TAGPerson: A Target-Aware Generation Pipeline for Person Re-identification

Kai Chen
chenkai2010.9@gmail.com
Tsinghua University & BNRist
Beijing, China

†Weihua Chen
kugang.cwh@alibaba-inc.com
Alibaba Group
Beijing, China

Tao He
kevin.92.he@gmail.com
Tsinghua University & BNRist
Beijing, China

Rong Du
dr193339@alibaba-inc.com
Alibaba Group
Shanghai, China

Fan Wang
fan.w@alibaba-inc.com
Alibaba Group
Sunnyvale, United States

Xiuyu Sun
xiuyu.sxy@alibaba-inc.com
Alibaba Group
Beijing, China

Yuchen Guo
yuchen.w.guo@gmail.com
BNRist
Beijing, China

†Guiguang Ding
dinggg@tsinghua.edu.cn
Tsinghua University & BNRist
Beijing, China

ABSTRACT

Nowadays, real data in person re-identification (ReID) task is facing privacy issues, e.g., the banned dataset DukeMTMC-ReID. Thus it becomes much harder to collect real data for ReID task. Meanwhile, the labor cost of labeling ReID data is still very high and further hinders the development of the ReID research. Therefore, many methods turn to generate synthetic images for ReID algorithms as alternatives instead of real images. However, there is an inevitable domain gap between synthetic and real images. In previous methods, the generation process is based on virtual scenes, and their synthetic training data can not be changed according to different target real scenes automatically. To handle this problem, we propose a novel Target-Aware Generation pipeline to produce synthetic person images, called TAGPerson. Specifically, it involves a parameterized rendering method, where the parameters are controllable and can be adjusted according to the target scenes. In TAGPerson, we extract information from target scenes and use them to control our parameterized rendering process to generate target-aware synthetic images, which would hold a smaller gap to the real images in the specific target domain. In our experiments, our target-aware synthetic images can achieve a much higher performance than the generalized synthetic images on MSMT17, i.e. 47.5% vs. 40.9% for rank-1 accuracy. We will release this toolkit for the ReID community to generate synthetic images at any desired taste. The code is available at https://github.com/tagperson/tagperson-blender.

CCS CONCEPTS

• **Information systems** → *Specialized information retrieval.*

KEYWORDS

Synthetic Datasets, Person Re-ID, Rendering

ACM Reference Format:
Kai Chen, et al.. 2022. TAGPerson: A Target-Aware Generation Pipeline for Person Re-identification. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22), October 10–14, 2022, Lisboa, Portugal.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3503161.3548013

1 INTRODUCTION

Person re-identification (ReID) technology is expected to determine if two persons across different views have the same identity. It plays an important role in intelligent applications [45]. Several large-scale annotated pedestrian datasets [40, 50] have been produced, and fully supervised methods for person ReID have achieved great success[27, 35]. However, with the increasing awareness of privacy protection, images containing human biology information are often along with ethical issues [41]. The identity labels are also difficult to label between cross-scene and cross-camera views, making the labeling cost very high [39].

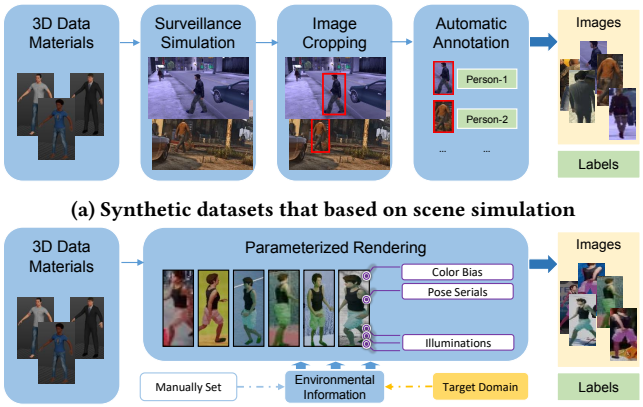
Several synthetic pedestrian datasets have been proposed to alleviate these problems [4, 5, 33, 39, 48]. They benefit from data diversification at low cost and the automatic generation of annotated labels. Recent works[33, 39, 48] are based on scene simulation. They construct virtual scenes and capture screenshots of the pedestrians who are walking around there. However, they suffer from the domain gap between virtual and real images. The scene-based rendering process is a snapshot sampling of the virtual scenes, and the synthesized images depend on the manually set scenes. The synthetic data can not be changed according to different target real scenes automatically. To make up for this shortcoming, this paper

*This work was supported by the National Natural Science Foundation of China (No. U1936202, 61925107, 62021002, 61971260, U21B2013) and the National Key R&D Program of China (No. 2020AAA0105500). This work is done during Kai Chen's internship at Alibaba Group. This work was supported by Alibaba Group through Alibaba Research Intern Program.

†Corresponding authors is Guiguang Ding and Weihua Chen .

MM '22, October 10–14, 2022, Lisboa, Portugal

ACM ISBN 978-1-4503-9203-7/22/10…$15.00
https://doi.org/10.1145/3503161.3548013



(a) Synthetic datasets that based on scene simulation

(b) TAGPerson pipeline

Figure 1: Comparison between the building process of current synthetic datasets and proposed TAGPerson. (a) A popular workflow of current synthetic datasets which are based on scene simulation. (b) Our proposed TAGPerson is based on parameterized rendering, and it can integrate information from the target domain during the rendering process.

proposes TAGPerson, a Target-Aware Generation (TAG) pipeline to generate the auto-labeling synthetic ReID datasets. Our method can integrate the target domain information during the rendering process to narrow the gap between synthetic and real data.

TAGPerson is composed of three stages. 3D data materials are firstly prepared to provide the basic person characters. Next, images are rendered under parameterized control by manual setting or the information of the target domain. It is achieved by parameter scope from prior knowledge or estimated distribution statistics of the target domain. The estimation models are trained to extract information from the target domain. The rendered images together with their labels serve as the training data to train the ReID model.

The difference between our workflow and the previous synthetic pipeline can be viewed in fig. 1. Compared to the previous synthetic pipelines, our method has three advantages. First, we integrate the target domain information during the rendering process. The rendering options can be constrained by the target domain, making the rendered images hold a smaller gap to the real images. Existing pipelines are oriented towards general scenarios. They do not take into account the utilization of possible available information about the target domain. Second, our parameterized rendering decouples different factors about the composition of the image, which reveals the most important environmental factors that affect the performance of the ReID model. Last, we can simulate more variables quantificationally like observation angle. This feature facilitates the rendering process to deal with some extreme scenarios.

