

---

# Deno-IF: Unsupervised Noisy Visible and Infrared Image Fusion Method

---

Han Xu<sup>1</sup>, Yuyang Li<sup>1</sup>, Yunfei Deng<sup>1</sup>, Jiayi Ma<sup>2</sup>, Guangcan Liu<sup>1\*</sup>

<sup>1</sup> School of Automation, Southeast University, Nanjing, China

<sup>2</sup> Electronic Information School, Wuhan University, Wuhan, China

{xu\_han, yuyangli}@seu.edu.cn, {yunfeideng455, jyma2010}@gmail.com,  
guangcanliu@seu.edu.cn

## Abstract

Most image fusion methods are designed for ideal scenarios and struggle to handle noise. Existing noise-aware fusion methods are supervised and heavily rely on constructed paired data, limiting performance and generalization. This paper proposes a novel unsupervised noisy visible and infrared image fusion method, comprising two key modules. First, when only noisy source images are available, a convolutional low-rank optimization module decomposes clean components based on convolutional low-rank priors, guiding subsequent optimization. The unsupervised approach eliminates data dependency and enhances generalization across various and variable noise. Second, a unified network jointly realizes denoising and fusion. It consists of both intra-modal recovery and inter-modal recovery and fusion, also with a convolutional low-rankness loss for regularization. By exploiting the commonalities of denoising and fusion, the joint framework significantly reduces network complexity while expanding functionality. Extensive experiments validate the effectiveness and generalization of the proposed method for image fusion under various and variable noise conditions. The code is publicly available at <https://github.com/hanna-xu/Deno-IF>.

## 1 Introduction

The widely used visible images, although rich in color and textures, are vulnerable to environmental factors, *e.g.*, illumination variations and occlusions. In contrast, infrared images can capture thermal radiation information under extreme-light or obscured conditions while they often suffer from poor details and low quality. Therefore, visible and infrared image fusion aims to combine the complementary and valuable information of these two modalities. The single fused image can retain the clarity and fine details of the visible image while incorporating the prominence of thermal targets in the infrared image for comprehensive representation. The fusion results can aid in both the human visual perception and subsequent widespread applications, such as object detection [26, 51, 7], semantic segmentation [20, 43], autonomous driving [47, 3].

With the rapid development of deep learning, many learning-based fusion methods have been proposed. These methods can be categorized into: methods based on auto-encoder (AE) [2, 40], convolutional neural network (CNN) [31, 21, 32], generative adversarial networks (GANs) [11, 42, 18], Transformer [16, 13], Mamba [29] and Diffusion [50, 35, 23]. Despite the progress in network architectures, these methods purely aim to preserve the original information in source images without distinguishing the information quality. In real-world scenarios, the quality of imaging is adversely affected by challenging conditions (*e.g.*, poor illumination) and the inherent constraints

---

\*Corresponding author

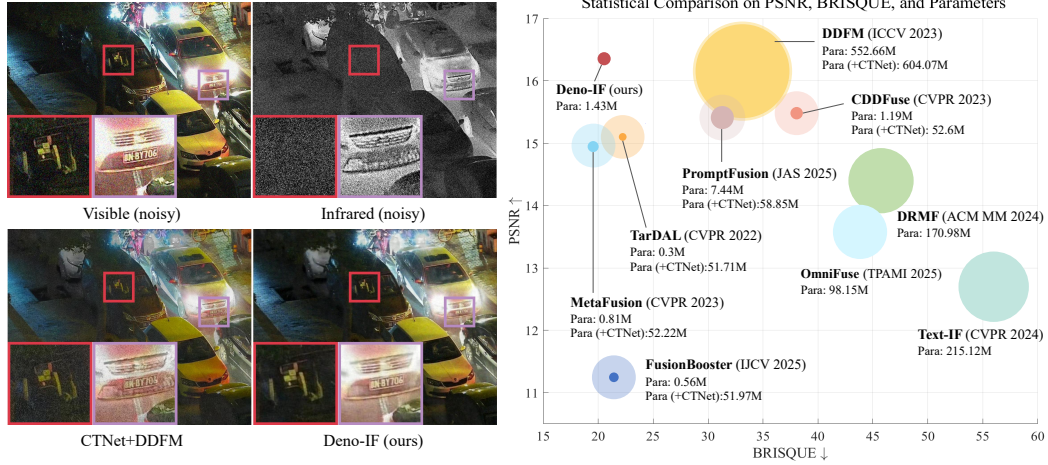


Figure 1: Comparison of Deno-IF with SOTA image denoising and fusion methods by: i) similarity with clean source images (measured by PSNR), ii) image quality assessment (measured by BRISQUE), and iii) complexity (measured by parameter numbers and represented by circle sizes). The high- and low-saturation circle regions represent parameters of fusion and denoising method [25], respectively.

of cost-effective multi-modal devices. These systems are prone to significant noise, stemming from degraded signal acquisition (especially in low-cost microbolometers and CMOS sensors) and ineffective embedded processing pipelines. Under interference conditions, there remain several critical challenges that need to be addressed to achieve robust high-quality fusion results [37, 33].

As most existing fusion methods are tailored for standard scenarios and ineffective at suppressing noise, an intuitive approach is to first denoise the source images before performing fusion. However, the framework taking denoising and fusion as independent tasks presents two limitations. On the one hand, denoising methods are typically deployed on a single source image, making it challenging to leverage complementary knowledge from the other modality. When denoising performance is limited, fusion methods are unable to handle the residual noise or over-smoothness, leading to its persistence in the fusion result, as shown in the result in Fig. 1. On the other hand, the mutually independent framework suffers from significant redundancy. The goal of image fusion is to extract vital and complementary information, which is then integrated into the fused images. Besides, image denoising aims to mine essential clean data from noisy inputs. From the perspective of essential information extraction and mining, their objectives are consistently aligned to some extent. The independent framework ignores the commonality, leading to redundancy as Fig. 1. There remains the potential to jointly realize denoising and fusion in a streamlined manner.

Although some degradation-aware image fusion methods exist [22, 36], they are supervised methods aimed at learning the mapping from degraded source images to clean fused images. The key lies in the construction of large amount of noisy and clean source image pairs. Their denoising effectiveness is limited by the fixed and limited mapping relationship between the constructed noisy and clean data, resulting in poor generalization. As a result, these methods are only effective for specific types of noise with particular variances. Facing various and variable noise encountered in real-world scenarios, these methods are highly susceptible to failure, as shown in their performances in Fig. 1.

To overcome these issues, this paper proposes an unsupervised noisy visible and infrared image fusion method, termed as Deno-IF. It consists of two modules, including a convolutional low-rank optimization module and a joint denoising and fusion module. First, based on the convolutional low-rank property of high-quality data, the convolutional low-rank optimization module decomposes clean component from noisy input through convolution nuclear norm minimization in an unsupervised manner. Then, the joint denoising and fusion module takes noisy source images as input and directly outputs the fused image. The network includes both intra-modal recovery and inter-modal recovery and fusion, with self- and cross-modal attention to deal with complex individual-modal and complementary multi-modal information. The decomposed data provides the optimization guidance for the joint denoising and fusion module. Besides, a convolution matrix-based regularization loss is

designed to further suppress noise in fused images. Thus, it can realize denoising and fusion jointly and generate clean fused images. The contributions are summarized as:

- We propose an unsupervised noisy visible and infrared image fusion method. Without the supervision of clean data, it can still realize denoising and fusion simultaneously with fewer network parameters, and is robust against various and variable noise conditions, ensuring effective across diverse scenarios.
- For noisy source images, we design a convolutional low-rank optimization module. As clean data usually exhibits convolutional low-rankness, we introduce the convolution nuclear norm minimization to decompose clean data from noisy inputs. It provides the optimization guidance for fusion network in the training phase.
- For joint denoising and fusion, a Transformer-based network consisting of intra-modal recovery and inter-modal recovery and fusion is designed. It leverages self- and cross-modal attention to collaboratively strive to approximate the optimization guidance. Besides, a convolution matrix-based regularization loss is designed to further suppress noise in fused images.

## 2 Related Work

**Learning-Based Image Fusion Methods.** These methods preserve the information in source images through the design of network architectures and loss functions. According to architectures, they can be divided into AE, CNN, GAN, Transformer, Mamba and Diffusion-based methods. AE-based methods design encoders and decoders for feature extraction and reconstruction, with traditional or learning-based [30] fusion strategies. CNN-based methods utilize end-to-end fusion [41, 6] and the key lies in pixel-level loss functions. GAN-based methods preserve unique information through adversarial games from the perspective of probability distribution while they always suffer from unstable training and model collapse [12, 17]. Transformer and Mamba-based methods build long-range dependency to solve the limited receptive field of CNN [13, 24, 29, 15]. Diffusion-based methods leverages the powerful generative properties of diffusion models to establish a stable fusion process [38, 4, 34]. However, these methods are designed for normal scenarios and ignore degradations in real world. Facing degraded images, they fail to tackle degradations, reducing the quality and usability of results.

**Degradation-Aware Image Fusion Methods.** Some fusion methods suppress degradations and achieve fusion in a unified network. Text-IF [36] leverages the text prompts with degradation information to guide the enhanced fusion. The network is trained with the manually constructed paired degraded and high-quality source images. DRMF [22] pretrains multiple degradation robust diffusion models and devises a combination module to integrate priors. Similarly, the performance also depends on paired constructed data. They can handle different types of degradations including noise. However, since the introduced noise has fixed variance, it demonstrates limited generalization. These methods heavily rely on paired constructed data. This paper proposes an unsupervised method that requires only degraded data for training. We infer the clean component to guide the fusion network, significantly reducing the data demand and the number of network parameters.

## 3 Proposed Method

### 3.1 Problem Formulation

Mathematically, a pair of observed noisy visible and infrared images  $\{\mathbf{V}, \mathbf{R}\}$  can be decomposed as:

$$\mathbf{V} = \mathbf{L}_v + \mathbf{S}_v, \quad \mathbf{R} = \mathbf{L}_r + \mathbf{S}_r, \quad (1)$$

where  $\mathbf{L}_v, \mathbf{L}_r$  is the clean data and  $\mathbf{S}_v, \mathbf{S}_r$  is the noise. As clean data,  $\mathbf{L}_v, \mathbf{L}_r$  should be of convolutional low-rankness [8].  $\mathbf{S}_v, \mathbf{S}_r$  should be small-scale perturbations with limited overall energy and amplitude to ensure the convolutional low-rankness of  $\mathbf{L}_v, \mathbf{L}_r$ .

Taking noisy source images  $\mathbf{V}, \mathbf{R}$  as input, the jointly denoising and fusion module generates a high-quality fused image  $\mathbf{F}$  free from noise. Based on the decomposition assumption in Eq. (1), the goal is equivalent to making  $\mathbf{F}$  contain vital information in  $\mathbf{L}_v, \mathbf{L}_r$  mined from  $\mathbf{V}, \mathbf{R}$ , respectively.

