Article | Received 1 September 2026; Revised 9 November 2026; Accepted 16 November 2026; Published 26 November 2026
https://doi.org/10.55092/aias20260008

IT-TextFusion: iterative text-image interaction with text-guided residual refinement for degradation-aware image fusion

Siyang Liu, Peiyi Zhou*, Tianle Jin, Rongrong Bian and Zheke Jin

Chair of Robotics, Artificial Intelligence and Real-time Systems, Technical University of Munich, Munich, Germany

* Correspondence author; E-mail: go25gik@mytum.de.

Highlights:

- Hierarchical text conditioning guides fusion across multiple decoder stages.
- Degradation-aware prompts provide scenario-specific global semantic guidance.
- Cross-Gate Fusion adaptively aggregates visible and infrared features at multiple hierarchical scales.
- Text-guided residual refinement improves high-resolution feature reconstruction under degradations.

Abstract: Text-guided image fusion has recently emerged as an effective paradigm for integrating multi-modal information while enabling flexible and task-oriented fusion control. However, existing text-guided fusion methods often rely on shallow semantic-visual interaction and limited attention mechanisms, which restrict their ability to robustly handle complex degradations and fully exploit textual guidance. In this paper, we propose an iterative text-guided image fusion framework that incorporates text-conditioned feature interaction across multiple fusion and refinement stages. The proposed method integrates deepest-level Cross-Attention, multi-scale Cross-Gate Fusion, and stage-specific text-conditioned modulation, allowing the global text embedding to condition hierarchical feature fusion and residual refinement. By repeatedly injecting the pooled text embedding across hierarchical decoder and refinement stages, the proposed framework provides degradation-aware global semantic conditioning while preserving complementary information from the visible and infrared modalities. Experiments on several benchmark datasets show that the proposed method improves several information-preservation and perceptual-quality metrics, while exhibiting metric-dependent trade-offs on some datasets.

Keywords: infrared-visible image fusion; text-guided image fusion; iterative text-image interaction; degradation-aware fusion; text-guided residual refinement

1. Introduction

In recent years, autonomous driving has attracted increasing attention due to rapid progress in sensing, learning, and system integration technologies. Autonomous driving systems aim to reduce driver workload,

Copyright©2026 by the authors. Published by ELSP. This work is licensed under Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium provided the original work is properly cited.

improve road safety, and mitigate traffic congestion. Among the three core components of autonomous driving systems—perception, prediction, and motion planning—perception plays a fundamental role by providing accurate environmental understanding to support reliable decision-making in advanced driver-assistance systems (ADAS) . Core perception tasks such as semantic segmentation [1] and object detection [2] rely heavily on visual sensing, with cameras remaining the most widely deployed sensors in autonomous vehicles due to their rich spatial and appearance information [3,4]. Cao *et al*. [5] further investigate fisheye object detection for autonomous driving, introducing a feature-aligned pyramid module and a location-aligned detection head.

Frame-based red-green-blue (RGB) images generally provide high visual fidelity, and recent advances in deep learning and end-to-end training have improved perception performance under favorable conditions. However, perception systems relying on a single sensing modality may become less reliable in challenging environments. Conventional cameras are particularly sensitive to adverse conditions such as low illumination, overexposure, motion blur, and sensor noise. In addition, complex traffic scenes frequently contain occlusions and large illumination variations, which can further affect the reliability of camera-based perception. These factors can degrade the performance of downstream perception tasks, including object detection and semantic segmentation, making robust perception more difficult to achieve. Multi-modal sensor fusion provides a way to address these problems by combining complementary information from different sensing modalities [2,3,6,7].

While conventional RGB cameras remain effective standalone sensors, combining different sensing modalities can provide a more complete and reliable view of the environment. Depth sensors provide geometric information, event cameras perform well in high-speed and high-dynamic-range scenarios [2,4,8], and light detection and ranging (LiDAR) systems provide accurate spatial and angular measurements [4]. Infrared cameras capture thermal radiation, which makes them useful for nighttime perception and detecting heat-emitting objects such as pedestrians and vehicles [9]. Although some studies investigate the simultaneous fusion of multiple modalities [10], this work focuses on visible and infrared image fusion, as the two modalities provide complementary information for safety-critical perception tasks.

One challenge in multi-modal perception is to effectively integrate information from heterogeneous sensors [11]. Existing fusion strategies are typically divided into early, middle, and late fusion. Early fusion combines raw sensor data at the input level but can be affected by noise. Late fusion integrates modality-specific outputs at the decision level, often through algebraic or probabilistic methods, but has limited ability to capture fine-grained cross-modal interactions at lower feature levels [11]. Middle fusion combines modality-specific features within intermediate network layers and is widely used in multi-modal perception. For example, NeuroGrasp [12] combines RGB and event information in a multi-modal neural network for robotic grasp pose estimation. However, extracting useful information from different modalities while reducing the influence of noise and inconsistencies remains challenging. In this work, we adopt a middle-fusion strategy and use semantic guidance from textual prompts to improve feature fusion under complex degradation conditions.

Recent progress in vision-language models has shown the potential of using textual semantics to guide visual representation learning. Large-scale vision-language models trained on diverse image-text pairs can align visual and textual information and provide semantic representations for downstream visual tasks [9]. In particular, Contrastive Language-Image Pre-training (CLIP) [13] provides transferable semantic embeddings that have been used in various vision applications, including image fusion [14]. Recent text-guided fusion methods such as Text-IF [9], TextFusion [15], and TITFormer [16] incorporate linguistic descriptions into the fusion pipeline to guide feature aggregation and refinement. These methods show that textual information can complement visual features during fusion, providing additional guidance beyond purely visual fusion strategies.

Nevertheless, many earlier fusion methods rely mainly on visual information. As illustrated in Figure 1, representative methods such as [17–19] use image-based fusion mechanisms without incorporating additional semantic information. Although these methods perform well under relatively clean imaging conditions, their performance can decline under complex real-world degradations, such as low illumination, reduced contrast, and sensor noise.

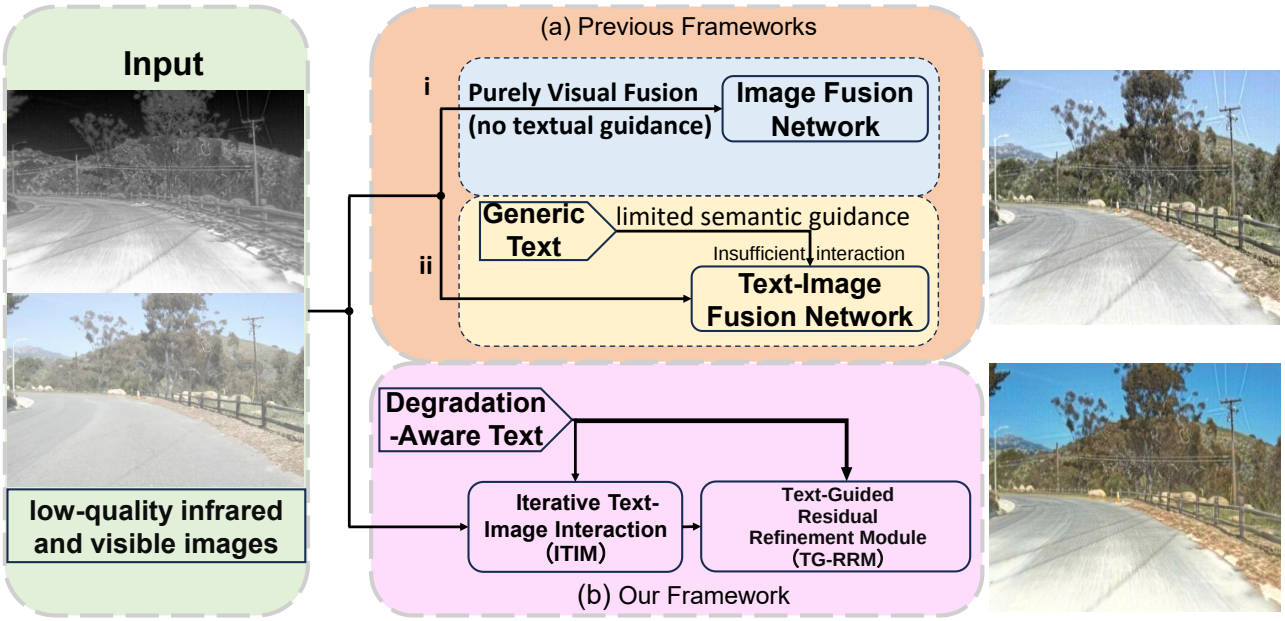

Figure 1. Comparison of representative image fusion paradigms. **(a-i)** Purely visual fusion without textual guidance; **(a-ii)** Representative text-guided fusion using generic text with limited semantic specificity and relatively shallow semantic-visual interaction across hierarchical feature levels; **(b)** The proposed IT-TextFusion framework, in which semantically enriched degradation-aware text provides global conditioning across four hierarchical decoding stages through the Iterative Text-Image Interaction Module (ITIM) and continues to condition the subsequent Text-Guided Residual Refinement Module (TG-RRM)-based residual refinement.

Despite recent progress, existing text-guided fusion methods often exhibit three closely related limitations. First, text is commonly represented as a generic semantic condition with limited degradation specificity. Because this condition is introduced through relatively shallow semantic–visual interaction, its effect may weaken as features are decoded at different resolutions. If the subsequent refinement stages

do not maintain the same textual condition, degradation-related information in the coarse fused features may not be adequately corrected before reconstruction. These issues are closely connected: limited text specificity reduces the semantic information available, shallow interaction makes it difficult to propagate this information across feature scales, and refinement without sufficient textual guidance may further reduce its influence before reconstruction.

To overcome these limitations, we propose a unified text-guided fusion framework that strengthens hierarchical text-conditioned feature fusion and residual refinement. The proposed approach improves the degradation- and scene-level specificity of the global textual condition through degradation-aware text engineering. In addition, the Cross-Gate Fusion module is introduced to regulate inter-modal feature aggregation through learnable cross-modal weighting, with the aim of reducing the influence of unreliable visual responses under degradation. To maintain textual conditioning across hierarchical scales, we design the Iterative Text-Image Interaction Module (ITIM), which repeatedly injects the pooled global text embedding into decoder features through independently parameterized stage-specific modulation functions. Furthermore, the Text-Guided Residual Refinement Module (TG-RRM) is developed to apply global text-conditioned modulation within the multi-scale refinement process, supporting high-resolution residual feature correction before final reconstruction. Overall, our contributions can be summarized as follows:

- We address two aspects of the fusion pipeline: degradation-aware text engineering provides more specific degradation- and scene-level global conditions, while the Cross-Gate Fusion module introduces learnable cross-modal weighting to support multi-scale feature aggregation and is intended to reduce the influence of unreliable visual responses.
- We propose ITIM, which incorporates textual guidance into visual features across multiple fusion stages, providing hierarchical global text conditioning throughout the fusion process.
- We introduce TG-RRM, which uses text-conditioned modulation during refinement to improve structural consistency and preserve fine details.
- In addition, we evaluate the proposed method on the Enhanced Multi-Spectral Various Scenarios (EMS) dataset. The dataset covers nine representative degradation types. The experimental results indicate improvements in some fusion and no-reference perceptual-quality metrics, but also reveal trade-offs in mutual information (MI), the summed Structural Similarity Index Measure ($SSIM_{sum}$), and CLIP-based Image Quality Assessment (CLIP-IQA).

2. Related work

2.1. General image fusion methods

Following the development of deep convolutional networks, encoder-decoder architectures and residual learning have been widely used in image fusion tasks [20,21]. With skip connections and parallel modality-specific backbones, deep fusion models can preserve complementary information from different sources [8,11,22]. More recently, transformer-based architectures have been used in image fusion to capture long-range dependencies through self-attention [23]. Vision Transformer (ViT) [24] represents images as sequences of patch tokens, allowing global context to be modeled in fusion tasks [9,10,16,25].

