

TAKING A CLOSER LOOK AT SYNTHESIS: FINE-GRAINED ATTRIBUTE ANALYSIS FOR PERSON RE-IDENTIFICATION

Suncheng Xiang¹, Yuzhuo Fu¹, Guanjie You², Ting Liu¹

¹School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University

²College of Intelligence Science and Technology, National University of Defense Technology

ABSTRACT

Person re-identification (re-ID) plays an important role in applications such as public security and video surveillance. Recently, learning from synthetic data, which benefits from the popularity of synthetic data engine, has achieved remarkable performance. However, in pursuit of high accuracy, researchers in the academic always focus on training with large-scale datasets at a high cost of time and label expenses, while neglect to explore the potential of performing efficient training from millions of synthetic data. To facilitate development in this field, we reviewed the previously developed synthetic dataset GPR and built an improved one (*GPR+*) with larger number of identities and distinguished attributes. Based on it, we quantitatively analyze the influence of dataset attribute on re-ID system. To our best knowledge, we are among the first attempts to explicitly dissect person re-ID from the aspect of attribute on synthetic dataset. This research helps us have a deeper understanding of the fundamental problems in person re-ID, which also provides useful insights for dataset building and future practical usage.

Index Terms— re-identification, synthetic dataset, fine-grained, attribute analysis

1. INTRODUCTION

Person re-ID aims to identify images of the same person from large number of cameras views in different places, which has attracted lots of interests and attention in both academia and industry. Encouraged by the remarkable success of deep learning methods [1, 2, 3, 4] and the availability of re-ID datasets [5, 6, 7], performance of person re-ID has been significantly boosted. Currently, these performance gains come only when a large diversity of training data is available, which is at the price of large amount of accurate annotations obtained by intensive human labor. Accordingly, real applications have to cope with challenges like complex lighting and scene variations, which current real datasets might fail to address [8].



Fig. 1. Illustration of the examples between Market-1501 (**upper-left**) and DukeMTMC-reID (**upper-right**). It can be easily observed that there exists obvious differences among different datasets in light, background or weather. In comparison, our new dataset *GPR+* (bottom) always has large variances, high resolution and different backgrounds.

To alleviate this problem, many successful person re-ID approaches [9, 10, 11, 12] have been proposed to take advantage of game engine to construct large-scale synthetic re-ID datasets, which can be used to pre-train or fine-tune CNN network. For example, Barbosa et al. [9] propose a synthetic dataset SOMAset created by photorealistic human body generation software. Sun et al. [11] introduce a synthetic data engine to dissect re-ID system on the viewpoints of pedestrian. Recently, Wang et al. [12] collect a virtual dataset RandPerson with 3D characters. However, current researches mainly concentrate on achieving satisfactory performance with large-scale data at the sacrifice of expensive time costs and intensive human labors, while neglect the potential of performing efficient training from millions of synthetic data. Besides, in the field of person re-ID, current synthetic datasets mainly provide no more than three attribute annotations, such as IDs, viewpoint or illumination, which cannot satisfy the need for this challenging fine-grained attribute analysis task.

Another challenge we observe is that, there exists serious scene difference [13] between synthetic and real dataset. For instance, as shown in Fig. 1, Market-1501 [5] only contains scenes that recorded in summer vacation, while DukeMTMC-

This work was supported by the National Natural Science Foundation of China under Project(Grant No.61977045). Preprint. Under review.