As shown in Fig. 2, the overall framework consists of the convolutional low-rank optimization module and the joint denoising and fusion module. The convolutional low-rank optimization module solves

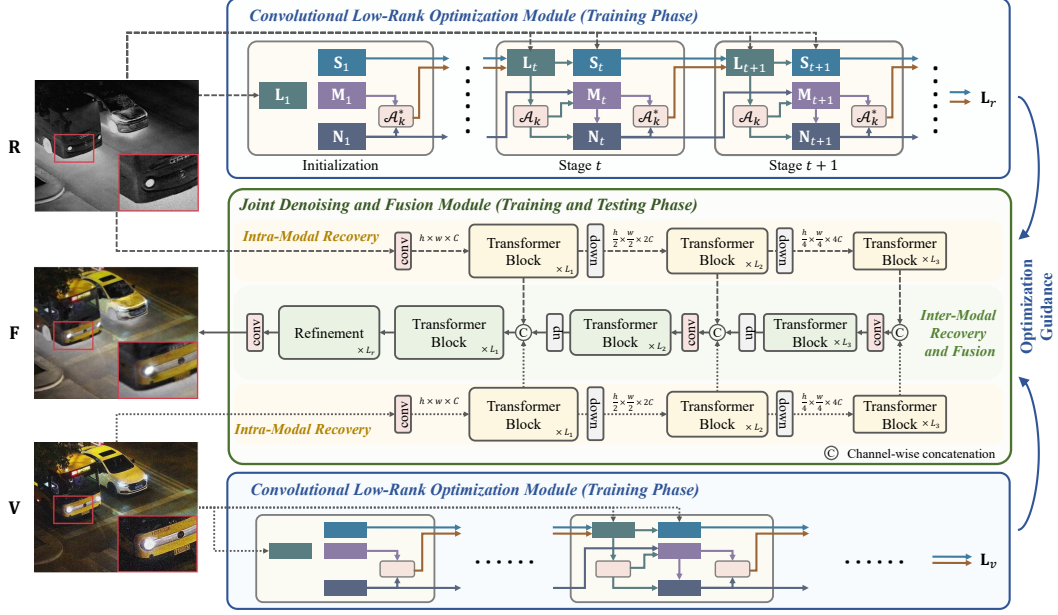


Figure 2: Overall framework of the unsupervised noisy infrared and visible image fusion method.

the ill-posed decomposition problem to provide precise guidance for network optimization. In the joint denoising and fusion module, a network  $\mathcal{F}$  takes noisy source image as input, consists of intra-modal recovery and inter-modal recovery and fusion. Based on Transformer, intra-modal recovery leverages the self-attention mechanism to enhance restoration by capturing long-range dependencies within individual modality. Inter-modal recovery and fusion exploits cross-modal attention, enabling the joint restoration and fusion of complementary multi-modal information.  $\mathcal{F}$  is optimized with the results of the convolutional low-rank optimization module to realize joint denoising and fusion.

### 3.2 Convolutional Low-Rank Optimization Module

For the decomposition problem in Eq. (1), we take the infrared image  $\mathbf{R} \in \mathbb{R}^{h \times w}$  as example (subscript  $r$  is temporarily omitted for simplification). To ensure the respective properties of  $\mathbf{L}$  and  $\mathbf{S}$ , the problem can be formulated as:

$$\min_{\mathbf{L}, \mathbf{S}} \|\mathcal{A}_k(\mathbf{L})\|_* + \beta \|\mathbf{S}\|_F^2, \quad s.t. \quad \mathbf{R} = \mathbf{L} + \mathbf{S}, \quad (2)$$

where the first term is based on convolution nuclear norm, which is the nuclear norm of its convolution matrix [8].  $\mathcal{A}_k(\cdot)$  is a linear map from  $\mathbb{R}^{h \times w}$  to  $\mathbb{R}^{h'w' \times k_1k_2}$  where  $k_1$  ( $1 \leq k_1 \leq h$ ) and  $k_2$  ( $1 \leq k_2 \leq w$ ) denote the kernel size in  $\mathcal{A}_k(\cdot)$ . It produces the convolution matrix of the input.  $\|\cdot\|_*$  and  $\|\cdot\|_F$  denote nuclear and Frobenius norms, respectively.  $\beta$  controls the trade-off between these terms.

Within the maximum a posteriori (MAP) inference framework,  $\mathbf{L}$  and  $\mathbf{S}$  can be obtained by minimizing the following energy function:

$$E(\mathbf{L}, \mathbf{S}) = \|\mathbf{R} - \mathbf{L} - \mathbf{S}\|_F^2 + \alpha \|\mathcal{A}_k(\mathbf{L})\|_* + \beta \|\mathbf{S}\|_F^2, \quad (3)$$

where the first term is fidelity term and the last two terms are regularization term denoting imposed priors over  $\mathbf{L}$  and  $\mathbf{S}$ .  $\alpha$  also controls the trade-off.

With [19] releasing nuclear norm as:

$$\|\mathbf{X}\|_* = \min_{\mathbf{A}, \mathbf{B}} \frac{1}{2} \|\mathbf{A}\|_F^2 + \frac{1}{2} \|\mathbf{B}\|_F^2, \quad s.t. \quad \mathbf{X} = \mathbf{AB}, \quad (4)$$

the energy function can be rewritten as:

$$E(\mathbf{L}, \mathbf{S}, \mathbf{M}, \mathbf{N}) = \|\mathbf{R} - \mathbf{L} - \mathbf{S}\|_F^2 + \frac{\alpha}{2} \|\mathbf{M}\|_F^2 + \frac{\alpha}{2} \|\mathbf{N}\|_F^2 + \gamma \|\mathcal{A}_k(\mathbf{L}) - \mathbf{MN}\|_F^2 + \beta \|\mathbf{S}\|_F^2, \quad (5)$$

where  $\mathbf{M} \in \mathbb{R}^{hw \times m}$ ,  $\mathbf{N} \in \mathbb{R}^{m \times k_1 k_2}$ .  $\gamma$  is a trade-off parameter.

This problem can be solved by iteratively addressing the subproblems related to  $\mathbf{L}$ ,  $\mathbf{S}$ ,  $\mathbf{M}$ ,  $\mathbf{N}$ . With  $t$  denotes the iteration step, the problem can be partitioned into the following four subproblems:

$$\mathbf{L}_t = \arg \min_{\mathbf{L}} \|\mathbf{R} - \mathbf{L} - \mathbf{S}_{t-1}\|_F^2 + \gamma \|\mathcal{A}_k(\mathbf{L}) - \mathbf{M}_{t-1} \mathbf{N}_{t-1}\|_F^2, \quad (6a)$$

$$\mathbf{S}_t = \arg \min_{\mathbf{S}} \|\mathbf{R} - \mathbf{L}_t - \mathbf{S}\|_F^2 + \beta \|\mathbf{S}\|_F^2, \quad (6b)$$

$$\mathbf{M}_t = \arg \min_{\mathbf{M}} \frac{\alpha}{2} \|\mathbf{M}\|_F^2 + \gamma \|\mathcal{A}_k(\mathbf{L}_t) - \mathbf{M} \mathbf{N}_{t-1}\|_F^2, \quad (6c)$$

$$\mathbf{N}_t = \arg \min_{\mathbf{N}} \frac{\alpha}{2} \|\mathbf{N}\|_F^2 + \gamma \|\mathcal{A}_k(\mathbf{L}_t) - \mathbf{M}_t \mathbf{N}\|_F^2. \quad (6d)$$

When  $t=1$ , we initially set the variables as  $\mathbf{L}_1 = \mathbf{R}$ .  $\mathbf{S}_1$  is set to randomly distributed noise.  $\mathbf{M}_1$  and  $\mathbf{N}_1$  are set to identity-like diagonal matrix. During the iteration process, the closed-form solutions for these four subproblems can then be obtained as:

$$\mathbf{L}_t = \frac{\mathbf{R} - \mathbf{S}_{t-1} + \gamma \mathcal{A}_k^*(\mathbf{M}_{t-1} \mathbf{N}_{t-1})}{\gamma + 1}, \quad (7)$$

where  $\mathcal{A}_k^*(\cdot)$  is the Hermitian adjoint of  $\mathcal{A}_k$ .

$$\mathbf{S}_t = \frac{\mathbf{R} - \mathbf{L}_t}{\beta + 1}, \quad \mathbf{M}_t = 2\gamma \mathcal{A}_k(\mathbf{L}_t) \mathbf{N}_{t-1}^\top (\alpha \mathbf{I} + 2\gamma \mathbf{N}_{t-1} \mathbf{N}_{t-1}^\top)^{-1}, \quad (8)$$

where  $\mathbf{I}$  is a  $m \times m$  identity matrix.

$$\mathbf{N}_t = 2\gamma (\alpha \mathbf{I} + 2\gamma \mathbf{M}_t^\top \mathbf{M}_t)^{-1} \mathbf{M}_t^\top \mathcal{A}_k(\mathbf{L}_t). \quad (9)$$

For the visible image  $\mathbf{V} \in \mathbb{R}^{h \times w \times c}$ , the primary distinction to infrared data lies in three channels. It results in slight differences in the construction of convolution matrix and the shapes of  $\mathbf{M}$ ,  $\mathbf{N}$ . Specifically,  $\mathcal{A}_k(\mathbf{V})$  is a linear map from  $\mathbb{R}^{h \times w \times c}$  to  $\mathbb{R}^{hwc \times k_1 k_2 k_3}$ , where  $k_1$  ( $1 \leq k_1 \leq h$ ),  $k_2$  ( $1 \leq k_2 \leq w$ ) and  $k_3$  ( $1 \leq k_3 \leq c$ ) are the kernel size.  $\mathbf{M} \in \mathbb{R}^{hwc \times n}$  and  $\mathbf{N} \in \mathbb{R}^{n \times k_1 k_2 k_3}$ .

### 3.3 Joint Denoising and Fusion Module

**Network Architecture.** Directly denoising multi-modal source images as a whole without distinction will overlook the modal differences and noise characteristics between multi-modal images, posing greater challenges to  $\mathcal{F}$ . Thus, the fusion network  $\mathcal{F}$  comprises two key components: intra-modal recovery and inter-modal recovery and fusion. In one source image, the modality is unique and the noise characteristics are consistent. We first perform intra-modal feature recovery in visible and infrared domains, respectively. Then, the inter-modal recovery and fusion integrates information across different modalities while effectively handling the complex noise characteristics among these modalities. The overall network is named as I2Former. The output is the clean fused image  $\mathbf{F}$ .

For the two components in  $\mathcal{F}$ , we adopt the Transformer block [39] for intra- and inter-modal feature extraction and integration. In intra-modal recovery, we respectively extract features from  $\mathbf{V}$ ,  $\mathbf{R}$  as  $\mathbf{f}_v$ ,  $\mathbf{f}_r$ . In each block, the intra-modal multi-Dconv head transposed attention ( $T_{ter}$ ) generates query, key and value projections for  $\mathbf{f}_v$ ,  $\mathbf{f}_r$  as:

$$\{\mathbf{Q}_v, \mathbf{K}_v, \mathbf{V}_v\} = T_{ter}^V(\mathbf{f}_v), \quad \{\mathbf{Q}_r, \mathbf{K}_r, \mathbf{V}_r\} = T_{ter}^R(\mathbf{f}_r). \quad (10)$$

Then, the output visible feature maps are  $\hat{\mathbf{f}}_v = \text{Softmax}(\frac{\mathbf{K}_v \cdot \mathbf{Q}_v}{\varsigma}) \mathbf{V}_v + \mathbf{f}_v$ , where  $\varsigma$  is a learnable scaling parameter. Infrared features are modulated similarly. The features are passed through gated-Dconv feed-forward network for controllable transformation. Multiple blocks generate multi-scale features.

In inter-modal recovery and fusion, the visible and infrared features of the smallest scale are concatenated, fed into a convolutional layer for channel reduction and upsampled. With  $\mathbf{f}_f$  and  $\mathbf{f}_v$ ,  $\mathbf{f}_r$  as input, the inter-modal multi-Dconv head transposed attention ( $T_{tra}$ ) generates projections as:

$$\{\mathbf{Q}_f, \mathbf{K}_f, \mathbf{V}_f\} = T_{tra}(W(\mathbf{f}_v, \mathbf{f}_r, \mathbf{f}_f)), \quad (11)$$

where  $W$  is a convolutional layer for channel reduction. The final output features can be obtained as  $\hat{\mathbf{f}}_f = \text{Softmax}(\frac{\mathbf{K}_f \cdot \mathbf{Q}_f}{\varsigma}) \mathbf{V}_f + W(\mathbf{f}_v, \mathbf{f}_r, \mathbf{f}_f)$ . The fused features of the original scale are finally passed through some blocks for refinement and a convolutional layer to generate the fused image  $\mathbf{F}$ .

