# IntensPure: Attack Intensity-aware Secondary Domain Adaptive Diffusion for Adversarial Purification

**Eun-Gi Lee**, **Moon Seok Lee**, **Jae Hyun Yoon** and **Seok Bong Yoo***
Department of Artificial Intelligence Convergence, Chonnam National University, Gwangju, Korea
sbyoo@jnu.ac.kr

## Abstract

Adversarial attacks pose a severe threat to the accuracy of person re-identification (re-ID) systems, a critical security technology. Adversarial purification methods are promising approaches for defending against comprehensive attacks, including unseen ones. However, re-ID testing identities (IDs) are unseen, requiring more sophisticated purification than other classification tasks for adversarial defense. We propose IntensPure, an adversarial purification method in person re-ID that quantifies attack intensity via ID stability and attribute inconsistency to customize purification strength. Based on the estimated attack intensity, IntensPure employs secondary domain adaptive diffusion focused on purifying the low- and mid-frequency coefficients vulnerable to re-ID attacks. This method significantly reduces computational costs compared to the conventional diffusion method. For elaborate purification, IntensPure performs a directional diffusion process and refinements, leveraging the directional characteristics of secondary images. The experimental results on diverse attacks demonstrate that IntensPure outperforms the existing methods in terms of rank-1 accuracy.

## 1 Introduction

Person re-identification (re-ID) retrieves individuals across non-overlapping camera views in various locations, which is crucial for enhancing urban safety. Moreover, person re-ID aids in locating fugitive criminals or missing children. In recent years, person re-ID systems have also been threatened by the influence of adversarial attacks [Bouniot *et al.*, 2020; Bai *et al.*, 2020; Sun *et al.*, 2024; Wang *et al.*, 2020].

Adversarial training methods [Bai *et al.*, 2020; Bouniot *et al.*, 2020; Yang *et al.*, 2022] can be commonly employed to defend against adversarial attacks in person re-ID systems (re-ID attacks). However, adversarial training approaches are limited in their ability to address unseen attacks, even when incorporating a comprehensive threat model into the training pipeline. Figure 1(a and b) illustrate the rank-1 to rank-5 results of the three defense models against Metric-FGSM [Bai *et al.*, 2020] with attack intensity ($\epsilon$) values of 8 and 16, on the Market1501 dataset [Zheng *et al.*, 2015]. The top row of Figure 1(a and b) displays the person re-ID results for GOAT [Bouniot *et al.*, 2020], revealing the failure case of the adversarial training approach against unseen attacks.

Moreover, adversarial purification methods, another adversarial defense approach, have actively been studied because they can effectively defend against unseen threats using generative models. As generative models have advanced, diffusion model-based adversarial purification methods [Lee and Kim, 2023; Nie *et al.*, 2022; Yoon *et al.*, 2021] have also emerged. Diffusion models have demonstrated robust performance against adversarial attacks by purifying images without destroying their label semantics. However, existing diffusion-based purification methods treat benign (before attack) and perturbed images equally and purify all images. Such indiscriminate purification is likely to result in re-ID performance degradation. This is because the re-ID is more sensitive to image changes than other classification tasks due to the presence of unseen re-ID queries during testing [Zheng *et al.*, 2016]. Figure 1(c) presents an example where existing purification is applied to a benign case, leading to an incorrect re-ID result with reduced rank-1 accuracy by $14.11\%p$.

In addition, existing adversarial purification methods do not consider attack intensity appropriately. For instance, existing diffusion-based purification methods lack robustness to varying attack intensities because they use fixed-time steps. The problem of fixed-time steps is revealed in the second row of Figure 1(a and b), which show that DiffPure successfully purifies perturbed images with $\epsilon = 8$ but fails to purify images properly with $\epsilon = 16$. An optimal diffusion time step exists for purification, and Figure 1(d) indicates that this optimal time step varies depending on the attack intensity.

Another observation is that low- and mid-frequency coefficients are vulnerable to person re-ID attacks compared to high-frequency ones. Figure 1(e) demonstrates the rank-1 accuracy on the Market1501 dataset under person re-ID attack, as the attack scope progressively extends from high- to low-frequency coefficients. An $8\times8$ block discrete cosine transform (BDCT) isolates the frequency coefficients, and the experiment was conducted by replacing the original benign BDCT coefficients with their adversarial counterparts. While replacing 1 to 58 high-frequency BDCT coefficients with adversarial components has little influence on person re-ID performance, replacing the remaining six BDCT coefficients causes significant person re-ID performance degradation. This outcome suggests that the six lowest-frequency

*Corresponding author

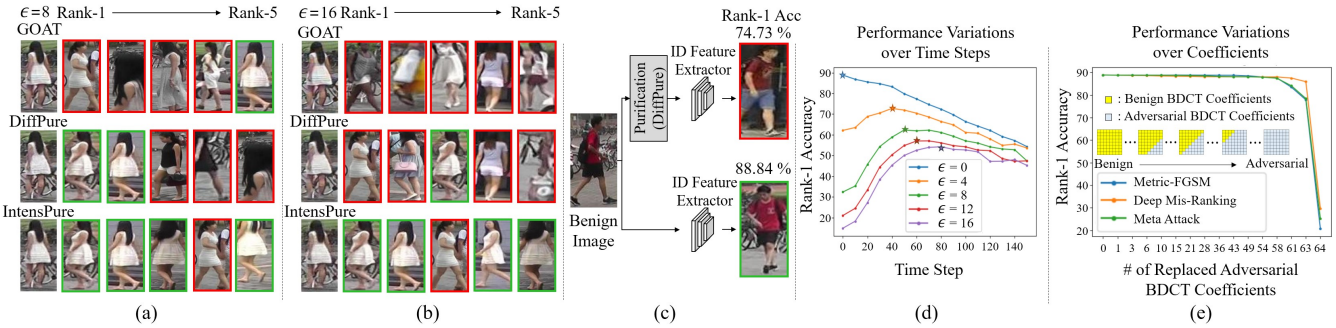


Figure 1: Person re-identification (re-ID) results from rank-1 to rank-5 under the unseen Metric-FGSM with an attack intensity of (a) $\epsilon = 8$ and (b) $\epsilon = 16$. First row: Adversarial training method results for GOAT. Second row: adversarial purification method results for DiffPure. Final row: IntensPure results. Red boxes denote false results. Green boxes denote correct results. (c) Influence on rank-1 accuracy by purifying the benign image. (d) Comparative analysis of the influence of diffusion time steps on person re-ID performance for each attack intensity $\epsilon$ of Metric-FGSM. Stars mark the highest performance points. (e) Performance variations when replacing the original block discrete cosine transform (BDCT) coefficients with the corresponding coefficients of adversarially attacked images from high to low frequencies.

BDCT coefficients are vulnerable to the person re-ID attack.

