Degradation-Resistant Unfolding Network for Heterogeneous Image Fusion

Chunming He[1,3,†], Kai Li[2*], Guoxia Xu[3,4,*], Yulun Zhang[5], Runze Hu[6], Zhenhua Guo[7], Xiu Li[1,*]
[1]Shenzhen International Graduate School, Tsinghua University, [2]NEC Laboratories America,
[3]Smart Vision, [4]Nanjing University of Posts and Telecommunications,
[5]ETH Zürich, [6]Beijing Institute of Technology, [7]Tianyi Traffic Technology
{chunminghe19990224, li.gml.kai, gxxu.re, yulun100, hrzlpk2015, cszguo}@gmail.com
li.xiu@sz.tsinghua.edu.cn

Abstract

Heterogeneous image fusion (HIF) techniques aim to enhance image quality by merging complementary information from images captured by different sensors. Among these algorithms, deep unfolding network (DUN)-based methods achieve promising performance but still suffer from two issues: they lack a degradation-resistant-oriented fusion model and struggle to adequately consider the structural properties of DUNs, making them vulnerable to degradation scenarios. In this paper, we propose a Degradation-Resistant Unfolding Network (DeRUN) for the HIF task to generate high-quality fused images even in degradation scenarios. Specifically, we introduce a novel HIF model for degradation resistance and derive its optimization procedures. Then, we incorporate the optimization unfolding process into the proposed DeRUN for end-to-end training. To ensure the robustness and efficiency of DeRUN, we employ a joint constraint strategy and a lightweight partial weight sharing module. To train DeRUN, we further propose a gradient direction-based entropy loss with powerful texture representation capacity. Extensive experiments show that DeRUN significantly outperforms existing methods on four HIF tasks, as well as downstream applications, with cheaper computational and memory costs.

1. Introduction

Heterogeneous image fusion (HIF) aims to integrate images acquired by different imaging sensors to generate a more robust and informative image. HIF has been investigated in various domains, such as infrared and visible image fusion (IVF) [61], medical image fusion (MIF) [45], and biological image fusion (BIF) [41], based on different sensors. By merging the complementary information from the heterogeneous images, the fused image is expected to have better quality than individual ones and thus serves better for decision-making or subsequent processing, especially in degradation scenarios such as extreme weather in IVF, magnetic turbulence in MIF, and phase noise in BIF.

Infrared Visible IVF-Net [17] DeRUN

Figure 1: IVF results under degradation scenarios, including low light and heavy smoke. DeRUN generates more visual-appealing results for its degradation resistance capacity.

Taking the IVF task as an example, visible images typically contain more details but are easily influenced by poor illumination, fog, and other factors. In contrast, infrared images are resistant to lighting and weather disturbances but usually have poor texture details. Hence, IVF aims to get fused images that preserve texture details from visible images and the robust thermal radiation from infrared images simultaneously. The fused images with enhanced quality and informative components can better resist degradation conditions and thus benefit high-level image analysis.

Existing HIF techniques can be classified into model-based and learning-based methods. Model-based fusion methods focus on iteratively solving the optimization-based fusion framework with manually designed fusion rules [30, 34]. Although guaranteed with strong interpretability, such methods are limited by hand-crafted feature extractors with poor generalizability, thus failing to cope with degradation conditions. Benefiting from the nonlinear capacity of convolutional neural network (CNN), learning-based fusion solutions have strong generalizability by learning an end-to-end fusion mapping [60, 65, 64]. However, such learning-

*Corresponding author, † Work done during the internship at Smart Vision.

based methods are criticized as black boxes [6] for lacking interpretability, which inevitably suppresses the fusion performance, especially in complex degradation scenarios.

Recently, a novel learning-based technique called deep unfolding network (DUN) has been proposed to unify the merits of model-based and deep learning-based methods. Specifically, DUNs unfold the iterative optimization steps of a model-based solution into a deep neural network for end-to-end training. DM-Fusion [44] and IVF-Net [17] have introduced DUN to the IVF task and achieved promising performance in most cases. However, existing DUN-based HIF techniques suffer from two issues. (i) They lack a fusion model dedicated to alleviating degradation, making it difficult to fully leverage the complementary information in complex degradation scenarios. (ii) These methods fail to entirely consider the structural characteristics of DUNs, thus neglecting to employ the valuable inter-module information interaction and lacking an efficient feature extraction module among the cascade structure. Consequently, as shown in Fig. 1, existing DUN-based methods struggle to generate visually appealing results in degradation scenarios.

To address the above problem, in this paper, we propose a novel deep unfolding network, Degradation-Resistant Unfolding Network (DeRUN), for the HIF task to generate high-quality fused images even in degradation scenarios (see Fig. 1). DeRUN is derived from a novel fusion model (HIFM) which improves the existing fusion model (Eq. (1)) in salient information emphasis and noise suppression, thus better resisting degradation conditions. Next, we frame the iterative optimization process of the fusion model into a multi-stage network, where each stage consists of three modules: data fusion module (DFM), visual fidelity module (VFM), and structure preservation module (SPM), with all the connections following the update procedure, thus accommodating interpretability and generalizability.

To ensure the robustness and efficiency of DeRUN, we incorporate a joint constraint strategy (JCS) and a partial weight sharing module (PWSM). Specifically, we propose JCS by imposing physical and denoising constraints, which are derived from DFM and VFM, on SPM to ensure the robustness of DeRUN in salient texture enhancement even in degradation scenarios. Additionally, to achieve efficient and effective feature extraction, we propose a lightweight network named PWSM, which can also mitigate the domain discrepancy problem of JCS. We further propose to train DeRUN with a novel gradient direction-based entropy loss. This loss is inspired by the group property of gradient directions for weak boundary recognition and noise removal, thus encouraging DeRUN to possess the texture representation capacity, even in degradation scenarios.

Our contributions are summarized as follows:

- We propose DeRUN for HIF, which unifies interpretability and generalization, to generate high-quality fused images even in degradation scenarios.
- We propose a series of valid techniques to ensure the robustness and efficiency of our DeRUN model, including the joint constraint strategy (JCS), the partial weight sharing module (PWSM), and a novel gradient direction-based entropy loss dubbed LGDE loss.
- Extensive experiments on four HIF tasks verify the superiority of our DeRUN to existing methods in terms of image quality, running efficiency, degradation resistance, and favorability in downstream applications.

2. Related Work

Heterogeneous image fusion. Early works address HIF in a model-based fashion by modeling physic priors with some optimization techniques, such as image decomposition [25] and saliency detection [34]. However, these methods usually suffer from poor generalizability for hand-crafted operators, thus failing to handle complex degradation. Learning-based methods have dominated this field recently. Some methods generated fused images with Generative Adversarial Networks from the concatenation of sources [63, 24, 62] or individually [33, 28]. DPCN [41] proposed a detail-preserving network with two branches corresponding to the two source inputs on the BIF task. U2Fusion [46] designed a VGG16-based unified fusion framework with saliency maps in IVF and MIF tasks. While these learning-based methods treat the neural network as a black box with poor interpretability, suppressing fusion performance.

To alleviate this problem, DM-Fusion [44] introduced DUN to the IVF task with the existing fusion model [35]. Moreover, IVF-Net [17] designed a novel fusion model to fuse infrared and visible images with different resolutions. However, current DUN-based HIF methods neither contain a degradation-resistant-oriented fusion model nor adequately consider the structural properties of DUNs, making it difficult to fully leverage the complementary information in complex degradation scenarios and thus struggling to generate visually appealing results in degradation scenarios.

Unlike existing DUN-based techniques, we propose DeRUN to generate high-quality results even in degradation conditions, which comprises a degradation-resistant model, a joint constraint strategy, a lightweight feature extraction module, and a gradient direction-based entropy loss.

Deep unfolding network. Early deep unfolding networks (DUNs) are devoted to unfolding the computational graph of an iterative algorithm to a deep neural network to train the predefined parameters and operators in an end-to-end manner, e.g., LSC [5] and ADMM-Net [50]. With the development of CNN, current existing DUNs tend to use off-the-shelf CNN modules to fit the physical priors in the optimization task [26, 4]. While increasing the nonlinear capacity of the framework, they neglect the significant constraint

of prior information on system imaging and weaken the interpretability of the DUNs. To overcome these problems, some DUNs only use the CNN modules, e.g., U-Net and DenseNet, for feature extraction [66, 57], which, however, leads to heavy storage burdens. Our DeRUN model directly unfolds the iterative minimization process of HIFM for interpretability and proposes the lightweight PWSM for efficient and effective feature extraction.

3. Methodology

In this section, we take the IVF task as an example to explain the working mechanism of DeRUN.

