Distilling Textual Priors from LLM to Efficient Image Fusion

Ran Zhang, Xuanhua He, Ke Cao, Liu Liu,
Li Zhang, Man Zhou, Jie Zhang, Dan Guo and Meng Wang *Fellow, IEEE*

***Abstract*—Multi-modality image fusion aims to synthesize a single, comprehensive image from multiple source inputs. Traditional approaches, such as CNNs and GANs, offer efficiency but struggle to handle low-quality or complex inputs. Recent advances in text-guided methods leverage large model priors to overcome these limitations, but at the cost of significant computational overhead, both in memory and inference time. To address this challenge, we propose a novel framework for distilling large model priors, eliminating the need for text guidance during inference while dramatically reducing model size. Our framework utilizes a teacher-student architecture, where the teacher network incorporates large model priors and transfers this knowledge to a smaller student network via a tailored distillation process. Crucially, our experiments demonstrate that this knowledge transfer is the primary driver of performance gains, rather than mere architectural optimization. Additionally, we introduce a spatial-channel cross-fusion module to enhance the model's ability to leverage textual priors across both spatial and channel dimensions. Our method achieves a favorable trade-off between computational efficiency and fusion quality. The distilled network, requiring only 10% of the parameters and inference time of the teacher network, retains 90% of its performance and outperforms existing SOTA methods. Extensive experiments demonstrate the effectiveness of our approach. Codes are available at https://github.com/Zirconium233/DTPF**

***Index Terms*—Image Fusion, Knowledge Distillation, Large Language Models, Multi-modality Learning.**

I. Introduction

IMAGE fusion plays an important role in visual enhancement within digital image processing. For example, visible-infrared image fusion combines color-based details from visible images, which are easily interpretable by humans, with radiation-based features from infrared images, which are highly effective for target detection and operations under low-light conditions. By integrating complementary information from both modalities, this fusion method produces high-quality images that enhance both human visual interpretation and machine-based detection performance. This work also shares similarities with pan-sharpening, a multi-modal image fusion task that has been extensively explored in this journal through various lenses, including universal Transformer architectures [1], semantic-aware fusion [2], and frequency domain analysis, as explored in recent studies such as [3].

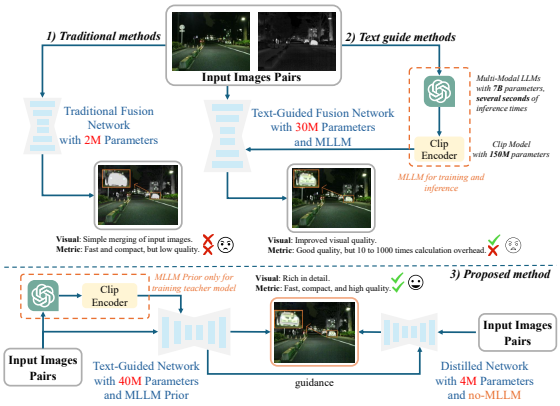


Fig. 1. Overview of different image fusion methods and their parameter efficiency. (1) Traditional methods use small fusion networks. (2) Text-guided methods significantly increasing computational demands with performance improvements. (3) Our proposed method leverages text-guided training and knowledge distillation to create a distilled network that achieves high-quality fused images without relying on LLMs during inference.

During the imaging process, environmental and device-related constraints often degrade the quality of source images. Visible images may suffer from low resolution, while infrared images are prone to various noise types. Traditional fusion methods, including FusionGAN [4], U2Fusion [5], and SwinFusion [6], fuse source images, often with a weak semantic understanding of the scene, without effectively distinguishing between degraded information and meaningful image content, leading to suboptimal performance. Some approaches [7] mitigate these issues by relying on manual pre-processing to enhance source images, but their flexibility is limited. Recently, text-guided methods, such as TextFusion [8] and Text-IF [9], have emerged as a promising solution by introducing high-level semantic priors to address these challenges. These methods leverage the capabilities of multimodal large language models (MLLMs) to generate text captions or degradation descriptions (e.g., low light) for source images, enabling the model to adaptively distinguish image content from degradation and enhance image quality. By integrating these priors, the model gains enhanced semantic understanding of the image content, which effectively promotes the fusion processes. For instance, Text-IF utilizes MLLMs [10] to generate task lists and incorporates CLIP [11] to guide the fusion process based on these descriptions. Similarly, TextFusion [8] employs

This work was supported in part by the Hefei Municipal Natural Science Foundation under Grant HZR2417, and in part by the National Natural Science Foundation of China under Grant 62302143. (Corresponding authors: Jie Zhang and Liu Liu; Project leader: Xuanhua He.)

R. Zhang and L. Liu are with Hefei University of Technology, Hefei, China (e-mail: 2023212219@mail.hfut.edu.cn; liuliu@hfut.edu.cn).

X. He, K. Cao, L. Zhang, and M. Zhou are with the University of Science and Technology of China, Hefei, China (e-mail: hexuanhua@mail.ustc.edu.cn; caoke200820@mail.ustc.edu.cn; zanly20@mail.ustc.edu.cn; manman@mail.ustc.edu.cn).

D. Guo and M. Wang are with Hefei University of Technology, Hefei, China (e-mail: guodan@hfut.edu.cn; wangmeng@hfut.edu.cn).

J. Zhang is with Hefei Institutes of Physical Science, Chinese Academy of Sciences, Hefei, China and Hefei Institutes of Innovation, Chinese Academy of Sciences (e-mail: zhangjie@iim.ac.cn).

text-based interactions to build models capable of focusing on specific image elements, such as emphasizing trees over people, based on captions generated by LLMs or human input.

Despite their advantages, incorporating LLMs into the fusion workflow introduces significant resource overhead. In most cases, users aim to fuse images for downstream tasks such as detection or segmentation, without the need for extensive interaction. LLMs dramatically increase computational demands, often occupying ten to thousands of times more resources than the fusion module itself. Even CLIP, a relatively lightweight vision-language model, requires approximately five times the parameters of the fusion module, creating a significant deployment bottleneck. This computational cost makes LLMs disproportionate to the task. Moreover, due to the limitations in the quality and quantity of datasets, the performance gains from integrating large models are marginal compared to the significant decrease in efficiency. Thus, we seek a method that enables the model to perform degradation-aware fusion and enhances semantic understanding without relying on large-scale models during inference.

In response to these issues, we propose a large-model prior distillation framework that eliminates the need for text guidance during inference while significantly reducing model size, as shown in Fig. 1. Our framework employs a teacher-student architecture, where the teacher network integrates MLLM priors to learn a set of powerful, semantically-aware fusion behaviors. To fully exploit textual priors across both spatial and channel dimensions, we introduce a spatial-channel cross-fusion module (SCFM). A specially tailored prior distillation loss is then used to transfer textual knowledge from both feature and output perspectives, enabling the student network to mimic the teacher's feature processing without requiring textual inputs. This knowledge transfer is the key to our method's success, allowing a compact network to achieve performance unattainable through architectural optimization alone. By leveraging this approach, our method achieves a 90% reduction in model size (excluding CLIP, which alone requires five times the parameters of the original fusion module) while retaining 90% of the teacher model's performance.

Our contributions are summarized as follows:

- We propose a novel teacher-student framework to distill textual information from MLLMs, effectively embedding their semantic priors into an efficient network, and enhance the image fusion process without interfering with inference. A tailored prior distillation loss transfers knowledge from the teacher network to the student, enabling the student network to internalize textual information while reducing model size.
- We introduce a spatial-channel cross-fusion module to fully exploit textual information across both spatial and channel dimensions, improving the model's ability to leverage textual priors.
- Extensive experiments on multiple datasets demonstrate that our method achieves state-of-the-art (SOTA) performance. Ablation studies further validate the effectiveness of our approach, highlighting its ability to achieve an optimal balance between efficiency and performance while providing valuable insights into model compression and the utilization of prior information.

II. Related Works

Image fusion has been a well-explored research area, and significant progress has been achieved with the advancements in deep learning. Early methods primarily utilized CNNs [5] to fuse images from different modalities. Subsequently, generative models [4], [12] and Transformer-based approaches [6], [13] were introduced to enhance fusion quality. Additionally, high-level vision tasks, such as object detection, have been incorporated to guide the fusion process [14]. However, these methods often overlook the degradation present in source images, resulting in suboptimal performance. To address this limitation, recent works, such as Text-IF [9] and Text-Fusion [8], leverage large model priors to achieve degradation-aware fusion. While effective, these approaches introduce substantial computational overhead, limiting their practicality for real-world applications.

Knowledge distillation has gained prominence as a strategy for compressing deep neural networks by transferring knowledge from a large "teacher" model to a smaller "student" model. This approach was initially popularized for reducing the computational overhead of deploying models on resource-constrained devices [15]. Distillation techniques include output softening, where the student mimics the probability distribution of the teacher, and intermediate-layer guidance, which aligns feature representations [16]. In multimodal domains, KD has enabled efficient training of smaller models for tasks involving text and vision, such as CLIP distillation [17]. Notably, KD has been explored in text-guided image generation and fusion, where computationally heavy models like Stable Diffusion serve as teachers for lightweight networks [18]. However, the distillation of semantic priors from LLMs and multimodal models in the domain of image fusion remains an open problem.

