wear short sleeve, Wear dress, Wear hat, Carry bag, etc.

![Feature importance at Identity-level with Feature Importance Score](image)

Figure 5. Feature importance at Identity-level with Feature Importance Score (higher is more important) on Market-1501.

ances is good at recognizing the pedestrians under different orientations.

We also provide feature distance distribution $D_{\text{total}}$ by our attribute optimization on Market-1501, which help us better understand the distance distribution based on our AOST with fine-grained FineGPR dataset. As shown in Fig. 6, those attributes whose value below horizontal line can be regarded as optimal items, which is more closer to real-world scenarios in re-ID task.

2.2. Visualization of Style Transfer Results

Fig. 11 demonstrates the translated images in high-resolution settings. To be more specific, we train the SP-GAN with MSCO dataset instead of original low-resolution images (e.g. Market-1501). Consequently, when compared with original synthetic data FineGPR, our method can generate more consistent and photo-realistic images in a even better fashion, which proves that using a unified real-world dataset with high-resolution can learn more discriminative embeddings and bring great benefits to the GAN models.

Since previous PersonX, Unreal and RandPerson datasets do not provide fine-grained attribute annotations, which fail to satisfy the need for AOST in generalizable re-ID task. To solve this problem, we instead randomly sample 124,200 images from these datasets individually, and then perform style transfer. The detailed analysis can be listed as following:

From a qualitative point of view, as shown in Fig. 12, Fig. 13 and Fig. 14, it can be obviously observed that the style transfer results of PersonX+R+ST, Unreal+R+ST and RandPerson+R+ST tend to loss some texture feature and produce some abnormal color (Marked with red rectangle). We argue that this is due to the samples in PersonX, Unreal and RandPerson are low-resolution, and GANs tend to have semantic shifts and create some noises inevitably, which seriously degrades the performance of re-ID model.

More importantly, from a quantitative point of view, we
Testing set → Market-1501 MSMT17 CUHK03
Training set ↓ Bboxes viewpoint Rank-1 Rank-5 mAP Rank-1 Rank-5 mAP Rank-1 Rank-5 mAP
FineGPR
225400 front-view 47.2 65.0 22.9 10.1 18.4 2.7 10.8 21.8 9.3
225400 side-view 46.8 64.2 22.5 8.8 16.9 2.3 8.1 17.9 7.7
225400 random-view 46.6 63.2 21.9 9.1 16.9 2.4 9.6 19.4 8.7

Table 1. Evaluation performance (%) of baseline model on Market-1501, MSMT17 and CUHK03 respectively in terms of viewpoint, e.g., front-view, side-view and random-view.

Testing set → Datasets Fusion Market-1501 MSMT17 CUHK03
Training set ↓
Market × 92.7 92.9 81.4 6.0 11.2 1.9 5.3 12.4 6.2
MSMT17 × 50.2 67.7 25.7 75.7 86.9 51.5 9.9 20.4 10.7
CUHK03 × 36.6 53.9 16.6 4.6 9.1 1.3 43.6 62.9 41.5
FineGPR+R × 39.9 57.1 18.3 4.6 9.7 1.4 5.9 15.4 5.4
FineGPR+R+Market ✓ 92.6 97.4 82.4 8.3 15.4 2.9 14.1 25.8 13.1
FineGPR+R+MSMT17 ✓ 56.0 71.4 30.3 71.2 83.3 47.0 14.2 27.3 14.5
FineGPR+R+CUHK03 ✓ 42.4 60.4 20.7 7.0 13.5 2.2 37.9 57.3 36.4

Table 2. Evaluation performance (%) of baseline model trained on various datasets and tested on Market-1501, MSMT17 and CUHK03 respectively. “R” indicates random sampling. Underline denotes supervised learning.

Figure 7. Comparison of FID (lower is better) to evaluate the realism of style transferred images of FineGPR, PersonX, Unreal and RandPerson respectively when comparing with real images, e.g., MSCO, Market-1501, MSMT17 and CUHK03 individually.

also provide a quantitative analysis on the quality of transferred images of FineGPR, PersonX, Unreal and RandPerson. As shown in Fig. 7, when compared to the real-world images in terms of realism, e.g., MSCO, Market-1501, MSMT17 and CUHK03, the transferred images of FineGPR always have lowest FID score than transferred samples of PersonX, Unreal and RandPerson, this significantly validates the priority of our transferred FineGPR dataset.

2.3. Visualization of Attribute Optimization Results

To further validate the effectiveness of our attribute optimization method, we present some visual exemplars qualitatively w.r.t different real target datasets. As shown in Fig. 8, it can be obviously observed that the selected samples from FineGPR are more similar to the real target samples in terms of style and content representation, demonstrating the effectiveness of our optimization method.

