D^{2}TNet: A Decoupled Dual-Teacher Framework for Robust Infrared and Visible Image Fusion in Diverse Degraded Scenes

Yujie Liu*, Peng Liu†, Zhen Chen*†, Feng Lu*, Congxuan Zhang*
*Nanchang Hangkong University, Nanchang, China †Beihang University, Beijing, China
Email: {a1822849697, dr_chenzhen, zcxdsg}@163.com, liupeng_97@buaa.edu.cn, lufeng@nchu.edu.cn

***Abstract*—Real-world image degradations present significant challenges for infrared-visible image fusion (IVIF), as most existing methods rely on the implicit assumption of pristine inputs. Conventional single-teacher frameworks struggle to balance image restoration and detail preservation, as denoising objectives inevitably suppress the high-frequency textures critical for fusion. Similarly, cascaded two-stage approaches suffer from error propagation and limited end-to-end efficacy. To resolve this fundamental conflict, we propose a novel Decoupled Dual-Teacher (D^{2}T) framework that disentangles restoration and fusion into dedicated expert networks. A lightweight student network distills this knowledge via a Quadruple-Supervision Strategy, aligning its intermediate features with robust representations from both teachers to achieve simultaneous restoration and fusion. Notably, our model operates in a fully blind manner, handling unknown degradations without requiring manual priors while maintaining high inference efficiency. Extensive experiments on challenging degraded datasets show that D^{2}T consistently outperforms state-of-the-art (SOTA) fusion methods, producing fused images with superior clarity and structural fidelity while effectively suppressing noise and artifacts from unknown degradations.**

***Index Terms*—Image Fusion, Task Conflict, Image Restoration, Teacher-Student Framework**

I. INTRODUCTION

Infrared-Visible Image Fusion (IVIF) is pivotal for safety-critical applications such as autonomous driving and video surveillance, where reliable perception across varying conditions is essential [1]. By integrating distinct thermal targets from infrared imagery with rich textural details from visible inputs, IVIF generates a unified representation that preserves comprehensive information for downstream vision tasks [2]. However, practical deployment is frequently hindered by complex real-world degradations, including sensor noise, motion blur, and adverse lighting. Such environmental volatility presents a significant impediment to current state-of-the-art (SOTA) fusion approaches [3], often rendering fused outputs unsuitable for high-level applications.

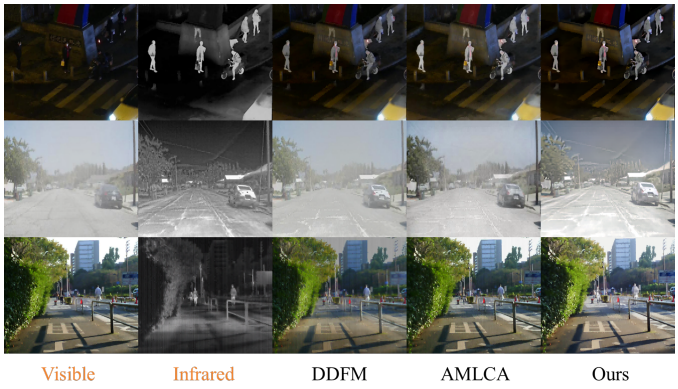


Fig. 1. Visual comparison of fusion performance under real-world degradations. From top to bottom: low-light visible inputs, hazy scenes, and infrared images corrupted by stripe noise.

Deep learning methods have become the mainstream in IVIF, evolving from early CNN-based models like DenseFuse [4] to advanced Transformer architectures like SwinFusion [5]. Despite their success, most existing methods typically assume that the input images are clean and ideal. However, this assumption rarely holds in real-world scenarios. Although some specific models like DIVFusion [6] are designed to handle low-light conditions, they struggle to generalize to complex scenes with unknown or mixed degradations. This limitation is clearly shown in Fig. 1. Under real-world degradations, even state-of-the-art methods fail to produce satisfactory results, leading to fused images with severe artifacts and structural loss.

Although a simple pipeline of separate restoration and fusion models provides a natural solution, this two-stage approach is limited by error propagation between stages [7]. Furthermore, cascading two complex models leads to high computational costs, making the method unsuitable for real-time edge deployment. It cannot optimize restoration and fusion in an end-to-end manner. An alternative is knowledge distillation, which offers an effective paradigm for end-to-end training. However, conventional single-teacher frameworks

This work was supported in part by the National Natural Science Foundation of China under Grant U25A20480, Grant U2441241, Grant 62272209, Grant 62401243, and Grant 62222206; in part by the Major Research and Development Project of Jiangxi Province under Grant 20232ACC01007, Grant 20242BAB20038, Grant 20253BAC280056; in part by the Key Research and Development Program of Jiangxi Province under Grant 20232BBE50006; in part by the Early-Career Young Scientists and Technologists Project of Jiangxi Province Grant 20244BCE52116, Grant 20244BCE52117.

Congxuan Zhang and Feng Lu are the corresponding authors.

like HKDNet [8] are hindered by a fundamental tension between the goals of restoration and fusion. There is a fundamental conflict between these two tasks. Image restoration aims to remove noise and artifacts, which often results in smoothing. However, image fusion requires preserving high-frequency details to maintain textures and structures. When a single teacher network performs both image restoration and fusion tasks, it learns conflicting objectives from the two supervision goals. Such conflicting objectives lead to oscillating gradients that force the network to encode both noise patterns and fine textures into its intermediate features, making learned features increasingly ambiguous. As a result, the student receives ambiguous guidance in knowledge distillation, weakening its ability to preserve structure while suppressing artifacts.

