VIPerson: Flexibly Generating Virtual Identity for Person Re-Identification

Supplementary Material

Figure S1. Images of virtual identities in VIPerson.

Figure S2. Images of hard identities in VIPerson.

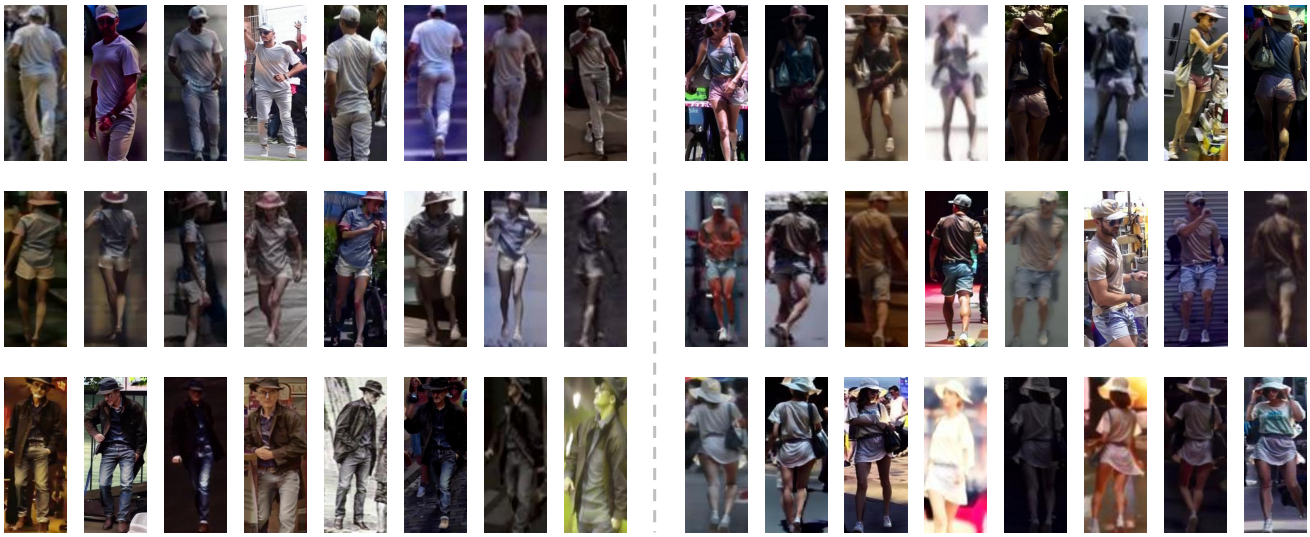

Figure S3. Images of intra-ID groups in VIPerson.

A. More Images in VIPerson

To showcase the diverse virtual identities, cross-camera variations, and camera-realistic style of our dataset, we present a collection of sample images from the dataset.

Diverse Virtual Identities. In Figure S1, we present fifteen randomly synthesized pedestrians. Our synthesized pedestrians exhibit diverse appearance variations while closely resembling the clothing styles of real-world individuals, which benefits the model in recognizing real-world pedestrians.

Hard Identities. We show eight groups of hard identities in Figure S2. In each group, the three identities share similar attire but exhibit slight variations in clothing colors or accessories. By providing the ReID model with hard identities, it can learn more discriminative features, thereby enhancing its performance on the test set.

Intra-ID Groups. As shown in Figure S3, we display intra-ID groups for six identities. In each intra-ID group, we obtain diverse variations of poses and scenes, and introduce a

