IntensPure: Attack Intensity-aware Secondary Domain Adaptive Diffusion for Adversarial Purification (Appendix)

Figure 1: Illustration of the top 10 retrieval results for query images on Market1501 dataset. The SIM and $INCON$ are components of ID stability and attribute inconsistency, respectively. The SIM sum and $INCON$ sum represent the summation of their respective result vectors. As the attack intensity increases, there is a tendency for the SIM value to decrease, while the $INCON$ value tends to increase.

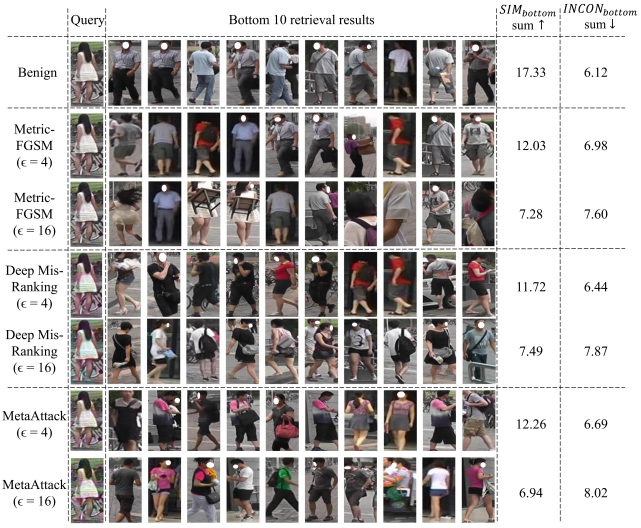

Figure 2: Illustration of the bottom 10 retrieval results for query images on Market1501 dataset. The SIM and $INCON$ are components of ID stability and attribute inconsistency, respectively. The SIM sum and $INCON$ sum represent the summation of their respective result vectors. As the attack intensity increases, there is a tendency for the SIM value to decrease, while the $INCON$ value tends to increase.

A Analysis of the Attack Intensity Estimator

A.1 Visualizing Retrieval Results

Figure 1 illustrates the inconsistency in retrieval results caused by adversarial attacks on the query set of the Market1501 dataset [Zheng *et al.*, 2015]. We covered the face area for privacy protection. In each row, the first image represents the query image, followed by the top 10 retrieved images. For benign queries, the top 10 retrieval results typically align with the query image identity (ID). Even if the retrieved image within the top 10 does not match the ID, it still returns a person who is visually very similar. However, adversarial attacks on person re-identification (Metric-FGSM [Bai *et al.*, 2020], Deep Mis-Ranking [Wang *et al.*, 2020], and MetaAttack [Yang *et al.*, 2022]), which can be abbreviated as person re-ID attacks, cause the retrieval system to return images unrelated to the query image in the top ranking list, such as retrieval results entirely different from the target person, or the wrong images, such as irrelevant objects like body parts or bicycles. Furthermore, the top 10 retrieved images not only become irrelevant to the query but also exhibit reduced consistency within the results.

The sum values of SIM_{top} and $INCON_{top}$, which are the components of ID stability and attribute inconsistency, are also shown in Figures 1 and 2. For the benign query, the SIM_{top} value is relatively higher than the values that come from perturbed queries and tends to decrease as the attack intensity increases. Further, the $INCON_{top}$ value, which represents the error in attribute recognition, is close to 0 for benign queries, whereas the $INCON_{top}$ value tends to have large values according to the attack intensity.

Figure 2 shows that the bottom 10 ranking list exhibits consistency between retrieval results for benign queries, similar to the top 10 retrieval results. However, when the query is perturbed, the bottom 10 ranking list exhibits inconsistency between retrieval results. The SIM_{bottom} and $INCON_{bottom}$ can be obtained from the bottom 10 ranking list and they show similar trends to SIM_{top} and $INCON_{top}$. By extracting features from these secondary effects of adversarial attacks, we can effectively detect attacks and estimate their intensities.