TAGPerson has opened up a new path to achieve integrating information from the target domain. Our main contributions can be summarized as follows:

- We propose a novel target-aware person ReID dataset construction process. The auto-labeled synthetic images can be rendered under parameterized control guided by the target domain information. Based on this, many person Re-ID tasks can be resolved without using real datasets.
- We come up with a simple yet effective solution to extract the information from the target domain, and use it to guide the rendering process. Images are rendered in a target-aware manner, which significantly improves the generalization ability and robustness of person ReID models in real scenarios.
- Experiments are conducted to explore the effects of key factors during the rendering process. The gap of performance between the usage of synthetic and real images can be narrowed, which enhances the availability of ReID models in the limited scenario where there exist data access restrictions and privacy issues.

2 RELATED WORK

2.1 Person ReID Tasks

Person ReID task is welcomed since it can be used to solve realistic problems in surveillance scenarios. Metrics-based learning methods are proposed to measure the similarity of the query and the gallery [19, 34, 37]. Supervised person ReID has achieved great success in recent years [23, 27, 30, 38], but they get degraded performance when deployed into another scenario. Unsupervised domain adaptation methods [3, 17, 20, 22, 42, 43, 49] are studied to improve the effectiveness on the target domain. These methods can be split into two categories. Some of them use the basic model trained from the source data to assign pseudo labels for the target images, and then iteratively update the model weights and the assigned labels [9, 10]. Other solutions use GAN-based methods to transfer the source images into styles of the target domain [8, 21, 51, 52], which alleviates the gap between the two domains. The above methods rely heavily on labeled or unlabeled data, which might not be satisfied under the circumstances of data privacy protection. In recent years, researchers have resorted to synthetic methods to generate pedestrian datasets in the absence of real data.

2.2 Synthetic Datasets

In consideration of the problems about data privacy and ethical issues, many researchers resort to constructing synthetic pedestrian datasets for ReID task. SOMAset [5] and SyRI [4] open the door to use synthetic datasets in deep learning ReID models. PersonX [33] is a large-scale person dataset constructed by Unity3D, which explores the influence of pedestrian rotation angle. RandPerson [39] proposes a texture generation method to produce masses of virtual persons and establishes customized environments to simulate the surveillance scenes. Recently, UnrealPerson [48] analyzes the strategies during 3D human data generation. It proves that hard samples in training data are important to improve performance. UnrealPerson uses Unreal Engine 4 and UnrealCV [31] to simulate real-world scenes, and achieves excellent results in several kinds of person ReID tasks. [44] uses an attribute descent approach to let the synthesized images approximate the attributes in the real-world dataset, which is based on the prior knowledge of vehicle ReID.

Previous scenario-oriented synthetic datasets have domain gaps between synthetic and real images. To solve this problem, we construct our TAGPerson dataset in a target-aware manner by integrating the target domain information during the rendering process.

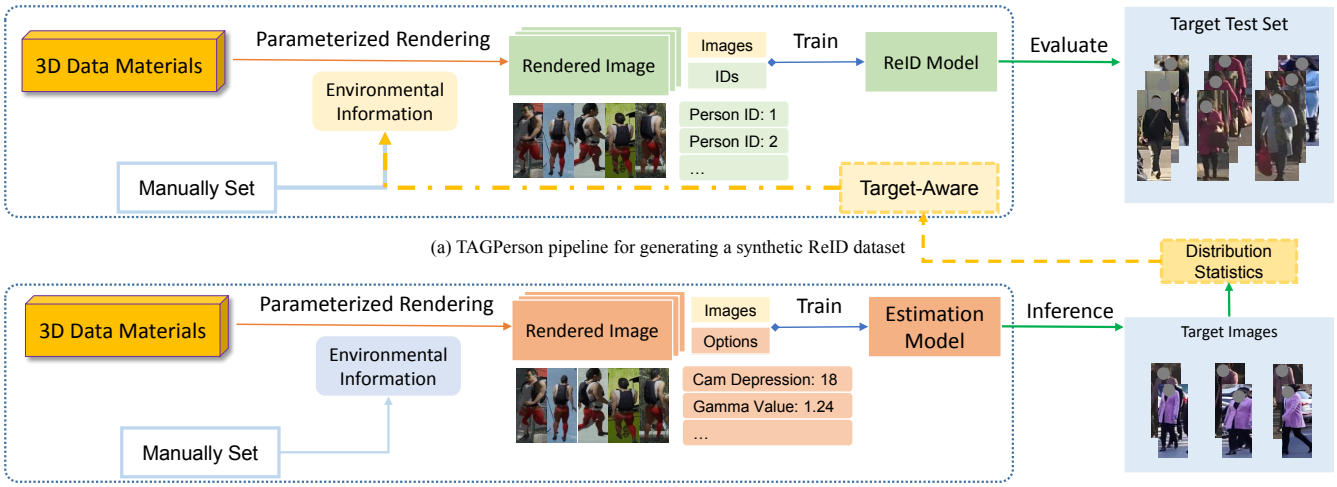


Figure 2: The illustration of the TAGPerson pipeline. Firstly, 3D data materials are prepared. Then the images are rendered under parameterized control. The rendered images together with their labels are used to train the ReID model or estimation models. (a) To generate a synthetic ReID dataset, the environmental information for parameterized rendering is set manually or from the target domain information. (b) The rendered images with their rendering option labels are used to train estimation models, which are used to extract information from the target images. The distribution statistics are passed to guide the rendering process and make it target-aware.

All identity-unrelated factors are constructed for a single target pedestrian, including different character poses, background images, camera perspectives, light conditions, etc. These rendering options can be under the guidance of the target domain information.

3 OUR METHOD

3.1 TAGPerson Pipeline

Existing synthetic person ReID datasets usually build virtual scenes and then capture person images. Different from them, we use a parameterized rendering process to directly generate images under specific rendering options. We avoid the process of scene simulation and data sampling, in order to obtain more control over the factors which dominate the quality of the person ReID dataset. Suppose the training dataset $S_{train} = \{\cup_{k=1}^{K}(I_k, y_k)\}$ consists of image I_k and its identity label y_k. From the parameterized perspective, the content of image data I_k can be decoupled into two parts: identity-related information and identity-unrelated information:

$$I_k = (P_k, \sum_{j=1}^{J} Fo_j) \tag{1}$$

where the P_k means the identity-related information for person k and Fo_j refers to the identity-unrelated environmental factor which is caused by the rendering option o_j. The rendering option o_j can be background, illumination, camera parameters, etc. That is what we can control dynamically during the rendering process.

Our proposed TAGPerson can be divided into three stages: prepare the 3D data materials, render the images, and train the ReID model. The overall pipeline can be viewed in fig. 2. During the construction, all the identity-related contents are determined by the 3D data, and other identity-unrelated contents are generated by parameterized rendering. The parameters can be set manually or from the target domain information. The rendered images with labels are used to train the ReID model or the estimation models.

As a prerequisite, the 3D data of humans are the fundamental content to build a person ReID dataset. They directly make up the appearances of humans. We control many factors to generate the various 3D data for one person, including its skin, face, height, obesity, muscle, etc. To distinguish one person from another, we dispatch different clothes, dresses, hairstyles, shoes, and optional accessories for each person. Next, the parameterized rendering process will be introduced in detail.