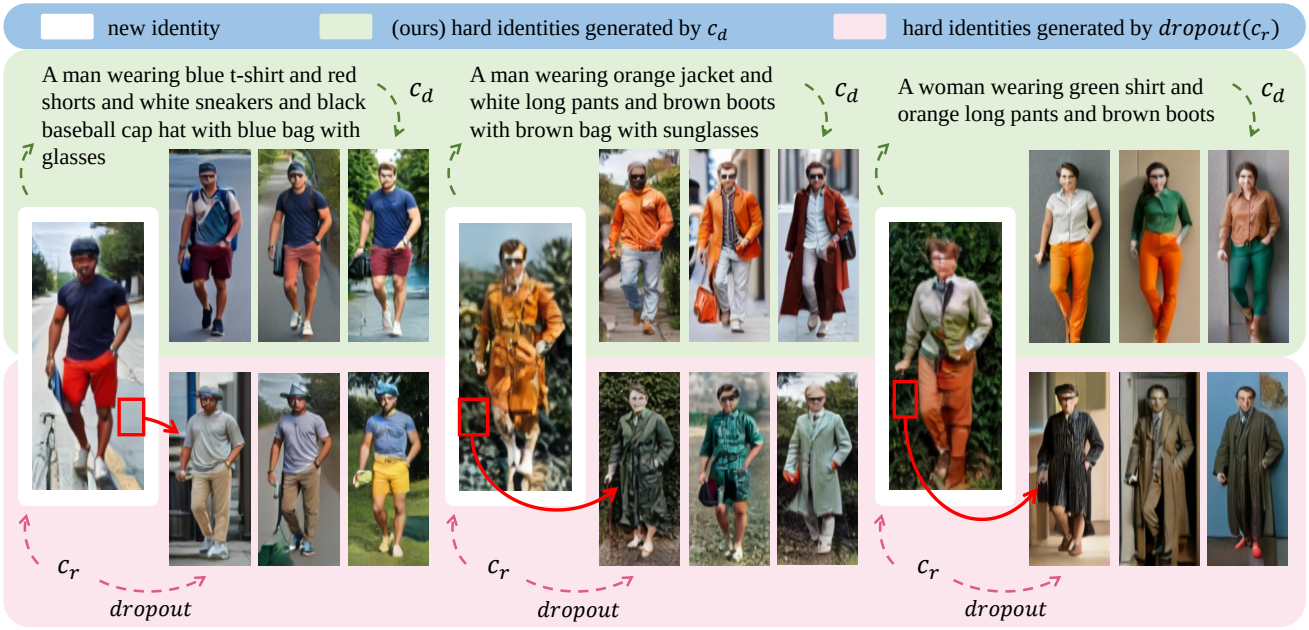

Figure S4. Effect of different methods to generate hard identities. In our method, the description of each identity contains four main attributes, including genders/ages, up dress, down dress, and shoes, along with three accessory attributes including hat, glasses, and bag.

variety of authentic camera styles. These cross-camera variations and realistic style make VIPerson more closed to the real-world scenarios.

B. Analysis of Hard Identity Generation

We conducted an ablation study on the method of synthesizing hard identities. A direct way to generate hard identities is to erase part of the random embedding c_r as the condition. We generate batches of hard identities in this way as shown in Figure S4, and hard sample identities generated by $dropout(c_r)$ show changes related to the background. This is due to c_r containing information about the image background, and direct injections of part erased c_r introduce disturbances during model sampling. Differently, in our method, we first obtain the description d_I of each new identity, and erase part of the textual feature as c_d to inject into the flexible identity generator and generate hard identities. As shown in Figure S4, in our method, the changes between hard identities mainly occur in the pre-defined attributes, as generating descriptions for new identities filters the unimportant information of the identity images.

C. Direct Transfer Results on More ReID Models

We also explore the performance of different datasets in the direct transfer task on other ReID methods, including CLIP-based CLIP-ReID [6] and ViT-based TransReID [3]. In CLIP-ReID, we resize the image to 256×128 and set the batch size as 128. We adopt the Adam optimizer for training, with the base learning rate set to 3.5×10^{-4}. For TransReID, we resize the image to 384×128 and set the batch size as 128. We adopt the SGD optimizer for training, with the base learning rate set to 0.001. The weight of triplet loss is set to 2.0.

Table S1. Direct transfer results on CLIP-ReID [6]. The fully-supervised results are shown in *italics*, the best direct transfer results are indicated in **bold**, and the second-best results are underlined.

Dataset Type	Training Dataset	MSMT17		Market1501	
		rank-1	mAP	rank-1	mAP
Real	MSMT17	*84.4*	*63.0*	52.7	27.2
	Market1501	10.0	3.3	*95.7*	*89.8*
Synthetic	SyRI	2.3	0.5	2.3	0.8
	PersonX	3.7	1.1	30.0	12.4
	RandPerson	10.3	3.1	50.0	24.7
	UnrealPerson	15.9	5.1	51.8	27.6
	VIPerson	23.6	**6.9**	**62.8**	34.9
	VIPerson*	**24.7**	**6.9**	62.6	**35.6**

As the result in Table S1 and Table S2 shows, our dataset outperforms other datasets, demonstrating its effectiveness across different ReID methods.

D. Scaling Law Research

The scaling law suggests that the performance gain from increasing dataset size is constrained by model capacity [1, 4, 5]. When the model size is fixed, adding more data

Table S2. Direct transfer results on TransReID [3]. The fully-supervised results are shown in *italics*, the best direct transfer results are indicated in **bold**, and the second-best results are underlined.

Dataset Type	Training Dataset	MSMT17		Market1501	
		rank-1	mAP	rank-1	mAP
Real	MSMT17	*84.2*	*65.8*	70.7	46.9
	Market1501	39.4	17.1	*94.9*	*87.6*
Synthetic	MALS	1.8	0.7	16.2	5.6
	SyRI	41.3	16.9	68.4	42.6
	PersonX	28.0	10.5	55.5	30.0
	RandPerson	43.6	18.9	70.7	45.7
	VIPerson	47.0	19.0	75.8	49.5
	VIPerson*	**48.8**	**19.2**	**76.9**	**51.7**

