# Feature-Level Adversarial Attacks and Ranking Disruption for Visible-Infrared Person Re-identification

**Xi Yang[1], Huanling liu[1], De Cheng[1]*, Nannan Wang[1], Xinbo Gao[2]**
[1]Xidian University, [2]Chongqing University of Posts and Telecommunications
yangx@xidian.edu.cn,huanlingliu@stu.xidian.edu.cn, dcheng@xidian.edu.cn
nnwang@xidian.edu.cn, gaoxb@cqupt.edu.cn

## Abstract

Visible-infrared person re-identification (VIReID) is widely used in fields such as video surveillance and intelligent transportation, imposing higher demands on model security. In practice, the adversarial attacks based on VIReID aim to disrupt output ranking and quantify the security risks of models. Although numerous studies have been emerged on adversarial attacks and defenses in fields such as face recognition, person re-identification, and pedestrian detection, there is currently a lack of research on the security of VIReID systems. To this end, we propose to explore the vulnerabilities of VIReID systems and prevent potential serious losses due to insecurity. Compared to research on single-modality ReID, adversarial feature alignment and modality differences need to be particularly emphasized. Thus, we advocate for feature-level adversarial attacks to disrupt the output rankings of VIReID systems. To obtain adversarial features, we introduce *Universal Adversarial Perturbations* (UAP) to simulate common disturbances in real-world environments. Additionally, we employ a *Frequency-Spatial Attention Module* (FSAM), integrating frequency information extraction and spatial focusing mechanisms, and further emphasize important regional features from different domains on the shared features. This ensures that adversarial features maintain consistency within the feature space. Finally, we employ an *Auxiliary Quadruple Adversarial Loss* to amplify the differences between modalities, thereby improving the distinction and recognition of features between visible and infrared images, which cause the system to output incorrect rankings. Extensive experiments on two VIReID benchmarks (i.e., SYSU-MM01, RegDB) and different systems validate the effectiveness of our method.

## 1 Introduction

VIReID is widely applied in key tasks such as security monitoring[1–3]. Suppose the law enforcement agency of a city uses ReID system to monitor public places for tracking criminal suspects. Internal personnel may attempt to deceive the system by modifying the images [4–6]of criminal suspects due to improper behavior or other reasons, in order to protect specific individuals. According to Figure1, infrared adversarial samples will erroneously match visible samples, while visible adversarial samples will also erroneously match infrared samples. The credibility and stability of VIReID are crucial in such special application scenarios. However, there is currently insufficient theoretical research on the security of VIReID. Therefore, this paper explores how to obtain better adversarial features and how to address the task characteristics of VIReID modality differences and re-ranking.

*Corresponding author.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

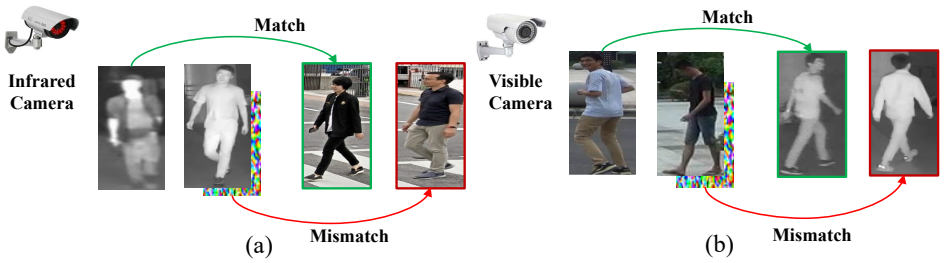



Figure 1: Security risks of VIReID in the physical world. Images with added noise are referred to as adversarial samples. Red indicates that adversarial samples will incorrectly match pedestrians. Green indicates that clean samples will correctly match pedestrians.

The current research on adversarial attacks in the fields of face recognition and person re-identification mainly focuses on digital attacks [7–9], which refer to the manipulation, distortion, or tampering of digital images to deceive, disrupt, or mislead systems, leading to erroneous identification results or reduced recognition accuracy[10]. Adversarial attacks based on ReID focus on crafting adversarial samples suitable for visible images and disrupting internal rankings. In contrast, VIReID requires more consideration regarding the generalizability of attack methods under different imaging mechanisms and how to align adversarial features between infrared and visible images. Moreover, dealing with two modalities necessitates comparing distances between modalities and within modalities. To address these challenges, we propose a feature-level adversarial attacks and ranking disruption for VIReID.

First, we adopt the method of universal adversarial perturbation (UAP)[11] to generate adversarial samples, which seeks a set of universal perturbations independent of the image and can generalize well in deep neural networks. At the same time, it significantly lowers the threshold for implementing adversarial attacks and adapts to different systems. Secondly, to make visible and infrared images more consistent in the feature space, we propose a frequency-spatial attention module, achieving adversarial feature alignment by unifying frequency and spatial features. Visible and infrared images are generated under different imaging conditions; the former provides rich texture information, while the latter contains significant pixel amplitude information. This module uses fast Fourier transform to decompose features into amplitude and phase, corresponding to the texture details and spatial position information of images, respectively. Since these features are closely related to the spatial domain, a spatial attention module is chosen for these two components to further emphasize or suppress different regions in the feature map. Additionally, a weighted spatial attention module is applied to the shared features to maintain consistency in the feature space. This method not only focuses on the frequency domain features of visible and infrared images from different imaging conditions but also emphasizes their spatial features for significant pedestrian poses, thereby achieving adversarial feature alignment. To disrupt the ranking of identification results, we propose an auxiliary quadruple adversarial loss function. The visible and infrared pedestrian features extracted by the first-stage model are used as auxiliary features in the calculation process of the features loss function extracted in the second stage. By pulling the distance between the same modality and the same person closer and pushing the distance between different modalities and the same person farther, while also ensuring that the intra-class distance is smaller than the inter-class distance, the differences between modalities are expanded and the ranking under different modalities is disrupted. By utilizing generated features containing multiple types of information, the network's ability to explore features at different levels is enhanced. That is, with four types of features, we ensure a double guarantee to achieve the goal of disrupting the ranking. The main contributions of this paper can be summarized as follows:

- We are the first to propose exploring the security of VIReID, considering the alignment of adversarial features across modalities in VIReID.
- We propose a frequency-spatial attention module that integrates frequency-domain features with spatial features to enhance the consistency and representation ability of adversarial features.

- We design an auxiliary quadruple adversarial loss function, which utilizes auxiliary features to amplify the differences within and between modalities, thereby disrupting the ranking results.