**Loss Functions.** From the aspect of network optimization, clean components derived with convolutional low-rank priors serve as i) physics-driven prior injector to regularize fused images, and ii) implicit teachers that enable knowledge distillation of denoising priors into the fusion network. It enables simultaneous noise suppression and information fusion. Thus, the loss functions of  $\mathcal{F}$  consist of some data fidelity terms and a regularization term.

*Data Fidelity Terms.* Sec. 3.2 obtains the estimation of  $\mathbf{L}_v$  and  $\mathbf{L}_r$ . The data fidelity terms constrain the fused image  $\mathbf{F}$  to keep similarity with  $\mathbf{L}_v$  and  $\mathbf{L}_r$  from multiple perspectives to preserve the scene information, including intensity, gradients, and chrominance components.

The intensity loss makes  $\mathbf{F}$  capture proper intensity distribution. One target is to preserve the thermal prominence in  $\mathbf{L}_r$ . Considering the low thermal radiation regions in  $\mathbf{L}_r$  and low-light regions in  $\mathbf{L}_v$ , the other target is to preserve the more prominent information. The intensity loss is denoted as  $\mathcal{L}_{in} = \|\mathbf{F}^y - \max(\mathbf{L}_v^y, \mathbf{L}_r^y)\|_1$  where  $\mathbf{L}_v^y, \mathbf{F}^y$  are the Y channels of  $\mathbf{L}_v, \mathbf{F}$  in YCbCr space, respectively.

The gradient loss preserves the prominent textures in  $\mathbf{L}_v, \mathbf{L}_r$  to present more contents. We generate a gradient mask  $M_g$  considering the gradient relationship between  $\mathbf{L}_v^y, \mathbf{L}_r^y$ . When  $|\nabla \mathbf{L}_{v,i,j}^y| > |\nabla \mathbf{L}_{r,i,j}^y|$ ,  $\mathbf{m}_{g,i,j} = 1$  where  $\nabla$  is Laplacian operator,  $i$  and  $j$  indicate the spatial position. Otherwise,  $\mathbf{m}_{g,i,j} = 0$ . The loss constrains the gradient similarity as:

$$\mathcal{L}_g = \|\nabla \mathbf{F}^y - [\mathbf{m}_g \cdot \nabla \mathbf{L}_v^y + (1 - \mathbf{m}_g) \cdot \nabla \mathbf{L}_r^y]\|_1. \quad (12)$$

The chrominance loss preserves the chrominance information in  $\mathbf{L}_v$  into  $\mathbf{F}$ .  $\mathbf{L}_v, \mathbf{F}$  are translated into YCbCr space and the constraint is performed on chrominance channels as:

$$\mathcal{L}_{chr} = \|\mathbf{F}^{Cb} - \mathbf{L}_v^{Cb}\|_1 + \|\mathbf{F}^{Cr} - \mathbf{L}_v^{Cr}\|_1. \quad (13)$$

*Regularization Term.* In the optimization process of  $\mathbf{F}$ , the performance gap in intra-modal recovery and the residual and cumulative deviation in inter-modal recovery and fusion will affect performance. To further correct deviations, the convolutional low rank of fused image is regularized as:

$$\mathcal{L}_{rank} = \|\mathcal{A}_k(\mathbf{F})\|_*, \quad (14)$$

where  $\mathcal{A}_k(\mathbf{F})$  computes the convolution matrix of  $\mathbf{F}$ .  $\mathcal{A}_k(\cdot)$  is a linear map from  $\mathbb{R}^{h \times w \times c}$  to  $\mathbb{R}^{hwc \times k_1 k_2 k_3}$  (defined at the end of Sec. 3.2). The nuclear norm imposes a low-rank constraint on the convolution matrix of  $\mathbf{F}$ , thereby reducing residual noise in  $\mathbf{F}$  by mining its intrinsic structure.

With hyper-parameters  $\lambda, \eta, \kappa$  to control the trade-off, the final loss function is defined as:

$$\mathcal{L}_{\mathcal{F}} = \mathcal{L}_{in} + \lambda \mathcal{L}_g + \eta \mathcal{L}_{chr} + \kappa \mathcal{L}_{rank}. \quad (15)$$

## 4 Experiments and Results

**Datasets and Implementation.** Existing datasets are collected for the target of high image quality, thus relying on cost-prohibitive imaging systems to ensure high signal-to-noise ratios. As there are no publicly available noisy infrared and visible datasets for direct use, We simulate real-life noisy source images by injecting two representative types of noise with various levels (additive Gaussian noise and multiplicative speckle noise) into clean data [27, 46]. The clean data is not used during training. We train Deno-IF on 2120 pairs of visible and infrared images across two datasets, including LLVIP [5] and M3FD [9]. For the introduced noise,  $\sigma$  randomly distributed in  $[10, 50]$ . In the training phase, images are randomly cropped into patches of  $128 \times 128$  for training. The evaluation is performed on image pairs introduced Gaussian or speckle noise with randomly distributed variance.

In convolutional low-rank optimization module,  $h$  and  $w$  are 128,  $c = 3$ . Kernel size of  $\mathcal{A}_k(\cdot)$  are  $k_1, k_2 = 12, k_3 = 2$ .  $m, n = 256$ . We perform 30 iterations for each patch. For parameters in Eq. (5),  $\beta = 2$ .  $\alpha, \gamma$  are empirically related to data characteristics. As an unsupervised method, for a patch  $\mathbf{x}$ , we roughly estimate its noise level as  $n = \mathbb{E}[\|\nabla(\mathbf{x}) - \nabla(G(\mathbf{x}))\|]$ , where  $G(\cdot)$  is Gaussian blur.  $\alpha$  is  $200n$  and  $80n$  for visible and infrared data respectively.  $\beta = 15n$ . These manually designed settings stem from the consideration of i) a principled trade-off for adaptive regularization (strong for high-level noise for noise suppression, weak for low-level noise to preserve details); and ii) computational efficiency for an optimal efficiency-accuracy balance.

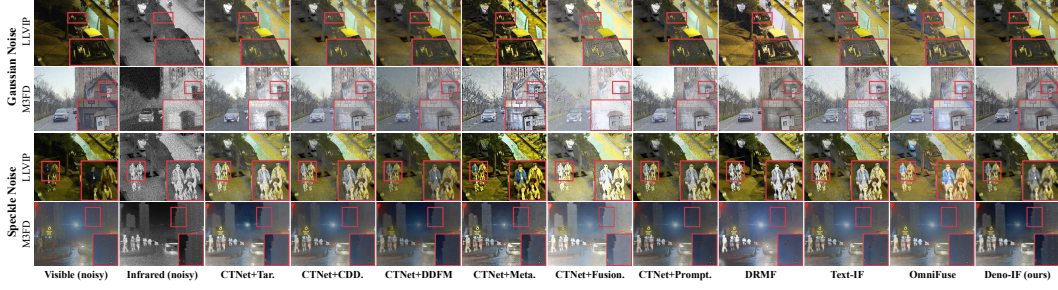


Figure 3: Qualitative results when both source images are subject to various-level **Gaussian/speckle** noise on LLVIP and M3FD datasets.

Table 1: Quantitative results when both source images suffer from various-level **Gaussian** or **Speckle** noise (**bold**: optimal, underline: suboptimal).

| Gaussian   | LLVIP Dataset                     |                                    |                                   |                                   |                                    | M3FD Dataset                      |                                    |                                   |                                   |                                     |
|------------|-----------------------------------|------------------------------------|-----------------------------------|-----------------------------------|------------------------------------|-----------------------------------|------------------------------------|-----------------------------------|-----------------------------------|-------------------------------------|
|            | SSIM $\uparrow$                   | PSNR $\uparrow$                    | FSIM $\uparrow$                   | CC $\uparrow$                     | BRISQUE $\downarrow$               | SSIM $\uparrow$                   | PSNR $\uparrow$                    | FSIM $\uparrow$                   | CC $\uparrow$                     | BRISQUE $\downarrow$                |
| CT+Tar.    | 0.421 $\pm$ 0.169                 | 15.600 $\pm$ 1.464                 | 0.792 $\pm$ 0.024                 | 0.680 $\pm$ 0.082                 | <u>19.519<math>\pm</math>8.577</u> | 0.469 $\pm$ 0.143                 | 14.597 $\pm$ 2.781                 | 0.756 $\pm$ 0.032                 | 0.513 $\pm$ 0.195                 | 24.898 $\pm$ 9.713                  |
| CT+CDD.    | 0.446 $\pm$ 0.206                 | 15.881 $\pm$ 1.366                 | 0.803 $\pm$ 0.028                 | 0.698 $\pm$ 0.072                 | 33.929 $\pm$ 20.019                | 0.466 $\pm$ 0.140                 | 15.076 $\pm$ 2.789                 | 0.772 $\pm$ 0.031                 | 0.501 $\pm$ 0.226                 | 42.196 $\pm$ 14.175                 |
| CT+DDFM    | 0.474 $\pm$ 0.178                 | <u>16.782<math>\pm</math>1.138</u> | <b>0.815<math>\pm</math>0.022</b> | <u>0.736<math>\pm</math>0.077</u> | 30.624 $\pm$ 16.596                | 0.489 $\pm$ 0.165                 | <u>15.525<math>\pm</math>2.513</u> | 0.767 $\pm$ 0.051                 | <b>0.587<math>\pm</math>0.191</b> | 35.771 $\pm$ 13.177                 |
| CT+Meta.   | <u>0.502<math>\pm</math>0.099</u> | 15.274 $\pm$ 1.193                 | 0.694 $\pm$ 0.045                 | 0.662 $\pm$ 0.083                 | 24.597 $\pm$ 12.281                | 0.533 $\pm$ 0.132                 | 14.606 $\pm$ 2.650                 | 0.675 $\pm$ 0.053                 | 0.548 $\pm$ 0.194                 | <b>14.430<math>\pm</math>5.869</b>  |
| CT+Fusion. | 0.270 $\pm$ 0.135                 | 11.694 $\pm$ 0.624                 | 0.760 $\pm$ 0.042                 | 0.679 $\pm$ 0.078                 | 26.381 $\pm$ 11.446                | 0.459 $\pm$ 0.094                 | 10.618 $\pm$ 1.040                 | 0.744 $\pm$ 0.047                 | <u>0.565<math>\pm</math>0.158</u> | <u>16.474<math>\pm</math>11.822</u> |
| CT+Prompt. | 0.433 $\pm$ 0.197                 | 15.811 $\pm$ 1.331                 | 0.803 $\pm$ 0.029                 | 0.702 $\pm$ 0.073                 | 36.876 $\pm$ 19.142                | 0.474 $\pm$ 0.125                 | 15.007 $\pm$ 2.679                 | <u>0.777<math>\pm</math>0.033</u> | 0.470 $\pm$ 0.243                 | 25.837 $\pm$ 13.808                 |
| DRMF       | 0.148 $\pm$ 0.081                 | 11.899 $\pm$ 1.242                 | 0.693 $\pm$ 0.031                 | 0.489 $\pm$ 0.156                 | 67.922 $\pm$ 12.462                | 0.280 $\pm$ 0.121                 | 13.482 $\pm$ 1.628                 | 0.738 $\pm$ 0.043                 | 0.334 $\pm$ 0.219                 | 44.186 $\pm$ 11.117                 |
| Text-IF    | 0.208 $\pm$ 0.069                 | 14.956 $\pm$ 1.125                 | 0.770 $\pm$ 0.025                 | 0.652 $\pm$ 0.071                 | 54.482 $\pm$ 10.614                | 0.225 $\pm$ 0.089                 | 13.837 $\pm$ 2.509                 | 0.710 $\pm$ 0.055                 | 0.404 $\pm$ 0.209                 | 37.228 $\pm$ 8.241                  |
| OmniFuse   | 0.377 $\pm$ 0.080                 | 12.642 $\pm$ 0.727                 | 0.731 $\pm$ 0.026                 | 0.642 $\pm$ 0.089                 | 41.046 $\pm$ 3.748                 | <u>0.540<math>\pm</math>0.098</u> | 14.488 $\pm$ 2.119                 | 0.763 $\pm$ 0.031                 | 0.459 $\pm$ 0.211                 | 46.787 $\pm$ 6.876                  |
| Deno-IF    | <b>0.616<math>\pm</math>0.064</b> | <b>17.070<math>\pm</math>1.426</b> | <u>0.811<math>\pm</math>0.023</u> | <b>0.738<math>\pm</math>0.074</b> | <b>16.725<math>\pm</math>4.091</b> | <b>0.585<math>\pm</math>0.081</b> | <b>15.637<math>\pm</math>2.899</b> | <b>0.788<math>\pm</math>0.029</b> | 0.563 $\pm$ 0.205                 | 24.298 $\pm$ 7.095                  |

