

# Dissecting Person Re-identification from the Viewpoint of Viewpoint

Xiaoxiao Sun Liang Zheng  
Australian National University

xxsunzrt@gmail.com liang.zheng@anu.edu.au

## Abstract

Variations in visual factors such as viewpoint, pose, illumination and background, are usually viewed as important challenges in person re-identification (re-ID). In spite of acknowledging these factors to be influential, quantitative studies on how they affect a re-ID system are still lacking. To derive insights in this scientific campaign, this paper makes an early attempt in studying a particular factor, viewpoint. We narrow the viewpoint problem down to the **pedestrian rotation angle** to obtain focused conclusions. In this regard, this paper makes two contributions to the community. First, we introduce a large-scale synthetic data engine, PersonX. Composed of hand-crafted 3D person models, the salient characteristic of this engine is “controllable”. That is, we are able to synthesize pedestrians by setting the visual variables to arbitrary values. Second, on the 3D data engine, we quantitatively analyze the influence of pedestrian rotation angle on re-ID accuracy. Comprehensively, the person rotation angles are precisely customized from  $0^\circ$  to  $360^\circ$ , allowing us to investigate its effect on the training, query, and gallery sets. Extensive experiment helps us have a deeper understanding of the fundamental problems in person re-ID. Our research also provides useful insights for dataset building and future practical usage, e.g., a person of a side view makes a better query.

## 1. Introduction

Viewpoint, pose of person, illumination, background and resolution are a few visual factors that are generally considered as influential problems in person re-identification (re-ID). Currently, major endeavor is devoted to algorithm design to mitigate their impact on the recognition system. Therefore, despite qualitatively acknowledging the factors as influential, it remains largely unknown how these factors affect the performance quantitatively.

In this paper, we study one of the most important factors, *i.e.*, viewpoint. Here, we denote viewpoint as the pedestrian rotation angle (Fig. 1). In what follows, we use viewpoint to replace pedestrian rotation angle unless specified. Since dif-

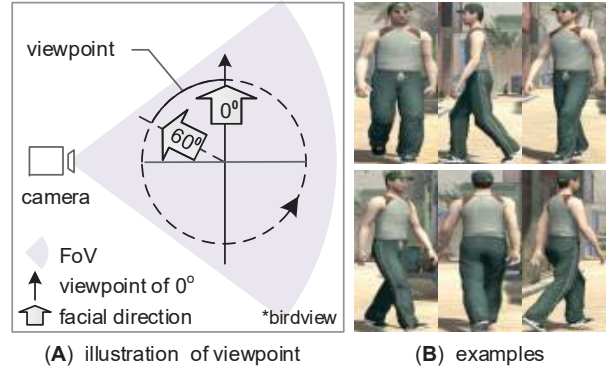


Figure 1. (A) Illustration of viewpoint of the birdview. Viewpoint is defined as the rotation angle of a person relative to a manually defined degree of  $0^\circ$ . The field of view (FoV) of a camera is shown. (B) Examples of persons under different viewpoints.

ferent views of a person contain different details, the viewpoint of a person influences the underlying visual data of an image, and thus the learning algorithm. Therefore, we aim to investigate the exact influence of viewpoint on the system. This study will benefit the community from two aspects. (1) The conclusions of this research can guide for building the training set effectively. For example, finding that certain angles are more important for learning models of identifying pedestrians. (2) It will advise for designing of query and gallery sets. By discovering viewpoints that are effective for re-ID accuracy, our research can potentially benefit the practical usage of re-ID systems.

In our attempt to reveal the influence of viewpoint, a notable obstacle is the lack of data. Existing datasets might have a biased and fixed distribution of environmental factors. In pedestrian viewpoint, for example, some angles might only have a few or even zero samples. In another example, when studying illumination on a real-world dataset, conclusions are less convincing because the dataset might only has a specific illumination condition. Further, a fixed/static data distribution forbids us from exploring how the impact of viewpoint relates to other visual factors. For example, the impact of viewpoint could be conditioned on the background, because background also affects feature

learning. To fully understand the role of viewpoint, we need to test its influence by changing the environment to either hard or easy modes. As such, without comprehensive and flexible data streams, we cannot make quantitative and scientific judgment of a visual factor’s significance.

This paper makes two contributions to the community. First, we build a large-scale data synthesis engine named PersonX. PersonX contains 1,266 manually designed identities and editable visual variables. It can simulate persons under various conditions. First of all, we demonstrate that existing re-ID models has **consistent accuracy trend** on both PersonX and real-world datasets [29, 30]. This observation suggests that **PersonX is indicative of the real world**. Moreover, as the name implies, the feature of PersonX is “controllable”. Persons take controllable poses and viewpoints, and the environment is controlled *w.r.t* the illumination, background, *etc.* Persons move by running, walking *etc.*, under the controlled camera view and scene. We can obtain the exact person bounding boxes without external detection tools and thus avoid the influence of detection errors on the system. Therefore, PersonX is indicative, flexible and extendable. It supports future research in not only algorithm design, but also scientific discoveries how environmental factors affect the system.

Second, we dissect a person re-ID system by quantitatively understanding the role of person viewpoint. Three questions are considered. (1) How does the *viewpoint of the training set* influence the system? (2) How does the *query viewpoint* influence the retrieval? (3) How does the re-ID accuracy change under different *viewpoint distributions of the testing set*? To answer these questions, we perform rigorous quantification on pedestrian images regarding viewpoints. We customize the viewpoints of persons in the PersonX engine from 0° to 360°. Both the control group and the experimental group are designed, so as to obtain convincing scientific conclusions. We also empirically study the real-world Market-1203 dataset where viewpoints of person are manually labeled. The empirical results are consistent with our findings on the synthetic data.