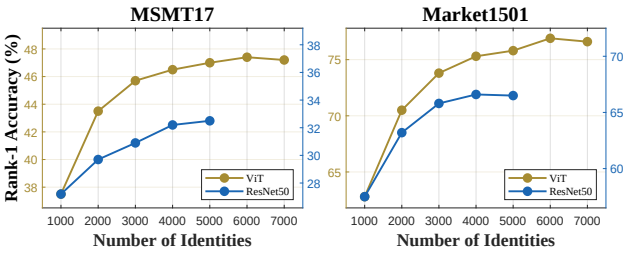

Figure S5. Effect of different VIPerson identities on different models. For ResNet50, we adopt BoT, and for ViT, we adopt TransReID.

yields diminishing returns beyond a certain point. To verify the impact of scaling law on our synthetic dataset, we evaluate the effect of the number of identities on performance using a larger model architecture. As the result in Figure S5 shown, when the number of identities reaches 5000, the model performance on ResNet50 has already plateaued, while the performance on ViT continues to improve. This indicates that for different model capacities, the scalability of our dataset provides a greater advantage in unlocking the potential of larger models.

E. Camera Diversity Research

To demonstrate the effect of diverse cross-camera variations, we modified the scene composition of the dataset. We generate 3000 virtual identities without hard identities and modify the scenes in each intra-ID group with different settings. For each image, we transfer the realistic style from Market1501. As shown in Figure S6, as the diversity in scene images increased, the performance keeps increasing. This result demonstrates the importance of diverse cross-camera variations in the ReID model training.

F. More Details of Experiments

Real ReID Dataset. To evaluate the generalization ability of our VIPerson, we train the ReID model on our synthetic dataset and conduct experiments on two real-world datasets,

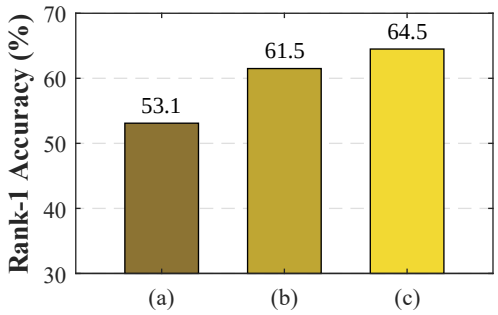

Figure S6. Effect of our dataset with different camera scales on Market1501. The experiment is based on BoT. In (a), we use scene images of 6 cameras from Market1501. In (b), we use scene images of 15 cameras from MSMT17. In (c), we use scene images sampled from LUPerson containing 46,260 cameras.

i.e. MSMT17[7] and Market1501[9]. MSMT17 contains 4,101 identities with 126,441 images under 15 scenes. The training set includes 1,041 identities with 32,621 images, and the test set includes 3,060 identities with 93,820 images. Market1501 contains 1,501 identities with 36,036 images under 6 real scenes. The training set includes 751 identities with 12,936 images and the test set includes 750 identities with 23,100 images. Besides, to obtain pose, scene, and camera-style information from the real world, we utilize the large-scale unlabeled ReID dataset LUPerson[2], which includes 4,180,243 unlabeled real-world person images under 46,260 scenes.

Training Details of Fine-tuning and Domain Adaptation. We performed fine-tuning and domain adaptation experiments based on the supervised fine-tune setting and the unsupervised domain adaptation setting. For supervised fine-tuning, we use BoT and first pre-train the model on the source domain following the direct transfer setting. When fine-tuning at MSMT17 and Market1501 training sets, the learning rate decreased by a factor of 0.1 at the 40^{th} and 70^{th} epochs. For unsupervised domain adaptation setting, we use ISE[8]. After pre-training on the source domain, when training at the real-world datasets, we resize the image to 256×128 and set the batch size as 256. We adopt Adam optimizer with the initial learning rate of 3.5×10^{-4}.

References

[1] Lijie Fan, Kaifeng Chen, Dilip Krishnan, Dina Katabi, Phillip Isola, and Yonglong Tian. Scaling laws of synthetic images for model training... for now. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7382–7392, 2024. 2

[2] Dengpan Fu, Dongdong Chen, Jianmin Bao, Hao Yang, Lu Yuan, Lei Zhang, Houqiang Li, and Dong Chen. Unsupervised pre-training for person re-identification. In *Proceedings*

of the IEEE/CVF conference on computer vision and pattern recognition, pages 14750–14759, 2021. 3

[3] Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. Transreid: Transformer-based object re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15013–15022, 2021. 2, 3

[4] Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md Mostofa Ali Patwary, Yang Yang, and Yanqi Zhou. Deep learning scaling is predictable, empirically. *arXiv preprint arXiv:1712.00409*, 2017. 2

[5] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020. 2

[6] Siyuan Li, Li Sun, and Qingli Li. Clip-reid: exploiting vision-language model for image re-identification without concrete text labels. In *Proceedings of the AAAI conference on artificial intelligence*, pages 1405–1413, 2023. 2

[7] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 79–88, 2018. 3

[8] Xinyu Zhang, Dongdong Li, Zhigang Wang, Jian Wang, Errui Ding, Javen Qinfeng Shi, Zhaoxiang Zhang, and Jingdong Wang. Implicit sample extension for unsupervised person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7369–7378, 2022. 3

[9] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision*, pages 1116–1124, 2015. 3