3.2 Parameterized Rendering

In previous works about synthetic datasets, the environmental information is determined by the virtual scenes. The rendering options are rarely mentioned. For the same 3D human data, different rendering options can produce different images. We dig into some rendering options that may affect the performance of the final model. The example rendered images can be seen in Fig. 3.

Rig Pose. We increase the variety of poses by introducing the change to the bones rig. Motion capture is the process of recording the movement of objects, and CMU Graphics Lab has published a free Motion Capture Database [11]. By applying the bvh files, we can make the person change its pose to a specific one, which can be an arbitrary moment of walking, standing, or others.

Camera Parameters. Existing methods use segmentation results to crop the pedestrian bounding boxes. We adopt another way by putting the target person as the anchor and setting different camera positions and orientations to capture the images.

Table 1: Detailed comparisons of synthesized datasets. TAGP-Base means the rendering options are set manually. TAGP-TA means some rendering options are rendered in a target-aware manner. We do not create virtual scenes and camera networks. By replacing background images and adding color bias, the count of our cameras can be regarded as infinite. The rank-1 accuracy on Market and MSMT17 datasets is the direct transfer performance of ReID models trained on the synthesized datasets.

Datasets	#Identities	#Cameras	#BBoxes	Parameterized Rendering	Scalabel	Rank-1 on Market	Rank-1 on MSMT
SyRI [4]	100	-	56,000	✗	✗	48.5%	21.8%
PersonX [33]	1266	6	273,456	✗	✓	58.7%	22.2%
RandPerson [39]	8000	19	228,655	✗	✓	64.7%	20.0%
UnrealPerson [48]	3000	34	120,000	✗	✓	79.0%	38.5%
TAGP-Base	2954	infinite	71,580	Manual	✓	79.9%	40.9%
TAGP-TA	2954	infinite	71,580	Target Aware	✓	81.6%	47.5%

Illumination. Illumination conditions affect the appearance of the person to a great extent. Previous works preset the lights in the virtual scenes, where each person's illumination condition is determined by its position. Instead, we change the position and intensity of the light source to produce controllable diverse illumination conditions from all angles for the target person.

Image Resolution. The distribution of image resolution is easily overlooked in previous datasets. If images have been resized to a fixed size, the intuitive representation is the degree of blurring of the images. In the real world, how far the person is from the camera directly determines the resolutions of the cropped images.

Background. The background is an important factor during the process of person ReID, which always contains a certain domain-specific message of the current dataset. Without creating a virtual scene, a trivial solution is to use images of diverse scenes as the background image to make the generated dataset robust to different complex scenarios. In our case, we use images from COCO [25] dataset as the background images. Each image is appropriately cropped according to the annotation of the person instance.

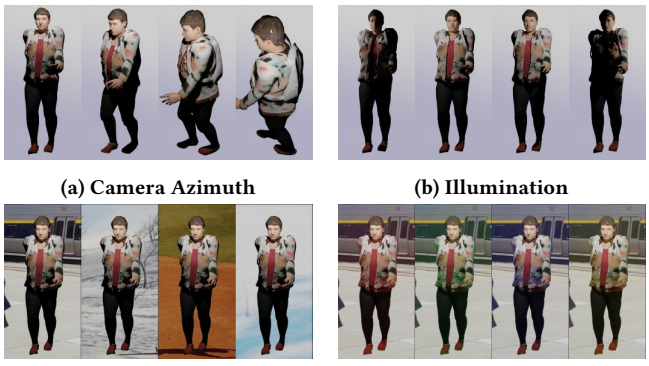
Color Bias. To make features distinguishable from the global perspective, many methods have attempted to eliminate the existing camera bias in images. We resolve the problem of camera style differences by simulating images with color bias quantitatively. We use a simple yet efficient strategy to add specific color bias to images according to the camera label currently assigned.

3.3 Target-Aware Generation

We intend to integrate the target domain information during the rendering process, towards making the rendered images more inclined to the style of the specific target domain.

We extract the target-aware information from the target domain, and then apply it in reverse to control the rendering options. We use $O = o_1, o_2, ..., o_J$ to represent the rendering options. For example, o_j can be the camera depression angle towards the person. We use $D^t(O)$ to describe the distribution of factors O on the target domain, and $D^s(O)$ as the one on the current synthetic domain. The Wasserstein Distance [36] can be used to measure the discrepancy between $D^s(O)$ and $D^t(O)$ as $W(D^s(O), D^t(O))$, and our goal is to minimize the discrepancy.

$$\min_{D^s(O)} W(D^s(O), D^t(O)) \tag{2}$$



(a) Camera Azimuth (b) Illumination (c) Background (d) Color Bias

Figure 3: Visualizations of the effects of some rendering options. (a) Different camera azimuths. (b) Different light conditions. (c) Changing background images. (d) Adding color bias. All the rendering options can be parametrically controlled, and a pedestrian individual can be rendered from all aspects.

We design a mechanism to automatically fetch the target-specific information. First, by manually setting the rendering options, we generate masses of images with their rendering option labels. These data are used to train an estimation model M_{r_j} which can be used to inference the value of rendering option o_j for an input image I_i^t:

$$\tilde{o_{ij}} = M_{r_j}(I_i^t) \tag{3}$$

Here I_i^t means the ith image of the target domain and $\tilde{o_{ij}}$ is the output value. We use all the available images of the target domain as the input fed to the model. The values of the inference output are collected to form distribution statistics.

$$D^t(o_j) \sim \sum_{i=1}^{N_t} \tilde{o_{ij}} \tag{4}$$

$$D^t(O) = \cup_{j=1}^{J} D^t(o_j) \tag{5}$$

Here N_t is the total number of collected target images, and $D^t(o_j)$ represents the distribution of factor o_j on the target domain. J is the number of rendering options. We fetch the distribution statistics for all J necessary rendering options to generate the distribution

Table 2: Detailed comparisons of rendering options. The check symbol means that the rendering option in the column has proper multiple values, otherwise it is set to a default value. Various background images are necessary and different resolutions are beneficial to ReID performance. Multiple illumination conditions and diversified poses help to improve the effect. Different camera depression angles can promote the performance. Slightly adding color bias makes the model more robust.

Background	Resolution	Illumination	Pose	Camera Depression Angle	Color Bias	Market		MSMT	
						R1	mAP	R1	mAP
-	-	-	-	-	-	22.1	8.6	3.9	1.1
✓	-	-	-	-	-	65.8	38.5	29.1	9.2
✓	✓	-	-	-	-	74.5	47.0	31.2	9.9
✓	✓	✓	-	-	-	75.2	47.1	34.0	11.2
✓	✓	✓	✓	-	-	76.1	49.4	35.0	12.0
✓	✓	✓	✓	✓	-	78.7	54.0	36.2	12.8
✓	✓	✓	✓	✓	✓	79.9	53.1	40.9	14.3

statistics $D^t(O)$. It is used as the target-aware information to guide the parameterized rendering process.

