A^{2}RNet: Adversarial Attack Resilient Network for Robust Infrared and Visible Image Fusion

Jiawei Li[1]*, Hongwei Yu[1]*, Jiansheng Chen[1]†, Xinlong Ding[1], Jinlong Wang[1], Jinyuan Liu[2], Bochao Zou[1], Huimin Ma[1]

[1]School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing, China
[2]School of Software Technology, Dalian University of Technology, Dalian, China
ljw19970218@163.com, yuhongwei22@xs.ustb.edu.cn, jschen@ustb.edu.cn, {dingxl22, M202320974}@xs.ustb.edu.cn, atlantis918@hotmail.com, {zoubochao, mhmpub}@ustb.edu.cn

Abstract

Infrared and visible image fusion (IVIF) is a crucial technique for enhancing visual performance by integrating unique information from different modalities into one fused image. Exiting methods pay more attention to conducting fusion with undisturbed data, while overlooking the impact of deliberate interference on the effectiveness of fusion results. To investigate the robustness of fusion models, in this paper, we propose a novel adversarial attack resilient network, called A^{2}RNet. Specifically, we develop an adversarial paradigm with an anti-attack loss function to implement adversarial attacks and training. It is constructed based on the intrinsic nature of IVIF and provide a robust foundation for future research advancements. We adopt a Unet as the pipeline with a transformer-based defensive refinement module (DRM) under this paradigm, which guarantees fused image quality in a robust coarse-to-fine manner. Compared to previous works, our method mitigates the adverse effects of adversarial perturbations, consistently maintaining high-fidelity fusion results. Furthermore, the performance of downstream tasks can also be well maintained under adversarial attacks. Code is available at https://github.com/lok-18/A2RNet.

Introduction

The purpose of infrared and visible image fusion (IVIF) aims to integrate salient information from different sensors for obtaining well-performing fused images, which can alleviate the imaging limitations of a single sensor. The fused image simultaneously contains thermal target information and texture contents from different modalities. In some computer vision tasks, *e.g.*, autonomous driving (Sun et al. 2022) and salient object detection (Wang et al. 2023), IVIF technology is applied to assist in achieving more accurate and detailed results.

The primary challenge of this task is how to effectively extract features from different modalities (Zhang et al. 2021). Early traditional methods employ techniques such as wavelet transforms (Li, Manjunath, and Mitra 1995) and sparse representation (Liu, Liu, and Wang 2015) to perform matrix operations on source images. However, their complex manual adjustment strategies are time-consuming and cumbersome to implement. Recently, deep learning-based methods have gradually replaced traditional approaches (Huang et al. 2022). These methods possess powerful feature extraction capabilities to learn salient information from source images (Li et al. 2022, 2023c; Liu et al. 2023a), which has entered an efficient and rapidly evolving stage.

In general, existing IVIF methods are based on nondestructive data to construct. They only focus on extracting useful information from source images without considering the potential presence of interference (Liu et al. 2024b). In other words, these networks become fragile under adversarial perturbations, which may obtain poor fusion results. Without any robust operations in the network, fused images generated by adversarial examples (AEs) exhibit noticeable artifact regions in Fig. 1. The segmentation accu-

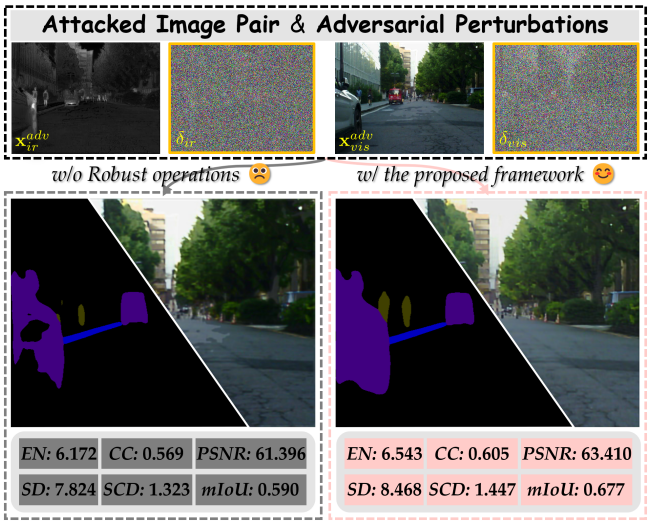

Figure 1: Schematic illustration of different adversarial operations. Clearly, fused images generated by attacked image pairs exhibit superior qualitative, quantitative and downstream task performance when conducting the proposed framework.

*These authors contributed equally.
†Corresponding author.

racy for some categories also deteriorates accordingly. Some researchers (Liu et al. 2023b) have utilized pre-trained segmentation models with adversarial training (AT) to defend fusion networks at a feature-wise level. However, it does not fundamentally design a fusion-oriented paradigm capable of formulating AT. In addition, employing a robust fusion network for resisting perturbations tailored to the IVIF task is also essential for maintaining robustness.

Based on the characteristics of IVIF and existing adversarial research, this paper first develops a novel paradigm with anti-attack loss to achieve AEs and facilitate AT. Then, we propose an adversarial attack resilient network named A^2RNet for accommodating this paradigm. Specifically, U-Net is employed as the pipeline, where its up-/downsampling operations help filter out noise attacks. To prevent U-Net from overlooking essential features, a transformer-based defensive refinement module (DRM) is implemented in the middle of U-Net, aiming to further refine feature learning and avoid the appearance of noise artifacts. Through the aforementioned network architecture and adversarial strategies, we are able to obtain robust fused images that perform well on both clean and adversarial samples. As depicted in Fig. 1, the proposed method performs excellent fused images and segmentation results under perturbations. In summary, the main contributions of this paper are as follows:

- To achieve a highly-robust fusion framework, we propose an adversarial strategy with a novel anti-attack loss to generate adversarial examples and conduct adversarial training. This approach is rooted in the essence of IVIF and advances the development of adversarial robustness in fusion tasks.
- The proposed adversarial attack resilient network (A^2RNet) uses U-Net as the pipeline for robust feature representation, leveraging its structural characteristics to defend against adversarial perturbations.
- Considering that using U-Net may result in texture missing, the defensive refinement module (DRM) is introduced to supplement the extracted information. Furthermore, it enhances the proposed network to resist noise, leading to more refined fusion results.
- Extensive experiments demonstrate that our method exhibits stronger robustness under adversarial perturbations. Meanwhile, it also outperforms other comparative methods in downstream task performance.