The Cross-Modal Fusion framework, CMX [4], uses a cross-modal feature rectification module to refine RGB features and features from other modalities before fusion, followed by cross-attention-based feature aggregation. Based on CMX, the arbitrary-modal semantic segmentation model CMNeXt [10] extends this framework to support an arbitrary number of modalities through dynamic feature selection and parallel pooling. Recent infrared–visible studies have also considered reliability-aware cross-modal weighting and alignment. IGIANet [26] uses illumination-guided frequency modulation and implicit alignment-driven dynamic fusion for infrared–visible UAV detection, while the Variance-Guided Attention Modulation Framework (VAMF) [27] uses variance-guided attention modulation to balance global contextual information and local detail extraction in infrared–visible image fusion. Although these two methods focus on detection and fusion, respectively, both suggest the need to consider illumination variations and modality reliability in cross-modal interaction. Other fusion strategies have also been studied. Adaptive Instance Normalization (AdaIN) [28] aligns feature statistics efficiently and has inspired lightweight modulation strategies in multi-modal representation learning. In contrast, Bayesian fusion approaches [11] perform late fusion at the detection level by combining independently trained detectors using probabilistic inference, which can improve robustness under modality misalignment. The E2E-MFD model [6] uses an Object-Region-Pixel Phylogenetic Tree (ORPPT) for hierarchical, coarse-to-fine fusion with a feedback mechanism. Finally, Text-IF [9] introduces semantic text guidance into image fusion, supporting interactive and degradation-aware fusion through transformer-based cross-attention and convolutional decoding.

2.2. Text-image models

Transformer-based representation learning has improved multi-modal understanding by allowing visual and textual data to be modeled together. One important development is CLIP [13], which aligns images and text in a shared embedding space through large-scale contrastive pretraining, supporting zero-shot recognition and semantic retrieval. Based on CLIP, DenseCLIP [29] extends vision-language pretraining to dense prediction tasks by introducing pixel-text similarity maps and context-aware prompting for semantic segmentation and detection. Diffusion-based models such as Stable Diffusion [30] also use textual conditions for fine-grained visual control by combining text encoders with generative architectures.

Textual semantics have also been introduced into image fusion. TextFusion [15] uses affine text-guided modulation within a transformer-based fusion framework to control fusion outcomes. TeRF [22] explores region-aware fusion by incorporating language commands into segmentation-guided pipelines. TITFormer [16] combines textual inputs with simulated infrared imagery for image enhancement, using cross-modal attention to extract contextual features from text. MGFusion [14] incorporates semantic information through CLIP-guided feature modulation to improve multi-modal representations without complex architectural designs. Text-IF [9] uses semantic text inputs to adjust fusion strategies according to different conditions.

Existing text-guided fusion methods differ mainly in where and how textual information is introduced during fusion. TextFusion uses text-conditioned feature modulation, TeRF focuses on region-aware language control, MGFusion uses CLIP-derived semantic features to modulate multi-modal representations, and Text-IF adjusts fusion according to degradation descriptions. In

comparison, IT-TextFusion maintains global text conditioning throughout hierarchical decoding and residual feature refinement. Our method introduces textual conditioning at different stages to combine multi-scale visual fusion with high-resolution residual refinement.

3. Method

This section describes the overall workflow of our framework, as shown in Figure 2. We first introduce Text Engineering, which generates degradation-aware prompts to guide image fusion. The Image Fusion Pipeline is then described, including the image encoders, Cross-Attention, Cross-Gate Fusion, and the ITIM. The Text Semantic Encoding and Feature Modulation subsection explains how the input text is encoded into a global embedding and used to condition visual features at each hierarchical stage. We also introduce the TG-RRM, which performs text-conditioned residual feature correction before image reconstruction. The Loss Functions subsection then presents the components of the overall training objective.

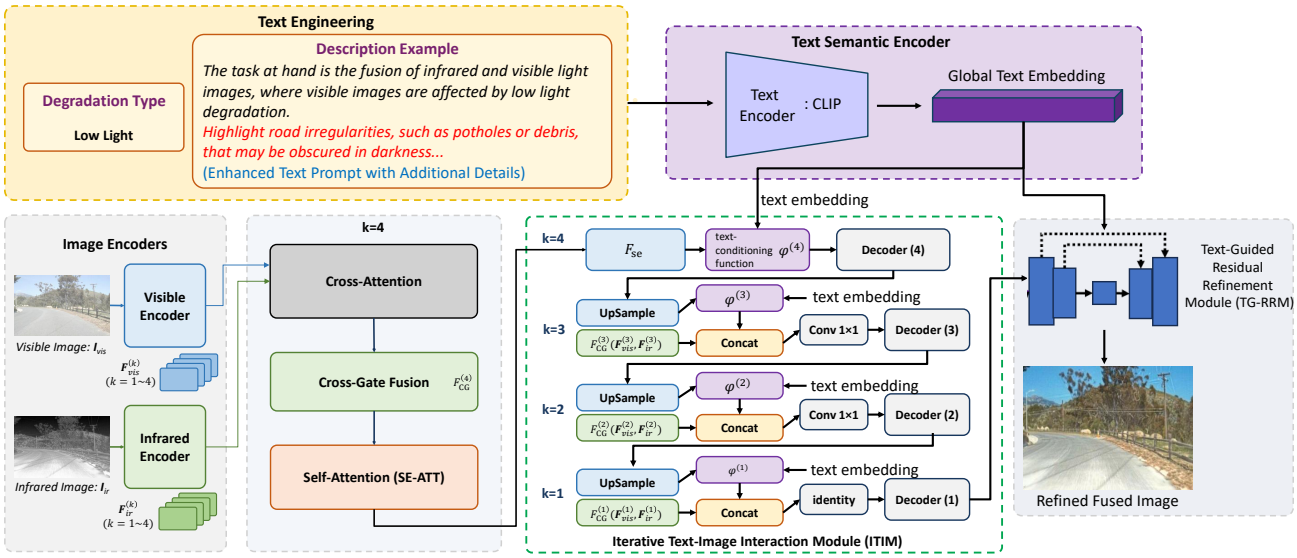

Figure 2. Overall workflow of our method. The framework includes two main parts: a text semantic encoder that extracts a global text embedding from the description generated by Text Engineering, and an image fusion pipeline containing Cross-Attention, Cross-Gate Fusion, Self-Attention (SE-ATT), ITIM, and TG-RRM.

3.1. Text engineering

Textual guidance has been used to guide multi-modal fusion. However, many existing text-guided fusion methods rely on coarse or generic task descriptions that mainly express high-level fusion objectives without explicitly specifying the affected modality, degradation category, or scenario-specific semantic priority. Such generic descriptions may not provide sufficient degradation-specific guidance, as the same or a similar prompt can be used under different input conditions. As a result, prompts that do not distinguish the affected modality and degradation type may produce similar global text embeddings, making it difficult to distinguish among fusion scenarios.

To address this issue, we manually construct degradation-aware prompt sets offline to provide more specific textual conditions. Rather than using simple task descriptions, the prompts are designed to include the fusion objective, dominant degradation characteristics, and application-relevant semantic priorities. Since the prompts are organized into predefined category-specific sets, the corresponding textual condition can be manually selected for each degradation scenario without changing the fusion-network architecture. In addition to describing the fusion task, the textual guidance includes manually specified semantic emphasis that varies with the target scenario. In safety-critical applications such as autonomous driving, objects including vehicles, pedestrians, and road structures need to remain clearly distinguishable under adverse imaging conditions. By including these priorities, the textual input provides an application-related scene-level condition along with degradation information. These object-related descriptions serve as global semantic priors rather than explicitly establishing correspondences between individual words and specific spatial regions. For example, the basic task description "The task at hand is the fusion of infrared and visible light images, where visible images are affected by low light degradation" can be extended with additional semantic guidance such as "... Highlight road irregularities, such as potholes or debris, that may be obscured in darkness." The resulting prompt contains a predefined global text-conditioning signal with more specific degradation and scene-level information than a generic fusion-task description. During training, the fusion network receives category-specific global text-conditioning signals through degradation-aware and semantically enriched prompts. The selected prompt is encoded by the frozen CLIP text encoder and used to condition hierarchical feature fusion and residual refinement.

3.2. Image fusion pipeline

Image Encoders: The visible and infrared encoders process their respective source images independently. To capture both spatial structures and high-level semantic information, we use Transformer-based blocks [31] as the backbone feature extractor.This allows the encoder to learn features that contain both local details and long-range dependencies across different scales. The encoding process is expressed as follows:

$$\boldsymbol{F}_{\mathrm{vis}} = \mathscr{F}_{\mathrm{v}}^{\mathrm{e}}\left(\boldsymbol{I}_{\mathrm{vis}}\right), \qquad \boldsymbol{F}_{\mathrm{ir}} = \mathscr{F}_{\mathrm{i}}^{\mathrm{e}}\left(\boldsymbol{I}_{\mathrm{ir}}\right), \tag{1}$$

where $\boldsymbol{I}_{\mathrm{vis}} \in \mathbb{R}^{H\times W\times 3}$ and $\boldsymbol{I}_{\mathrm{ir}} \in \mathbb{R}^{H\times W\times 3}$ denote the visible and infrared network inputs, respectively. In the implementation, the infrared image is loaded in three-channel RGB format before being passed to the encoder. H and W represent the height and width of the input image. $\mathscr{F}_{\mathrm{v}}^{\mathrm{e}}$ and $\mathscr{F}_{\mathrm{i}}^{\mathrm{e}}$ denote the visible encoder and infrared encoder. $\boldsymbol{F}_{\mathrm{vis}}$ and $\boldsymbol{F}_{\mathrm{ir}}$ collectively denote the four-level visible and infrared feature pyramids, respectively: $\boldsymbol{F}_{\mathrm{vis}} = \{\boldsymbol{F}_{\mathrm{vis}}^{(k)}\}_{k=1}^{4}$ and $\boldsymbol{F}_{\mathrm{ir}} = \{\boldsymbol{F}_{\mathrm{ir}}^{(k)}\}_{k=1}^{4}$. Here, $\boldsymbol{F}_{m}^{(k)}$ denotes the feature of modality $m \in \{\mathrm{vis}, \mathrm{ir}\}$ at the k-th encoder level. In our framework, multi-level features are extracted at four different stages of the encoder, corresponding to hierarchical representations at progressively reduced spatial resolutions. These multi-scale features are later utilized as inputs to different stages of the fusion and decoding modules.

Cross-Attention: The Cross-Attention layer performs bidirectional cross-modal interaction at the deepest encoder level by using cross-modal affinities to aggregate modality-specific value features. Before generating the query, key, and value representations, the two features are independently normalized using 16-group

GroupNorm. The normalized features are then projected by modality-specific 1×1 convolutions:

$$[\boldsymbol{Q}_m, \boldsymbol{K}_m, \boldsymbol{V}_m] = \mathrm{Reshape}_h\left(\mathrm{Split}_3\left[\mathrm{Conv}^{1\times1}_{\mathrm{qkv},m}\left(\mathrm{GN}_m\left(\boldsymbol{F}^{(4)}_m\right)\right)\right]\right), \quad m \in \{\mathrm{vis}, \mathrm{ir}\}, \tag{2}$$

where $\boldsymbol{Q}_m, \boldsymbol{K}_m, \boldsymbol{V}_m \in \mathbb{R}^{B\times h\times d\times N_4}$, $N_4 = H_4W_4$, and $d = C_4/h$. In our implementation, $h = 1$ and $d = C_4$.

Subsequently, bidirectional cross-modal interaction is performed. In the visible branch, the infrared query is matched against the visible keys and values, and the attended response is projected and added back to the visible feature. Symmetrically, the visible query is used to retrieve information from the infrared keys and values for the infrared branch. The implemented Cross-Attention can therefore be written as

$$\boldsymbol{F}^{\mathrm{ca}}_{\mathrm{v}} = \boldsymbol{F}^{(4)}_{\mathrm{vis}} + \mathscr{P}_{\mathrm{v}}\left[\mathrm{Softmax}\left(\frac{\boldsymbol{Q}_{\mathrm{i}}\boldsymbol{K}^{\top}_{\mathrm{v}}}{\sqrt{d}}\right)\boldsymbol{V}_{\mathrm{v}}\right], \quad \boldsymbol{F}^{\mathrm{ca}}_{\mathrm{i}} = \boldsymbol{F}^{(4)}_{\mathrm{ir}} + \mathscr{P}_{\mathrm{i}}\left[\mathrm{Softmax}\left(\frac{\boldsymbol{Q}_{\mathrm{v}}\boldsymbol{K}^{\top}_{\mathrm{i}}}{\sqrt{d}}\right)\boldsymbol{V}_{\mathrm{i}}\right], \tag{3}$$

where $\sqrt{d}$ is the scaling factor and $\mathscr{P}_{\mathrm{v}}(\cdot)$ and $\mathscr{P}_{\mathrm{i}}(\cdot)$ denote the modality-specific output projections. The single-head attention responses are reshaped into spatial feature maps, passed through modality-specific 1×1 output projections, and added to the original, unnormalized level-4 encoder features through residual connections.