| Speckle    | LLVIP Dataset                     |                                    |                                   |                                   |                                    | M3FD Dataset                      |                                    |                                   |                                   |                                    |
|------------|-----------------------------------|------------------------------------|-----------------------------------|-----------------------------------|------------------------------------|-----------------------------------|------------------------------------|-----------------------------------|-----------------------------------|------------------------------------|
|            | SSIM $\uparrow$                   | PSNR $\uparrow$                    | FSIM $\uparrow$                   | CC $\uparrow$                     | BRISQUE $\downarrow$               | SSIM $\uparrow$                   | PSNR $\uparrow$                    | FSIM $\uparrow$                   | CC $\uparrow$                     | BRISQUE $\downarrow$               |
| CT+Tar.    | 0.423 $\pm$ 0.073                 | 15.125 $\pm$ 1.464                 | 0.776 $\pm$ 0.024                 | 0.687 $\pm$ 0.072                 | 13.375 $\pm$ 9.263                 | 0.553 $\pm$ 0.129                 | 15.358 $\pm$ 2.393                 | 0.775 $\pm$ 0.039                 | 0.553 $\pm$ 0.206                 | <u>13.950<math>\pm</math>9.283</u> |
| CT+CDD.    | 0.419 $\pm$ 0.096                 | 15.527 $\pm$ 1.515                 | 0.789 $\pm$ 0.026                 | 0.681 $\pm$ 0.063                 | 45.497 $\pm$ 13.496                | 0.539 $\pm$ 0.145                 | 15.906 $\pm$ 2.520                 | 0.789 $\pm$ 0.041                 | 0.558 $\pm$ 0.215                 | 25.231 $\pm$ 14.022                |
| CT+DDFM    | 0.474 $\pm$ 0.086                 | <u>16.191<math>\pm</math>1.218</u> | <b>0.808<math>\pm</math>0.022</b> | <u>0.723<math>\pm</math>0.069</u> | 32.897 $\pm$ 12.572                | 0.584 $\pm$ 0.125                 | <u>16.232<math>\pm</math>2.259</u> | <u>0.806<math>\pm</math>0.047</u> | <u>0.641<math>\pm</math>0.157</u> | 18.320 $\pm$ 11.767                |
| CT+Meta.   | <u>0.538<math>\pm</math>0.072</u> | 15.102 $\pm$ 1.220                 | 0.697 $\pm$ 0.043                 | 0.669 $\pm$ 0.075                 | <u>10.860<math>\pm</math>9.531</u> | 0.615 $\pm$ 0.115                 | 15.606 $\pm$ 2.526                 | 0.710 $\pm$ 0.046                 | 0.611 $\pm$ 0.164                 | <b>10.412<math>\pm</math>6.587</b> |
| CT+Fusion. | 0.338 $\pm$ 0.055                 | 11.909 $\pm$ 0.694                 | 0.737 $\pm$ 0.032                 | 0.683 $\pm$ 0.066                 | 29.446 $\pm$ 7.250                 | 0.516 $\pm$ 0.106                 | 10.950 $\pm$ 0.807                 | 0.774 $\pm$ 0.034                 | 0.612 $\pm$ 0.150                 | 21.178 $\pm$ 9.460                 |
| CT+Prompt. | 0.381 $\pm$ 0.085                 | 15.212 $\pm$ 1.313                 | 0.787 $\pm$ 0.026                 | 0.681 $\pm$ 0.066                 | 48.286 $\pm$ 12.442                | 0.522 $\pm$ 0.146                 | 15.881 $\pm$ 2.447                 | 0.784 $\pm$ 0.041                 | 0.521 $\pm$ 0.244                 | 21.837 $\pm$ 12.825                |
| DRMF       | 0.475 $\pm$ 0.047                 | 12.417 $\pm$ 1.377                 | 0.722 $\pm$ 0.021                 | 0.567 $\pm$ 0.137                 | 21.406 $\pm$ 10.566                | <u>0.649<math>\pm</math>0.100</u> | 14.147 $\pm$ 2.561                 | 0.790 $\pm$ 0.026                 | 0.423 $\pm$ 0.246                 | 43.180 $\pm$ 17.155                |
| Text-IF    | 0.400 $\pm$ 0.054                 | 13.953 $\pm$ 1.119                 | 0.759 $\pm$ 0.019                 | 0.665 $\pm$ 0.060                 | 27.072 $\pm$ 9.667                 | 0.554 $\pm$ 0.128                 | 15.941 $\pm$ 2.402                 | 0.780 $\pm$ 0.037                 | 0.496 $\pm$ 0.235                 | 18.908 $\pm$ 10.381                |
| OmniFuse   | 0.479 $\pm$ 0.052                 | 12.859 $\pm$ 0.775                 | 0.741 $\pm$ 0.017                 | 0.665 $\pm$ 0.074                 | 39.696 $\pm$ 4.430                 | 0.585 $\pm$ 0.104                 | 15.221 $\pm$ 2.089                 | 0.767 $\pm$ 0.033                 | 0.523 $\pm$ 0.213                 | 44.850 $\pm$ 7.131                 |
| Deno-IF    | <b>0.633<math>\pm</math>0.052</b> | <b>16.528<math>\pm</math>1.370</b> | <u>0.805<math>\pm</math>0.022</u> | <b>0.738<math>\pm</math>0.065</b> | <b>9.280<math>\pm</math>5.918</b>  | <b>0.656<math>\pm</math>0.075</b> | <b>16.525<math>\pm</math>2.800</b> | <b>0.813<math>\pm</math>0.026</b> | <b>0.648<math>\pm</math>0.155</b> | 16.514 $\pm$ 8.460                 |

In the joint denoising and fusion module, I2Former is updated with Adam Optimizer with batch size as 4 and epoch as 80. Learning rate is  $2e-4$  with exponential decay. Hyper-parameters are  $\lambda = 1e3$ ,  $\eta = 30$ .  $\kappa$  increases during training and equals the multiplication of  $1e-6$  and epoch. Numbers of blocks in Fig. 2 are  $L_r, L_3 = 2, L_1, L_2 = 4$ . Experiments are conducted on an NVIDIA 3090 GPU.

**Comparison Results.** We consider two typical types of noise (additive and multiplicative noise) during imaging process, *e.g.*, Gaussian and speckle noise in source images. For noisy source images, we compare Deno-IF with both the combination of SOTA denoising and fusion methods and degradation-aware image fusion methods. As some fusion methods (TarDAL [9], CDDFuse [49], DDFM [50], MetaFusion [48], FusionBooster [1], and PromptFusion [10]) cannot handle noise, CTNet [25] (supervised denoising method) is used for pre-denoising before these fusion methods are performed. The rationale of selecting CTNet for comparison is two-fold: i) validating the unsupervised generalization of the proposed method against the supervised denoising performance of CTNet; ii) validating the improvements of architecturally relevant frameworks and objectives as CTNet also designs a Transformer-based network. The supervised degradation-aware fusion methods (DRMF [22], Text-IF [36], and OmniFuse [44]) directly fuse noisy source images.

**Qualitative Results.** Qualitative results where both source images suffer from various-level Gaussian or speckle noise across two datasets are shown in Fig. 3. First, as Deno-IF jointly realizes denoising and fusion, it avoids the residual noise in competitors caused by the limited pre-denoising performance. Second, as an unsupervised method, Deno-IF can infer clean data from source images which both suffer from noise and present clear fused images. On the other hand, it shows robust

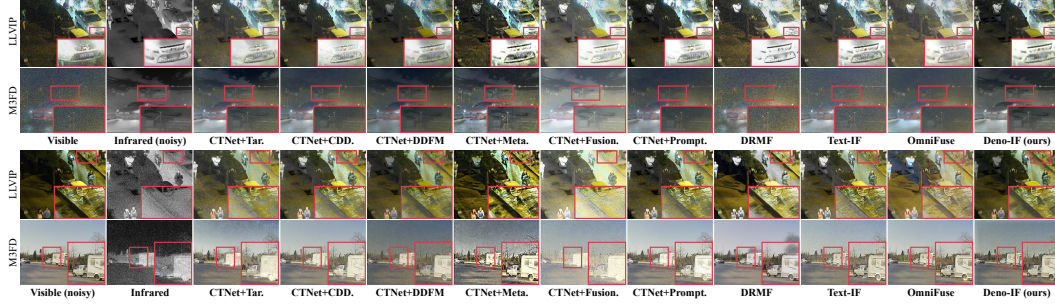


Figure 4: Qualitative results when only one source image is subject to various-level noise.

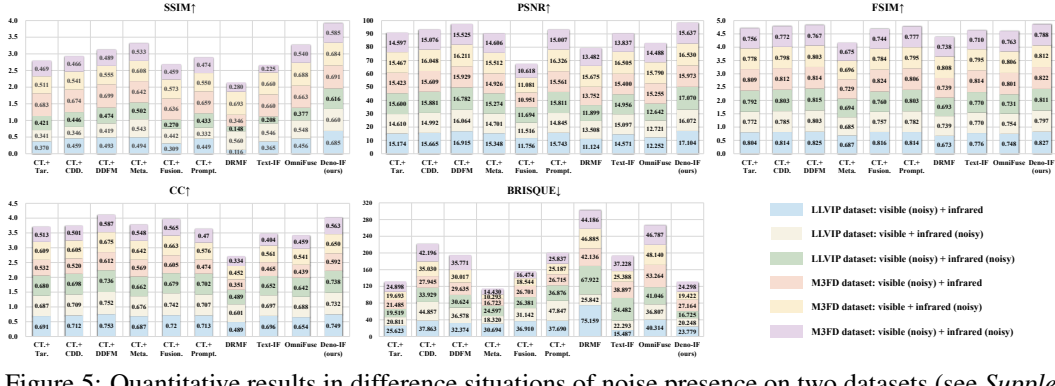


Figure 5: Quantitative results in difference situations of noise presence on two datasets (see *Supplementary Material* for results in specific situations).

performance across various noise types and levels, addressing the low generalization of existing supervised denoising and fusion methods. Finally, the inference on clean data in Deno-IF is based on convolutional low-rankness, avoiding excessive distortion of meaningful scene content in competitors.

**Quantitative Results.** Traditional fusion metrics fall into similarity-based metrics (evaluating similarity between fused and source images) and statistics-based metrics (measuring image characteristics, e.g., SD, SF, EN, etc.). However, when fusing noisy images, statistics-based metrics may produce misleading results as noise may artificially inflate these values (e.g., higher EN/SF caused by noise rather than meaningful information). Thus, in this work, we prioritize similarity-based full-reference metrics (*i.e.*, SSIM [28], PSNR, feature similarity index (FSIM) [45] and correlation coefficient (CC)) and a quality-based no-reference metric BRISQUE [14] to evaluate denoising and fusion performances on 30 image pairs.

As reported in Tab. 1, Deno-IF achieves optimal performances on most metrics. The results on full-reference metrics indicate that our results show high similarity with clean source images. The results on the non-reference metric indicate that our results have less distortion. Finally, the comprehensive results and small variances indicate the generalization of Deno-IF for various scenes, noise types, and noise levels.