ID stability (SIM)			Attribute inconsistency ($INCON$)			Acc ↑	MAE ↓
Q-R	top	$bottom$	Q-$R1$	top	$bottom$		
✓	-	-	-	-	-	92.13	4.094
✓	✓	-	-	-	-	94.89	3.508
✓	✓	✓	-	-	-	96.91	3.154
-	-	-	✓	-	-	77.14	7.887
-	-	-	✓	✓	-	89.18	5.337
-	-	-	✓	✓	✓	90.82	5.253
✓	-	-	✓	-	-	93.74	2.019
✓	✓	-	✓	✓	-	97.61	1.195
✓	✓	✓	✓	✓	✓	**98.85**	**0.806**

Table 1: Ablation study on attack intensity estimator in terms of detection accuracy (Acc) and mean absolute error (MAE) on Market1501 dataset attacked by Metric-FGSM. The hyphen - indicates without, and the check ✓indicates with the method. The best results are in bold.

Number of rank images	1	5	10	15	20
Acc ↑	81.65	97.02	**98.85**	98.77	98.49
MAE ↓	3.113	1.214	**0.806**	0.820	0.856

Table 2: Comparison of adversarial detection accuracy (Acc) and mean absolute error (MAE) according to the number of top and bottom rank images on Market1501 dataset attacked by Metric-FGSM. The best results are in bold.

A.2 Ablation Study of ID Stability and Attribute Inconsistency Components

Table 1 illustrates an ablation study to evaluate the contribution of the components used for adversarial attack intensity estimation. First, we investigated the performance of ID stability and attribute inconsistency individually. Detection accuracy refers to the binary classification accuracy for detecting whether an image is an adversarial example and mean absolute error is calculated using the estimated attack intensity and ground truth attack intensity. In both individual and joint experiments, the performance is more improved by additionally considering features derived from the consistency between retrieval results themselves than considering only the features between the query and the top retrieval results ($Q-R$ and $Q-R1$). Using the features derived from the top ranking list, the detection accuracy is achieved at 94.89%, 89.18%, and 97.61%, respectively. Additionally, using both the features derived from the bottom ranking list and the top ranking list, the detection performance is improved by 2.02%p, 1.64%p, and 1.24%p, respectively, and the attack intensity estimation error is also reduced by 0.354, 0.084, and 0.389, respectively. This suggests the effectiveness of leveraging the bottom ranking list. Attribute inconsistency exhibits relatively lower performance than ID stability when used alone due to the inherent variability of person attributes. However, when used in conjunction with ID stability, attribute inconsistency significantly improves the performance of both adversarial detection and attack intensity estimation.

Number of BDCT coefficients	1	3	6	10
Rank-1 accuracy (%)	49.12	60.87	**66.65**	65.63

Table 3: Rank-1 accuracy (%) variation with the number of BDCT coefficients used for secondary image generation of IntensPure, evaluated on Market1501 dataset attacked by Metric-FGSM with $\epsilon = 8$. The best results are in bold.

BDCT block size	Rank-1 accuracy (%) ↑	FLOPs (G) ↓	Params (M) ↓	Time (ms) ↓
4×4	66.58	154	**751**	98
8×8	**66.65**	39	**751**	59
16×16	60.87	**10**	**751**	**34**

Table 4: Rank-1 accuracy (%) and complexity variation with BDCT block size of IntensPure, evaluated on Market1501 dataset attacked by Metric-FGSM with $\epsilon = 8$. The complexity is assessed using a single diffusion time step without the estimator. The best results are in bold.

B Estimator Performance Comparison for the Number of Retrieval Results

Table 2 examines the influence of rank range on the performance of both adversarial detection and attack intensity estimation. Rank range 1 refers to using only the top 1 and bottom 1 retrieval results. In general, performance improves as the range increases up to 10. We achieve the highest accuracy of 98.85% and the lowest estimation error of 0.806, when the rank range is 10. However, extending the rank range beyond 10 leads to a decline in performance due to a weakened correlation between the retrieved images.

C Performance Evaluation of BDCT Coefficient Counts and Block Size

Table 3 represents the variation of rank-1 accuracy according to the varying number of block discrete cosine transform (BDCT) coefficients employed for generating secondary images within the Market1501 dataset attacked by Metric-FGSM with attack intensity $\epsilon = 8$. The rank-1 accuracy is highest when the six lowest-frequency BDCT coefficients were used for generating secondary images. This result indicates that there is a trade-off between using sufficient coefficients to maintain benign information and using as few coefficients as possible to minimize the impact of attacks. Based on the experimental results, we selected the 6 BDCT coefficients for secondary domain purification.