3.1. HIFM Model

Given an infrared image $\mathbf{U}$ and a visible image $\mathbf{V}$, early model-based IVF methods typically obtain the fused image $\mathbf{X}$ by optimizing the objective function [31, 56]:

$$\min_{\mathbf{X}} \frac{1}{2}\|\mathbf{X}-\mathbf{U}\|_2^2 + \lambda\|\nabla\mathbf{X}-\nabla\mathbf{V}\|_p^p, \tag{1}$$

where λ, $\|\cdot\|_p$, ∇ represent a trade-off parameter, ℓ_p-norm, the gradient operator. Eq. (1) encourages $\mathbf{X}$ to preserve the global appearance from $\mathbf{U}$ and texture details from $\mathbf{V}$.

Inspired by the success of residual priors in related fields [51, 38], some model-based IVF methods also introduce the cross-field residual map $\mathbf{M} = \mathbf{U} - \mathbf{V}$ to enhance the pixel-level salient information by the fidelity term [35] and rewrite Eq. (1) with the residual variable $\mathbf{Y} = \mathbf{X} - \mathbf{V}$:

$$\min_{\mathbf{Y}} \frac{1}{2}\|\mathbf{Y}-\mathbf{M}\|_2^2 + \sum_{l=1}^{L}\lambda_l\psi\left(\nabla_l\mathbf{Y}\right), \tag{2}$$

where L, λ_l, $\psi(\cdot)$, and ∇_l denote the filter number, trade-off parameter, regularizer, and Sobel operator [29]. The final result $\mathbf{X}$ can be calculated by $\mathbf{X} = \mathbf{Y}^{\text{final}} + \mathbf{V}$. For smoothness, existing fusion methods set $\psi(\cdot)$ as ℓ_2-norm [35]. However, since $\mathbf{V}$ is sensitive to imaging scenarios, simple subtraction of $\mathbf{U}$ and $\mathbf{V}$ struggles to eradicate the degradation, such as heavy fog or complex noise, which leads to the presence of undesired noise and artifacts in the residual map $\mathbf{M}$ and further diminishes the quality of $\mathbf{Y}$. To overcome this, we set $\psi(\cdot)$ as ℓ_1-norm regularizer, which can suppress undesired degradation in $\mathbf{Y}$ and serve as a denoiser.

In addition to pixel-level saliency, salient texture details are also critical for maintaining salient information. Therefore, we propose a sparse texture constraint for $\mathbf{Y}$ to focus on the salient texture in $\mathbf{M}$ under sparse scenarios. By incorporating the salient residual map and introducing the salient texture constraint, the final energy-based HIF model (**HIFM**) is designed to emphasize salient information at both pixel and texture levels while effectively suppressing undesired noise. Our **HIFM** is formulated as follows:

$$\min_{\mathbf{Y}} \frac{1}{2}\|\mathbf{Y}-\mathbf{M}\|_2^2 + \sum_{l=1}^{L}\lambda_l\psi(\nabla_l\mathbf{Y}) + \sum_{k=1}^{K}\mu_k\phi[\nabla_k(\mathbf{Y}-\mathbf{M})], \tag{3}$$

where μ_k is a trade-off parameter, ∇_k is a texture extractor, and $\phi(\cdot)$ is a sparsity-oriented non-convex potential function [40]: $\phi(\mathbf{a}) = \sum_i \log\left(1+\theta\mathbf{a}_i^2\right)$, where $\mathbf{a}_i$ denotes the i^{th} element of $\mathbf{a}$ and θ is the sparsity controlled parameter.

Model optimization. By introducing two auxiliary variables $\mathbf{E}_l$ and $\mathbf{H}_k$, Eq. (3) can be reformulated as follows:

$$\begin{aligned}&\min_{\mathbf{Y},\mathbf{E}_l,\mathbf{H}_k} \frac{1}{2}\|\mathbf{Y}-\mathbf{M}\|_2^2 + \sum_{l=1}^{L}\lambda_l\psi\left(\mathbf{E}_l\right) + \sum_{k=1}^{K}\mu_k\phi\left(\mathbf{H}_k\right), \\ &\text{s.t. } \mathbf{E}_l = \nabla_l\mathbf{Y}, \mathbf{H}_k = \nabla_k(\mathbf{Y}-\mathbf{M}).\end{aligned} \tag{4}$$

We solve Eq. (4) with alternative direction method of multipliers (ADMM) [1] by introducing two dual variables $\mathbf{F}_l$ and $\mathbf{G}_k$ to control the step size of $\mathbf{E}_l$ and $\mathbf{H}_k$. The solutions are presented as follows (see derivations in **Supp**):

$$\begin{cases}\mathbf{Y}^{(n)} = \left(\mathbf{Q_a}\right)^{-1}\left(\mathbf{M}+\mathbf{Q_b}\right), \\ \mathbf{E}_l^{(n)} = S\left(\nabla_l\mathbf{Y}^{(n)} + \mathbf{F}_l^{(n-1)}; \frac{\lambda_l}{\rho_l}\right), \\ \mathbf{F}_l^{(n)} = \mathbf{F}_l^{(n-1)} + \varphi_l\left(\nabla_l\mathbf{Y}^{(n)} - \mathbf{E}_l^{(n)}\right), \\ \mathbf{H}_k^{(n)} = \mathbf{H}_k^{(n-1)} - \sigma_k\mathbf{Q_c}, \\ \mathbf{G}_k^{(n)} = \mathbf{G}_k^{(n-1)} + \omega_k\left(\nabla_k\left(\mathbf{Y}^{(n)}-\mathbf{M}\right) - \mathbf{H}_k^{(n)}\right),\end{cases} \tag{5}$$

where n denotes n^{th} iteration. $\mathbf{Q_a} = \mathbf{I} + \sum_{l=1}^{L}\rho_l\nabla_l^2 + \sum_{k=1}^{K}\tau_k\nabla_k^2$, $\mathbf{Q_b} = \sum_{l=1}^{L}\rho_l\nabla_l^T(\mathbf{E}_l^{(n-1)} - \mathbf{F}_l^{(n-1)}) + \sum_{k=1}^{K}\tau_k\nabla_k^T(\nabla_k\mathbf{M} + \mathbf{H}_k^{(n-1)} - \mathbf{G}_k^{(n-1)})$, and $\mathbf{Q_c} = \tau_k\nabla_k(\mathbf{Y}^{(n)}-\mathbf{M}) + \frac{2\mu_k\mathbf{H}_k^{(n-1)}}{1+\theta(\mathbf{H}_k^{(n-1)})^2} + \tau_k\mathbf{G}_k^{(n-1)} + \tau_k\mathbf{H}_k^{(n-1)}$. ρ_l and τ_k are penalty weights. $\mathbf{I}$ is the identity matrix. $S(\cdot)$ is a nonlinear proximal operator. φ_l and ω_k control the training rates. Note that $\mathbf{H}_k$ is optimized by the gradient descent strategy for its non-convex property [4].

3.2. DeRUN

3.2.1 Deep Unfolding Mechanism

In this section, we propose a Degradation-Resistant Unfolding Network (**DeRUN**), which incorporates the HIFM-based optimization into the deep network architecture. As shown in Fig. 2 (i), DeRUN is unfolded into N stages that correspond to N iterative optimization steps. Each stage comprises three decoupled modules, namely, the data fusion module (DFM), visual fidelity module (VFM), and structure preservation module (SPM), whose progression process follows Eq. (5) with the learnable variables and parameters for adaptive image fusion. Refer to Fig. 2 (ii) for more details.