**Validation on Single-Modal Noise.** To validate the generalization under different noise presence situations especially specific-modal noise, we perform experiments where only one source image suffers from noise. We take source images with Gaussian noise for example and qualitative results are shown in Fig. 4. On the one hand, Deno-IF effectively removes noise in infrared or visible modalities without being affected by modal differences, avoiding residual noise in results. On the other hand, it indicates the generalization of Deno-IF in different noise presence situations. As reported in the quantitative results of each situation in Fig. 5, Deno-IF shows the optimal results in SSIM, PSNR and FSIM, and suboptimal results in CC and BRISQUE in total amount. From the perspective of variance across situations and datasets, Deno-IF achieves advantageous and balanced performances.

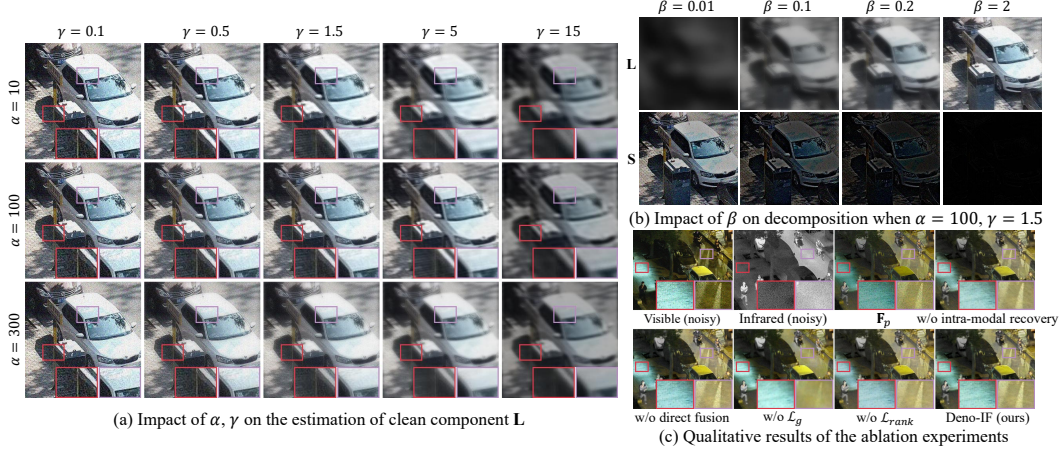


Figure 6: Impacts of decomposition parameter settings and results of ablation experiments.

Table 2: Quantitative results of ablation study and performances on dealing with different-level noise.

| Metrics                  | SSIM $\uparrow$ | PSNR $\uparrow$ | FSIM $\uparrow$ | CC $\uparrow$ | PSNR $\uparrow$ |        |        |        |        | SSIM $\uparrow$ |       |       |       |       | FSIM $\uparrow$ |       |       |       |       | CC $\uparrow$ |       |        |        |        | BRISQUE $\downarrow$ |      |       |      |       |
|--------------------------|-----------------|-----------------|-----------------|---------------|-----------------|--------|--------|--------|--------|-----------------|-------|-------|-------|-------|-----------------|-------|-------|-------|-------|---------------|-------|--------|--------|--------|----------------------|------|-------|------|-------|
|                          |                 |                 |                 |               | $\sigma$        | DRMF   | Text.  | Omn.   | Deno.  | DRMF            | Text. | Omn.  | Deno. | DRMF  | Text.           | Omn.  | Deno. | DRMF  | Text. | Omn.          | Deno. | DRMF   | Text.  | Omn.   | Deno.                | DRMF | Text. | Omn. | Deno. |
| w/o intra-modal          | 0.606           | 16.960          | 0.799           | 0.730         |                 |        |        |        |        |                 |       |       |       |       |                 |       |       |       |       |               |       |        |        |        |                      |      |       |      |       |
| w/o direct fusion        | 0.613           | 16.965          | 0.797           | 0.730         | 10              | 11.997 | 16.141 | 12.522 | 16.878 | 0.309           | 0.431 | 0.484 | 0.675 | 0.716 | 0.814           | 0.747 | 0.818 | 0.557 | 0.702 | 0.667         | 0.740 | 26.855 | 33.060 | 41.297 | 15.364               |      |       |      |       |
|                          |                 |                 |                 |               | 20              | 11.874 | 15.736 | 12.551 | 17.097 | 0.182           | 0.239 | 0.427 | 0.650 | 0.705 | 0.793           | 0.738 | 0.816 | 0.528 | 0.681 | 0.657         | 0.738 | 58.717 | 45.843 | 41.610 | 16.493               |      |       |      |       |
| w/o $\mathcal{L}_g$      | 0.600           | 16.160          | 0.789           | 0.685         | 30              | 11.949 | 14.433 | 12.701 | 17.199 | 0.170           | 0.148 | 0.384 | 0.632 | 0.700 | 0.737           | 0.732 | 0.813 | 0.510 | 0.629 | 0.648         | 0.735 | 66.016 | 58.087 | 40.411 | 17.989               |      |       |      |       |
| w/o $\mathcal{L}_{rank}$ | 0.611           | 16.554          | 0.802           | 0.727         | 40              | 11.935 | 14.182 | 12.803 | 17.262 | 0.138           | 0.109 | 0.350 | 0.617 | 0.688 | 0.725           | 0.728 | 0.809 | 0.467 | 0.604 | 0.636         | 0.733 | 74.548 | 61.441 | 38.531 | 20.514               |      |       |      |       |
| Deno-IF (ours)           | <b>0.616</b>    | <b>17.070</b>   | <b>0.811</b>    | <b>0.738</b>  | 50              | 11.719 | 13.477 | 12.848 | 17.242 | 0.106           | 0.084 | 0.318 | 0.602 | 0.675 | 0.693           | 0.722 | 0.805 | 0.423 | 0.562 | 0.624         | 0.732 | 75.814 | 66.258 | 35.604 | 24.256               |      |       |      |       |

(a) Results of ablation study.

(b) Results on dealing with different-level noise.

**Optimization Parameter Settings.** We validate impacts of  $\alpha, \beta, \gamma$  in Eq. (5) on the estimation of clean component  $\mathbf{L}$ . We first analysis impacts of  $\alpha, \gamma$  and then analyze the impact of  $\beta$  with fixed  $\alpha, \gamma$ . Taking visible data for example, when  $\beta = 2$ , Fig. 6a shows results with varying  $\alpha, \gamma$ .  $\alpha$  constrains the Frobenius norm of  $\mathbf{M}, \mathbf{N}$ .  $\gamma$  constrains similarity between  $\mathcal{A}_k(\mathbf{L})$  and  $\mathbf{MN}$ . Small  $\gamma$  fails to ensure similarity. Regardless of varying  $\alpha$ ,  $\mathbf{L}$  tends to be noisy. As  $\gamma$  increases,  $\mathcal{A}_k(\mathbf{L})$  tends to be low-rank for clarify. Excessively large  $\gamma$  disrupts the similarity between  $\mathbf{L} + \mathbf{S}$  and noisy image, leading to brightness and content distortions. With appropriate  $\gamma$ , a small  $\alpha$  leads to texture distortion. A large  $\alpha$  leads to smoothing and hazy appearance. By carefully tuning  $\alpha$ , the details can be adjusted for balance. Then, the decomposition results with varying  $\beta$  are shown in Fig. 6b. When  $\beta$  is small, the constraint on  $\mathbf{S}$  is weak and many contents are mistakenly decomposed into  $\mathbf{S}$ . When  $\beta$  is large enough,  $\mathbf{S}$  contains little information, resulting in residual noise in  $\mathbf{L}$ .

**Ablation Study.** We validate some architecture designs and loss functions. For architectures, we consider: i) intra-modal recovery in I2Former. We remove it and modify I2Former to a Restormer-like structure with similar parameters. ii) direct fusion. Rather than learning a residual image to noisy inputs, I2Former directly generates fused image. It learns a residual image  $\mathbf{F}_r$  to a pre-fused image  $\mathbf{F}_p$  as fused image  $\mathbf{F} = \mathbf{F}_r + \mathbf{F}_p$ , where  $\mathbf{F}_p = c(\max(\mathbf{V}^y, \mathbf{R}), \mathbf{V}^{Cb}, \mathbf{V}^{Cr})$ . For loss functions, we remove the gradient loss  $\mathcal{L}_g$  and regularization term  $\mathcal{L}_{rank}$ . As in Fig. 6c,  $\mathbf{F}_p$  contains noise. Without intra-modal recovery, the result shows diagonal effects. If learning a residual image to  $\mathbf{F}_p$ , the unstable training process results in slight intensity and color casts. The lack of  $\mathcal{L}_g$  results in blur contents. Without  $\mathcal{L}_{rank}$ , the result cannot infer more contents according to convolutional low-rank property. Deno-IF generates clear results with more contents. The results in Tab. 2a are also consistent with qualitative results. Without  $\mathcal{L}_g$ , it shows the worst performance. By introducing intra-modal recovery, direct fusion, and  $\mathcal{L}_{rank}$ , our method achieves the optimal performance among all the settings.

**Performances on Different-Level Noise.** Experiments on different-level noise Gaussian noise (as a representative) are reported in Tab. 2b. As  $\sigma$  increases, the performances of the competitors significantly decrease on most metrics. Although it is an increasingly challenging trend for recovering under more severe noise, our method shows smaller fluctuations, showing its robustness.

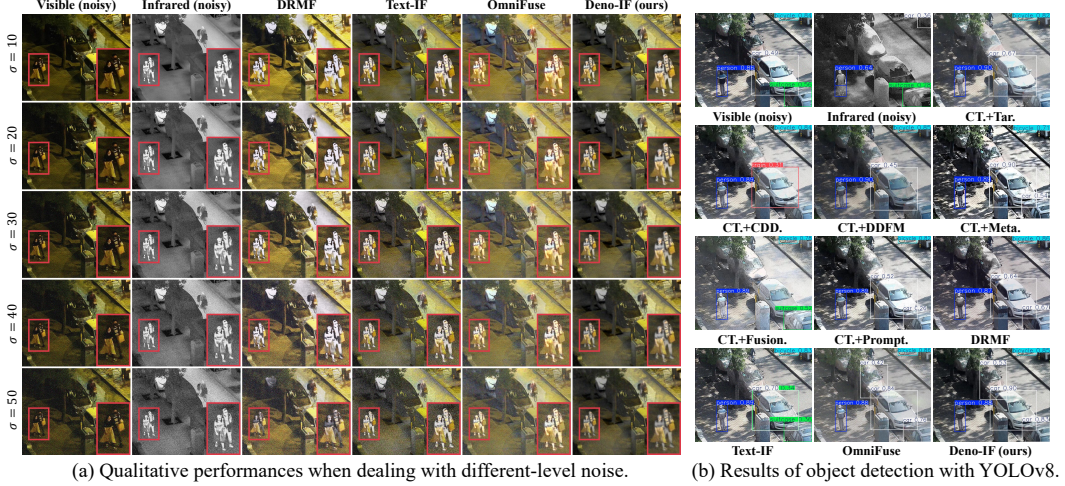


Figure 7: Results of different-level noise and external verification on high-level task.

Table 3: Computational efficiency comparison of different methods.

| Methods     | Tar. | CDD. | DDFM   | Meta. | Fusion. | Prompt. | CT.+Tar.     | CT.+CDD. | CT.+DDFM | CT.+Meta. | CT.+Fusion. | CT.+Prompt. | DRMF   | Text-IF     | Omn.  | Deno-IF      |
|-------------|------|------|--------|-------|---------|---------|--------------|----------|----------|-----------|-------------|-------------|--------|-------------|-------|--------------|
| Para (M)    | 0.30 | 1.19 | 552.66 | 0.81  | 0.56    | 7.44    | <u>51.71</u> | 52.60    | 604.07   | 52.22     | 51.97       | 58.91       | 170.98 | 215.12      | 98.15 | <b>1.433</b> |
| FLOPs (G)   | 91   | 548  | 5221   | 159   | 241     | 36877   | 41695        | 42152    | 46824    | 41763     | 41845       | 78481       | 4576   | <u>1519</u> | 5471  | <b>115</b>   |
| Runtime (s) | 0.03 | 0.23 | 68.79  | 0.04  | 2.49    | 1.08    | 13.50        | 13.70    | 82.26    | 13.51     | 15.96       | 14.55       | 4.47   | <u>0.31</u> | 4.53  | <b>0.06</b>  |

**External Verification on High-Level Task.** The denoising and fusion performances are externally validated by a subsequent high-level task—object detection. In Fig. 7b, it is challenging to accurately detect some targets in a single noisy visible/infrared image. By denoising and fusing noisy source images, some competitors can boost detection accuracy while some methods exhibit notable misclassification. Our result eliminates noise interference and preserve essential information, improving both detection accuracy and precision. Quantitative results are reported in *Supplementary Material*.