To address this limitation, we design a D^2T framework that separates restoration and fusion into specialized networks. Each teacher focuses on a single task, providing cleaner supervision and reducing interference between competing objectives. This decoupling enables more effective knowledge transfer, allowing the student to learn both artifact removal and texture preservation under real-world conditions. Most importantly, our student network works in a blind setting, handling unknown degradations efficiently without needing manual priors. Our main contributions to this work are as follows:

- We introduce a D^2T framework that explicitly separates image restoration and multimodal fusion at the supervision source, mitigating conflicting gradients and stabilizing end to end training.
- We introduce a quadruple-supervision strategy that delivers targeted guidance at multiple levels, helping the student recover fine details while maintaining structural consistency.
- Experimental results show that our method achieves strong performance on public benchmarks, with clear improvements in visual quality and robustness across diverse degradation types.

To ensure the reproducibility of our work, the source code and the evaluation protocols will be made available to the community upon the acceptance of this paper.

II. RELATED WORK

In this section, we review the evolution of IVIF from general deep learning frameworks to advanced generative and robust paradigms, and finally discuss the role of knowledge distillation.

A. General Fusion Frameworks

Deep learning has become the dominant approach in the field of IVIF. Early approaches, such as DenseFuse [4], adopted Auto-Encoders for feature extraction. However, these methods typically rely on manual fusion rules, such as simple addition, rather than learnable weights. This separation between feature extraction and fusion often results in information loss. To overcome this limitation, Convolutional Neural Networks (CNNs) emerged to enable end-to-end training. For instance, IFCNN [9] applies a unified framework to fuse images directly. Recent works have focused on improving feature interaction across different scales. For example, MMFI-Net [10] proposes a multi-level injection strategy. This method combines deep semantic features with shallow texture details, ensuring that both thermal targets and background clarity are well preserved. To capture features at different scales, researchers have also explored nest-based architectures. Methods like NestFuse [11] and RFN-Nest [12] use dense skip connections to link the encoder and decoder. This design preserves intermediate features that might otherwise be lost in deep networks, allowing the final image to retain more background details.

Although CNNs are effective in extracting local features, they are limited in capturing the global context. To address this issue, Transformer-based architectures have gained prominence. Models like SwinFusion [5] and CDDFuse [13] utilize self-attention mechanisms to model long-range dependencies. This capability allows them to preserve more semantic information in complex scenes than standard CNNs.

Despite these architectural advancements, a common limitation remains: most methods assume that the input images are high-quality. However, in real-world applications, images are often degraded by noise, blur, or poor lighting. When applied to such data, general frameworks tend to mistake noise for useful texture and fuse it into the final result. This leads to severe artifacts and poor visual quality, limiting their practical value.

B. Generative and Robust Paradigms

Generative models have become a popular choice for modeling complex data distributions. Among them, Generative Adversarial Networks (GANs), such as FusionGAN [14], introduce adversarial learning to image fusion. Unlike traditional methods that rely only on pixel-level loss functions, these methods create a competition between a generator and a discriminator. The discriminator tries to distinguish the fused result from the source images. This process drives the network to generate images with sharper edges and richer textures, avoiding the blurring problem often caused by traditional losses. However, GANs are difficult to train and often suffer from stability issues like mode collapse. Recently, Diffusion Models have gained attention due to their high generation quality. For example, DDFM [15] uses a diffusion process to generate fused images based on source inputs. Although these models produce excellent visual details, their multi-step sampling process is computationally expensive, which limits their application in real-time scenarios.

Beyond improving visual quality, recent research has also focused on the performance of downstream tasks. Works such as SeAFusion [16] and TarDAL [17] jointly train the fusion network with detection or segmentation models. By using semantic losses, these methods force the network to preserve features that are meaningful for machine recognition, rather than just focusing on visual appeal. In addition to generative models, recent studies have focused on improving the robust-