## 2. Related Work

We first review re-ID methods that improve the robustness against variations in pose, illumination, and background. We then review methods based on synthetic data.

**Against pose variance.** Some works [7, 28, 6, 20, 18] learn pose invariant representations for persons. For example, Farenza *et al.* [7] utilizes body symmetry on the x-axis and asymmetry on the y-axis two axes to design a descriptor with pose invariance. Cho *et al.* [6] quantize person poses into one of four canonical directions (front, right, back, left) to facilitate feature learning. Zheng *et al.* [28] design the PoseBox to align different persons along the body parts.

**Against background variance.** Some works reduce the

influence of background [4, 25, 5, 19, 22, 32, 34]. For instance, Chen *et al.* [5] fuse the descriptors from the foreground person and the original image, such that the foreground is paid more attention to by the network. In [19], Song *et al.* use binary segmentation masks to separate foreground from the background. They then learn representations from the foreground and background regions, respectively. Zheng *et al.* [31] apply STN to align pedestrian images, which reduces background noise and scale variances.

**Against resolution variance.** Resolution denotes the level of information granularity of an image. High resolution is typically preferred. But usually, the resolution level differs significantly across images. It thus affects the effectiveness of the learned features. To solve this problem, Jing *et al.* [11] design a mapping function that converts the features of low-resolution images into discriminative high-resolution features. alignment In [23], features from the bottom and top layers are concatenated during training and testing. Supervision signals are incorporated at each layer to train the multi-resolution features.

**Against viewpoint variance.** Learning viewpoint invariance is another focus [9, 24, 2, 12, 26]. For example, Both [9][12] regard viewpoint variations as the most prominent problem. In this area, Gray *et al.* [9] investigate the properties of localized features, while Karanam *et al.* [12] propose to learn dictionaries that can match person images captured under different viewpoints.

**Learning from synthetic data.** Leveraging synthetic data is a useful idea to alleviate the reliance on large-scale datasets. This strategy has been applied in problems like semantic segmentation [17], object tracking [8], traffic vision research [14] *etc.* In the person re-ID domain, SOMAset [3] is a synthetic dataset with 50 person models and 11 types of outfits. Barbosa *et al.* use SOMAset for training and test on real-world datasets. The accuracy was competitive. Bak *et al.* [1] also introduce a synthetic dataset SyRI including 100 characters. This dataset is featured by rich lighting conditions. A domain adaptation method is designed based on this dataset to fit real-world illumination distributions. Departing significantly from previous objectives of using synthetic dataset, this paper lays emphasis on quantitatively analyzing how visual factors influence the re-ID system. We derive useful insights by precisely controlling the simulator. This is a very early attempt of this kind in the community.

## 3. A Controllable Person Generation Engine

### 3.1. Description

**Software.** The PersonX engine<sup>1</sup> is built on Unity [15]. We create a 3D controllable world containing 1,266 person models. As a controllable system, it can satisfy various data

<sup>1</sup>The PersonX data engine, including pedestrian models, scene assets, project and script files *etc.*, are released at link.



Figure 2. The PersonX dataset. **A**: Background. In each background, a person can face toward a manually denoted direction, thus generating a controlled viewpoint. (1) - (3) represent backgrounds with uniform colors and (4) - (6) use street scenes as the background. **B**: Sample pedestrians bounding boxes in background (4). Various persons wearing various clothes are shown.

requirements. In PersonX, the characters and objects look realistic, because the texture and materials of these models are mapped from the real world by scanning real people and objects. The values of visual variables, *e.g.*, illumination, scenery and background, are designed to be editable. Therefore, PersonX is highly flexible and extendable.

**Identities.** PersonX has 1,266 hand-crafted identities including 547 females and 719 males. To ensure diversity, we hand-crafted the human models with different skin colors, ages, body forms (height and weight), hair styles, *etc.* The clothes of these identities include jeans, pants, shorts, slacks, skirts, T-shirts, dress shirts, maxiskirt, *etc.*, and some of these identities have a backpack, shoulder bag, glasses or hat. The materials of the clothes (color and texture) are mapped from images of real-world clothes. The motion of these characters can be walking, running, idling (standing), having a dialogue *etc.* Therefore, the 3D models in PersonX look realistic. Figure 2 (B) presents examples of identities of various ages, clothes, body shapes and poses.

### 3.2. Visual Factors in PersonX

PersonX is featured by editable environmental factors such as illumination, cameras, backgrounds and viewpoints. Details of these factors are described below.

**Illumination.** Illumination can be directional light (sun-light), point light, spotlight, area light, *etc.* Parameters like color and intensity can be modified for each illumination type. By editing the values of these terms, various kinds of illumination environment can be created.

**Camera.** The configuration of cameras in PersonX is subject to different values of image resolution, projection, focal length, and height.