During the process of synthetic dataset generation, the rendering options are constrained by the above distribution $D^t(o)$ to fit the style of the target domain. We use a simple yet effective Monte Carlo [13] sampling method to ensure the limitation of $D^s(O)$. To be precise, the values of rendering options are randomly sampled from the distribution statistics $D^t(o)$ with equal probability. This strategy makes the images rendered in a target-aware manner, and the rendered images are closer to the target domain in terms of the distribution of corresponding environmental factors. We assumed that these rendering options are less affected by the cross-domain problem compared to identification labels during the estimation model training, and the estimation error can be inessential.

Note that although we need the distribution statistic information of target image, our approach does NOT require target images during training. In real scenarios where the target data is not available for training, we can deliver the extraction methods to the data owner, and they give back the statistics extracted from the data for our training rather than the whole original data itself. This process can avoid data privacy issues.

4 EXPERIMENTS

In this section, we introduce our implementation details and the experimental results. The TAGPerson dataset is generated by manually set or target-aware rendering options. The overview of generated TAGPerson dataset can be seen in table 1. Some experiments are conducted to explore the importance of rendering options and demonstrate the effectiveness of our target-aware mechanism.

4.1 Implementation Details

During the data generation, we utilize the MakeHuman [2] Python API to generate thousands of 3D data of humans. We render images by Blender Python Library [1]. In the training stage, We use the Fastreid [16], a toolkit based on PyTorch [29], as the basic training framework. We use ResNet-50 [15] structure as the backbone, which is pre-trained on ImageNet [7]. We train the model based on the labeled synthetic images with cross-entropy loss and triplet loss, and then directly evaluate it on the target datasets. The input images are resized to 256×128. We use SGD as the optimizer with a momentum of 0.9 and weight decay of 0.0005. We choose ColorJitter and AugMix [18] as data augmentation. More experimental details can be found in our source code.

Table 3: Direct transfer performance of some real datasets and synthetic datasets.

Source	Training Data	Market		MSMT	
		R1	mAP	R1	mAP
Real	Market	94.7	86.2	25.7	9.6
	MSMT17	74.4	45.4	74.7	50.5
Synthetic	SyRI	48.5	22.6	21.8	5.7
	PersonX	58.7	32.7	22.2	7.9
	RandPerson	64.7	39.3	20.0	6.8
	UnrealPerson	79.0	54.3	38.5	15.3
Synthetic	TAGP-Base	79.9	53.1	40.9	14.3
	TAGP-TA	**81.6**	**54.8**	**47.5**	**17.7**

4.2 Direct Transfer Evaluation

Direct transfer evaluation is the setting closest to the actual application scenario where the target domain is unavailable during training. To prove the validity and practicality, we apply direct transfer evaluation for the model trained by our proposed TAGPerson dataset on two real-world person ReID datasets. Market-1501 [50], MSMT17 [40] are used as the test sets. The direct transfer performance of TAGPerson and other datasets is compared in table 3.

For short we use Market and MSMT to represent Market-1501 and MSMT17. TAGP-Base represents that the rendering options are manually set and TAGP-TA represents that the rendering options are guided by the target domain information. From the table, we can see that without relying on the construction of multiple virtual scenes, TAGP-TA can achieve competitive performance compared to the state-of-the-art method in synthetic datasets. The performance on the Market dataset also surpasses the performance of the large-scale real-world dataset MSMT. Specifically, we boost the rank-1 accuracy and mAP on Market to 81.6% and 54.8%, respectively. On the MSMT dataset, our TAGP-TA boosts the rank-1 accuracy and mAP to 47.5% and 17.7%, surpassing all previous synthetic datasets.

4.3 Ablation Study on Rendering Options

Rendering options have significant impacts on the performance of the ReID model. We explore the effects of several rendering options and find the most important factors that contribute to good synthetic images. The results can be viewed in table 2.

The experimental results show that complex background information is the most important necessity. By using the background images cropped from the COCO [22] dataset, the performance has a huge improvement for rank-1 accuracy from 3.9% to 29.0% on the MSMT dataset, compared to the case of using empty background. Different sizes of the resolution also bring large improvements to the performance, especially on the Market dataset. The mAP on the Market dataset is boosted from 38.5% to 47.2% because there are many blurred images there. Multiple illuminations have positive effects, especially on the MSMT dataset where some images are under extreme lighting conditions. Applying various poses to the person increases the diversity and improves the performance, which is reasonable. Multiple camera depression angles can deal with the situation when the heights of cameras are different, and they have positive effects when introduced. The improvement is not obvious since the range is set manually, without considering specific scenarios. Adding the color bias factor boosts the rank-1 accuracy and mAP on the MSMT dataset over 4.7% and 1.5% respectively.

4.4 Ablation Study on Target-Aware Rendering

Camera and illumination are important factors in person ReID task [26, 28, 51]. However, from the experiments in table 2, we find that adding multiple camera parameters and illumination conditions does not bring obvious improvement. We suppose that these factors are strongly correlated with the dataset. they differ in different datasets and it is hard to control the value range manually. That hinders the improvement of the performance on real-world datasets.

To solve this problem, we conduct experiments to demonstrate the effectiveness of our proposed target-aware generation for the rendering process. We adopt camera depression angle and gamma value as the representatives of target-aware information to control the rendering options. Synthetic datasets with different rendering options are generated. We compare the performance of ReID models trained based on them.

4.4.1 Camera Depression angle. Few studies have looked at the effect of camera depression angle. We find that in the actual scene, there is a big difference in the heights of the installation position of the cameras, which has a great impact on the appearances of persons. For example, most of the camera shots of the MSMT dataset are in a flat direction, while the average value of camera depression angles in the GRID [26] dataset is larger. We train an estimation model for camera depression angle by the target-aware information extraction method. The model is used to estimate the depression angle value of each image in three datasets, and the distribution statistics are visualized in fig. 4a.

We compare the results of rendered TAGPerson with or without target domain information about the camera depression angle in table 4. The column TAGP-C represents the rendering option value of camera depression angle is target-aware. We can find that compared to the basic setting if the distribution of the camera depression angle coincides with the specific domain, the performance of the trained ReID model can be improved. For example, the images of dataset GRID are captured underground, and the depression angle of the camera is obviously larger. The performance of TAGP-C rendered in a target-aware manner for GRID outperforms the basic method by 37.2% vs. 28.8% and 45.2% vs. 38.2% for rank-1 accuracy and mAP, respectively.