Large Vision-Language Models in Image Processing The advent of Large Language Models (LLMs) and Large Vision-Language Models (VLMs) has catalyzed a paradigm shift in computer vision. Models like CLIP [11] demonstrated the power of learning transferable visual representations from natural language supervision. This has been extended by numerous powerful VLMs, such as the GPT series [19], LLaVA, and Qwen-VL [10], [20], [21], which can perform complex reasoning and generate detailed textual descriptions of visual content. In image fusion, these models offer a unique opportunity to guide the process with high-level semantic priors. As pioneering work, Text-IF [9] leverages VLM-generated text to make the fusion process degradation-aware and interactive. However, this comes at a significant computational cost. In this work, we select Qwen-VL as our primary teacher model due to its strong open-source availability, versatile vision-language capabilities, and its proficiency in generating detailed, relevant descriptions for image fusion tasks, as we will demonstrate in our experiments. Our framework aims to distill this powerful but expensive textual guidance into an efficient network, a direction that remains underexplored.

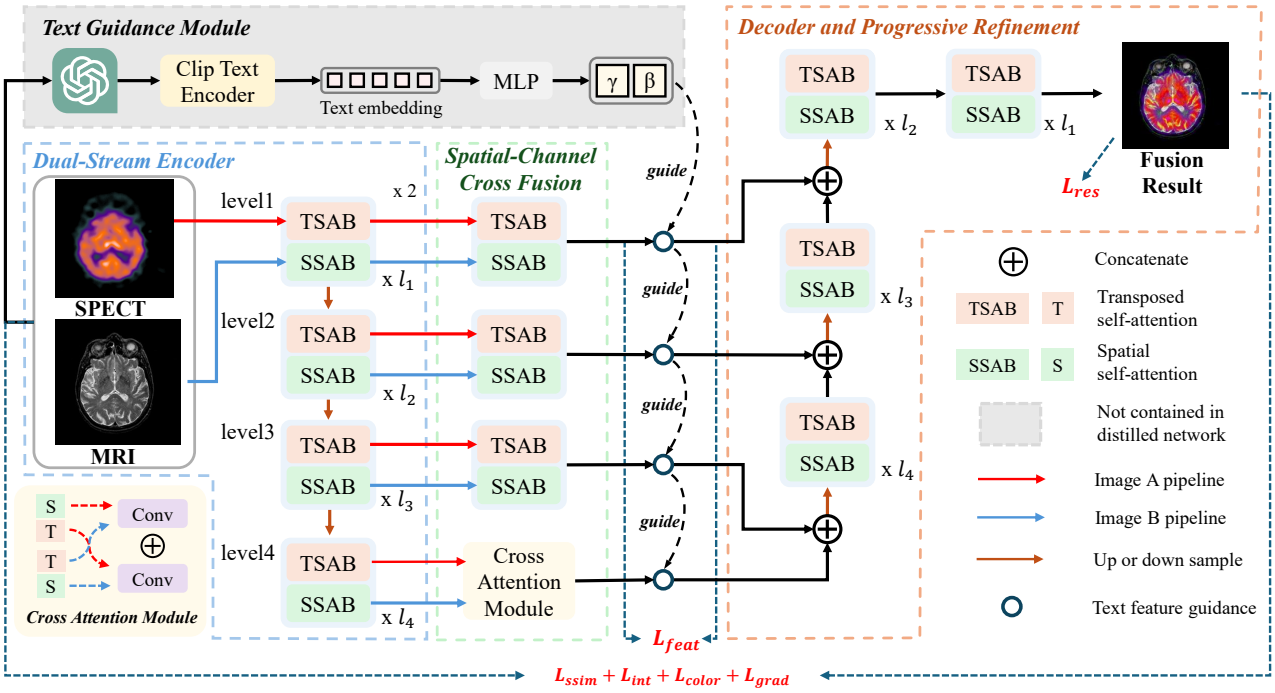


Fig. 2. Overview of our text-guided image fusion framework. The architecture consists of three main components: (1) Text Guidance Module that leverages LLMs and CLIP to generate semantic guidance; (2) Encoder that processes visible and infrared inputs through dual-stream transformers with TSAB and SSAB blocks, followed by cross-modal fusion; (3) Decoder and Refinement that progressively reconstructs the fused image with text-guided feature modulation. The full architecture represents the teacher network used during training. The distilled student network shares the same core structure but excludes the gray-shaded components (Text Guidance Module and text feature guidance), and operates with reduced channel dimensions.

III. Method

A. Framework

Traditional image fusion methods formulate the task as mapping two source images (e.g., $\mathbf{I_{vis}}$, $\mathbf{I_{ir}}$) to a fused output $\mathbf{I_f}$ through a fusion network $\Phi(.)$:

$$\mathbf{I_f} = \Phi(\mathbf{I_{vis}}, \mathbf{I_{ir}}) \tag{1}$$

Text-guided methods enhance this process by incorporating additional textual input, generated by large language models and encoded by CLIP to produce $\mathbf{T}$. The fused image $\mathbf{I_f}$ is then generated using $\mathbf{T}$ as guidance:

$$\mathbf{I_f} = \Phi(\mathbf{I_{vis}}, \mathbf{I_{ir}}, \mathbf{T}) \tag{2}$$

However, the computational cost of LLMs and CLIP is substantial. To mitigate this, we propose a prior distillation framework consisting of a text-guided teacher network Φ_t and its distilled student network Φ_s. During training, Φ_t utilizes textual information from MLLMs, which is then transferred to Φ_s through our tailored prior distillation loss. Since semantic information is encoded across both spatial and channel dimensions, we introduce spatial-channel cross fusion blocks within the network. As illustrated in Fig. 2, this framework enables the student network to internalize textual information while significantly reducing model size. Both networks share similar structural blocks such as dual-stream encoder, spatial-channel cross fusion module, decoder and refinement module, with the student network being more compact.

B. Network Architecture

Our framework consists of two main components: a text-guided teacher network and a lightweight student network. The teacher network serves as a proxy, leveraging priors from the LLM and transferring this information to the student network in the form of features. The student network mimics the feature processing of the teacher, thus acquiring the semantic knowledge indirectly from the LLM. This enables the student network to maintain a small model size while being adaptive to image degradation, without requiring direct text input.

1) Text-Guided Teacher Network: The teacher network adopts a hierarchical structure with four levels of feature processing. Given its critical role in leveraging textual priors, the core operator aims to effectively model features in spatial and channel dimensions [22]. Additionally, we introduce spatial-channel cross fusion blocks as the core operator, enabling the exploration of correlations between spatial and channel dimensions. Text information is fused through the sequential feature modulation, ensuring effective integration of textual priors. This network is organized into three primary components:

1) **Dual-Stream Encoder**: Processes visible and infrared inputs through parallel branches, each consisting of overlapped patch embedding layers and spatial and channel self-attention blocks. The hierarchical feature transformation at level l is expressed as:

$$F_l^i = \mathcal{F}_n \circ \mathcal{F}_{n-1} \circ \cdots \circ \mathcal{F}_1(F_{l-1}^i) \tag{3}$$

where $\mathcal{F}_k$ represents the k-th spatial-channel self-attention block for modality $i \in \{\text{vis}, \text{ir}\}$, and $\circ$ denotes func-

tion composition. Each block integrates sequential Windowed Self-Attention Blocks (SSAB) and Transposed Self-Attention Blocks (TSAB) for comprehensive feature encoding:

$$\text{SSAB}(\mathbf{F}) = \mathbf{F} + \text{LN}(\text{MHA}_s(\mathbf{F})) \tag{4}$$

$$\text{TSAB}(\mathbf{F}) = \mathbf{F} + \text{LN}(\text{MHA}_c(\mathbf{F})) \tag{5}$$

Here, $\text{MHA}_c(.)$ and $\text{MHA}_s(.)$ denote multi-head attention mechanisms for channel and spatial dimensions [23], [24], respectively, and LN represents layer normalization.

2) **Spatial-Channel Cross Fusion**: To fully leverage prior information and establish the connection between two modalities, we propose the spatial-channel cross fusion module. At the upper level, features are concatenated and fused; at the lower level, features from both modalities are fused:

$$\hat{\mathbf{F}_l^{vis}} = \text{Conv}_{1\times1} \circ C(\text{SSAB}(\mathbf{F}_l^{vis}), \text{TSAB}(\mathbf{F}_l^{ir})), \tag{6}$$

$$\hat{\mathbf{F}_l^{ir}} = \text{Conv}_{1\times1} \circ C(\text{SSAB}(\mathbf{F}_l^{ir}), \text{TSAB}(\mathbf{F}_l^{vis})), \tag{7}$$

$$\mathbf{F}_l^{fused} = \hat{\mathbf{F}_l^{vis}} + \hat{\mathbf{F}_l^{ir}}. \tag{8}$$

where $C(.)$ denotes concatenation. The fused features are then modulated by the text embeddings:

$$\hat{\mathbf{F}_l^{fused}} = (1 + \gamma(\mathbf{T})) \odot \mathbf{F}_l^{fused} + \beta(\mathbf{T}) \tag{9}$$

where $\gamma(\mathbf{T})$ and $\beta(\mathbf{T})$ are learnable modulation parameters derived from the text embeddings $\mathbf{T}$, and $\odot$ denotes element-wise multiplication.