## 2 Related Work

**Visible-Infrared Person Re-identification.** VIReID [12–15]refers to the technique of identifying and matching pedestrians from one modality to another using visible or infrared images. Moreover, VIReID finds wide application in the field of security, enhancing the intelligence level of security systems. The significant differences between different modalities make VIReID challenging. To alleviate the modality discrepancy at the feature level, some methods adopt single-stream, dual-stream, or multi-stream networks[16–19], extracting shared features from different modalities by designing various attention mechanisms and loss functions[20–23]. Ye et al.[16] proposed the concept of metric learning, jointly optimizing modality-specific and modality-shared matrices. Subsequently, Zhu et al.[18] proposed a hetero-center loss, which for the first time shortens the distance between feature centers of the same identity, bridging the gap between features of the same pedestrian across different modalities. In addition, Ling et al.[24] devised a multi-constraint similarity learning approach to comprehensively explore the relationships between cross-modal information. Meanwhile, as the posture and shape of pedestrians provide important information in the recognition process, these methods mainly focus on spatially enhancing feature representation. However, there are also differences in frequency information between visible and infrared images. Li et al.[25] proposed a novel frequency-domain modality-invariant feature learning framework to reduce modality differences from a frequency-domain perspective.

**Adversarial Attacks.** In the fields of computer graphics and pattern recognition, adversarial attacks[26–29] are an important area of research. Many studies have revealed the vulnerability of deep models to carefully crafted small perturbation, resulting in significant errors with high confidence in predictions. Adversarial attacks aim to explore adversarial noise that causes deep learning models to behave abnormally. FGSM[30] belongs to the single-step attack algorithm, which optimizes the loss by quickly determining the direction of perturbation for input samples and calculates the adversarial perturbation through backpropagation. The concept of iterative thinking was subsequently incorporated into FGSM, thereby leading to the development of Projected Gradient Descent [31] (PGD). Subsequently, adversarial attacks on ReID were also studied, with the core idea of generating well-crafted adversarial examples or disturbing ranking results. LTA[32] used local grayscale iteration to generate adversarial examples, mainly focusing on disturbing the color of the original image. A mis-ranking formula was proposed by DMR[33] to increase the distance between images of the same pedestrian while decreasing the distance between images of different pedestrians, effectively disrupting the ranking results.

**Cross-modality Attacks.** Adversarial instances are widely present across various domains of visible images. In recent years, exploration of adversarial instances in the field of infrared pedestrian detection has begun. Osahor et al.[34] discussed perturbation by altering pixel values within infrared images. Subsequently, Zhu et al.[35] attempted for the first time to alter the infrared radiation distribution of the human body by simulating additional heat sources using a set of small bulbs, generating physical adversarial examples. To be more easily implemented in the physical world, Wei et al.[36, 37] considered the different imaging mechanisms of visible and infrared sensors, proposing a unified adversarial patch to execute cross-modality physical attacks. Meanwhile, we chose to perform security evaluation on visible-infrared person re-identification models proposed in recent years in the digital world.

## 3 Proposed Method

As shown in Figure 2, the overall structure of the proposed method adopts a dual-stream ResNet network as the backbone. Firstly, *Universal Adversarial Perturbation* (UAP) are added separately to visible and infrared images. In the first stage, the *Frequency-Spatial Attention Module* (FSAM) is embedded to extract frequency-domain spatial correlated features as auxiliary features for images of two different modalities, visible and infrared. Subsequently, through a shared module, further focus is applied to the temporal domain features. This completes the focus on spatial features in both frequency and normal domains, making the learned information more diversified, thereby completing

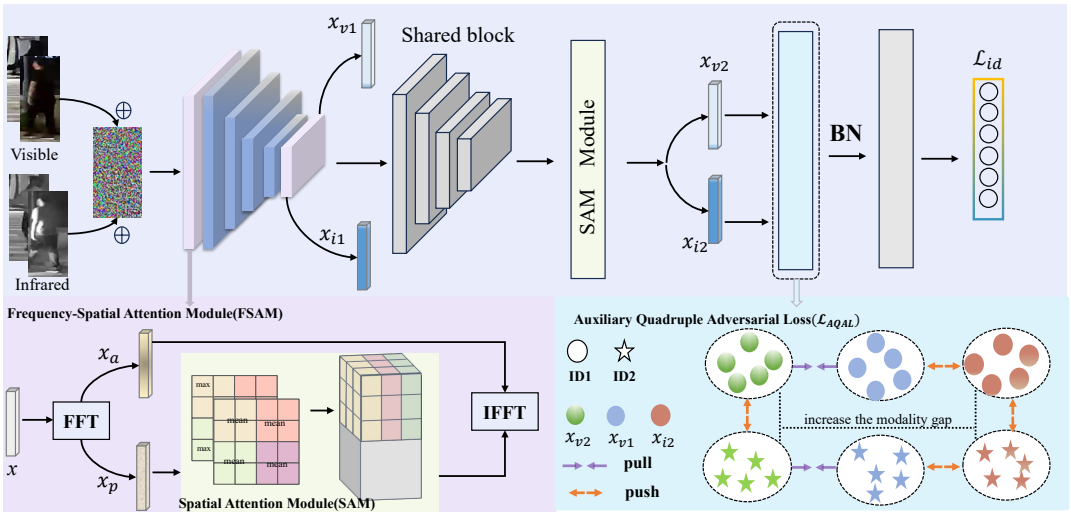



Figure 2: Overview of the proposed method. FSAM consists of FFT and IFFT along with a spatial focusing module, focusing on the frequency-domain spatial characteristics of the image. We propose a auxiliary quadruple adversarial loss function and provide a simple illustration of its operation.

the learning of features in the second stage. During the training phase, the features from the first stage are used as auxiliary features and combined with the features from the second stage, all of which are inputted into the *Auxiliary Quadruple Adversarial Loss* for optimizing the entire module.

### 3.1 Universal Adversarial Perturbation

Universal Adversarial Perturbation (UAP) aims to generate a single perturbation that can be added to any image from the same distribution, resulting in mis-classification when added. Deep neural networks are highly susceptible to this type of UAP, yet they remain imperceptible to the human eye. Defining a set of images $I$ that satisfies distribution $\mu$, where $g(I)$ fits the output function, and after perturbation $\delta$, the labels are not equal. That is:

$$g(I+\delta) \neq g(I). \tag{1}$$

When $\delta$ is constrained by two conditions simultaneously, the optimization problem of finding universal perturbation can be described as follows:

$$s.t. ||\delta||_p \leq \varepsilon, \quad P_{I\sim\mu}(g(I+\delta) \neq g(I)) \geq 1-\epsilon, \tag{2}$$

where $P(\cdot)$ represents probability, $||\cdot||_p$ denotes the p-norm, $\varepsilon$ indicates the magnitude of the perturbation to ensure that the adversarial perturbation is visually imperceptible, $\mu$ represents the data distribution, $\epsilon \in (0,1]$ denotes the success rate of deception to ensure the attack's success rate. The goal is to find an adversarial perturbation $\delta$ that can be added to all sample points and will result in misclassification of adversarial samples with a probability of 1 - $\epsilon$. The UAP algorithm does not require solving optimization problems or gradients of the model and is applied in scenarios where a large number of adversarial samples need to be quickly generated.