Building on these observations, we propose an attack intensity estimator that quantifies the adversarial attack intensity beyond mere detection. We further introduce an adversarial purification method that seamlessly integrates this estimator. To the best of our knowledge, IntensPure is the first adversarial purification method for re-ID systems. The proposed estimator assesses the attack intensity via identity (ID) stability and attribute inconsistency. This approach optimizes re-ID performance by selectively applying purification. Benign images remain intact, whereas perturbed images undergo tailored purification via adaptive time step adjustments in the diffusion process. In addition, we propose secondary domain adaptive diffusion, focusing on purifying the low- and mid-frequency coefficients that are strongly affected by person re-ID attacks. Thus, the spatial dimension is reduced by 64-fold, leading to significant computational cost savings. Further, leveraging the directional characteristics of BDCT-based secondary images, directional diffusion achieves effective purification. Our source code and appendix are available at https://github.com/st0421/IntensePure. We summarize the contributions below:

- We propose an attack intensity estimator by leveraging ID stability and attribute inconsistency by integrating an inner auxiliary network.
- We observe a meaningful correlation between the adversarial attack intensity and the optimal purification strength. We customize the diffusion time step to the estimated attack intensity to achieve optimal purification.
- We propose a secondary domain adaptive diffusion model. The diffusion process considers only a few secondary images; therefore, the computational complexity is reduced significantly compared to the existing methods while improving the re-ID accuracy.
- We propose two refinement strategies: an inter-block directional filter to enhance the inter-block correlation in the secondary domain and a perturbation constraint set to restrict excessive variations in pixel values.

## 2 Related Work

### 2.1 Adversarial Attack Detection

Adversarial attack detection [Feinman *et al.*, 2017; Hendrycks and Gimpel, 2017; Kim *et al.*, 2023; Liu *et al.*, 2019; Wang *et al.*, 2023; Yin *et al.*, 2021] is an approach to distinguishing adversarial examples from benign ones. [Li *et al.*, 2020] proposed using auto-encoders to identify inconsistencies among region proposals and detect adversarial attacks in object detection. Moreover, [Deng *et al.*, 2021] introduced LiBRe, which is a task- and attack-agnostic adversarial example detector. [Zhang *et al.*, 2023b] proposed EPS-AD, which performs adversarial detection based on a new statistic called the expected perturbation score. For re-ID, unlike other classification or detection tasks, the training and testing sets do not share the same categories. [Wang *et al.*, 2021a] proposed MEAAD for detecting inconsistencies in context, using multiple re-ID networks with different structures. However, despite their capacity to identify the presence or absence of adversarial attacks, the estimation of attack intensities is limited.

### 2.2 Adversarial Attacks in Person Re-ID

Some studies have addressed attack methods to analyze the vulnerability to adversarial attack [Bouniot *et al.*, 2020; Kang *et al.*, 2023; Liu *et al.*, 2023; Lu *et al.*, 2023; Wang *et al.*, 2021b; Wang *et al.*, 2021c; Zhang *et al.*, 2023a]. For the person re-ID attack, [Zheng *et al.*, 2018a] presented an opposite-direction feature attack to manipulate features in reverse directions using adversarial gradients. Moreover, [Bai *et al.*, 2020] introduced Metric-FGSM to maximize the metric distance between perturbed image features and reference features by extending classification attacks to re-ID attacks. [Wang *et al.*, 2020] proposed deep mis-ranking, which disrupts person re-ID system rankings with a multi-stage network, displaying substantial success in black-box attacks. Recently, [Yang *et al.*, 2022] introduced MetaAttack, combining combinatorial adversarial attacks and meta-learning to improve attack universality. Furthermore, [Sun *et al.*, 2024] proposed a backdoor attack for person re-ID models, using

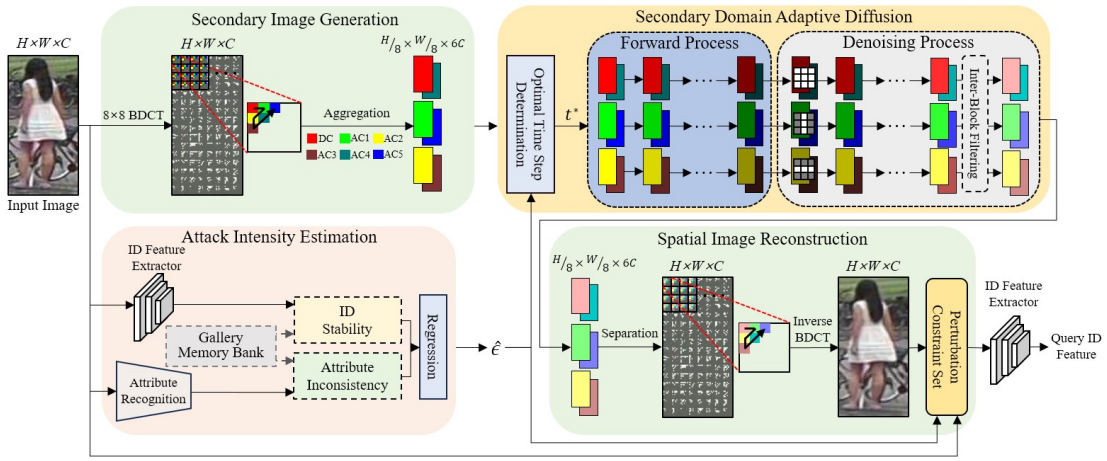


Figure 2: Overall architecture of the proposed framework.

dynamic triggers to change target identities, proving effective in open-set scenarios. Adversarial queries are sufficient to deceive the re-ID system and relatively cheap to generate [Zheng *et al.*, 2018a]. Although it is possible to attack the gallery set, it is generally much larger and more time-consuming to attack, making adversarial queries a more practical option [Wang *et al.*, 2021a]. Therefore, we focus on query image manipulation attacks in a white-box setting, where the attacker has full knowledge of the re-ID model.

### 2.3 Adversarial Defense

**Adversarial Training.** Adversarial attacks cause a significant security problem because they can induce unwanted behavior in deep learning models, drawing considerable attention to defense methods. Adversarial training [Gowal *et al.*, 2020; Kang *et al.*, 2021; Madry *et al.*, 2018] is a powerful countermeasure against such attacks. [Yu *et al.*, 2022] introduced the loss stationary condition for controlled weight perturbation, enhancing its robustness by avoiding excessive perturbations. Additionally, [Gong *et al.*, 2022] introduced a color-based local transformation attack and a joint adversarial defense, improving the re-ID model robustness across domains. However, as [Bai *et al.*, 2021] indicated, adversarial training has a trade-off between robustness and generalization and still struggles to counter unseen attacks effectively.