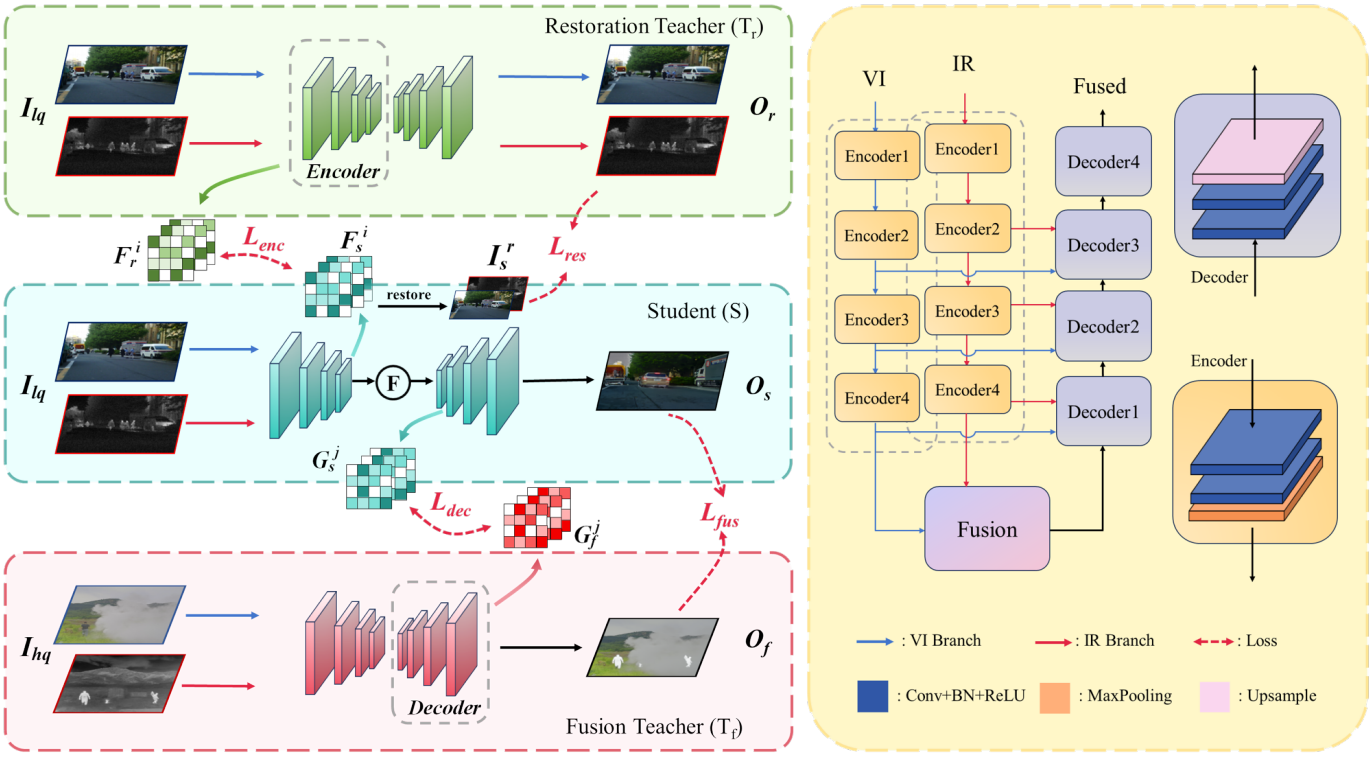


Fig. 2. Overview of our proposed D^{2}T framework. The student network (S) learns to fuse degraded images by distilling knowledge from two specialist teachers: a Restoration Teacher (T_r) that provides degradation-resilient features, and a Fusion Teacher (T_f) that provides a high-fidelity fusion target.

ness of fusion methods. For example, PAIF [18] proposes a perception-aware framework to resist adversarial attacks during semantic segmentation. Similarly, MoEFuse [19] uses a Mixture-of-Experts strategy to handle different types of image degradation effectively. However, these methods still face a major limitation. There is a conflict between image restoration and image fusion. Restoration typically requires smoothing to reduce noise, while fusion aims to preserve sharp details. It is difficult for a single network to balance these two opposing goals. As a result, the output images often contain residual noise or suffer from blurred textures.

C. Knowledge Distillation for Efficient Fusion

Knowledge Distillation (KD) offers a practical trade-off between performance and efficiency. In the field of IVIF, methods such as HKDNet [8] transfer knowledge from a heavy teacher to a lightweight student, enabling real-time inference on edge devices.

In addition to feature-based distillation, researchers have also investigated attention transfer. By mimicking the teacher's attention maps, the student network learns to focus on salient regions, such as thermal targets and texture details. This strategy allows the lightweight model to maintain high performance with much lower computational costs. More recently, TSDM-Fusion [20] introduced a degradation simulation module. In this method, the teacher guides the student to restore high-quality images from degraded inputs.

However, most existing KD frameworks follow a single-teacher paradigm. As analyzed earlier, a single network struggles to balance noise removal and texture preservation due to their conflicting gradient directions. When such a network acts as the teacher, it provides coupled and suboptimal guidance to the student. Consequently, the student mimics a teacher that still contains compromised features, limiting its final performance. To address this issue, we propose a Decoupled Dual-Teacher strategy. Instead of relying on a single model, we employ two specialized teachers to decouple the conflicting objectives. Specifically, one teacher focuses on degradation removal, while the other concentrates on texture preservation. This design provides the student with clear and distinct guidance, enabling robust fusion in an end-to-end manner.

III. METHOD

A. Network Architecture

As shown in Fig. 2, our proposed D^{2}T framework consists of three main components: a pre-trained Restoration Teacher (T_r), a pre-trained Fusion Teacher (T_f), and a trainable Student network (S). To address the conflict between restoration and fusion, we freeze the weights of both teachers during training. This allows them to provide stable, decoupled supervision for degradation removal and texture preservation, respectively.

1) Teacher Networks: The Restoration Teacher (T_r) is a pre-trained network designed for image restoration. In our framework, we employ its encoder to extract stable features from degraded inputs. Since T_r is optimized to suppress

noise and artifacts, its intermediate features (F_r^i) capture clean structural details. We use these features to supervise the student's encoder, forcing it to learn a degradation-invariant feature representation. The Fusion Teacher (T_f) acts as a reference for the fusion task. It takes clean source images as input to generate a high-quality fused result (O_f). In contrast to T_r, T_f emphasizes the decoding stage. Its decoder features (G_f^j) contain essential semantic information required for image synthesis. We distill this generative knowledge into the student's decoder, enabling the student to reconstruct fused images with high visual quality.

2) Student Network: The Student network (S) adopts a dual-branch U-Net architecture to process degraded visible and infrared images. As shown in Fig. 2, the network consists of a dual-stream encoder, a fusion bottleneck, and a decoder. The two encoder branches function in parallel, with each branch containing four stages of stacked convolutional blocks (Conv+BN+ReLU). We apply Max Pooling between stages to increase the receptive field and reduce spatial resolution. This hierarchical design allows the network to capture features ranging from low-level textures to high-level semantics. At the bottleneck, feature maps from both modalities are concatenated and processed by 3×3 convolutions. This step compresses the channel dimension and integrates cross-modal information before the decoding stage. Finally, the decoder mirrors the encoder structure and uses Upsampling layers to restore the image resolution.

To improve image reconstruction, we integrate two additional mechanisms into the backbone. First, we use multi-modal skip connections to bridge the encoder and decoder. These connections fuse features from both encoder branches and transfer them to the corresponding decoder layers. This design retains spatial details that are often lost during down-sampling, leading to sharper fused images. Second, we add lightweight auxiliary heads to each encoder branch. These heads map the encoder features to intermediate restored images (I_s^r). This provides direct supervision, guiding the network to extract clean features at an early stage before the fusion process.