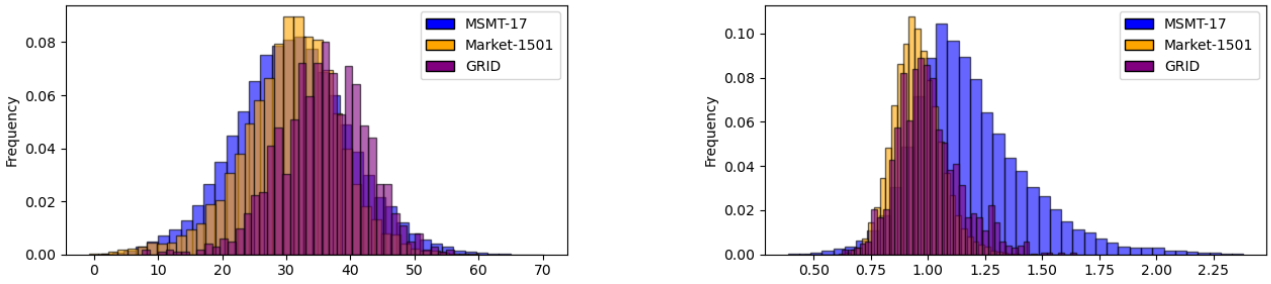
Table 4: Direct transfer performance on the Market, MSMT, and GRID datasets. The first row without checkmark symbols in TAGP-C and TAGP-G columns represents the manual setting for TAGPerson. TAGP-C and TAGP-G represent that the rendering option about camera depression angle and gamma value are guided by the target domain information, respectively. The last row means that both the two options are controlled by the target domain information.

TAGP-C	TAGP-G	Market		MSMT		GRID	
		R1	mAP	R1	mAP	R1	mAP
		79.9	53.1	40.9	14.3	28.8	38.2
✓		81.2	54.5	46.2	17.2	37.2	45.2
	✓	81.3	54.0	46.3	17.1	38.0	45.9
✓	✓	**81.6**	**54.8**	**47.5**	**17.7**	**38.8**	**47.3**

4.4.2 Gamma Value. Gamma correction is a nonlinear operation used to encode and decode luminance or tristimulus values in video or still image systems. For devices with different gamma correction settings, the captured images may appear in different brightness. The images captured by one camera may also show a wide range of brightness as the light changes from morning to evening. There are differences in gamma value parameters between different datasets because of the camera devices and acquisition time. The different distribution statistics of gamma value can be viewed in fig. 4b. For different distribution statistics, we construct the TAGPerson dataset in a target-aware manner towards the gamma value, which is constrained by information extracted from the target domain and controls the range of the rendering option values.

From table 4 we can also find that the performance of the ReID model can be improved by integrating the gamma information of the specific domains. The improvements to MSMT and GRID datasets are obvious. For the MSMT dataset, the rank-1 accuracy is improved from 40.9% to 46.3%. For the GRID dataset, the target-aware setting can obtain a 7.7% mAP improvement compared to the manual setting. That makes sense because the images of the GRID dataset are captured in an underground scene and the illumination condition is poor and varies a lot. Meanwhile, many images in the MSMT dataset are taken in the situation of reverse light. Thus the acquisition of the gamma value from the target images can be useful to narrow the domain gap.

4.4.3 Gamma Value for Extreme Scenarios. Integrating gamma value has been proved to have a big boost for extreme scenarios like the GRID dataset. To explore the effects more deeply, we conducted extension experiments. Due to the lack of similar large ReID datasets in extreme scenarios, we decide to simulate possible actual scenarios where the gamma values are changed. We create the variant datasets of gamma by applying different gamma values to the original images. For the Market dataset, we create Market-G1



(a) Camera Depression Angle **(b) Gamma Value**

Figure 4: Visualization of the distribution statistics of the camera depression angle and gamma value in the Market-1501, MSMT17, and GRID datasets. From the statistical histogram we can see that, the MSMT17 has a wider range of depression angles than Market-1501. The average value of the camera depression angles in the GRID dataset is larger.

(a) The mAP(%) performance on Market-variant datasets. **(b) The mAP(%) performance on MSMT-variant datasets**

Figure 5: The mAP performance of models which are trained from different training data is drawn in different colors. TAGP-Base represents images rendered by manual setting and TAGP-G represents images rendered according to the gamma value distribution. (a) Test sets are Market variant datasets. (b) Test sets are MSMT variant datasets. Market-Base and MSMT-Base represent the original Market and MSMT datasets. When the gamma values are disturbed, the model trained from supervised learning and UnrealPerson datasets deteriorate significantly. TAGP-Base also drops slightly. The model trained from the TAGP-G dataset which is rendered in a target-aware manner keeps a steady performance on different target domains.

datasets where the gamma values are randomized from 0.5 to 1.5, and Market-G2 datasets where the gamma values are randomized from 0.5 to 2.0. The same operation is performed for the MSMT dataset and we can get MSMT-G1 and MSMT-G2 variant datasets.

fig. 5a shows the performance of different training datasets on these test sets. TAGP-Base represents the TAGPerson dataset rendered under manually set parameters. TAGP-G represents the TAGPerson dataset rendered in a target-aware manner about the gamma value. We can see that, with only a little gamma jitter on the Market dataset, the mAP of the model trained from supervised learning significantly drops from 86.2% to 43.0%. The performance of synthetic datasets like UnrealPerson or TAGP-Base also declines seriously. It seems that the gamma discrepancy introduces a large domain shift. The TAGP-G dataset is rendered in a target-aware manner by integrating the distribution statistics of the target domain, so it can alleviate the problem to some extent and achieve better performance in its corresponding case. Compared to the dataset generated with manually set rendering options, the target-aware one gains 4.5% mAP improvement on the Market-G2 dataset. The results of TAGP-G on Market-Base, Market-G1, and Market-G2 are not from the same model. There are three target domains and the TAGP-G serials are de facto three models on the target domain, respectively.

The same phenomenon can be observed for the MSMT dataset. The changes in gamma value significantly affect the performance of the ReID model. The results can be seen in fig. 5b. By rendering images in a target-aware manner, the TAGP-G obtain 17.1%, 16.4%, and 14.6% mAP on MSMT-Base, MSMT-G1, and MSMT-G2 datasets, respectively. Compared to using the manual setting, the TAGP-G reduces the model deterioration in extreme scenarios.

Table 5: Results of how much the number of target images affects the quality of target-aware rendering.

Image number	Market		MSMT	
	R1	mAP	R1	mAP
0%(TAGP-Base)	79.9	53.1	40.9	14.3
1%	81.5	54.5	46.0	17.2
10%	80.6	54.0	46.5	17.4
50%	81.5	54.6	46.5	17.2
100%(TAGP-TA)	**81.6**	**54.8**	**47.5**	**17.7**

Table 6: Results of Domain Adaptation (DA) experiments based on Multi-view Evolutionary Training (MET)[12] and supervised fine-tuning based on Pretraining.