**Background.** Currently PersonX has six different backgrounds (Fig. 2). In each experiment, we set 2-3 different backgrounds/cameras views. In each background/camera view, a person moves freely in arbitrary directions, exhibiting arbitrary viewpoints relative to the camera. In Fig. 2, backgrounds (4), (5) and (6) depict different street scenes. Among the three scenes, backgrounds (4) and (5) share the

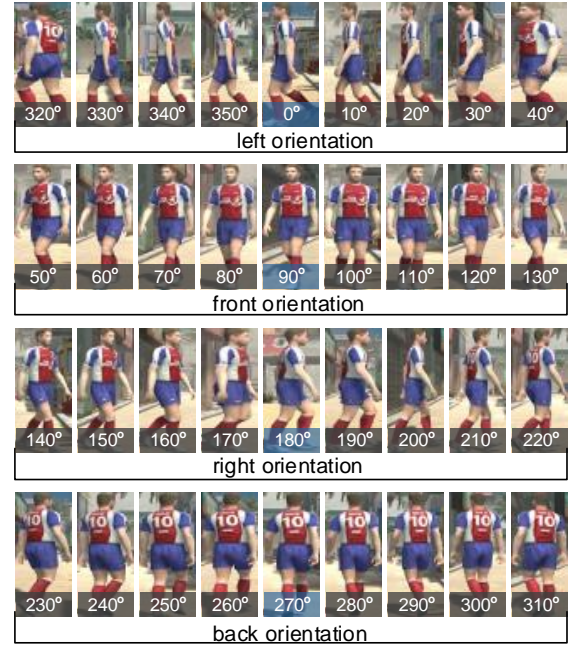


Figure 3. Definition of different viewpoints. Viewpoints of one identity are sampled at an interval of  $10^\circ$ . Left orientation represents the set of the viewpoints that contains more information on the left side of the person, *i.e.*,  $320^\circ - 40^\circ$ . Similarly, other orientations of the pedestrian represent the sets of viewpoints containing front ( $50^\circ - 130^\circ$ ), right ( $140^\circ - 220^\circ$ ) and back ( $230^\circ - 310^\circ$ ) information. The viewpoints with blue tags represent the due left ( $0^\circ$ ), front ( $90^\circ$ ), right ( $180^\circ$ ) and back ( $270^\circ$ ) of a person.

same illumination and ground color, while background (6) is a shadowed region and the ground color is gray. Meanwhile, backgrounds (1), (2) and (3) are pure colors and are used when background influence needs to be eliminated. Because we simplify our system into two cameras, we use various combinations of these six cameras to create different re-ID environments. When not specified, all the cameras have a high resolution of  $1024 \times 768$ .

**Viewpoint.** Figure 3 presents image examples under specified viewpoints. Those images are sampled during normal walking. Specifically, a person image is sampled ev-

	dataset	#identity	#box	#cam.	view
real data	Market-1501 [29]	1,501	32,668	6	N
	Market-1203 [29]	1,203	8,569	2	Y
	MARS [27]	1,261	1,191,003	6	N
	CUHK03 [13]	1,467	14,096	2	N
	Duke [16]	1,404	36,411	8	N
synthetic data	SOMAsset [3]	50	100,000	250	N
	SyRI [1]	100	1,680,000	—	N
	PersonX	1,266	273,456	6	Y
	PersonX <sub>123,456</sub>	1,266	136,728	3	Y
	PersonX <sub>12,13</sub>	1,266	91,152	2	Y
	PersonX <sub>45,46</sub>	1,266	91,152	2	Y

Table 1. Comparison of real-world and synthetic re-ID datasets. “View” denotes whether the dataset has viewpoint labels.

ery  $10^\circ$  from  $0^\circ$  to  $350^\circ$  (36 different viewpoints in total). Each viewpoint has 1 image, so each person has 36 images. The entire PersonX engine thus has  $36$  (viewpoints)  $\times$   $1,266$  (identities)  $\times$   $6$  (cameras) = 273,456 images. For each person, the 36 viewpoints are divided into 4 groups, representing four orientations: left, front, right and back. We use “left” and “due left” to represent images with viewpoints from  $320^\circ$  to  $40^\circ$  (*i.e.*, left orientation), the image of  $0^\circ$ , respectively. This convention applies for other orientations.

Comparisons of PersonX and some existing re-ID datasets are presented in Table 1. There are two existing synthetic datasets, SyRI [1] and SOMAsset [3]. SyRI is used as an alternative data source for domain adaptation and does not have the concept of cameras. SOMAsset contains 250 cameras, which are uniformly distributed along a hemisphere around each person. Neither datasets are freely editable by the public. In comparison, PersonX has configurable backgrounds and much more identities. Importantly, it can be edited/extended not only for this study, but also for future research in this area.

## 4. Benchmarking and Dataset Validation

In this section, we aim to validate that PersonX is indicative of the real world, such that conclusions derived from this dataset can be of value to practice.

### 4.1. Methods and Subsets

We use IDE+ [33], triplet feature [10] and PCB [21] for our purpose. IDE+ is implemented on ResNet50. During training, the batch size is set to 64 and the model is trained for 50 epochs. The learning rate is initialized to 0.1 and decays to 0.01 after 40 epochs. The model parameters are initialized with the model pre-trained on ImageNet. For triplet feature, the number of identities per batch is set to 32 and number of images per identity is set to 4. So the batch size is  $32 \times 4 = 128$ . The learning rate is initialized to  $2 \times 10^{-4}$  and decays after 150 epochs (300 epochs in total). Training of PCB follows the standard setup described in [21].

Through combinations of the six backgrounds described in Section 3.2, PersonX has the following subsets.

- PersonX<sub>12</sub>. It has backgrounds (1) and (2). Both are pure color backgrounds; the colors are similar.
- PersonX<sub>13</sub>. The two cameras face backgrounds (1) and (3). The color difference between the two backgrounds is significant than that in PersonX<sub>12</sub>.
- PersonX<sub>123</sub>. This is a three-camera system, comprising backgrounds (1), (2) and (3).
- PersonX<sub>45</sub>. It contains backgrounds (4) and (5) of street scenes. The two backgrounds are close in scene and illumination.
- PersonX<sub>46</sub>. It consists of backgrounds (4) and (6). The two backgrounds have larger disparity than PersonX<sub>45</sub>.
- PersonX<sub>456</sub>. It is a three-camera system consisting of backgrounds (4), (5) and (6).