Cross-Gate Fusion: While Cross-Attention enables effective information exchange between visible and infrared features, it does not explicitly regulate the reliability of the transferred information. Therefore, a Cross-Gate Fusion module is introduced to adaptively modulate the cross-attended features, allowing the network to selectively preserve informative cues and suppress degraded or irrelevant responses. At the deepest level, the Cross-Gate Fusion module takes the cross-attended features as inputs and can be expressed as $\boldsymbol{F}^{(4)}_{\mathrm{cg}} = \mathscr{F}^{(4)}_{\mathrm{CG}}(\boldsymbol{F}^{\mathrm{ca}}_{\mathrm{v}}, \boldsymbol{F}^{\mathrm{ca}}_{\mathrm{i}})$, where $\mathscr{F}^{(4)}_{\mathrm{CG}}(\cdot,\cdot)$ denotes the Cross-Gate Fusion operation at the deepest level. The application of Cross-Gate Fusion at the finer hierarchical levels is described in the subsequent formulation of ITIM. The detailed architecture is illustrated in Figure 3. First, we calculate the gating weights for the two image feature maps from the Cross-Attention layer:

$$\mathrm{gate}_{\mathrm{i}} = \sigma\left(\mathrm{Conv}^{1\times1}_{\mathrm{i}}\left(\boldsymbol{F}^{\mathrm{ca}}_{\mathrm{v}}\right)\right), \quad \mathrm{gate}_{\mathrm{v}} = \sigma\left(\mathrm{Conv}^{1\times1}_{\mathrm{v}}\left(\boldsymbol{F}^{\mathrm{ca}}_{\mathrm{i}}\right)\right), \tag{4}$$

where σ denotes the sigmoid function. The learned gating weights are then used to adaptively regulate the contribution of each modality at every spatial location, enabling selective preservation of informative features while suppressing less reliable responses, as formulated below:

$$\widehat{\boldsymbol{F}}^{(4)}_{\mathrm{cg}} = \boldsymbol{F}^{\mathrm{ca}}_{\mathrm{v}} \odot \mathrm{gate}_{\mathrm{v}} + \boldsymbol{F}^{\mathrm{ca}}_{\mathrm{i}} \odot \mathrm{gate}_{\mathrm{i}}, \tag{5}$$

where $\odot$ denotes the Hadamard product. For the final output, an additional convolutional layer is introduced to further refine the fused features and enhance their representational capability:

$$\boldsymbol{F}^{(4)}_{\mathrm{cg}} = \mathrm{Conv}^{(4)}_{1\times1}(\widehat{\boldsymbol{F}}^{(4)}_{\mathrm{cg}}).$$

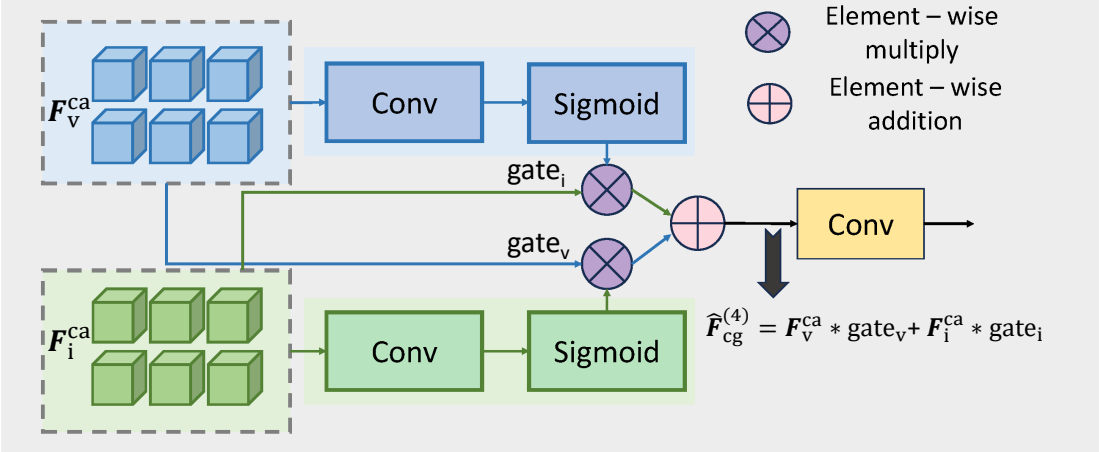

Figure 3. Architecture of the Cross-Gate Fusion module, illustrated using its application at the deepest level ($k = 4$). Cross-attended visible and infrared features generate cross-modal gates to regulate the opposite modality, after which the cross-gated features are added and projected through a convolution.

ITIM: The deepest-level Cross-Gate Fusion output is first processed by a single-head SE-ATT module. Before attention, the feature is normalized using 16-group GroupNorm, and a 1×1 convolution generates the query, key, and value representations. The attention response is output-projected and combined with a residual connection:

$$
\begin{aligned}
[\boldsymbol{Q}_{cg}, \boldsymbol{K}_{cg}, \boldsymbol{V}_{cg}] &= \mathrm{Tokenize}\left(\mathrm{Split}_3\left[\mathrm{Conv}_{qkv}^{1\times1}\left(\mathrm{GN}\left(\boldsymbol{F}_{cg}^{(4)}\right)\right)\right]\right), \\
\boldsymbol{F}_{se} &= \boldsymbol{F}_{cg}^{(4)} + \mathscr{P}_{se}\left[\mathrm{Softmax}\left(\frac{\boldsymbol{Q}_{cg}\boldsymbol{K}_{cg}^{\top}}{\sqrt{d}}\right)\boldsymbol{V}_{cg}\right],
\end{aligned}
\tag{6}
$$

where $\boldsymbol{Q}_{cg}$, $\boldsymbol{K}_{cg}$, and $\boldsymbol{V}_{cg}$ denote the query, key, and value projections of $\boldsymbol{F}_{cg}^{(4)}$, respectively, and $\mathscr{P}_{se}(\cdot)$ denotes the output projection. After this early fusion stage, the framework enters a stage where textual information actively participates in the fusion process.

Before applying text-conditioned feature modulation, the global text embedding is projected into a stage-specific semantic conditioning vector. At each hierarchical stage, the visual feature is represented as spatial tokens, while the single global text embedding is projected to generate a semantic conditioning signal. This text-conditioned representation is then mapped to affine modulation parameters that adjust the corresponding visual features.

Since a single pooled sentence-level text embedding is used rather than a sequence of token-level text features, the operation provides global semantic conditioning at each hierarchical stage instead of token-wise language–vision attention. It does not explicitly establish correspondences between individual words and specific image regions. This design repeatedly introduces the global prompt condition into visual features during hierarchical decoding, while spatial structure and local content remain encoded in the visual feature stream. The independently parameterized conditioning functions allow the same global text condition to modulate features at different semantic resolutions.

At the deepest encoder level ($k = 4$), the visible and infrared features first undergo the Cross-Attention

described above, followed by Cross-Gate Fusion and SE-ATT. The fused feature $\boldsymbol{F}_{\mathrm{se}}$ is then conditioned on the global text embedding using the stage-specific text-conditioning function $\varphi^{(4)}(\cdot,\cdot)$ and processed by a Transformer-based decoder block [31]:

$$\boldsymbol{F}_{\mathrm{d}}^{(4)} = \mathscr{F}_{\mathrm{D}}^{(4)}\left(\varphi^{(4)}\left(\boldsymbol{F}_{\mathrm{se}}, \boldsymbol{F}_{\mathrm{text}}\right)\right). \tag{7}$$

For the following levels $k \in \{3,2,1\}$, the decoded feature from the immediately coarser level is first upsampled to the current spatial resolution and then modulated by the same global text embedding. In parallel, the visible and infrared encoder features at the corresponding level are reintroduced and fused using a stage-specific Cross-Gate Fusion module. The text-conditioned decoder feature and the cross-gated encoder feature are then concatenated along the channel dimension:

$$\begin{aligned} \widetilde{\boldsymbol{F}}_{\mathrm{u}}^{(k)} &= \varphi^{(k)}\left(\mathscr{U}^{(k)}\left(\boldsymbol{F}_{\mathrm{d}}^{(k+1)}\right), \boldsymbol{F}_{\mathrm{text}}\right), \quad \boldsymbol{F}_{\mathrm{e}}^{(k)} = \mathscr{F}_{\mathrm{CG}}^{(k)}\left(\boldsymbol{F}_{\mathrm{vis}}^{(k)}, \boldsymbol{F}_{\mathrm{ir}}^{(k)}\right), \\ \boldsymbol{F}_{\mathrm{d}}^{(k)} &= \mathscr{F}_{\mathrm{D}}^{(k)}\left(\mathscr{R}^{(k)}\left([\widetilde{\boldsymbol{F}}_{\mathrm{u}}^{(k)}; \boldsymbol{F}_{\mathrm{e}}^{(k)}]\right)\right), \quad k = 3,2,1. \end{aligned} \tag{8}$$

Here, $\mathscr{U}^{(k)}(\cdot)$ represents the implemented upsampling operator, $[\cdot;\cdot]$ denotes channel-wise concatenation, and $\mathscr{R}^{(k)}(\cdot)$ denotes the channel-adjustment operation. Specifically, $\mathscr{R}^{(3)}(\cdot)$ and $\mathscr{R}^{(2)}(\cdot)$ are implemented with 1×1 convolutions, while $\mathscr{R}^{(1)}(\cdot)$ is the identity mapping. Thus, the concatenated level-1 feature is fed directly to the final decoder block without an additional channel-reduction convolution.

After the initial level-4 fusion, the encoder features used again during the remaining three interactions are the visible and infrared features at levels 3, 2, and 1, respectively. At the deepest level, Cross-Attention enables bidirectional high-level semantic interaction between the visible and infrared features. At the finer levels, the framework combines an upsampled text-conditioned decoder feature with a cross-gated encoder feature at the same resolution. In this way, the global text embedding is repeatedly injected into the hierarchical decoding process while modality-specific visual information is progressively recovered from the encoder branches.

3.3. Text semantic encoding and feature modulation

This subsection describes the stage-specific text-conditioning function $\varphi^{(k)}(\cdot,\cdot)$ used in Equations (7) and (8). Specifically, the visual feature provides the query representation, whereas the pooled CLIP text embedding provides the key and value representations. The resulting text-conditioned context is then projected to the visual channel dimension and used to generate feature-wise affine modulation parameters.

Text Semantic Encoder: The text semantic encoder transforms the input prompt into a pooled sentence-level text embedding. With large-scale vision-language pretraining, CLIP [13] provides text representations with strong semantic discrimination. To maintain linguistic consistency and reduce the risk of overfitting the text encoder to the fusion datasets, the CLIP text encoder is kept frozen throughout training. The text feature extraction process is expressed as

$$\boldsymbol{F}_{\text{text}} = \mathscr{F}_{\text{clip}}\left(T_{\text{text}}\right), \tag{9}$$

where $\boldsymbol{F}_{\text{text}} \in \mathbb{R}^{B\times 512}$ represents the pooled global text embedding obtained from the frozen CLIP ViT-B/32 text encoder. Only this single 512-dimensional embedding is retained for each prompt.

Stage-Specific Visual-text Conditioning: Let $\boldsymbol{F}_{\text{f}}^{(k)} \in \mathbb{R}^{B\times C_k\times H_k\times W_k}$ represent the visual feature entering the text-conditioning module at the k-th interaction stage, where C_k, H_k, and W_k denote the number of channels, height, and width, respectively. The visual feature is first flattened into $N_k = H_k W_k$ spatial tokens:

$$\boldsymbol{X}_{\text{f}}^{(k)} = \text{reshape}\left(\boldsymbol{F}_{\text{f}}^{(k)}\right) \in \mathbb{R}^{B\times N_k\times C_k}. \tag{10}$$

The visual tokens and the global text embedding are independently normalized and projected to a common embedding dimension $D = 512$:

$$\boldsymbol{Z}_{\text{f}}^{(k)} = \boldsymbol{W}_{\text{F}}^{(k)}\,\text{LN}_{\text{F}}^{(k)}\left(\boldsymbol{X}_{\text{f}}^{(k)}\right), \quad \boldsymbol{Z}_{\text{text}}^{(k)} = \boldsymbol{W}_{\text{T}}^{(k)}\,\text{LN}_{\text{T}}^{(k)}\left(\boldsymbol{F}_{\text{text}}\right). \tag{11}$$

Here, $\boldsymbol{Z}_{\text{f}}^{(k)} \in \mathbb{R}^{B\times N_k\times D}$. The projected text embedding is regarded as a single text token, such that $\boldsymbol{Z}_{\text{text}}^{(k)} \in \mathbb{R}^{B\times 1\times D}$. Each hierarchical stage uses an independent set of projection parameters.