B. Quadruple-Supervision Strategy

We formulate the training of the student network S, parameterized by θ_S, as a multi-objective optimization problem. To address the challenges of joint restoration and fusion, we employ a Quadruple-Supervision Strategy. This strategy decomposes the overall task into four complementary sub-objectives, imposing constraints on both the latent feature space and the signal reconstruction space. By regulating the network from these different perspectives, we ensure both robust feature extraction and high-quality image synthesis. The total objective function $\mathcal{J}(\theta_S)$ is defined as a weighted sum of these components:

$$\min_{\theta_S} \mathcal{J}(\theta_S) = \lambda_{enc}\mathcal{L}_{enc} + \lambda_{dec}\mathcal{L}_{dec} + \lambda_{res}\mathcal{L}_{res} + \lambda_{fus}\mathcal{L}_{fus}, \tag{1}$$

where $\lambda_{\{\cdot\}}$ are hyperparameters that modulate the trade-off between intermediate representation learning and final output synthesis.

1) Encoder Loss ($\mathcal{L}_{enc}$): The primary objective of the encoder is to map degraded inputs onto a clean, degradation-invariant latent manifold. We treat the pre-trained Restoration Teacher as a robust feature prior. To enforce the student network to learn this optimal representation, we minimize the divergence between the student's feature maps F_s^i and the teacher's expert features F_r^i. Formally, assuming the feature residuals follow a Laplacian distribution, we formulate the distillation objective using the channel-wise ℓ_1-norm. The loss is defined as the expectation over all encoder stages:

$$\begin{aligned}\mathcal{L}_{enc} &= \sum_{i=1}^{N_e} \frac{1}{C_i H_i W_i} \left\| F_s^i - F_r^i \right\|_1 \\ &= \sum_{i=1}^{N_e} \frac{1}{\Omega_i} \sum_{c=1}^{C_i} \sum_{\mathbf{p}\in\mathcal{P}} \left| F_s^i(\mathbf{p}, c) - F_r^i(\mathbf{p}, c) \right|,\end{aligned} \tag{2}$$

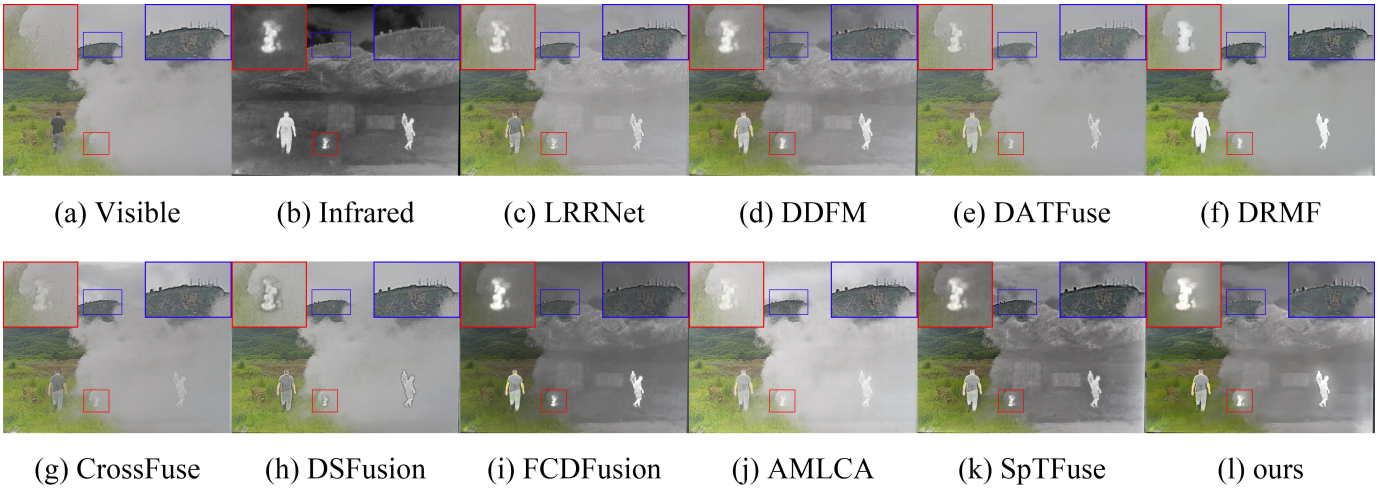
where N_e denotes the number of encoder stages, and $\Omega_i = C_i \times H_i \times W_i$ is the normalization factor for the i-th stage. $\mathbf{p} = (x, y)$ represents the spatial coordinate indices within the feature map domain $\mathcal{P}$. This strictly supervised constraint compels the student to suppress noise artifacts and align its latent space with the robust structural priors provided by the teacher.

2) Decoder Loss ($\mathcal{L}_{dec}$): While the encoder focuses on feature extraction, the decoder is responsible for the complex inversion from the latent space back to the image space. To ensure the student inherits the generative semantics required for high-quality fusion, we align its decoding trajectory with that of the Fusion Teacher. We minimize the semantic divergence between the student's decoder features G_s^j and the teacher's expert features G_f^j. Similar to the encoder strategy, this is formulated as a stage-wise optimization problem across the feature manifold:

$$\begin{aligned}\mathcal{L}_{dec} &= \sum_{j=1}^{N_d} \frac{1}{\Omega_j} \left\| G_s^j - G_f^j \right\|_1 \\ &= \sum_{j=1}^{N_d} \frac{1}{C_j H_j W_j} \sum_{c=1}^{C_j} \sum_{\mathbf{p}\in\mathcal{P}_j} \left| G_s^j(\mathbf{p}, c) - G_f^j(\mathbf{p}, c) \right|,\end{aligned} \tag{3}$$

where N_d denotes the number of decoder stages, and $\mathcal{P}_j$ represents the spatial coordinate domain at the j-th level. By minimizing this objective, we strictly regularize the student's generative process, forcing it to reconstruct salient details and texture information with the same fidelity as the Fusion Teacher.

3) Restoration Loss ($\mathcal{L}_{res}$): This loss directly supervises intermediate restored images I_s^r, generated by lightweight auxiliary heads attached to each encoder branch, against the Restoration Teacher's clean output O_r'. It judiciously combines



(a) Visible (b) Infrared (c) LRRNet (d) DDFM (e) DATFuse (f) DRMF

(g) CrossFuse (h) DSFusion (i) FCDFusion (j) AMLCA (k) SpTFuse (l) ours

Fig. 3. Fusion results on the M^{3}FD dataset. Red and blue boxes highlight the enlarged details of visible textures and infrared targets, respectively. Our method achieves the best visual fidelity.