**Adversarial Purification.** As a counterpart to adversarial training, adversarial purification methods [Samangouei *et al.*, 2018; Song *et al.*, 2018; Yang *et al.*, 2024] have demonstrated robustness against unseen attacks. [Du and Mordatch, 2019] and [Grathwohl *et al.*, 2020] demonstrated the effectiveness of energy-based models trained with Markov chain Monte Carlo methods in purifying adversarial examples. Moreover, [Hill *et al.*, 2021] presented the robustness of long-run Markov chain Monte Carlo methods with energy-based models for the purification of perturbed images, using Langevin dynamics to enhance the robustness and efficacy. Similarly, [Yoon *et al.*, 2021] used the denoising score-based model for purification. In recent work [Ankile *et al.*, 2023; Shi *et al.*, 2023; Sun *et al.*, 2023; Wu *et al.*, 2022; Xiao *et al.*, 2023], diffusion models have been considered for purification. Specifically, [Nie *et al.*, 2022] optimized the computations for solving reverse-time stochastic differential equations using the adjoint sensitivity method. Moreover, [Carlini *et al.*, 2023] introduced a diffusion-denoised smoothing method to achieve adversarial robustness and efficiency. Moreover, [Lee and Kim, 2023] proposed a robustness measurement guideline and introduced an improved purification strategy with a gradual noise-scheduling technique. Despite advancements in adversarial purification methods, they did not adaptively purify perturbed images based on the intensity of attacks, often leading to unnecessary or insufficient purification efforts.

## 3 Proposed Method

### 3.1 Overview

As depicted in Figure 2, we propose IntensPure, an adversarial purification method that adaptively purifies perturbed images using attack intensity estimation and secondary domain adaptive diffusion. IntensPure estimates the attack intensity using ID stability and attribute inconsistency. This method determines the optimal diffusion time step and generates secondary images by aggregating the six lowest-frequency BDCT coefficients. This secondary domain purification focuses on low- and mid-frequency coefficients, which are particularly vulnerable to person re-ID attacks. Further, employing directional diffusion leverages the BDCT coefficient directional characteristics. Inter-block directional filtering is employed to improve the diffusion process further, enhancing inter-block correlations within the secondary image. Once the diffusion process is complete, the secondary images revert to its original spatial domain. Overpurification is restrained through the perturbation constraint set to ensure the robustness of the ID feature extractor against adversarial attacks.

### 3.2 Attack Intensity Estimation

Inspired by the basis that inner models that do not output to users are undetectable to attackers [Wang *et al.*, 2021a], we leveraged an auxiliary task model that acts as an undetectable inner model that is robust to primary task attacks. Figure 3(a) presents the accuracy of ResNet50 for attribute recognition (auxiliary task: $\star$) and person re-ID (primary task: $\bullet$) under re-ID attacks (Metric-FGSM, Deep-Mis-Ranking, and

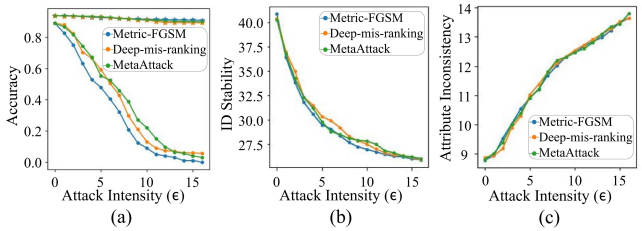


Figure 3: Influence of person re-ID attack intensity on attribute recognition (auxiliary task) and person re-ID (primary task) and related measures of ID stability and attribute inconsistency, as observed on the Market1501 dataset. (a) Attribute recognition accuracy and re-ID rank-1 accuracy vs. attack intensity. (b) ID stability results according to attack intensity. (c) Attribute inconsistency results according to attack intensity.

MetaAttack) with varying intensity. Both primary and auxiliary models are on the same network, but the rank-1 accuracy decreased significantly due to the re-ID attack, whereas the attribute recognition results were virtually unaffected. This empirical study supports the validity of auxiliary task models that are insensitive to primary task attacks. Accordingly, we estimate the person re-ID attack intensity by leveraging attacked retrieval results and the results from attribute recognition as an auxiliary task.

Additionally, the re-ID attack causes a retrieval system to return images irrelevant to the query in the top ranking list [Li *et al.*, 2019]. As a secondary effect, the consistency between retrieval results is compromised. In addition, re-ID attacks can also affect the consistency of the bottom retrieval results because the retrieval process relies on similarity metrics. Therefore, we extract features from not only the top 10 retrieval results but also the bottom 10 retrieval results to estimate the attack intensity. In Appendix $A$, we provide the inconsistent retrieval results due to person re-ID attacks and the analysis of the effectiveness of the bottom list.

**ID Stability.** We define ID stability with a dimension of $1\times100$ as a concatenated set of $SIM_{top}$, $SIM_{bottom}$, and $SIM_{Q-R1}$, containing cosine similarities between embedded vectors extracted by ID feature extractors. The $SIM_{top}$ represents the cosine similarities between the top 10 retrieval results themselves and is obtained as follows:

$$SIM_{top} = \{F_i \cdot F_j \mid F_i, F_j \in \{F_1, F_2, \ldots, F_{10}\}_{\text{Top10}}, i \neq j\}, \quad (1)$$

where $F_i$ and $F_j$ are embedding vectors of the rank-$i$ and rank-$j$ images selected from the top 10 retrieval results in the gallery. The middle dot symbol represents the inner product, employed for cosine similarity calculation. In addition, $\{\cdot\}_{\text{Top10}}$ denotes a set of 10 retrieval result features that have the highest cosine similarity to the query feature. Similarly, the $SIM_{bottom}$ can be obtained from the same operation with Equation 1, replacing the top 10 with the bottom 10. Then, the dimensions of both $SIM_{top}$ and $SIM_{bottom}$ are $1\times45$. The $SIM_{Q-R}$ with a dimension of $1\times10$ is obtained as follows:

$$SIM_{Q-R} = \{F_Q \cdot F_R \mid F_R \in \{F_1, F_2, \ldots, F_{10}\}_{\text{Top10}}\}, \quad (2)$$

where $F_Q$ and $F_R$ represent the embedding vectors of the query image and the top 10 retrieval results.

The ID stability sum in Figure 3(b) is the sum of 100 elements ($SIM_{top}$, $SIM_{bottom}$, and $SIM_{Q-R1}$) that make up ID stability for observing the correlation between ID stability and the attack intensity. Figure 3(b) demonstrates the monotonic decline of ID stability sum in response to increasing attack intensity, suggesting the potential for estimating the attack intensity based on ID stability.

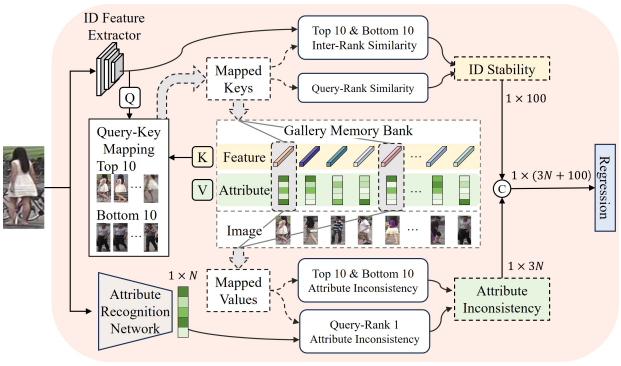


Figure 4: Illustration of the attack intensity estimator.