Table 4 shows the variation of rank-1 accuracy and complexity variation according to the BDCT block sizes (4×4, 8×8, 16×16). As the block size becomes smaller, the number of blocks increases, leading to higher resolution in the secondary image. Therefore, when the block size is 4×4, the complexity is the largest, resulting in a slight decline in re-ID performance. On the other hand, a block size of 16×16 reduces complexity but marginally lowers re-ID performance, as it relies on limited information. Selecting an 8×8 block size maximizes the enhancement in rank-1 accuracy, as it optimally utilizes information.

Figures 3 and 4 show the visualization of this secondary domain adaptive diffusion process for the Market1501 dataset

Filter size	3	5	7	9
Rank-1 accuracy (%)	66.37	**66.65**	66.60	66.47

Table 5: Rank-1 accuracy (%) variation with the inter-block directional filter size of IntensPure, evaluated on Market1501 dataset attacked by Metric-FGSM with $\epsilon = 8$. The best results are in bold.

σ_r / σ_s	0.05 / 1	0.10 / 2	0.15 / 3	0.20 / 4
Rank-1 accuracy (%)	66.46	66.51	**66.65**	66.43

Table 6: Rank-1 accuracy (%) variation with the inter-block directional filter sigma values of IntensPure, evaluated on Market1501 dataset attacked by Metric-FGSM with $\epsilon = 8$. The best results are in bold.

and the DukeMTMC dataset [Ristani *et al.*, 2016], respectively. Along the diffusion process, the perturbed secondary images pass through the forward diffusion process, where random noise is injected. Subsequently, the denoising process is applied to remove both random noise and perturbations. In this process, the color change induced by the adversarial attack is erased. Finally, inter-block directional filters enhance the correlation between blocks, effectively removing remaining perturbations. After undergoing these diffusion processes, the purified secondary images are reconstructed into a spatial domain image via an inverse BDCT process and further enhanced using a perturbation constraint set. The reconstructed image is subsequently fed into the re-ID model.

D Parameters Analysis of Inter-block Filter

We investigate the influence of inter-block directional filter size and sigma values (σ_r and σ_s) on performance. Concretely, the experiments are conducted on the Market1501 dataset attacked by Metric-FGSM with attack intensity $\epsilon = 8$.

Table 5 presents rank-1 accuracy variations according to the inter-block directional filter size (3, 5, 7, and 9). The experimental results indicate that the highest rank-1 accuracy is achieved with the size 5. Through these empirical results, we determined 5 as the optimal filter size.

Table 6 evaluates effects by varying the setting of the range and spatial standard deviations (σ_r and σ_s) included in inter-block directional filter. In this experiment, the highest accuracy of 66.65% is achieved with σ_r and σ_s at 0.15 and 3. Based on these results, we experimentally choose these optimal settings.

E Additional Ablation Study

We conduct an additional ablation study to evaluate the effectiveness of each module by removing only one module at a time on the Market1501 dataset attacked by Metric-FGSM with various ϵ values (0, 4, 8, 12, and 16). Table 7 shows the rank-1 accuracy of IntensPure with different configurations. The first row shows the accuracy of the complete model, in which all modules are activated. The second row shows the rank-1 accuracy when only the attack intensity estimation module is deactivated and the time step is fixed at 50. Additionally, the perturbation constraint set is not available when the attack intensity estimator is deactivated, as it relies

Attack Intensity Estimation	Secondary Image Generation	Directional Diffusion	Inter-Block Directional Filter	Perturbation Constraint Set	Rank-1 Accuracy
✓	✓	✓	✓	✓	70.04
	✓	✓	✓		64.01
✓				✓	64.20
✓	✓		✓	✓	68.64
✓	✓	✓		✓	69.26
✓	✓	✓	✓		68.31