3) **Decoder and Progressive Refinement**: The decoder progressively refines features through:

$$\mathbf{F}_{dec}^l = \mathcal{D}_n \circ \text{Up}(\text{C}(\mathbf{F}_{dec}^{l+1}, \hat{\mathbf{F}_{fused}^l})) \tag{10}$$

where $\mathcal{D}_n$ represents a sequence of n decoder blocks similar to encoder. Specially, for the lowest level($l = 1$), we add extra decoder blocks as refinement:

$$\mathbf{F}_{out} = \mathcal{R}_m \circ \mathcal{R}_{m-1} \circ \ldots \circ \mathcal{R}_1(\mathbf{F}_{dec}^1) \tag{11}$$

Each block $\mathcal{R}_k$ consists of SSAB and TSAB operations sharing block number with the encoder at the corresponding level. At the end, the $\mathbf{F}_{out}$ is the final result $\mathbf{I}_{fused}$.

2) Lightweight Student Network: The student network maintains a similar hierarchical structure but achieves 90% parameter reduction through:

- Dimension reduction in transformer blocks (from 48 to 16 channels)
- Removal of text guidance module(the gray-shaded component in 2), directly using $\mathbf{F}_{fused}^l$ instead of $\hat{\mathbf{F}_{fused}^l}$

C. Prior Knowledge Distillation

To achieve efficient inference without relying on MLLMs while preserving their semantic prior and reducing model complexity, we propose a two-stage knowledge distillation method to progressively transfer knowledge from MLLMs to a lightweight student network.

1) Teacher Network Training: In the first stage, we transfer knowledge from the MLLM to a teacher network, enabling the teacher network to acquire semantic understanding and degradation-aware fusion capabilities. The teacher network is trained with text guidance using the following loss:

$$\mathcal{L}_{tea} = \lambda_{ssim}^t \mathcal{L}_{ssim} + \lambda_{int}^t \mathcal{L}_{int} + \lambda_{grad}^t \mathcal{L}_{grad} + \lambda_{color}^t \mathcal{L}_{color} \tag{12}$$

where $\mathcal{L}_{ssim}$ measures structural similarity through SSIM index, $\mathcal{L}_{int}$ preserves maximum intensity information between source images, $\mathcal{L}_{grad}$ ensures gradient consistency using Sobel operators, and $\mathcal{L}_{color}$ maintains color fidelity in YCbCr space. The weights λ^t are dynamically adjusted based on text guidance. The detailed expressions for each loss component are:

$$\mathcal{L}_{int} = \frac{1}{HW} \|I_f - \max(I_{vis}^g, I_{ir}^g)\|_1 \tag{13}$$

where H and W are the height and width of the image, I_f is the fused image, and I_{vis}^g and I_{ir}^g represent the visible and infrared guidance images respectively.

$$\mathcal{L}_{SSIM}(t) = (1-SSIM(I_f, I_{vis}^g)) + \delta_{ir}(t)(1-SSIM(I_f, I_{ir}^g)) \tag{14}$$

where $SSIM(\cdot)$ calculates the structural similarity index, and $\delta_{ir}(t)$ is a text-dependent weight for infrared guidance.

$$\mathcal{L}_{grad} = \frac{1}{HW} \|\nabla I_f - \max(\nabla I_{vis}^g, \nabla I_{ir}^g)\|_1 \tag{15}$$

where ∇ represents the gradient operator implemented using Sobel filters in both horizontal and vertical directions.

$$\mathcal{L}_{color} = \frac{1}{HW} \|F_{CbCr(I_f)} - F_{CbCr(I_{vis}^g)}\|_1 \tag{16}$$

where F_{CbCr} denotes the transfer function of RGB to YCbCr.

These loss terms work together to ensure high-quality image fusion: $\mathcal{L}_{int}$ maintains the maximum intensity information from both source images, $\mathcal{L}_{SSIM}$ preserves structural details through the SSIM [25] metric, $\mathcal{L}_{grad}$ ensures edge consistency through gradient preservation, and $\mathcal{L}_{color}$ maintains natural color appearance. The text-dependent weights λ^t allow dynamic adjustment of each loss component's contribution based on the specific fusion requirements described in the text prompt.

2) Progressive Knowledge Transfer: In the second stage, the student network is trained to progressively acquire knowledge from the teacher network by mimicking its feature processing through three complementary loss functions:

1) Feature consistency loss:

$$\mathcal{L}_{feat} = \sum_{l=1}^{L} \|\text{Down}(F_t^l) - F_s^l\|_1 \tag{17}$$

where Down represents learnable dimension reduction to match teacher-student feature dimensions.

2) Output reconstruction loss:

$$\mathcal{L}_{res} = \|I_t^{fused} - I_s^{fused}\|_1 \tag{18}$$

3) Base fusion loss $\mathcal{L}_{base}$ inherited from teacher but with fixed weights.

The final distillation objective is:

$$\mathcal{L}_{distill} = \alpha_1 \mathcal{L}_{base} + \alpha_2 \mathcal{L}_{feat} + \alpha_3 \mathcal{L}_{res} \tag{19}$$

This progressive distillation strategy enables knowledge transfer while reducing computational efficiency.

IV. Experiments

We conduct extensive experiments to evaluate our proposed method on both infrared-visible fusion (IVF) and medical image fusion tasks. In this section, we first introduce the implementation details and datasets, then present comprehensive comparisons with state-of-the-art methods, followed by detailed ablation studies.

A. Implementation Details and Datasets

Implementation Details: Both teacher and student networks are trained using AdamW optimizer with a learning rate of 0.0001. Input images are cropped to 128×128 patches during training. For the teacher network, the hyper-parameters $\{\lambda_{int}, \lambda_{ssim}, \lambda_{grad}, \lambda_{color}\}$ are set to $\{24, 40, 48, 12\}$ respectively. When text guidance is enabled, task-specific parameters are provided in the supplementary material. We employ QWen2VL [20], an open-source large vision-language model, to generate descriptive text labels for input image pairs. All experiments are conducted on four NVIDIA GeForce RTX 4090 GPUs using PyTorch framework.

Datasets: For IVF tasks, we evaluate on three widely-used datasets: MSRS [26], M3FD [27], and RoadScene [28]. These datasets cover diverse scenarios including urban scenes, indoor environments, and road conditions with varying lighting and weather conditions. We use the training split of MSRS for training and evaluate on its test split as well as the entire M3FD and RoadScene datasets. For medical image fusion, we utilize the Harvard Medical Image Fusion Dataset, which contains three modality pairs: SPECT-MRI, CT-MRI, and PET-MRI. Each modality pair is trained and evaluated separately.

Evaluation Metrics: For IVF tasks, we employ three complementary metrics: Information Entropy (EN) [29] to quantify information density, Visual Information Fidelity (VIF) [30] to evaluate perceptual quality, and Quality of Gradient-based Fusion ($Q^{AB/F}$) [31] to assess edge preservation. For medical image fusion, we include Structural Similarity (SSIM) [25] due to its critical role in medical applications, which is calculated as the sum of the SSIM values between each source image and the fused image. The metrics are chosen for their strong correlation with the source images, as they take all of the images as input.

A specific note on the use of the Structural Similarity (SSIM) metric is warranted. While SSIM is a crucial metric for medical image fusion, where preserving fine structural details is paramount, its application to Infrared-Visible Fusion (IVF) tasks can sometimes be misleading. Aggressively optimizing such as high weight of L_{ssim} can lead to high scores at the expense of noticeable color artifacts and an unnatural appearance 7. This observation informs our decision not to use SSIM as a primary metric for IVF, an approach consistent with prior works like Text-IF [9], which also favored other metrics over SSIM for this task. Therefore, we prioritize VIF and $Q^{AB/F}$ for assessing perceptual quality and gradient preservation in IVF. For medical tasks, however, we continue to report SSIM as it remains a reliable indicator of fusion performance in that domain.

Benchmark Methods: We compare our teacher and distilled networks with several state-of-the-art methods, including SwinFusion [6], U2Fusion [5], CDDFuse [32], FusionGAN [4], PIAFusion [7], SuperFusion [33], IFCNN [34], DDFM [12], Text-IF [9], TC-MOA [35], PSLPT [36], EMFusion [37], Zero [38], and MSRPAN [39]. These methods include transformer-based, unsupervised learning, dual-branch, generative models, illumination- and semantic-aware fusion, CNN-based fusion, diffusion models, and text-guided fusion (Text-IF). Beyond performance comparisons, we focus on analyzing the impact of different components and parameters on both model effectiveness and inference efficiency through comprehensive ablation studies.