The query representation is generated from the projected visual tokens, whereas the key and value representations are generated from the projected text token:

$$\boldsymbol{Q}^{(k)} = \boldsymbol{W}_{\text{Q}}^{(k)}\boldsymbol{Z}_{\text{f}}^{(k)}, \quad \boldsymbol{K}^{(k)} = \boldsymbol{W}_{\text{K}}^{(k)}\boldsymbol{Z}_{\text{text}}^{(k)}, \quad \boldsymbol{V}^{(k)} = \boldsymbol{W}_{\text{V}}^{(k)}\boldsymbol{Z}_{\text{text}}^{(k)}. \tag{12}$$

Single-head attention is then computed using $\boldsymbol{Q}^{(k)}$, $\boldsymbol{K}^{(k)}$, and $\boldsymbol{V}^{(k)}$ to obtain the text-conditioned context $\boldsymbol{C}_{\text{ctx}}^{(k)} \in \mathbb{R}^{B\times N_k\times C_k}$. Because the key and value representations are derived from a single pooled text token, the resulting context represents a global prompt condition rather than an explicit token-to-region correspondence. Spatial structure and local content remain encoded by the visual feature stream and the subsequent visual processing operations.

Semantic Feature Modulation: The text-conditioned context $\boldsymbol{C}_{\text{ctx}}^{(k)}$ obtained above contains the textual information used to regulate the visual feature at the k-th interaction stage. To apply this context to the visual feature, it is first mapped to a pair of affine modulation parameters:

$$[\boldsymbol{\gamma}^{(k)}, \boldsymbol{\beta}^{(k)}] = \text{Split}_2\left[\text{GELU}\left(\boldsymbol{W}_{\gamma\beta}^{(k)}\boldsymbol{C}_{\text{ctx}}^{(k)} + \boldsymbol{b}_{\gamma\beta}^{(k)}\right)\right], \tag{13}$$

where $\boldsymbol{W}_{\gamma\beta}^{(k)} : \mathbb{R}^{C_k} \to \mathbb{R}^{2C_k}$. The $2C_k$-dimensional output is divided equally into the scaling parameter $\boldsymbol{\gamma}^{(k)}$ and the bias parameter $\boldsymbol{\beta}^{(k)}$. The two parameters are generated together by the same linear projection followed by GELU, while different hierarchical stages use independent parameter sets.

After reshaping $\boldsymbol{\gamma}^{(k)}$ and $\boldsymbol{\beta}^{(k)}$ from token form into spatial feature maps, they are applied to the input visual feature $\boldsymbol{F}_{\text{f}}^{(k)}$ as

$$\widehat{\boldsymbol{F}}_{\text{f}}^{(k)} = \varphi^{(k)}\left(\boldsymbol{F}_{\text{f}}^{(k)}, \boldsymbol{F}_{\text{text}}\right) = \left(1 + \text{reshape}\left(\boldsymbol{\gamma}^{(k)}\right)\right) \odot \boldsymbol{F}_{\text{f}}^{(k)} + \text{reshape}\left(\boldsymbol{\beta}^{(k)}\right). \tag{14}$$

Here, $\boldsymbol{\gamma}^{(k)}$ controls the feature scaling, while $\boldsymbol{\beta}^{(k)}$ provides an additive feature shift. $\widehat{\boldsymbol{F}}_{\text{f}}^{(k)}$ denotes

the text-modulated output feature, whereas $\boldsymbol{F}_{\mathrm{f}}^{(k)}$ denotes the visual feature before text conditioning. The function $\varphi^{(k)}(\cdot,\cdot)$ denotes the complete text-conditioning operation at the k-th stage, and its parameters are not shared across hierarchical stages.

3.4. TG-RRM

After four hierarchical text-image interactions, the network produces a high-resolution fused feature representation containing cross-modal information guided by text. Although this intermediate feature captures the global fusion structure, some local ambiguities and degradation-related residual responses may remain before final image reconstruction.

To address these issues, refinement stages have been used in image restoration and fusion frameworks [31,32]. These methods improve visual quality by correcting prediction errors and preserving fine details, often using convolutional or transformer-based layers. Refinement has also been studied in event-based stereo depth estimation, where URNet [33] combines local-global refinement with uncertainty modeling to improve disparity prediction. However, refining infrared-visible fused features under complex degradation conditions is still challenging. Coarse outputs may contain structural ambiguities and inconsistent responses across regions, making it difficult for shallow or single-scale refinement modules to recover fine details and coherent structures.

To address these problems, we introduce TG-RRM to refine local details and improve structural coherence. As shown in Figure 4, TG-RRM uses a U-Net architecture that takes the high-resolution fused feature as input and predicts a residual feature correction. This correction is added to the input feature, and the network's output layer then reconstructs the three-channel fused image. The encoder-decoder structure progressively extracts and reconstructs features, while skip connections pass low-level spatial information directly to the decoder to preserve fine textures and sharp boundaries.

The refinement process also uses the global CLIP text embedding $\boldsymbol{F}_{\mathrm{text}}$ to guide residual correction. Specifically, textual information is injected into both encoder and decoder stages through a feature-wise modulation mechanism. The text embedding passes through a two-layer multilayer perceptron (MLP) to generate affine parameters $(\boldsymbol{\gamma}_{\mathrm{r}}, \boldsymbol{\beta}_{\mathrm{r}})$, which modulate intermediate feature responses according to $\boldsymbol{x} \leftarrow (1+\boldsymbol{\gamma}_{\mathrm{r}}) \odot \boldsymbol{x} + \boldsymbol{\beta}_{\mathrm{r}}$. The text-conditioned affine transformations used at different encoder stages, the bottleneck, decoder stages, and the residual output stage are independently parameterized and do not share parameters across feature levels. At each modulation site, a separate two-layer MLP with dimensions $512 \rightarrow 1024 \rightarrow 2C$ and a LeakyReLU activation between the two linear layers jointly produces the $2C$-dimensional affine vector, which is subsequently split into the scaling term $\boldsymbol{\gamma}_{\mathrm{r}}$ and bias term $\boldsymbol{\beta}_{\mathrm{r}}$. Thus, the two affine parameters share the same MLP within one site, while different scales and refinement sites use independent MLP parameters.

With textual guidance, TG-RRM adjusts feature responses according to the global prompt condition while reducing degradation-related artifacts. The combination of multi-scale U-Net refinement and global semantic conditioning helps TG-RRM maintain visual consistency and preserve structural details. The spatial variation of the refined features mainly depends on the visual feature stream rather than explicit text-to-region alignment.

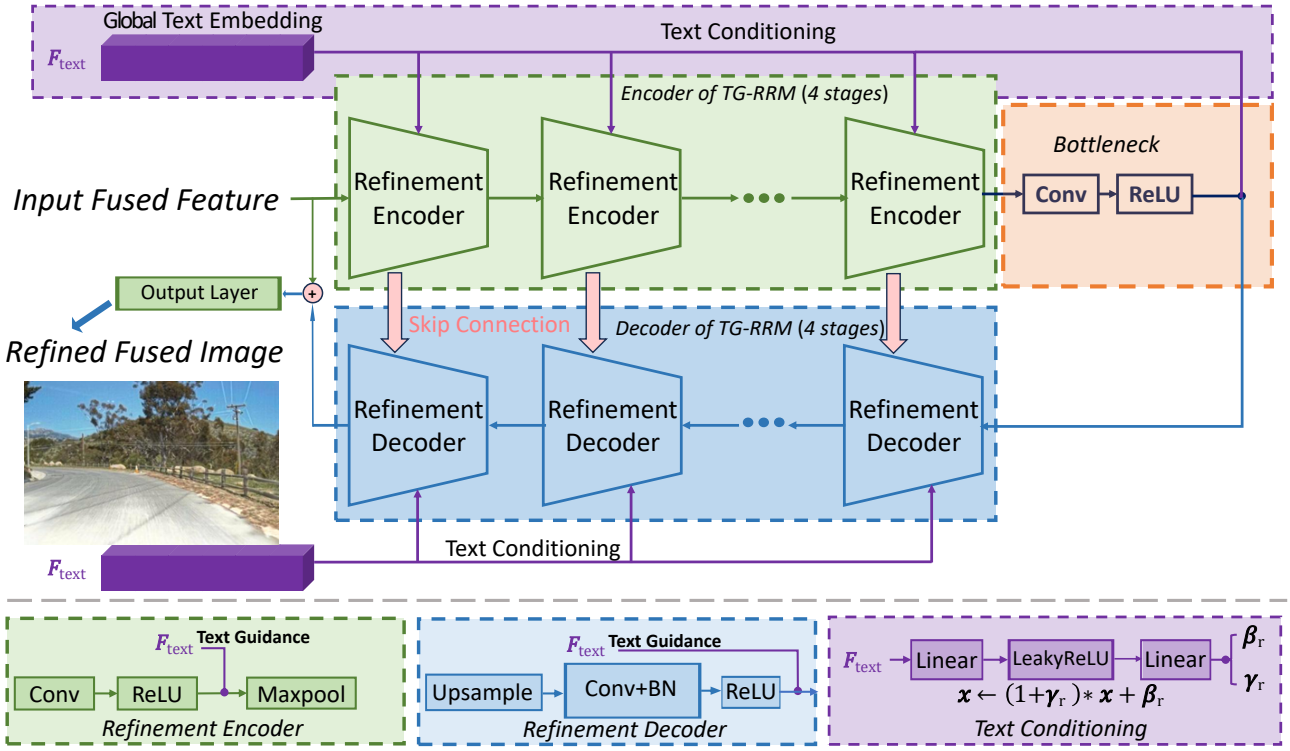

Figure 4. The U-Net structured TG-RRM, which consists of four encoder-decoder stages with text-conditioned modulation.

3.5. Loss functions

The loss function largely determines what source information is preserved and how different modalities contribute to the final fused result. Similar to previous text-guided fusion methods, the fusion-related losses include intensity loss, SSIM loss [34], maximum gradient loss, and color consistency loss. Since ground-truth fused images are unavailable, the fusion network is trained with multiple loss terms to preserve visual saliency, structural information, and color consistency. The overall training objective is defined as a weighted sum of these losses.

Intensity Loss: The intensity loss preserves the source pixel chosen according to a luminance comparison while retaining the three-channel value of the selected modality. Let

$$\bar{\boldsymbol{I}}_m = \frac{1}{3}\sum_{c=1}^{3} \boldsymbol{I}_{m,c}, \quad \boldsymbol{M} = \mathbf{1}\left[\bar{\boldsymbol{I}}_{\mathrm{ir}} > \bar{\boldsymbol{I}}_{\mathrm{vis}}\right], \tag{15}$$

where $m \in \{\mathrm{vis}, \mathrm{ir}\}$ and $\mathbf{1}[\cdot]$ denotes the indicator function. The binary mask $\boldsymbol{M}$ is obtained by comparing the channel-averaged infrared and visible intensities at each spatial location and is broadcast across the three image channels. The intensity target and loss are

$$\boldsymbol{I}_{\max} = \boldsymbol{M} \odot \boldsymbol{I}_{\mathrm{ir}} + (1 - \boldsymbol{M}) \odot \boldsymbol{I}_{\mathrm{vis}}, \quad \mathscr{L}_{\mathrm{int}} = \frac{1}{3BHW} \left\| \boldsymbol{I}_{\mathrm{fused}} - \boldsymbol{I}_{\max} \right\|_1, \tag{16}$$

where $\|\cdot\|_1$ denotes the ℓ_1 norm, computed as the sum of the absolute values of all tensor elements. For samples from the EMS dataset, the corresponding high-quality source pair provided by the dataset is used to construct the fusion losses. For standard datasets without separate clean counterparts, the available

infrared and visible source images are used directly.

Structural Similarity Loss: To preserve structural information from both modalities, the implementation computes a visible-image SSIM term in RGB space and an infrared SSIM term after grayscale conversion. The structural similarity loss is defined as

$$\mathscr{L}_{\text{SSIM}} = w_{\text{ssim-vis}}[1 - \text{SSIM}(\boldsymbol{I}_{\text{fused}}, \boldsymbol{I}_{\text{vis}})] + w_{\text{ssim-ir}}[1 - \text{SSIM}(\text{G}(\boldsymbol{I}_{\text{fused}}), \text{G}(\boldsymbol{I}_{\text{ir}}))], \tag{17}$$

where $\text{G}(\cdot)$ denotes RGB-to-grayscale conversion, and $w_{\text{ssim-vis}}$ and $w_{\text{ssim-ir}}$ control the contributions of the visible and infrared structural similarity terms, respectively. The SSIM index is computed in its standard form:

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + \text{C}_1)(2\sigma_{xy} + \text{C}_2)}{(\mu_x^2 + \mu_y^2 + \text{C}_1)(\sigma_x^2 + \sigma_y^2 + \text{C}_2)}, \tag{18}$$

where μ, σ, and σ_{xy} denote local means, standard deviations, and cross-covariance computed within a sliding window. C_1 and C_2 are small constants that prevent division by zero.