Table 7: Ablation study of IntensPure, removing one module at a time. Model performances are evaluated on Market1501 dataset attacked by Metric-FGSM with various ϵ values (0, 4, 8, 12, and 16). The check ✓ indicates with the method.

on the estimated attack intensity. As expected, the accuracy decreases when the attack intensity estimation module is deactivated due to its fixed-step purification of all images, including benign ones, similar to previous diffusion-based purification methods. However, as presented in the main paper, Diffpure [Nie *et al.*, 2022] achieves 55.48%, GNSP [Lee and Kim, 2023] achieves 62.53%, and IntensPure still outperforms these state-of-the-art models with a rank-1 accuracy of 64.01%. This demonstrates that IntensPure can achieve notable performance gains with secondary domain adaptive diffusion alone, without relying on optimal purification strength estimation. The third row shows the performance of IntensPure when only the attack intensity estimation module for the time step optimization and perturbation constraint set are activated. This shows that IntensPure can achieve significant performance gains by adjusting purification strength. The difficulty in performing directional diffusion in the spatial domain comes from the deactivation of the secondary image generation module, because individual channels in the spatial domain is difficult to be considered that they have specific directional characteristics. Furthermore, the inter-block directional filter is not applicable to the spatial domain images because they do not need to consider the correlation between blocks. The fourth and fifth rows show the performance of IntensPure when directional diffusion and inter-block directional filter are deactivated, respectively. Directional diffusion leads to a performance improvement of 1.36%p, and inter-block directional filter leads to a performance improvement of 0.74%p. The sixth row shows that the perturbation constraint set further improves the performance by 1.79%p by suppressing over-purification and preserving the original information.

F Visual Comparison of Retrieval Results through Purification

For a qualitative comparison, we present the visual retrieval results of IntensPure and state-of-the-art models. Figures 5, 6, 7, 8, 9, and 10 illustrate the top 10 retrieval results on Market1501 dataset, which is attacked by Metric-FGSM, Deep Mis-Ranking, and MetaAttack with various attack intensities. The results demonstrate that IntensPure consistently outperforms other state-of-the-art models in terms of rank-1 accuracy, under various attack types and intensities. Its superiority is further presented via the secondary domain adaptive diffusion. While other models struggle to restore original colors, leading to inconsistent retrieval results, IntensPure effec-

Method	Person Re-ID Attack		Person Re-ID Attack + Attribute Recognition Attack	
	Acc ↑	MAE ↓	Acc ↑	MAE ↓
MEAAD	98.50	3.189	99.20	2.714
Ours	99.55	0.769	99.13	1.850

Table 8: Comparison of adversarial detection accuracy (Acc) and mean absolute error (MAE) according to the adaptive attack for proposed attack intensity estimator on Market1501 dataset attacked by Deep Mis-Ranking.

tively purifies color perturbations, maintaining consistency in the retrieval outcomes. Furthermore, compared to other models, IntensPure generally exhibits superior visual consistency among the top 10 retrieval results. This consistency enhances the reliability of the retrieval process.

G Discussion

G.1 Influence of Adaptive Attack on Attack Intensity Estimation

To evaluate the robustness of the attack intensity estimator against adaptive attacks, we generated images subjected to both re-ID attacks on the primary task model and classification (attribute recognition) attacks on the auxiliary task model. Subsequently, experiments are conducted on attack detection and intensity estimation using these manipulated images. We perturbed the query set in the Market1501 dataset using Deep Mis-Ranking, which not only enables re-ID attacks but also offers the additional option of attribute recognition attacks. The intensity of attribute attacks and re-ID attacks was set to be equal and the rest of the setting is the same as the experiment in the main paper.

As shown in Table 8, the incorporation of attribute recognition attacks marginally enhanced the detection performance of MEAAD [Wang *et al.*, 2021], the comparative model. This improvement is attributed to the fact that the perturbation caused by attribute recognition attacks does not serve as an adaptive attack against MEAAD but aids in its detection. Although the proposed attack intensity estimator exhibited a slight performance decline, it still demonstrated excellent capabilities in attack detection. The detection accuracy for adversarial examples subjected to both types of attacks reached 99.13%, reflecting a marginal decrease of 0.42%p, while the error in attack intensity estimation increased to 1.850, marking a rise of 1.081. While the estimation performance is slightly lower than that of existing general attacks for adaptive attacks, it is still meaningful because it maintains superior performance compared to the comparative model and exhibits a relatively lower error. This is because features are extracted based on the relationships between the retrieval results. However, there is still room for improvement in accurate attack intensity estimation for adversarial attacks that also affect auxiliary tasks.