So we choose universal perturbation for addition, which can be used on both visible and infrared images and can also adapt to scenarios with a large number of pedestrian samples in real-world environments. The introduction of UAP can make the model more independent of specific data distributions and training sets, thereby reducing the model's dependence on specific data and making the system more flexible and adaptable to challenges in different environments and scenarios. Visible and infrared images generate their respective adversarial samples as follows:

$$\hat{Q}_{vi} = Q_{vi} + \delta_{vi}, \tag{3}$$

$$\hat{Q}_{ir} = Q_{ir} + \delta_{ir}. \tag{4}$$

For the input $Q_{vi}$ and $Q_{ir}$, adding perturbation($\delta_{vi}$,$\delta_{ir}$) related to the data distribution to each, the generated adversarial sample $\hat{Q}_{vi}$, $\hat{Q}_{ir}$ can deceive the system by exploiting visual similarity attacks.

### 3.2 Frequency-Spatial Attention Module

#### 3.2.1 Fast Fourier Transform

The Fast Fourier Transform (FFT) is widely utilized in the field of image processing to convert images into the frequency domain, enabling the analysis of the frequency components of the image. This aids in understanding the overall structure, texture, and edge information of the image. Therefore, we combine the Fast Fourier Transform with a spatial attention module to focus on the unique features in the frequency domain of the image, which aids in enhancing feature representation capabilities. First, we provide a brief introduction to the basic concepts of the FFT. Given the feature $\boldsymbol{x} \in \mathbb{R}^{C\times H\times W}$ output by network, its FFT can be expressed as follows:

$$\mathcal{F}(\boldsymbol{x})(u,v) = \frac{1}{\sqrt{HW}} \sum_{h=0}^{H-1} \sum_{w=0}^{W-1} x(h,w) e^{-2j\pi(u\frac{h}{H}+v\frac{w}{W})}, \tag{5}$$

where $j$ represents the imaginary unit, $u$ and $v$ are the horizontal and vertical coordinates of the $\boldsymbol{x}$, and $\mathcal{F}(\cdot)$ denotes the Fourier transform, $C$, $H$ and $W$ denote the number of the channel, height and width of features. The frequency-domain feature $\mathcal{F}(\boldsymbol{x})$ is represented as $\mathcal{F}(\boldsymbol{x}) = \mathcal{R}(\boldsymbol{x}) + j\mathcal{I}(\boldsymbol{x})$, where $\mathcal{R}(\boldsymbol{x})$ and $\mathcal{I}(\boldsymbol{x})$ represent the real and imaginary part of $\mathcal{F}(\boldsymbol{x})$. These real and imaginary parts can be converted to amplitude and phase spectrums, which can be formulated as follows:

$$\mathcal{A}(\boldsymbol{x})(u,v) = \left[\mathcal{R}^2(\boldsymbol{x})(u,v) + \mathcal{I}^2(\boldsymbol{x})(u,v)\right]^{1/2}, \tag{6}$$

$$\mathcal{P}(\boldsymbol{x})(u,v) = \arctan\left[\frac{\mathcal{I}(\boldsymbol{x})(u,v)}{\mathcal{R}(\boldsymbol{x})(u,v)}\right]. \tag{7}$$

As shown in Figure 3, in the task of VI-ReID, the amplitude component captures the overall brightness and contrast of pedestrian images, reflecting the luminance and color information of the image, while the phase component captures the structural information and details of the pedestrians, including their shape and outline, to help distinguish between different pedestrians' details and features. The combination of these components allows for the effective extraction of global and local features of pedestrians. By focusing further on spatial information characteristics in the phase component, attention to spatial information in the frequency domain is increased, enhancing the ability to express distinguishing features. In the context of the features $\boldsymbol{x}$ extracted by the network, we represent the amplitude component of the FFT as $\boldsymbol{x}_a$, and the phase component as $\boldsymbol{x}_p$. Given that the dual-branch network separately extracts visible and infrared pedestrian features,

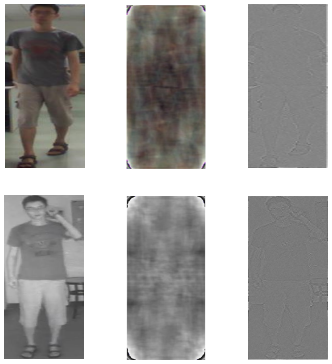



Figure 3: Decomposition and reconstruction of visible and infrared image in the frequency domain. (a) denote visible and infrared images of pedestrian; (b)present the reconstructed images with amplitude information only; (c) are the reconstructed images with phase information only.

the amplitude component of the visible pedestrian features after FFT is represented as $\boldsymbol{x}_{va}$ and the phase component as $\boldsymbol{x}_{vp}$, while the amplitude component of the infrared pedestrian features after FFT is represented as $\boldsymbol{x}_{ia}$and the phase component as $\boldsymbol{x}_{ip}$.

#### 3.2.2 Spatial Attention Module

The spatial attention module generates spatial attention maps using the spatial relationships within the features. It focuses on the distribution of information, locating the position and shape of the subject while reducing background information interference. We apply max pooling and average pooling operations along the channel axis. These two pooling operations respectively extract the maximum and average value information from the feature map and concatenate them to form a richer feature descriptor. Pooling is performed along the channel axis, which helps highlight information-rich

regions in the feature map and better locate key information. In summary, the specific calculation process is as follows:

$$M(\boldsymbol{x}) = Sigmoid(f([MaxPool(\boldsymbol{x}); AvgPool(\boldsymbol{x})])), \tag{8}$$

where $M(\boldsymbol{x})$ denotes the weight distribution generated by applying the convolution layer and *Sigmoid* denotes the sigmoid function. $f$ represents a convolution operation with the filter size of $7 \times 7$. $MaxPool(\boldsymbol{x})$ and $AvgPool(\boldsymbol{x})$ represent the features after passing through the max-pooling layer and average-pooling layer, respectively.