B. Comparison on IVF Tasks

Table I presents the quantitative comparison with SOTA methods on three IVF datasets. Our teacher network consistently outperforms existing methods across all metrics and datasets, demonstrating the effectiveness of text-guided fusion. On the MSRS dataset, our teacher network achieves significant improvements in $Q^{AB/F}$ compared to the previous best method Text-IF. More remarkably, our distilled

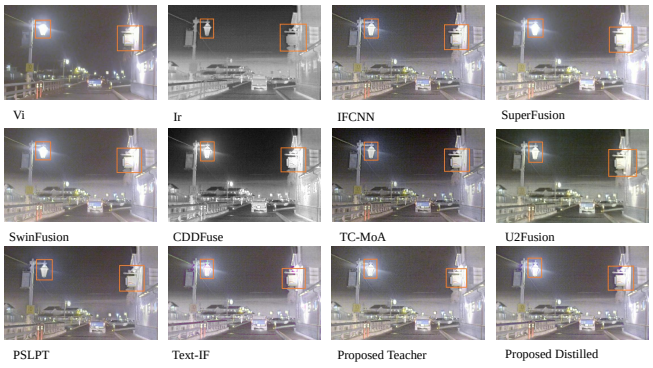


Fig. 3. Qualitative comparison of different image fusion methods on a challenging scene (FLIR_05767.jpg) from the RoadScene dataset. Our method better preserves thermal information from infrared images while maintaining visible details and natural appearance, especially in scenarios with extreme lighting conditions or complex textures. For a more exhaustive visual comparison across various scenarios and methods, please refer to Fig. 9.

student network not only maintains but sometimes exceeds the teacher's performance, particularly in EN and $Q^{AB/F}$ metrics, while reducing parameters by 90%. This suggests that the knowledge distillation process effectively transfers the teacher's knowledge to a much smaller model. The slight performance variations between teacher and student networks can be attributed to the extended training during the distillation process, where the student benefits from both the teacher's guidance and task-specific objectives. Another reason is that

TABLE I
QUANTITATIVE COMPARISON WITH SOTA METHODS ON IVF TASKS ACROSS MSRS, M3FD, AND ROADSCENE DATASETS. **BOLD** AND UNDERLINED VALUES INDICATE THE BEST AND SECOND-BEST RESULTS RESPECTIVELY. TRA.P. = TRAINABLE PARAMETERS, FIX.P. = FIXED PARAMETERS.

Dataset	Method	Tra.P. (M)	Fix.P. (M)	EN↑	MI↑	SF↑	VIF↑	$Q^{AB/F}$ ↑
MSRS	SuperFusion [33]	11.23	-	6.587	3.596	10.783	0.813	0.557
	CDDFuse [32]	2.40	-	6.701	3.657	12.083	0.819	0.548
	IFCNN [34]	0.08	-	5.975	1.706	12.734	0.579	0.479
	U2Fusion [5]	0.63	-	5.246	2.183	9.242	0.512	0.391
	SwinFusion [6]	13.04	-	6.619	3.652	11.038	0.825	0.558
	PSLPT [36]	3.06	-	6.307	2.284	10.419	0.753	0.553
	TC-MOA [35]	7.24	(+329)	6.633	3.251	9.370	0.811	0.565
	Text-IF [9]	63.8	(+151)	6.729	**5.406**	17.384	1.051	0.690
	Ours-teacher	40.3	(+151)	<u>6.749</u>	<u>4.883</u>	<u>17.914</u>	<u>1.060</u>	<u>0.732</u>
	Ours-distilled	4.10	-	**6.763**	4.867	**18.299**	**1.075**	**0.734**
M3FD	SuperFusion [33]	11.23	-	6.726	4.345	11.748	0.664	0.522
	CDDFuse [32]	2.40	-	<u>7.070</u>	3.994	17.578	0.802	0.613
	IFCNN [34]	0.08	-	6.935	2.630	16.250	0.685	0.590
	U2Fusion [5]	0.63	-	6.872	2.683	14.248	0.673	0.578
	SwinFusion [6]	13.04	-	6.844	4.020	14.415	0.746	0.616
	PSLPT [36]	3.06	-	**7.204**	4.563	6.439	**0.958**	0.321
	TC-MOA [35]	7.24	(+329)	6.747	2.856	11.221	0.579	0.508
	Text-IF [9]	63.8	(+151)	6.849	**5.553**	14.484	0.780	0.550
	Ours-teacher	40.3	(+151)	6.965	4.780	<u>17.949</u>	0.896	**0.706**
	Ours-distilled	4.10	-	6.989	<u>4.898</u>	**18.507**	<u>0.927</u>	<u>0.704</u>
RoadScene	SuperFusion [33]	11.23	-	6.990	<u>3.562</u>	12.185	0.608	0.452
	CDDFuse [32]	2.40	-	**7.475**	3.001	19.779	0.610	0.450
	IFCNN [34]	0.08	-	7.222	2.842	15.998	0.591	0.536
	U2Fusion [5]	0.63	-	6.739	2.578	15.282	0.564	0.506
	SwinFusion [6]	13.04	-	7.000	3.334	12.161	0.614	0.450
	PSLPT [36]	3.06	-	7.077	2.001	9.172	0.134	0.171
	TC-MOA [35]	7.24	(+329)	<u>7.387</u>	2.853	12.786	0.577	0.477
	Text-IF [9]	63.8	(+151)	7.332	**5.009**	14.199	0.739	0.578
	Ours-teacher	40.3	(+151)	7.248	3.454	**20.891**	<u>0.743</u>	**0.639**
	Ours-distilled	4.10	-	7.279	3.328	<u>20.082</u>	**0.751**	<u>0.634</u>

current domain models are generally over-parameterized compared to relatively small datasets, which is detailed in the ablation experiments.

C. Comparison on Medical Datasets

To demonstrate the generalization capability of our method, we evaluate its performance on medical image fusion tasks. As shown in Table II, both our teacher and student networks achieve superior performance across different modality pairs. The improvement is particularly significant in structural preservation metrics (SSIM and $Q^{AB/F}$), which is crucial for medical applications. For SPECT-MRI fusion, our method shows notable advantages in preserving functional information while maintaining anatomical details. In CT-MRI fusion, where structural alignment is critical, our approach achieves the highest SSIM scores, indicating better structural preservation. The PET-MRI results further confirm our method's effectiveness in handling multi-modal medical images with different characteristics. Fig. 4 presents visual examples from different medical modalities. Our fusion results exhibit better detail preservation and contrast enhancement, which is essential for clinical applications. The distilled student network maintains these advantages while significantly reducing computational requirements, making it more practical for clinical deployment.

D. Ablation Study

To thoroughly evaluate the effectiveness of our proposed method, we conduct comprehensive ablation studies on both the teacher and student networks using the MSRS test set. These experiments examine the impact of text guidance, various loss components, and model architectures. Furthermore, we introduce new analyses to isolate the contribution of knowledge distillation against architectural improvements and to assess the framework's robustness to different Large Vision-Language Models. Specifically, we address key questions such as: Why does the distilled network sometimes achieve a higher score than the teacher? Is it more effective to train a small model directly instead of using distillation, and how does our distilled model compare against other state-of-the-art lightweight architectures?

1) Analysis of Teacher Network:

a) Effect of Text Guidance.: To validate the effectiveness of text guidance, we compare our full model with its variant without text information. Removing text guidance leads to a significant performance drop across all metrics (EN: -0.023, VIF: -0.083, $Q^{AB/F}$: -0.05). This demonstrates that text descriptions provide valuable semantic information that helps the model better understand and fuse image features. The text guidance acts as an advanced form of reference information, offering rich contextual information beyond pixel-level features.