Related Works

Infrared and Visible Image Fusion

Thanks to the continuous advancements in deep learning technology, infrared and visible image fusion has gradually transitioned from using traditional methods to employing various network architectures(Ma, Ma, and Li 2019), *e.g.*, CNN (Cao et al. 2023), Transformer (Rao, Xu, and Wu 2023), GNN (Li et al. 2023b) and Diffusion (Yue et al. 2023). They can effectively extract features from source images and fuse them together, avoiding the hassle of manual adjustment strategies inherent in traditional methods. As a representation, (Li and Wu 2018) proposed an Auto-Encoder-based fusion network with DenseNet (Huang et al. 2017), which has been widely adopted in subsequent methods. (Liu et al. 2021) proposed a network structure based on a modified GAN, aiming to address the instability in GAN training while retaining the discriminator to enhance the fidelity of the generated results. In the fusion task, using Transformers to build models allows for a greater focus on global features (Ma et al. 2022). Recently, the surge in the popularity of generative models has led researchers to incorporate them into IVIF tasks (Zhao et al. 2023). In addition, some key priors or information that help improve fusion results have also been incorporated into networks to assist in feature learning. For instance, (Zhao et al. 2024b) leveraged large language models, such as GPT3 (Brown et al. 2020), to provide detailed descriptions of source images. Text-IF (Yi et al. 2024) also employed a text-guided architecture to construct the fusion network.

Adversarial Attack and Defence

By adding perturbations to input samples, adversarial attacks aim to mislead deep neural networks (DNN) for producing incorrect outputs. These perturbations are typically difficult for the human visual system (HVS) to detect. As a representative, Fast Gradient Sign Method (FGSM) was proposed by (Goodfellow, Shlens, and Szegedy 2014). The process of FGSM can be quantified as:

$$\delta = \epsilon \cdot \operatorname{sign}\left(\bigtriangledown_{\mathbf{x}} \mathcal{L}\left(f\left(\mathbf{x};\theta\right),\mathbf{y}\right)\right), \tag{1}$$

where $\mathbf{x}$, $\mathbf{y}$ and $f\left(\cdot;\theta\right)$ mean a clean input, its ground truth (GT) and a DNN, respectively. $\mathcal{L}$ is the loss function. $\bigtriangledown$ and $\operatorname{sign}(\cdot)$ represent gradients and their direction. ϵ denote the magnitude of applied perturbation under the specified constraint. With a small step size α for a fixed number of gradient iterations, (Madry et al. 2017) proposed a multi-step optimization variant of FGSM called Projected Gradient Descent (PGD). Similarly, it can be defined as:

$$\begin{aligned}\delta_{k+1} = \delta_k + \alpha \cdot \operatorname{sign}\left(\bigtriangledown_{\mathbf{x}+\delta_k} \mathcal{L}\left(f\left(\mathbf{x}+\delta_k;\theta\right),\mathbf{y}\right)\right)\\ \text{s.t.}||\delta||_p \leq \epsilon.\end{aligned} \tag{2}$$

This restriction ensures that α remains within ϵ around $\mathbf{x} + \delta_k$. p represents the norm type. In recent years, an increasing number of adversarial attack methods based on PGD have been proposed in different fields (Ding et al. 2024; Yu et al. 2024).

In adversarial defence, (Yu et al. 2022) employed an enumeration method to conduct robustness analysis on popular models and loss functions in the deraining task, and combined them into a more robust architecture. As a direct way to improve adversarial robustness, AT involves feeding both clean samples and AEs into the network during training (Shafahi et al. 2020; Gokhale et al. 2021; Jia et al. 2022). Researchers have quantified the process of AT as a min-max optimization problem (Madry et al. 2017):

$$\min_{\theta} \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}\left[\max_{||\delta||_p\leq\epsilon} \mathcal{L}\left(f(\mathbf{x}+\delta;\theta),\mathbf{y}\right)\right], \tag{3}$$

where $\mathcal{D}$ denotes the data distribution. For instance, (Jiang et al. 2024) proposed a robust image stitching algorithm with

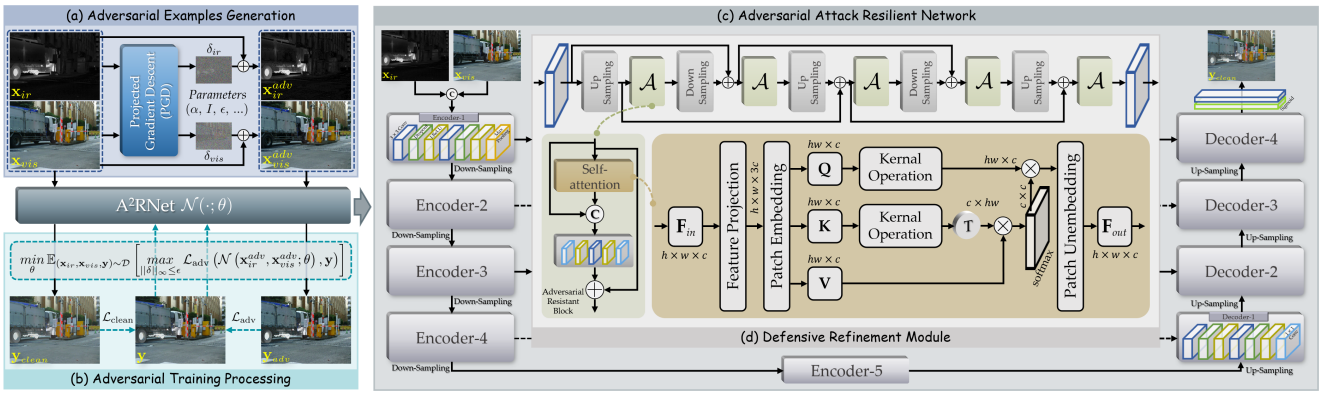

Figure 2: Framework of the proposed A^{2}RNet. In specific, (a) and (b) represent the adversarial examples generation and adversarial training processing, respectively. (c) is the adversarial attack resilient network, which contains the defensive refinement module (DRM) as shown in (d).

adaptive AT, which enables to resist adversarial attacks and achieve better stitching results. PAIF (Liu et al. 2023b) was introudced to conduct IVIF for more robust semantic segmentation. It is the first work related to adversarial robustness in the field of image fusion. Unlike PAIF, our proposed method focuses on the robustness of fused images, then ameliorating the performance of downstream tasks.

Proposed Method

Overview

For the IVIF task, adversarial attacks seek to disrupt the original representation of source images, causing undesired artifacts and halos in fusion results. To achieve more stable and robust fusion images, we formulate the adversarial attack and training process in IVIF. Note that it is designed based on existing adversarial robustness and the characteristics of IVIF, and exclusively targets the fusion stage. With this formulation, we propose a novel network and a loss function called adversarial attack resilient network (A^{2}RNet) and anti-attack loss ($\mathcal{L}_{\mathrm{a}}$) resepctively to facilitate more robust feature learning.