Given that the phase component involves more spatial information, we sequentially apply the spatial attention module to obtain the new phase component($\boldsymbol{x}_{vp}^{'}$,$\boldsymbol{x}_{ip}^{'}$), which can be expressed as follows:

$$\boldsymbol{x}_{vp}^{'} = M_v(\boldsymbol{x}_{vp}) \otimes \boldsymbol{x}_{vp}, \tag{9}$$

$$\boldsymbol{x}_{ip}^{'} = M_i(\boldsymbol{x}_{ip}) \otimes \boldsymbol{x}_{ip}, \tag{10}$$

where $\otimes$ denotes element-wise multiplication, $M_v(\boldsymbol{x}_{vp})$ and $M_i(\boldsymbol{x}_{ip})$ represent the attention weights generated by the feature phase components. The combination of the original amplitude component and the new phase component can reconstruct the original feature information through inverse fast Fourier transform(IFFT). At this point, the network outputs features $\boldsymbol{x}_{v1}$ and $\boldsymbol{x}_{i1}$, which will serve as auxiliary features. This process can be described as:

$$x_{v1} = IFFT(x_{vp}^{'}, x_{va}), \tag{11}$$

$$x_{i1} = IFFT(x_{ip}^{'}, x_{ia}). \tag{12}$$

The features are then further input into a shared module $\mathcal{T}$ in the model to complete feature extraction. At this point, choosing to pass through the spatial attention module allows for focusing on spatial information in the original domain. This two-stage process completes the focus on both frequency domain and original domain spatial information, enhancing the representation capability of distinguishing features. Thus, the final extracted visible and infrared features can be expressed as:

$$\boldsymbol{x}_{v2} = M_v^{'}(\mathcal{T}(\boldsymbol{x}_{v1})) \otimes \mathcal{T}(\boldsymbol{x}_{v1}), \tag{13}$$

$$\boldsymbol{x}_{i2} = M_i^{'}(\mathcal{T}(\boldsymbol{x}_{i1})) \otimes \mathcal{T}(\boldsymbol{x}_{i1}), \tag{14}$$

where $M_v^{'}$ and $M_i^{'}$ indicates the weight distribution generated by the second-stage SAM module, resulting in the final stage features, namely $\boldsymbol{x}_{v2}$ and $\boldsymbol{x}_{i2}$.

### 3.3 Auxiliary Quadruple Adversarial Loss

We propose an auxiliary quadruplet adversarial loss function to disrupt the system's output ranking. This method effectively adapts to ReID issues by attacking the predicted ranking results. Additionally, considering the modality differences involved in VIReID, we introduce auxiliary features to fully leverage the information disparities between modalities, erroneously outputting cross-modality ranking results. In this process, there are both modality differences and identity differences. Therefore, we approach it from two aspects: under the condition of the same modality, minimizing the distance between the same identities and maximizing the distance between different identities; under the condition of the same identity, minimizing the distance between the same modalities and maximizing the distance between different modalities. Let's start by controlling the same modality case:

$$\mathcal{L}(\boldsymbol{x}_{v2}, \boldsymbol{x}_{v1}) = \left[ D(\boldsymbol{x}_{v2}^{j}, \boldsymbol{x}_{v1}^{j}) - D(\boldsymbol{x}_{v2}^{j}, \boldsymbol{x}_{v2}^{k}) \right]_{+}, \tag{15}$$

where $\boldsymbol{x}_{v1}^{j}$ represents the auxiliary feature, $j$ and $k$ denote different pedestrians, $D(\cdot, \cdot)$ represents the Euclidean distance between two feature vectors and $[x]_{+} = max(x, 0)$. When the modalities are the same, the task is transformed into a single-modality ReID. This problem discusses the disruption of matching pairs across modalities, while still maintaining the original requirement for the same modality, namely, encouraging that the maximum distance between the most easily identifiable pairs of images across identities is still less than the minimum distance between the most easily identifiable pairs of images within an identity, ensuring that the sorting output within the same modality is normal.

This part ensures that the original order and matching relationships within the modality are not disrupted.

When the identities are consistent across modalities:

$$\mathcal{L}(\boldsymbol{x}_{v2}, \boldsymbol{x}_{v1}, \boldsymbol{x}_{i2}) = \left[ D(\boldsymbol{x}_{v2}^{j}, \boldsymbol{x}_{v1}^{j}) - D(\boldsymbol{x}_{i2}^{j}, \boldsymbol{x}_{v1}^{j}) \right]_{+}, \tag{16}$$

we encourage reducing the distance between pedestrians in the same modality while increasing the distance between pedestrians in different modalities. This strategy aims to narrow the distance within the same modality while widening the distance between different modalities, thereby increasing the modalities differences. It disperses the features in the feature space, making it more challenging for the model to cluster features. This is intended to impact the matching and sorting results across modalities. The loss function augmented with visible features can be expressed as:

$$\mathcal{L}(\boldsymbol{x}_{v2}, \boldsymbol{x}_{i2}, \boldsymbol{x}_{v1}) = \sum_{\substack{j,k=1 \\ j \neq k}}^{N} \left[ D(\boldsymbol{x}_{v2}^{j}, \boldsymbol{x}_{v1}^{j}) - D(\boldsymbol{x}_{i2}^{j}, \boldsymbol{x}_{v1}^{j}) - D(\boldsymbol{x}_{v2}^{j}, \boldsymbol{x}_{v2}^{k}) + \alpha \right]_{+}, \tag{17}$$

where $N$ is the number of person ID in a mini-batch. Meanwhile, using more discriminative embedding centers($\boldsymbol{c}_{v1}$, $\boldsymbol{c}_{v2}$, $\boldsymbol{c}_{i1}$, $\boldsymbol{c}_{i2}$) for each class, we introduce a margin term $\alpha$ to balance the two terms. Thus, the loss function augmented with visible features and infrared features can be expressed as follows:

$$\mathcal{L}(\boldsymbol{c}_{v2}, \boldsymbol{c}_{i2}, \boldsymbol{c}_{v1}) = \sum_{\substack{j,k=1 \\ j \neq k}}^{N} \left[ D(\boldsymbol{c}_{v2}^{j}, \boldsymbol{c}_{v1}^{j}) - D(\boldsymbol{c}_{i2}^{j}, \boldsymbol{c}_{v1}^{j}) - D(\boldsymbol{c}_{v2}^{j}, \boldsymbol{c}_{v2}^{k}) + \alpha \right]_{+}, \tag{18}$$

$$\mathcal{L}(\boldsymbol{c}_{i2}, \boldsymbol{c}_{v2}, \boldsymbol{c}_{i1}) = \sum_{\substack{j,k=1 \\ j \neq k}}^{N} \left[ D(\boldsymbol{c}_{i2}^{j}, \boldsymbol{c}_{i1}^{j}) - D(\boldsymbol{c}_{v2}^{j}, \boldsymbol{c}_{i1}^{j}) - D(\boldsymbol{c}_{i2}^{j}, \boldsymbol{c}_{i2}^{k}) + \alpha \right]_{+}. \tag{19}$$

Finally, the auxiliary quadruplet adversarial loss function ($\mathcal{L}_{AQAL}$) is ultimately formulated as:

$$\mathcal{L}_{AQAL} = \mathcal{L}(\boldsymbol{c}_{v2}, \boldsymbol{c}_{i2}, \boldsymbol{c}_{v1}) + \mathcal{L}(\boldsymbol{c}_{i2}, \boldsymbol{c}_{v2}, \boldsymbol{c}_{i1}). \tag{20}$$

The $\mathcal{L}_{AQAL}$ forces the distance between modalities to increase, preventing them from easily clustering together, thereby disrupting the overall ranking results.