2.4. Impact of Viewpoint

Based on the FineGPR, we further evaluate the impact of viewpoint on generalizable person re-ID quantitatively. The experiment is based on the re-ID baseline in the main manuscript. We note that each subset contains 225,400 images with 1,150 identities of 4 kinds of different angles. The detailed viewpoint settings are designed as following:

- Experimental group 1. Random-view Setting. This subset includes samples randomly selected from FineGPR with angles of 100°, 150°, 190° and 350°.
- Experimental group 2. Side-view Setting. This subset includes samples from FineGPR with angles of 40°, 130°, 220° and 310°.
- Experimental group 3. Front-view Setting. This subset includes samples from FineGPR with angles of 0°, 90°, 180° and 270°.

According to the Table 1, the major observation is that...
there seems to exist obvious correlation between viewpoint distribution and performance results. To be more specific, using samples with Front-view setting can always bring significant performance gains to the re-ID system, no matter which dataset is adopted as testing-set, (e.g. Market-1501, MSMT17 and CUHK03). This observation is intuitive because the samples with front-view contain more discriminative information and clothes appearance of human body, which help to learn a discriminative re-ID model. Consequently, the performance can be greatly improved when identifying pedestrians under other different orientations.

2.5. Evaluation of Dataset Fusion

To further verify the superiority of our proposed FineGPR, we perform another round of experiments with dataset fusion strategy. To be more specific, we combine our synthetic dataset FineGPR with single real-world datasets during the experiment, e.g., Market-1501, MSMT17 and CUHK03. In particular, we randomly select 124,200 images from our synthetic dataset (Denoted as FineGPR+R). In this case, as reported in Table 2, there are two observations which can be made.

First, when compared to training with single real-world dataset, dataset fusion strategy can always boost the generalizable re-ID performance into a satisfactory level. For example, as illustrated in Table 2, we can achieve a rank-1 accuracy of 56.0% and 42.4% on Market-1501 when trained with FineGPR+R+MSMT17 and FineGPR+R+CUHK03 respectively, outperforming the single-dataset training (MSMT17 or CUHK03) by +5.8% and +5.8%, respectively.

Second and importantly, compared to only using FineGPR+R for training, as can be observed in Table 2, dataset fusion strategy can significantly boost the performance from 4.6% to 8.3% and 7.0% on MSMT17 when combining FineGPR+R with Market-1501 and CUHK03 respectively. It can be concluded that combining synthetic data with real-world data for training can bridge the domain gap between synthetic and real-world datasets and help to learn a robust and discriminative model.

2.6. Pre-training

Does FineGPR can replace ImageNet for pre-training?

To find the answer to this question, we use different pre-trained models (ImageNet pre-train and FineGPR pre-train) on the re-ID baseline in a traditional pre-train manner. As shown in Fig. 9, it can be seen that, pre-training with FineGPR achieves slightly inferior performance than ImageNet pre-training, but still acceptable. One potential reason is the large domain gap between real and synthetic images. Consequently, we suspect that this fine-grained attribute dataset is not applicable to the traditional pre-training of coarse-grained retrieval tasks in some degree.

Surprisingly, recent research work by Xiang et al. [3] has given re-ID community some insights on pre-training, which demonstrated that a pure semantic-based pre-training approach VTBR\(^2\), by leveraging the fine-grained attributes of FineGPR, can outperform the traditional ImageNet pre-training by a clear margin (e.g., mAP 85.3% vs. 85.0% on Market-1501 in a supervised manner), so FineGPR can still reveal its potential in re-ID pretraining. Further details related to pretraining may be found in [3].

3. Extendibility

3.1. More Tasks

Accompanied with our FineGPR, we also provide some human body masks and keypoint locations of all characters during the annotation. The standard skeleton provided among with our FineGPR has 17 keypoints, which is obtained by AlphaPose [1], and segmentation for human mask is extracted by PSPNet [4]. More pose estimation results and segmentation results are shown in Fig. 15. For example, we have adopted human body masks to reproduce style transfer results with PTGAN. To sum up, we hope that our synthetic dataset FineGPR can not only contribute a lot to the development of generalizable person re-ID, but also advance the research of other computer vision tasks, such as human part segmentation and pose estimation.