We define A^{2}RNet as $\mathcal{N}(\cdot;\theta)$ parameterized with θ. First, we need to generate AEs with PGD for AT. Different from traditional tasks, the IVIF task involves dual inputs and has no real GT for reference. Therefore, we introduce fair pseudo-labels in conjunction with $\mathcal{L}_{\mathrm{a}}$ to guide the adversarial attacks and training. The inclusion of pseudo-labels ensures that the entire adversarial process is more rational and targeted. After generating the AEs, both clean and attacked results are fed into $\mathcal{N}(\cdot;\theta)$ simultaneously for training. It is worth noting that our proposed network excels well during AT, which can produce well-performed fusion results under attacks. The specific procedure of the proposed framework is illustrated in Fig. 2.

Adversarial Attacks and Training in IVIF

In the initial stage, given a clean image pair $(\mathbf{x}_{ir}, \mathbf{x}_{vis})$ to generate the corresponding AE pair $(\mathbf{x}_{ir}^{adv}, \mathbf{x}_{vis}^{adv}) \leftarrow (\mathbf{x}_{ir} + \delta_{ir}, \mathbf{x}_{vis} + \delta_{vis})$ as illustrated in Fig. 2(a). We employ PGD as the AE generator, and its attack process is formulated according to Eq. 2 (taking attacks on infrared images as an example, the generation of visible perturbation δ_{vis}^{k+1} is similar to δ_{ir}^{k+1}) :

$$\begin{aligned}\delta_{ir}^{k+1} = \delta_{ir}^{k} + \alpha \cdot \mathrm{sign}\Big(\nabla_{\mathbf{x}_{ir}+\delta_{ir}^{k}} \mathcal{L}_{\mathrm{adv}}\big(\mathcal{N}\big((\mathbf{x}_{ir}+ \\ \delta_{ir}^{k}, \mathbf{x}_{vis} + \delta_{vis}^{k}\big);\theta\big), \mathbf{y}\big)\Big) \quad \mathrm{s.t.}||\delta||_{\infty} \leq \epsilon,\end{aligned} \tag{4}$$

where $\mathcal{L}_{\mathrm{adv}}$ denotes a part of $\mathcal{L}_{\mathrm{a}}$, and $\mathbf{y}$ represents the introduced pseudo-label. l_{∞}-norm is chosen to constrain ϵ. In PGD, we conduct $\mathcal{L}_{\mathrm{adv}}$ for gradient backpropagation and employ pseudo-labels to address the challenge of achieving AEs without GT in IVIF. To ensure relative fairness, we use a common CNN and loss functions, *e.g.*, l_1 and SSIM loss, for training and inference to obtain the labels. More details can be found in the supplementary material. Note that not directly using the results of other SOTA methods as pseudo-labels is intended to prevent overfitting. In addition, it can also avoid biasing fusion results towards any particular SOTA method. Therefore, applying this "moderate" pseudo-labels to the entire process is relatively fair.

To obtain a robust fusion model against attacks, we redesign the AT process based on the characteristics of the IVIF task. Compared to existing methods (Liu et al. 2023b), we focus more on the robustness of fusion, which can first enhance the performance of fused results, then improve the outcomes of downstream tasks. Similarly, Eq. 3 is reformulated as:

$$\min_{\theta} \mathbb{E}_{(\mathbf{x}_{ir},\mathbf{x}_{vis},\mathbf{y})\sim\mathcal{D}} \left[\max_{||\delta||_{\infty}\leq\epsilon} \mathcal{L}_{\mathrm{adv}}\left(\mathcal{N}\left(\mathbf{x}_{ir}^{adv}, \mathbf{x}_{vis}^{adv};\theta\right), \mathbf{y}\right)\right] \tag{5}$$

where $\mathbf{y}$ is also incorporated into AT. As depicted in Fig. 2(b), clean $(\mathbf{x}_{ir}, \mathbf{x}_{vis})$ and adversarial examples $(\mathbf{x}_{ir}^{adv}, \mathbf{x}_{vis}^{adv})$ are fed into $\mathcal{N}(\cdot;\theta)$ to achieve corresponding results separately, *i.e.*, $\mathbf{y}_{clean}$ and $\mathbf{y}_{adv}$. Subsequently, we calculate the loss value with $\mathcal{L}_{\mathrm{clean}}$ and $\mathcal{L}_{\mathrm{adv}}$, which are backpropagated features through the fusion network $\mathcal{N}(\cdot;\theta)$. The distribution of clean and adversarial examples

is ensured to be balanced with this setting, resulting in preventing the robustness bias.

For the unsupervised IVIF task, generating AEs and performing AT is challenging, so that we employ pseudo-labels to construct a "supervision" manner. It not only facilitates the effective generation of AEs for AT, but also enhances robustness while ensuring the quality of fusion results. Specifically, the mean squared error (MSE) loss is introduced to estimate the error magnitude between fusion results and labels. Meanwhile, the structural similarity index (SSIM) loss (Wang et al. 2004) is used to measure the similarity between them. The basic form of the training loss can be quantified as:

$$\mathcal{L} = \beta \mathcal{L}_{\mathrm{MSE}}\left(\mathbf{y}', \mathbf{y}\right) + \gamma\left(1 - \mathcal{L}_{\mathrm{SSIM}}\left(\mathbf{y}', \mathbf{y}\right)\right), \quad (6)$$

where β and γ are hyperparameters. $\mathbf{y}'$ means fusion results. In PGD, we obtain $(\mathbf{x}_{ir}^{adv}, \mathbf{x}_{vis}^{adv})$ by computing $\mathcal{L}_{\mathrm{adv}}$ with $\mathbf{y}_{\mathrm{adv}}$. Note that $\mathcal{L}_{\mathrm{adv}}$ is derived by replacing $\mathbf{y}'$ with $\mathbf{y}_{\mathrm{adv}}$. In the stage of AT, $\mathcal{L}_{\mathrm{adv}}$ is used to backpropagate features learned from AEs. Hence, the anti-attack loss for the entire architecture can be expressed as:

$$\mathcal{L}_{\mathrm{a}} = \mathcal{L}_{\mathrm{clean}} + \mathcal{L}_{\mathrm{adv}}, \quad (7)$$

where $\mathcal{L}_{\mathrm{clean}}$ denotes the loss incurred during normal training with clean samples.