### 3.4 Objective Function

Besides the auxiliary quadruple adversarial loss $\mathcal{L}_{AQAL}$, we also have the identity loss $\mathcal{L}_{id}$. The training process of VIReID is considered an image classification problem, where each identity is a distinct class. During the testing phase, the output from the pooling layer or embedding layer is used as the feature extractor. Given an input image $x_i$ with label $y_i$, the probability of $x_i$ being recognized as class $y_i$is encoded using the softmax function and denoted as $p(y_i|x_i)$. The identity loss is then computed by the cross-entropy

$$\mathcal{L}_{id} = -\frac{1}{N} \sum_{i=1}^{N} log(p(y_i|x_i)), \tag{21}$$

where N represents the number of training samples within each batch.

## 4 Experiments

### 4.1 Implementation Details

The experiments are conducted on an NVIDIA GeForce 3090 GPU with Pytorch. We chose the powerful baseline model AGW[38], which is a ResNet-50 pretrained on ImageNet, as the backbone network. During training, we randomly sampled 16 identities, each with 4 images, to form a mini-batch of size 64. Pedestrian images are resized to 288 $\times$ 144. Data augmentation included random

Table 1: Comparison of CMC (%) and mAP (%) with the state-of-the-art methods on SYSU-MM01 and RegDB datasets. Our results show the best results in terms of Rank-1 accuracy and mAP .

| methods | SYSU-MM01 | | | | | | | | RegDB | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | All-search | | | | Indoor-search | | | | Visible to Thermal | | | Thermal to Visible | | |
| | Rank-1 | Rank-10 | mAP | mINP | Rank-1 | Rank-10 | mAP | mINP | Rank-1 | mAP | mINP | Rank-1 | mAP | mINP |
| Before Attack | 47.50 | 84.39 | 47.65 | 35.30 | 54.17 | 91.14 | 62.97 | 59.23 | 70.05 | 66.37 | - | 70.49 | 65.90 | - |
| FGSM [30] | 36.02 | 59.22 | 31.80 | 19.72 | 33.25 | 69.17 | 46.31 | 30.36 | 45.87 | 44.39 | 36.59 | 46.15 | 46.12 | 43.24 |
| PGD [31] | 26.60 | 46.65 | 25.67 | 16.90 | 36.89 | 75.19 | 43.50 | 27.20 | 29.37 | 26.87 | 17.83 | 29.05 | 29.12 | 17.14 |
| SMA [39] | 21.72 | 39.80 | 21.34 | 20.39 | 25.77 | 51.24 | 30.91 | 29.96 | 19.85 | 17.37 | 9.49 | 16.57 | 18.23 | 10.84 |
| UAP [11] | 17.59 | 56.81 | 25.35 | 20.89 | 35.34 | 47.86 | 30.75 | 19.13 | 29.51 | 24.42 | 16.64 | 19.61 | 18.95 | 12.67 |
| LTA [32] | 15.47 | 30.39 | 17.71 | 13.44 | 21.68 | 38.56 | 26.61 | 20.59 | 11.60 | 10.86 | 6.07 | 14.56 | 13.11 | 7.75 |
| DMR [33] | 9.20 | 24.43 | 10.21 | 4.71 | 13.62 | 27.14 | 14.94 | 6.27 | 4.97 | 4.80 | 2.12 | 6.09 | 5.26 | 3.54 |
| Ours | **0.79** | **9.83** | **2.81** | **1.69** | **1.68** | **17.35** | **6.73** | **5.65** | **0.49** | **0.85** | **0.57** | **0.71** | **0.89** | **0.60** |

Table 2: Comparison of CMC (%) and mAP (%) of different VIReID systems before and after attack. Bold numbers indicate values after attack.

| methods | SYSU-MM01 | | | | | | RegDB | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | All-search | | | Indoor-search | | | Visible to Thermal | | | Thermal to Visible | | |
| | Rank-1 | Rank-10 | mAP | Rank-1 | Rank-10 | mAP | Rank-1 | Rank-10 | mAP | Rank-1 | Rank-10 | mAP |
| HCLoss[18] | 56.96/**1.09** | 91.50/**10.37** | 54.95/**2.91** | 59.74/**2.09** | 92.07/**18.85** | 64.91/**7.17** | 86.02/**2.67** | 96.36/**10.73** | 74.80/**2.82** | 87.28/**1.02** | 97.04/**4.66** | 78.30/**2.07** |
| CAJ[40] | 69.88/**1.04** | 95.71/**10.32** | 66.89/**2.99** | 76.26/**2.45** | 97.88/**18.93** | 80.37/**7.66** | 85.0/**1.04** | 95.5/**10.32** | 84.6/**2.99** | 88.3/**2.45** | 98.5/**18.93** | 81.9/**7.66** |
| MMN [41] | 70.60/**1.39** | 96.2/**4.55** | 66.9/**3.81** | 76.2/**4.26** | 97.2/**27.81** | 79.6/**8.16** | 91.6/**0.49** | 97.7/**0.97** | 84.1/**1.71** | 87.5/**3.50** | 96.0/**14.08** | 80.5/**3.72** |
| DEEN [42] | 74.70/**1.71** | 97.60/**10.49** | 71.80/**3.55** | 80.30/**2.04** | 99.00/**17.96** | 83.30/**7.34** | 91.1/**3.15** | 97.8/**12.48** | 85.1/**4.05** | 89.5/**2.85** | 96.8/**16.34** | 83.4/**6.21** |

horizontal flipping and random erasing with a probability of 0.5. We optimize using the stochastic gradient descent (SGD) optimizer, with a weight decay set to 0.0005 and a momentum parameter set to 0.9. The initial learning rate for both datasets was set to 0.1, and it was decayed by a factor of 0.1 at the 20th and 50th epochs, respectively. Finally, the margin $\alpha$ in the auxiliary quadruplet adversarial loss function is set to 0.2. We adopt a warm-up learning rate scheme, with a total of 60 training epochs.