G.2 Influence of Fitting Error on Optimal Diffusion Time Step Selection

In the main paper, we derive some graphs for the optimal time step t^* according to various attack intensities of each attack type and fit it to a single logarithmic curve. All of the graphs show similar and consistent logarithmic trends, so the fitted curve can approximately cover all variations. As the fitted curve is represented by a single line, discrepancies arise between t^* and the time step calculated using the fitted curve, particularly at specific outlying points. We define the error occurring between t^* and the calculated time step as fitting error, and Table 9 presents the rank-1 accuracy error based on the fitting error and their average values for Market1501 and DukeMTMC dataset. Nevertheless, the performance of the re-ID model is rarely impacted by the observed fitting error below 0.08

Dataset	Average fitting error (time step)	Average rank-1 accuracy error (%)
Market1501	3.6	0.05
DukeMTMC	2.9	0.08

Table 9: Rank-1 accuracy error according to the fitting error and their average values, evaluated on Market1501 and DukeMTMC dataset .

References

[Bai *et al.*, 2020] Song Bai, Yingwei Li, Yuyin Zhou, Qizhu Li, and Philip HS Torr. Adversarial metric attack and defense for person re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(6):2119–2126, 2020.

[Lee and Kim, 2023] Minjong Lee and Dongwoo Kim. Robust evaluation of diffusion-based adversarial purification. *ICCV*, 2023.

[Nie *et al.*, 2022] Weili Nie, Brandon Guo, Yujia Huang, Chaowei Xiao, Arash Vahdat, and Anima Anandkumar. Diffusion models for adversarial purification. *ICML*, 2022.

[Ristani *et al.*, 2016] 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*, 2016.

[Wang *et al.*, 2020] Hongjun Wang, Guangrun Wang, Ya Li, Dongyu Zhang, and Liang Lin. Transferable, controllable, and inconspicuous adversarial attacks on person re-identification with deep mis-ranking. In *CVPR*, 2020.

[Wang *et al.*, 2021] Xueping Wang, Shasha Li, Min Liu, Yaonan Wang, and Amit K Roy-Chowdhury. Multi-expert adversarial attack detection in person re-identification using context inconsistency. In *ICCV*, 2021.

[Yang *et al.*, 2022] Fengxiang Yang, Juanjuan Weng, Zhun Zhong, Hong Liu, Zheng Wang, Zhiming Luo, Donglin Cao, Shaozi Li, Shin'ichi Satoh, and Nicu Sebe. Towards robust person re-identification by defending against universal attackers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):5218–5235, 2022.

[Zheng *et al.*, 2015] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *ICCV*, 2015.