**Attribute Inconsistency.** We define attribute inconsistency with a dimension of $1\times3N$ as a merged set of $INCON_{top}$, $INCON_{bottom}$, and $INCON_{Q-R1}$, containing absolute errors between attribute recognition results. The $N$ is the number of the person attribute classes provided by [Lin *et al.*, 2019]. Concretely, $INCON_{top}$ with a dimension of $1\times N$ denotes absolute errors between the attribute recognition results of the query image and the mean of the attribute recognition results of the top 10 retrieval results and is obtained as follows:

$$INCON_{top} = \left| Attr(I_Q) - \frac{1}{10}\sum_{i=1}^{10} Attr(I_{t_i}) \right|, \quad (3)$$

where $Attr(I)$ represents the results of the attribute recognition for image $I$. The $I_Q$ denotes the query image and $I_t$ denotes the top 10 retrieval results. Similarly, the $INCON_{bottom}$ can be obtained from the same operation with Equation 3 using the bottom 10 retrieval results instead of $I_t$. Then, the dimensions of both $INCON_{top}$ and $INCON_{bottom}$ are $1\times N$. The analysis of the number of retrieval results is provided in Appendix $B$. Further, $INCON_{Q-R1}$ with a dimension of $1\times N$ consists of absolute errors between the attribute recognition results of $I_Q$ and them of rank-1 image. The attribute inconsistency sum in Figure 3(c) is the sum of $3N$ elements that make up attribute inconsistency to examine the relationship between attribute inconsistency and attack intensity. Figure 3(c) depicts the steady increase in attribute inconsistency sum as the attack intensity escalates. The monotonic nature of this increase implies the potential of attribute inconsistency as a tool for attack intensity estimation.

**Implementation of Estimator.** Figure 4 illustrates the estimator framework, exploiting a pre-built gallery memory bank. The gallery memory bank is a key-value store where keys ($K$) are image ID features and values ($V$) are attribute recognition results. The estimator extracts the ID feature of an input image as a query ($Q$) and maps it to the 10 most and least similar keys in the memory bank. Using this mapping, the estimator obtains ID stability and attribute inconsistency, and concatenates them. Finally, a multi-layer perceptron regression model uses these information to estimate the attack

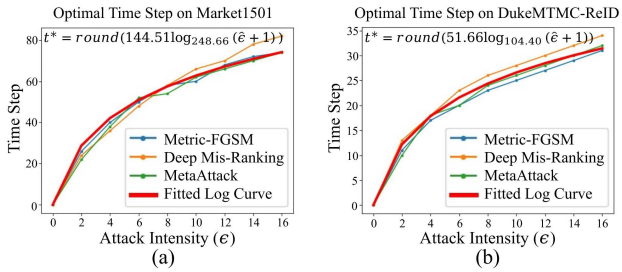


Figure 5: Optimal time steps according to attack intensity on the (a) Market1501 and (b) DukeMTMC-ReID datasets.

intensity. The regression model is implemented with two hidden layers with 512 and 256 nodes, respectively. Rectified linear unit functions are applied after each hidden layer.

### 3.3 Secondary Image Generation

As depicted in the upper left part of Figure 2, IntensPure conducts an 8×8 BDCT on the input image with a dimension of $H \times W \times C$. For each patch, the six lowest frequency coefficients, including DC, are extracted, and the coefficients are aggregated into six secondary images with a dimension of $\frac{H}{8} \times \frac{W}{8} \times 6C$. After the secondary image generation process, the input dimensions are reduced to 6/64-fold ($\simeq$0.093) compared to the initial input dimensions. Through this, the computational complexity of IntensPure is significantly reduced compared to the conventional diffusion method.

### 3.4 Secondary Domain Adaptive Diffusion

**Optimal Time Step Determination.** For the diffusion process, Figure 5 indicates the variation of empirically selected time steps optimal for purifying according to the attack intensities. For each attack method employed in the experiments, the graphs illustrating the optimal time step for the most effective diffusion-based purification have consistent logarithmic trends across attack intensities. Therefore, we fit a logarithmic curve that can cover the variations in Figure 5 using SciPy [Virtanen *et al.*, 2020]. The prominent red curve in the figure indicates the fitted curve. Then, IntensPure determines the approximately optimal time step $t^*$ using the formula derived from the fitted curve and the estimated attack intensity.

**Secondary Domain Diffusion.** Figure 6 illustrates the secondary domain adaptive diffusion process with a determined time step $t^*$. IntensPure operates in the frequency domain on secondary images with dimensions of $\frac{H}{8} \times \frac{W}{8} \times 6C$. IntensPure addresses the six lowest frequency coefficients from DC to AC5, which are vulnerable to re-ID attacks, excluding the other high-frequency coefficients. Experiments exploring the influence of BDCT coefficient counts are presented in Appendix $C$. Additionally, IntensPure uses secondary image directional characteristics in the diffusion process to address coefficients adaptively according to the isotropic, vertical, and horizontal diffusion processes. Specifically, IntensPure performs forward and reverse processes in a direction parallel to each coefficient during $t^*$. For DC and AC4 coefficients, IntensPure employs an isotropic diffusion process with 3×3 convolutional layers, considering all directions within the kernel. For AC1 and AC5 coefficients with vertical directions, IntensPure employs a vertical diffusion process with

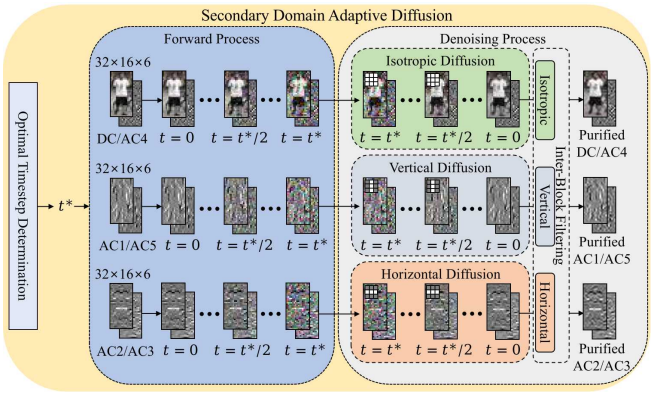


Figure 6: Illustration of secondary domain adaptive diffusion.

3×1 convolutional layers. For AC2 and AC3 coefficients with horizontal directions, 1×3 convolutional layers are employed. These directional layers enable the effective detection of perturbations along vertical or horizontal directions, facilitating the efficient purification of perturbations.