### 4.2 Datasets

SYSU-MM01[1] is a cross-modality pedestrian re-identification dataset proposed in 2017. There are 287,628 visible images and 15,792 infrared images in total. Cameras 1 and 2 are installed in well-lit environments, cameras 3 and 6 operate under infrared conditions, and cameras 4 and 5 are placed in outdoor scenes. The dataset comprises two testing modes: all-search and indoor-search. The all-search mode is more challenging because the gallery includes images from all cameras.

The RegDB [43]dataset comprises 412 individuals, with each person having 20 images, including 10 visible images and 10 infrared images. Among the 412 individuals, there are 254 females and 158 males, with 156 individuals captured in frontal views and 256 individuals captured in rear views. During the testing phase, RegDB offers two modes: visible to infrared and infrared to visible. In the visible to infrared mode, visible images serve as query images, while infrared images are used as gallery images. The infrared to visible mode operates in the opposite manner.

### 4.3 Comparison with State-of-the-Art Methods

We select six different methods to evaluate the security of the AGW model. Among them, FGSM and PGD are typical gradient-based methods for generating adversarial samples, while UAP generates universal perturbations that are independent of the image. Additionally, we chose SMA, DMR, and LTA, three pedestrian re-identification attack methods which primarily disrupt ranking and generate adversarial samples through local color iteration. The results, as shown in Table1, indicate a sharp decline in all evaluation metrics after attacking AGW across two datasets and two different tests. In the RegDB dataset, under two testing modes, our mAP decreased dramatically from 66.37% and 65.90% to 0.85% and 0.89%, nearly approaching 0. This demonstrates that our method is more suitable for testing robustness against VIReID tasks.

To demonstrate the universality of our method, we select three approaches to improve the recognition accuracy of VIReID systems: HCLoss enhances intra-class cross-modality similarity through a heterogeneous center loss function, CAJ improves recognition accuracy based on data augmentation techniques, and MMN and DEEN utilize different network designs to better extract effective shared features. These methods are commonly used to improve VIReID accuracy at the feature level. As

Table 3: Analysis about the influence of each component in terms of Rank-1 (%) and mAP (%).

| Noise | FSAM | SAM | $\mathcal{L}_{AQAL}$ | SYSU-MM01 | | RegDB | |
|---|---|---|---|---|---|---|---|
| | | | | Rank-1 | mAP | Rank-1 | mAP |
| | | | | 46.73 | 45.78 | 82.24 | 76.52 |
| ✓ | | | | 16.67 | 18.28 | 33.00 | 30.60 |
| ✓ | ✓ | | | 12.75 | 12.17 | 15.74 | 12.91 |
| ✓ | | ✓ | | 11.04 | 13.19 | 20.73 | 19.81 |
| ✓ | | | ✓ | 14.62 | 15.52 | 15.86 | 15.53 |
| ✓ | ✓ | ✓ | | 5.10 | 6.35 | 7.58 | 7.51 |
| ✓ | ✓ | | ✓ | 1.07 | 3.27 | 0.58 | 1.35 |
| ✓ | | ✓ | ✓ | 1.74 | 3.81 | 6.70 | 5.77 |
| ✓ | ✓ | ✓ | ✓ | **0.79** | **2.81** | **0.49** | **0.85** |

Table 4: Performance comparison of different feature extraction methods in terms of CMC (%) and mAP (%) on RegDB. (Setting: Baseline + SFM + $\mathcal{L}_{AQAL}$.)

| Settings + FSAM | SYSU-MM01 | | | | RegDB | | | |
|---|---|---|---|---|---|---|---|---|
| | Rank-1 | Rank-10 | mAP | mINP | Rank-1 | Rank-10 | mAP | mINP |
| block0 | 1.05 | 10.60 | 3.15 | 2.04 | 0.49 | 4.17 | 1.07 | 1.27 |
| block1 | 1.05 | 10.52 | 4.29 | 5.33 | 0.53 | 1.69 | 1.68 | 2.41 |
| block2 | 1.05 | 10.07 | 3.56 | 2.25 | 0.54 | 2.94 | 1.23 | 1.11 |
| block3 | 1.34 | 10.41 | 4.07 | 4.72 | 0.49 | **1.46** | 1.67 | 2.41 |
| block4 | 1.01 | 10.12 | 2.91 | **1.68** | 0.49 | 4.71 | 0.93 | 0.67 |
| block0-block4(ours) | **0.79** | **9.83** | 2.**81** | 1.69 | **0.48** | 4.64 | **0.85** | **0.57** |

shown in Table2, our experimental results show that after applying our attack method, the mAP of all systems dramatically decreased across different datasets and test modes (for example, the Rank-1 of DEEN drops from 74.70% to 1.71%, and mAP falls from 71.80% to 3.55% in all-search mode), indicating that these systems cannot withstand our attacks. This demonstrates that our method can effectively test the robustness of various VIReID systems.

### 4.4 Ablation Study

As shown in Table3, the effectiveness of different modules is validated on two datasets. Using the SYSU-MM01 dataset as an example, after adding noise to generate adversarial samples, the mAP decreases from 45.75% to 18.28%. Subsequently, the three modules are individually tested on this basis, and all show a decrease in accuracy, with the FSAM module being the most effective, reducing the mAP to 12.17%. Additionally, since both FSAM and SAM are designed to enhance adversarial feature representation capabilities, the combined effect of these two modules is tested, further reducing the mAP to 6.35%. The experiments demonstrate that all three proposed modules are highly effective. Although targeting different aspects, their combined usage enhances the effectiveness of method.

Determining which stage of ResNet-50 should have the FSAM module inserted. In this experiment, we use ResNet-50 as the backbone, which has five stages: block 0 to block 4. We study the impact on model performance by inserting the FSAM module after different stages of ResNet-50. This experiment controls only the insertion position of FSAM as the variable, keeping all other factors constant. As shown in Table4, inserting FSAM after block 0 and block 4 results in lower accuracy, indicating better performance. This suggests that in the shallow layers of the network, image processing captures more effective information from the initial frequency domain features, while in the deeper layers of the network, the feature representation capability is further enhanced, and spatial information becomes more important. Based on these results, we integrate the proposed FSAM module after block 0 and block 4 of the ResNet-50 model.

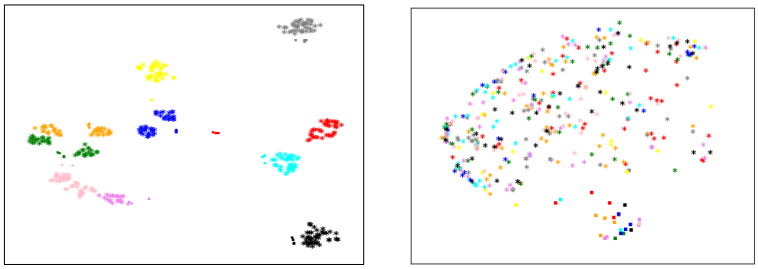
As shown in Table 5, we compared our frequency domain attention module with channel attention mechanisms like SE-Net [44] and ECA-Net[45]. The results reveal that while SE-Net and ECA-Net excel in certain areas, our frequency domain attention module is highly competitive across various metrics. It achieves the best Rank-1 and mAP scores on both datasets. Channel attention mechanisms often focus on global features, which may overlook local details and spatial relationships in cross-modal VIReID tasks, leading to potential information loss and reduced performance.