Adversarial Attack Resilient Network

As essential as adversarial strategies, the design of fusion networks also determines the robustness of fused images. Inspired by techniques such as image restoration (Ma et al. 2023; Zheng and Wu 2024), Unet-based networks exhibit the excellent capability of feature decoupling. Therefore, we select Unet as the pipeline for our A^{2}RNet. However, Unet often experiences some information missing during the decoupling process, which may lead to undesirable artifacts in fusion results. Thanks to the versatility of Unet, we can incorporate flexible modules within it to prevent the aforementioned issues. In short, A^{2}RNet is employed to conduct robust feature learning for the IVIF task.

The details of our proposed network are presented in Fig. 2(c). In specific, we construct our pipeline by referencing the classical Unet network (Ronneberger, Fischer, and Brox 2015). Unlike the common Unet, certain parts of the encoder and decoder, *e.g.*, up/downsampling operation are fine-tuned for better feature extraction. To prevent from missing important details, we propose the defensive refinement module (DRM) in the middle of our Unet pipeline. Considering trade-off, DRM is only connected from Encoder-1/3 to Decoder-4/2. The connections from Encoder-2/4 to Decoder-3/1 conduct the conventional "copy & crop" operation.

DRM contains five adversarial resistant blocks (ARBs), which leverage features extracted from the Unet for more robust self-attention learning. Multiple sampling and residual operations can also help keep effective contents from source images and filter out attack perturbations during the AT stage. Note that PixelShuffle/Unshuffle (Shi et al. 2016) is employed here for up/downsampling. As shown in Fig. 2(d),

Algorithm 1: Adversarial training in A^{2}RNet

Require: dataset $(\mathbf{x}_{ir}, \mathbf{x}_{vis}) \sim \mathcal{D}$, pseudo-labels $\mathbf{y}$, total epoch T, network parameters θ

for epoch from 1 to T **do**
  for minibatch $b = 4$ **do**
    % AEs generation with Eq. 4
    for $iteration$ form 1 to I **do**
      $(\delta_{ir}^{i}, \delta_{vis}^{i}) \leftarrow \mathrm{PGD}(\alpha, \epsilon, I, \mathbf{x}_{ir}, \mathbf{x}_{vis}, \mathbf{y})$
      Update $(\delta_{ir}, \delta_{vis}) \leftarrow (\delta_{ir}^{i}, \delta_{vis}^{i})$
    end for
    Generate $(\mathbf{x}_{ir}^{adv}, \mathbf{x}_{vis}^{adv}) \leftarrow (\mathbf{x}_{ir} + \delta_{ir}, \mathbf{x}_{vis} + \delta_{vis})$
    % AT in $\mathcal{N}$ with Eq. 5
    Compute $\mathcal{L}_{\mathrm{clean}}$ on $(\mathbf{x}_{ir}, \mathbf{x}_{vis})$ with θ
    Compute $\mathcal{L}_{\mathrm{adv}}$ on $(\mathbf{x}_{ir}^{adv}, \mathbf{x}_{vis}^{adv})$ with θ
    Update $\theta \leftarrow \mathcal{L}_{\mathrm{a}} = \mathcal{L}_{\mathrm{clean}} + \mathcal{L}_{\mathrm{adv}}$
  end for
end for

the ARB is illustrated. Similar to the typical transformer architecture, it is also composed of a self-attention layer and a feed-forward layer (including 1×1 Conv, 3×3 Conv and LeakyReLU layers). In the self-attention layer, given an input feature $\mathbf{F}_{in} \in \mathbb{R}^{h \times w \times c}$ with height, width and channel dimensions of h, w and c, feature projection first transforms it into $\mathbf{F}_{in} \in \mathbb{R}^{h \times w \times 3c}$. Next, $\{\mathbf{Q}, \mathbf{K}, \mathbf{V}\} \in \mathbb{R}^{hw \times c}$ are obtained through patch embedding. We utilize Mercer's theorem (Mercer 1909) to contruct a Mercer-based kernel operation for robust feature representation, which reconstructs $\mathbf{Q}$ and $\mathbf{K}$ through the corresponding projection mapping. The Pearson correlation coefficient (Cohen et al. 2009) is introduced to measure the correlation r between $\mathbf{Q}$ and $\mathbf{K}$, and to validate the kernel operation $m(\cdot)$. It can be defined through Taylor expansion as follow:

$$K\left(\mathbf{Q}, \mathbf{K}\right) = \sum_{i=0}^{\infty} \frac{\left(\mathbf{Q} - \bar{\mathbf{Q}}\right)^{2i}}{\sigma^{\frac{1}{2}i}\sqrt{i!}} \frac{\left(\mathbf{K} - \bar{\mathbf{K}}\right)^{2i}}{\sigma^{\frac{1}{2}i}\sqrt{i!}}, \quad (8)$$

where $\bar{\mathbf{Q}}$ and $\bar{\mathbf{K}}$ represent the means of $\mathbf{Q}$ and $\mathbf{K}$, respectively. The mapping function is expressed as (taking $\mathbf{Q}$ as an example):

$$m(\mathbf{Q}) = (1, \frac{(\mathbf{Q} - \bar{\mathbf{Q}})^2}{\sigma^{\frac{1}{2}}}, \frac{(\mathbf{Q} - \bar{\mathbf{Q}})^4}{\sigma^{\frac{1}{2}}}, \cdots, \frac{(\mathbf{Q} - \bar{\mathbf{Q}})^{2i}}{\sigma^{\frac{1}{2}}}). \quad (9)$$

The kernel operation enable to improve the robust representation of self-attention, building a resilient fusion model at the feature-wised level. Moreover, we alter the multiplication order in self-attention by first multiplying $\mathbf{K}$ and $\mathbf{V}$, which reduces the complexity from $O(N^2)$ to $O(N)$ and promotes the efficiency of adversarial training. Therefore, the self-attention matrix is computed as:

$$\mathrm{Att}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathrm{softmax}\left(\frac{\mathbf{K} \cdot m(\mathbf{V})^{\mathrm{T}}}{\sqrt{d_s}}\right) m(\mathbf{Q}), \quad (10)$$

where d_s and T are the scaling factor and transpose operation. Finally, $\mathbf{F}_{out} \in \mathbb{R}^{h \times w \times c}$ is achieved through patch unembedding and fed into Decoder-4/2. We detail the entire process in Algorithm.1, including the steps for adversarial example generation, robust feature learning, and adversarial training.