**Inter-block Directional Filter.** As depicted on the right in Figure 6, the inter-block directional filters are employed after the diffusion process to achieve two goals: removing remaining perturbations and enhancing inter-block correlation. This approach employs distinct directional filters for DC and AC coefficients, optimized to address their specific directional characteristics. These filters are defined as follows:

$$D'_p = \frac{1}{W} \sum_{p_i \in \Omega} D_{p_i} G_{\sigma_r}(||D_{p_i} - D_p||) G_{\sigma_s}(||p_i - p||), \quad (4)$$

where $D'$ denotes filtered coefficients, while $D$ denotes input coefficients. Further, $p$ represents the pixel index, and $\Omega$ is the set of pixels $p_i$ in a window. The window size is set to 5×5 for isotropic filtering, 5×1 for vertical filtering, and 1×5 for horizontal filtering, corresponding to the respective directions. Moreover, $W$ is the normalization term, and $G_{\sigma_r}$ and $G_{\sigma_s}$ denote the range and spatial Gaussian functions, respectively, with $\sigma_r$ and $\sigma_s$ as their standard deviations (set to 0.15 and 3.0). The investigation of inter-block filter size and $\sigma$ values is presented in Appendix $D$.

### 3.5 Spatial Image Reconstruction

In the bottom-right of Figure 2, the spatial image reconstruction involves the transformation process from the secondary domain to the original spatial domain, following the diffusion. Following an arrangement of purified secondary images into their designated coefficient positions on zero-padded 8×8 blocks, an inverse BDCT process is applied. Through this process, the image is reconstructed to its original dimensions of $H \times W \times C$. Inadvertently, the reconstructed image can lose essential details or experience excessive variation in pixel level due to the diffusion process. We introduce a perturbation constraint set restricting pixel variations more than the attack intensity to address this problem. This perturbation constraint set is defined as follows:

$$PCS(I_p, O_p, \hat{\epsilon}) = \begin{cases} O_p & |O_p - I_p| \leq \hat{\epsilon} \\ I_p + \hat{\epsilon} & O_p - I_p > \hat{\epsilon} \\ I_p - \hat{\epsilon} & O_p - I_p < -\hat{\epsilon} \end{cases}, \quad (5)$$

where $O$ denotes the image purified from diffusion, $I$ denotes the input image, and $\hat{\epsilon}$ represents the estimated attack intensity. The image is effectively purified while maintaining image integrity without excessive alterations via the perturbation constraint set. Moreover, when $\hat{\epsilon}$ equals 0, it maintains the initial state of the input image.

# 4 Experiments

## 4.1 Experimental Setup

**Datasets and Person Re-ID Network.** For the experiments, we used two real-world datasets: Market1501 [Zheng *et al.*, 2015] and DukeMTMC-reID [Ristani *et al.*, 2016]. Market1501 includes 32,668 images of 1,501 identities. In addition, 750 IDs are labeled for the training set, and 751 IDs are labeled for the testing set. The DukeMTMC-reID set consists of 36,411 images of 1,404 identities, with 702 IDs used for both the training and testing sets. For person re-ID performance evaluation, we leveraged the ResNet50 [He *et al.*, 2016] architecture and pre-trained weights provided by [Zheng *et al.*, 2018b] for fair comparisons.

**Adversarial Attacks.** To evaluate the performance of IntensPure against a variety of person re-ID attack methods, we applied Metric-FGSM [Bai *et al.*, 2020], Deep Mis-Ranking [Wang *et al.*, 2020], and MetaAttack [Yang *et al.*, 2022] on the Market1501 and DukeMTMC-reID datasets. To demonstrate the superiority of the proposed method, we considered various attack intensities, ranging from $\epsilon = 0$ to 16.

**Evaluation Metrics.** We evaluated the attack intensity estimator using the area under the receiver operating characteristic (AUROC) and classification accuracy for adversarial detection comparison and the mean absolute error (MAE) for attack intensity estimation. In addition, we considered rank-1 accuracy to assess the overall performance of IntensPure.

## 4.2 Implementation Details

**Attack Intensity Estimator.** The attack intensity estimator uses an ID feature extractor and an attribute recognition network as the inner networks. The ID feature extractor shares the weight with the re-ID model, and the attribute recognition model uses pre-trained weights, as provided by [Lin *et al.*, 2019]. For training the estimator, the training set is selected in the same way as by [Wang *et al.*, 2021a], perturbing query samples in the training set with only Metric-FGSM and Deep Mis-Ranking with attack intensities from $\epsilon = 0$ to 16. We extracted the ID stability and attribute inconsistency from the samples, and the attack intensity is designated as a label.

**Secondary Domain Adaptive Diffusion.** We adopted the U-Net architecture used in Palette [Saharia *et al.*, 2022]. For the experiments, we set the total time steps to 1,000. The initial and final values of the linear noise schedule are set to 1e-6 and 9e-2, respectively. The training process encompassed 700 epochs, with a learning rate of 1e-4. The IntensPure network was trained on a single A100 GPU.

## 4.3 Comparison of Person Re-ID Attack Detection

First, we compared the estimator to previous studies in terms of adversarial detection because no previous studies exist on estimating the attack intensity. We transformed the estimator into a detector by distinguishing between 0 and non-zero outcomes of the rounded estimated result.

| Method | Metric-FGSM ($\epsilon$ = 4) | | Deep Mis-Ranking ($\epsilon$ = 16) | | MetaAttack ($\epsilon$ = 8) | |
|---|---|---|---|---|---|---|
| | Acc | AUROC | Acc | AUROC | Acc | AUROC |
| LiBRe | - | 0.933 | - | 0.962 | - | 0.913 |
| EPS-AD | - | 0.955 | - | 0.972 | - | 0.941 |
| MEAAD | 97.30 | 0.980 | 98.50 | 1.000 | 94.02 | 0.944 |
| Ours | 98.85 | 1.000 | 99.55 | 1.000 | 95.88 | 0.985 |

Table 1: Comparison of detection accuracy (Acc) and AUROC of adversarial attack detection on the Market1501 dataset.

| Method | Metric-FGSM ($\epsilon$ = 4) | Deep Mis-Ranking ($\epsilon$ = 16) | MetaAttack ($\epsilon$ = 8) |
|---|---|---|---|
| MEAAD | 3.340 | 3.189 | 3.912 |
| Ours | 0.806 | 0.769 | 1.071 |

Table 2: Comparison on the mean absolute error of the attack intensity estimator for person re-ID attacks on the Market1501 dataset.

We compared the adversarial detection performance with three state-of-the-art detection methods: LiBRe [Deng *et al.*, 2021], MEAAD [Wang *et al.*, 2021a], and EPS-AD [Zhang *et al.*, 2023b]. These methods do not require additional modifications to be used with person re-ID attack detection. In addition, LiBRe and EPS-AD are only evaluated using AUROC, making it challenging to determine their exact accuracy and adjust the network for attack intensity estimation.