**Efficiency Comparison.** The efficiency is evaluated by parameter numbers, FLOPs, and runtime. Due to the large parameters, high computational complexity, and runtime of denoising method CTNet, we also report the efficiency of pure fusion methods. As reported in Tab. 3, our method shows higher efficiency than the competitors. It also achieves simultaneous denoising and fusion with an efficiency that is comparable or even superior to those of pure fusion networks.

## 5 Conclusion

This paper proposes an unsupervised noisy visible and infrared image fusion method, addressing the limitations of existing methods that cannot handle noise interference or rely on paired data restricted by noise characteristics. The proposed network, featuring a convolutional low-rank optimization module and a joint denoising and fusion module, eliminates data dependency and improves generalization while reducing complexity. Experimental results demonstrate its effectiveness and robustness.

## Acknowledgments and Disclosure of Funding

This work was supported by the National Natural Science Foundation of China (NSFC) under Grant U21B2027 and 62502085, the Natural Science Foundation of Jiangsu Province of China under Grant BK20241280, and the Start-up Research Fund of Southeast University under Grant RF1028623006 and RF1028624061.

## References

- [1] C. Cheng, T. Xu, X.-J. Wu, H. Li, X. Li, and J. Kittler. Fusionbooster: A unified image fusion boosting paradigm. *International Journal of Computer Vision*, 133(5):3041–3058, 2025.
- [2] F. Duffhauss, N. A. Vien, H. Ziesche, and G. Neumann. Fusionvae: A deep hierarchical variational autoencoder for rgb image fusion. In *Proceedings of the European Conference on Computer Vision*, pages 674–691. Springer, 2022.
- [3] Y. Gong, J. Lu, W. Liu, Z. Li, X. Jiang, X. Gao, and X. Wu. Sifdrivenet: Speed and image fusion for driving behavior classification network. *IEEE Transactions on Computational Social Systems*, 2023.
- [4] H. Huang, W. He, H. Zhang, Y. Xia, and L. Zhang. Stfdiff: Remote sensing image spatiotemporal fusion with diffusion models. *Information Fusion*, page 102505, 2024.
- [5] X. Jia, C. Zhu, M. Li, W. Tang, and W. Zhou. Llvip: A visible-infrared paired dataset for low-light vision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3496–3504, 2021.
- [6] H. Li, X.-J. Wu, and J. Kittler. Rfn-nest: An end-to-end residual fusion network for infrared and visible images. *Information Fusion*, 73:72–86, 2021.
- [7] X. Li, X. Li, H. Tan, and J. Li. Samf: small-area-aware multi-focus image fusion for object detection. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3845–3849. IEEE, 2024.
- [8] G. Liu and W. Zhang. Recovery of future data via convolution nuclear norm minimization. *IEEE Transactions on Information Theory*, 69(1):650–665, 2022.
- [9] J. Liu, X. Fan, Z. Huang, G. Wu, R. Liu, W. Zhong, and Z. Luo. 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*, pages 5802–5811, 2022.
- [10] J. Liu, X. Li, Z. Wang, Z. Jiang, W. Zhong, W. Fan, and B. Xu. Promptfusion: Harmonized semantic prompt learning for infrared and visible image fusion. *IEEE/CAA Journal of Automatica Sinica*, 12(3):502–515, 2025.
- [11] J. Ma, H. Xu, J. Jiang, X. Mei, and X.-P. Zhang. Ddcgan: A dual-discriminator conditional generative adversarial network for multi-resolution image fusion. *IEEE Transactions on Image Processing*, 29:4980–4995, 2020.
- [12] J. Ma, H. Zhang, Z. Shao, P. Liang, and H. Xu. Ganmcc: A generative adversarial network with multiclassification constraints for infrared and visible image fusion. *IEEE Transactions on Instrumentation and Measurement*, 70:1–14, 2020.
- [13] J. Ma, L. Tang, F. Fan, J. Huang, X. Mei, and Y. Ma. Swinfusion: Cross-domain long-range learning for general image fusion via swin transformer. *IEEE/CAA Journal of Automatica Sinica*, 9(7):1200–1217, 2022.
- [14] A. Mittal, A. K. Moorthy, and A. C. Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, 21(12):4695–4708, 2012.
- [15] S. Peng, X. Zhu, H. Deng, L.-J. Deng, and Z. Lei. Fusionmamba: Efficient remote sensing image fusion with state space model. *IEEE Transactions on Geoscience and Remote Sensing*, 2024.
- [16] L. Qu, S. Liu, M. Wang, and Z. Song. Transmef: A transformer-based multi-exposure image fusion framework using self-supervised multi-task learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2126–2134, 2022.
- [17] D. Rao, T. Xu, and X.-J. Wu. Tgfuse: An infrared and visible image fusion approach based on transformer and generative adversarial network. *IEEE Transactions on Image Processing*, 2023.

- [18] Y. Rao, D. Wu, M. Han, T. Wang, Y. Yang, T. Lei, C. Zhou, H. Bai, and L. Xing. At-gan: A generative adversarial network with attention and transition for infrared and visible image fusion. *Information Fusion*, 92:336–349, 2023.
- [19] P. Sprechmann, A. M. Bronstein, and G. Sapiro. Learning efficient sparse and low rank models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(9):1821–1833, 2015.
- [20] L. Tang, J. Yuan, and J. Ma. 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, 2022.
- [21] L. Tang, J. Yuan, H. Zhang, X. Jiang, and J. Ma. Piafusion: A progressive infrared and visible image fusion network based on illumination aware. *Information Fusion*, 83:79–92, 2022.
- [22] L. Tang, Y. Deng, X. Yi, Q. Yan, Y. Yuan, and J. Ma. Drmf: Degradation-robust multi-modal image fusion via composable diffusion prior. In *Proceedings of the ACM International Conference on Multimedia*, pages 8546–8555, 2024.
- [23] L. Tang, C. Li, and J. Ma. Mask-difuser: A masked diffusion model for unified unsupervised image fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.
- [24] W. Tang, F. He, Y. Liu, Y. Duan, and T. Si. Datfuse: Infrared and visible image fusion via dual attention transformer. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(7):3159–3172, 2023.
- [25] C. Tian, M. Zheng, W. Zuo, S. Zhang, Y. Zhang, and C.-W. Lin. A cross transformer for image denoising. *Information Fusion*, 102:102043, 2024.
- [26] D. Wang, J. Liu, R. Liu, and X. Fan. An interactively reinforced paradigm for joint infrared-visible image fusion and saliency object detection. *Information Fusion*, 98:101828, 2023.
- [27] Y. Wang, X. Geng, W. Huang, B. Huang, and M. Gong. Generator identification for linear sdes with additive and multiplicative noise. *Advances in Neural Information Processing Systems*, 36:64103–64138, 2023.
- [28] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.
- [29] X. Xie, Y. Cui, T. Tan, X. Zheng, and Z. Yu. Fusionmamba: Dynamic feature enhancement for multimodal image fusion with mamba. *Visual Intelligence*, 2(1):37, 2024.
- [30] H. Xu, H. Zhang, and J. Ma. Classification saliency-based rule for visible and infrared image fusion. *IEEE Transactions on Computational Imaging*, 7:824–836, 2021.
- [31] H. Xu, J. Ma, J. Jiang, X. Guo, and H. Ling. U2fusion: A unified unsupervised image fusion network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1):502–518, 2022.
- [32] H. Xu, J. Yuan, and J. Ma. Murf: Mutually reinforcing multi-modal image registration and fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):12148–12166, 2023.
- [33] H. Xu, X. Yi, C. Lu, G. Liu, and J. Ma. Urfusion: Unsupervised unified degradation-robust image fusion network. *IEEE Transactions on Image Processing*, 34:5803–5818, 2025.
- [34] B. Yang, Z. Jiang, D. Pan, H. Yu, G. Gui, and W. Gui. Lfdt-fusion: a latent feature-guided diffusion transformer model for general image fusion. *Information Fusion*, 113:102639, 2025.
- [35] X. Yi, L. Tang, H. Zhang, H. Xu, and J. Ma. Diff-if: Multi-modality image fusion via diffusion model with fusion knowledge prior. *Information Fusion*, 110:102450, 2024.
- [36] X. Yi, H. Xu, H. Zhang, L. Tang, and J. Ma. 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*, pages 27026–27035, 2024.

- [37] X. Yi, Y. Ma, Y. Li, H. Xu, and J. Ma. Artificial intelligence facilitates information fusion for perception in complex environments. *The Innovation*, 2025.
- [38] J. Yue, L. Fang, S. Xia, Y. Deng, and J. Ma. Dif-fusion: Towards high color fidelity in infrared and visible image fusion with diffusion models. *IEEE Transactions on Image Processing*, 2023.
- [39] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and M.-H. Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5728–5739, 2022.
- [40] G. Zhang, R. Nie, and J. Cao. Ssl-waeie: Self-supervised learning with weighted auto-encoding and information exchange for infrared and visible image fusion. *IEEE/CAA Journal of Automatica Sinica*, 9(9):1694–1697, 2022.
- [41] H. Zhang, H. Xu, Y. Xiao, X. Guo, and J. Ma. Rethinking the image fusion: A fast unified image fusion network based on proportional maintenance of gradient and intensity. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12797–12804, 2020.
- [42] H. Zhang, J. Yuan, X. Tian, and J. Ma. Gan-fm: Infrared and visible image fusion using gan with full-scale skip connection and dual markovian discriminators. *IEEE Transactions on Computational Imaging*, 7:1134–1147, 2021.
- [43] H. Zhang, X. Zuo, H. Zhou, T. Lu, and J. Ma. A robust mutual-reinforcing framework for 3d multi-modal medical image fusion based on visual-semantic consistency. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 7087–7095, 2024.
- [44] H. Zhang, L. Cao, X. Zuo, Z. Shao, and J. Ma. Omnifuse: Composite degradation-robust image fusion with language-driven semantics. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(9):7577–7595, 2025.
- [45] L. Zhang, L. Zhang, X. Mou, and D. Zhang. Fsim: A feature similarity index for image quality assessment. *IEEE Transactions on Image Processing*, 20(8):2378–2386, 2011.
- [46] M. Zhang, K. Yue, B. Li, J. Guo, Y. Li, and X. Gao. Single-frame infrared small target detection via gaussian curvature inspired network. *IEEE Transactions on Geoscience and Remote Sensing*, 62:5005013, 2024.
- [47] X. Zhang, L. Wang, G. Zhang, T. Lan, H. Zhang, L. Zhao, J. Li, L. Zhu, and H. Liu. Ri-fusion: 3d object detection using enhanced point features with range-image fusion for autonomous driving. *IEEE Transactions on Instrumentation and Measurement*, 72:1–13, 2022.
- [48] W. Zhao, S. Xie, F. Zhao, Y. He, and H. Lu. Metafusion: Infrared and visible image fusion via meta-feature embedding from object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13955–13965, 2023.
- [49] Z. Zhao, H. Bai, J. Zhang, Y. Zhang, S. Xu, Z. Lin, R. Timofte, and L. Van Gool. Cddfuse: Correlation-driven dual-branch feature decomposition for multi-modality image fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5906–5916, 2023.
- [50] Z. Zhao, H. Bai, Y. Zhu, J. Zhang, S. Xu, Y. Zhang, K. Zhang, D. Meng, R. Timofte, and L. Van Gool. Ddfm: denoising diffusion model for multi-modality image fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8082–8093, 2023.
- [51] A. Zheng, J. Liu, Z. Wang, L. Huang, C. Li, and B. Yin. Visible-infrared person re-identification via specific and shared representations learning. *Visual Intelligence*, 1(1):29, 2023.