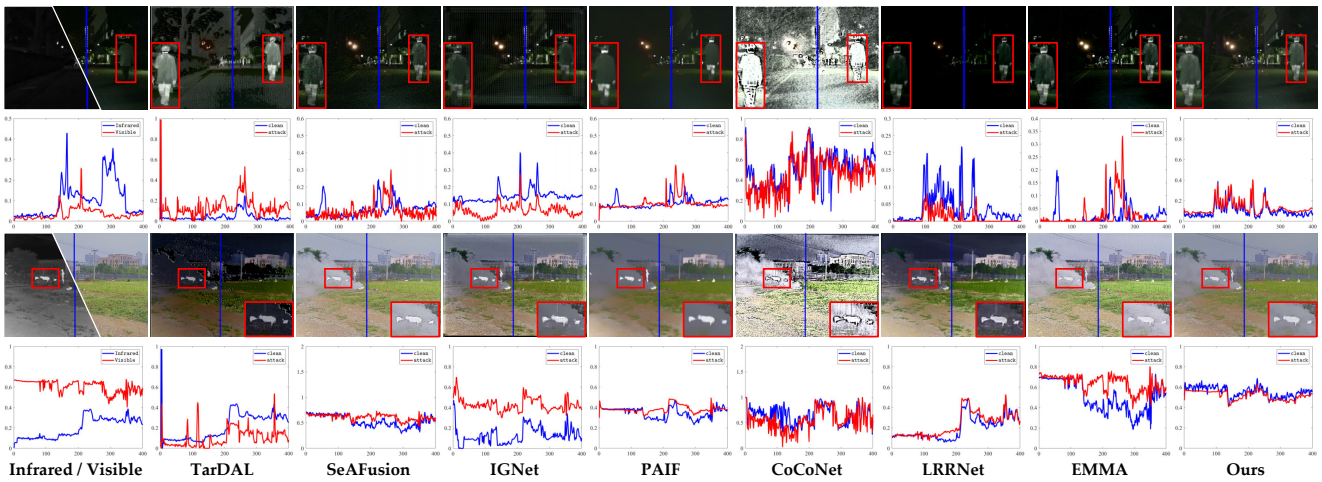

Figure 3: Fusion comparisons with SOTA methods in MFNet and M^{3}FD datasets. We apply PGD to clean samples and add perturbations with $\epsilon = 4/255$ to generate AEs. The signal maps are also provided for clean and attack states. The closer the waveform, the stronger the robustness.

Experiments

Experimental Setup Details

Before adversarial training, we first need to generate adversarial examples by using PGD. Specifically, a moderate perturbation is set with a total iteration I of 3, a perturbation strength ϵ of 4/255, and a step size α of 1/255. This configuration helps to avoid excessive time spent on getting adversarial examples. l_∞-norm constrains perturbations. During the adversarial training phase, the Adam optimizer is chosen to adjust θ with a 0.001 learning rate. We set the batch size and total epochs to 4 and 50, respectively. In $\mathcal{L}$, β and γ are 100. The number of clean and adversarial examples is kept at 1:1 for balance. M^{3}FD (Liu et al. 2022) and MFNet (Ha et al. 2017) datasets are introduced for training and testing. In the adversarial inference stage, we set I to 20 with unchanged ϵ and α to generate AEs and use them to obtain fused images. It is noticed that all experiments are conducted on an Intel(R) Xeon(R) Gold 6271C CPU and a NVIDIA Tesla A100 GPU with PyTorch.

To demonstrate the superiority of our method, we conduct comparisons in both qualitative and quantitative results. Seven SOTA methods are selected for comprehensive comparison, including TarDAL (Liu et al. 2022), SeAFusion (Tang, Yuan, and Ma 2022), IGNet (Li et al. 2023b), PAIF (Liu et al. 2023b), CoCoNet (Liu et al. 2024a), LRRNet (Li et al. 2023a) and EMMA (Zhao et al. 2024a). Except for PAIF, none of the other methods have investigated adversarial robustness. To ensure fairness, we apply the same adversarial settings as our method to the open-source code of these approaches. In the quantitative comparison, we choose Entropy (EN), Standard Deviation (SD), Peak Signal-to-Noise Ratio (PSNR), Correlation Coefficient (CC) (Shah, Merchant, and Desai 2013) and the Sum of the Correlations of Differences (SCD) (Aslantas and Bendes 2015). Higher values indicate better image performance. In the comparison of downstream tasks, mean average precision (mAP@.5) and mean intersection over union (mIoU) are used to evaluate detection and segmentation, respectively. The experimental details of the downstream tasks are provided in the supplementary materials.

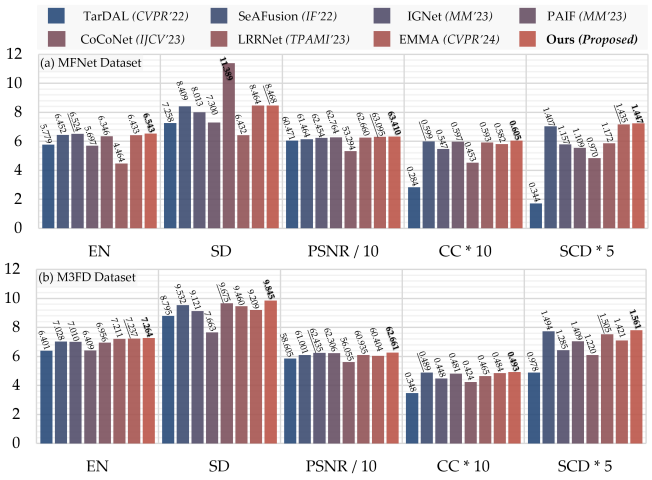

Figure 4: Bar charts of the fusion comparison metrics. For better visualization, we have scaled the values of certain metrics.

Comparison Results

Comparison of Fusion Results Fig. 3 presents the qualitative comparison results of our method and other SOTA approaches under adversarial attacks. The performance of TarDAL and CoCoNet is noticeably inconsistent with HVS, exhibiting evident attacked regions such as noisy spots on the ground (first set) and color distortions on the grass (second set). SeAFusion, IGNet and EMMA all exhibit varying degrees of noisy textures, which significantly impact the visual quality. For instance, in the magnified patch of the