Table 5: Performance of the FSAM module replacing different attention mechanisms on CMC(%) and mAP(%).

| Attention | SYSU-MM01 | | | | RegDB | | | |
|---|---|---|---|---|---|---|---|---|
| | Rank-1 | Rank-10 | mAP | mINP | Rank-1 | Rank-10 | mAP | mINP |
| SENet | 1.66 | 12.15 | 3.30 | 1.76 | 0.53 | 5.00 | 1.15 | 0.64 |
| ECA-Net | 1.60 | 11.65 | 3.24 | **1.65** | 0.68 | **3.64** | 1.20 | 0.68 |
| Ours) | **0.79** | **9.83** | **2.81** | 1.69 | **0.48** | 4.64 | **0.85** | **0.57** |

### 4.5 Visualization Analysis

We compare the t-SNE visualization results on the baseline and the proposed method. To ensure fairness, we randomly select several images of ten identities from all cameras. For each individual, 20 visible images and 20 infrared images are randomly chosen. As shown in Figure4, after the attack, the clustering of all pedestrians became more dispersed, with the visible and infrared modalities each forming their own clusters, and the relative distance between them increasing. The visualization experiments validated the effectiveness of our attack method, as the attack further exaggerated the gap between the two modalities, effectively suppressing the output performance. These results demonstrate that our method is highly effective for security testing in VIReID.



Baseline Attacked

Figure 4: t-SNE visualization comparison before and after the attack. Different colors represent different identities. The 'asterisks' and 'rectangles' denote the infrared person features and visible person features, respectively. After the attack, the features are dispersed, enlarging the distance between modalities.

## 5 Conclusion

In this paper, our objective is to conduct security validation of VIReID systems and propose a novel attack method suitable for cross-modality tasks. We introduce the frequency-spatial attention module, which are used at two stages of feature extraction, focusing on spatial information in both frequency and source domains to enhance the representation capability of adversarial features and strengthen the effectiveness of adversarial samples. Additionally, we propose an auxiliary quadruple adversarial loss function considering the modalities differences involved in VIReID tasks to interfere with the ranking of system outputs, completing the robustness test of the current VIReID system. Extensive experiments not only deepens the understanding of the security of cross-modality ReID systems but also provides a new direction for the development of VIReID and emphasizes the importance of ensuring their reliability and protection in practical applications.

## 6 Acknowledgement

This work was supported in part by the National Natural Science Foundation of China under Grant 62372348, Grant 62441601, Grant 62176195, Grant 62176198, Grant U22A2096, Grant U21A20514; in part by the Key Research and Development Program of Shaanxi under Grant 2024GX-ZDCYL-02-10; in part by Shaanxi Outstanding Youth Science Fund Project under Grant 2023-JC-JQ-53; in part by the Shaanxi Province Core Technology Research and Development Project under Grant 2024QY2-GJHX-11; in part by the Fundamental Research Funds for the Central Universities under Grant QTZX24080 and Grant QTZX23042.

## References


[1] Ancong Wu, Wei-Shi Zheng, Hong-Xing Yu, Shaogang Gong, and Jianhuang Lai. Rgb-infrared cross-modality person re-identification. In *ICCV*, pages 5380–5389, 2017.

[2] Qiong Wu, Pingyang Dai, Jie Chen, Chia-Wen Lin, Yongjian Wu, Feiyue Huang, Bineng Zhong, and Rongrong Ji. Discover cross-modality nuances for visible-infrared person re-identification. In *CVPR*, pages 4330–4339, 2021.

[3] Jiangming Shi, Yachao Zhang, Xiangbo Yin, Yuan Xie, Zhizhong Zhang, Jianping Fan, Zhongchao Shi, and Yanyun Qu. Dual pseudo-labels interactive self-training for semi-supervised visible-infrared person re-identification. In *ICCV*, pages 11218–11228, 2023.

[4] Yinpeng Dong, Hang Su, Baoyuan Wu, Zhifeng Li, Wei Liu, Tong Zhang, and Jun Zhu. Efficient decision-based black-box adversarial attacks on face recognition. In *CVPR*, pages 7714–7722, 2019.

[5] Yaoyao Zhong and Weihong Deng. Towards transferable adversarial attack against deep face recognition. *IEEE Transactions on Information Forensics and Security*, 16:1452–1466, 2020.

[6] Dong Yin, Raphael Gontijo Lopes, Jon Shlens, Ekin Dogus Cubuk, and Justin Gilmer. A fourier perspective on model robustness in computer vision. In *NeurIPS*, 2019.

[7] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[8] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *TCSP*, pages 39–57, 2017.

[9] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. In *NeurIPS*, 2019.

[10] Minghui Chen, Zhiqiang Wang, and Feng Zheng. Benchmarks for corruption invariant person re-identification. In *NeurIPS*, 2021.

[11] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *CVPR*, pages 1765–1773, 2017.

[12] Cuiqun Chen, Mang Ye, Meibin Qi, Jingjing Wu, Jianguo Jiang, and Chia-Wen Lin. Structure-aware positional transformer for visible-infrared person re-identification. *IEEE Transactions on Image Processing*, 31:2352–2364, 2022.

[13] Yunhao Du, Cheng Lei, Zhicheng Zhao, Yuan Dong, and Fei Su. Video-based visible-infrared person re-identification with auxiliary samples. *IEEE Transactions on Information Forensics and Security*, 19:1313–1325, 2023.

[14] Hao Yu, Xu Cheng, Wei Peng, Weihao Liu, and Guoying Zhao. Modality unifying network for visible-infrared person re-identification. In *ICCV*, pages 11185–11195, 2023.

[15] Yukang Zhang, Yang Lu, Yan Yan, Hanzi Wang, and Xuelong Li. Frequency domain nuances mining for visible-infrared person re-identification. *arXiv preprint arXiv:2401.02162*, 2024.

[16] Mang Ye, Xiangyuan Lan, Jiawei Li, and Pong Yuen. Hierarchical discriminative learning for visible thermal person re-identification. In *AAAI*, 2018.