## Technical Appendices and Supplementary Material

### 1. Details of Transformer Block and Variation of Losses under Different Network Settings

The network architecture of the Transformer Block in Fig. 2 is shown in Fig. 8a, where the left part is multi-dconv head transposed attention and the right part is the gated dconv feed-forward network.

In the ablation study introduced in Sec. 4, we perform the experiments by removing the intra-modal recovery and learning a residual image to a pre-fused image (w/o direct fusion). The variation of losses under these settings during the training phase is shown in Fig. 8b. By comparison, the I2Former in Deno-IF achieves both faster convergence speed and higher convergence accuracy.

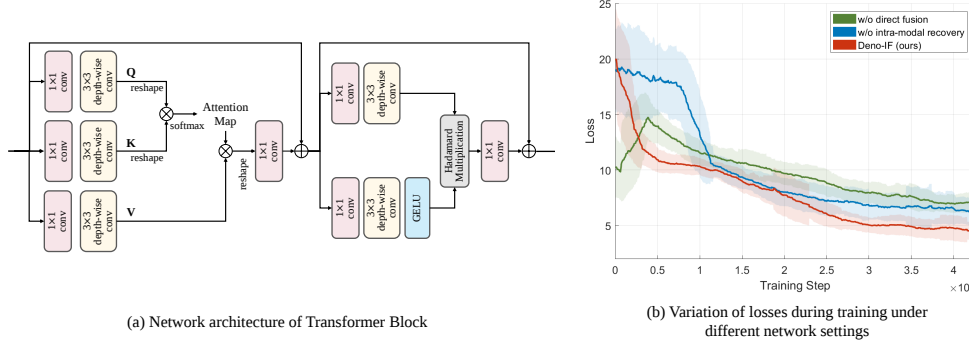


Figure 8: Details of Transformer block and variation of losses under different network settings.

### 2. Quantitative Results of Generalization in More Situations of Noise Presence

The quantitative results on five metrics across two datasets when fusing i) noisy visible and clean infrared images; ii) clean visible and noisy infrared images are reported in Tabs. 4 and 5, respectively. In these situations, our method can still achieve optimal or comparable performances, indicating its generalization for various situation of noise presence.

Table 4: Quantitative results when fusing noisy visible and clean infrared images (**bold**: optimal, underline: suboptimal).

| Datasets    | LLVIP                             |                                    |                                   |                                   |                                    | M3FD                              |                                    |                                   |                                   |                                    |
|-------------|-----------------------------------|------------------------------------|-----------------------------------|-----------------------------------|------------------------------------|-----------------------------------|------------------------------------|-----------------------------------|-----------------------------------|------------------------------------|
|             | SSIM $\uparrow$                   | PSNR $\uparrow$                    | FSIM $\uparrow$                   | CC $\uparrow$                     | BRISQUE $\downarrow$               | SSIM $\uparrow$                   | PSNR $\uparrow$                    | FSIM $\uparrow$                   | CC $\uparrow$                     | BRISQUE $\downarrow$               |
| CT.+Tar.    | 0.370 $\pm$ 0.133                 | 15.174 $\pm$ 1.444                 | 0.804 $\pm$ 0.028                 | 0.691 $\pm$ 0.073                 | 25.623 $\pm$ 8.442                 | 0.683 $\pm$ 0.111                 | 15.423 $\pm$ 2.617                 | 0.809 $\pm$ 0.034                 | 0.532 $\pm$ 0.199                 | 21.485 $\pm$ 5.250                 |
| CT.+CDD.    | 0.459 $\pm$ 0.150                 | 15.665 $\pm$ 1.510                 | 0.814 $\pm$ 0.026                 | 0.712 $\pm$ 0.068                 | 37.863 $\pm$ 19.525                | 0.674 $\pm$ 0.125                 | 15.609 $\pm$ 2.704                 | 0.812 $\pm$ 0.043                 | 0.520 $\pm$ 0.224                 | 27.945 $\pm$ 6.977                 |
| CT.+DDFM    | 0.493 $\pm$ 0.112                 | <u>16.915<math>\pm</math>1.353</u> | <u>0.825<math>\pm</math>0.020</u> | <b>0.753<math>\pm</math>0.068</b> | 32.374 $\pm$ 15.354                | <b>0.699<math>\pm</math>0.111</b> | <u>15.929<math>\pm</math>2.368</u> | 0.814 $\pm$ 0.056                 | <b>0.612<math>\pm</math>0.189</b> | 29.635 $\pm$ 6.150                 |
| CT.+Meta.   | <u>0.494<math>\pm</math>0.072</u> | 15.348 $\pm$ 1.285                 | 0.687 $\pm$ 0.033                 | 0.687 $\pm$ 0.078                 | 30.694 $\pm$ 7.736                 | 0.642 $\pm$ 0.114                 | 14.926 $\pm$ 2.402                 | 0.729 $\pm$ 0.052                 | 0.569 $\pm$ 0.185                 | <b>16.723<math>\pm</math>9.889</b> |
| CT.+Fusion. | 0.309 $\pm$ 0.100                 | 11.756 $\pm$ 0.663                 | 0.816 $\pm$ 0.020                 | 0.720 $\pm$ 0.075                 | 36.910 $\pm$ 13.231                | 0.636 $\pm$ 0.075                 | 10.951 $\pm$ 0.983                 | <b>0.824<math>\pm</math>0.038</b> | <u>0.605<math>\pm</math>0.151</u> | 26.701 $\pm$ 7.944                 |
| CT.+Prompt. | 0.449 $\pm$ 0.145                 | 15.743 $\pm$ 1.410                 | 0.814 $\pm$ 0.025                 | 0.713 $\pm$ 0.069                 | 37.690 $\pm$ 17.569                | 0.659 $\pm$ 0.128                 | 15.561 $\pm$ 2.593                 | 0.806 $\pm$ 0.045                 | 0.474 $\pm$ 0.243                 | 26.715 $\pm$ 7.489                 |
| DRMF        | 0.116 $\pm$ 0.045                 | 11.124 $\pm$ 0.688                 | 0.673 $\pm$ 0.019                 | 0.489 $\pm$ 0.150                 | 75.159 $\pm$ 4.613                 | 0.346 $\pm$ 0.144                 | 13.752 $\pm$ 2.144                 | 0.739 $\pm$ 0.043                 | 0.351 $\pm$ 0.232                 | 42.136 $\pm$ 15.254                |
| Text-IF     | 0.365 $\pm$ 0.112                 | 14.571 $\pm$ 1.222                 | 0.776 $\pm$ 0.022                 | 0.696 $\pm$ 0.071                 | <b>15.487<math>\pm</math>8.137</b> | 0.660 $\pm$ 0.137                 | 15.400 $\pm$ 3.025                 | 0.814 $\pm$ 0.044                 | 0.465 $\pm$ 0.225                 | 38.897 $\pm$ 8.459                 |
| OmniFuse    | 0.456 $\pm$ 0.061                 | 12.252 $\pm$ 0.852                 | 0.748 $\pm$ 0.014                 | 0.654 $\pm$ 0.093                 | 40.314 $\pm$ 2.856                 | 0.663 $\pm$ 0.101                 | 15.255 $\pm$ 2.486                 | 0.801 $\pm$ 0.033                 | 0.439 $\pm$ 0.237                 | 53.264 $\pm$ 5.029                 |
| Deno-IF     | <b>0.685<math>\pm</math>0.040</b> | <b>17.104<math>\pm</math>1.691</b> | <b>0.827<math>\pm</math>0.022</b> | <u>0.749<math>\pm</math>0.076</u> | 23.779 $\pm$ 4.270                 | <u>0.691<math>\pm</math>0.082</u> | <b>15.973<math>\pm</math>2.905</b> | <u>0.822<math>\pm</math>0.032</u> | 0.592 $\pm$ 0.174                 | 27.164 $\pm$ 5.110                 |

Table 5: Quantitative results when fusing clean visible and noisy infrared images.

| Datasets    | LLVIP                             |                                    |                                   |                                   |                                     | M3FD                              |                                    |                                   |                                   |                                    |
|-------------|-----------------------------------|------------------------------------|-----------------------------------|-----------------------------------|-------------------------------------|-----------------------------------|------------------------------------|-----------------------------------|-----------------------------------|------------------------------------|
|             | SSIM $\uparrow$                   | PSNR $\uparrow$                    | FSIM $\uparrow$                   | CC $\uparrow$                     | BRISQUE $\downarrow$                | SSIM $\uparrow$                   | PSNR $\uparrow$                    | FSIM $\uparrow$                   | CC $\uparrow$                     | BRISQUE $\downarrow$               |
| CT.+Tar.    | 0.341 $\pm$ 0.138                 | 14.610 $\pm$ 1.300                 | 0.772 $\pm$ 0.021                 | 0.687 $\pm$ 0.077                 | 20.811 $\pm$ 8.946                  | 0.511 $\pm$ 0.177                 | 15.467 $\pm$ 2.243                 | 0.778 $\pm$ 0.036                 | 0.609 $\pm$ 0.161                 | 19.693 $\pm$ 9.830                 |
| CT.+CDD.    | 0.346 $\pm$ 0.154                 | 14.992 $\pm$ 1.316                 | 0.785 $\pm$ 0.021                 | 0.709 $\pm$ 0.073                 | 44.857 $\pm$ 11.512                 | 0.541 $\pm$ 0.165                 | 16.048 $\pm$ 2.466                 | 0.798 $\pm$ 0.026                 | 0.605 $\pm$ 0.188                 | 35.030 $\pm$ 17.913                |
| CT.+DDFM    | 0.419 $\pm$ 0.137                 | <u>16.064<math>\pm</math>1.220</u> | <b>0.803<math>\pm</math>0.017</b> | <b>0.752<math>\pm</math>0.073</b> | 36.578 $\pm$ 11.318                 | 0.555 $\pm$ 0.177                 | 16.211 $\pm$ 2.278                 | 0.803 $\pm$ 0.042                 | <b>0.675<math>\pm</math>0.148</b> | 30.017 $\pm$ 15.380                |
| CT.+Meta.   | 0.543 $\pm$ 0.060                 | 14.701 $\pm$ 1.189                 | 0.685 $\pm$ 0.034                 | 0.676 $\pm$ 0.084                 | <b>18.320<math>\pm</math>10.509</b> | 0.608 $\pm$ 0.100                 | 15.512 $\pm$ 2.471                 | 0.696 $\pm$ 0.043                 | 0.642 $\pm$ 0.150                 | <b>10.293<math>\pm</math>7.004</b> |
| CT.+Fusion. | 0.442 $\pm$ 0.062                 | 11.516 $\pm$ 0.652                 | 0.757 $\pm$ 0.029                 | <u>0.742<math>\pm</math>0.072</u> | 31.142 $\pm$ 5.761                  | 0.573 $\pm$ 0.083                 | 11.081 $\pm$ 0.835                 | 0.784 $\pm$ 0.037                 | <u>0.663<math>\pm</math>0.130</u> | <u>18.544<math>\pm</math>5.514</u> |
| CT.+Prompt. | 0.332 $\pm$ 0.146                 | 14.845 $\pm$ 1.311                 | 0.782 $\pm$ 0.022                 | 0.707 $\pm$ 0.073                 | 47.847 $\pm$ 12.074                 | 0.550 $\pm$ 0.143                 | 16.326 $\pm$ 2.311                 | 0.795 $\pm$ 0.026                 | 0.576 $\pm$ 0.208                 | 25.187 $\pm$ 14.365                |
| DRMF        | <u>0.560<math>\pm</math>0.055</u> | 13.508 $\pm$ 1.760                 | 0.739 $\pm$ 0.027                 | 0.601 $\pm$ 0.133                 | 25.842 $\pm$ 7.661                  | <b>0.693<math>\pm</math>0.078</b> | 15.675 $\pm$ 2.150                 | <u>0.808<math>\pm</math>0.026</u> | 0.452 $\pm$ 0.221                 | 46.885 $\pm$ 10.449                |
| Text-IF     | 0.546 $\pm$ 0.089                 | 15.097 $\pm$ 1.355                 | 0.770 $\pm$ 0.026                 | 0.697 $\pm$ 0.066                 | 22.293 $\pm$ 7.881                  | 0.660 $\pm$ 0.113                 | 16.505 $\pm$ 2.464                 | 0.795 $\pm$ 0.032                 | 0.561 $\pm$ 0.211                 | 25.388 $\pm$ 6.248                 |
| OmniFuse    | 0.548 $\pm$ 0.050                 | 12.721 $\pm$ 0.912                 | 0.754 $\pm$ 0.021                 | 0.688 $\pm$ 0.076                 | 36.807 $\pm$ 4.436                  | <u>0.688<math>\pm</math>0.091</u> | 15.790 $\pm$ 2.374                 | 0.806 $\pm$ 0.023                 | 0.541 $\pm$ 0.215                 | 48.140 $\pm$ 4.642                 |
| Deno-IF     | <b>0.660<math>\pm</math>0.052</b> | <b>16.072<math>\pm</math>1.183</b> | <u>0.797<math>\pm</math>0.017</u> | 0.732 $\pm$ 0.064                 | 20.248 $\pm$ 7.930                  | 0.684 $\pm$ 0.072                 | <b>16.530<math>\pm</math>2.496</b> | <b>0.812<math>\pm</math>0.023</b> | 0.650 $\pm$ 0.172                 | 19.422 $\pm$ 9.141                 |