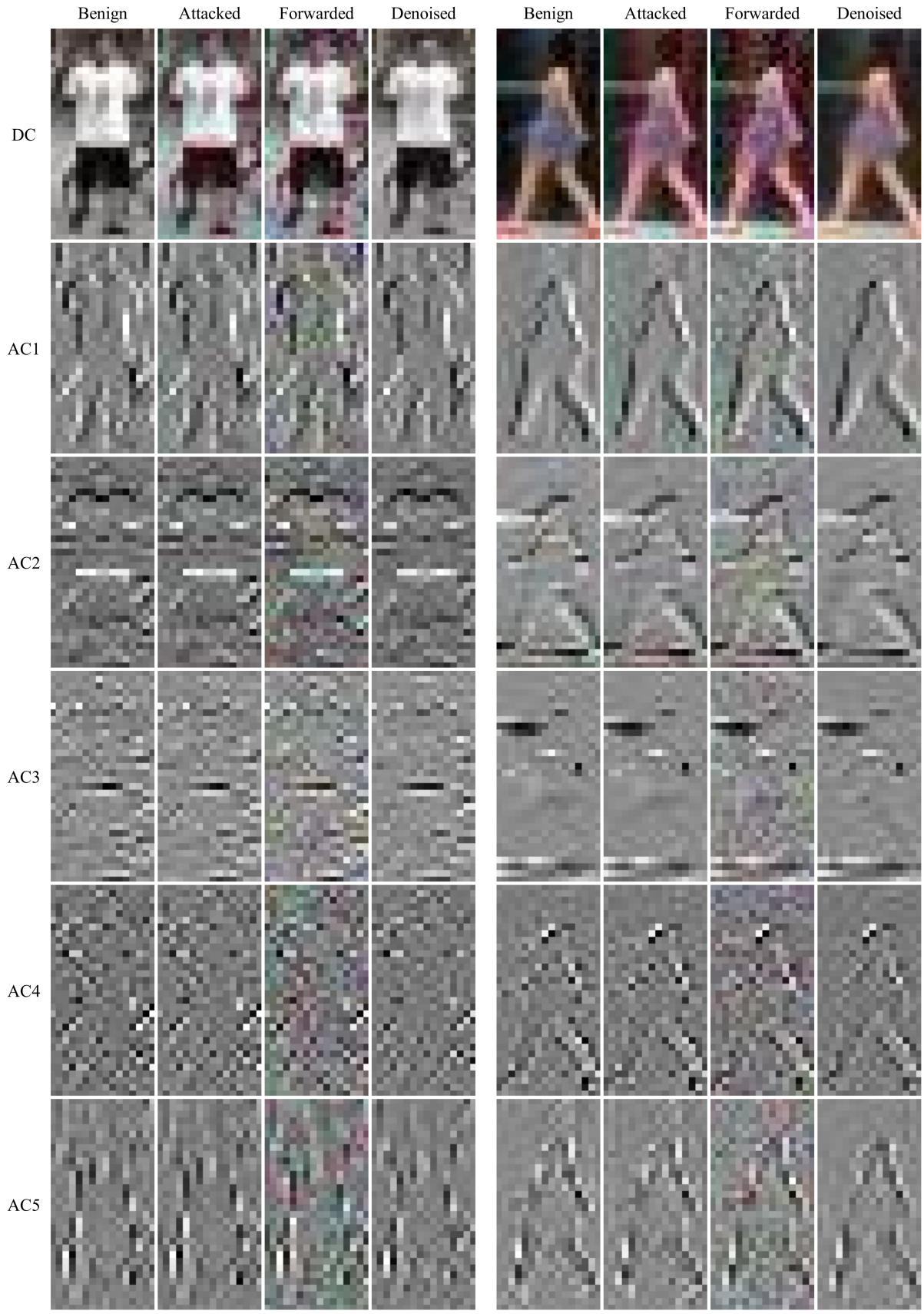

Figure 3: Illustration of the secondary diffusion process on **Market1501 dataset** attacked by Deep Mis-Ranking with $\epsilon = 16$.