Data fusion module. DFM is derived from the data term in Eq. (4) and serves to consolidate the extracted information from the previous stage. Therefore, DFM is instrumental in revealing the intrinsic essence of fusion tasks and guiding network training. Given $\mathbf{E}_l^{(n-1)}$, $\mathbf{F}_l^{(n-1)}$, $\mathbf{H}_k^{(n-1)}$, $\mathbf{G}_k^{(n-1)}$, the DFM at the n^{th} stage is defined as:

$$\mathbf{Y}^{(n)} = (\mathbf{Q_a})^{-1}(\mathbf{M}+\mathbf{Q_b}), \tag{6}$$

where $\mathbf{Q_a}$ and $\mathbf{Q_b}$ keep the same formulations in Eq. (5). In

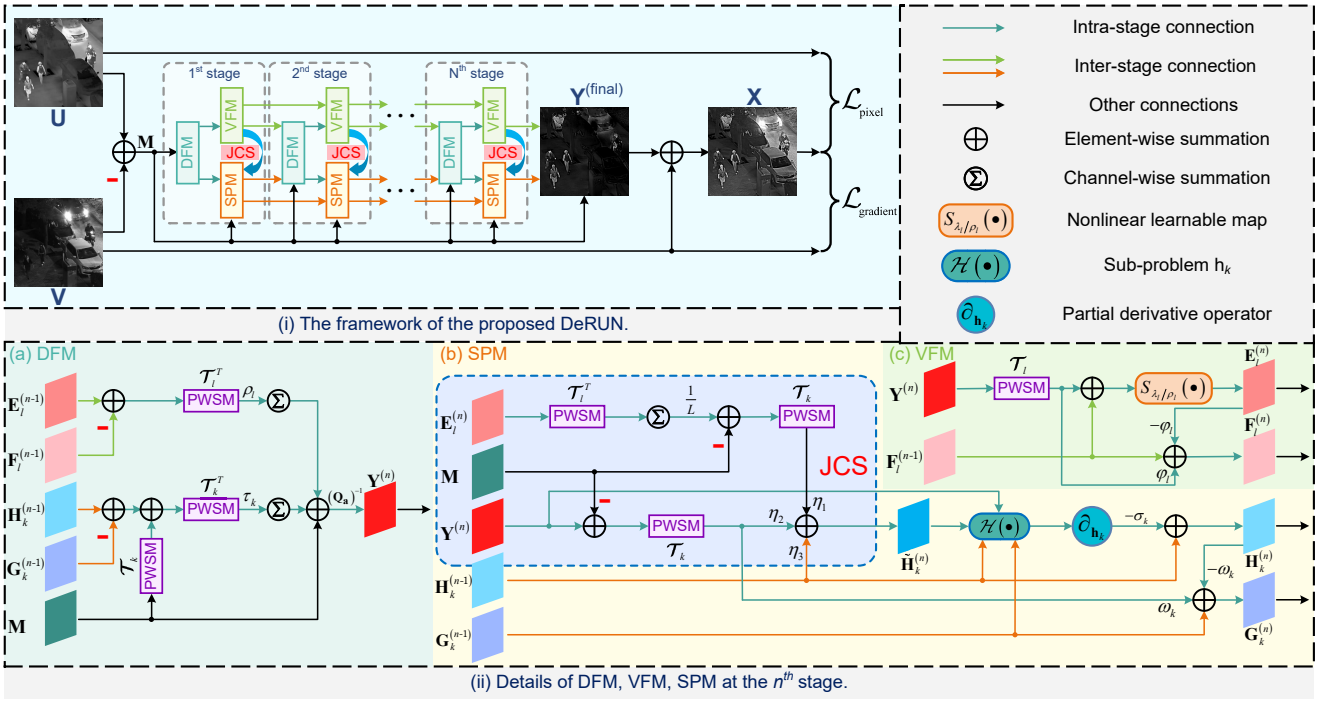


Figure 2: The framework of the proposed DeRUN with the detailed architecture at n^{th} stage, where we take IVF task as an example.

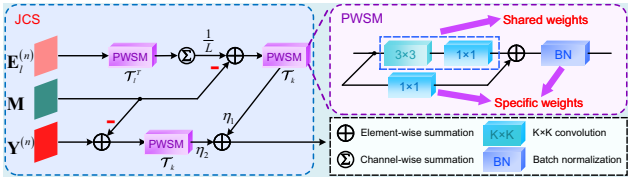


Figure 3: The detailed information of JCS and PWSM.

the first stage, $\mathbf{E}_l^{(n-1)}$, $\mathbf{F}_l^{(n-1)}$, $\mathbf{H}_k^{(n-1)}$, $\mathbf{G}_k^{(n-1)}$ are initialized as zeros. For consistency in mathematical representation, we relax the variables $\{\rho_l, \nabla_l, \tau_k, \nabla_k\}$ to be learnable parameters, with the stage number omitted.

Visual fidelity module. VFM is evolved from the ℓ_1-norm-based denoising regularization in Eq. (4) with a noise suppression variable $\mathbf{E}_l$ and its dual variable $\mathbf{F}_l$ for step size. Therefore, VFM can serve as a denoiser by suppressing undesired noise and artifacts in $\mathbf{Y}$ and provide cleaner feature maps. Given $\mathbf{F}_l^{(n-1)}$ and $\mathbf{Y}^{(n)}$, the solution of denoising variable $\mathbf{E}_l^{(n)}$ at n^{th} stage is presented as follows:

$$\mathbf{E}_l^{(n)} = S_{\lambda_l/\rho_l}\left(\nabla_l \mathbf{Y}^{(n)} + \mathbf{F}_l^{(n-1)}; \{\theta_{l,i}\}_{i=1}^{I_t}\right), \quad (7)$$

where $S_{\lambda_l/\rho_l}(\cdot)$ is a nonlinear map with the learnable parameters $\{\theta_{l,i}\}_{i=1}^{I_t}$. Given $\mathbf{E}_l^{(n)}$, we update $\mathbf{F}_l^{(n)}$ as:

$$\mathbf{F}_l^{(n)} = \mathbf{F}_l^{(n-1)} + \varphi_l\left(\nabla_l \mathbf{Y}^{(n)} - \mathbf{E}_l^{(n)}\right), \quad (8)$$

where φ_l is a learnable parameter to control the step size.

Structure preservation module. SPM comes from the salient texture constraint from Eq. (4) with a salient texture enhancement variable $\mathbf{H}_k$ and a dual variable $\mathbf{G}_k$. $\mathbf{H}_k^{(n)}$ is updated with the gradient descent strategy [4]:

$$\mathbf{H}_k^{(n)} = \mathbf{H}_k^{(n-1)} - \sigma_k \mathbf{Q}_\mathbf{c}, \quad (9)$$

where σ_k is a learnable parameter and $\mathbf{Q}_\mathbf{c}$ follows the definition in Eq. (5). Then, we acquire the dual variable $\mathbf{G}_k^{(n)}$:

$$\mathbf{G}_k^{(n)} = \mathbf{G}_k^{(n-1)} + \omega_k\left(\nabla_k\left(\mathbf{Y}^{(n)} - \mathbf{M}\right) - \mathbf{H}_k^{(n)}\right), \quad (10)$$

where ω_k is a learnable parameter.

Overall, in each stage, DeRUN updates the five variables, namely, $\{\mathbf{Y}, \mathbf{E}_l, \mathbf{F}_l, \mathbf{H}_k, \mathbf{G}_k\}$, with eight sets of learnable weights, *i.e.*, $\{\rho_l, \nabla_l, \tau_k, \nabla_k, \{\theta_{l,i}\}_{i=1}^{I_t}, \varphi_l, \sigma_k, \omega_k\}$. By integrating the degradation-resistance HIF model (HIFM) into the deep network architecture with the deep unfolding mechanism, DeRUN is expected to adaptively generate high-quality fused images even in degradation scenarios.

3.2.2 Joint Constraint Strategy

To enhance the robustness of DeRUN in salient texture enhancement, **joint constraint strategy (JCS)** is proposed for SPM. JCS imposes extra constraints, including the physical constraint and the denoising constraint, on the salient texture enhancement variable $\mathbf{H}_k^{(n)}$ to facilitate salient texture maintenance and noise suppression, ensuring robustness even in degradation scenarios. Specifically, the physic constraint, derived from the textural subtraction of $\mathbf{Y}^{(n)}$ (in DFM) and $\mathbf{M}$, provides physical guidance for salient texture maintenance. Meanwhile, the denoising constraint transfers cleaner information from $\mathbf{E}_l^{(n)}$ (in VFM), which

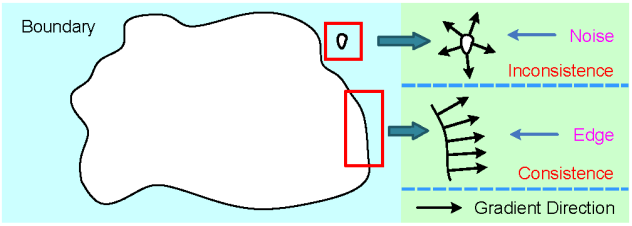


Figure 4: Distributions of gradient directions at edge and noise.

