Text-IF: Leveraging Semantic Text Guidance for Degradation-Aware and Interactive Image Fusion

Xunpeng Yi, Han Xu, Hao Zhang, Linfeng Tang, Jiayi Ma*
Electronic Information School, Wuhan University, Wuhan 430072, China
{yixunpeng, xu_han}@whu.edu.cn, {zhpersonalbox, linfeng0419, jyma2010}@gmail.com

Abstract

Image fusion aims to combine information from different source images to create a comprehensively representative image. Existing fusion methods are typically helpless in dealing with degradations in low-quality source images and non-interactive to multiple subjective and objective needs. To solve them, we introduce a novel approach that leverages semantic text guidance image fusion model for degradation-aware and interactive image fusion task, termed as Text-IF. It innovatively extends the classical image fusion to the text guided image fusion along with the ability to harmoniously address the degradation and interaction issues during fusion. Through the text semantic encoder and semantic interaction fusion decoder, Text-IF is accessible to the all-in-one infrared and visible image degradation-aware processing and the interactive flexible fusion outcomes. In this way, Text-IF achieves not only multi-modal image fusion, but also multi-modal information fusion. Extensive experiments prove that our proposed text guided image fusion strategy has obvious advantages over SOTA methods in the image fusion performance and degradation treatment. The code is available at https://github.com/XunpengYi/Text-IF.

1. Introduction

Image fusion is a prominent field within the domain of digital image processing [15, 27, 35]. Single-modal images can only capture partial representation of the scene. Multi-modal images allow for the effective acquisition of more comprehensive representation. As an important representative, visible images provide the reflectance-based visual information, akin to human vision. Infrared images provide thermal radiation-based information, more valuable for detecting thermal targets and observing nighttime activities. The infrared and visible image fusion focuses on fusing the complementary information of infrared and visible images, yielding high-quality fusion images [18–20, 28, 38, 39, 43].

*Corresponding author

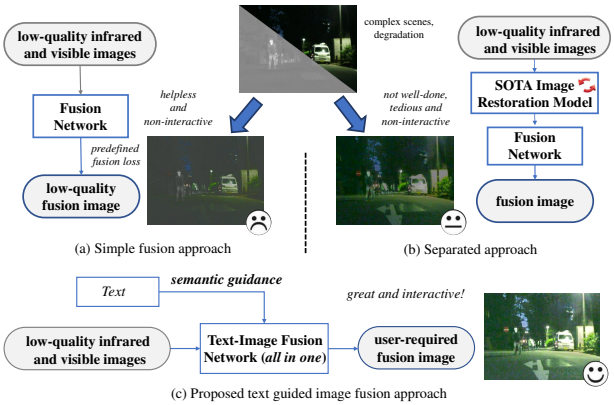

Figure 1. Fusion approaches for complex scenes with degradations. (a) simple fusion approach: treating image fusion with predefined fusion loss and not applicable to complex scenes with degradations. (b) separated approach: requiring frequent restoration methods switching according to the type of degradations, which is troublesome and not well-done. (c) proposed text guided image fusion approach: achieving interactive and high-quality fusion image without tedious replacement of models.

Limited by the conditions of environments, the originally acquired infrared and visible images may suffer from degradations and show low fusion image quality. The visible images are susceptible to degradation issues, *e.g.*, low light, over exposure, *etc*. The infrared images are inevitably affected by noise (including thermal, electronic, and environmental noise), diminished contrast, and other associated effects. Current fusion methods lack the capability to adaptively solve the degradations, leading to the low-quality fusion image. Furthermore, relying on manual pre-processing to enhance the image has the problems of flexibility and efficiency [29]. Therefore, it is of practical interest to study a model that harmonises degradation-aware processing and interactive fusion.

Designing a model for individualized degradation to achieve image enhancement and fusion is feasible. However, most of image fusion tasks need to be carried out in various complex conditions around the clock. As shown in

Fig. 1, it means that multiple image restoration models are needed to match the requirements, which requires frequent models switching and brings great consumption and trouble. In addition, the separation approach has the problem to achieve harmony between the enhancement and fusion, resulting in unsatisfactory overall performance.

In addition, the real-world image fusion is complex, flexible and task-oriented. The requirements of the image fusion may change according to the subjective needs of users and objective application tasks. In all scenarios, if the method is non-interactive and produces a relatively fixed fusion result, it usually falls short for various and flexible requirements of the users.

As an important way of human and machine interaction, text is widely used in the model of specifying requirements. Recent research in large-scale visual language has achieved amazing results in image generation [13, 16, 26, 30], demonstrating the potential of this paradigm. The interaction between semantic text and image processing procedures can achieve the goal of customized image processing. In addition, PromptIR [24] proposed learnable visual prompts and implemented various degradation removals, but not realizes text guided and lacks the design for multi-modal degradations and fusion. Therefore, it is of great significance to implement the degradation-aware processing and user interactivity in image fusion by text.