Source Domain	Methods	Market		MSMT17	
		R1	mAP	R1	mAP
Unreal	UDA (MET)	92.0	81.0	71.6	42.2
TAGP-Base		92.5	80.9	71.7	43.0
TAGP-TA		**92.6**	**81.3**	**72.7**	**43.9**
Unreal	Pretraining and Finetuning	94.0	84.7	74.5	46.0
ImageNet		94.4	86.1	74.1	50.2
TAGP-Base		94.8	**87.0**	75.3	50.7
TAGP-TA		**94.8**	86.9	**75.7**	**50.7**

4.5 Effect of the Amount of Target Images

Our target-aware information comes from the statistics of the target domain. In previous experiments, the statistics are abstracted from the output of the estimation models on target images. Characterizing the distribution of a rendering factor does not require all images, and we conduct experiments to explore how much the number of target images affects the quality of target-aware rendering.

The results are shown in table 5. The conclusion is that only 1% (around 100+) images are necessary to provide the target-ware information, since these samples are enough to define the distribution. Compared with the UDA methods or GAN-based methods which require a large number of target images to achieve domain adaptation, our target-aware information only needs a small part of the images of the target domain. This significantly reduces the risk of data privacy concerns when the target domain contains sensitive information, like the human face and body.

4.6 UDA and Pretraining Evaluation

We have also conducted experiments for downstream tasks, including Unsupervised Domain Adaptation (UDA) methods and Pre-training settings. The results can be viewed in table 6. The TAGP-TA outperforms current optimal synthetic datasets UnrealPerson in UDA methods and Pre-training settings, but there is only a slight advantage. From the results, we can find that although our target-aware rendering has the overlap effect as DA methods, it can be used as a complementary approach to achieve domain adaptation effect beyond the common UDA methods. When they are used together, the target-aware rendering strategy can achieve slight improvement.

4.7 Discussion about Privacy Issues

We should declare that our research intends to resolve the tasks at a lower risk of privacy issues. Although we use the information from the target domain as the current UDA methods, the difference lies in the following points. Firstly, our target-aware rendering only requires statistics from a small quantity instead of multiple images from the target domain. Second, the target-aware information is not involved in the training procedure. It is just used for rendering and it is more difficult to decode the target information from the model. Nowadays, some transfer learning methods can use the pre-trained model to infer the original images, and UDA models are at risk of being decoded out the human body images [6, 14, 24, 32, 46, 47]. Finally, on the other side of data privacy, technology privacy is also important. UDA methods may have to be provided to the client when the training data is private. Our estimation methods can be delivered to the data owner, and they give back the statistics extracted from the data. Then we can train the ReID model without releasing the training technology and solution.

4.8 Limitation

One of the drawbacks of our TAGPerson is that there is a natural lack of situations of occlusion and multiple persons because the rendering process is person-centered. We have tried to deliberately add the occlusion by using annotations from COCO dataset [25]. With the instance segmentation annotations, the rendered images have a realistic occlusion effect. We also created the images containing multiple persons by placing another person nearby, and they can be rendered to be like partners. However, adding images of these two scenarios has not obtained improvements.

Broader Impact. Person ReID technology may inevitably infringe on the privacy of pedestrians. Our work attempts to reduce this infringement from two aspects. In the pre-training stage, the synthetic TAGPerson dataset can be used to replace real datasets, e.g., DukeMTMC-ReID, which has been taken down due to ethics issues. In terms of target domain information utilization, we use statistical information rather than raw data to avoid accessing images directly. However, real images from surveillance data are necessary for the test stage when applying the model to actual scenarios. This may potentially raise privacy issues because not all humans know and permit that they are being recorded. We urge that users should follow strict regulations and laws to use the person ReID models.

5 CONCLUSION

In this paper, we propose a target-aware generation pipeline named TAGPerson to resolve the person ReID task. Without establishing complex virtual scenes, we can directly render person images under desired parameters, to serve as an effective training dataset. If the information of the target domain can be extracted, we can render the images in a target-aware manner by integrating the target domain information to guide the rendering options. This novel idea explores a new path to utilize the target domain when the images can not be accessed directly. TAGPerson provides a strategy to estimate environmental factors from the target images and an effective way to minimize the gap between synthesized datasets and real-world scenarios. In the future, we will study how to mine the potential key factors behind the rendering procedure.

REFERENCES

[1] 2021. https://www.blender.org/. https://www.blender.org/.

[2] 2021. MakeHuman Community. http://www.makehumancommunity.org/.

[3] Zechen Bai, Zhigang Wang, Jian Wang, Di Hu, and Errui Ding. 2021. Unsupervised Multi-Source Domain Adaptation for Person Re-Identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12914–12923.

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

[5] Igor Barros Barbosa, Marco Cristani, Barbara Caputo, Aleksander Rognhaugen, and Theoharis Theoharis. 2018. Looking beyond appearances: Synthetic training data for deep cnns in re-identification. *Computer Vision and Image Understanding* 167 (2018), 50–62.

[6] Hanting Chen, Yunhe Wang, Chang Xu, Zhaohui Yang, Chuanjian Liu, Boxin Shi, Chunjing Xu, Chao Xu, and Qi Tian. 2019. Data-free learning of student networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3514–3522.

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

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

[9] Yixiao Ge, Dapeng Chen, and Hongsheng Li. 2019. Mutual Mean-Teaching: Pseudo Label Refinery for Unsupervised Domain Adaptation on Person Re-identification. In *International Conference on Learning Representations*.

[10] Yixiao Ge, Feng Zhu, Dapeng Chen, Rui Zhao, et al. 2020. Self-paced contrastive learning with hybrid memory for domain adaptive object re-id. *Advances in Neural Information Processing Systems* 33 (2020), 11309–11321.

[11] Ralph Gross and Jianbo Shi. 2001. The cmu motion of body (mobo) database. (2001).

[12] Jianyang Gu, Weihua Chen, Hao Luo, Fan Wang, Hao Li, Wei Jiang, and Weijie Mao. 2022. Multi-view Evolutionary Training for Unsupervised Domain Adaptive Person Re-Identification. *IEEE Transactions on Information Forensics and Security* (2022).

[13] John Hammersley. 2013. *Monte carlo methods*. Springer Science & Business Media.

[14] Matan Haroush, Itay Hubara, Elad Hoffer, and Daniel Soudry. 2020. The knowledge within: Methods for data-free model compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8494–8502.

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

[16] Lingxiao He, Xingyu Liao, Wu Liu, Xinchen Liu, Peng Cheng, and Tao Mei. 2020. Fastreid: A pytorch toolbox for general instance re-identification. *arXiv preprint arXiv:2006.02631* (2020).

[17] Tao He, Leqi Shen, Yuchen Guo, Guiguang Ding, and Zhenhua Guo. 2022. SECRET: Self-Consistent Pseudo Label Refinement for Unsupervised Domain Adaptive Person Re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 879–887.