serves as a denoiser, to $\mathbf{H}_k^{(n)}$ (in SPM) for noise suppression. To maintain the original optimization process, we introduce an optimization auxiliary variable $\widetilde{\mathbf{H}}_k^{(n)}$ [15] with the joint constraints imposed, which is defined as follows:

$$\begin{aligned}\widetilde{\mathbf{H}}_k^{(n)} &= \eta_1 \mathbf{H}_k^{(n-1)} + \eta_2 \nabla_k (\mathbf{Y}^{(n)} - \mathbf{M}) \\ &+ \eta_3 \nabla_k \left(\frac{1}{L} \sum_{l=1}^{L} \nabla_l^T \mathbf{E}_l^{(n)} - \mathbf{M} \right),\end{aligned} \tag{11}$$

where η_1, η_2, η_3 are learnable trade-off parameters. With the assistance of $\widetilde{\mathbf{H}}_k^{(n)}$, $\mathbf{H}_k^{(n)}$ can be re-updated as follows:

$$\mathbf{H}_k^{(n)} = \mathbf{H}_k^{(n-1)} - \sigma_k \partial_{\mathbf{H}_k} \mathcal{H} \left(\mathbf{H}_k^{(n-1)}, \mathbf{G}_k^{(n-1)}, \mathbf{Y}^{(n)}, \widetilde{\mathbf{H}}_k^{(n)} \right), \tag{12}$$

where $\partial_{\mathbf{H}_k}$ denote the partial derivative operator for $\mathbf{H}_k^{(n-1)}$ and $\mathcal{H}(\mathbf{H}_k^{(n-1)}, \mathbf{G}_k^{(n-1)}, \mathbf{Y}^{(n)}, \widetilde{\mathbf{H}}_k^{(n)}) = \mu_k \phi(\mathbf{H}_k^{(n-1)}) + \lambda_{\mathbf{H}_k} \|\widetilde{\mathbf{H}}_k^{(n)} - \mathbf{H}_k^{(n-1)}\|_2^2 + \tau_k \|\nabla_k(\mathbf{Y}^{(n)} - \mathbf{M}) - \mathbf{H}_k^{(n-1)} + \mathbf{G}_k^{(n-1)}\|_2^2$. $\lambda_{\mathbf{H}_k}$ is a learnable step size parameter.

The second and third terms in Eq. (11) correspond to the physical constraint and denoising constraint, which provide the physical guidance and cleaner information for variables in SPM, *i.e.*, $\mathbf{H}_k$ (Eq. (12)) and $\mathbf{G}_k$ (Eq. (10)). These constraints ensure salient texture maintenance and noise suppression, thereby improving the robustness of DeRUN in salient texture enhancement, even in degradation scenarios.

3.2.3 Partial Weight Sharing Module

DUNs typically employ CNN modules [10, 11] for feature extraction, owing to their powerful nonlinear capacity. However, due to the cascade structure of DUNs, the utilization of large-scale CNN modules for feature extraction can result in heavy storage burdens [66, 12]. To address this issue, recent DUN-based techniques have opted to reduce the number of parameters by using Siamesed structures, such as weight sharing in one stage or across the entire pipeline [26, 57]. Although decreasing memory costs, such Siamesed structures can inevitably weaken the feature extraction capacity and thus degrade the performance of DUNs. Therefore, a lightweight but effective feature extraction module is critical for DUN-based frameworks.

Considering that overly independent feature extractors can introduce domain discrepancies between VFM and SPM and thus limit the effect of the denoising constraint in JCS, we propose a lightweight network, named **partial weight sharing module (PWSM)** $\mathcal{T}$, to replace ∇_l and ∇_k for powerful feature extraction. Specifically, as shown in Fig. 3, the proposed PWSM consists of a weight-specific branch and a weight-sharing branch. The weight-specific branch ensures the feature extraction capacity and thus compensates for the deficiency of the Siamesed structure, while the weight-sharing branch simultaneously decreases parameters and mitigates the domain discrepancy problem of JCS. The two branches cooperate with each other and jointly form a lightweight but powerful feature extractor. Following [57], the feature extractors share the same structure as their transposes, i.e., $\mathcal{T}_l^T$ and $\mathcal{T}_k^T$, with different weights for weight-specific branches.

3.2.4 Local Gradient Directional Entropy Loss

In HIF, degradation scenarios, *e.g.*, the low-light illumination in IVF, the magnetic turbulence in MIF, and the phase noise in BIF, often manifest as blurred images with low contrast and undesired artifacts, which weaken the boundary information and introduce complex noise. In this case, existing gradient magnitude-based loss functions struggle to effectively grasp the weak boundary information and resist the complex noise, because gradient magnitude is sensitive to pixel-level numerical variations (individual property). However, distinct from gradient magnitude, gradient direction exhibits the group property, *i.e.*, gradient directions present a consistent trend around the texture and are cluttered around noise (see Fig. 4). Therefore, gradient direction is more favorable than gradient magnitude in weak boundary recognition and noise removal, and thus owns advantages to resist those degraded images.

Inspired by local entropy that measures the local consistency of pixel values for image segmentation [13], we quantize the gradient direction into I directions with $\frac{2\pi}{I}$ intervals and propose a Shannon entropy-based locally consistent measurement for quantified gradient direction, termed **local gradient directional entropy (LGDE)**. Given a $B \times B$ local block, the LGDE of the center point (x, y) is defined as: $E\left(D\left(\mathbf{X}\right)\right)_{(x,y)} = -\sum_{i=1}^{I} p_i \log p_i$, where $D\left(\mathbf{X}\right)$ is the quantized gradient directional map of the input image $\mathbf{X}$. I is the number of quantized gradient directions, which is set as 8 to balance the performance and efficiency. p_i denotes the probability of the i^{th} gradient direction in the $B \times B$ block with the center point of (x, y). RelectionPad2d [36] is used to pad the boundaries. Note that LGDE is sensitive to the block size, where a larger size can ignore some detailed texture and a smaller size can suppress the diversity of entropy. Inspired by dilated convolution [53], a multi-scale LGDE is proposed with the following definition:

$$E\left(D\left(\mathbf{X}\right)\right) = \frac{1}{2}\left(E_{B=3}\left(D\left(\mathbf{X}\right)\right) + E_{B=5}\left(D\left(\mathbf{X}\right)\right)\right), \tag{13}$$

where LGDE of the block with $B = 5$ is a dilated operation

Methods	*TNO* (100 pairs)								*M*^{3}*FD* (4200 pairs)								*LLVIP* (3463 pairs)							
	SSIM	PSNR	AG	FMI	VIF	EN	UIQI	Q_P	SSIM	PSNR	AG	FMI	VIF	EN	UIQI	Q_P	SSIM	PSNR	AG	FMI	VIF	EN	UIQI	Q_P
U2Fusion [46]	0.73	16.73	6.05	0.90	1.27	6.97	0.80	0.70	0.72	16.89	6.04	0.91	1.30	6.96	0.80	0.72	0.69	17.13	6.13	0.90	0.75	6.59	0.80	0.66
SDNet [54]	0.76	17.24	6.26	0.88	1.30	6.64	0.82	0.71	0.74	17.13	6.22	0.90	0.94	6.83	0.78	0.71	0.68	17.26	6.09	0.89	1.13	6.72	0.78	0.65
IVF-Net [17]	0.77	18.52	6.73	0.90	1.32	7.12	0.83	0.75	0.79	**18.57**	6.68	**0.92**	1.16	7.15	0.83	0.71	0.72	17.15	6.62	**0.92**	1.42	6.87	0.85	0.67
DM-Fusion [44]	0.75	**18.62**	6.61	0.89	1.31	**7.34**	0.81	0.75	0.78	18.38	6.43	0.91	**1.33**	7.22	0.82	0.73	0.73	17.32	6.50	0.91	1.36	7.04	0.81	0.65
TarDAL [28]	0.72	17.19	6.58	0.89	0.98	7.28	0.84	0.73	0.73	18.21	5.13	0.90	0.89	7.16	0.77	0.70	0.70	17.09	5.57	0.91	0.91	**7.22**	0.79	**0.70**
RFNet [47]	0.77	18.53	6.87	**0.92**	**1.42**	6.83	**0.88**	0.75	0.80	18.40	**7.30**	0.91	1.30	7.07	0.82	**0.79**	0.72	17.35	**6.99**	0.91	1.41	7.13	0.85	**0.70**
DeFusion [27]	**0.79**	18.46	**7.02**	0.90	1.35	6.95	0.85	**0.79**	**0.81**	18.51	7.18	0.90	1.27	**7.28**	**0.87**	**0.77**	**0.75**	**17.42**	6.68	0.90	**1.54**	7.10	**0.88**	0.69
DeRUN	**0.79**	**18.83**	**7.18**	**0.91**	**1.54**	**7.45**	**0.88**	**0.78**	**0.81**	**18.73**	**7.47**	**0.93**	**1.42**	**7.34**	**0.86**	**0.77**	**0.75**	**17.58**	**6.98**	**0.92**	**1.58**	**7.17**	**0.86**	**0.72**

Table 1: Quantitative evaluation in IVF task on five datasets. The best results and the second-best are marked in red and blue.