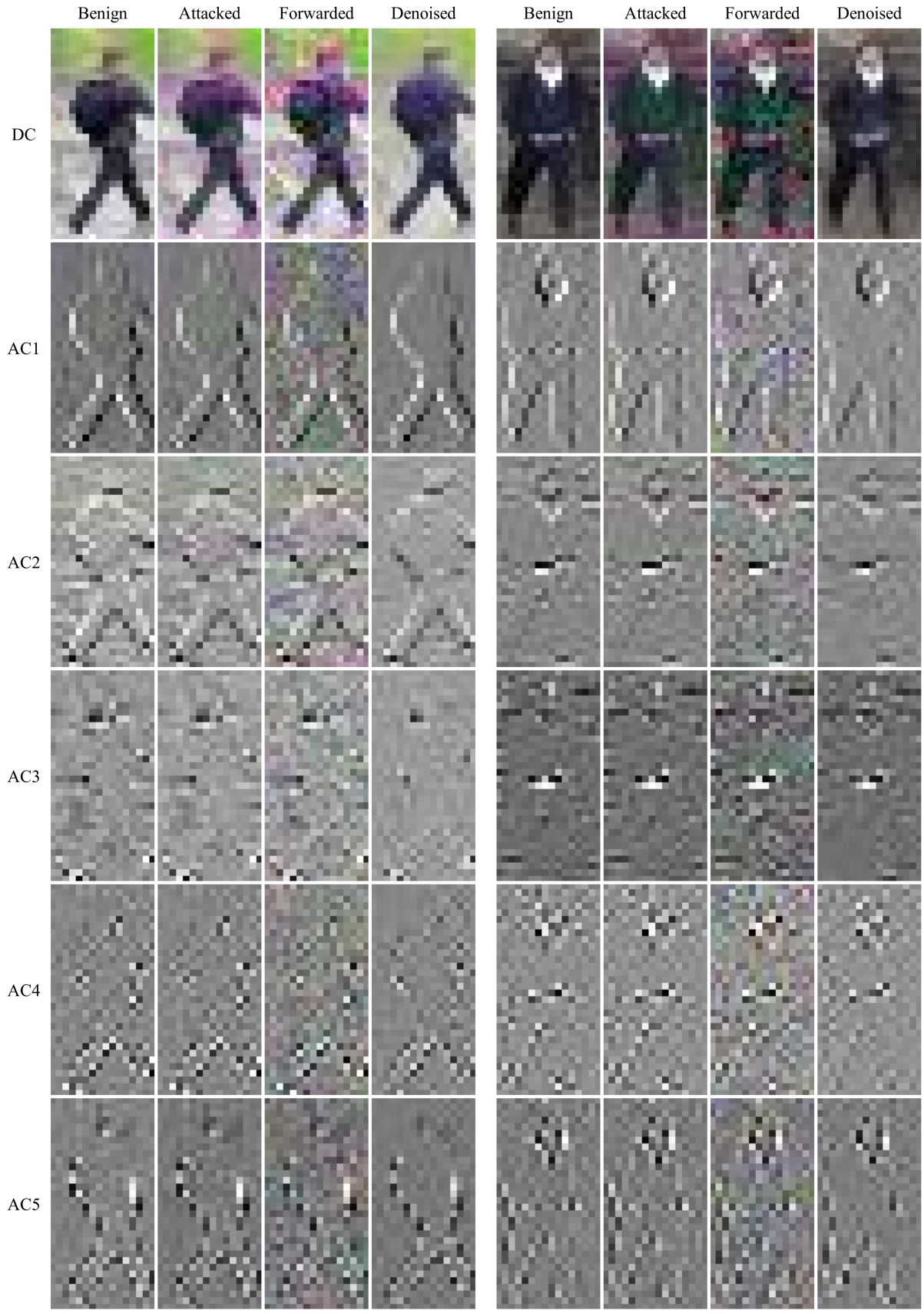

Figure 4: Illustration of the secondary diffusion process on **DukeMTMC dataset** attacked by Deep Mis-Ranking with $\epsilon = 16$.

Figure 5: Illustration of the top 10 retrieval results for images purified on Market1501 dataset attacked by Metric-FGSM with $\epsilon = 4$. Red boxes denote false results. Green boxes denote correct results.

Figure 6: Illustration of the top 10 retrieval results for images purified on Market1501 dataset attacked by Metric-FGSM with $\epsilon = 16$. Red boxes denote false results. Green boxes denote correct results.

Figure 7: Illustration of the top 10 retrieval results for images purified on Market1501 dataset attacked by Deep Mis-Ranking with $\epsilon = 4$. Red boxes denote false results. Green boxes denote correct results.

Figure 8: Illustration of the top 10 retrieval results for images purified on Market1501 dataset attacked by Deep Mis-Ranking with $\epsilon = 16$. Red boxes denote false results. Green boxes denote correct results.

Figure 9: Illustration of the top 10 retrieval results for images purified on Market1501 dataset attacked by MetaAttack with $\epsilon = 4$. Red boxes denote false results. Green boxes denote correct results.

Figure 10: Illustration of the top 10 retrieval results for images purified on Market1501 dataset attacked by MetaAttack with $\epsilon = 16$. Red boxes denote false results. Green boxes denote correct results.