Overall, PersonX<sub>12</sub>, PersonX<sub>13</sub> and PersonX<sub>123</sub> are simple subsets, while PersonX<sub>45</sub>, PersonX<sub>46</sub> and PersonX<sub>456</sub> are more complex ones. Moreover, we introduce low-resolution subsets to create more challenging settings. We edit the image resolution of PersonX<sub>45</sub>, PersonX<sub>46</sub> and PersonX<sub>456</sub> from  $1024 \times 768$  to  $512 \times 242$  (for images of FoV). We use “-lr” to represent low-resolution subsets.

For benchmarking, we randomly sample 410 identities for training and the rest 856 identities for testing. In each camera, an identity has 36 images, *i.e.*, 36 viewpoints, from which one image is selected as the query during testing. Therefore, the three-camera subsets, *i.e.*, PersonX<sub>456</sub> and PersonX<sub>123</sub>, contain 44,280 ( $410 \times 36 \times 3$ ) training and 92,448 ( $856 \times 36 \times 3$ ) testing images. The two-camera subsets have 29,520 ( $410 \times 36 \times 2$ ) training and 61,632 ( $856 \times 36 \times 2$ ) testing images.

### 4.2. System Validation

We evaluate the three algorithms on both real-world and synthetic datasets. We use the standard evaluation protocols [29, 30]. Results are reported in Fig 4. We observe three characteristics of PersonX.

First, **eligibility**. We find the performance trend of the three algorithms is similar between PersonX and real-world datasets. On Market-1501 and DukeMTMC, for example, PCB has the best accuracy, and the performance of IDE+ and triplet feature is close. That is,  $\text{PCB} \succ \text{triplet} \approx \text{IDE+}$ . This is consistent with findings in [21]. On the synthetic PersonX subsets, the performance trend is similar: IDE+ and triplet feature have similar accuracy; PCB is usually 2%-3% higher than them. These observations suggest that PersonX is indicative of the real-world and that future conclusions derived from PersonX can be of real-world value.

Second, **purity**. The re-ID accuracy on PersonX subsets (Fig. 4 (B)) are relatively high compared to the real-world datasets (Fig. 4 (A)). It does not mean these subsets are “easy”. In fact, the high accuracy is what we design for, as it excludes the influence of the environmental factors. In other words, these subsets are *oracle*: images are



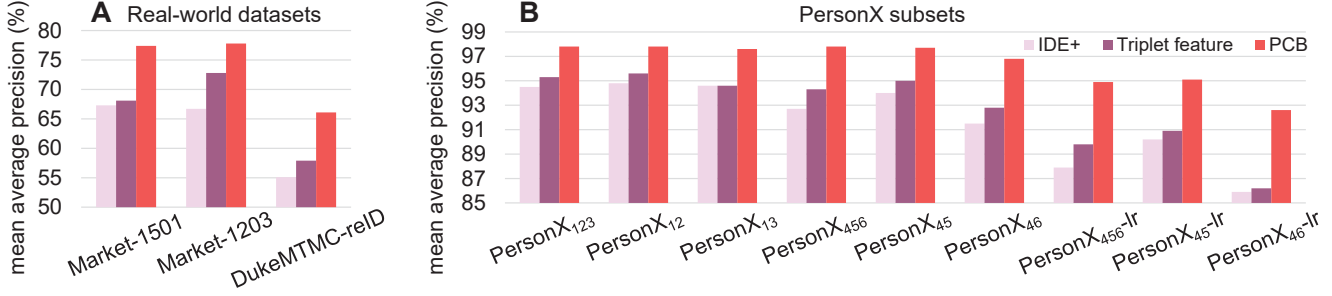


Figure 4. Re-ID mean average accuracy (mAP, %) of IDE+, triplet feature, and PCB on (A) real-world datasets and (B) the PersonX subsets. “lr” means the frames are low resolution of  $512 \times 242$  instead of the original resolution  $1024 \times 768$ .

high-resolution, and the scenes have normal sunlight and relatively consistent backgrounds. These subsets are thus ideal ones for studying the impact of viewpoints.

Third, **sensitivity**. We show that these subsets are sensitive to the changes in the environment. For example, background variation in PersonX<sub>46</sub> is much larger than PersonX<sub>45</sub>. As such, we observe that mAP in PersonX<sub>46</sub> is lower by 1% - 3% for different algorithms. Similarly, the background in PersonX<sub>12</sub> is much simpler than PersonX<sub>46</sub>, which causes mAP on PersonX<sub>46</sub> to be lower than on PersonX<sub>12</sub>. Further, when these subsets are manually edited to be low resolution, we observe a significant mAP drop. For example, the mAP drop from PersonX<sub>46</sub> to PersonX<sub>46-lr</sub> is about 6%. The above comparisons demonstrate that PersonX subsets are sensitive to background complexity, variation between cameras, and image resolution. This is consistent with our intuition and indicates that PersonX is useful in studying the influence of visual factors.

The above discussions indicate that PersonX is indicative of the real-world trend, has strictly controlled environment variables, and is reasonably sensitive to environmental changes. We believe PersonX will be a useful tool for the community and encourage the development of robust algorithms and scientific analysis.

## 5. Evaluation of Viewpoint

We evaluate the impact of viewpoint on person re-ID. The experiment is based on PCB [21]. We note that other standard re-ID methods (*e.g.*, IDE+) can draw similar conclusions. Three questions will be investigated in the following subsections: how does the viewpoint in (1) the training set, (2) the query set, and (3) the gallery set affect the re-ID accuracy? For clearer understanding, we mainly show figures in this section. Detailed numbers are provided in the supplementary material.