3.2. More Identities

According to the Table 2 in the main manuscript, we observe that our synthetic FineGPR can achieve a satisfactory performance in most cases for generalizable re-ID task. However, it is still not as good as RandPerson where 132,145 person images of 8,000 identities are adopted for training. We suspect that the success of RandPerson can be

\( ^2 \)The VTBR adopts another total different re-ID backbone.
Table 3. Performance comparison between RandPerson and our enhanced FineGPR_V2 dataset.

| Dataset    | Market-1501 | MSMT17 | CUHK03 |
|------------|-------------|--------|--------|
| RandPerson | 55.6        | 28.8   | 20.1   | 6.3    | 13.4   | 10.8  |
| FineGPR_V2 | 53.7        | 26.6   | 18.8   | 5.4    | 12.5   | 9.8   |

Figure 10. Some failure cases in the style transfer.

largely attributed to the huge number of identities, which can greatly enhance the discrimination capability of backbone network. Note that for fairness, we manually construct a ID-enhanced version of FineGPR dataset (named as FineGPR_V2), which contains 137,808 samples of 6506 identities. As illustrated in Table 3, it can be easily observed that the FineGPR_V2 can achieve competitive performance with RandPerson when Identities are highly promoted, which again verified the attribute importance of Identity. (Consistent with the conclusion in the Section 5.4 of the main manuscript).

4. Limitations

The proposed method in this paper can dynamically mine some critical pedestrian samples which is closer to the specific domain, then perform style transfer. Despite its promising performance on person re-ID, we note that there are several limitations in AOST, including (1) the premise of our AOST is that the attribute distribution of target domain are fixed, this means the sample selection process might be repeated when target domain is gradually updated; (2) the GAN model will generate semantic shifts in some extremely cases, as shown in Fig. 10, which may bring negative impacts on downstream tasks. We believe addressing these challenges are promising directions of our work for future research.

References

[1] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. Rmpe: Regional multi-person pose estimation. In ICCV, pages 2334–2343, 2017.
[2] Yutian Lin, Liang Zheng, Zhedong Zheng, Yu Wu, Zhilan Hu, Chenggang Yan, and Yi Yang. Improving person re-identification by attribute and identity learning. Pattern Recognition, 95:151–161, 2019.
[3] Suncheng Xiang, Zirui Zhang, Mengyuan Guan, Hao Chen, Binjie Yan, Ting Liu, and Yuzhuo Fu. Vtbr: Semantic-based pretraining for person re-identification. arXiv preprint arXiv:2110.05074, 2021.
[4] Hengshuang Zhao, Jianping Shi, Xiaojian Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In CVPR, pages 2881–2890, 2017.
| Attribute                    | Category name | Label ID |
|------------------------------|---------------|----------|
| Gender                       | male          | 01       |
|                              | female        | 02       |
| Hair length                  | long          | 01       |
|                              | short         | 02       |
| Age                          | teenager      | 01       |
|                              | adult         | 02       |
|                              | older         | 03       |
| Wear glass                   | yes           | 01       |
|                              | no            | 02       |
| Wear short sleeve            | yes           | 01       |
|                              | no            | 02       |
| Length of top clothes        | long          | 01       |
|                              | short         | 02       |
| Wear dress                   | yes           | 01       |
|                              | no            | 02       |
| Wear boot                    | yes           | 01       |
|                              | no            | 02       |
| Wear hat                     | yes           | 01       |
|                              | no            | 02       |
| Carry bag                    | yes           | 01       |
|                              | no            | 02       |
| Color of shoes               | dark          | 01       |
|                              | light         | 02       |
| Colors of upper-body clothes | black         | 01       |
|                              | white         | 02       |
|                              | red           | 03       |
|                              | purple        | 04       |
|                              | gray          | 05       |
|                              | yellow        | 06       |
|                              | blue          | 07       |
|                              | green         | 08       |
|                              | brown         | 09       |
| Colors of lower-body clothes | black         | 01       |
|                              | white         | 02       |
|                              | red           | 03       |
|                              | purple        | 04       |
|                              | gray          | 05       |
|                              | yellow        | 06       |
|                              | blue          | 07       |
|                              | green         | 08       |
|                              | brown         | 09       |

Table 4. List of all the ID-level attribute annotations in FineGPR. The relative class id is also provided.
Figure 11. Some visual samples of transferred person images from synthetic FineGPR dataset to photo-realistic style by SPGAN. (FineGPR→MSCO). It can be easily observed that there exist obvious differences between original synthetic FineGPR and our transferred results with photo-realistic style.

Figure 12. Some visual samples of transferred person images from synthetic PersonX dataset to photo-realistic style by SPGAN (PersonX→MSCO). Some cases tend to produce some abnormal color (Marked with red rectangle). Please zoom in for better observation.
Figure 13. Some visual samples of transferred person images from Unreal dataset to photo-realistic style by SPGAN (Unreal→MSCO). Some cases tend to produce some abnormal color (Marked with red rectangle). Please zoom in for better observation.

Figure 14. Some visual samples of transferred person images from RandPerson dataset to photo-realistic style by SPGAN (RandPerson→MSCO). Some cases tend to produce some abnormal color (Marked with red rectangle). Please zoom in for better observation.
Figure 15. Some visual results with human body segmentation (Middle) and pose estimation (Bottom) based on our FineGPR dataset.
Figure 16. The position of each location of 9 backgrounds from our proposed FineGPR dataset in GTA5 world.