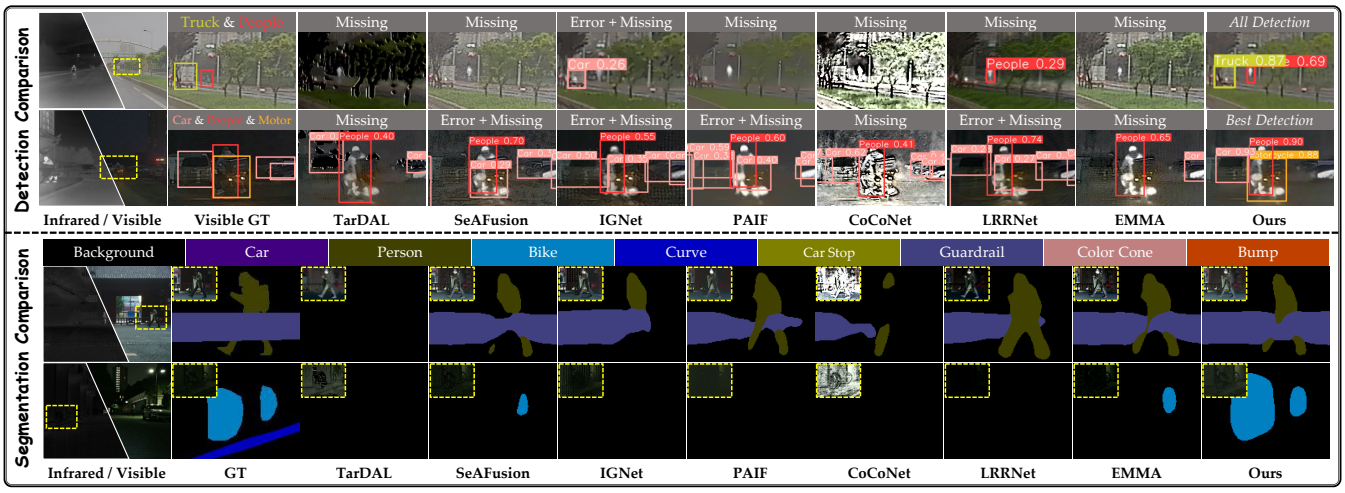

Figure 5: Detection and segmentation comparisons of fused images. Under adversarial conditions, our method yields better performance in downstream tasks.

Metric	Method							
	TarDAl	SeAFusion	IGNet	PAIF	CoCoNet	LRRNet	EMMA	Ours
mAP@.5	$0.462_{\downarrow 0.161}$	$0.732_{\downarrow 0.042}$	$0.557_{\downarrow 0.154}$	$0.715_{\downarrow 0.046}$	$0.311_{\downarrow 0.093}$	$0.704_{\downarrow 0.032}$	$0.696_{\downarrow 0.038}$	$0.781_{\downarrow 0.024}$
mIoU	$0.415_{\downarrow 0.069}$	$0.618_{\downarrow 0.078}$	$0.514_{\downarrow 0.184}$	$0.578_{\downarrow 0.134}$	$0.409_{\downarrow 0.080}$	$0.623_{\downarrow 0.027}$	$0.649_{\downarrow 0.086}$	$0.677_{\downarrow 0.068}$

Table 1: Quantitative results of detection (mAP@.5) and segmentation (mIoU). Red and blue denote the optimal and suboptimal results, respectively. The subscripts indicate the change compared adversarial conditions with the clean.

person in the first set, the details on the clothing are not as smooth as the clean image. Although LRRNet does not contain excessive noise, it compromises on brightness. In the first set of results, the details on the road are not clearly observed, meanwhile in the second set, the sky appears overly dark, which does not align with realistic natural scenes. It indicates that adversarial examples disrupt the balance of information extraction in the original network. In addition, we also provide signal maps for each method under different inference conditions, *i.e.*, clean and attack. Since PAIF incorporates robustness operations, it can withstand certain levels of attacks. However, it exhibits an over-smooth phenomenon or even blurry texture, which is an undesirable outcome. Thanks to the effective adversarial training strategy and network architecture in our method, we achieve more stable and robust fusion results. They not only capture the desired detailed features but also show minimal differences compared to the clean results. As shown in Fig. 4, we present the quantitative comparison results with bar charts. It can also prove that our method obtains superior results.

Comparison of Detection Results The fusion results obtained from AEs also lead to some changes in downstream tasks. In the detection task, we provide two sets of enlarged patches on the M^3FD dataset as shown in Fig. 5. In the first set, the patch contains a truck and a person. Due to the poor visual quality of TarDAL and CoCoNet, the detector fails to identify any targets. SeAFusion and EMMA are affected by noisy spots, misleading the detector into making incorrect judgments. As the edge features of each target are not distinct in PAIF, the detection network is unable to capture the necessary information for detection. The results from IGNet and LRRNet also exhibit some errors or missing detections with low confidence. However, the proposed method can accurately detect all targets with high confidence. There are more objects to detect in second comparisons. Compared to other methods, we obtain the best detection results. In the quantitative comparison, Table. 1 presents the mAP@.5 for all methods. The subscripts indicate the difference between clean and attack detection results. From the subscripts of mAP@.5, it can be seen that our results not only achieve the highest scores but also maintain the smallest difference compared to the clean. The specific AP@.5 values for each category are provided in the supplementary materials.

Comparison of Segmentation Results Similarly, we present the qualitative and quantitative segmentation results on the MFNet dataset in Fig. 5 and Table. 1, respectively. The detailed AP@.5 values for each category are presented in the supplementary materials. In the daytime scenes, fused images with perturbations exhibit inaccurate regions in the segmentation results. For instance, SeAFusion, IGNet and CoCoNet perform poorly on the "Person" category. LRRNet shows undesirable results at the boundary between "Guardrail" and "Person", struggling to accurately differentiate foreground information. Although EMMA achieves

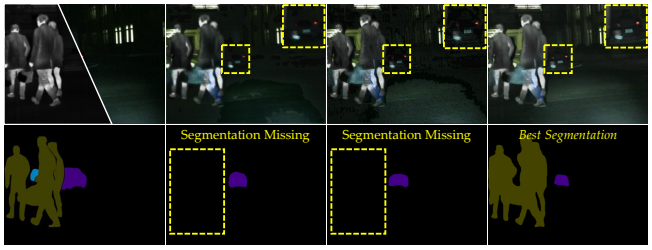

Figure 6: Ablation analysis in different modules.

Model	$\mathcal{U}$	DRM	Dataset:MFNet					
			EN	SD	PSNR	CC	SCD	mIoU
M1	✗	✔	5.231	6.942	60.673	0.593	1.403	0.594
M2	✔	✗	5.742	7.586	63.157	0.587	1.325	0.669
M3	✔	✔	6.543	8.468	63.410	0.605	1.447	0.745

Table 2: Quantitative ablation results of different modules.

relatively good segmentation results, it still falls short compared to ours. It can be substantiated by both the mIoU values and ground truth. In the nighttime scenes, only SeAFusion and EMMA are able to segment parts of the "Bike" category. TarDAL fails to depict any meaningful information in all results. However, our method achieves superior visual performance and quantitative metrics compared to all SOTA methods, which proves that our fusion results can maintain a robust state in the segmentation task.