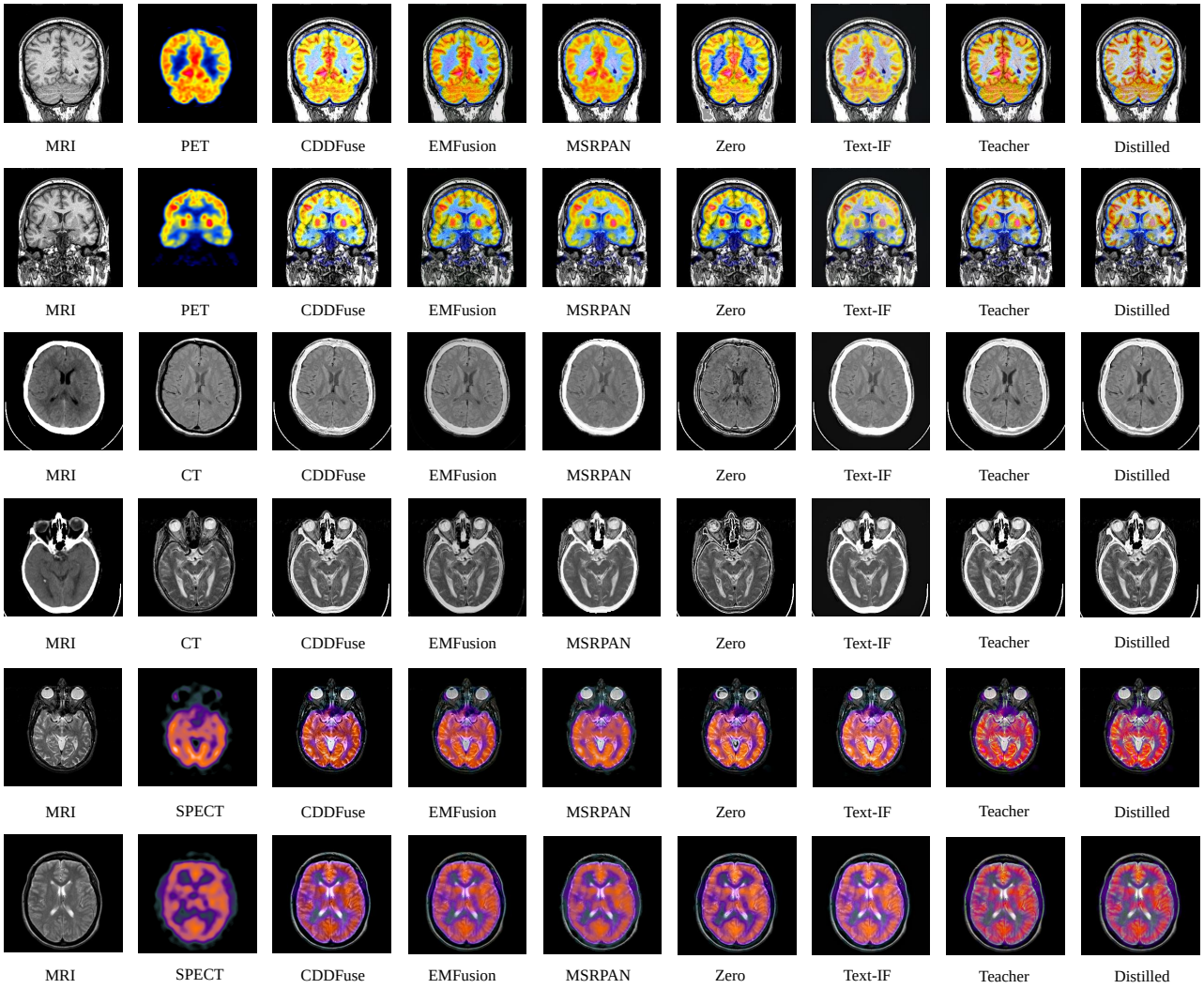


Fig. 4. Comprehensive visual comparison of different image fusion methods on Harvard Medical Image Fusion Datasets (PET-MRI, CT-MRI, SPECT-MRI). For each set of results, from left to right: Modality 1 Input, Modality 2 Input, Text-IF Output, Ours-Teacher Output, Ours-Distilled Output.

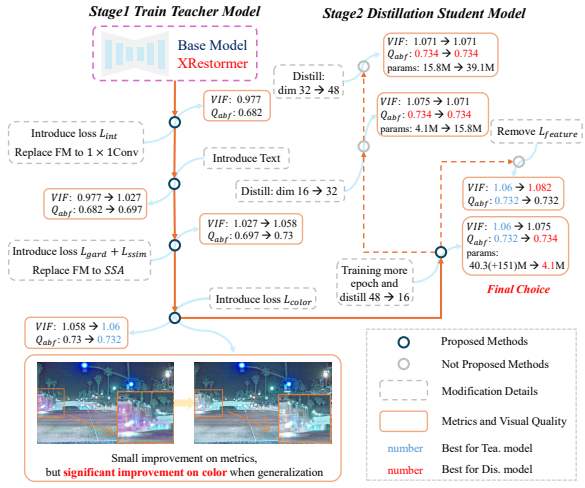


Fig. 5. Overview of our ablation study.

b) Impact of Different Loss Components.: We evaluate the contribution of each loss component by systematically removing them from the full model. The results, as shown in Table III, indicate that all loss components positively impact the final performance, although their importance varies. The intensity loss ($\mathcal{L}_{int}$) and gradient loss ($\mathcal{L}_{grad}$) have the most significant influence, highlighting their essential role in preserving structural information. The SSIM loss ($\mathcal{L}_{ssim}$) aids in maintaining perceptual quality, while the color loss ($\mathcal{L}_{color}$) ensures the natural appearance of fused results.

Due to the nature of our method, which reconstructs the entire image rather than only modifying the Y-axis in YCbCr color space, $\mathcal{L}_{color}$ plays a particularly important role. Although most metrics convert images to grayscale for evaluation, making $\mathcal{L}_{color}$ appear to have minimal effect on metric scores, it is critical for producing visually appealing results in the final fused images.

c) Feature Fusion Methods.: We investigate different fusion embedding strategies to combine features from infrared and visible images. The comparison includes concatenation &

1×1 conv, unlearnable weighted attention, dynamic weighted conv, multiscale conv, and our proposed adaptive fusion method. As shown in Table IV, our adaptive fusion approach achieves superior performance by dynamically adjusting the fusion weights based on spatial and channel features.

2) Analysis of Student Network:

a) Effect of Model Size: A notable finding in our distillation experiments is that reducing the model size by up to 90% (from 40.3M to 4.1M parameters if not calculate CLIP) does not significantly impact performance. As shown in Table V, the compact student model maintains comparable or even slightly better results across different metrics. This suggests that the original model might be over-parameterized for the current fusion task, and the knowledge distillation process effectively transfers essential fusion capabilities to a much smaller architecture.

b) Impact of Distillation Losses: We evaluate the contribution of different components in our distillation framework. The results in Table VI show that result-level supervision (matching the teacher's output) provides strong guidance, significantly improving the student model's performance. The feature distillation loss, on the other hand, plays a key role in preserving feature-level consistency, ensuring that the student model captures important intermediate representations from the teacher. By introducing feature distillation, the model learns high-frequency details and intermediate features, enriching the information. This improves metrics like EN and $Q^{AB/F}$, though it may slightly reduce the VIF score due to imperfect alignment with perceptual fidelity.

c) Text Prior Analysis: To better understand how textual priors influence the fusion process and how they are distilled, we conducted a detailed analysis of their effects at different network levels. As shown in Fig. 6, text guidance significantly impacts the model's attention to different image regions. For instance, when processing a low-contrast nighttime scene, the text prior helps the model better distinguish the vehicle's contours from the background, evident in the feature maps showing more defined object boundaries after prompt guidance. Different text descriptions (e.g., "Low Contrast" vs. "Low Light") lead to distinct feature emphasis patterns. The distilled results show that our student network successfully inherits these text-guided attention patterns without explicit textual input, validating the effectiveness of our prior distillation approach in preserving semantic understanding.

d) Model Design and Knowledge Distillation Analysis: To unequivocally demonstrate that the performance gains stem from our knowledge distillation framework rather than solely from architectural design, we conducted two crucial sets of experiments. First, we trained our student network from scratch using only the base fusion loss ($\mathcal{L}_{base}$), without any guidance from the teacher network. Second, we compare our distilled model with recent lightweight fusion architectures, including LVT [40] and MobileNetV3 [41], ensuring a similar parameter count for a fair comparison. The results, presented in Table VII, clearly show that the distilled student

TABLE II

QUANTITATIVE COMPARISON WITH SOTA METHODS ON MEDICAL IMAGE FUSION TASKS. **BOLD** AND UNDERLINED VALUES INDICATE THE BEST AND SECOND-BEST RESULTS RESPECTIVELY. THE ADDITIONAL PARAMETERS (IN PARENTHESES) ARE NON-TRAINABLE BUT USED DURING INFERENCE.

Dataset	Method	Parameters (M)	SSIM↑	VIF↑	$Q^{AB/F}$ ↑	MI↑	EN↑
PET-MRI	PSLPT [36]	3.06	0.815	0.548	0.373	2.641	5.492
	EMFusion [37]	0.78	1.221	0.685	0.783	3.207	**5.646**
	MSRPAN [39]	0.39	1.182	0.581	**0.799**	4.119	5.073
	SwinFusion [6]	13.04	0.725	0.703	0.683	3.662	5.817
	Zero [38]	19.1	1.162	0.635	0.774	3.786	5.495
	U2Fusion [5]	0.63	0.494	0.460	0.292	2.785	5.532
	CDDFuse [32]	2.40	1.227	0.650	0.765	3.572	5.149
	Text-IF [9]	63.8 (+151)	1.232	0.640	0.690	3.718	5.443
	Ours-teacher	40.3 (+151)	1.223	0.909	0.782	4.248	5.611
	Ours-distilled	4.10	**1.243**	**0.929**	0.784	**4.318**	5.474
CT-MRI	PSLPT [36]	3.06	0.810	0.502	0.432	2.392	4.730
	EMFusion [37]	0.78	1.266	0.552	0.475	3.116	4.785
	MSRPAN [39]	0.39	1.261	0.436	0.455	**4.126**	4.202
	SwinFusion [6]	13.04	0.579	0.522	0.545	3.190	5.144
	Zero [38]	19.1	1.199	0.320	0.582	3.358	4.406
	U2Fusion [5]	0.63	0.042	0.074	0.489	1.694	4.896
	CDDFuse [32]	2.40	1.224	0.526	0.530	3.683	**5.733**
	Text-IF [9]	63.8 (+151)	**1.313**	0.542	0.561	3.211	4.356
	Ours-teacher	40.3 (+151)	**1.313**	0.641	**0.657**	3.246	4.462
	Ours-distilled	4.10	1.312	**0.653**	**0.657**	3.233	4.494
SPECT-MRI	PSLPT [36]	3.06	0.933	0.359	0.325	2.747	5.140
	EMFusion [37]	0.78	**1.212**	0.665	0.692	3.210	4.911
	MSRPAN [39]	0.39	1.153	0.525	0.560	4.334	4.753
	SwinFusion [6]	13.04	0.684	0.744	0.720	3.795	**5.401**
	Zero [38]	19.1	1.180	0.582	0.681	3.564	4.997
	U2Fusion [5]	0.63	0.479	0.419	0.696	2.870	4.539
	CDDFuse [32]	2.40	1.169	0.786	0.719	4.109	4.396
	Text-IF [9]	63.8 (+151)	1.200	0.747	0.715	3.994	4.807
	Ours-teacher	40.3 (+151)	1.210	**0.888**	0.746	4.248	5.035
	Ours-distilled	4.10	1.202	0.887	**0.747**	**4.371**	5.361

TABLE III
ABLATION STUDY ON LOSS COMPONENTS FOR THE TEACHER NETWORK. RESULTS ARE REPORTED ON THE MSRS TEST SET.