Methods	**PMF** task in **MIF** (200 pairs)								**SMF** task in **MIF** (250 pairs)								**GPF** task in **BIF** (90 pairs)							
	SSIM	PSNR	AG	FMI	VIF	EN	UIQI	Q_P	SSIM	PSNR	AG	FMI	VIF	EN	UIQI	Q_P	SSIM	PSNR	AG	FMI	VIF	EN	UIQI	Q_P
U2Fusion [46]	0.73	20.17	4.92	0.86	0.72	4.55	0.85	0.76	0.75	22.41	5.41	0.84	0.68	3.47	0.77	0.74	0.77	17.90	4.71	0.85	0.66	4.31	0.78	0.68
SDNet [54]	0.76	**20.54**	5.92	0.87	0.75	3.83	0.82	0.74	0.74	22.48	6.13	0.86	0.65	3.43	0.85	0.71	0.75	18.07	5.51	0.86	0.86	4.67	0.77	0.67
IVF-Net [17]	0.79	20.41	6.73	0.90	**0.77**	4.68	0.88	0.75	0.76	22.54	7.16	**0.89**	0.68	4.76	0.88	0.74	**0.78**	18.96	5.17	0.88	0.94	5.04	0.81	0.70
DM-Fusion [44]	0.80	20.35	6.64	0.89	0.72	**4.96**	0.87	0.74	0.79	22.35	6.44	0.88	0.66	**5.42**	0.86	0.73	0.76	18.52	4.96	0.87	0.73	4.10	0.83	0.72
TarDAL [28]	0.76	20.04	6.50	0.88	0.73	3.81	0.87	0.76	0.47	22.61	6.83	0.84	0.64	3.66	**0.91**	0.75	0.73	18.20	5.25	0.86	0.77	4.33	0.82	0.69
RFNet [47]	0.80	20.41	6.81	**0.92**	**0.78**	4.77	0.90	0.76	0.80	23.94	**7.94**	**0.89**	**0.72**	4.39	0.89	0.78	**0.78**	**19.74**	6.67	**0.90**	**1.10**	5.12	0.87	**0.73**
DeFusion [27]	**0.81**	20.46	**7.18**	0.90	0.75	4.27	**0.91**	**0.77**	**0.81**	**24.18**	7.87	0.88	0.62	3.81	0.90	**0.80**	0.77	19.66	**6.80**	0.89	1.04	**5.37**	**0.89**	**0.73**
DeRUN	**0.81**	**20.59**	**7.60**	**0.91**	**0.77**	**5.14**	**0.93**	**0.80**	**0.83**	**24.32**	**8.01**	**0.89**	**0.83**	**5.39**	**0.92**	**0.81**	**0.79**	**19.79**	**6.73**	**0.91**	**1.18**	**5.53**	**0.89**	**0.75**

Table 2: Quantitative evaluation for the PMF, SMF, and GPF tasks. The best results and the second-best are marked in red and blue.

with the dilated rate $r = 1$, which has the same number of pixels as the block with $B = 3$ but with larger perceptual fields. Therefore, the total loss is presented as follows:

$$\begin{aligned} L_t &= L_{pixel} + \gamma L_{lgde}, \\ &= MSE(\mathbf{X}, \mathbf{U}) + \gamma MSE(E(D(\mathbf{X})), E(D(\mathbf{V}))), \end{aligned} \tag{14}$$

where L_{pixel} is the content loss [9]. $MSE(\cdot)$ denotes mean square error and γ is the trade-off parameter.

4. Experiment

Implementation details. The proposed DeRUN is implemented by PyTorch on two RTX3090TI GPUs and is optimized by Adam optimizer with momentum terms (0.9, 0.999), 150 epochs, and a batch size of 4. The learning rate is initialized as 1×10^{-4} and decays with a factor of 0.9 every ten epochs after the first 50 epochs. The stage number N is set as 9 and the trade-off parameter γ in the loss is 10.

Compared methods and evaluation metrics. We select seven state-of-the-art (SOTA) methods for comparison, including U2Fusion [46], SDNet [54], DM-Fusion [44], IVF-Net [17], TarDAL [28], RFNet [47], and DeFusion [27].

Eight metrics are employed to quantitatively evaluate the fusion performance, including SSIM, PSNR, average gradient (AG) [21], FMI, visual fidelity (VIF) [7], entropy (EN) [39], universal image quality index (UIQI) [43], and phase congruency-based metric Q_P [59]. For all the metrics, larger metrics indicate better fusion performance. For fairness, all methods are evaluated with the same toolbox.

4.1. Comparative Results

Natural image fusion (NIF). Following [58], we employ the first 90 pairs from *TNO* [42] to train DeRUN and evaluate it on five datasets, *i.e.*, the rest in *TNO*, *M*^{3}*FD* [28], *LLVIP* [16], *INO* [30], *RoadScene* [46], where *M*^{3}*FD* and *LLVIP* are the two largest IVF datasets. For space limitation, we only report the results of *TNO*, *M*^{3}*FD* [28], *LLVIP* [16], and present the rest in **Supp**. Tab. 1 shows the quantitative results, where DeRUN achieves seventeen best results and seven second-best results in the three datasets. Fig. 5 presents the qualitative results, where DeRUN generates fusion images with visual fidelity and texture enhancement even in extreme conditions while existing methods fail to achieve this, such as the slogan carved on the stone.

Medical image fusion (MIF). Following [52], we evaluate two medical image fusion tasks, i.e., PET and MRI image fusion (PMF), and SPECT and MRI image fusion (SMF). We train DeRUN with HWBA dataset [3], where the training set has 144 pairs of images. In Tab. 2, our DeRUN model achieves thirteen best metrics and three second-best metrics on the two tasks. The first two rows of Fig. 6 are the fusion results of PMF and SMF, respectively. We can see that DeRUN can ensure the functional information, i.e., color intensity, from PET/SPECT, and preserve the anatomical details from MRI, which is attributed to our task-specific HIFM and the multi-scale LGDE loss.

Biological image fusion (BIF). ATC dataset [18] of green fluorescent protein (GFP) and phase contrast (PC) image fusion (GPF) is used for training, which includes 60 pairs. As shown in Tab. 2, we achieve leading performance on almost all metrics, which demonstrates the superiority of our DeRUN. We present a qualitative comparison in the last row of Fig. 6. With the proposed JCS and PWSM, DeRUN can preserve the structure feature from PC while retaining the high contrast of the green fluorescent from GFP. Furthermore, DeRUN can resist the phase noise from PC and generate a cleaner output by LGDE loss. Although

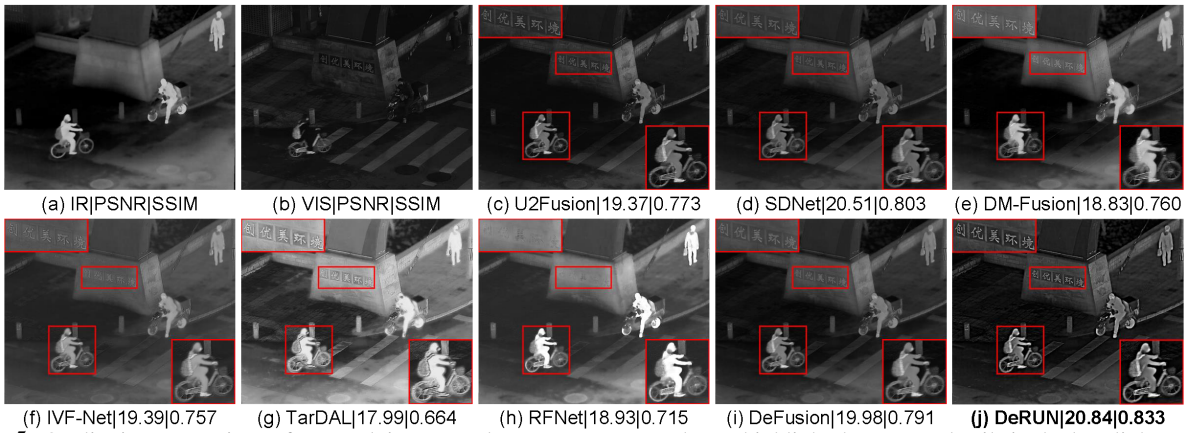


Figure 5: Qualitative comparison of IVF task in *LLVIP* dataset. DeRUN can better highlight the texture details in the low-light scenario.