### 3. Results on the Original Clean Visible and Infrared Image Fusion Task

We conduct the experiment to fuse the original clean visible and infrared images. The experiment merely compares the fusion performances of Deno-IF and SOTA image fusion methods and verifies whether the proposed method will blur clean data. As shown in Figs. 9-10, our method can preserve the content of two source images in a more balanced manner, thus reflecting more scene contents in the fused image and assisting in human visual perception with comparable contrast.

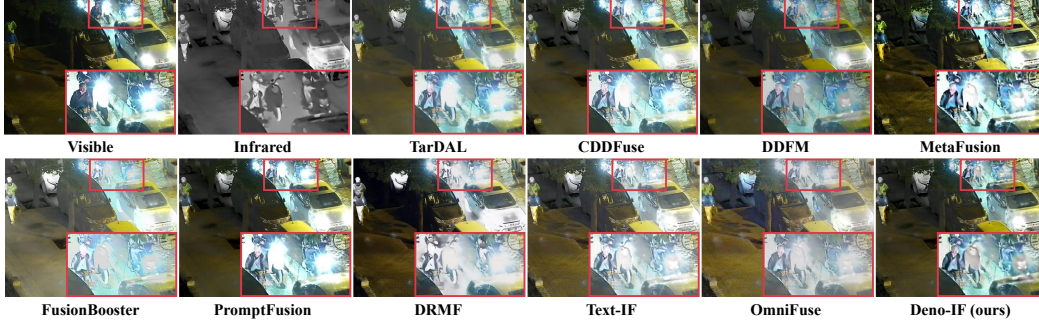


Figure 9: Qualitative results on the LLVIP dataset when fusing clean visible and infrared images.

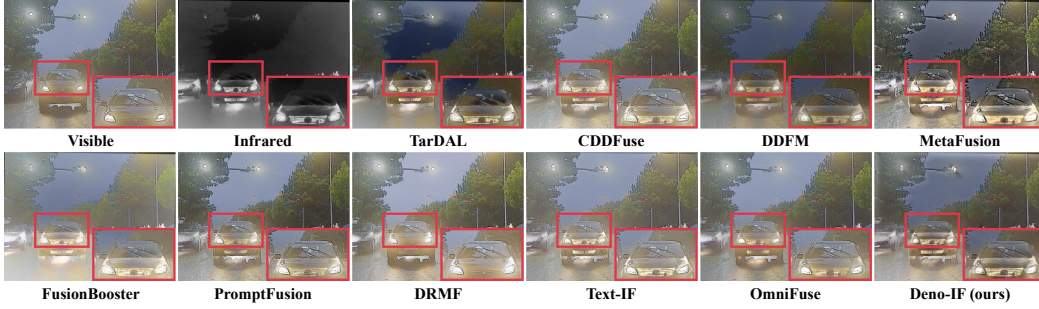


Figure 10: Qualitative results on the M3FD dataset when fusing clean visible and infrared images.

### 4. Quantitative Results of Detection under Noise Interference

Tab. 6 reports the quantitative results of detection on the fusion results of noisy source images. The metrics include precision, recall, average precision (AP), and mean AP (mAP). mAP is the average of all APs at IoU thresholds from 0.5 to 0.95 in steps of 0.05.

Table 6: Object detection results when source images suffer from noise.

|           | CT.+Tar. | CT.+CDD.     | CT.+DDFM     | CT.+Meta.    | CT.+Fusion. | CT.+Prompt. | DRMF  | Text-IF | OmniFuse | Deno-IF      |
|-----------|----------|--------------|--------------|--------------|-------------|-------------|-------|---------|----------|--------------|
| Precision | 0.888    | <u>0.917</u> | 0.877        | <b>0.917</b> | 0.857       | 0.883       | 0.877 | 0.902   | 0.897    | 0.878        |
| Recall    | 0.619    | 0.670        | <u>0.701</u> | 0.653        | 0.657       | 0.645       | 0.601 | 0.688   | 0.697    | <b>0.724</b> |
| AP@0.5    | 0.763    | 0.787        | <b>0.811</b> | 0.772        | 0.758       | 0.769       | 0.722 | 0.786   | 0.802    | <u>0.807</u> |
| mAP       | 0.488    | 0.516        | <u>0.535</u> | 0.496        | 0.489       | 0.496       | 0.437 | 0.507   | 0.525    | <b>0.536</b> |

### 5. Validation of Convergence during Iterative Process

The first module is the convolutional low-rank module, which aims to decompose the noisy input into clean and noisy components. The decomposition problem is rewritten as the energy function in Eq. (5). To validate the convergence during the iterative process, we document several terms of this energy function during iteration.

As summarized in Tab. 7, the iterative process of low-rank optimization module demonstrates convergence. For  $\mathbf{M}$  and  $\mathbf{N}$ , they are initialized as rectangular identity-like matrices (all elements

on the main diagonal are 1 and other elements are 0). As the initialized matrices do not contain much information,  $\|\mathbf{M}\|_F$  and  $\|\mathbf{N}\|_F$  initially increase to incorporate more information, followed by a gradual stabilization as iteration progresses.

Table 7: Changes of various terms in the energy function during the iteration process.

| Iterations                                      | 2        | 4        | 6        | 8       | 10      | 12      | 14      | 16      |
|-------------------------------------------------|----------|----------|----------|---------|---------|---------|---------|---------|
| $\ \mathbf{R} - \mathbf{L} - \mathbf{S}\ _F$    | 227.970  | 87.555   | 44.188   | 33.726  | 31.791  | 31.446  | 31.390  | 31.390  |
| $\ \mathcal{A}_k(\mathbf{L}) - \mathbf{MN}\ _F$ | 3821.812 | 2395.634 | 1220.628 | 940.686 | 889.409 | 880.317 | 878.878 | 878.862 |
| $\ \mathbf{S}\ _F$                              | 0.456    | 0.175    | 0.088    | 0.067   | 0.064   | 0.063   | 0.063   | 0.063   |

From the aspect of network optimization, clean components derived with convolutional low-rank priors serve as i) physics-driven prior injector to regularize fused images, and ii) implicit teachers that enable knowledge distillation of denoising priors into the fusion network. It enables simultaneous noise suppression and information fusion

## 6. Analysis of Computational Training Cost

As the training latency is dominated by the first stage (convolutional low-rank optimization module), we quantify the computational bottleneck within this module. In this module, computational complexity scales primarily with i) convolution kernel size ( $k_1, k_2$ ) and ii) number of iterations ( $T$ ) for solving the low-rank subproblem. The training costs of different settings are reported in Tab. 8. It is observed that increasing  $k_1, k_2$  incurs significant increase in training time while the number of iteration scales linearly with computational cost. Nevertheless, the entire training process remains practical, completing within several days.

Table 8: Changes of various terms in the energy function during the iteration process.

| $k_1, k_2$                | 12    | 18    | 24    | 30    | 36     | 42     |
|---------------------------|-------|-------|-------|-------|--------|--------|
| Training Time (m) / Epoch | 41.31 | 54.35 | 71.51 | 93.28 | 121.22 | 151.68 |
| $T$                       | 10    | 20    | 30    | 40    | 50     | 60     |
| Training Time (m) / Epoch | 20.58 | 30.85 | 41.31 | 51.61 | 61.62  | 71.62  |

## 7. Differences with Prior Low-Rank Methods in Other Domains

While low-rank methods exist in some image processing (e.g., image smoothing, restoration, denoising), the proposed method differs from existing works in task formulation and low-rank theory.

For task formulation, the proposed method differs with prior low-rank methods in input complexity and objective shift. For input complexity, we handle multi-modal data rather than single-modal inputs. For objective, we focus on joint denoising and fusion rather than a single image processing task.

For low-rank theory, the proposed theory differs in energy function formulation and optimization strategy. For energy function, the regularization term is based on convolutional low-rank prior by the convolution nuclear norm. Compared with standard low-rank theory, it can avoid excessive smoothing and structural damage for better local structure preservation, noise robustness, and physical interpretability. For optimization strategy, previous researches directly implement the objective function as the loss of network. Differently, we design a two-phase optimization strategy. In the first phase, we rewrite the energy function and iteratively address the subproblems with analytical solutions. In the second phase, we design a deep network to distill the physics-driven priors and realize denoising and fusion jointly.

## 8. Limitations

The convolutional low-rank optimization module serves as the theoretical cornerstone of the joint denoising and fusion network. It ensures physically-consistent guidance through iterative refinement during training but requires iterative optimization. Although it is not applied in the testing phase, it lengthens the overall training time. In future work, we plan to investigate more efficient optimization schemes to preserve performance while reducing training costs.

## 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 and introduction of the paper clearly state the claims made, including the contributions made in the paper and important assumptions and limitations.

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: [\[Yes\]](#)

Justification: The paper discusses the limitations of the work performed by the authors, providing a balanced view of the study's scope and potential areas for improvement.

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: All the theorems, formulas, and proofs in the paper have been numbered and cross-referenced, and all assumptions have been clearly stated or referenced in the statement of any theorems.

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: The paper has fully disclosed all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and conclusions of the paper.

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: [Yes]

Justification: We will open access to the code after the paper is accepted.

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, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We have specified all the training and test details necessary to understand the results.

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: [Yes]

Justification: The paper reports error bars and other appropriate information about the statistical significance of the experiments, ensuring the reliability and validity of the results.

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: The paper provides sufficient information on the computer resource for experiments.

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: The research conducted in the paper conforms in every respect with the NeurIPS Code of Ethics, ensuring ethical standards are upheld throughout the study.

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: The paper discusses both potential positive societal impacts and negative societal impacts of the work performed, providing a comprehensive evaluation of the study's broader implications.

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: The paper describes safeguards that have been put in place for the responsible release of data or models that have a high risk for misuse, ensuring that appropriate measures are taken to mitigate potential risks.

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: The creators or original owners of assets used in the paper are properly credited, and the license and terms of use are explicitly mentioned and properly respected.

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](https://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: [\[Yes\]](#)

Justification: The new assets introduced in the paper are well documented, and the documentation is provided alongside the 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: [\[NA\]](#)

Justification: The paper does not involve crowdsourcing nor research with human subjects.

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: [\[NA\]](#)

Justification: The paper does not involve crowdsourcing nor research with human subjects.

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.

#### 16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: The core method development in the paper does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.