Maximum Gradient Loss: The maximum gradient loss is introduced to preserve sharp edges from both modalities. In the implementation, the visible, infrared, and fused images are first converted to grayscale. Horizontal and vertical absolute Sobel responses are then calculated separately, and the fused response is matched to the element-wise maximum source response in each direction:

$$\begin{aligned} \mathscr{L}_{\text{grad}} = {} & \frac{1}{BHW} \left\| \text{D}_x(\text{G}(\boldsymbol{I}_{\text{fused}})) - \max\left(\text{D}_x(\text{G}(\boldsymbol{I}_{\text{vis}})), \text{D}_x(\text{G}(\boldsymbol{I}_{\text{ir}}))\right) \right\|_1 \\ & + \frac{1}{BHW} \left\| \text{D}_y(\text{G}(\boldsymbol{I}_{\text{fused}})) - \max\left(\text{D}_y(\text{G}(\boldsymbol{I}_{\text{vis}})), \text{D}_y(\text{G}(\boldsymbol{I}_{\text{ir}}))\right) \right\|_1, \end{aligned} \tag{19}$$

where $\text{D}_x(\cdot)$ and $\text{D}_y(\cdot)$ denote the absolute horizontal and vertical Sobel responses after grayscale conversion. The Sobel filters are applied after one-pixel replicate padding, and the absolute horizontal and vertical responses are matched separately.

Color Consistency Loss: The color consistency loss encourages the fused image to preserve the chrominance information of the visible image. Both images are transformed to YCbCr space, and the Cb and Cr channels are constrained as follows:

$$\mathscr{L}_{\text{color}} = \frac{1}{BHW} \left\| \text{C}_\text{b}(\boldsymbol{I}_{\text{fused}}) - \text{C}_\text{b}(\boldsymbol{I}_{\text{vis}}) \right\|_1 + \frac{1}{BHW} \left\| \text{C}_\text{r}(\boldsymbol{I}_{\text{fused}}) - \text{C}_\text{r}(\boldsymbol{I}_{\text{vis}}) \right\|_1. \tag{20}$$

Overall Training Objective: The overall training objective is composed of the intensity, structural similarity, color consistency, and maximum gradient losses:

$$\mathscr{L}_{\text{total}} = w_{\text{int}}\mathscr{L}_{\text{int}} + w_{\text{ssim}}\mathscr{L}_{\text{SSIM}} + w_{\text{color}}\mathscr{L}_{\text{color}} + w_{\text{grad}}\mathscr{L}_{\text{grad}}, \tag{21}$$

where w_{int}, w_{ssim}, w_{color}, and w_{grad} denote the coefficients of the corresponding loss terms.

Prompt-Conditioned Coefficient Generation: For each degradation category t, the loss function is associated with a predefined base coefficient vector $\boldsymbol{w}_{\text{base}}^{(t)}$. The frozen CLIP sentence-level representation is mapped to a bounded coefficient offset vector:

$$\boldsymbol{g} = P_{\text{coef}}(\boldsymbol{F}_{\text{text}}), \quad \boldsymbol{g} \in [-1,1]^6, \tag{22}$$

where $P_{\text{coef}}(\cdot)$ denotes the coefficient-mapping function. The resulting loss coefficients are calculated as

$$\boldsymbol{w}^{(t)} = \text{ReLU}\left(\boldsymbol{w}_{\text{base}}^{(t)} + \boldsymbol{\alpha} \odot \boldsymbol{g}\right), \tag{23}$$

where $\boldsymbol{\alpha}$ controls the bounded variation range, $\odot$ denotes element-wise multiplication, and ReLU ensures nonnegative coefficients. The degradation-specific base coefficients and the predefined scaling vector $\boldsymbol{\alpha}$ are reported in Section 4.1.

4. Results

In this section, we first describe the implementation details and experimental configurations. We then evaluate the effectiveness of the proposed method through both qualitative and quantitative comparisons, with a particular focus on text-guided image fusion. Finally, ablation experiments are conducted to analyze the contributions of different components in the proposed framework.

4.1. Implementation details

IT-TextFusion is implemented in PyTorch and trained on NVIDIA RTX A5000 GPUs.

We use the AdamW optimizer with an initial learning rate of 1×10^{-4}, weight decay of 5×10^{-2}, and batch size of 8. Training is conducted for 120 epochs (50 epochs for the Low-Light Visible and Infrared Paired (LLVIP) dataset), and the checkpoint achieving the lowest validation loss is selected for final evaluation. The model is independently trained for each dataset used in this study. During training, input images are randomly cropped to a spatial resolution of 96×96. Horizontal and vertical flips are applied with a probability of 0.5 to improve generalization.

The visible and infrared encoders share the same channel configuration, with $E_1 = 48$, $E_2 = 96$, $E_3 = 192$, and $E_4 = 384$. The hierarchical decoder contains two Transformer blocks at each level. From the deepest to the shallowest level, the decoder uses 384, 192, 96, and 96 channels with 8, 4, 2, and 1 self-attention heads, respectively. For the prompt-conditioned coefficient generation described in Section 3.5, using the coefficient order $[w_{\text{int}}, w_{\text{ssim-vis}}, w_{\text{ssim-ir}}, w_{\text{ssim}}, w_{\text{color}}, w_{\text{grad}}]$, the predefined base coefficient vectors are set to $[8,1,1,1,12,10]$ for Low Light and Infrared Low Contrast, $[4,1,1,0,12,2]$ for Overexposure, $[6,1,1,1,12,10]$ for Infrared Random Noise and Infrared Stripe Noise, and $[5,1,1,1,12,8]$ for Visible Blur, Visible Haze, Rain, and Visible Random Noise. For all datasets, the corresponding coefficient configuration is selected according to the degradation type of the data. The predefined scaling vector is $\boldsymbol{\alpha} = [3, 0.5, 0.5, 1, 4, 4]$.

At test time, the prompt is manually specified for each evaluation dataset or degradation subset according to its predefined degradation category. The current implementation does not include an automatic degradation classifier or prompt-selection module. Therefore, the degradation-aware evaluation assumes that the degradation category is available in advance. When such information is unavailable, a fixed generic prompt may instead be manually supplied through the same text-input interface, as evaluated in Section 4.4, although it does not provide degradation-specific guidance.

4.2. Datasets

Standard Datasets: We evaluate our model on three infrared-visible fusion datasets: Multi-Spectral Road Scenes (MSRS) [35], LLVIP [36], and RoadScene [17]. These datasets mainly cover single-degradation conditions. MSRS and LLVIP primarily contain low-light visible images. RoadScene includes outdoor traffic scenes with illumination imbalance and exposure variations, offering another real-world evaluation scenario.

For MSRS, we use 1083 training image pairs and 361 evaluation image pairs. LLVIP contains 12,025 training image pairs and 3463 evaluation image pairs, while RoadScene includes 177 training image pairs and 44 evaluation image pairs in our experiments. Since MSRS and LLVIP do not provide ground-truth fused images, the original infrared and visible image pairs are used as supervision targets during training and loss computation.

EMS Dataset: Besides the standard fusion datasets, we use the EMS dataset released by Text-IF [9] to assess robustness under different degradation conditions. The dataset includes nine degradation types: Low Light, Rain, Visible Blur, Overexposure, Visible Haze, Visible Random Noise, Infrared Low Contrast, Infrared Random Noise, and Infrared Stripe Noise. For EMS, we use 2934 degraded RGB–IR image pairs for training and 968 image pairs for evaluation. Each training pair is accompanied by a corresponding clean RGB–IR pair. Additionally, the dataset provides multiple text prompts describing different degradation types to support text-guided image fusion. For degradation types that did not originally include textual descriptions, matching prompts are manually constructed using the sentence patterns and semantic structure of the existing prompts to maintain consistency across all degradation scenarios.

4.3. Evaluation metrics and comparison methods

Comparison Methods: We compare the proposed method with several state-of-the-art (SOTA) methods on multiple datasets, which are described in Section 4.2. In particular, the proposed method is mainly compared with Text-IF [9], a recent representative framework for text-guided image fusion. Other methods for comparison include U2Fusion [17], MetaFusion [18], DDFM [19], MURF [37], ReCoNet [38], TarDAL [39], and UMF-CMGR [40], which are conventional image fusion approaches relying purely on visual information without involving any text or language-based guidance. For the EMS dataset described in Section 4.2, we compare only with Text-IF. Among the text-guided methods discussed in the related work, only Text-IF was selected for quantitative comparison because it directly addresses degradation-aware infrared–visible fusion and provides the closest prompt setting and evaluation protocol. In contrast, TextFusion and TeRF focus on task-, object-, or region-controllable fusion under different prompting and supervision settings; MGFusion uses semantic features extracted from a frozen CLIP image encoder to guide fusion under a different training setup; and TITFormer uses simulated infrared information for image enhancement rather than paired infrared–visible fusion. Because these methods differ in task formulation, input construction, supervision, and training protocol, their reported results are not directly comparable under the experimental setting adopted in this study. We therefore restrict our quantitative conclusions to the methods evaluated under the current protocol.

Evaluation Metrics: Following the evaluation protocol of Text-IF [9], we employ a comprehensive

set of classical fusion metrics, including Information Entropy (EN) [41], Standard Deviation (SD) [42], Spatial Frequency (SF) [43], Mutual Information (MI) [44], Sum of the Correlations of Differences (SCD) [45], Visual Information Fidelity (VIF) [46], Quality of Gradient-based Fusion $Q^{AB/F}$ [44], and a fusion-specific summed metric constructed based on the SSIM [47], denoted as $\mathrm{SSIM}_{\mathrm{sum}}$.

Since no ground-truth fused image is available, the structural similarity evaluation is computed between the fused image and each of the two source images and then summed:

$$\mathrm{SSIM}_{\mathrm{sum}} = \mathrm{SSIM}(\boldsymbol{I}_{\mathrm{f}}, \boldsymbol{I}_{\mathrm{ir}}) + \mathrm{SSIM}(\boldsymbol{I}_{\mathrm{f}}, \boldsymbol{I}_{\mathrm{vis}}), \tag{24}$$

where $\boldsymbol{I}_{\mathrm{f}}$, $\boldsymbol{I}_{\mathrm{ir}}$, and $\boldsymbol{I}_{\mathrm{vis}}$ denote the grayscale fused, infrared, and visible images, respectively. Each pairwise SSIM term is computed using the standard definition in Equation (18), with the image intensity range set to 255. Therefore, $\mathrm{SSIM}_{\mathrm{sum}}$ is not bounded by 1 and may take values greater than 1. A higher value indicates that the fused image jointly preserves more structural information from the two source modalities.

To complement the source-based fusion metrics, we additionally report CLIP-IQA [48], the Natural Image Quality Evaluator (NIQE) [49], the Multi-scale Image Quality Transformer (MUSIQ) [50], and the Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE) [51], which assess no-reference perceptual quality and image naturalness using different statistical or learned representations. Although CLIP-IQA uses CLIP-derived representations, the adopted implementation receives only the fused image and does not use the input degradation-aware prompt. Therefore, it is treated as a no-reference perceptual-quality metric rather than a direct measure of image-prompt semantic alignment. Lower NIQE and BRISQUE values generally suggest better perceptual quality, while higher EN, $Q^{AB/F}$, VIF, and $\mathrm{SSIM}_{\mathrm{sum}}$ values reflect better information content, gradient preservation, visual fidelity, and joint structural preservation, respectively. For consistency with previous studies, all metrics are calculated using publicly available implementations commonly used in image-fusion research.

4.4. Comparison with a fixed generic prompt

To evaluate fusion performance without degradation-specific textual guidance, we first test the model using a fixed generic prompt, just like "This is an infrared and visible image fusion task." Under this setting, the textual branch remains active, while the input prompt contains no degradation-aware or scene-specific semantic information. This setting is also consistent with our primary baseline Text-IF [9]. Therefore, rather than treating this experiment as a text-free visual-only baseline, we regard it as an experiment using a fixed generic prompt and use it for controlled comparison. This allows us to examine whether the proposed architecture improves fusion performance without explicit degradation-aware text guidance. The fixed-prompt comparison results are presented in Table 1.

Table 1 shows that the proposed method performs competitively with a fixed generic prompt, although its advantages vary across datasets and metrics. On MSRS, the proposed method achieves higher SCD, EN, VIF, and $Q^{AB/F}$ than Text-IF, but its SD is lower. On LLVIP, the proposed method improves SD, EN, VIF, and $Q^{AB/F}$, while SCD decreases slightly. On RoadScene, the proposed method achieves higher SD and $Q^{AB/F}$, whereas Text-IF has higher SCD and VIF, and both methods obtain the same EN. Overall, the architectural changes lead to different trade-offs among fusion properties rather than

consistent improvements across all metrics.