TABLE I
QUANTITATIVE COMPARISON ON THE M3FD AND ROADSCENE DATASETS. THE BEST AND SECOND-BEST RESULTS ARE HIGHLIGHTED.

Methods	**M^{3}FD Dataset**					**RoadScene Dataset**				
	EN ↑	CC ↑	MS-SSIM ↑	PSNR ↑	MSE ↓	EN ↑	CC ↑	MS-SSIM ↑	PSNR ↑	MSE ↓
LRRNet [21]	6.9415	0.5264	0.8432	14.4838	0.0419	7.1455	0.5895	0.9573	11.9534	0.0656
DDFM [15]	6.9958	**0.5649**	**0.9810**	**15.0753**	**0.0352**	7.2108	0.4866	0.2508	12.6075	0.0577
DATFuse [22]	6.8780	0.4777	0.8910	12.8374	0.0589	6.6743	0.5754	0.9442	14.4164	0.0421
DRMF [23]	7.2239	0.4141	0.9008	10.9458	0.0929	7.2748	0.4647	0.9680	12.4807	0.0653
CrossFuse [24]	7.0701	0.3960	0.8708	13.0217	0.0623	7.1603	0.5072	0.9405	12.1368	0.0710
DSFusion [25]	7.2128	0.4118	0.9040	13.0577	0.0614	7.0180	0.5289	0.9168	14.6342	0.0447
FCDFusion [26]	6.6250	**0.5647**	0.9222	14.6032	0.0382	7.1202	**0.6049**	**1.0762**	**15.3112**	**0.0336**
AMLCA [27]	7.2789	0.4478	0.8858	11.8483	0.0779	**7.3455**	0.5460	0.9477	14.1494	0.0465
SpTFuse [28]	6.8972	0.4661	0.8568	14.2023	0.0448	7.2647	0.5497	0.9640	14.9353	0.0377
Ours	**7.2954**	0.5550	**0.9846**	**15.2163**	**0.0349**	**7.1733**	**0.6009**	**1.0450**	**15.8576**	**0.0310**

pixel-accurate L1 with perceptually aware SSIM to explicitly enhance the encoders' restoration capacity.

$$\mathcal{L}_{res} = \mathbb{E}\left[(1-\alpha)\mathcal{L}^1_{\text{pixel}}(I^r_s, O'_r) + \alpha(1-\mathcal{L}_{\text{SSIM}}(I^r_s, O'_r))\right], \tag{4}$$

where $\mathcal{L}^1_{\text{pixel}}(\cdot)$ denotes the standard pixel-wise ℓ_1 distance. The Structural Similarity Index Measure (SSIM) is formulated as:

$$\mathcal{L}_{\text{SSIM}}(I^r_s, O'_r) = \frac{1}{2}\left(\mathcal{S}(I^r_{s,VI}, O'_{r,VI}) + \mathcal{S}(I^r_{s,IR}, O'_{r,IR})\right), \tag{5}$$

Here, $I^r_{s,k}$ denotes the restored image from student branch k (where $k \in \{VI, IR\}$ for visible and infrared modalities), and $O'_{r,k}$ denotes the clean output for modality k from the Restoration Teacher. The expectation $\mathbb{E}[\cdot]$ is taken over the training batch.

4) Fusion Loss ($\mathcal{L}_{fus}$): The ultimate objective of our framework is to generate a single, high-fidelity fused image, O_S, that is both informative and visually compelling. We guide this final synthesis step by supervising the student's output, O_S, against the pristine target image from the Fusion Teacher, O_f. Our designed loss function strategically incorporates a photometric intensity term, $\mathcal{L}_{int}$, to precisely maintain the essential intensity distribution from the source images.

$$\mathcal{L}_{fus} = \mathbb{E}\left[\beta\mathcal{L}_{int}(O_S, O_f) + (1-\beta)\mathcal{L}_{grad}(O_S, O_f)\right], \tag{6}$$

Simultaneously, it includes a structural gradient term, $\mathcal{L}_{grad}$, which vigorously preserves fine-grained textural details and sharp edges. The intensity loss $\mathcal{L}_{int}$ is a pixel-wise L1 loss:

$$\mathcal{L}_{int}(O_S, O_f) = \mathcal{L}^1_{\text{pixel}}(O_S, O_f), \tag{7}$$

and the gradient loss $\mathcal{L}_{grad}$ measures the L1 difference in image gradients:

$$\mathcal{L}_{grad}(O_S, O_f) = \mathcal{L}^1_{\text{pixel}}(\partial_x O_S, \partial_x O_f) + \mathcal{L}^1_{\text{pixel}}(\partial_y O_S, \partial_y O_f), \tag{8}$$

This dual-component supervision strikes a balance between quantitative fidelity and perceptual realism in the fused image. ∂_x and ∂_y are the horizontal and vertical gradient operators.