Method	EN↑	VIF↑	$Q^{AB/F}$ ↑
$\mathcal{L}_{int}$	6.695	1.027	0.697
$\mathcal{L}_{int} + \mathcal{L}_{color}$	6.696	1.025	0.698
$\mathcal{L}_{int} + \mathcal{L}_{grad} + \mathcal{L}_{ssim}$	6.731	1.058	0.730
$\mathcal{L}_{int} + \mathcal{L}_{color} + \mathcal{L}_{grad}$	6.733	1.059	0.733
$\mathcal{L}_{color} + \mathcal{L}_{grad} + \mathcal{L}_{ssim}$	6.676	1.063	0.728
$\mathcal{L}_{color} + \mathcal{L}_{grad} + \mathcal{L}_{ssim} + \mathcal{L}_{int}$	**6.763**	**1.075**	**0.734**

TABLE IV
COMPARISON OF DIFFERENT FUSION MODULES. RESULTS ON MSRS TEST SET.

Method	EN↑	VIF↑	$Q^{AB/F}$ ↑
Cat & 1×1 conv	6.738	1.059	0.728
Unlearn weight atten	6.667	0.971	0.655
Dynamic weight conv	6.737	1.058	0.729
Multiscale conv	6.743	1.053	0.720
Ours	**6.749**	**1.060**	**0.732**

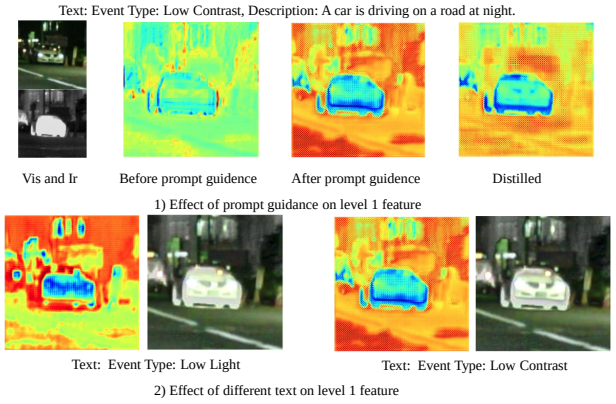


Fig. 6. Effect of text guidance on level 1 and fused image. (a) Original visible and infrared input images. (b) Feature maps before text guidance. (c) Feature maps after text guidance. (d) Final fused images under different text descriptions ("Low Contrast" and "Low Light"). (e) Distilled student network's output.

network ("Ours-distilled") significantly outperforms both the student trained from scratch and other lightweight models. This substantial performance gap (e.g., an improvement of over 0.08 in VIF on MSRS) validates that knowledge transfer is the key contributor to our model's effectiveness. While our XRestormer-based [42] backbone shows a slight advantage over generic lightweight models, this architectural benefit is minor compared to the substantial leap provided by the distillation process.

e) Knowledge Distillation Analysis: An intriguing observation from our experiments is that the distilled student network occasionally outperforms its teacher model. This phenomenon, illustrated in Fig. 7, suggests that the distillation process can act as a form of regularization. When training the teacher network with an extremely high weight for $\mathcal{L}_{ssim}$ (10× the baseline), the teacher becomes highly specialized in edge

TABLE V
IMPACT OF MODEL SIZE ON DISTILLED NETWORK. RESULTS ON MSRS TEST SET.

Model	Params(M)	EN↑	VIF↑	$Q^{AB/F}$ ↑
48-dim	39.1	**6.780**	1.071	**0.734**
32-dim	15.8	6.768	1.071	**0.734**
16-dim	**4.1**	6.763	**1.075**	**0.734**

TABLE VI
ABLATION STUDY ON DISTILLATION LOSSES (MSRS DATASET).

Method	EN↑	VIF↑	$Q^{AB/F}$ ↑
$\mathcal{L}_{base}$	6.728	0.989	0.681
$\mathcal{L}_{base} + \mathcal{L}_{res}$	6.761	**1.082**	0.732
$\mathcal{L}_{base} + \mathcal{L}_{feat} + \mathcal{L}_{res}$	**6.763**	1.075	**0.734**

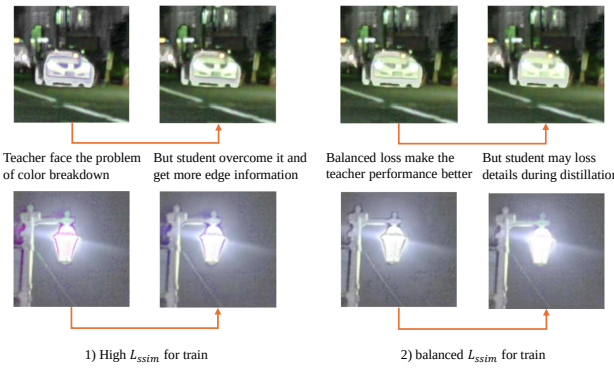


Fig. 7. When teacher performs extremely, the student learns more information. This figure illustrates how a student network learns to balance structural and color information even when the teacher over-prioritizes structural details, leading to color breakdown in the teacher's output.

detection, sacrificing color fidelity. Surprisingly, the student, when distilled from such a teacher, learns to balance these competing objectives, maintaining edge clarity while preserving natural color representation. This indicates that extreme specialization in the teacher can provide a clearer supervision signal for certain features, enabling the student to learn more comprehensive fusion strategies. In our final implementation, a moderately increased weight for $\mathcal{L}_{ssim}$ proves sufficient for the student to achieve superior performance across multiple metrics with stable training.

Our ablation studies reveal several key insights:

- Text guidance significantly enhances fusion performance by embedding rich semantic context into the process.
- The integration of multiple loss terms is crucial for achieving balanced and robust fusion results.
- Model compression via knowledge distillation achieves comparable performance with drastically fewer parameters, as the teacher model and some text-guided methods exhibit excessive over-parameterization compared to small datasets.

f) Robustness to LLM Choice: To address the concern that our framework's performance might be overly sensitive to the choice of the teacher LLM, we conducted an analysis

TABLE VII
ABLATION STUDY ON THE CONTRIBUTION OF KNOWLEDGE DISTILLATION. ALL LIGHTWEIGHT MODELS HAVE PARAMETERS IN THE < 10M RANGE. RESULTS ARE REPORTED ACROSS THREE IVF DATASETS.

Method	Params(M)	MSRS Dataset			M3FD Dataset			RoadScene Dataset		
		MI↑	VIF↑	$Q^{AB/F}$ ↑	MI↑	VIF↑	$Q^{AB/F}$ ↑	MI↑	VIF↑	$Q^{AB/F}$ ↑
Switch to LVT [40] Backbone	7.9	4.093	0.971	0.691	3.983	0.762	0.611	3.122	0.611	0.554
Switch to MobileNetV3 [41] Backbone	1.3	3.998	0.922	0.662	3.822	0.747	0.581	3.089	0.589	0.514
Ours-student (from scratch)	4.1	4.110	0.989	0.682	4.154	0.782	0.615	3.144	0.655	0.536
Ours-student (from distillation)	4.1	**4.867**	**1.075**	**0.734**	**4.898**	**0.927**	**0.704**	**3.328**	**0.751**	**0.634**

with different state-of-the-art VLMs, including the newest Qwen2.5-VL [21] and GPT-4V [19]. As shown in Fig. 8, all tested VLMs provide highly consistent and relevant textual priors for the fusion task, identifying key degradation events such as "Low Light" and "Overexposure". The textual responses from our chosen Qwen2-VL-7B [20] model are closely aligned with those from GPT-4V. Furthermore, the final fusion metrics on the MSRS dataset remain remarkably stable across different teacher models, with VIF and $Q^{AB/F}$ scores showing minimal variation. This demonstrates that our framework is robust to the choice of the VLM. The semantic knowledge required for this task is foundational and can be provided by any sufficiently powerful VLM, making our approach generalizable and not dependent on a single proprietary model.