Table 1. Quantitative comparison on the MSRS, LLVIP, and RoadScene datasets. Text-IF and IT-TextFusion use the same fixed generic prompt without degradation-aware or scene-specific information. The conventional visual baselines do not include textual input (**Bold**: best performance).

Methods	MSRS Dataset					LLVIP Dataset					RoadScene Dataset				
	SCD↑	SD↑	EN↑	VIF↑	$Q^{AB/F}$ ↑	SCD↑	SD↑	EN↑	VIF↑	$Q^{AB/F}$ ↑	SCD↑	SD↑	EN↑	VIF↑	$Q^{AB/F}$ ↑
MURF [37]	0.868	16.431	5.047	0.413	0.327	0.514	21.834	6.051	0.386	0.206	1.576	36.788	6.992	0.484	0.432
ReCoNet [38]	1.191	44.374	3.895	0.438	0.367	1.345	41.234	5.514	0.513	0.364	1.589	37.580	6.822	0.504	0.354
TarDAL [39]	1.484	35.460	6.347	0.673	0.426	0.817	39.070	5.349	0.330	0.252	1.415	42.609	7.054	0.525	0.391
UMF-CMGR [40]	0.981	20.819	5.600	0.430	0.266	1.029	31.501	6.569	0.509	0.352	1.613	36.251	6.973	0.554	0.429
U2Fusion [17]	1.182	23.541	5.246	0.506	0.372	0.757	23.614	5.972	0.552	0.341	1.498	30.969	6.739	0.513	0.467
MetaFusion [18]	1.486	39.432	6.368	0.726	0.478	1.317	42.446	6.823	0.833	0.493	1.581	**50.613**	7.223	0.512	0.338
DDFM [19]	1.550	32.749	5.693	0.622	0.431	1.414	38.346	6.979	0.549	0.220	**1.864**	44.925	7.226	0.544	0.413
Text-IF [9]	1.681	**44.564**	6.789	1.046	0.676	**1.591**	48.834	7.325	1.011	0.616	1.572	48.962	7.332	**0.739**	0.578
IT-TextFusion (Ours)	**1.721**	42.889	**6.882**	**1.064**	**0.725**	1.520	**50.451**	**7.443**	**1.042**	**0.760**	1.102	49.102	**7.332**	0.733	**0.636**

Compared with Text-IF, our method uses an iterative text-image interaction mechanism, where the same global text embedding modulates features at all four hierarchical decoding stages and is introduced again during residual refinement. With a fixed generic prompt, this repeated conditioning allows textual information to be used at multiple feature resolutions and increases the interaction between progressively decoded fusion features and reintroduced encoder features. The results also suggest that the performance gains do not depend entirely on degradation-aware textual semantics, as hierarchical text-conditioned interaction and subsequent residual refinement contribute to fusion performance.

4.5. Comparison with degradation-aware text guidance

4.5.1. Evaluation on standard datasets

After testing the model with a fixed generic prompt, we further examine its performance using degradation-aware text guidance, where the prompts describe the degradation characteristics of the source images. In real-world scenes, source images may be affected by different degradations, including poor illumination, noise, and low contrast. Among the compared methods introduced above, Text-IF [9] is the only framework that explicitly incorporates textual guidance and is designed to handle diverse degradation conditions in a unified manner. In contrast, the other comparison methods are conventional image fusion approaches that rely solely on visual information and are not specifically designed to address degraded inputs. For the conventional visual fusion baselines, we construct two-stage degradation-handling pipelines by applying URetinex-Net [52] to low-light inputs, AirNet [53] to low-contrast inputs, the masked-training-based denoising approach [54] to noisy inputs, and Learning Multi-Scale Photo Exposure Correction (LMPEC) [55] to overexposed inputs before fusion. Text-IF and IT-TextFusion instead process the degraded infrared–visible pair directly. Therefore, Table 2 should be interpreted as a system-level comparison of complete degradation-handling pipelines rather than a controlled architecture-level comparison in which all fusion backbones directly receive identical degraded inputs without preprocessing.

Table 2. System-level quantitative comparison on degraded source images (MSRS and LLVIP: low-light visible images; RoadScene: overexposed visible images). Each pipeline receives the same degraded infrared–visible pair as its external input. Conventional visual fusion methods are preceded by a degradation-specific image-restoration model, denoted by "*eir.*," while Text-IF and IT-TextFusion process the degraded pair directly. The table thus compares complete degradation-handling pipelines rather than isolated fusion backbones receiving identical internal inputs. (**Bold**: best performance).

Methods	MSRS Dataset			LLVIP Dataset			RoadScene Dataset		
	CLIP-IQA↑	EN↑	NIQE↓	EN↑	NIQE↓	MUSIQ↑	SF↑	NIQE↓	BRISQUE↓
eir.+UMF-CMGR	0.101	6.316	3.738	7.087	3.891	47.543	11.047	3.792	32.485
eir.+TarDAL	0.082	5.855	4.750	7.042	3.659	41.735	11.789	3.667	32.436
eir.+ReCoNet	0.117	7.216	5.769	7.109	4.695	44.187	10.312	4.785	37.775
eir.+MURF	0.111	5.872	4.199	6.757	4.177	50.589	15.605	3.779	30.594
eir.+U2Fusion	0.127	6.724	3.997	7.439	3.969	48.481	18.006	4.215	34.577
eir.+MetaFusion	0.106	**7.302**	3.584	7.495	3.722	49.620	**26.653**	3.473	29.500
eir.+DDFM	0.094	6.723	**3.465**	7.150	5.184	35.933	10.493	3.717	32.334
Text-IF [9]	0.132	7.172	3.708	7.391	3.502	48.625	17.766	3.342	29.021
IT-TextFusion (Ours)	**0.145**	6.689	3.906	**7.502**	**3.336**	**54.067**	15.332	**3.330**	**27.571**

Quantitative Comparison: Table 2 presents the quantitative results on the three standard infrared-visible fusion datasets. Overall, the proposed method shows different strengths and trade-offs across datasets and evaluation metrics, rather than outperforming all competing methods on every metric.

On MSRS, IT-TextFusion obtains the highest CLIP-IQA score (0.145), showing its advantage in this no-reference perceptual-quality metric. However, its EN and NIQE are not the best among the compared methods, reflecting differences in performance across perceptual-quality, information-content, and naturalness measures.

On LLVIP, IT-TextFusion achieves the highest EN (7.502) and MUSIQ (54.067), as well as the lowest NIQE (3.336). These results show better information content and performance on selected no-reference perceptual-quality metrics for this dataset.

On RoadScene, the proposed method obtains the lowest NIQE (3.330) and BRISQUE (27.571), while its SF (15.332) remains below the best competing score. The method therefore performs well on selected no-reference image-quality metrics but does not lead in all fusion metrics.

Qualitative Comparison: Figure 5 shows the visual comparison between the proposed method and Text-IF [9] on the three standard datasets (LLVIP, MSRS, and RoadScene). The visual results are generally consistent with the quantitative findings in Table 2. On LLVIP, where visible images are affected by severe low-light degradation, our method produces fused images with more suitable brightness, better color saturation, and clearer structural details than Text-IF. On RoadScene, our method provides more balanced brightness and clearer textures in the illustrated overexposed scenes. On MSRS, however, both methods produce visually comparable fusion results, suggesting a relatively small performance difference under this degradation setting.

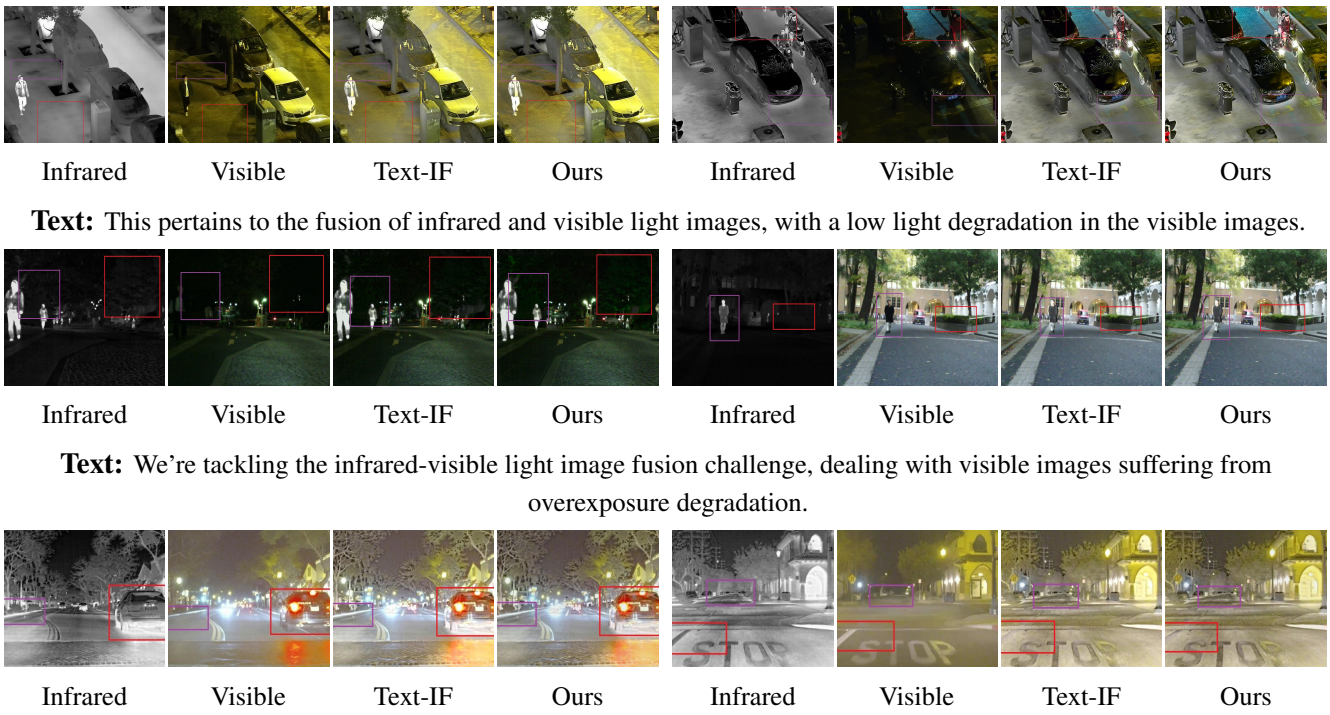

Figure 5. Qualitative comparison with the baseline method Text-IF [9] on the three standard datasets. Degradations from top to bottom: Low Light (LLVIP), Low Light (MSRS), and Overexposure (RoadScene). For each dataset, Text-IF and our method use the same degradation-aware text prompt to ensure a fair comparison.

4.5.2. Evaluation on the EMS dataset

We evaluate our method on the EMS dataset under diverse degradation conditions. Following the evaluation protocol adopted by Text-IF [9], we report eight fusion-related metrics: EN, SD, SF, MI, SCD, VIF, $Q^{AB/F}$, and SSIM_{sum}. These metrics assess different aspects of the fused images, including information content, contrast, spatial activity, source-information preservation, structural consistency, and visual fidelity.

Quantitative Comparison: Table 3 shows that our method achieves better results than Text-IF on several fusion metrics, including EN, SD, SF, SCD, $Q^{AB/F}$, and VIF. The higher SCD, VIF, and $Q^{AB/F}$ scores indicate better preservation of complementary source information and structural details under different degradation conditions. However, the improvements vary across metrics: Text-IF obtains slightly higher MI and SSIM_{sum}, while the difference in EN remains small. Overall, our method shows advantages in several fusion metrics but does not outperform Text-IF on every criterion.

Table 3. Average quantitative comparison of our method with Text-IF [9] on the EMS dataset. For each metric, results are first averaged within each degradation type and then equally averaged across the nine degradation types. ↑ indicates that a higher value is better, and ↓ indicates that a lower value is better. (**Bold**: better result).