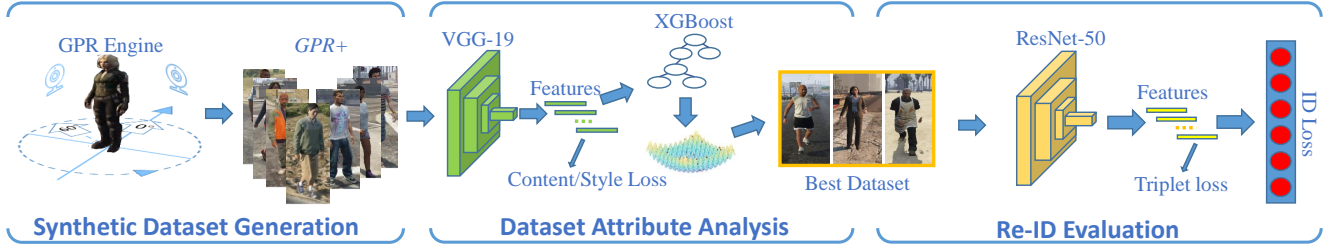


Fig. 2. The procedure of our proposed end-to-end systematic framework, which consists of 1) synthetic dataset generation (*GPR+*), 2) dataset attribute analysis, and 3) re-ID evaluation period. Best viewed in color.

reID [6, 14] is set in the blizzard scenes. Consequently, pre-training with all synthetic datasets may lead to negative domain adaptation and deteriorate performance on target domain, which is not practical in real-world scenarios.

In this paper, we circumvent above issues by exploring a new direction, that is, analyzing the influences of different attributes in a fine-grained manner. To our knowledge there is no work in the existing literatures that comprehensively study the impacts of multiple attributes on re-ID system. So a natural question then come to our attention: *how does these attributes influence the retrieval performance? Which one is most critical to our re-ID system?* To answer these questions, we perform rigorous quantification on pedestrian images regarding different attributes. In this paper, our baseline system is built with commonly used loss functions [15, 16] on vanilla ResNet-50 with no bells and whistles. Moreover, we have summarized a systematic framework for evaluating the importance of attributes, as shown in Fig. 2, which can be applied to the construction of datasets with high quality for other computer vision related tasks.

To this end, our work contributes to the research of re-ID community in two different ways: 1) We upgrade the previous dataset *GPR* to *GPR+*, which is both densely annotated and visually coherent with real world. It also has more identities and decoupled attributes. This is non-trivial, especially when it comes to the data annotation. 2) We conduct in-depth studies on top of *GPR+* to dissect a person re-ID system comprehensively, then quantitatively analyze the feature importance of different attributes. The empirical results is significant for us to construct high-quality dataset.

2. PROPOSED METHOD

2.1. The *GPR+* dataset

In previous *GPR* [17] dataset, we found that there exists serious redundancy in attribute distribution, for example, the time distribution between “21~24” and “00~03” is highly correlated that inevitably introduces some interferences in attribute analysis. To address the correlation problem and enhance the

orthogonality of intra-attributes, we redefined them in the upgraded version *GPR+*, which provides a solid foundation for fine-grained attribute analysis. In comparison with *GPR*, we have 1) More identities and bounding boxes; 2) More distinguished and complementary attribute distribution. We summarize the new features in *GPR+* into the following aspects:

Identity. 808 identities and 475,104 bounding boxes, including more high-quality attribute annotations.

Viewpoint. 12 different types of viewpoints for one identity: every 30° from $0^\circ \sim 330^\circ$.

Weather. 7 different types of weather: *sunny, clouds, overcast, foggy, neutral, blizzard, snowlight*.

Illumination. 7 different types of illumination: *midnight, dawn, forenoon, noon, afternoon, dusk, night*.

On the whole, these aspects together make *GPR+* a rich dataset for research. Detailed information and results about *GPR+* can be found at <https://JeremyXSC.github.io/GPR/>.

