Supplementary Materials of "Transferable, Controllable, and Inconspicuous Adversarial Attacks on Person Re-identification With Deep Mis-Ranking"

Hongjun Wang[1]* Guangrun Wang[1]* Ya Li[2] Dongyu Zhang[2] Liang Lin[1,3]†
[1]Sun Yat-sen University [2]Guangzhou University [3]DarkMatter AI
[1]{wanghq8,wanggrun,zhangdy27}@mail2.sysu.edu.cn [2]liya@gzhu.edu.cn [3]linliang@ieee.org

1. Appendix A: More Visualization of the Destruction to the ReID system

We present more results from both Market1501 and CUHK03 datasets by exhibiting the Rank-10 matches from the target ReID system before and after an adversarial attack in Figure 1.

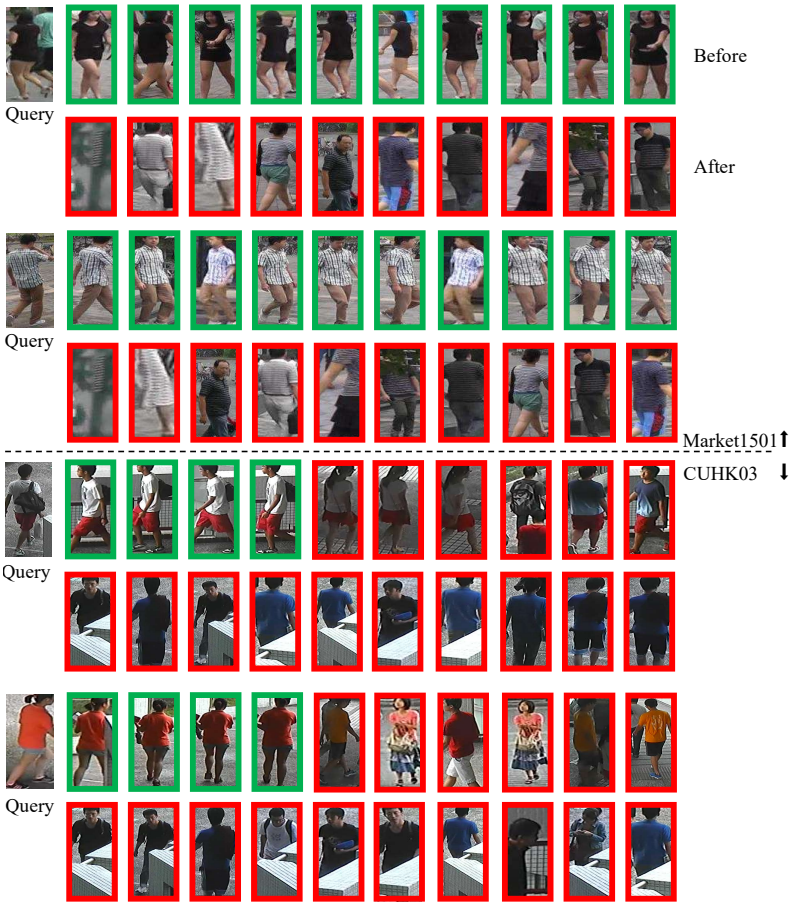


Figure 1. The rank-10 predictions of AlignedReID [2] (one of the state-of-the-art ReID models) before and after our attack on Market-1501 and CUHK03. We display the gallery images according to their rank accuracies returned from AlignedReID [3] model. The green boxes are the correctly matched images, while the red boxed are the mismatched images. Only top-10 gallery images are visualized.

*Equal contribution
†Corresponding author

2. Appendix B: Visualization of Sampling Mask

We further visualize the noise layout in Figure 2 to explore the interpretability of our attack in ReID. However, it is hard to get intuitive information from a few sporadic samples.

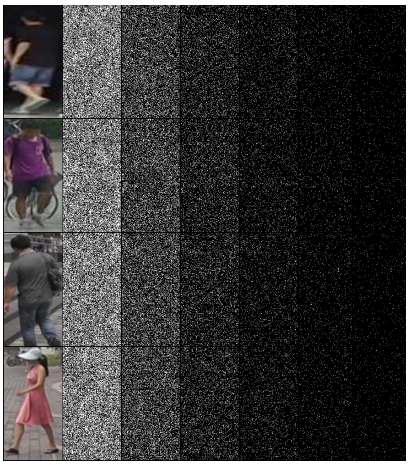


Figure 2. The layout of the noise. Column 1 shows the query images. Column 2-7 are the noise layouts with different noise numbers, i.e., $\{1/2, 1/4, 1/8, 1/16, 1/32, 1/64\} \times HW$.

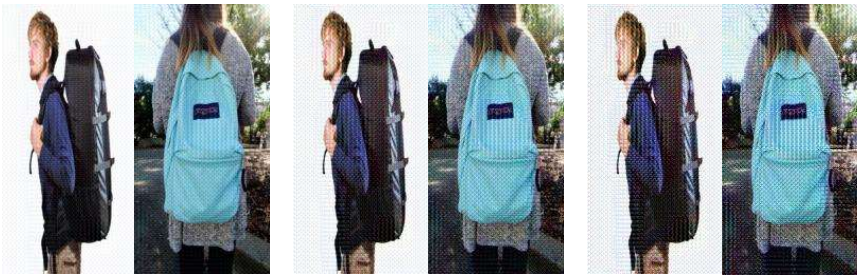
3. Appendix C: More Details of Attacking

The experiments are performed on a workstation with an NVIDIA Titan X GPU with 12GB GPU memory. The protocols of target models are the same as their official releases. The minibatch size is 32 for 50 epochs. The filters in each multi-column network in our multi-stage discriminator are 4×4. The weights were initialized from a truncated normal distribution with mean 0 and standard deviation 0.02 if the layer is not followed by the spectral normalization. The whole network is trained in an end-to-end fashion using Adam [1] optimizer with the default setting, and the learning rate is set to 2e-4.

As for the balanced factors in our full objective, the selection of β mainly depends on how worse the visual quality we can endure. In our case, we fix $\alpha = 2$ and $\beta = 5$ in the main experiments but β ranging from 5 to 0 when we gradually cut down the number of attacked pixels.

4. Appendix D: Attacking to the Real-world System

To the best of our knowledge, no ReID API is available online for the public to attack. Fortunately, we find a real-world system, Human Detection and Attributes *API provided by Baidu (https://ai.baidu.com/tech/body/attr)*, that can examine the capacity of our attacker. Actually, our attacker is able to perform an attribute attack, which is crucial in ReID. We randomly pick up person images *with backpack* from Google and generate noise to them using our attacker. This results in the adversarial examples listed in Fig. 3. When uploading these adversarial examples to the Baidu API, the API misclassified them as persons *without backpack*, implying the real-world system has been fooled by our attacker.



(a) eps=20 (b) eps=30 (c) eps=40

Figure 3. Visualization of several adversarial examples with ε=20,20,30,30,40,40 respectively which successfully attack Baidu AI service.

References

[1] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 2

[2] 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*, pages 480–496, 2018. 1

[3] Xuan Zhang, Hao Luo, Xing Fan, Weilai Xiang, Yixiao Sun, Qiqi Xiao, Wei Jiang, Chi Zhang, and Jian Sun. Alignedreid: Surpassing human-level performance in person re-identification. *CoRR*, 2017. 1