To this end, we propose a model that leverages the semantic text guidance for degradation-aware and interactive image fusion, termed as Text-IF. It integrates the text and image fusion to meet the needs of harmonious degradation-aware processing and interactivity fusion. Especially, it allows the text to provide the flexible semantic guidance to deal with various degradations, which is a type of multi-modal information fusion. In general, Text-IF contains the image pipeline, and text interaction guidance architecture, including the text semantic encoder and the semantic interaction guidance module. In the image fusion pipeline, we meticulously design the Transformer-based image extraction module and the cross fusion layer for high-quality fusion. In the text semantic encoder, we aggregate the text semantic extraction capabilities of powerful pre-trained vision-language models. Through the semantic interaction guidance module, the semantic features of text and image fusion features are coupled together to achieve the goal of text guided image fusion. It solves the problem that the existing image fusion methods are difficult to adapt to the fusion of complex scenes with degradations, and can only output relatively fixed results without interactivity. It provides a feasible direction for the subsequent research of text guided image fusion tasks.

Overall, our contributions can be summarized as follows:

- To adapt complex degradation conditions, we address the integrated problem of image fusion and degradation-aware processing. It breaks through the limitation of quality improvement in image fusion.
- We introduce a semantic interaction guidance module to fuse the information of text and images. The proposed method achieves not only multi-modal image fusion, but also multi-modal information fusion.
- The proposed method ultimately increases freedom of customized fusion results. It provides the interactive fusion and can generate more flexible, high-quality and user-required results without prior expertise or predefined rules.

2. Related Work

General Image Fusion Methods. General Image fusion methods have made significant advancements with the advent of deep learning. During the early phase, fusion strategies based on pre-trained autoencoders were extensively employed. CSR [17] adopts convolutional sparse representation for image fusion, extracts multi-layer features, and utilizes these features to generate fusion images. To eliminate the need for laborious manual design, an end-to-end fusion structure based on CNN is proposed, rendering the fusion process more flexible and straightforward. U2Fusion [34] adopts the densely connected network to generate the fusion image conditioned on source images. And the weight block is used to obtain two data-driven weights, which are used as the retention of features in different source images to measure the quality and information of the image. Furthermore, it combines continuous learning and other technologies to achieve multi-task fusion. It is the first all-in-one image fusion method. In addition, in recent years, the image fusion with high-level visual task has made great progress [14]. Recently, the diffusion-based image fusion came into people's view. DDFM [42] utilizes the bayesian theory, score-matching and pretrained diffusion model to get the awesome results.

Text-Image Models. With advancements in Transformers and representation learning, coupled with the support of large datasets, multi-modal text guided image models have achieved success. CLIP [25] is built upon two neural network-based encoders that use a contrastive loss for aligning pairs of image and text. Owing to extensive data and unsupervised training, it has the powerful zero-shot recognition and the robust text, image feature extraction capabilities. Numerous methods for text-driven image generation and processing have been proposed with the support of the CLIP model. Style-CLIP [23] designs a text guided interface to the StyleGAN [9], allowing the alteration of real images using text prompts. In addition to GAN models, the diffusion model with text conditions also catch a lot of attentions. DiffusionCLIP [11] proposes the diffusion model with the CLIP for text-driven image processing. Besides, stable diffusion [26] combines the diffusion model with the