Table 1 presents a comparative analysis of adversarial detection methods. It shows accuracy and AUROC under the proposed default attack intensity $\epsilon$ values for Metric-FGSM, Deep Mis-Ranking, and MetaAttack. Across the three evaluated person re-ID attacks, the proposed detector consistently demonstrated superior performance in terms of accuracy and AUROC values. Notably, the MetaAttack is an unseen attack method that was not included in the training data. This result highlights the effectiveness of the detector for unseen threats.

To evaluate the performance of attack intensity estimation, we converted MEAAD to an estimator and compared it to the proposed estimator after learning under the same conditions. Table 2 indicates that the proposed estimator outperforms MEAAD in terms of MAE for all three attacks. The results demonstrate that ID stability and attribute inconsistency are more effective than the context feature proposed by MEAAD for attack detection and intensity estimation.

## 4.4 Comparison of Adversarial Purification

Table 3 compares the rank-1 accuracy of existing purification methods on the Market1501 dataset. The results were evaluated against adversarial attacks: Metric-FGSM, Deep Mis-Ranking, and MetaAttack, each with $\epsilon$ values (0, 4, 8, 12, and 16) to represent the attack intensity. For comparison with existing purification methods, the default settings were applied based on each paper. The table reveals that our IntensPure consistently outperforms other purification methods in all cases and is the most effective against adversarial attacks while maintaining its performance on benign examples ($\epsilon = 0$). As the attack intensity increases, the re-ID accuracy of the ResNet50 re-ID model decreases, and this trend continues even after purification. This is because the purification strength is insufficient to remove stronger perturbations. IntensPure adaptively increases the purification strength, but it also shows a similar trend of decline due to the trade-off of losing benign information. However, our IntensPure is superior because, the rate of decline is much more gradual compared to other purification models when attack intensity increases from $\epsilon = 0$ to 16.

| Attack Method | - | Metric-FGSM | | | | Deep Mis-Ranking | | | | MetaAttack | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Attack Intensity | $\epsilon = 0$ | $\epsilon = 4$ | $\epsilon = 8$ | $\epsilon = 12$ | $\epsilon = 16$ | $\epsilon = 4$ | $\epsilon = 8$ | $\epsilon = 12$ | $\epsilon = 16$ | $\epsilon = 4$ | $\epsilon = 8$ | $\epsilon = 12$ | $\epsilon = 16$ |
| ResNet50 (Baseline) | 88.84 | 52.95 | 20.86 | 4.59 | 0.00 | 66.48 | 29.63 | 11.67 | 5.70 | 67.13 | 38.69 | 8.14 | 3.00 |
| ADP [Yoon *et al.*, 2021] | 69.61 | 63.42 | 49.51 | 43.72 | 35.42 | 67.26 | 42.51 | 19.24 | 9.21 | 68.57 | 50.54 | 38.72 | 27.17 |
| GDMP [Wang *et al.*, 2022] | 70.75 | 67.97 | 61.79 | 55.36 | 52.54 | 70.84 | 48.62 | 20.17 | 5.86 | 68.75 | 54.31 | 31.80 | 12.30 |
| DiffPure [Nie *et al.*, 2022] | 74.73 | 65.08 | 51.54 | 47.12 | 38.93 | 70.54 | 52.02 | 21.29 | 13.00 | 71.59 | 66.33 | 46.02 | 39.73 |
| GNSP [Lee and Kim, 2023] | 73.25 | 68.17 | 60.77 | 56.04 | 54.44 | 68.43 | 53.15 | 23.16 | 19.71 | 70.98 | 59.59 | 55.61 | 48.75 |
| Ours | 88.36 | 72.74 | 66.65 | 62.05 | 60.42 | 76.51 | 74.52 | 56.41 | 49.52 | 78.50 | 74.52 | 70.99 | 65.88 |

Table 3: Rank-1 accuracy on the **Market1501** dataset against person re-ID attacks, evaluated at varying attack intensities ($\epsilon = 0, 4, 8, 12, 16$).

| Attack Method | - | Metric-FGSM | | | | Deep Mis-Ranking | | | | MetaAttack | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Attack Intensity | $\epsilon = 0$ | $\epsilon = 4$ | $\epsilon = 8$ | $\epsilon = 12$ | $\epsilon = 16$ | $\epsilon = 4$ | $\epsilon = 8$ | $\epsilon = 12$ | $\epsilon = 16$ | $\epsilon = 4$ | $\epsilon = 8$ | $\epsilon = 12$ | $\epsilon = 16$ |
| ResNet50 (Baseline) | 79.35 | 54.09 | 15.72 | 2.01 | 0.00 | 49.69 | 18.27 | 5.79 | 2.06 | 55.07 | 19.21 | 1.17 | 0.40 |
| ADP [Yoon *et al.*, 2021] | 66.71 | 57.43 | 46.26 | 39.31 | 33.74 | 59.58 | 45.67 | 37.93 | 20.84 | 61.26 | 59.58 | 54.34 | 50.61 |
| GDMP [Wang *et al.*, 2022] | 68.36 | 63.33 | 50.25 | 43.08 | 38.29 | 62.34 | 49.03 | 20.15 | 14.06 | 63.20 | 60.34 | 55.69 | 51.62 |
| DiffPure [Nie *et al.*, 2022] | 70.69 | 62.79 | 52.52 | 46.99 | 42.29 | 69.39 | 49.33 | 42.55 | 21.86 | 63.51 | 63.33 | 57.05 | 53.59 |
| GNSP [Lee and Kim, 2023] | 69.40 | 63.89 | 50.39 | 45.60 | 41.79 | 64.95 | 51.66 | 43.20 | 35.84 | 68.31 | 64.86 | 59.91 | 56.94 |
| Ours | 78.95 | 64.99 | 57.59 | 54.62 | 54.13 | 71.32 | 58.35 | 44.83 | 44.61 | 70.60 | 66.42 | 60.57 | 59.69 |

Table 4: Rank-1 accuracy on the **DukeMTMC** dataset against person re-ID attacks, evaluated at varying attack intensities ($\epsilon = 0, 4, 8, 12, 16$).

| Purification Method | FLOPs (G) ↓ | Params (M) ↓ | Time (ms) ↓ |
|---|---|---|---|
| ADP | 420 | 94 | 1210 |
| GDMP | 835 | 553 | 993 |
| DiffPure | 583 | 190 | 366 |
| GNSP | 530 | 95 | 249 |
| Ours | 50 | 833 | 64 |
| (Estimator + Purifier) | (11+39) | (82+751) | (5+59) |

Table 5: Complexity of diffusion-based purification methods on the Market1501 dataset.

| Attack Intensity Estimation | Secondary Image Generation | Directional Diffusion | Inter-Block Directional Filter | Perturbation Constraint Set | Rank-1 Accuracy |
|---|---|---|---|---|---|
| | | | | | 54.52 |
| ✓ | | | | | 63.14 |
| ✓ | ✓ | | | | 66.55 |
| ✓ | ✓ | ✓ | | | 67.97 |
| ✓ | ✓ | ✓ | ✓ | | 68.31 |
| ✓ | ✓ | ✓ | ✓ | ✓ | 70.04 |

Table 6: Rank-1 accuracy of IntensPure on the Market1501 dataset according to each proposed module.