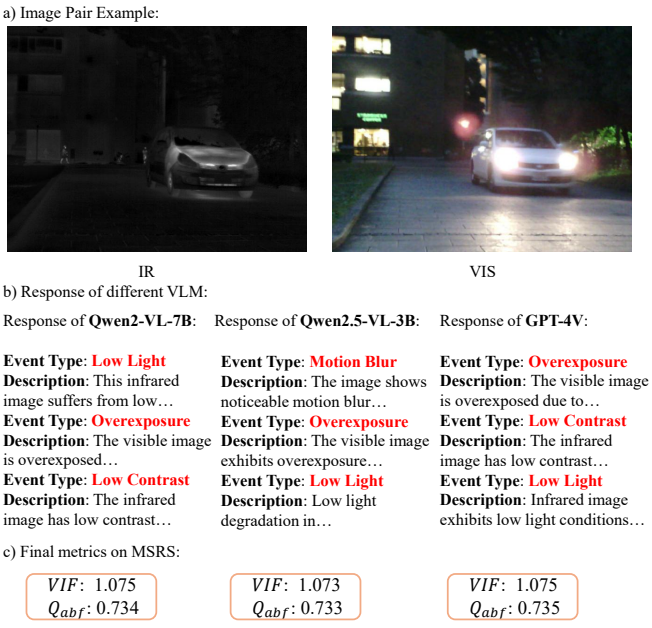


Fig. 8. Comparison of textual priors generated by different VLMs and their impact on final fusion performance. (a) An example IR/VIS image pair. (b) The generated text from three different VLMs all identify similar key issues. (c) The final metrics on the MSRS dataset for teacher networks trained with these priors are highly stable, demonstrating the robustness of our framework.

E. Inference Time Analysis

To evaluate the computational efficiency, we analyze the inference times of different components across the previous state-of-the-art (SoTA) text-guided method Text-IF, our teacher network, and the distilled student network. For text generation, we use Qwen2-VL [20] as the large language model (LLM), which operates on two NVIDIA RTX 4090 GPUs with the fast attention library [43]. Other timing experiments were conducted on a separate single NVIDIA RTX 4090 GPU, running with a batch size of 1 and an image resolution of 128×128 pixels. Since these results may vary across different machines or even different states of the same machine, this comparison should be considered non-rigorous.

As shown in Table VIII, the LLM component dominates the computational cost, requiring over 2 seconds per image. CLIP encoding adds another 110-115ms overhead. While the teacher network's fusion module takes 133.4ms, our distilled student network reduces this to just 46.1ms—a 65% reduction in fusion time. More importantly, by eliminating the need for LLM and CLIP during inference, our student network achieves a dramatic 98% reduction in total inference time (from 2556.91ms to 46.11ms).

TABLE VIII
DETAILED INFERENCE TIME COMPARISON (IN MILLISECONDS) FOR DIFFERENT COMPONENTS. LLM REFERS TO THE QWEN2-VL MODEL USED FOR TEXT GENERATION.

Method	Data Load	LLM	CLIP	Fusion	Total
Text-IF	0.01	2261.2	115.7	152.1	2529.01
Teacher	0.01	2311.1	112.4	133.4	2556.91
Student	0.01	-	-	**46.1**	46.11

F. Limitations

While our distilled model achieves significant inference-time efficiency, we acknowledge a limitation in training overhead. The knowledge distillation process introduces additional computational costs during the training phase.

TABLE IX
TRAINING TIME COMPARISON FOR DIFFERENT EPOCHS. ALL EXPERIMENTS WERE CONDUCTED ON A SINGLE NVIDIA 4090 GPU.

Method	100 epochs	500 epochs
Text-IF	1.9h	11.1h
Teacher	2.0h	12.1h
Student (distill)	2.1h	13.5h

As shown in Table IX, our distillation approach requires slightly longer training time compared to both the baseline

Fig. 9. Comprehensive visual comparison with different methods on various IVF datasets (MSRS, M3FD, RoadScene). For each set of results, from left to right: Visible Input, Infrared Input, Text-IF Output, Ours-Teacher Output, Ours-Distilled Output.

Text-IF and the teacher model. For a 100-epoch training cycle, the student network's distillation process takes 2.1 hours, approximately 10.5% longer than Text-IF (1.9h) and 5% longer than the teacher model (2.0h). This pattern persists for longer training durations, with 500-epoch training requiring 13.5 hours for distillation compared to 11.1 hours for Text-IF and 12.1 hours for the teacher model.

This increased training time is primarily attributed to two factors: (1) the two-stage training process where the teacher must be trained first, and (2) the additional computational overhead from the distillation loss calculations. However, we consider this a reasonable trade-off given the substantial inference-time benefits, as the training process is typically a one-time cost while inference efficiency directly impacts real-world applications.

V. Conclusion

We address the weak semantic understanding of traditional image fusion methods and the high cost of text-guided approaches by distilling textual priors and eliminating the need for text guidance during inference. Our approach, which leverages a teacher-student architecture and tailored prior distillation, significantly reduces the model size while retaining high performance. Extensive experiments and ablation studies validate the effectiveness of our method, highlighting its ability to achieve a strong trade-off between computational efficiency and fusion quality. Specifically, our findings confirm that the performance of the distilled student network is primarily attributed to the knowledge transfer process, as it substantially outperforms an identical network trained from scratch. This demonstrates that our framework provides a robust method for embedding large-model intelligence into compact architectures, offering a practical path toward deploying high-performance fusion models in resource-constrained environments.

References

[1] Y. Zhao, Q. Zheng, P. Zhu, X. Zhang, and W. Ma, "Tufusion: A transformer-based universal fusion algorithm for multimodal images," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 3, pp. 1712–1725, 2024.

[2] X. Li, G. Zhang, W. Chen, L. Cheng, Y. Xie, and J. Ma, "An infrared and visible image fusion method based on semantic-sensitive mask selection and bidirectional-collaboration region fusion," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2024.

[3] J. Zhang, K. Cao, K. Yan, Y. Lin, X. He, Y. Wang, R. Li, C. Xie, J. Zhang, and M. Zhou, "Frequency decoupled domain-irrelevant feature learning for pan-sharpening," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 35, no. 2, pp. 1237–1250, 2025.

[4] 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

[5] 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, 2020.

[6] 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.

[7] L. Tang, J. Yuan, H. Zhang, X. Jiang, and J. Ma, "Piafusion: A progressive infrared and visible image fusion network based on illumination aware," *Information Fusion*, vol. 83-84, pp. 79–92, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S156625352200032X

[8] C. Cheng, T. Xu, X.-J. Wu, H. Li, X. Li, Z. Tang, and J. Kittler, "Textfusion: Unveiling the power of textual semantics for controllable image fusion," *arXiv preprint arXiv:2312.14209*, 2023.

[9] X. Yi, H. Xu, H. Zhang, L. Tang, and J. Ma, "Text-if: Leveraging semantic text guidance for degradation-aware and interactive image fusion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[10] J. Bai, S. Bai, S. Yang, S. Wang, S. Tan, P. Wang, J. Lin, C. Zhou, and J. Zhou, "Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond," *arXiv preprint arXiv:2308.12966*, 2023.

[11] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," *CoRR*, vol. abs/2103.00020, 2021. [Online]. Available: https://arxiv.org/abs/2103.00020

[12] Z. Zhao, H. Bai, Y. Zhu, J. Zhang, S. Xu, Y. Zhang, K. Zhang, D. Meng, R. Timofte, and L. Van Gool, "Ddfm: Denoising diffusion model

for multi-modality image fusion," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023, pp. 8082–8093.

[13] M. Zhou, N. Zheng, X. He, D. Hong, and J. Chanussot, "Probing synergistic high-order interaction for multi-modal image fusion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[14] B. Cao, Y. Sun, P. Zhu, and Q. Hu, "Multi-modal gated mixture of local-to-global experts for dynamic image fusion," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 23 555–23 564.

[15] G. Hinton, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[16] L. Yuan, F. E. Tay, G. Li, T. Wang, and J. Feng, "Revisiting knowledge distillation via label smoothing regularization," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 3903–3911.

[17] K. Wu, H. Peng, Z. Zhou, B. Xiao, M. Liu, L. Yuan, H. Xuan, M. Valenzuela, X. S. Chen, X. Wang *et al.*, "Tinyclip: Clip distillation via affinity mimicking and weight inheritance," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 21 970–21 980.

[18] Y. Zhao, Y. Xu, Z. Xiao, and T. Hou, "Mobilediffusion: Subsecond text-to-image generation on mobile devices," *arXiv preprint arXiv:2311.16567*, 2023.

[19] OpenAI, "Gpt-4 technical report," 2024. [Online]. Available: https://arxiv.org/abs/2303.08774

[20] P. Wang, S. Bai, S. Tan, S. Wang, Z. Fan, J. Bai, K. Chen, X. Liu, J. Wang, W. Ge, Y. Fan, K. Dang, M. Du, X. Ren, R. Men, D. Liu, C. Zhou, J. Zhou, and J. Lin, "Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution," *arXiv preprint arXiv:2409.12191*, 2024.