[18] Dan Hendrycks, Norman Mu, Ekin Dogus Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. 2019. AugMix: A Simple Data Processing Method to Improve Robustness and Uncertainty. In *International Conference on Learning Representations*.

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

[20] Xin Jin, Cuiling Lan, Wenjun Zeng, and Zhibo Chen. 2020. Global distance-distributions separation for unsupervised person re-identification. In *European Conference on Computer Vision*. Springer, 735–751.

[21] Xin Jin, Cuiling Lan, Wenjun Zeng, Zhibo Chen, and Li Zhang. 2020. Style normalization and restitution for generalizable person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3143–3152.

[22] Jianing Li and Shiliang Zhang. 2020. Joint visual and temporal consistency for unsupervised domain adaptive person re-identification. In *European Conference on Computer Vision*. Springer, 483–499.

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

[24] Yuhang Li, Feng Zhu, Ruihao Gong, Mingzhu Shen, Xin Dong, Fengwei Yu, Shaoqing Lu, and Shi Gu. 2021. Mixmix: All you need for data-free compression are feature and data mixing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4410–4419.

[25] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*. Springer, 740–755.

[26] Chen Change Loy, Tao Xiang, and Shaogang Gong. 2009. Multi-camera activity correlation analysis. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 1988–1995.

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

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

[29] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems* 32 (2019), 8026–8037.

[30] Angelo Porrello, Luca Bergamini, and Simone Calderara. 2020. Robust re-identification by multiple views knowledge distillation. In *European Conference on Computer Vision*. Springer, 93–110.

[31] Weichao Qiu and Alan Yuille. 2016. Unrealcv: Connecting computer vision to unreal engine. In *European Conference on Computer Vision*. Springer, 909–916.

[32] James Smith, Yen-Chang Hsu, Jonathan Balloch, Yilin Shen, Hongxia Jin, and Zsolt Kira. 2021. Always be dreaming: A new approach for data-free class-incremental learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 9374–9384.

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

[34] Yifan Sun, Changmao Cheng, Yuhan Zhang, Chi Zhang, Liang Zheng, Zhongdao Wang, and Yichen Wei. 2020. Circle loss: A unified perspective of pair similarity optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6398–6407.

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

[36] SS Vallender. 1974. Calculation of the Wasserstein distance between probability distributions on the line. *Theory of Probability & Its Applications* 18, 4 (1974), 784–786.

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

[38] Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi Zhou. 2018. Learning discriminative features with multiple granularities for person re-identification. In *Proceedings of the 26th ACM international conference on Multimedia*. 274–282.

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

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

[41] Guile Wu and Shaogang Gong. 2021. Decentralised Learning from Independent Multi-Domain Labels for Person Re-Identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 2898–2906.

[42] Shiyu Xuan and Shiliang Zhang. 2021. Intra-Inter Camera Similarity for Unsupervised Person Re-Identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11926–11935.

[43] Fengxiang Yang, Zhun Zhong, Zhiming Luo, Yuanzheng Cai, Yaojin Lin, Shaozi Li, and Nicu Sebe. 2021. Joint Noise-Tolerant Learning and Meta Camera Shift Adaptation for Unsupervised Person Re-Identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4855–4864.

[44] Yue Yao, Liang Zheng, Xiaodong Yang, Milind Naphade, and Tom Gedeon. 2020. Simulating content consistent vehicle datasets with attribute descent. In *European Conference on Computer Vision*. Springer, 775–791.

[45] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. 2021. Deep learning for person re-identification: A survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021).

[46] Hongxu Yin, Pavlo Molchanov, Jose M Alvarez, Zhizhong Li, Arun Mallya, Derek Hoiem, Niraj K Jha, and Jan Kautz. 2020. Dreaming to distill: Data-free knowledge transfer via deepinversion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8715–8724.

[47] Jaemin Yoo, Minyong Cho, Taebum Kim, and U Kang. 2019. Knowledge extraction with no observable data. *Advances in Neural Information Processing Systems* 32 (2019).

[48] Tianyu Zhang, Lingxi Xie, Longhui Wei, Zijie Zhuang, Yongfei Zhang, Bo Li, and Qi Tian. 2021. UnrealPerson: An Adaptive Pipeline towards Costless Person Re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11506–11515.

[49] Kecheng Zheng, Wu Liu, Lingxiao He, Tao Mei, Jiebo Luo, and Zheng-Jun Zha. 2021. Group-aware label transfer for domain adaptive person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5310–5319.

[50] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. 2015. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision*. 1116–1124.
[51] Zhun Zhong, Liang Zheng, Zhedong Zheng, Shaozi Li, and Yi Yang. 2018. Camera style adaptation for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5157–5166.
[52] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. In *Computer Vision (ICCV), 2017 IEEE International Conference on*.

Algorithm 1 Parameterized Rendering in TAGPerson

Input:

- 3D human data: human model files and texture files
- Auxiliary media: background, motion capture, etc.
- Parameterized assignment configure

Output:

- Rendered images
- Corresponding ID labels, rendering option labels

Initialize rendering environment.
Load auxiliary media resources.
for $mesh_i$ in $mesh\_list$ **do**
    Load human 3d model from $mesh_i$
    Generate rendering options
    if Target-aware information exists **then**
        Restrain rendering option values
    Render the images for $person_i$
    Save ID labels and rendering option labels
return Images and corresponding labels

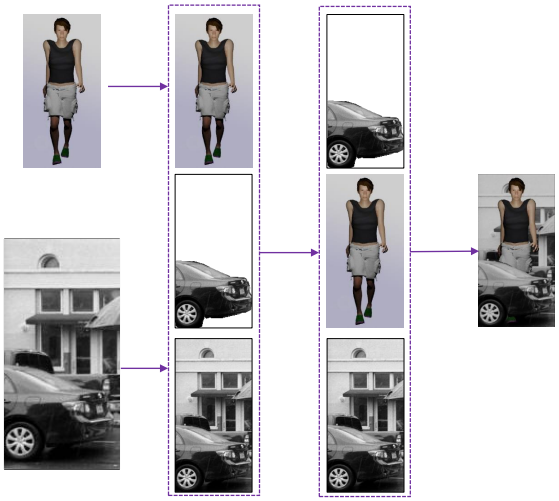


Figure 6: The demonstration of making shield cases.

A APPENDIX

In the Appendix section, we introduce more details about our TAGPerson generation process.

A.1 3D Human Models

We briefly introduce how to generate the 3D human models for rendering. Rendering or image synthesis is the process of generating a photorealistic or non-photorealistic image from a 2D or 3D model. To generate a person ReID dataset, we need to prepare the basic 3D human models firstly. The 3D human model consists of two parts: the structure part and the texture part, which define the shape information and appearance information, respectively.

MakeHuman [2] is an open-source toolkit for creating 3D data of human bodies. It supports setting the basic attributes of the human model like skin, age, weight, height, hair, and so on. It also supports equipping the human model with extra objects such as clothes, shoes, and accessories. In the MakeHuman community, many free assets have been provided. Besides, it provides plugins for Blender [1] to load the exported human models.