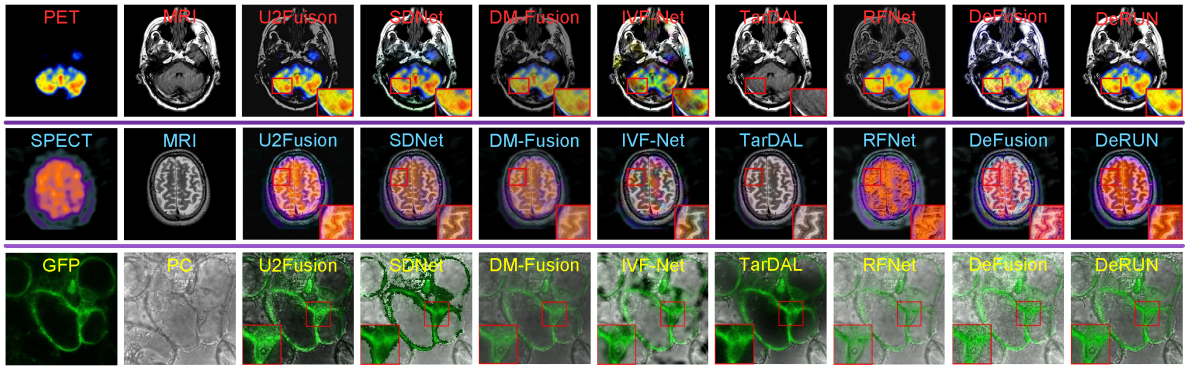
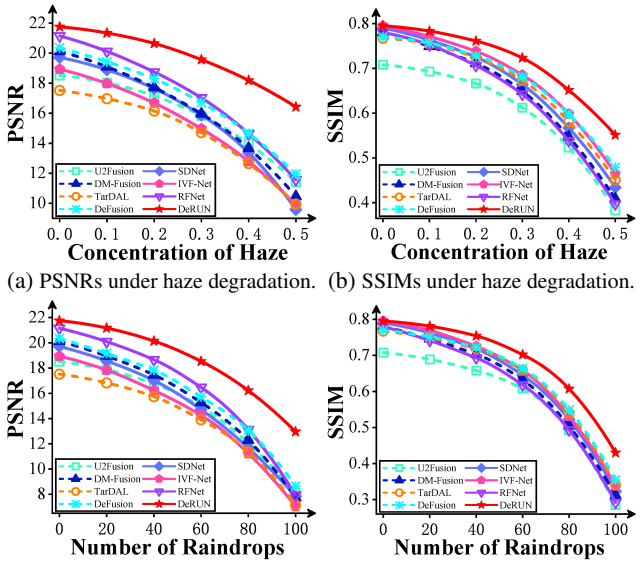


Figure 6: Qualitative comparison of PMF, SMF, GPF tasks. DeRUN generates visual-appealing results with texture preservation.



(a) PSNRs under haze degradation. (b) SSIMs under haze degradation.

(c) PSNRs under rain degradation. (d) SSIMs under rain degradation.

Figure 7: Robustness analysis in degraded scenarios on IVF task. Haze and rain are simulated following [20] and [14].

PET/SPECT/GFP are color images, information fusion is only conducted in their Y channel, and the outputs are colorized by a common post-processing operation [46].

4.2. Ablation Study

We evaluate the effect of DeRUN with four metrics, *i.e.*, SSIM, PSNR, AG, and FMI. See **Supp** for more details, *e.g.*, parameter analyses for the stage number N, and the hyper-parameters I and B in the LGDE loss.

Effect of HIFM. To verify the effect of HIFM, we compare DeRUN with other combinations of fusion models and solutions. As shown in Tab. 3a, better results can be obtained with our fusion model (HIFM), whether in ADMM or DUN, which demonstrates the superiority of HIFM.

Effect of JCS. In Tab. 3b, the proposed JCS outperforms other ablation methods, i.e., w/o JCS, w/o physical constraint (PC), and w/o denoising constraint (DC), which is credited to our joint physical and denoising constraints, in salient texture maintenance and noise suppression.

Effect of PWSM. To verify the effect of PWSM, we compare PWSM with a Siamese extractor [50], and a symmetrical extractor, where the symmetrical extractor shares a similar parameter number with PWSM. Tab. 3c illustrates the superiority of PWSM, which attributes to the favorable cooperation between the weight-specific branch and the weight-sharing branch.

Effect of LGDE loss L_{lgde}. We verify the effect of LGDE loss by comparing it with existing gradient-based losses,

Metrics	CM+OS	PM+OS	CM+DL	PM+DL
SSIM	0.70	0.73	0.78	**0.81**
PSNR	15.76	16.23	18.46	**18.73**
AG	5.83	6.37	6.58	**7.47**
FMI	0.87	0.88	0.91	**0.93**

(a) Effect of HIFM.

Metrics	w/o JCS	w/o PC	w/o DC	w/ JCS
SSIM	0.79	0.80	0.80	**0.81**
PSNR	18.14	18.53	18.29	**18.73**
AG	6.82	7.08	7.25	**7.47**
FMI	0.91	0.91	0.92	**0.93**

(b) Effect of JCS.

Metrics	Siam	Symm	PWSM
SSIM	0.79	**0.81**	**0.81**
PSNR	18.24	18.57	**18.73**
AG	6.96	7.23	**7.47**
FMI	0.90	0.92	**0.93**

(c) Effect of PWSM.

Metrics	w/o L_{lgde}	L_{gp}	L_{jg}	w/ L_{lgde}
SSIM	0.78	0.80	0.80	**0.81**
PSNR	17.14	18.67	18.58	**18.73**
AG	6.87	7.28	7.32	**7.47**
FMI	0.91	0.91	0.91	**0.93**

(d) Effect of LGDE loss.

Table 3: Ablation study in the IVF task on M^3FD. "w/" and "w/o" denote with and without. (a) CM, PM, OS, and DL are short for conventional model (Eq. (2)), proposed model (HIFM), optimization solution (ADMM), and deep learning (DUN). (b) JCS contains PC and DC. (c) Siam and Symm are short for Siamesed and Symmetrical. The best results are marked in **bold**.

Datasets	Metrics	Infrared	Visible	DM-Fusion	IVF-Net	TarDAL	RFNet	DeFusion	DeRUN
LLVIP	mAP@.5 ↑	0.853	0.799	0.871	0.789	**0.876**	0.867	0.869	**0.884**
	mAP@.75 ↑	0.521	0.394	**0.562**	0.497	0.558	0.553	0.538	**0.564**
	mAP@.5:.95 ↑	0.502	0.417	0.525	0.487	**0.531**	0.524	0.516	**0.529**
M^3FD	mAP@.5 ↑	0.764	0.756	0.784	0.745	0.781	0.776	**0.807**	**0.819**
	mAP@.75 ↑	0.584	0.581	0.604	0.533	0.616	0.614	**0.629**	**0.644**
	mAP@.5:.95 ↑	0.496	0.489	0.507	0.427	0.523	**0.524**	0.516	**0.553**

Table 4: Object detection comparison with partial IVF methods.

Trackers	Metrics	Infrared	Visible	DM-Fusion	IVF-Net	TarDAL	RFNet	DeFusion	DeRUN
LADCF	Accuracy ↑	0.501	0.537	0.542	0.533	0.547	0.546	**0.578**	**0.596**
	Failures ↓	59.47	43.71	41.15	42.33	**37.15**	39.16	39.45	**36.11**
	EAO ↑	0.201	0.257	0.245	0.224	0.263	**0.267**	0.259	**0.284**
GFSDCF	Accuracy ↑	0.522	0.571	0.583	0.591	0.554	0.607	**0.623**	**0.635**
	Failures ↓	57.39	35.42	30.16	33.72	35.41	**30.13**	30.36	**28.33**
	EAO ↑	0.221	0.287	0.271	0.288	**0.302**	0.297	0.301	**0.311**

Table 5: Object tracking comparison with partial IVF methods.

i.e., gradient penalty-based L_{gp} [23] and joint gradient-based L_{jg} [55]. To do so, we replace L_{lgde} with these losses and retrain the DeRUN model. Tab. 3d indicates the effect of our LGDE loss, which is attributed to the strong capacity of L_{lgde} in weak boundary recognition and noise removal.

4.3. Downstream Tasks and Running Time

Robustness analysis in degraded scenarios. A significant application for IVF is the intelligent transportation systems (ITS), where the core competency is the ability to resist degradation from visible images [32, 2]. To rank the capacity in degradation resistance, we simulate some inference (haze and rain) for the visible images in *RoadScene* dataset following [20, 14]. In Fig. 7, our DeRUN not only shows a strong performance but also exhibits a more robust trend of change with the increasing degradation, which attributes to the robustness of DeRUN with HIFM, JCS, and LGDE.

Object detection with fused images. The fused images with enhanced quality are expected to have better downstream performance than the individual ones. We first verify this on object detection [37, 8] with IVF task. Following TarDAL, all the compared methods are performed on *LLVIP* and M^3FD datasets with YOLOv5. As shown in Tab. 4, DeRUN acquires five best results and one second-best result, which indicates that we can generate detection-friendly fused results owing to the powerful capability in salient information emphasis and degradation resistance.

Object tracking with fused images. Object tracking is also conducted on the IVF task with VOT-RGBT2019 benchmark [19]. Following [22], two trackers, i.e., LADCF [49] and GFSDCF [48] are used for evaluation. The best performance exhibited in Tab. 5 proves the superiority of DeRUN, attributing to our salient information preservation capacity.