Method	Fusion Metrics								Perceptual Metrics			
	EN↑	SD↑	SF↑	MI↑	SCD↑	VIF↑	$Q^{AB/F}$ ↑	SSIM_{sum} ↑	NIQE↓	MUSIQ↑	BRISQUE↓	CLIP-IQA↑
Text-IF [9]	7.287	48.372	15.899	**14.481**	1.447	0.784	0.566	**1.198**	3.405	46.999	26.929	**0.224**
IT-TextFusion (Ours)	**7.296**	**48.875**	**16.402**	14.474	**1.525**	**0.809**	**0.577**	1.195	**3.373**	**47.125**	**25.747**	0.212

NIQE, MUSIQ, BRISQUE, and CLIP-IQA assess no-reference perceptual quality and image naturalness from different statistical or learned perspectives and are not direct measures of multi-modal information preservation. In particular, although CLIP-IQA uses CLIP-derived representations, the adopted implementation receives only the fused image and therefore does not directly measure alignment between the fusion result and the input degradation-aware prompt. Their values may also be affected by the statistical characteristics of natural images and may not fully reflect the effectiveness of infrared-visible image fusion. Accordingly, these metrics are reported as complementary evaluation criteria and are not combined with the fusion-related metrics into a single aggregate score. Compared with Text-IF, our method obtains lower NIQE and BRISQUE values and a higher MUSIQ score, while Text-IF achieves a higher CLIP-IQA score. These results indicate that the proposed method improves several aspects of perceptual quality but also shows trade-offs across different no-reference evaluation criteria.

Across Tables 1–3, the differences among metrics reflect the properties they measure in fused images. MI measures the statistical dependence between the fused image and the source images, while $\mathrm{SSIM}_{\mathrm{sum}}$ assesses how well source-image structures are preserved. Cross-gating and residual refinement may change local luminance and contrast or suppress weak degradation-related responses. These changes can reduce MI or $\mathrm{SSIM}_{\mathrm{sum}}$ even when the fused image has clearer salient structures and fewer degradation artifacts. VIF measures multiscale visual information fidelity and may also penalize the removal of weak responses considered as source information, even if this improves perceptual quality. On the EMS dataset, MI and $\mathrm{SSIM}_{\mathrm{sum}}$ decrease by 0.048% and 0.250%, respectively, while SCD, VIF, and $Q^{AB/F}$ increase by 5.39%, 3.19%, and 1.94%. NIQE and BRISQUE decrease by 0.94% and 4.39%, respectively, while CLIP-IQA decreases by 5.36%. Overall, the proposed modules affect source-information preservation, structural fidelity, salient-detail enhancement, and perceptual naturalness differently, rather than improving every metric.

Qualitative Comparison: The qualitative results on the EMS dataset under six representative degradation scenarios are shown in Figure 6.

Compared with Text-IF [9], the proposed method produces more visually consistent fusion results under the illustrated degradation conditions.

For the Visible Blur, Low Light, Rain, and Overexposure degradation scenarios, visible images are affected by changes in contrast, brightness, and texture, making fusion more difficult. Text-IF can improve scene visibility to some extent, but its fusion results may still show insufficient brightness, saturation artifacts, or reduced color vividness in degraded regions, suggesting that some visible information remains unrecovered. In comparison, our method produces fused images with more balanced brightness, clearer structural details, and better color consistency, helping salient targets and scene content remain distinguishable under severe degradations. These examples show better visual robustness under several visible-side degradation conditions. For low-contrast infrared images, thermal targets are less distinguishable from the background, which makes fusion more challenging. In this case, our method produces results visually similar to those of Text-IF [9], with limited contrast enhancement in both methods. One possible explanation is that the degradation reduces the inherent contrast of the infrared modality, leaving weak thermal cues that are difficult to recover through fusion alone. Without explicit priors for infrared contrast

enhancement, both methods tend to preserve the original infrared intensity distribution, leading to similar fusion results under low-contrast infrared conditions.

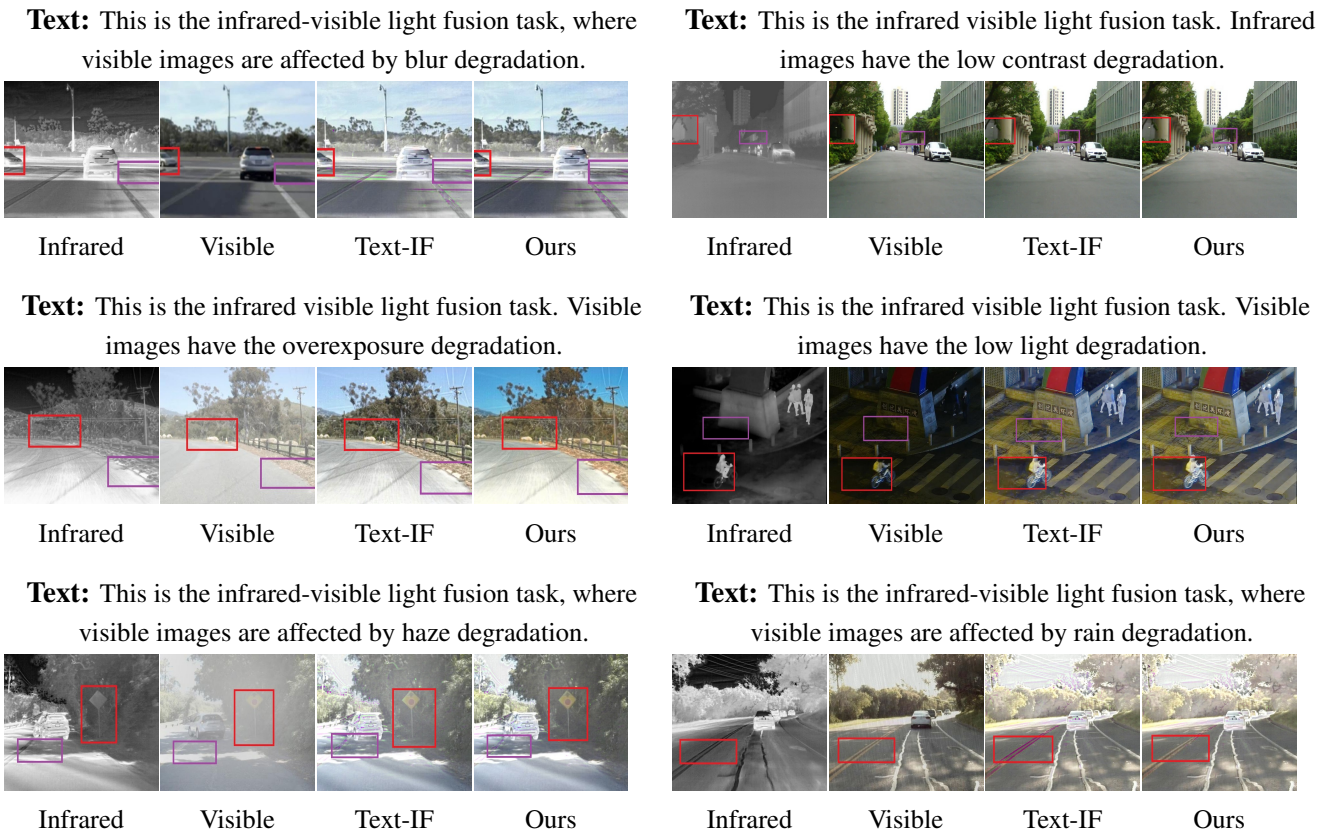

Figure 6. Qualitative comparison on the EMS dataset under six degradation types. Each group shows the infrared input, visible input, the result of Text-IF [9], and our result. The text above each group indicates the degradation-aware prompt used for fusion.

4.6. Ablation study

To examine the contribution of each component in our framework, we perform ablation experiments on the LLVIP dataset. Starting with a basic fusion backbone, we gradually add the key modules, including text engineering, ITIM, Cross-Gate Fusion, and TG-RRM. The results are presented in Table 4 and Figure 7.

Table 4. Ablation study of the proposed method on the LLVIP dataset. (**Bold** indicates the best performance)

TG-RRM	Text Engineering	Cross-Gate Fusion	ITIM	EN	SF	VIF	$Q^{AB/F}$	NIQE
×	×	×	×	7.443	16.153	**1.045**	0.751	3.600
×	×	×	✓	7.444	16.296	1.040	0.752	3.472
✓	×	×	✓	7.440	16.411	1.041	**0.765**	3.353
✓	×	✓	✓	7.439	16.394	1.043	0.758	3.363
✓	✓	✓	✓	**7.502**	**16.464**	1.037	**0.765**	**3.336**

The quantitative results show that the proposed components affect the evaluation metrics differently. Compared with the basic fusion backbone, adding ITIM slightly increases EN, SF, and $Q^{AB/F}$ and lowers NIQE, although VIF decreases slightly. ITIM therefore improves some metrics but not all of them. Adding TG-RRM further increases SF and $Q^{AB/F}$ and reduces NIQE, consistent with its role in structural refinement and artifact suppression. Cross-Gate Fusion slightly increases VIF from 1.041 to 1.043 relative to the preceding variant, but EN, SF, and $Q^{AB/F}$ decrease and NIQE increases. Its effect therefore varies across the reported metrics. Compared with the preceding variant, text engineering brings relatively small changes: EN, SF, and $Q^{AB/F}$ increase by 0.063, 0.070, and 0.007, respectively, while NIQE decreases by 0.027 and VIF decreases by 0.006. The complete model achieves the highest EN and SF, the lowest NIQE, and ties for the highest $Q^{AB/F}$ score, although its VIF remains below that of the baseline model. These differences show that the module combinations affect information preservation, structural fidelity, and perceptual naturalness in different ways, rather than improving every metric. The visual comparisons also show differences among components, while the complete model produces more balanced fusion results with clearer structures and better visual consistency. Overall, the ablation results suggest that the components provide complementary benefits on LLVIP, although their effects vary across metrics. Since the experiments cover only one dataset, the same trends cannot be assumed to hold across other datasets.

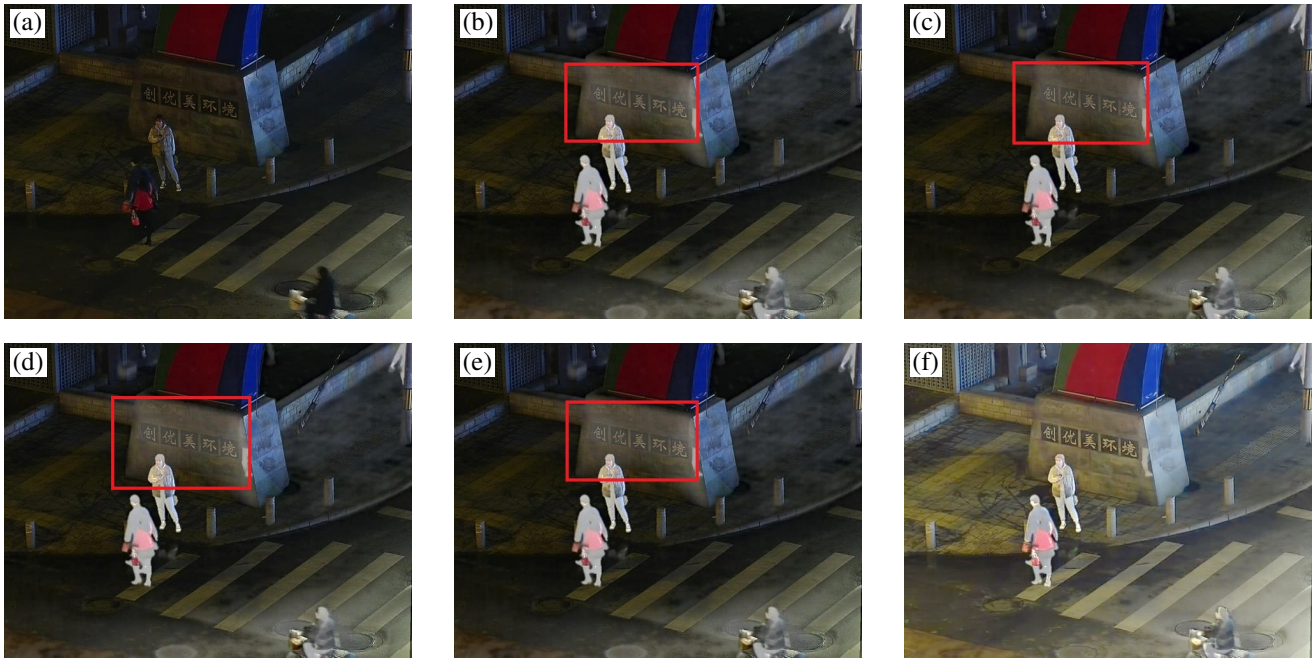

Figure 7. Qualitative comparison of different ablation variants on the LLVIP dataset. **(a)** Visible; **(b)** Original; **(c)** Only ITIM; **(d)** ITIM+TG-RRM; **(e)** w/o Text Engineering; **(f)** Complete Model.