[21] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang, H. Zhong, Y. Zhu, M. Yang, Z. Li, J. Wan, P. Wang, W. Ding, Z. Fu, Y. Xu, J. Ye, X. Zhang, T. Xie, Z. Cheng, H. Zhang, Z. Yang, H. Xu, and J. Lin, "Qwen2.5-vl technical report," 2025. [Online]. Available: https://arxiv.org/abs/2502.13923

[22] X. Chen, Z. Li, Y. Pu, Y. Liu, J. Zhou, Y. Qiao, and C. Dong, "A comparative study of image restoration networks for general backbone network design," in *European Conference on Computer Vision*. Springer, 2025, pp. 74–91.

[23] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and M.-H. Yang, "Restormer: Efficient transformer for high-resolution image restoration," in *CVPR*, 2022.

[24] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[25] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[26] L. Tang, J. Yuan, H. Zhang, X. Jiang, and J. Ma, "MSRS: Multi-Spectral Road Scenarios for Practical Infrared and Visible Image Fusion," https://github.com/Linfeng-Tang/MSRS, 2022.

[27] J. Liu, X. Fan, Z. Huang, G. Wu, R. Liu, W. Zhong, and Z. Luo, "Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5802–5811.

[28] H. Xu, J. Ma, Z. Le, J. Jiang, and X. Guo, "Fusiondn: A unified densely connected network for image fusion," in *proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*, 2020.

[29] J. W. Roberts, J. A. van Aardt, and F. B. Ahmed, "Assessment of image fusion procedures using entropy, image quality, and multispectral classification," *Journal of Applied Remote Sensing*, vol. 2, no. 1, p. 023522, 2008. [Online]. Available: https://doi.org/10.1117/1.2945910

[30] Y. Han, Y. Cai, Y. Cao, and X. Xu, "A new image fusion performance metric based on visual information fidelity," *Information Fusion*, vol. 14, no. 2, pp. 127–135, 2013. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S156625351100056X

[31] G. Piella and H. Heijmans, "A new quality metric for image fusion," in *Proceedings 2003 International Conference on Image Processing (Cat. No.03CH37429)*, vol. 3, 2003, pp. III–173.

[32] 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*, 2023, pp. 5906–5916.

[33] L. Tang, Y. Deng, Y. Ma, J. Huang, and J. Ma, "Superfusion: A versatile image registration and fusion network with semantic awareness," *IEEE/CAA Journal of Automatica Sinica*, vol. 9, no. 12, pp. 2121–2137, 2022.

[34] 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.

[35] P. Zhu, Y. Sun, B. Cao, and Q. Hu, "Task-customized mixture of adapters for general image fusion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[36] W. Wang, L.-J. Deng, and G. Vivone, "A general image fusion framework using multi-task semi-supervised learning," *Information Fusion*, p. 102414, 2024.

[37] H. Xu and J. Ma, "Emfusion: An unsupervised enhanced medical image fusion network," *Information Fusion*, 2021.

[38] F. Lahoud and S. Süsstrunk, "Zero-learning fast medical image fusion," in *2019 22th International Conference on Information Fusion (FUSION)*, 2019, pp. 1–8.

[39] J. Fu, W. Li, J. Du, and Y. Huang, "A multiscale residual pyramid attention network for medical image fusion," *Biomedical Signal Processing and Control*, vol. 66, p. 102488, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1746809421000859

[40] C. Yang, Y. Wang, J. Zhang, H. Zhang, Z. Wei, Z. Lin, and A. Yuille, "Lite vision transformer with enhanced self-attention," 2021. [Online]. Available: https://arxiv.org/abs/2112.10809

[41] S. N. Wadekar and A. Chaurasia, "Mobilevitv3: Mobile-friendly vision transformer with simple and effective fusion of local, global and input features," 2022. [Online]. Available: https://arxiv.org/abs/2209.15159

[42] X. Chen, Z. Li, Y. Pu, Y. Liu, J. Zhou, Y. Qiao, and C. Dong, "A comparative study of image restoration networks for general backbone network design," 2024. [Online]. Available: https://arxiv.org/abs/2310.11881

[43] T. Dao, "FlashAttention-2: Faster attention with better parallelism and work partitioning," in *International Conference on Learning Representations (ICLR)*, 2024.

[44] O. Patashnik, Z. Wu, E. Shechtman, D. Cohen-Or, and D. Lischinski, "Styleclip: Text-driven manipulation of stylegan imagery," *CoRR*, vol. abs/2103.17249, 2021. [Online]. Available: https://arxiv.org/abs/2103.17249

[45] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," 2021.

[46] V. Petrovic and C. Xydeas, "Objective image fusion performance characterisation," in *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, vol. 2, 2005, pp. 1866–1871 Vol. 2.

[47] X. X. Zhu and R. Bamler, "A sparse image fusion algorithm with application to pan-sharpening," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 51, no. 5, pp. 2827–2836, 2013.

[48] D. Guo, K. Li, B. Hu, Y. Zhang, and M. Wang, "Benchmarking micro-action recognition: Dataset, methods, and applications," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 7, pp. 6238–6252, 2024.

[49] H. Albarqaan, R. Qin, and Y. Teng, "Image Fusion in Remote Sensing: An Overview and Meta-Analysis," *Photogrammetric Engineering & Remote Sensing*, vol. 90, no. 12, pp. 755–775, December 2024. [Online]. Available: https://doi.org/10.14358/PERS.24-00110R1

[50] 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. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1566253517307972

[51] T. Dao, D. Y. Fu, S. Ermon, A. Rudra, and C. Ré, "FlashAttention: Fast and memory-efficient exact attention with IO-awareness," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[52] W. Zhao, D. Wang, and H. Lu, "Multi-focus image fusion with a natural enhancement via a joint multi-level deeply supervised convolutional neural network," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 4, pp. 1102–1115, 2019.

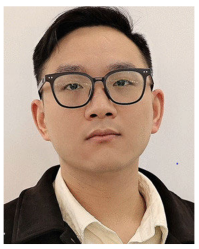
Ran Zhang received the B.S. degree (expected) in computer science and technology from the School of Computer Science and Information Engineering, Hefei University of Technology in 2027. He is currently a junior at Hefei University of Technology. His current research interests include computer vision and image processing.

Liu Liu received the B.S. degree from the Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 2015, and the Ph.D. degree from the University of Science and Technology of China, Hefei, China, in 2020. He was a Postdoctoral Researcher with Shanghai Jiao Tong University, China, from 2020 to 2022. He is currently a Associate Professor with Hefei University of Technology. His research interests include computer vision, deep learning, and embodied AI.

Xuanhua He received the B.E. degree in software engineering from Xiamen University, Xiamen, China, in 2022. He is currently pursuing the M.S. degree with the University of Science and Technology of China, Hefei, China. His research interests include image restoration and video generation.

Ke Cao received the B.E. degree in automation from East China Jiaotong University, Nanchang, China, in 2023. He is currently pursuing the M.S. degree with the University of Science and Technology of China, Hefei, China. His research interests include image restoration and pansharpening.

Li Zhang is currently pursuing the Ph.D. degree in computer science and application with the University of Science and Technology of China, Hefei. His research interests include computer vision, machine learning, and embodied AI. In particular, he is interested in the rigid and non-rigid object pose perception. He has published several top-tier conference papers at ECCV, NeurIPS, ICML etc.



Man Zhou received the Ph.D. degree from the University of Science and Technology of China, Hefei, China, in 2022. Particularly, he focuses on geography information systems, machine/deep-learning-based satellite image processing, and multi-source information fusion. He was a recipient of the Baidu Scholarship (top ten globally) in 2022 and the WAIC Yunfan Award in 2023 (top 15 globally).

Jie Zhang received the M.S. degree from Hefei University of Technology, Hefei, China, in 2009, and the Ph.D. degree from the University of Science and Technology of China, Hefei, in 2014. He is currently an Associate Professor with the Institute of Intelligent Machines, Hefei Institutes of Physical Science, Chinese Academy of Sciences, Hefei. His current research interests include image processing, pattern recognition, and artificial intelligence.

Dan Guo received the B.E. degree in computer science and technology from Yangtze University, China, in 2004, and the Ph.D. degree in system analysis and integration from Huazhong University of Science and Technology, China, in 2010. She is currently an Associate Professor with the School of Computer Science and Information Engineering, Hefei University of Technology, China. Her research interests include computer vision, machine learning, and intelligent multimedia content analysis.

Meng Wang (Fellow, IEEE) received the BE and PhD degrees from USTC, in 2003 and 2008, respectively. He is a professor with HFUT. His current research interests include multimedia content analysis, computer vision, and pattern recognition. He has authored more than 200 book chapters, journal, and conference papers in these areas. He is the recipient of the ACM SIGMM Rising Star Award 2014. He is an associate editor of the IEEE Transactions on Knowledge and Data Engineering, the IEEE Transactions on Circuits and Systems for Video Technology, and the IEEE Transactions on Neural Networks and Learning Systems. He is an IEEE Fellow and IAPR Fellow.