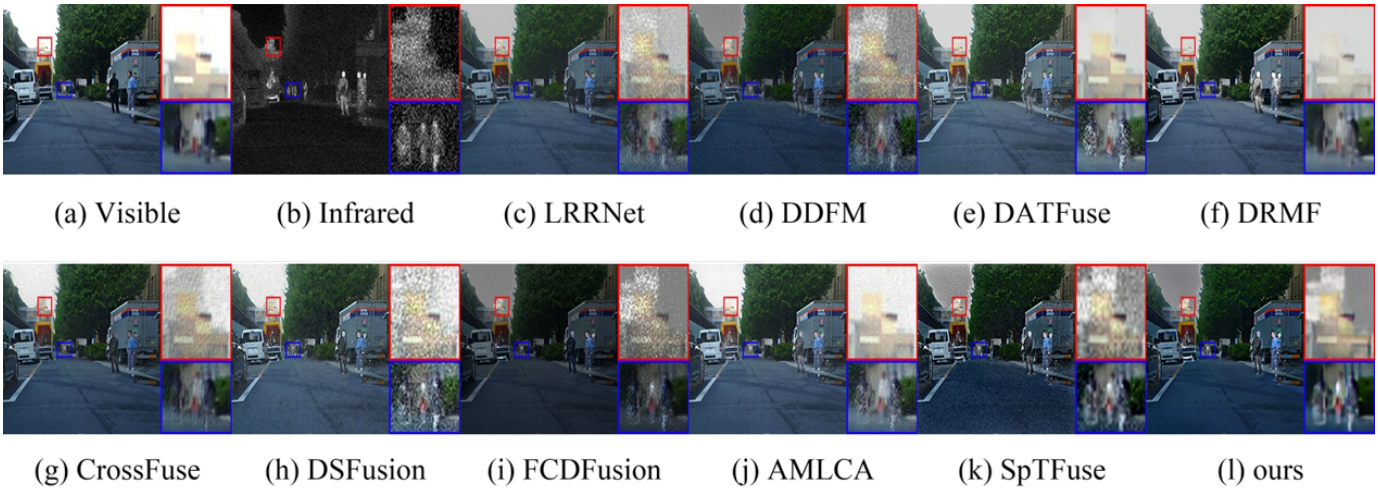


Fig. 4. Fusion results on the EMS dataset under noisy degraded scenarios. Red and blue boxes highlight the enlarged details of background textures and thermal targets, respectively. Our method effectively suppresses noise while maintaining sharp structural fidelity.

TABLE II
QUANTITATIVE COMPARISON ON THE EMS DATASETS. THE BEST AND SECOND-BEST RESULTS ARE HIGHLIGHTED.

Methods	EN ↑	CC ↑	MS-SSIM ↑	PSNR ↑	MSE ↓
LRRNet [21]	6.9415	0.5088	0.8466	14.0454	0.0444
DDFM [15]	6.9740	0.5730	0.8234	**14.6906**	0.0386
DATFuse [22]	6.8780	0.5254	0.9159	12.8681	0.0557
DRMF [23]	7.2239	0.4697	0.8979	11.4774	0.0757
CrossFuse [24]	7.0701	0.4637	0.8958	13.2795	0.0544
DSFusion [25]	7.2128	0.4976	0.8891	13.2117	0.0534
FCDFusion [26]	6.6250	**0.5760**	0.9058	14.6549	**0.0363**
AMCLA [27]	7.2789	0.5183	0.9489	13.0455	0.0548
SpTFuse [28]	**7.2954**	0.5148	**0.9330**	14.1534	0.0418
Ours	**6.8972**	**0.5713**	**0.9478**	**14.8216**	**0.0360**

IV. EXPERIMENTS

A. Experimental Setup

Datasets and Protocols. We evaluate our framework on three public benchmarks to cover diverse scenarios. For standard fusion tasks, we use the M^{3}FD [29] and RoadScene [17] datasets. M^{3}FD contains 4,200 image pairs captured in various environments, while RoadScene focuses on traffic scenarios. For robustness testing, we use the EMS dataset [30], which includes images with real-world degradations such as low light and bad weather. We adopt two evaluation protocols. In standard scenarios, we test the models on the original test sets. In degraded scenarios (EMS), we follow a "blind" evaluation protocol, where the model must process the inputs without prior knowledge of the degradation type. To ensure a fair and rigorous comparison, we follow the benchmarking strategy proposed in DAFusion [31] and establish two-stage baselines by combining specialized restoration models with fusion methods. Specifically, we use NeurBR [32] for low-light enhancement, DTR [33] for rain removal, and NDR [34] for denoising.

Evaluation Metrics. We use five quantitative metrics to evaluate the fused results, categorized into two groups. (1) *Information Retention*: We use Entropy (EN) and Correlation Coefficient (CC) to measure how much information is preserved from the source images. Higher scores indicate richer information content. (2) *Signal Fidelity*: We use Multi-scale Structural Similarity (MS-SSIM), Peak Signal-to-Noise Ratio (PSNR), and Mean Squared Error (MSE). These metrics assess the structural consistency and noise suppression capabilities of the model. For PSNR and MS-SSIM, higher values are better, while for MSE, lower values indicate less distortion.

Implementation Details. Our framework is implemented in PyTorch and trained on a single NVIDIA A100 GPU. During training, we randomly crop the input images into patches of size 128×128. We train the student network for 200 epochs with a batch size of 8. We use the AdamW optimizer with an initial learning rate of 1×10^{-4} and a weight decay of 1×10^{-5}. The learning rate is adjusted using a cosine annealing scheduler. The loss function weights are determined via grid search: $\lambda_{enc} = 0.5$, $\lambda_{dec} = 0.3$, $\lambda_{res} = 0.1$, and $\lambda_{fus} = 1.0$. The internal weights for the texture and intensity terms are set to $\alpha = 0.8$ and $\beta = 0.5$, respectively. Regarding the teacher networks, they are pre-trained on task-specific datasets. The Restoration Teacher is trained on degraded data to learn feature denoising, while the Fusion Teacher is trained on clean data to capture optimal fusion semantics. Both teachers are frozen during the student's training phase to ensure stable guidance.

B. Comparison with SOTA Methods

1) Performance on Standard Datasets: Table I presents the quantitative results on standard benchmarks. On the M^{3}FD dataset, our D^{2}T framework achieves the best scores in four out of five metrics. The high PSNR and MS-SSIM values indicate that our model preserves structural details and reduces spectral distortion more effectively than competing methods.

Results on the RoadScene dataset further confirm this generalization capability. Here, our method achieves the highest PSNR and MSE, and ranks within the top two for the other metrics. These consistent results demonstrate the effectiveness of our quadruple-supervision strategy in guiding the network to learn accurate feature representations.