Table 4 presents a comparison of rank-1 accuracy on the DukeMTMC-ReID dataset in the same setting as Table 3 to evaluate the robustness and effectiveness of the defense methods. IntensPure demonstrates outstanding performance against various adversarial attacks, with a consistent rank-1 accuracy across $\epsilon$ values. These results exhibit a reliable ability to respond to diverse adversarial environments and variations in attack intensity effectively.

Table 5 compares the complexity of diffusion-based purification models and IntensPure in terms of the FLOPs, parameters, and inference time on the Market1501 dataset. To ensure a fair comparison, we evaluated comparison models using their respective default settings with a single diffusion time step. While IntensPure exhibits a higher parameter count than its counterparts due to the separated directional diffusion networks, it achieves a marked reduction in inference time and FLOPs by leveraging only a few BDCT coefficients instead of the complete set. Furthermore, the proposed estimator operates only once regardless of the diffusion time step and is remarkable in all types of complexity, making it more practical for real-world applications.

### 4.5 Ablation Study

Table 6 presents the analysis of the individual contributions of each IntensPure component to the overall rank-1 accuracy on the Market1501 dataset, attacked by Metric-FGSM with various $\epsilon$ values (0, 4, 8, 12, and 16). Each row represents a distinct framework configuration, incrementally activating modules following the IntensPure procedure to isolate their respective effects on rank-1 accuracy.

The first row establishes a baseline with no activated modules and a time step $t$=150, which is the default setting of DiffPure. The activation of the attack intensity estimation module yields notable performance improvements, showing its crucial role in adaptive purification tailored to attack intensity. Subsequently, the secondary image generation module leads to a performance increase, presenting its individual effectiveness. Furthermore, the directional diffusion module assists in a slight increase in re-ID accuracy. Finally, the inter-block directional filter and the perturbation constraint set provide incremental yet valuable improvements. These results demonstrate that each of the suggested modules plays a role in enhancing the overall performance. Appendix E presents an ablation study that removes one module at a time to demonstrate the individual effect of each module.

## 5 Conclusion

This paper addresses the limitations of existing adversarial purification methods, such as their high computational cost and inflexible purification strength, by proposing IntensPure. The proposed method introduces the attack intensity-aware adaptive diffusion model, incorporating a network that estimates attack intensity using ID stability and attribute inconsistency. IntensPure efficiently operates by adjusting the purification strength and using the secondary domain to focus on a few low- and mid-frequencies vulnerable to adversarial attacks. Additionally, we introduce two refinement strategies, including inter-block directional filtering to enhance the inter-block correlation and a perturbation constraint set to preserve the benign information. Consequently, IntensPure performs computationally efficiently and achieves state-of-the-art re-ID performance against adversarial attacks.

## Acknowledgments

This work was supported by the Industrial Fundamental Technology Development Program (No. 20018699) funded by MOTIE of Korea and the IITP grant funded by the Korea government (MSIT) (No.2021-0-02068, RS-2023-00256629, RS-2022-00156287).

## Contribution Statement

Eun-Gi Lee, Moon Seok Lee, and Jae Hyun Yoon contributed equally to this work.

## References

[Ankile *et al.*, 2023] Lars Lien Ankile, Anna Midgley, and Sebastian Weisshaar. Denoising diffusion probabilistic models as a defense against adversarial attacks. *arXiv :2301.06871*, 2023.

[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.

[Bai *et al.*, 2021] Tao Bai, Jinqi Luo, Jun Zhao, Bihan Wen, and Qian Wang. Recent advances in adversarial training for adversarial robustness. In *IJCAI*, 2021.

[Bouniot *et al.*, 2020] Quentin Bouniot, Romaric Audigier, and Angelique Loesch. Vulnerability of person re-identification models to metric adversarial attacks. In *CVPRW*, 2020.

[Carlini *et al.*, 2023] Nicholas Carlini, Florian Tramer, Krishnamurthy Dj Dvijotham, Leslie Rice, Mingjie Sun, and J Zico Kolter. (certified!!) adversarial robustness for free! In *ICLR*. 2023.

[Deng *et al.*, 2021] Zhijie Deng, Xiao Yang, Shizhen Xu, Hang Su, and Jun Zhu. Libre: A practical bayesian approach to adversarial detection. In *CVPR*, 2021.

[Du and Mordatch, 2019] Yilun Du and Igor Mordatch. Implicit generation and modeling with energy based models. *NeurIPS*, 32, 2019.

[Feinman *et al.*, 2017] Reuben Feinman, Ryan R Curtin, Saurabh Shintre, and Andrew B Gardner. Detecting adversarial samples from artifacts. *arXiv :1703.00410*, 2017.

[Gong *et al.*, 2022] Yunpeng Gong, Liqing Huang, and Lifei Chen. Person re-identification method based on color attack and joint defence. In *CVPR*, 2022.

[Gowal *et al.*, 2020] Sven Gowal, Chongli Qin, Jonathan Uesato, Timothy Mann, and Pushmeet Kohli. Uncovering the limits of adversarial training against norm-bounded adversarial examples. *arXiv :2010.03593*, 2020.

[Grathwohl *et al.*, 2020] Will Grathwohl, Kuan-Chieh Wang, Jörn-Henrik Jacobsen, David Duvenaud, Mohammad Norouzi, and Kevin Swersky. Your classifier is secretly an energy based model and you should treat it like one. *ICLR*, 2020.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[Hendrycks and Gimpel, 2017] Dan Hendrycks and Kevin Gimpel. Early methods for detecting adversarial images. *ICLR Workshop*, 2017.

[Hill *et al.*, 2021] Mitch Hill, Jonathan Mitchell, and Song-Chun Zhu. Stochastic security: Adversarial defense using long-run dynamics of energy-based models. *ICLR*, 2021.

[Kang *et al.*, 2021] Qiyu Kang, Yang Song, Qinxu Ding, and Wee Peng Tay. Stable neural ode with lyapunov-stable equilibrium points for defending against adversarial attacks. *NeurIPS*, 2021.

[Kang *et al.*, 2023] Mintong Kang, Dawn Song, and Bo Li. Diffattack: Evasion attacks against diffusion-based adversarial purification. In *NeurIPS*, 2023.

[Kim *et al.*, 2023] Hannah Kim, Celia Cintas, Girmaw Abebe Tadesse, and Skyler Speakman. Spatially constrained adversarial attack detection and localization in the representation space of optical flow networks. In *IJCAI*, 2023.

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

[Li *et al.*, 2019] Jie Li, Rongrong Ji, Hong Liu, Xiaopeng Hong, Yue Gao, and Qi Tian. Universal perturbation attack against image retrieval. In *ICCV*, pages 4899–4908, 2019.