### 5.1. How Do Viewpoint Distributions in the Training Set Affect Model Learning?

**Experiment design.** Initially, the subsets contain all the viewpoints for the training and testing IDs. That is, a person has 36 images under each camera. In this section, to study

the influence of missing viewpoints in the training set, we remove specific orientations from the training set. The orientations refer to left, front, right and back shown in Fig. 3. We design the following training sets.

- Control group 1. We *randomly* select half (18 out of 36) or a quarter (9 out of 36) images of each identity for training.
- Control group 2. The training set is constituted by *randomly* selecting half (18 out of 36) or a quarter (9 out of 36) viewpoints for each identity.
- Experimental group 1. Train with two orientations. The training images exhibit two orientations, left+right or front+back. The training set is thus half of the original training set.
- Experimental group 2. Train with one orientation. The training set has one orientation, *i.e.*, left, right, front, or back. The training set becomes a quarter of the size of the original training set.

**Discussions.** The experimental groups are used to assess the impact of missing viewpoints in the training set. To cancel out the result influence of reduced training images and the non-uniform viewpoint distributions of the experimental groups, we further design two control groups, where the number of images used for training is the same with the above two experimental groups. The first control group removes images randomly, and the second control group removes viewpoints randomly. For control group 1 and 2, we repeat our experiment 5 times and report the average re-ID accuracy. Using the two control groups, we highlight the impact of the missing viewpoints in the training set.

**Result analysis.** Using the experimental groups and control groups designed above, we summarize the key experimental results in Fig. 5. These results are reported on two synthetic datasets, PersonX<sub>12</sub> and PersonX<sub>46</sub>, and a real-world dataset, Market-1203. We mainly use the mean average precision (mAP) for evaluation, as it provides a comprehensive assessment of the system’s ability to retrieve all the relevant images. From these results, we have several observations as follows.

First, the two control training sets have similar accuracy, and control group 2 is slightly inferior. Control groups 2

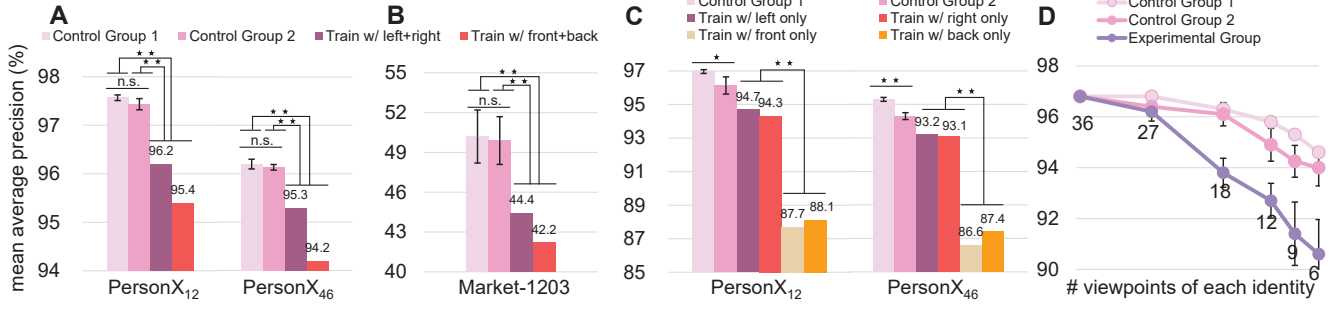


Figure 5. Re-ID accuracy (mAP, %) when the training set has missing orientations/viewpoints. **A** and **B**: we use two orientations for training. For example, we can train with left and right orientations only (see Fig. 3 for the definition of orientation). **C**: we train with one orientation *only*, i.e., left, right, front, or back orientation. For each dataset, we have two control groups. **D**: Impact of missing continuous viewpoints on PersonX<sub>46</sub>. The horizontal axis is the remaining number of viewpoints and vertical axis is the mAP. In the experimental group, continuous viewpoints are removed. The number on this curve denotes the remaining number of viewpoints. “n.s.” represents that the difference between results is **not statistically significant** (i.e.,  $p$ -value  $> 0.05$ ).  $\star$  corresponds to **statistically significant** (i.e.,  $0.01 < p$ -value  $< 0.05$ ).  $\star\star$  means the difference between results is **statistically very significant** (i.e.,  $0.001 < p$ -value  $< 0.01$ ).

has some specific viewpoint missing, while control group 1 has images randomly missing. This indicates that viewpoint comprehensiveness is important for a training set.

Second, removing continuous viewpoints in the training set causes more accuracy drop than removing random viewpoints or random images. In control group 2, the viewpoints are randomly removed. In the two experimental groups, continuous viewpoints are removed. The inferior accuracy of the experimental groups (see Fig. 5 (A), (B), (C) and (D)) indicate that continuous viewpoints are more important. Further, Fig. 5 (D) demonstrates an increasing performance gap as more viewpoints are removed. This observation is intuitive because continuous viewpoints encode appearance cues that cannot be recovered by other viewpoints and once lost, will cause system degradation.

Third, from Fig. 5 (A) and (B), when the training set only has two orientations (left+right or front+back), we observe a significant accuracy drop compared with the control groups. Similarly, Fig. 5 (C) indicates that a training set with only one orientation also deteriorates the re-ID accuracy when compared with the control groups.