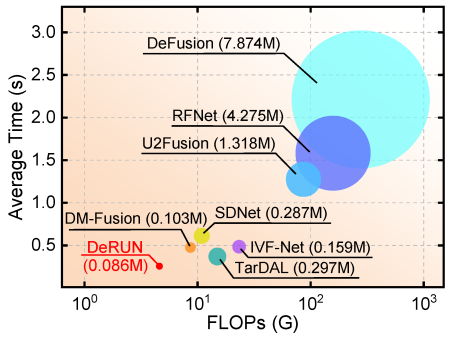


Figure 8: Efficiency analysis with the image size of 256×256

Efficiency analysis. We compare the average time, parameters, and FLOPs on the IVF task in Fig. 8. As shown in Fig. 8, our DeRUN achieves the fastest average time, the smallest FLOPs, and the most minor parameters for its DUN-based framework and the lightweight PWSM.

5. Conclusions

In this paper, we first design a novel HIF model (HIFM) for degradation resistance. Then we unfold the ADMM solution of HIFM and propose a Degradation-Resistant Unfolding Network (DeRUN) for HIF. To ensure the robustness and efficiency of DeRUN, we propose JCS and PWSM. Inspired by the group property of gradient direction, we propose a gradient direction-based entropy loss (LGDE) with powerful texture representation capacity. Extensive experiments comprehensively verify the advancement of our DeRUN over SOTAs from image quality, computational efficiency, and downstream applications.

Acknowledgements: This work is supported by National Key R&D Program of China (Grant No. 2020AAA0108303), Shenzhen Science and Technology Project (Grant No. JCYJ20200109143041798) & Shenzhen Stable Supporting Program (WDZC20200820200655001) & Shenzhen Key Laboratory of next generation interactive media innovative technology (Grant No. ZDSYS 20210623092001004), National Natural Science Foundation of China (No. 62192712), Beijing Institute of Technology Research Fund Program for Young Scholars.

References

[1] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, Jonathan Eckstein, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. Trends Mach. Learn.*, 3(1):1–122, 2011. 3

[2] Lizhen Deng, Chunming He, Guoxia Xu, Hu Zhu, and Hao Wang. Pcgan: A noise robust conditional generative adversarial network for one shot learning. *IEEE Trans. Intell. Transp. Syst.*, 23(12):25249–25258, 2022. 8

[3] Jiao Du, Weisheng Li, Ke Lu, and Bin Xiao. An overview of multi-modal medical image fusion. *Neurocomputing*, 215:3–20, 2016. 6

[4] Xueyang Fu, Xi Wang, Aiping Liu, Junwei Han, and Zheng-Jun Zha. Learning dual priors for jpeg compression artifacts removal. In *ICCV*, pages 4086–4095, 2021. 2, 3, 4

[5] Karol Gregor and Yann LeCun. Learning fast approximations of sparse coding. In *ICML*, pages 399–406, 2010. 2

[6] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM Comput. Surv.*, 51(5):1–42, 2018. 2

[7] Yu Han, Yunze Cai, Yin Cao, and Xiaoming Xu. A new image fusion performance metric based on visual information fidelity. *Inf. Fusion*, 14(2):127–135, 2013. 6

[8] Yizeng Han, Dongchen Han, Zeyu Liu, Yulin Wang, Xuran Pan, Yifan Pu, Chao Deng, Junlan Feng, Shiji Song, and Gao Huang. Dynamic perceiver for efficient visual recognition. In *International Conference on Computer Vision*, 2023. 8

[9] Chunming He, Kai Li, Guoxia Xu, Jiangpeng Yan, Longxiang Tang, Yulun Zhang, Xiu Li, and Yaowei Wang. Hqg-net: Unpaired medical image enhancement with high-quality guidance. *arXiv preprint arXiv:2307.07829*, 2023. 6

[10] Chunming He, Kai Li, Yachao Zhang, Longxiang Tang, Yulun Zhang, Zhenhua Guo, and Xiu Li. Camouflaged object detection with feature decomposition and edge reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22046–22055, 2023. 5

[11] Chunming He, Kai Li, Yachao Zhang, Guoxia Xu, Longxiang Tang, Yulun Zhang, Zhenhua Guo, and Xiu Li. Weakly-supervised concealed object segmentation with sam-based pseudo labeling and multi-scale feature grouping. *arXiv preprint arXiv:2305.11003*, 2023. 5

[12] Chunming He, Kai Li, Yachao Zhang, Yulun Zhang, Zhenhua Guo, Xiu Li, Martin Danelljan, and Fisher Yu. Strategic preys make acute predators: Enhancing camouflaged object detectors by generating camouflaged objects. *arXiv preprint arXiv:2308.03166*, 2023. 5

[13] Chunming He, Xiaobo Wang, Lizhen Deng, and Guoxia Xu. Image threshold segmentation based on glle histogram. In *CPSCom*, pages 410–415. IEEE, 2019. 5

[14] Dennis Hospach, Stefan Mueller, Wolfgang Rosenstiel, and Oliver Bringmann. Simulation of falling rain for robustness testing of video-based surround sensing systems. In *DATE*, pages 233–236. IEEE, 2016. 7, 8

[15] Fukeng Huang, Jie Shen, and Zhiguo Yang. A highly efficient and accurate new scalar auxiliary variable approach for gradient flows. *SIAM J. Sci. Comput.*, 42(4):A2514–A2536, 2020. 5

[16] Xinyu Jia, Chuang Zhu, Minzhen Li, Wenqi Tang, and Wenli Zhou. Llvip: A visible-infrared paired dataset for low-light vision. In *ICCV*, pages 3496–3504, 2021. 6

[17] Mingye Ju, Chunming He, Juping Liu, Bin Kang, Jian Su, and Dengyin Zhang. Ivf-net: An infrared and visible data fusion deep network for traffic object enhancement in intelligent transportation systems. *IEEE Transactions Intell Transp Syst*, 2022. 1, 2, 6

[18] Olga A Koroleva, Matthew L Tomlinson, David Leader, Peter Shaw, and John H Doonan. High-throughput protein localization in arabidopsis using agrobacterium-mediated transient expression of gfp-orf fusions. *Plant J.*, 41(1):162–174, 2005. 6

[19] Matej Kristan, Jiri Matas, Ales Leonardis, Michael Felsberg, Roman Pflugfelder, Joni-Kristian Kamarainen, Luka ˇCehovin Zajc, Ondrej Drbohlav, Alan Lukezic, Amanda Berg, et al. The seventh visual object tracking vot2019 challenge results. In *ICCVW*, pages 0–0, 2019. 8

[20] Boyi Li, Wenqi Ren, Dengpan Fu, Dacheng Tao, Dan Feng, Wenjun Zeng, and Zhangyang Wang. Benchmarking single-image dehazing and beyond. *IEEE Trans. Image Process.*, 28(1):492–505, 2018. 7, 8

[21] Huafeng Li, Yueliang Cen, Yu Liu, Xun Chen, and Zhengtao Yu. Different input resolutions and arbitrary output resolution: a meta learning-based deep framework for infrared and visible image fusion. *IEEE Trans. Image Process.*, 30:4070–4083, 2021. 6

[22] Hui Li, Xiao-Jun Wu, and Josef Kittler. Mdlatlrr: A novel decomposition method for infrared and visible image fusion. *IEEE Trans. Image Process.*, 29:4733–4746, 2020. 8

[23] Jing Li, Hongtao Huo, Chang Li, Renhua Wang, and Qi Feng. Attentionfgan: Infrared and visible image fusion using attention-based generative adversarial networks. *IEEE Trans. Multimedia*, 23:1383–1396, 2020. 8

[24] Kai Li, Yulun Zhang, Kunpeng Li, and Yun Fu. Adversarial feature hallucination networks for few-shot learning. In *CVPR*, pages 13470–13479, 2020. 2

[25] Shutao Li, Xudong Kang, and Jianwen Hu. Image fusion with guided filtering. *IEEE Trans. Image process.*, 22(7):2864–2875, 2013. 2

[26] Yuqi Li, Qiang Fu, and Wolfgang Heidrich. Multispectral illumination estimation using deep unrolling network. In *ICCV*, pages 2672–2681, 2021. 2, 5

[27] Pengwei Liang, Junjun Jiang, Xianming Liu, and Jiayi Ma. Fusion from decomposition: A self-supervised decomposition approach for image fusion. In *ECCV*, pages 719–735. Springer, 2022. 6

[28] Jinyuan Liu, Xin Fan, Zhanbo Huang, Guanyao Wu, Risheng Liu, Wei Zhong, and Zhongxuan Luo. Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection. In *CVPR*, 2022. 2, 6