[Li *et al.*, 2020] Shasha Li, Shitong Zhu, Sudipta Paul, Amit Roy-Chowdhury, Chengyu Song, Srikanth Krishnamurthy, Ananthram Swami, and Kevin S Chan. Connecting the dots: Detecting adversarial perturbations using context inconsistency. In *ECCV*, 2020.

[Lin *et al.*, 2019] Yutian Lin, Liang Zheng, Zhedong Zheng, Yu Wu, Zhilan Hu, Chenggang Yan, and Yi Yang. Improving person re-identification by attribute and identity learning. *Pattern recognition*, 95:151–161, 2019.

[Liu *et al.*, 2019] Jiayang Liu, Weiming Zhang, Yiwei Zhang, Dongdong Hou, Yujia Liu, Hongyue Zha, and Nenghai Yu. Detection based defense against adversarial examples from the steganalysis point of view. In *CVPR*, 2019.

[Liu *et al.*, 2023] Han Liu, Xingshuo Huang, Xiaotong Zhang, Qimai Li, Fenglong Ma, Wei Wang, Hongyang Chen, Hong Yu, and Xianchao Zhang. Boosting decision-based black-box adversarial attack with gradient priors. In *IJCAI*, 2023.

[Lu *et al.*, 2023] Zhuoran Lu, Zhuoyan Li, Chun-Wei Chiang, and Ming Yin. Strategic adversarial attacks in ai-assisted decision making to reduce human trust and reliance. In *IJCAI*, 2023.

[Madry *et al.*, 2018] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *ICLR*, 2018.

[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.

[Saharia *et al.*, 2022] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH*, pages 1–10, 2022.

[Samangouei *et al.*, 2018] Pouya Samangouei, Maya Kabkab, and Rama Chellappa. Defense-gan: Protecting classifiers against adversarial attacks using generative models. *ICLR*, 2018.

[Shi *et al.*, 2023] Yucheng Shi, Mengnan Du, Xuansheng Wu, Zihan Guan, Jin Sun, and Ninghao Liu. Black-box backdoor defense via zero-shot image purification. *NeurIPS*, 2023.

[Song *et al.*, 2018] Yang Song, Taesup Kim, Sebastian Nowozin, Stefano Ermon, and Nate Kushman. Pixeldefend: Leveraging generative models to understand and defend against adversarial examples. *ICLR*, 2018.

[Sun *et al.*, 2023] Jiachen Sun, Jiongxiao Wang, Weili Nie, Zhiding Yu, Zhuoqing Mao, and Chaowei Xiao. A critical revisit of adversarial robustness in 3D point cloud recognition with diffusion-driven purification. In *ICML*, 2023.

[Sun *et al.*, 2024] Wenli Sun, Xinyang Jiang, Shuguang Dou, Dongsheng Li, Duoqian Miao, Cheng Deng, and Cairong Zhao. Invisible backdoor attack with dynamic triggers against person re-identification. *IEEE Transactions on Information Forensics and Security*, 19:307–319, 2024.

[Virtanen *et al.*, 2020] Pauli Virtanen, Ralf Gommers, Travis E Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, et al. Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature methods*, 17(3):261–272, 2020.

[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.*, 2021a] 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.

[Wang *et al.*, 2021b] Yajie Wang, Shangbo Wu, Wenyi Jiang, Shengang Hao, Yu-an Tan, and Quanxin Zhang. Demiguise attack: Crafting invisible semantic adversarial perturbations with perceptual similarity. *IJCAI*, 2021.

[Wang *et al.*, 2021c] Zeyuan Wang, Chaofeng Sha, and Su Yang. Reinforcement learning based sparse black-box adversarial attack on video recognition models. *IJCAI*, 2021.

[Wang *et al.*, 2022] Jinyi Wang, Zhaoyang Lyu, Dahua Lin, Bo Dai, and Hongfei Fu. Guided diffusion model for adversarial purification. *ICML*, 2022.

[Wang *et al.*, 2023] Qian Wang, Yongqin Xian, Hefei Ling, Jinyuan Zhang, Xiaorui Lin, Ping Li, Jiazhong Chen, and Ning Yu. Detecting adversarial faces using only real face self-perturbations. In *IJCAI*, 2023.

[Wu *et al.*, 2022] Quanlin Wu, Hang Ye, and Yuntian Gu. Guided diffusion model for adversarial purification from random noise. *arXiv :2206.10875*, 2022.

[Xiao *et al.*, 2023] Chaowei Xiao, Zhongzhu Chen, Kun Jin, Jiongxiao Wang, Weili Nie, Mingyan Liu, Anima Anandkumar, Bo Li, and Dawn Song. Densepure: Understanding diffusion models for adversarial robustness. In *ICLR*, 2023.

[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.

[Yang *et al.*, 2024] Zhaoyuan Yang, Zhiwei Xu, Jing Zhang, Richard Hartley, and Peter Tu. Adversarial purification with the manifold hypothesis. In *AAAI*, 2024.

[Yin *et al.*, 2021] Mingjun Yin, Shasha Li, Zikui Cai, Chengyu Song, M Salman Asif, Amit K Roy-Chowdhury, and Srikanth V Krishnamurthy. Exploiting multi-object relationships for detecting adversarial attacks in complex scenes. In *ICCV*, 2021.

[Yoon *et al.*, 2021] Jongmin Yoon, Sung Ju Hwang, and Juho Lee. Adversarial purification with score-based generative models. In *ICML*, 2021.

[Yu *et al.*, 2022] Chaojian Yu, Bo Han, Mingming Gong, Li Shen, Shiming Ge, Bo Du, and Tongliang Liu. Robust weight perturbation for adversarial training. *IJCAI*, 2022.

[Zhang *et al.*, 2023a] Jianping Zhang, Yung-Chieh Huang, Weibin Wu, and Michael R Lyu. Towards semantics-and domain-aware adversarial attacks. In *IJCAI*, 2023.

[Zhang *et al.*, 2023b] Shuhai Zhang, Feng Liu, Jiahao Yang, Yifan Yang, Changsheng Li, Bo Han, and Mingkui Tan. Detecting adversarial data by probing multiple perturbations using expected perturbation score. In *ICML*, 2023.

[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.

[Zheng *et al.*, 2016] Liang Zheng, Yi Yang, and Alexander G Hauptmann. Person re-identification: Past, present and future. *arXiv :1610.02984*, 2016.

[Zheng *et al.*, 2018a] Zhedong Zheng, Liang Zheng, Zhilan Hu, and Yi Yang. Open set adversarial examples. *arXiv :1809.02681*, 3, 2018.

[Zheng *et al.*, 2018b] Zhedong Zheng, Liang Zheng, and Yi Yang. A discriminatively learned cnn embedding for person reidentification. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 14(1):13, 2018.