Fourth and importantly, left/right orientations make a better training set than front/back orientations. From Fig. 5 (A), (B) and (C), when the training set is composed of left/right orientations, the re-ID accuracy is higher than training sets with front/back orientations. For example, when using a training set composed of “front+back” orientations, the mAP score in Fig. 5 (A) is 0.8%-1.1% lower than a model trained with “left+right” orientations. On the real-world dataset, Market-1203 in Fig. 5 (B), the mAP of a model trained with left and right orientations is 2.2% higher than learning from the viewpoint from front and back. It indicates that data synthesis is indicative of the real world to some extent. Similarly, when the training set only has one orientation (Fig. 5 (C)), the left or right orientations are

significantly more beneficial than the front or back orientations. The mAP gap can be as large as 6%. Note that for evaluating training sets with only one orientation, we do not use Market-1203. This is because Market-1203 does not have sufficient training samples under each orientation.

Regarding the observation that left/right orientations are more useful than front/back orientations, we provide a plausible reason below. For pedestrians, the left or right orientations reflect important general information, such as color, outfit (e.g., long or short sleeve, pants, shorts) *etc.* In comparison, the front and back views capture more detailed appearance cues such as prints of clothes, face, *etc.* As such, a model trained with left/right viewpoints encodes the general appearance knowledge about pedestrians; a model trained with front/back viewpoints somehow has abilities that are useful in recognizing specifically the front/back views but might be abundant for the side views. In other words, if a true match is of the left or right orientation that does not present as much texture details, a model trained with front or back orientations may not work well. On the other hand, a model trained with left or right viewpoints is good at recognizing the clothes appearance, so its performance will not deteriorate much when identifying pedestrians under front or back orientations.

**A further study.** To further understand the superiority of left/right orientations in training, we quantify the query viewpoint and the true match viewpoint into four orientations, too. Results are shown in Fig. 6. Here, training sets are constructed with only one orientation. First, when the query images exhibit only one orientation, and when the true match viewpoint distributes uniformly in the gallery, a training set with left or right orientations is superior to that with front or back orientations. Second, we assume a single viewpoint for all the true matches in the gallery. We also assume a single true match for each query. Four models

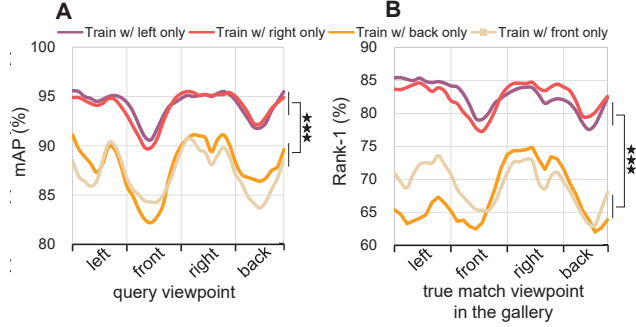


Figure 6. Evaluation of models trained on one orientation only. **A:** query viewpoint change vs. mAP. Query viewpoint changes from left, front, right to back. True matches in the gallery are uniformly distributed. Since a query has multiple true matches, we use mAP to measure accuracy. **B:** true match viewpoint vs. rank-1. The true match viewpoint changes from left, front, right to back. Query viewpoints are uniformly distributed. *Each query has only one true match in the gallery*, so the rank-1 accuracy is used for evaluation. \*\*\* means that the difference between results of models trained on left/right and front/back orientations is **statistically extremely significant** (i.e.,  $p$ -value  $< 0.001$ ).

are trained with images solely from one of the four orientations. We show the rank-1 accuracy of the four models in Fig. 6 (B). Note that the viewpoint distribution in query set is uniform. In our observation, for a true match to be retrieved, using models trained on the left or right orientations yields higher accuracy than models trained on front or back orientations. Therefore, regardless of the viewpoint distribution in the gallery or query set, a person re-ID model trained with left or right orientations performs favorably than trained with front or back orientations.

#### Subsection conclusions

- Missing viewpoint compromises training.
- Missing continuous viewpoints are more detrimental than missing randomly viewpoints.
- When limited training viewpoints are available, left/right orientations allow models to be better trained than front/back orientations.

## 5.2. How Does Query Viewpoint Affect Retrieval?

We study how query viewpoint influences re-ID results.

**Experiment design.** We train a model on the original training set comprised of every viewpoint. We modify the query viewpoints to see its effect during testing. Specifically, the viewpoint of a probe image can be set to the *due left* ( $0^\circ$ ), *due front* ( $90^\circ$ ), *due right* ( $180^\circ$ ) or *due back* ( $270^\circ$ ) to represent different sides of person. During retrieval, we assume only one true match in gallery; the true match contains the same person as the query, and its viewpoint is between  $0^\circ$  and  $350^\circ$ . Viewpoints of the distractor gallery images are images of all other persons.

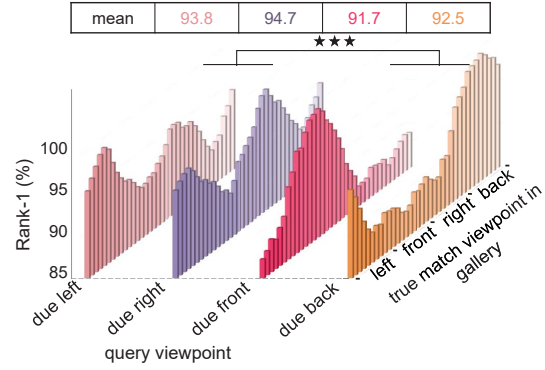


Figure 7. Impact of query viewpoint on system performance on PersonX<sub>45</sub>. Four viewpoints are evaluated, i.e., due left ( $0^\circ$ ), right ( $180^\circ$ ), front ( $90^\circ$ ) and back ( $270^\circ$ ). In the gallery, there is only one true match for each query. The true match viewpoint varies from  $0^\circ$  to  $350^\circ$  (deep axis). Under each query viewpoint, we report 36 rank-1 scores obtained by the query to retrieve 36 types of true match viewpoints. \*\*\* means the difference between retrieval results of due left/right and due front/back is **statistically extremely significant** (i.e.,  $p$ -value  $< 0.001$ ). On the top, we show the average rank-1 accuracy for each query viewpoint.