Ablation Analysis

Analysis of Modules In the proposed method, the synergy between the Unet pipeline and DRM is a key reason for maintaining the robustness of network. We conduct ablation studies by progressively disabling these two modules. The corresponding qualitative and quantitative comparison results are presented in Fig. 6 and Table. 2. Note that we introduce $\mathcal{U}$ to represent Unet. Without $\mathcal{U}$, undesirable patches appear in the fusion results, most notably on the ground. Omitting DRM may cause prominent noisy areas, particularly at the edges of cars and windows. The segmentation results of them completely fail to capture the "Person" category. Obviously, our results not only avoid artifacts but also prevent the occurrence of objectionable noise. It indicates that using $\mathcal{U}$ allows the architecture to filter out most perturbations. Additionally, DRM can further resist noise attacks and refine feature expressions to achieve robust fusion images. In the segmentation performance, our results are the most similar to the ground truth. The quantitative results in Table. 2 also demonstrate that our architecture plays a positive role in constructing a robust IVIF network.

Analysis of Adversarial Training Apart from the contribution of modules, AT is also significant in enhancing the robustness of our network. We keep the original model and experimental settings unchanged with different training strategies, *i.e.*, AT *vs.* non-AT. When the network conducts without AT and subjected to attacks, we can observe in Fig. 7 that the fusion results exhibit noticeable blurriness. Moreover, not all targets in the scene are detected well. It proves that the robustness of the network without AT is still compromised by perturbations, leading to less-than-ideal fusion quality and performance in downstream tasks. In contrast, we can achieve more robust and stable results with the proposed adversarial strategy designed for the fusion task. More details and targets are able to be observed and detected in our method. The metrics with AT are also significantly higher than those without, which is shown in Fig. 7.

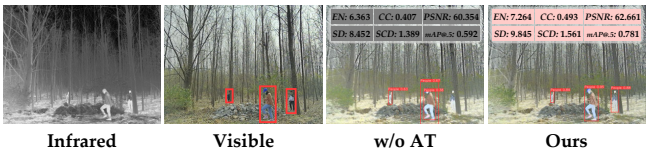

Figure 7: Ablation analysis of adversarial training.

Figure 8: Ablation analysis of $\mathcal{L}_{\mathrm{a}}$. Error maps between original and ablation results are given to observe differences.

Analysis of Loss Function Another ablation study is also to investigate whether the robustness of network would degrade if $\mathcal{L}_{\mathrm{a}}$ is not used. In Fig. 8, we present the ablation results without $\mathcal{L}_{\mathrm{a}}$. Instead, we employ the weighted average of source images as $\mathbf{y}$ and common loss functions (Tang, Yuan, and Ma 2022) into Eq. 4 and 5 for ablation experiments. From the error map (a), the fused image exhibits noticeable noise compared to the clean image. In addition, we embed $\mathbf{y}$ and $\mathcal{L}_{\mathrm{a}}$ into PAIF and retrain the model to verify whether they can also enhance the existing robustness. Apart from slight changes in luminance, the texture features and details of the targets do not improve. Therefore, it can be concluded that $\mathcal{L}_{\mathrm{a}}$ is not a plug-and-play loss function, which needs to work with the proposed network to achieve more robust representations.

Conclusion

This paper proposed a robust method for infrared and visible image fusion that is designed to endure adversarial disturbances. Based on the intrinsic nature of the fusion task, we conducted the adversarial attack and training processes by using the proposed anti-attack loss. During the training phase, we employed a Unet-based architecture and a transformer-based defensive refinement module to equip the network with a coarse-to-fine noise filtering capability. The defensive refinement modelue also comliemented missing features to refine textures of fused images. Compared to existing methods, A^2RNet demonstrates strong resilience against perturbations. Furthermore, it maintains a high level of performance in downstream tasks under attack.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (62376024, 62206015), and by the Fundamental Research Funds for the Central Universities (FRF-TP-22-043A1).

References

Aslantas, V.; and Bendes, E. 2015. A new image quality metric for image fusion: The sum of the correlations of differences. *Aeu-international Journal of electronics and communications*, 69(12): 1890–1896.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Cao, B.; Sun, Y.; Zhu, P.; and Hu, Q. 2023. Multi-modal gated mixture of local-to-global experts for dynamic image fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 23555–23564.

Cohen, I.; Huang, Y.; Chen, J.; Benesty, J.; Benesty, J.; Chen, J.; Huang, Y.; and Cohen, I. 2009. Pearson correlation coefficient. *Noise reduction in speech processing*, 1–4.

Ding, X.; Chen, J.; Yu, H.; Shang, Y.; Qin, Y.; and Ma, H. 2024. Transferable Adversarial Attacks for Object Detection Using Object-Aware Significant Feature Distortion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 1546–1554.

Gokhale, T.; Anirudh, R.; Kailkhura, B.; Thiagarajan, J. J.; Baral, C.; and Yang, Y. 2021. Attribute-guided adversarial training for robustness to natural perturbations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 7574–7582.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Ha, Q.; Watanabe, K.; Karasawa, T.; Ushiku, Y.; and Harada, T. 2017. MFNet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 5108–5115. IEEE.

Huang, G.; Liu, Z.; Van Der Maaten, L.; and Weinberger, K. Q. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4700–4708.

Huang, Z.; Liu, J.; Fan, X.; Liu, R.; Zhong, W.; and Luo, Z. 2022. Reconet: Recurrent correction network for fast and efficient multi-modality image fusion. In *European conference on computer Vision*, 539–555. Springer.

Jia, X.; Zhang, Y.; Wu, B.; Ma, K.; Wang, J.; and Cao, X. 2022. LAS-AT: adversarial training with learnable attack strategy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13398–13408.

Jiang, Z.; Li, X.; Liu, J.; Fan, X.; and Liu, R. 2024. Towards Robust Image Stitching: An Adaptive Resistance Learning against Compatible Attacks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 2589–2597.

Li, H.; Manjunath, B.; and Mitra, S. K. 1995. Multisensor image fusion using the wavelet transform. *Graphical models and image processing*, 57(3): 235–245.

Li, H.; and Wu, X.-J. 2018. DenseFuse: A fusion approach to infrared and visible images. *IEEE Transactions on Image Processing*, 28(5): 2614–2623.

Li, H.; Xu, T.; Wu, X.-J.; Lu, J.; and Kittler, J. 2023a. Lrrnet: A novel representation learning guided fusion network for infrared and visible images. *IEEE transactions on pattern analysis and machine intelligence*, 45(9): 11040–11052.