[29] Ying Lu, Chunming He, Yu-Feng Yu, Guoxia Xu, Hu Zhu, and Lizhen Deng. Vector co-occurrence morphological edge detection for colour image. *IET Image Processing*, 15(13):3063–3070, 2021. 3

[30] Jiayi Ma, Chen Chen, Chang Li, and Jun Huang. Infrared and visible image fusion via gradient transfer and total variation minimization. *Inf. Fusion*, 31:100–109, 2016. 1, 6

[31] Jiayi Ma, Yong Ma, and Chang Li. Infrared and visible image fusion methods and applications: A survey. *Inf. Fusion*, 45:153–178, 2019. 3

[32] Jiayi Ma, Yong Ma, and Chang Li. Infrared and visible image fusion methods and applications: a survey. *Inf. Fusion*, 45:153–178, 2019. 8

[33] Jiayi Ma, Han Xu, Junjun Jiang, Xiaoguang Mei, and Xiao-Ping Zhang. Ddcgan: A dual-discriminator conditional generative adversarial network for multi-resolution image fusion. *IEEE Trans. Image Process.*, 29:4980–4995, 2020. 2

[34] Jinlei Ma, Zhiqiang Zhou, Bo Wang, and Hua Zong. Infrared and visible image fusion based on visual saliency map and weighted least square optimization. *Infrared Phys. Technol.*, 82:8–17, 2017. 1, 2

[35] Yong Ma, Jun Chen, Chen Chen, Fan Fan, and Jiayi Ma. Infrared and visible image fusion using total variation model. *Neurocomputing*, 202:12–19, 2016. 2, 3

[36] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *NIPS*, 32, 2019. 5

[37] Yifan Pu, Yiru Wang, Zhuofan Xia, Yizeng Han, Yulin Wang, Weihao Gan, Zidong Wang, Shiji Song, and Gao Huang. Adaptive rotated convolution for rotated object detection. In *International Conference on Computer Vision*, 2023. 8

[38] Man Qin, Chao Ren, Hong Yang, Xiaohai He, and Zhengyong Wang. Blind image denoising via deep unfolding network with degradation information guidance. *IEEE Trans. Circuits and Syst. II Express Briefs*, 2023. 3

[39] J Wesley Roberts, Jan A van Aardt, and Fethi Babikker Ahmed. Assessment of image fusion procedures using entropy, image quality, and multispectral classification. *J. Appl. Remote Sens.*, 2(1):1–28, 2008. 6

[40] Stefan Roth and Michael J Black. Fields of experts. *Int. J. Comput. Vis.*, 82(2):205–229, 2009. 3

[41] Wei Tang, Yu Liu, Juan Cheng, Chang Li, and Xun Chen. Green fluorescent protein and phase contrast image fusion via detail preserving cross network. *IEEE Trans. Comput. Imaging*, 7:584–597, 2021. 1, 2

[42] Alexander Toet. The tno multiband image data collection. *Data Brief*, 15:249, 2017. 6

[43] Zhou Wang and Alan C Bovik. A universal image quality index. *IEEE Signal Processing Lett.*, 9(3):81–84, 2002. 6

[44] Guoxia Xu, Chunming He, Hao Wang, Hu Zhu, and Weiping Ding. Dm-fusion: Deep model-driven network for heterogeneous image fusion. *IEEE Trans. Neural Networks Learn. Syst.*, 2023. 2, 6

[45] Han Xu and Jiayi Ma. Emfusion: An unsupervised enhanced medical image fusion network. *Inf. Fusion*, 76:177–186, 2021. 1

[46] Han Xu, Jiayi Ma, Junjun Jiang, Xiaojie Guo, and Haibin Ling. U2fusion: A unified unsupervised image fusion network. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(1):502–518, 2022. 2, 6, 7

[47] Han Xu, Jiayi Ma, Jiteng Yuan, Zhuliang Le, and Wei Liu. Rfnet: Unsupervised network for mutually reinforcing multi-modal image registration and fusion. In *CVPR*, pages 19679–19688, 2022. 6

[48] Tianyang Xu, Zhen-Hua Feng, Xiao-Jun Wu, and Josef Kittler. Joint group feature selection and discriminative filter learning for robust visual object tracking. In *ICCV*, pages 7950–7960, 2019. 8

[49] Tianyang Xu, Zhen-Hua Feng, Xiao-Jun Wu, and Josef Kittler. Learning adaptive discriminative correlation filters via temporal consistency preserving spatial feature selection for robust visual object tracking. *IEEE Trans. Image Process.*, 28(11):5596–5609, 2019. 8

[50] Yan Yang, Jian Sun, Huibin Li, Zongben Xu, et al. Deep admm-net for compressive sensing mri. *NIPS*, 29, 2016. 2, 7

[51] Qiaosi Yi, Juncheng Li, Qinyan Dai, Faming Fang, Guixu Zhang, and Tieyong Zeng. Structure-preserving deraining with residue channel prior guidance. In *ICCV*, pages 4238–4247, 2021. 3

[52] Ming Yin, Xiaoning Liu, Yu Liu, and Xun Chen. Medical image fusion with parameter-adaptive pulse coupled neural network in nonsubsampled shearlet transform domain. *IEEE Trans. Instrum. Meas.*, 68(1):49–64, 2019. 6

[53] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015. 5

[54] Hao Zhang and Jiayi Ma. Sdnet: A versatile squeeze-and-decomposition network for real-time image fusion. *Int. J. Comput. Vis.*, 129:2761–2785, 2021. 6

[55] Hao Zhang, Jiteng Yuan, Xin Tian, and Jiayi Ma. Gan-fm: Infrared and visible image fusion using gan with full-scale skip connection and dual markovian discriminators. *IEEE Trans. Comput. Imaging*, 7:1134–1147, 2021. 8

[56] Qiheng Zhang, Yuli Fu, Haifeng Li, and Jian Zou. Dictionary learning method for joint sparse representation-based image fusion. *Opt. Eng.*, 52(5):057006–057006, 2013. 3

[57] Xuanyu Zhang, Yongbing Zhang, Ruiqin Xiong, Qilin Sun, and Jian Zhang. Herosnet: Hyperspectral explicable reconstruction and optimal sampling deep network for snapshot compressive imaging. In *CVPR*, pages 17532–17541, 2022. 3, 5

[58] Yu Zhang, Yu Liu, Peng Sun, Han Yan, Xiaolin Zhao, and Li Zhang. Ifcnn: A general image fusion framework based on convolutional neural network. *Inf. Fusion*, 54:99–118, 2020. 6

[59] Jiying Zhao, Robert Laganiere, and Zheng Liu. Performance assessment of combinative pixel-level image fusion based on an absolute feature measurement. *Int. J. Innov. Comput. Inf. Control*, 3(6):1433–1447, 2007. 6

[60] Zixiang Zhao, Haowen Bai, Jiangshe Zhang, Yulun Zhang, Shuang Xu, Zudi Lin, Radu Timofte, and Luc Van Gool. Cddfuse: Correlation-driven dual-branch feature decomposition for multi-modality image fusion. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5906–5916, 2023. 1

[61] Zixiang Zhao, Haowen Bai, Jiangshe Zhang, Yulun Zhang, Kai Zhang, Shuang Xu, Dongdong Chen, Radu Timofte, and Luc Van Gool. Equivariant multi-modality image fusion. *arXiv preprint arXiv:2305.11443*, 2023. 1

[62] Zixiang Zhao, Haowen Bai, Yuanzhi Zhu, Jiangshe Zhang, Shuang Xu, Yulun Zhang, Kai Zhang, Deyu Meng, Radu Timofte, and Luc Van Gool. Ddfm: denoising diffusion model for multi-modality image fusion. In *2023 IEEE/CVF International Conference on Computer Vision*, 2023. 2

[63] Zixiang Zhao, Shuang Xu, Chunxia Zhang, Junmin Liu, Jiangshe Zhang, and Pengfei Li. Didfuse: Deep image decomposition for infrared and visible image fusion. In *2020 International Joint Conferences on Artificial Intelligence*, 2021. 2

[64] Zixiang Zhao, Jiangshe Zhang, Xiang Gu, Chengli Tan, Shuang Xu, Yulun Zhang, Radu Timofte, and Luc Van Gool. Spherical space feature decomposition for guided depth map super-resolution. In *2023 IEEE/CVF International Conference on Computer Vision*, 2023. 1

[65] Zixiang Zhao, Jiangshe Zhang, Shuang Xu, Zudi Lin, and Hanspeter Pfister. Discrete cosine transform network for guided depth map super-resolution. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5697–5707, 2022. 1

[66] Chuanjun Zheng, Daming Shi, and Wentian Shi. Adaptive unfolding total variation network for low-light image enhancement. In *ICCV*, pages 4439–4448, 2021. 3, 5