**Result analysis.** Figure 7 presents the results obtained by the above query and gallery images. We use PersonX<sub>45</sub> for training and testing. We have several observations.

First, when the viewpoint of the true match is similar to the query, the highest re-ID accuracy can be achieved. For example, the maximum rank-1 values of due left queries correspond exactly to the due left true match in the gallery. Under the same viewpoint, the query and true match are different only in illumination and background. This indicates that viewpoint differences between two to-be-matched images cause performance drop.

Meanwhile, queries of the due left and the due right viewpoint lead to a higher average rank-1 accuracy than queries of due front and due back viewpoints. For example, the accuracy of the due left queries and the due front queries is 93.8% and 91.7%, respectively. It is noteworthy that in Section 5.1 and Fig. 6, we can have a similar observation regarding the superiority of left/right viewpoints in the training and query sets.

#### Subsection conclusions

- The query viewpoint of left/right generally leads to higher re-ID accuracy than front/back viewpoints.

## 5.3. How Do True Match Viewpoints in the Gallery Affect Retrieval?

Finally, we study how the gallery viewpoint distribution affects re-ID accuracy. Specifically, we study the viewpoint disparity between the query and its true matches.

**Experiment design.** We denote the viewpoint of a query and its true match as  $\theta_q$  and  $\theta_t$ , respectively. The experi-



Figure 8. The impact of viewpoint disparity between a query and its true matches. For training, we use the original training sets (balanced viewpoints). Results are reported on **A** PersonX subsets and **B** Market-1203. From PersonX<sub>45</sub>, PersonX<sub>46</sub>, PersonX<sub>46</sub>-lr to Market-1203, the environmental difficulty is in increasing order.

mental groups for PersonX subsets are designed as below.

- Experimental group 1. The three true matches whose  $\theta_t \in [\theta_q \pm 10^\circ]$  are removed (set as “junk”).
- Control group 1. Three true matches are randomly removed from the gallery.
- Experimental group 2. The nine true matches whose  $\theta_t \in [\theta_q \pm 20^\circ]$  are removed.
- Control group 2. Randomly removing 5 true matches. More difficult situations:
- Experimental group 3. The nine true matches whose  $\theta_t \in [\theta_q \pm 40^\circ]$  are removed.
- Control group 3. Randomly removing 9 true matches.

Since Market-1203 only contains eight viewpoint types, the two experimental groups remove three or five true matches that have the most similar viewpoints to the query. The corresponding control groups randomly remove the same number of true match images.

**Result analysis.** From Fig. 8, we have two observations.

The major observation is that if true matches with similar viewpoints are not present, there will be a non-trivial performance drop. In other words, if true matches in the gallery have large viewpoint disparity with the query, the retrieval accuracy will be negatively affected. For instance, the mAP of removing 9 true matches (experimental group) on PersonX<sub>45</sub> is 96.6%, and there is a decrease of about 1.0% on mAP compared to the control group 3. Consistent observation can be made on Market-1203. For example, compared with the control group, there is a decrease of about 3.0% on mAP when 3 or 5 true matches with similar viewpoints to the query are removed from the gallery.

Figure 9 shows some re-ID results on the Market-1203 dataset. For the first query images in Fig. 9, the true match is ranked to the highest position. This is because the first true match is similar to the query in both appearance and viewpoint. After removing it, the highly ranked images are mostly false matches that have a similar viewpoint with the query. Similarly, for other example query images that do

Query Rank-1 → Rank-10



Figure 9. Sample re-ID results on Market-1203. Images in the first column are queries. The retrieved images are sorted according to their similarity to the query (high to low) from left to right. The similarity is calculated by using feature extracted from the PCB model. True matches and false matches are in green and red rectangles, respectively.

not have true matches of similar viewpoints in gallery, the false matches of distinctive appearance (*e.g.*, different styles and colors of clothes and bags) but similar viewpoints to the query will be ranked higher than the true matches.

Moreover, the accuracy decrease caused by viewpoint disparity between a query and its true match becomes more obvious when the environment becomes more challenging. For example, the mAP drop of the experimental groups on the PersonX<sub>46</sub>-lr dataset is almost twice as large as the performance decline on the PersonX<sub>46</sub> dataset.

#### Subsection conclusions

- True matches whose viewpoints are dissimilar to the query are harder to be retrieved than true matches with a similar viewpoint to the query.
- The above problem becomes more severe when the environment is challenging, *e.g.*, complex background, extreme illumination, and low resolution.

## 6. Conclusion

This paper makes a step from engineering new technologies to science new discoveries. We make two contributions to the community. First, we build a synthetic data engine PersonX that can generate images under controllable cameras and environments. Subsets of PersonX are shown to be indicative of the real world. Second, based on PersonX, we conduct comprehensive experiments to quantitatively assess the influence of pedestrian viewpoint on person re-ID accuracy. Interesting and constructive insights are derived, *e.g.*, it is better to use a query image capturing the side view of a person. In the future, visual factors such as illumination and background will be studied with this new engine.