2.2. Attribute Analysis

One of the key problems in attribute analysis is to find the best representation, which can be addressed from two levels:

Content representation. In this work, aiming at visualizing the content representation [18] of a image in different layers of CNN, we reconstruct the input image through the feature map of a certain layer in the VGG network [19]. Under this condition, we adopt squared-error loss between two feature representations F_{ij}^l and P_{ij}^l of the i^{th} filter at position j in layer l as content loss,

$$\mathcal{L}_{\text{content}} = \frac{1}{2} \sum_{i,j} (F_{ij}^l - P_{ij}^l)^2 \quad (1)$$

Style representation. Formally, to obtain the style representation [20, 21] of input image, we use Gram Matrix to measure the feature correlations between the different filter responses, which is built on top of the CNN responses in each layer of the network. And the style loss is written as,

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

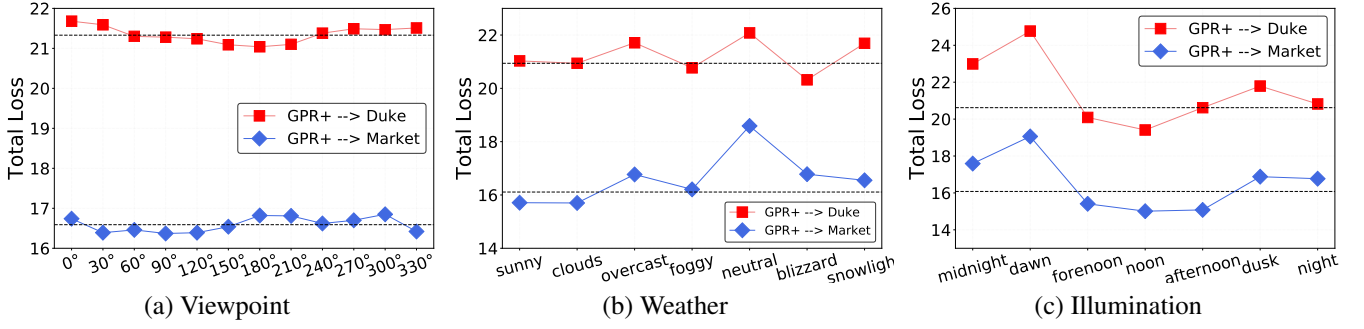


Fig. 3. Loss distribution on Duke and Market. Attributes whose value below horizontal line can be regarded as optimal items.

where w_l is a hyper-parameter that controls the importance of each layer to the style loss $\mathcal{L}_{\text{style}}$. N_l represents number of filters and M_l is height \times width of the feature map. G_{ij} and A_{ij} denote the Gram Matrix of real images and synthetic images with style representation in layer l . Then the total loss (illustrated in Fig. 3) for our attribute metric is represented as

$$\mathcal{L}_{\text{total}} = \alpha * \mathcal{L}_{\text{style}} + \beta * \mathcal{L}_{\text{content}} \quad (3)$$

where α and β are two hyper-parameters which control the relative importance of style loss and content loss separately. Then we adopt a scalable end-to-end tree boosting system XGBoost[22] in terms of **gain** to analyze the feature importance score of different attributes to re-ID system (see Fig. 4).

2.3. Re-ID evaluation

Re-ID evaluation is the third step of the proposed pipeline, which is same as supervised method. Intuitively, since we mainly focus on **1) dataset generation** and **2) attribute analysis**, we follow a widely used open-source¹ as our standard baseline, and adopt global features provided by backbone ResNet-50 [2] to perform feature learning. Note that we only modify the output dimension of the last fully-connected layer to the number of training identities. During testing, we extract the 2,048-dim pool-5 vector for retrieval under the commonly used Euclidean distance.

3. EXPERIMENTAL RESULTS

3.1. Datasets and Implementation Details

In this paper, we evaluate our method on two benchmark datasets Market-1501 [5] and DukeMTMC-reID [6, 14]. Market-1501 has 32,668 person images of 1,501 identities which is collected in the campus of Tsinghua University. DukeMTMC-reID contains 1,812 identities. 702 identities are used as the training set and the remaining 1,110 identities as