[17] Mang Ye, Jianbing Shen, David J. Crandall, Ling Shao, and Jiebo Luo. Dynamic dual-attentive aggregation learning for visible-infrared person re-identification. In *ECCV*, pages 229–247, 2020.

[18] Yuanxin Zhu, Zhao Yang, Li Wang, Sai Zhao, Xiao Hu, and Dapeng Tao. Hetero-center loss for cross-modality person re-identification. *Neurocomputing*, 386:97–109, 2020.

[19] Haijun Liu, Xiaoheng Tan, and Xichuan Zhou. Parameter sharing exploration and hetero-center triplet loss for visible-thermal person re-identification. *IEEE Transactions on Multimedia*, 23:4414–4425, 2020.

[20] Mang Ye, Zheng Wang, Xiangyuan Lan, and Pong C Yuen. Visible thermal person re-identification via dual-constrained top-ranking. In *IJCAI*, page 2, 2018.

[21] Mang Ye, Xiangyuan Lan, Zheng Wang, and Pong C Yuen. Bi-directional center-constrained top-ranking for visible thermal person re-identification. *IEEE Transactions on Information Forensics and Security*, 15:407–419, 2019.

[22] Chaoyou Fu, Yibo Hu, Xiang Wu, Hailin Shi, Tao Mei, and Ran He. Cm-nas: Cross-modality neural architecture search for visible-infrared person re-identification. In *ICCV*, pages 11823–11832, 2021.

[23] Yehansen Chen, Lin Wan, Zhihang Li, Qianyan Jing, and Zongyuan Sun. Neural feature search for rgb-infrared person re-identification. In *CVPR*, pages 587–597, 2021.

[24] Yongguo Ling, Zhun Zhong, Zhiming Luo, Paolo Rota, Shaozi Li, and Nicu Sebe. Class-aware modality mix and center-guided metric learning for visible-thermal person re-identification. In *ACM MM*, pages 889–897, 2020.

[25] Yulin Li, Tianzhu Zhang, and Yongdong Zhang. Frequency domain modality-invariant feature learning for visible-infrared person re-identification. *arXiv preprint arXiv:2401.01839*, 2024.

[26] 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.

[27] Mo Zhou, Zhenxing Niu, Le Wang, Qilin Zhang, and Gang Hua. Adversarial ranking attack and defense. In *ECCV*, pages 781–799, 2020.

[28] Zhedong Zheng, Liang Zheng, Yi Yang, and Fei Wu. Query attack via opposite-direction feature: Towards robust image retrieval. *arXiv preprint arXiv:1809.02681*, 2018.

[29] Fengxiang Yang, Zhun Zhong, Hong Liu, Zheng Wang, Zhiming Luo, Shaozi Li, Nicu Sebe, and Shin'ichi Satoh. Learning to attack real-world models for person re-identification via virtual-guided meta-learning. In *AAAI*, pages 3128–3135, 2021.

[30] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[31] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[32] Yunpeng Gong, Liqing Huang, and Lifei Chen. Person re-identification method based on color attack and joint defence. In *CVPR*, pages 4313–4322, 2022.

[33] 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*, pages 342–351, 2020.

[34] Uche M Osahor and Nasser M Nasrabadi. Deep adversarial attack on target detection systems. In *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications*, pages 620–628, 2019.

[35] Xiaopei Zhu, Xiao Li, Jianmin Li, Zheyao Wang, and Xiaolin Hu. Fooling thermal infrared pedestrian detectors in real world using small bulbs. In *AAAI*, pages 3616–3624, 2021.

[36] Xingxing Wei, Jie Yu, and Yao Huang. Physically adversarial infrared patches with learnable shapes and locations. In *CVPR*, pages 12334–12342, 2023.

[37] Xingxing Wei, Yao Huang, Yitong Sun, and Jie Yu. Unified adversarial patch for cross-modal attacks in the physical world. In *ICCV*, pages 4445–4454, 2023.

[38] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44:2872–2893, 2021.

[39] Quentin Bouniot, Romaric Audigier, and Angelique Loesch. Vulnerability of person re-identification models to metric adversarial attacks. In *CVPR*, pages 794–795, 2020.

[40] Mang Ye, Weijian Ruan, Bo Du, and Mike Zheng Shou. Channel augmented joint learning for visible-infrared recognition. In *ICCV*, pages 13567–13576, 2021.

[41] Yukang Zhang, Yan Yan, Yang Lu, and Hanzi Wang. Towards a unified middle modality learning for visible-infrared person re-identification. In *ACM MM*, pages 788–796, 2021.

[42] Yukang Zhang and Hanzi Wang. Diverse embedding expansion network and low-light cross-modality benchmark for visible-infrared person re-identification. In *CVPR*, pages 2153–2162, 2023.

[43] Dat Tien Nguyen, Hyung Gil Hong, Ki Wan Kim, and Kang Ryoung Park. Person recognition system based on a combination of body images from visible light and thermal cameras. *Sensors*, 17:605, 2017.

[44] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *CVPR*, pages 7132–7141, 2018.

[45] Qilong Wang, Banggu Wu, Pengfei Zhu, Peihua Li, Wangmeng Zuo, and Qinghua Hu. Eca-net: Efficient channel attention for deep convolutional neural networks. In *CVPR*, pages 11534–11542, 2020.

## 7 NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract succinctly summarizes the core contributions and scope of the paper, while the introduction elaborates on the research background, problem definition, proposed methods, and main results obtained.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [No]

   Justification: Our approach was tested on only a few datasets and systems, demonstrating the feasibility of the theory without exploring its application in real-world scenarios.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: Detailed theorems and formulas are provided in Section 3, with the theorems relied upon being properly cited.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We contribute a novel architecture, and the entire process is detailed in Section 3 through images and formulas.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: We will consider making the code details publicly available in the future.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide a detailed explanation in Section 4

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Our experimental results do not include error bars, confidence intervals, or tests for statistical significance.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We present the relevant computational information in Section 4.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have reviewed the relevant requirements and ensured compliance with ethical guidelines.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: While our method can be applied for robustness testing of models, it also implies that certain groups might use this method to attack systems, resulting in potentially destructive impacts. These are considerations that need to be taken into account in practical deployments.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.

- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: We use publicly available datasets and do not involve scraping data from the internet.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We ensure that all assets used are properly credited and adhere to the relevant licenses and terms of use.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: This paper does not introduce any new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [No]

    Justification: The pedestrian dataset used in this paper is publicly available, and it is stated by the publisher that each pedestrian captured in the dataset has signed a privacy release allowing the use of the images for scientific research and display in research papers.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [No]

    Justification: The pedestrian dataset used in this paper is publicly available, and it is stated by the publisher that each pedestrian captured in the dataset has signed a privacy release allowing the use of the images for scientific research and display in research papers.

    Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.