A.2 Parameterized Rendering

The parameterized rendering process in the proposed TAGPerson is based on the Blender Python Library [1]. Here we dig into some details about the rendering process.

A.2.1 Description for the Process. We describe the workflow of our parameterized rendering process in algorithm 1. Firstly, we initialize the rendering environment for the Blender toolkit and load some auxiliary media resources, such as the background images and the motion capture files. Since our method does not need to construct a virtual scene, we direct render the images for all persons in the 3d human data list. For each person, we first generate the rendering options for it according to the given configuration. If the target-aware information can be fetched, we use this guidance to restrain the value of relevant option values. Lastly, the images are rendered according to the rendering options, and corresponding identity labels and option labels are saved.

A.2.2 More Description of Rendering Options. The description for part of the rendering options are listed in table 7. We attempt to change as many identity-unrelated factors as possible in person ReID datasets.

A.2.3 Logic of Background Image Generation. We directly render the image without using scenario simulation in order to remove the interference of irrelevant factors introduced by scenario construction, pedestrian walking, image cropping, and so on.

We use MS COCO [25] which contains instance annotation information. We crop the area around the pedestrian instance as a background image candidate. The illustration for logic of background image generation can be seen in fig. 7. First, we collect all images containing person instances in the COCO dataset. For each person instance, we check whether its left or right side has overlapped persons. If not, we crop it out as the background image candidates, otherwise, it is dropped.

We arrange the cropped images as the basic auxiliary media data for the rendering process. We can see that, without creating complex virtual scenarios, we can obtain comparable performance with the state-of-the-art method.

A.2.4 Occlusion Production. One of the shortcomings of TAGPerson is the lack of occlusion cases in the rendered images. Since there is no virtual scene construction, pedestrians will not be blocked by any object. We attempt to create these cases by using object detection datasets. Suppose there is an image I_i with object segmentation annotation. we split the image I_i into two parts: the foreground part $Front_i$ and background parts $Back_i$. By rendering the separate parts $Back_i$ and $Front_i$ orderly, we can rebuild the image I_i into its original appearance. On this basis, we insert the person subject $Person_k$ between $Back_i$ and $Fore_i$ to create a blocking effect.

$$I_i = (Back_i, Person_k, Fore_i) \tag{6}$$

Table 7: The description for part of the rendering options.

Module	Option Name	Meaning Description	Data Type	Remark
Background	Background Image	Background image when rendered	Enum	Preset candidates
Pose	Pose serial ID	Motion capture applied to the bones rig	Enum	Preset candidates
Camera	Camera Azimuth	Azimuth of camera	Integer	[0, 359]
	Camera Depression Angle	Depression angle of camera	Integer	[0, 90]
	Camera Distance	Distance from camera to human model	Float	$[0, +\infty]$
	Camera Rotation X	Rotation euler x value of camera	Float	$[-\pi, \pi]$
	Camera Rotation Y	Rotation euler y value of camera	Float	$[-\pi, \pi]$
	Camera Rotation Z	Rotation euler z value of camera	Float	$[-\pi, \pi]$
Light	Light Azimuth	Azimuth of light	Integer	[0, 359]
	Light Depression Angle	Depression angle of light	Integer	[0, 90]
	Light Distance	Distance from light to human model	Float	$[0, +\infty]$
Image	Resolution X	Width of rendered image	Integer	$[0, +\infty]$
	Resolution Y	Height of rendered image	Integer	$[0, +\infty]$
Composite	Gamma Value	Color Correction Value	Float	[0, 10]
	Color Bias	Color bias on human model	RGBA	(R, G, B, A)
	Shield	Shield objects blocking human model	Enum	Preset candidates

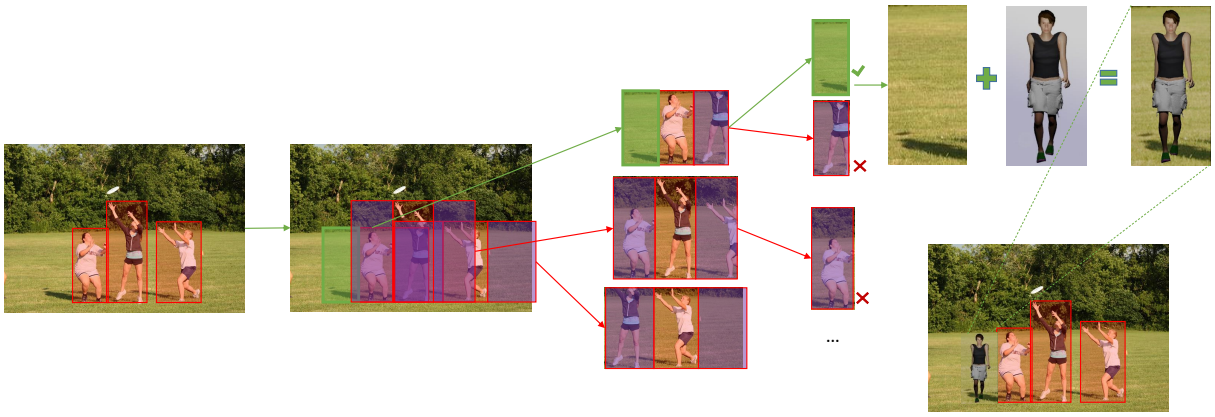


Figure 7: Illustration of logic about background image generation. We crop proper bounding boxes from images in the COCO dataset with its annotation information. For each person instance, we check whether its left or right side has overlapped persons. If not, we crop it out as the background image candidates, otherwise, it is dropped.

Table 8: Results of applying data augmentation and gamma-aware settings on the MSMT17 dataset. Gamma-aware represents the images rendered under the guidance of the gamma value information. Brightness and Contrast represent data augmentation operations for brightness and contrast, respectively.

Gamma-aware	Augmentation		MSMT17	
	Brightness	Contrast	R1	mAP
✓	✓	✓	**46.3**	**17.1**
✓	✓	-	43.0	15.7
✓	-	✓	44.0	16.0
✓	-	-	38.7	13.6
-	✓	✓	40.9	14.3
-	✓	-	35.8	12.5
-	-	✓	37.9	12.9
-	-	-	30.7	10.3

The illustration can be seen in fig. 6. However, we could not obtain improvement by adding occlusion cases. We suspect that this composite strategy can not simulate enough diverse cases.

A.2.5 Gamma vs. Data Augmentation. Gamma value is used to correct the color balance of the image. It changes the appearance of the image like brightness and contrast. Whether the effect of gamma adaptive can be replaced by data augmentation? From the table 8 we can see that the gamma value and data augmentation are both necessary. Removing any of them causes the degradation of the performance. Thus the gamma value guided by the target domain is not in conflict with the effect of data augmentation.