¹<https://github.com/Cysu/open-reid>

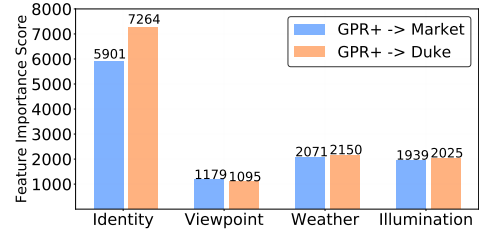


Fig. 4. Feature importance of XGBoost with Feature Importance Score (higher is more important).

the testing set. For attribute metric, we use the feature space provided by the 19-layer VGG network, and empirically set $w_l=0.2$ in Eq. 2, $\alpha=0.9$ and $\beta=1$ in Eq. 3. During the re-ID evaluation, we adopt a initializing model pre-trained on ImageNet [23] and follow the training procedure in [24, 25]. All timings for training use one Nvidia Tesla P100 GPU on Pytorch framework [26].

3.2. Evaluation of Attribute Importance

In this section, we evaluate the importance of different attributes on a basic re-ID system. According to Fig. 4, it can be easily observed that **identity** accounts for the largest proportion when performing cross-domain re-ID task from *GPR+* to Market-1501, following by **weather**, **illumination** and **viewpoint**. Typically, this conclusion is consistent with results on DukeMTMC-reID dataset, which helps us fully understand the role of different attributes.

3.3. Fine-grained Attribute Analysis

In this part, we further explore the impacts of different attributes on a basic re-ID system. There are several observations which can be made as follows.

First, we can easily observe that using more IDs will noticeably improve the re-ID performance. For example, as presented in Table 1 and Table 2, we can only achieve a per-

Table 1. Ablation study results (%) on Market dataset. ✓ indicates using all conditions in each attribute. **Bold** denotes the best.

| #identity | #box | #viewpoint | #weather | #illumination | Time (h)↓ | mAP↑ | R@1↑ | R@5↑ |
|-----------|---------|---------------------------------|--------------|---------------------------|-------------|-------------|-------------|-------------|
| 100 | 58,800 | ✓ | ✓ | ✓ | 8.3 | 4.8 | 15.4 | 29.3 |
| 400 | 235,200 | ✓ | ✓ | ✓ | 31.0 | 13.5 | 35.7 | 53.8 |
| 800 | 470,400 | ✓ | ✓ | ✓ | <u>61.5</u> | 17.4 | 41.8 | 60.6 |
| 800 | 134,400 | ✓ | sunny,clouds | ✓ | 18.0 | 19.7 | 43.3 | 59.8 |
| 800 | 201,600 | ✓ | ✓ | forenoon,noon,afternoon | 26.5 | 18.6 | 41.4 | 58.8 |
| 800 | 235,200 | 30°, 60°, 90°, 120°, 150°, 330° | ✓ | ✓ | 30.5 | 19.3 | 44.1 | 62.4 |
| 800 | 28,800 | 30°, 60°, 90°, 120°, 150°, 330° | sunny,clouds | forenoon, noon, afternoon | 4.5 | 17.4 | 40.3 | 56.7 |

Table 2. Ablation study results (%) on Duke dataset. ✓ indicates using all conditions in each attribute. **Bold** denotes the best.

| #identity | #box | #viewpoint | #weather | #illumination | time (h)↓ | mAP↑ | R@1↑ | R@5↑ |
|-----------|---------|------------------------------|----------------|-------------------------|-------------|-------------|-------------|-------------|
| 100 | 58,800 | ✓ | ✓ | ✓ | 8.3 | 4.3 | 13.8 | 24.2 |
| 400 | 235,200 | ✓ | ✓ | ✓ | 30.6 | 10.7 | 26.3 | 38.2 |
| 800 | 470,400 | ✓ | ✓ | ✓ | <u>60.7</u> | 15.1 | 33.5 | 48.0 |
| 800 | 134,400 | ✓ | foggy,blizzard | ✓ | 18.0 | 17.8 | 33.8 | 48.3 |
| 800 | 201,600 | ✓ | ✓ | forenoon,noon,afternoon | 26.5 | 18.8 | 38.2 | 52.3 |
| 800 | 235,200 | 60°,90°,120°,150°,180°, 210° | ✓ | ✓ | 30.6 | 17.2 | 37.7 | 52.2 |
| 800 | 28,800 | 60°,90°,120°,150°,180°, 210° | foggy,blizzard | forenoon,noon,afternoon | 4.4 | 13.3 | 25.7 | 39.1 |