Li, J.; Chen, J.; Liu, J.; and Ma, H. 2023b. Learning a graph neural network with cross modality interaction for image fusion. In *Proceedings of the 31st ACM International Conference on Multimedia*, 4471–4479.

Li, J.; Liu, J.; Zhou, S.; Zhang, Q.; and Kasabov, N. K. 2022. Learning a coordinated network for detail-refinement multiexposure image fusion. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(2): 713–727.

Li, J.; Liu, J.; Zhou, S.; Zhang, Q.; and Kasabov, N. K. 2023c. Gesenet: A general semantic-guided network with couple mask ensemble for medical image fusion. *IEEE Transactions on Neural Networks and Learning Systems*.

Liu, J.; Fan, X.; Huang, Z.; Wu, G.; Liu, R.; Zhong, W.; and Luo, Z. 2022. Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5802–5811.

Liu, J.; Fan, X.; Jiang, J.; Liu, R.; and Luo, Z. 2021. Learning a deep multi-scale feature ensemble and an edge-attention guidance for image fusion. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(1): 105–119.

Liu, J.; Lin, R.; Wu, G.; Liu, R.; Luo, Z.; and Fan, X. 2024a. Coconet: Coupled contrastive learning network with multi-level feature ensemble for multi-modality image fusion. *International Journal of Computer Vision*, 132(5): 1748–1775.

Liu, J.; Liu, Z.; Wu, G.; Ma, L.; Liu, R.; Zhong, W.; Luo, Z.; and Fan, X. 2023a. Multi-interactive feature learning and a full-time multi-modality benchmark for image fusion and segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 8115–8124.

Liu, R.; Liu, Z.; Liu, J.; Fan, X.; and Luo, Z. 2024b. A task-guided, implicitly-searched and metainitialized deep model for image fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Liu, Y.; Liu, S.; and Wang, Z. 2015. A general framework for image fusion based on multi-scale transform and sparse representation. *Information fusion*, 24: 147–164.

Liu, Z.; Liu, J.; Zhang, B.; Ma, L.; Fan, X.; and Liu, R. 2023b. Paif: Perception-aware infrared-visible image fusion for attack-tolerant semantic segmentation. In *Proceedings of the 31st ACM International Conference on Multimedia*, 3706–3714.

Ma, J.; Ma, Y.; and Li, C. 2019. Infrared and visible image fusion methods and applications: A survey. *Information fusion*, 45: 153–178.

Ma, J.; Tang, L.; Fan, F.; Huang, J.; Mei, X.; and Ma, Y. 2022. SwinFusion: Cross-domain long-range learning for general image fusion via swin transformer. *IEEE/CAA Journal of Automatica Sinica*, 9(7): 1200–1217.

Ma, L.; Jin, D.; An, N.; Liu, J.; Fan, X.; Luo, Z.; and Liu, R. 2023. Bilevel fast scene adaptation for low-light image enhancement. *International Journal of Computer Vision*, 1–19.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2017. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*.

Mercer, J. 1909. Xvi. functions of positive and negative type, and their connection the theory of integral equations. *Philosophical transactions of the royal society of London. Series A, containing papers of a mathematical or physical character*, 209(441-458): 415–446.

Rao, D.; Xu, T.; and Wu, X.-J. 2023. TGFuse: An infrared and visible image fusion approach based on transformer and generative adversarial network. *IEEE Transactions on Image Processing*.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, 234–241. Springer.

Shafahi, A.; Najibi, M.; Xu, Z.; Dickerson, J.; Davis, L. S.; and Goldstein, T. 2020. Universal adversarial training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 5636–5643.

Shah, P.; Merchant, S. N.; and Desai, U. B. 2013. Multifocus and multispectral image fusion based on pixel significance using multiresolution decomposition. *Signal, Image and Video Processing*, 7: 95–109.

Shi, W.; Caballero, J.; Huszár, F.; Totz, J.; Aitken, A. P.; Bishop, R.; Rueckert, D.; and Wang, Z. 2016. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1874–1883.

Sun, Y.; Cao, B.; Zhu, P.; and Hu, Q. 2022. Drone-based RGB-infrared cross-modality vehicle detection via uncertainty-aware learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(10): 6700–6713.

Tang, L.; Yuan, J.; and Ma, J. 2022. Image fusion in the loop of high-level vision tasks: A semantic-aware real-time infrared and visible image fusion network. *Information Fusion*, 82: 28–42.

Wang, D.; Liu, J.; Liu, R.; and Fan, X. 2023. An interactively reinforced paradigm for joint infrared-visible image fusion and saliency object detection. *Information Fusion*, 98: 101828.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4): 600–612.

Yi, X.; Xu, H.; Zhang, H.; Tang, L.; and Ma, J. 2024. Text-IF: Leveraging Semantic Text Guidance for Degradation-Aware and Interactive Image Fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 27026–27035.

Yu, H.; Chen, J.; Ding, X.; Zhang, Y.; Tang, T.; and Ma, H. 2024. Step Vulnerability Guided Mean Fluctuation Adversarial Attack against Conditional Diffusion Models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 6791–6799.

Yu, Y.; Yang, W.; Tan, Y.-P.; and Kot, A. C. 2022. Towards robust rain removal against adversarial attacks: A comprehensive benchmark analysis and beyond. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6013–6022.

Yue, J.; Fang, L.; Xia, S.; Deng, Y.; and Ma, J. 2023. Dif-fusion: Towards high color fidelity in infrared and visible image fusion with diffusion models. *IEEE Transactions on Image Processing*.

Zhang, H.; Xu, H.; Tian, X.; Jiang, J.; and Ma, J. 2021. Image fusion meets deep learning: A survey and perspective. *Information Fusion*, 76: 323–336.

Zhao, Z.; Bai, H.; Zhang, J.; Zhang, Y.; Zhang, K.; Xu, S.; Chen, D.; Timofte, R.; and Van Gool, L. 2024a. Equivariant multi-modality image fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 25912–25921.

Zhao, Z.; Bai, H.; Zhu, Y.; Zhang, J.; Xu, S.; Zhang, Y.; Zhang, K.; Meng, D.; Timofte, R.; and Van Gool, L. 2023. DDFM: denoising diffusion model for multi-modality image fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8082–8093.

Zhao, Z.; Deng, L.; Bai, H.; Cui, Y.; Zhang, Z.; Zhang, Y.; Qin, H.; Chen, D.; Zhang, J.; Wang, P.; et al. 2024b. Image Fusion via Vision-Language Model. *arXiv preprint arXiv:2402.02235*.

Zheng, Z.; and Wu, C. 2024. U-shaped vision mamba for single image dehazing. *arXiv preprint arXiv:2402.04139*.