Fig. 3 presents a visual comparison of different fusion methods. Autoencoder-based methods, such as LRRNet and DATFuse, often fail to preserve thermal intensity. Consequently, the fused images suffer from low global contrast, where salient infrared targets are attenuated and merged into the background. On the other hand, methods like DRMF and AMLCA tend to aggressively enhance the infrared features. This leads to artificial ringing artifacts (halos) around edges and distorts the natural color distribution of the visible background. In contrast, our method achieves a superior balance between spectral consistency and texture fidelity. It highlights salient infrared objects with high contrast while strictly preserving the fine-grained textures and natural colors of the visible scene, producing fused images with high perceptual quality.

2) Performance on Degraded Datasets: Table II shows the results on the degraded EMS test set. Our end-to-end D^{2}T framework consistently outperforms the two-stage baselines. Note that these baselines theoretically have an advantage because they use specific restoration models (NeurBR, DTR, NDR) for known degradation types. However, our unified model still achieves higher scores in key metrics like PSNR and MSE. This difference highlights a major limitation of two-stage methods: error propagation. Artifacts produced during the restoration step are often amplified by the fusion network. In contrast, our approach avoids this issue by optimizing restoration and fusion jointly, allowing the model to learn robust features directly from the input.

Visual results in Fig. 4 further confirm this robustness. In challenging scenarios involving rain or low light, two-stage methods often exhibit noticeable degradation. For instance, the restoration pre-processing tends to over-smooth high-frequency details or introduce unnatural color shifts due to the disjoint optimization goals. In contrast, our D^{2}T framework effectively handles unknown degradations without manual intervention. It directly generates clean, structurally consistent fused images, preserving sharper edges and richer textures while maintaining natural color gradients. This validates that our decoupled supervision strategy successfully guides the network to separate valid semantic information from complex environmental noise.

C. Ablation Study

To comprehensively validate the effectiveness of our decoupled supervision strategy and the necessity of each component, we conducted a progressive ablation study on the EMS dataset. The student network architecture remained fixed across all variants to ensure a fair comparison of the supervision strategies.

First, we evaluated the baseline performance in stage (a), where the student network was supervised solely by the Fusion Teacher using the fusion loss. As shown in Table III, this configuration yields suboptimal performance, with a PSNR of 13.98 dB. This result indicates that without explicit guidance on degradation removal, the network struggles to handle noisy inputs effectively, as the fusion objective alone cannot distinguish between high-frequency noise and valid textural details.

Fig. 5. Visualization of the ablation study on the EMS dataset with noisy infrared inputs. From (a) to (c), the progressive integration of restoration and coupled distillation losses leads to cleaner backgrounds and sharper target edges.

TABLE III
ABLATION STUDY OF OUR PROPOSED SUPERVISION STRATEGY ON THE EMS DATASET. THE **BEST** RESULTS ARE IN BOLD.

Model Variant	MS-SSIM ↑	PSNR ↑	EN ↑
(a) Baseline ($+\mathcal{L}_{fus}$)	0.8851	13.9875	6.7512
(b) w/ Restoration ($+\mathcal{L}_{res}$)	0.9189	14.3521	6.8245
(c) **Ours (Full Model)**	**0.9478**	**14.8216**	**6.8972**

Next, in stage (b), we introduced the Restoration Teacher and enabled the restoration loss. This step implements our core decoupling mechanism. By delegating the restoration supervision to a dedicated expert branch, the student network is guided to explicitly suppress noise artifacts before the fusion stage. This modification leads to a significant performance gain of 0.37 dB in PSNR. This improvement confirms that decoupling the restoration task from the fusion process effectively mitigates the inherent conflict between noise removal and detail preservation, allowing the network to learn cleaner representations.

Finally, in stage (c), we incorporated the feature-level distillation losses for both the encoder and decoder. Unlike stage (b), which only constrains the final output, these losses enforce a strict multi-level feature alignment. They compel the intermediate features of the student to mimic the noise-robust representations of the Restoration Teacher and the structure-rich semantics of the Fusion Teacher. Table III demonstrates that this strategy further boosts PSNR by 0.47 dB and MS-SSIM by 0.029. This validates that feature-level constraints are essential for recovering fine-grained textures and sharp edges that might be under-optimized by pixel-level supervision alone.

These quantitative findings are visually corroborated by Fig. 5. The baseline model (a) exhibits noticeable residual noise, while model (b) improves image clarity but results in slightly blurred details. The full model (c) achieves the best visual quality, effectively balancing noise suppression with the preservation of sharp structural edges.

V. CONCLUSION

In this work, we proposed the D^2T framework, a dual-teacher distillation approach designed to resolve the inherent conflict between degradation removal and image fusion. By decoupling these distinct objectives, we assigned noise suppression to a Restoration Teacher and texture preservation to a Fusion Teacher, which effectively stabilizes the learning process. Comprehensive evaluations on public benchmarks confirm that D^2T achieves competitive performance and strong robustness to unseen degradations while maintaining high efficiency for end-to-end deployment.

REFERENCES

[1] Y. Zhou, K. He, D. Xu, D. Tao, X. Lin, and C. Li, "Asfusion: Adaptive visual enhancement and structural patch decomposition for infrared and visible image fusion," *Engineering Applications of Artificial Intelligence*, vol. 132, p. 107905, 2024.

[2] J. Liu, Z. Liu, G. Wu, L. Ma, R. Liu, W. Zhong, Z. Luo, and X. Fan, "Multi-interactive feature learning and a full-time multi-modality benchmark for image fusion and segmentation," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 8081–8090.

[3] J. Ma, Y. Ma, and C. Li, "Infrared and visible image fusion methods and applications: A survey," *Information Fusion*, vol. 45, pp. 153–178, 2019.