formance of 4.8% and 4.3% in mAP accuracy when tested on Market-1501 and DukeMTMC-reID, respectively. Moreover, adding IDs to 800 as supervised information notably improves the re-ID accuracy, leading to **+12.6%** and **+10.8%** improvement in mAP accuracy.

Second, we perform greedy search for the objective of **smaller loss** (see Fig. 3) to obtain optimal attribute. For instance, with constraint of 800 IDs, using *sunny&clouds* and *foggy&blizzard* bring about +2.3% and +2.7% more improvement than using all weather conditions when tested on Market-1501 and DukeMTMC-reID respectively; using *forenoon,noon,afternoon* as induction can lead an additional improvement of +1.2% and +3.7% in mAP accuracy. Furthermore, we can achieve a significant improvement of **+2.3%** and **+4.2%** in rank-1 accuracy with some critical viewpoints. Intuitively, by taking all attributes into consideration, we can obtain the rank-1 accuracy to **40.3%** and **25.7%** on Market-1501 and DukeMTMC-reID respectively. Meanwhile, fast training is our second main advantage, e.g., training time will be considerably reduced by **13×** (61.5 vs. 4.5 hours) and **14×** (60.7 vs. 4.4 hours), respectively, leading a more computationally efficient training on re-ID backbone.

3.4. Discussion

To go even further, we gave an explanation about several interesting phenomena observed during the experiment.

Firstly, as depicted in Table 1 and Table 2, it can be observed easily that with only one individual attribute constraint can obtain a more satisfactory performance compared with using all images, no matter which attribute is adopted. It probably because that our attribute analysis strategy can find im-

portant attributes and reduce the redundancy of training-set.

Secondly, a simply combination of several attributes cannot always guarantee the most optimal attributes for re-ID task, and the mutual influence of multiple attributes should be considered for fine-grained attribute analysis in the future.

Third and importantly, by simply increasing the scale of training examples does not automatically bring notable performance gains. However, using more IDs as training samples is always beneficial to the system. Based on this observation, we can drastically improve our performance by enhancing the diversity of train-set instead of the scale of dataset. In summary, our fine-grained attribute analysis strategy is similar to [13], but our solution is entirely parameter free, which makes it more flexible and adaptable.

4. CONCLUSION

In this paper, we addressed a critically important problem in person re-identification which has received little attention thus far - fine-grained attribute analysis. First, we upgrade and enrich the previous GPR dataset to *GPR+*, which provide orthogonal distribution and eliminate the correlation between different attributes. Based on *GPR+*, we introduce a fine-grained analysis strategy to quantitatively assess the importance of attributes, then conduct comprehensive experiments to explore the influence of various attributes on re-ID task. Nevertheless, this research is very meaningful since it will provide guidance for us to construct a high-quality re-ID dataset. In closing, we hope that our new experimental evidence about *GPR+* and its role will shed light into potential future directions for the community to move forward.