5. Conclusion

In this paper, we study text-guided infrared-visible image fusion under different degradation conditions and propose a framework that incorporates global textual guidance into hierarchical feature fusion and refinement. While some existing text-guided fusion methods rely on relatively shallow semantic-visual interaction, our method maintains global text conditioning throughout hierarchical decoding and residual

feature refinement. Specifically, we use task- and degradation-aware text engineering to make prompts more informative, introduce stage-specific text-conditioned modulation to inject the global text representation into hierarchical decoder features, and design a Cross-Gate Fusion block that uses learnable cross-modal weighting for content-adaptive multi-modal feature integration. We also incorporate a Text-Guided Residual Refinement Module to refine details and maintain visual consistency under severe degradations. Experiments on standard fusion benchmarks and the multi-degradation EMS dataset show that the proposed method performs competitively, with improvements on several fusion and perceptual-quality metrics and trade-offs on others. The LLVIP ablation results show that the components affect different metrics and can provide complementary benefits. Future work will explore token-level language–vision interaction for more fine-grained semantic guidance and investigate ways to reduce computational cost and inference latency for practical deployment. We will also study automatic degradation recognition and prompt selection to reduce reliance on manually specified textual conditions and evaluate the model's sensitivity to mismatched or incorrect prompts.

Data availability statement

The EMS dataset used in this study is publicly accessible through its official repository (https://github.com/XunpengYi/EMS). The remaining datasets are accessible from their respective official sources.

Declaration of generative AI and AI-assisted technologies

During the preparation of this manuscript, the authors used generative AI tools only to improve language and readability. Specifically, the authors used ChatGPT for rewriting in substantial portions of the manuscript. The authors take full responsibility for the content of the manuscript.

Acknowledgments

The authors would like to thank the Chair of Robotics, Artificial Intelligence and Real-Time Systems, Technical University of Munich, for providing hardware support.

Authors' contribution

Conceptualization, methodology and writing—original draft, S.L.; data curation, investigation, software, formal analysis and writing—review and editing, P.Z.; formal analysis and visualization, T.J.; validation, R.B.; writing—review and editing, Z.J.; All authors have read and agreed to the published version of the manuscript.

Conflicts of interest

The authors declare no conflicts of interest.

References

[1] Cao H, Chen G, Zhao H, Jiang D, Zhang X, *et al.* SDPT: semantic-aware dimension-pooling transformer for image segmentation. *IEEE Trans. Intell. Transp. Syst.* 2024, 25(11):15934–15946.

[2] Cao H, Chen G, Xia J, Zhuang G, Knoll A. Fusion-based feature attention gate component for vehicle detection based on event camera. *IEEE Sens. J.* 2021, 21(21):24540–24548.

[3] Huang K, Shi B, Li X, Li X, Huang S, *et al.* Multi-modal sensor fusion for auto driving perception: a survey. *arXiv* 2022, arXiv:2202.02703.

[4] Zhang J, Liu H, Yang K, Hu X, Liu R, *et al.* CMX: cross-modal fusion for RGB-X semantic segmentation with transformers. *IEEE Trans. Intell. Transp. Syst.* 2023, 24(12):14679–14694.

[5] Cao H, Sun D, Song R, Xia Y, Li X, *et al.* Feature-aligned fisheye object detection network for autonomous driving. In *Proceedings of the 2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Hangzhou, China, October 19–25, 2025, pp. 7781–7787.

[6] Zhang J, Cao M, Xie W, Lei J, Li D, *et al.* E2E-MFD: towards end-to-end synchronous multimodal fusion detection. *Adv. Neural Inf. Process. Syst.* 2024, 37:52296–52322.

[7] Cui C, Ma Y, Cao X, Ye W, Zhou Y, *et al.* A survey on multimodal large language models for autonomous driving. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*, Waikoloa, USA, January 4–8, 2024, pp. 958–979.

[8] Cao H, Zhang Z, Xia Y, Li X, Xia J, *et al.* Embracing events and frames with hierarchical feature refinement network for object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, Milan, Italy, September 29–October 4, 2024, pp. 161–177.

[9] Yi X, Xu H, Zhang H, Tang L, Ma J. 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)*, Seattle, USA, June 17–21, 2024, pp. 27026–27035.

[10] Zhang J, Liu R, Shi H, Yang K, Reiß S, *et al.* Delivering arbitrary-modal semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Vancouver, Canada, June 18–22, 2023, pp. 1136–1147.

[11] Chen YT, Shi J, Ye Z, Mertz C, Ramanan D, *et al.* Multimodal object detection via probabilistic ensembling. In *Proceedings of the European Conference on Computer Vision (ECCV)*, Tel Aviv, Israel, October 23–27, 2022, pp. 139–158.

[12] Cao H, Chen G, Li Z, Hu Y, Knoll A. NeuroGrasp: multimodal neural network with euler region regression for neuromorphic vision-based grasp pose estimation. *IEEE Trans. Instrum. Meas.* 2022, 71:1–11.

[13] Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, *et al.* Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*, Online, July 18–24, 2021, pp. 8748–8763.

[14] Yang Z, Li Y, Tang X, Xie M. MGFusion: a multimodal large language model-guided information perception for infrared and visible image fusion. *Front. Neurorob.* 2024, 18:1521603.

[15] Cheng C, Xu T, Wu X, Li H, Li X, *et al.* TextFusion: unveiling the power of textual semantics for controllable image fusion. *Inf. Fusion* 2025, 117:102790.

[16] Li K, Li H, Cui M, Li J, Lv P, *et al.* TITFormer: combining textual modality and simulating infrared modality based on transformer for image enhancement. *IEEE Trans. Multimedia* 2025, 27:4725–4735.

[17] Xu H, Ma J, Jiang J, Guo X, Ling H. U2Fusion: a unified unsupervised image fusion network. *IEEE Trans. Pattern Anal. Mach. Intell.* 2022, 44(1):502–518.

[18] Zhao W, Xie S, Zhao F, He Y, Lu H. MetaFusion: infrared and visible image fusion via meta-feature embedding from object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Vancouver, Canada, June 18–22, 2023, pp. 13955–13965.

[19] Zhao Z, Bai H, Zhu Y, Zhang J, Xu S, *et al.* DDFM: denoising diffusion model for multi-modality image fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Paris, France, October 2–6, 2023, pp. 8082–8093.

[20] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, USA, June 7–12, 2015, pp. 3431–3440.

[21] Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, Munich, Germany, October 5–9, 2015, pp. 234–241.

[22] Wang H, Zhang H, Yi X, Xiang X, Fang L, *et al.* TeRF: text-driven and region-aware flexible visible and infrared image fusion. In *Proceedings of the 32nd ACM International Conference on Multimedia (ACM MM)*, Melbourne, Australia, October 28–November 1, 2024, pp. 935–944.

[23] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, *et al.* Attention is all you need. *Adv. Neural Inf. Process. Syst.* 2017, 30:5998–6008.

[24] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, *et al.* An image is worth 16 $\times$ 16 words: transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, Online, May 3–7, 2021.

[25] Cao H, Qu Z, Chen G, Li X, Thiele L, *et al.* GhostViT: expediting vision transformers via cheap operations. *IEEE Trans. Artif. Intell.* 2024, 5(6):2517–2525.

[26] Chen X, Zhang D, Zhao L, Yang C, Chen Z, *et al.* IGIANet: illumination guided implicit alignment network for infrared–visible uav detection. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, Singapore, January 20–27, 2026, pp. 3056–3064.

[27] Mustafa HT, Asad M, Jiang H, Zheng Z, Shamsolmoali P, *et al.* VAMF: variance-guided attention modulation framework for infrared and visible image fusion. *Expert Syst. Appl.* 2026, 327:132653.

[28] Huang X, Belongie S. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, October 22–29, 2017, pp. 1501–1510.

[29] Rao Y, Zhao W, Chen G, Tang Y, Zhu Z, *et al.* DenseCLIP: language-guided dense prediction with context-aware prompting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, USA, June 19–24, 2022, pp. 18082–18091.

[30] Rombach R, Blattmann A, Lorenz D, Esser P, Ommer B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, USA, June 19–24, 2022, pp. 10684–10695.

[31] Zamir SW, Arora A, Khan S, Hayat M, Khan FS, *et al.* Restormer: efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, USA, June 19–24, 2022, pp. 5728–5739.

[32] Sun S, Ren W, Gao X, Wang R, Cao X. Restoring images in adverse weather conditions via histogram transformer. In *European Conference on Computer Vision (ECCV)*, Milan, Italy, September 29–October 4, 2024, pp. 111–129.

[33] Cheng Y, Knoll A, Cao H. URNet: uncertainty-aware refinement network for event-based stereo depth estimation. *Visual Intell.* 2025, 3(1):18.

[34] Zhao H, Gallo O, Frosio I, Kautz J. Loss functions for image restoration with neural networks. *IEEE Trans. Comput. Imaging* 2017, 3(1):47–57.

[35] Tang L, Yuan J, Zhang H, Jiang X, Ma J. PIAFusion: a progressive infrared and visible image fusion network based on illumination awareness. *Inf. Fusion* 2022, 83:79–92.

[36] Jia X, Zhu C, Li M, Tang W, Zhou W. LLVIP: a visible-infrared paired dataset for low-light vision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, Online, October 11–17, 2021, pp. 3496–3504.

[37] Xu H, Yuan J, Ma J. MURF: mutually reinforcing multi-modal image registration and fusion. *IEEE Trans. Pattern Anal. Mach. Intell.* 2023, 45(10):12148–12166.

[38] Huang Z, Liu J, Fan X, Liu R, Zhong W, *et al.* ReCoNet: recurrent correction network for fast and efficient multi-modality image fusion. In *Proceedings of the European Conference on Computer Vision (ECCV)*, Tel Aviv, Israel, October 23–27, 2022, pp. 539–555.

[39] Liu J, Fan X, Huang Z, Wu G, Liu R, *et al.* 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 (CVPR)*, New Orleans, USA, June 19–24, 2022, pp. 5802–5811.

[40] Wang D, Liu J, Fan X, Liu R. Unsupervised misaligned infrared and visible image fusion via cross-modality image generation and registration. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI)*, Vienna, Austria, July 23–29, 2022, pp. 3508–3515.

[41] Roberts JW, van Aardt JA, Ahmed FB. Assessment of image fusion procedures using entropy, image quality, and multispectral classification. *J. Appl. Remote Sens.* 2008, 2(1):023522.

[42] Rao YJ. In-fibre bragg grating sensors. *Meas. Sci. Technol.* 1997, 8(4):355–375.

[43] Eskicioglu AM, Fisher PS. Image quality measures and their performance. *IEEE Trans. Commun.* 1995, 43(12):2959–2965.

[44] Qu G, Zhang D, Yan P. Information measure for performance of image fusion. *Electron. Lett.* 2002, 38(7):313–315.

[45] Aslantas V, Bendes E. A new image quality metric for image fusion: the sum of the correlations of differences. *AEU Int. J. Electron. Commun.* 2015, 69(12):1890–1896.

[46] Han Y, Cai Y, Cao Y, Xu X. A new image fusion performance metric based on visual information fidelity. *Inf. Fusion* 2013, 14(2):127–135.

[47] Wang Z, Bovik AC, Sheikh HR, Simoncelli EP. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.* 2004, 13(4):600–612.

[48] Wang J, Chan KCK, Loy CC. Exploring CLIP for assessing the look and feel of images. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, Washington, USA, February 7–14, 2023, pp. 2555–2563.

[49] Mittal A, Soundararajan R, Bovik AC. Making a "completely blind" image quality analyzer. *IEEE Signal Process Lett.* 2013, 20(3):209–212.

[50] Ke J, Wang Q, Wang Y, Milanfar P, Yang F. MUSIQ: multi-Scale image quality transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Online, October 11–17, 2021, pp. 5148–5157.

[51] Mittal A, Moorthy AK, Bovik AC. No-reference image quality assessment in the spatial domain. *IEEE Trans. Image Process.* 2012, 21(12):4695–4708.

[52] Wu W, Weng J, Zhang P, Wang X, Yang W, *et al.* URetinex-Net: retinex-based deep unfolding network for low-light image enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, USA, June 19–24, 2022, pp. 5901–5910.

[53] Li B, Liu X, Hu P, Wu Z, Lv J, *et al.* All-in-one image restoration for unknown corruption. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, USA, June 18–24, 2022, pp. 17452–17462.

[54] Chen H, Gu J, Liu Y, Magid SA, Dong C, *et al.* Masked image training for generalizable deep image denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Vancouver, Canada, June 17–24, 2023, pp. 1692–1703.

[55] Afifi M, Derpanis KG, Ommer B, Brown MS. Learning multi-scale photo exposure correction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Online, June 19–25, 2021, pp. 9157–9167.