[4] H. Li and X.-J. Wu, "Densefuse: A fusion approach to infrared and visible images," *IEEE Transactions on Image Processing*, vol. 28, no. 5, pp. 2614–2623, 2019.

[5] 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*, vol. 9, no. 7, pp. 1200–1217, 2022.

[6] L. Tang, X. Xiang, H. Zhang, M. Gong, and J. Ma, "Divfusion: Darkness-free infrared and visible image fusion," *Information Fusion*, vol. 91, pp. 477–493, 2023.

[7] X. Wang, Z. Guan, W. Qian, J. Cao, R. Ma, and C. Bi, "A degradation-aware guided fusion network for infrared and visible image," *Information Fusion*, vol. 118, p. 102931, 2025.

[8] W. Xiao, Y. Zhang, H. Wang, F. Li, and H. Jin, "Heterogeneous knowledge distillation for simultaneous infrared-visible image fusion and super-resolution," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–15, 2022.

[9] Y. Zhang, Y. Liu, P. Sun, H. Yan, X. Zhao, and L. Zhang, "Ifcnn: A general image fusion framework based on convolutional neural network," *Information Fusion*, vol. 54, pp. 99–118, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1566253518305505

[10] R. Du, J. Cheng, H. Zhao, and Y. Guo, "Mmfi-net: Multilevel multimodal feature injection network for infrared and visible image fusion," in *2024 International Joint Conference on Neural Networks (IJCNN)*, 2024, pp. 1–8.

[11] H. Li, X.-J. Wu, and T. Durrani, "Nestfuse: An infrared and visible image fusion architecture based on nest connection and spatial/channel attention models," *IEEE Transactions on Instrumentation and Measurement*, vol. 69, no. 12, pp. 9645–9656, 2020.

[12] H. Li, X.-J. Wu, and J. Kittler, "Rfn-nest: An end-to-end residual fusion network for infrared and visible images," *Information Fusion*, vol. 73, pp. 72–86, 2021.

[13] 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 (CVPR)*, June 2023, pp. 5906–5916.

[14] J. Ma, W. Yu, P. Liang, C. Li, and J. Jiang, "Fusiongan: A generative adversarial network for infrared and visible image fusion," *Information Fusion*, vol. 48, pp. 11–26, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1566253518301143

[15] 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 *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 8048–8059.

[16] L. Tang, J. Yuan, H. Zhang, X. Jiang, and J. Ma, "Seafusion: Overcoming the ambiguity of features in infrared and visible image fusion by semantic-aware fusion," *Information Fusion*, vol. 83, pp. 1–13, 2022.

[17] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 5792–5801.

[18] Z. Liu, J. Liu, B. Zhang, L. Ma, X. Fan, and R. Liu, "Paif: Perception-aware infrared-visible image fusion for attack-tolerant semantic segmentation," *ACM MM*, 2023.

[19] A. Yu, Y. Li, and W. Feng, "Moefuse: Degradation-aware fusion of infrared and visible images using mixture of experts," in *2025 International Joint Conference on Neural Networks (IJCNN)*, 2025, pp. 1–8.

[20] W. Xue, Y. Liu, F. Wang, G. He, and Y. Zhuang, "A novel teacher–student framework with degradation model for infrared–visible image fusion," *IEEE Transactions on Instrumentation and Measurement*, vol. 73, pp. 1–12, 2024.

[21] H. Li, T. Xu, X.-J. Wu, J. Lu, and J. Kittler, "Lrrnet: A novel representation learning guided fusion network for infrared and visible images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 9, pp. 11 040–11 052, 2023.

[22] 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*, vol. 33, no. 7, pp. 3159–3172, 2023.

[23] 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 32nd ACM International Conference on Multimedia*, 2024, p. 8546–8555.

[24] H. Li and X.-J. Wu, "Crossfuse: A novel cross attention mechanism based infrared and visible image fusion approach," *Information Fusion*, vol. 103, p. 102147, 2024.

[25] K. Liu, M. Li, C. Chen, C. Rao, E. Zuo, Y. Wang, Z. Yan, B. Wang, C. Chen, and X. Lv, "Dsfusion: Infrared and visible image fusion method combining detail and scene information," *Pattern Recognition*, vol. 154, p. 110633, 2024.

[26] H. Li and Y. Fu, "Fcdfusion: A fast, low color deviation method for fusing visible and infrared image pairs," *Computational Visual Media*, vol. 11, no. 1, pp. 195–211, 2025.

[27] D. Wang, C. Huang, H. Pan, Y. Sun, J. Dai, Y. Li, and Z. Ren, "Amlca: Additive multi-layer convolution-guided cross-attention network for visible and infrared image fusion," *Pattern Recognition*, vol. 163, p. 111468, 2025.

[28] L. Guo, X. Luo, Y. Liu, Z. Zhang, and X. Wu, "Sam-guided multi-level collaborative transformer for infrared and visible image fusion," *Pattern Recognition*, vol. 162, p. 111391, 2025.

[29] 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*, vol. 44, no. 1, pp. 502–518, 2022.

[30] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 27 016–27 025.

[31] X. Wang, Z. Guan, W. Qian, J. Cao, R. Ma, and C. Bi, "A degradation-aware guided fusion network for infrared and visible image," *Information Fusion*, vol. 118, p. 102931, 2025.

[32] Z. Zhao, H. Lin, D. Shi, and G. Zhou, "A non-regularization self-supervised retinex approach to low-light image enhancement with parameterized illumination estimation," *Pattern Recognition.*, vol. 146, no. C, Feb. 2024.

[33] P. W. Patil, S. Gupta, S. Rana, S. Venkatesh, and S. Murala, "Multi-weather image restoration via domain translation," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 21 639–21 648.

[34] M. Yao, R. Xu, Y. Guan, J. Huang, and Z. Xiong, "Neural degradation representation learning for all-in-one image restoration," *IEEE Transactions on Image Processing*, vol. 33, pp. 5408–5423, 2024.