5. REFERENCES

- [1] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang, “Deepreid: Deep filter pairing neural network for person re-identification,” in *CVPR*, 2014, pp. 152–159.
- [2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, “Deep residual learning for image recognition,” in *CVPR*, 2016, pp. 770–778.
- [3] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna, “Rethinking the inception architecture for computer vision,” in *CVPR*, 2016, pp. 2818–2826.
- [4] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger, “Densely connected convolutional networks,” in *CVPR*, 2017, pp. 4700–4708.
- [5] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian, “Scalable person re-identification: A benchmark,” in *ICCV*, 2015, pp. 1116–1124.
- [6] 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*. Springer, 2016, pp. 17–35.
- [7] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian, “Person transfer gan to bridge domain gap for person re-identification,” in *CVPR*, 2018, pp. 79–88.
- [8] Suncheng Xiang, Yuzhuo Fu, Mingye Xie, Zefang Yu, and Ting Liu, “Unsupervised person re-identification by hierarchical cluster and domain transfer,” *MULTIMEDIA TOOLS AND APPLICATIONS*, 2020.
- [9] Igor Barros Barbosa, Marco Cristani, Barbara Caputo, Aleksander Rognhaugen, and Theoharis Theoharis, “Looking beyond appearances: Synthetic training data for deep cnns in re-identification,” *Computer Vision and Image Understanding*, vol. 167, pp. 50–62, 2018.
- [10] Slawomir Bak, Peter Carr, and Jean-Francois Lalonde, “Domain adaptation through synthesis for unsupervised person re-identification,” in *ECCV*, 2018, pp. 189–205.
- [11] Xiaoxiao Sun and Liang Zheng, “Dissecting person re-identification from the viewpoint of viewpoint,” in *CVPR*, 2019, pp. 608–617.
- [12] Yanan Wang, Shengcai Liao, and Ling Shao, “Surpassing real-world source training data: Random 3d characters for generalizable person re-identification,” 2020.
- [13] Yue Yao, Liang Zheng, Xiaodong Yang, Milind Naphade, and Tom Gedeon, “Simulating content consistent vehicle datasets with attribute descent,” in *ECCV*, 2020.
- [14] Zhedong Zheng, Liang Zheng, and Yi Yang, “Unlabeled samples generated by gan improve the person re-identification baseline in vitro,” in *ICCV*, 2017, pp. 3754–3762.
- [15] Alexander Hermans, Lucas Beyer, and Bastian Leibe, “In defense of the triplet loss for person re-identification,” *arXiv preprint arXiv:1703.07737*, 2017.
- [16] Zhilu Zhang and Mert Sabuncu, “Generalized cross entropy loss for training deep neural networks with noisy labels,” in *NIPS*, 2018, pp. 8778–8788.
- [17] Suncheng Xiang, Yuzhuo Fu, Guanjie You, and Ting Liu, “Unsupervised domain adaptation through synthesis for person re-identification,” in *ICME*. IEEE, 2020, pp. 1–6.
- [18] Leon A Gatys, Alexander S Ecker, and Matthias Bethge, “Image style transfer using convolutional neural networks,” in *CVPR*, 2016, pp. 2414–2423.
- [19] Karen Simonyan and Andrew Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556*, 2014.
- [20] Leon A Gatys, Alexander S Ecker, Matthias Bethge, Aaron Hertzmann, and Eli Shechtman, “Controlling perceptual factors in neural style transfer,” in *CVPR*, 2017, pp. 3985–3993.
- [21] Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang, “Universal style transfer via feature transforms,” in *NIPS*, 2017, pp. 386–396.
- [22] Tianqi Chen and Carlos Guestrin, “Xgboost: A scalable tree boosting system,” in *KDD*, 2016, pp. 785–794.
- [23] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *CVPR*, 2009, pp. 248–255.
- [24] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang, “Bag of tricks and a strong baseline for deep person re-identification,” in *CVPRW*, 2019, pp. 0–0.
- [25] Hao Luo, Wei Jiang, Youzhi Gu, Fuxu Liu, Xingyu Liao, Shenqi Lai, and Jianyang Gu, “A strong baseline and batch normalization neck for deep person re-identification,” *IEEE Transactions on Multimedia*, 2019.
- [26] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al., “Pytorch: An imperative style, high-performance deep learning library,” in *NIPS*, 2019, pp. 8026–8037.