## References

- [1] Slawomir Bak, Peter Carr, and Jean-Francois Lalonde. Domain adaptation through synthesis for unsupervised person re-identification. *arXiv preprint arXiv:1804.10094*, 2018.
- [2] Slawomir Bak, Sofia Zaidenberg, Bernard Boulay, and François Bremond. Improving person re-identification by viewpoint cues. In *AVSS*, pages 175–180, 2014.
- [3] Igor Barros 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.
- [4] Loris Bazzani, Marco Cristani, Alessandro Perina, and Vittorio Murino. Multiple-shot person re-identification by chromatic and epitomic analyses. *Pattern Recognition Letters*, 33(7):898–903, 2012.
- [5] Di Chen, Shanshan Zhang, Wanli Ouyang, Jian Yang, and Ying Tai. Person search via a mask-guided two-stream cnn model. In *ECCV*, 2018.
- [6] Yeong-Jun Cho and Kuk-Jin Yoon. Improving person re-identification via pose-aware multi-shot matching. In *CVPR*, pages 1354–1362, 2016.
- [7] Michela Farenzena, Loris Bazzani, Alessandro Perina, Vittorio Murino, and Marco Cristani. Person re-identification by symmetry-driven accumulation of local features. In *CVPR*, 2010.
- [8] Adrien Gaidon, Qiao Wang, Yohann Cabon, and Eleonora Vig. Virtual worlds as proxy for multi-object tracking analysis. In *CVPR*, 2016.
- [9] Douglas Gray and Hai Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *ECCV*, 2008.
- [10] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017.
- [11] Xiao-Yuan Jing, Xiaoke Zhu, Fei Wu, Xinge You, Qinglong Liu, Dong Yue, Ruimin Hu, and Baowen Xu. Super-resolution person re-identification with semi-coupled low-rank discriminant dictionary learning. In *CVPR*, 2015.
- [12] Srikrishna Karanam, Yang Li, and Richard J Radke. Person re-identification with discriminatively trained viewpoint invariant dictionaries. In *ICCV*, 2015.
- [13] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deep-reid: Deep filter pairing neural network for person re-identification. In *CVPR*, 2014.
- [14] Xuan Li, Kunfeng Wang, Yonglin Tian, Lan Yan, Fang Deng, and Fei-Yue Wang. The paralleleye dataset: A large collection of virtual images for traffic vision research. *IEEE Transactions on Intelligent Transportation Systems*, (99):1–13, 2018.
- [15] John Riccitiello. John riccitiello sets out to identify the engine of growth for unity technologies (interview). *VentureBeat. Interview with Dean Takahashi. Retrieved January, 18, 2015*.
- [16] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *ECCV*, pages 17–35, 2016.
- [17] Swami Sankaranarayanan, Yogesh Balaji, Arpit Jain, Ser Nam Lim, and Rama Chellappa. Learning from synthetic data: Addressing domain shift for semantic segmentation. In *CVPR*, 2018.
- [18] M Saquib Sarfraz, Arne Schumann, Andreas Eberle, and Rainer Stiefelhagen. A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking. In *CVPR*, 2018.
- [19] Chunfeng Song, Yan Huang, Wanli Ouyang, and Liang Wang. Mask-guided contrastive attention model for person re-identification. In *CVPR*, 2018.
- [20] Chi Su, Jianing Li, Shiliang Zhang, Junliang Xing, Wen Gao, and Qi Tian. Pose-driven deep convolutional model for person re-identification. In *ICCV*, 2017.
- [21] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *ECCV*, 2018.
- [22] Maoqing Tian, Shuai Yi, Hongsheng Li, Shihua Li, Xuesen Zhang, Jianping Shi, Junjie Yan, and Xiaogang Wang. Eliminating background-bias for robust person re-identification. In *CVPR*, 2018.
- [23] Yan Wang, Lequn Wang, Yurong You, Xu Zou, Vincent Chen, Serena Li, Gao Huang, Bharath Hariharan, and Kilian Q Weinberger. Resource aware person re-identification across multiple resolutions. In *CVPR*, 2018.
- [24] Ziyang Wu, Yang Li, and Richard J Radke. Viewpoint invariant human re-identification in camera networks using pose priors and subject-discriminative features. *IEEE transactions on pattern analysis and machine intelligence*, 37(5):1095–1108, 2015.
- [25] Tong Xiao, Shuang Li, Bochao Wang, Liang Lin, and Xiaogang Wang. Joint detection and identification feature learning for person search. In *CVPR*, 2017.
- [26] Kang Zheng, Xiaochuan Fan, Yuewei Lin, Hao Guo, Hongkai Yu, Dazhou Guo, and Song Wang. Learning view-invariant features for person identification in temporally synchronized videos taken by wearable cameras. In *ICCV*, 2017.
- [27] Liang Zheng, Zhi Bie, Yifan Sun, Jingdong Wang, Chi Su, Shengjin Wang, and Qi Tian. Mars: A video benchmark for large-scale person re-identification. In *European Conference on Computer Vision*, 2016.
- [28] Liang Zheng, Yujia Huang, Huchuan Lu, and Yi Yang. Pose invariant embedding for deep person re-identification. *arXiv preprint arXiv:1701.07732*, 2017.
- [29] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *ICCV*, 2015.
- [30] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *ICCV*, 2017.
- [31] Zhedong Zheng, Liang Zheng, and Yi Yang. Pedestrian alignment network for large-scale person re-identification. *TCSVT*, 2018.

- [32] Zhun Zhong, Liang Zheng, Zhiming Luo, Shaozi Li, and Yi Yang. Invariance matters: Exemplar memory for domain adaptive person re-identification. In *CVPR*, 2019.
- [33] Zhun Zhong, Liang Zheng, Zhedong Zheng, Shaozi Li, and Yi Yang. Camera style adaptation for person re-identification. In *CVPR*, 2018.
- [34] Zhun Zhong, Liang Zheng, Zhedong Zheng, Shaozi Li, and Yi Yang. Camstyle: A novel data augmentation method for person re-identification. *IEEE Transactions on Image Processing*, 28(3):1176–1190, 2019.