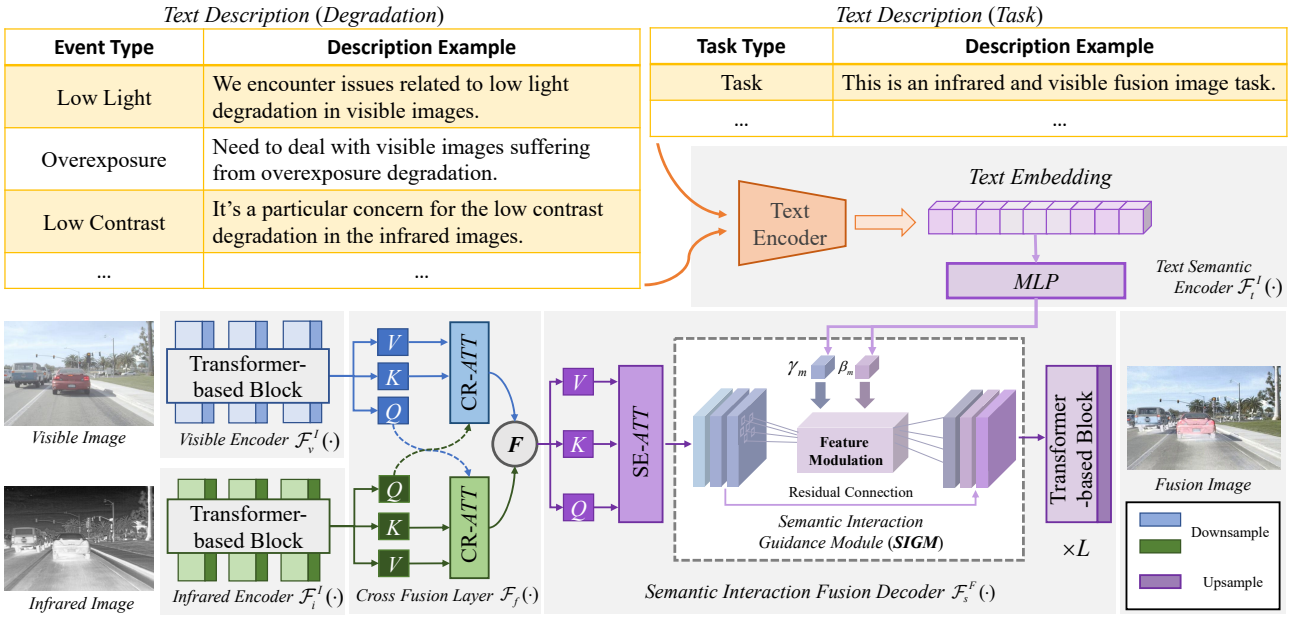

Figure 2. The workflow of Text-IF. It contains two important parts, which are the image fusion pipeline and the text semantic feature encoder. Text semantic features are used to guide image fusion through the Semantic Interaction Guidance Module (**SIGM**).

text encoder and the attention mechanism to achieve the effect of text controlling image generation. Through text guidance, it can customize the effect of image generation, image processing and other tasks, and realize interactive multi-modal fusion control.

Existing image fusion methods are helpless in complex scene facing degradations. It is troublesome and not well-done even if equipped with the SOTA image restoration models. Furthermore, it is difficult to achieve interactive high-quality image fusion for users without professional knowledge. Thus, it is necessary to innovatively introduce the text guided image fusion framework for simply using.

3. The Proposed Method

This section describes the workflow of Text-IF, as shown in Fig. 2. We introduce from the perspective of the image fusion pipeline and the text interaction guidance architecture, including the text semantic encoder and the semantic interaction guidance module.

3.1. Problem Formulation

General image fusion methods formulate the image fusion task into taking two source images (*e.g.*, I_{vis}, I_{ir}) as the input and a fusion network (*e.g.*, θ_n) to obtain an immobilized image fusion result. The network is designed to learn the mapping predefined fusion function $\mathcal{F}_{if}$ corresponding to the fusion task. In simple terms, it can be described as:

$$I^f = \mathcal{F}_{if}(I_{vis}, I_{ir}; \theta_n). \tag{1}$$

It means that the fusion network tends to learn a relatively fixed fusion strategy. Moreover, in complex environments, such as the source images suffering from degradation, this kind of task paradigm is helpless. We study leveraging the text for breaking the traditional single fusion result along with difficulty in the quality improvement on degradations, and explore the novel text guided image fusion paradigm. Due to the introduction of text semantics, this fusion task is rewritten as:

$$I^f = \mathcal{F}_{s-if}(I_{vis}, I_{ir}, T_{text}; \theta_{n-s}). \tag{2}$$

The origin mapping fusion function $\mathcal{F}_{if}$ is extended to $\mathcal{F}_{s-if}$ with text semantic information guidance. Through the interaction of text semantics T_{text}, the image fusion network θ_{n-s} can achieve a more customized and flexible fusion effect according to the text given by the users. At the same time, it can also restore and fuse images freely in the face of various source image degradation.

3.2. Image Fusion Pipeline

Image Encoder. The image encoder takes the source visible and infrared images as the input, respectively. Considering the spatial and deep information extraction, to obtain a comprehensive and accurate representation, we adopted Transformer/Restormer[37]-based blocks as the base feature extractor. In simple terms, it can be stated as follows:

$$F_{vis} = \mathcal{F}_v^I(I_{vis}), F_{ir} = \mathcal{F}_i^I(I_{ir}), \tag{3}$$

where $I_{vis} \in \mathbb{R}^{H\times W\times 3}$ and $I_{ir} \in \mathbb{R}^{H\times W\times 1}$ represent the visible and infrared images. H, W denote the height and

width of the image. $\mathcal{F}_v^I$ and $\mathcal{F}_i^I$ are the visible image and infrared image encoder, respectively.

Cross Fusion Layer. The cross fusion layer aims to integrate the feature information from different modalities. In order to comprehensively integrate the features across all dimensions, the cross-attention (CR-ATT) is firstly used to interact the features of different modalities. Specifically, it can be expressed as follows:

$$\{Q_v, K_v, V_v\} = \mathcal{F}_{qkv}^v(F_{vis}), \quad \{Q_i, K_i, V_i\} = \mathcal{F}_{qkv}^i(F_{ir}), \quad (4)$$

where F_{vis}, F_{ir} denote features from the visible encoder and infrared encoder. Subsequently, we exchange the queries Q of two modalities for spatial interaction:

$$F_f^i = softmax(\frac{Q_v K_i}{d_k})V_i, \quad F_f^v = softmax(\frac{Q_i K_v}{d_k})V_v, \quad (5)$$

where d_k is the scaling factor. Finally, we concatenate the results obtained by the cross-attention calculation through $F_f^0 = Concat(F_f^i, F_f^v)$ to get the fusion features.

Semantic Interaction Fusion Decoder. The features of the cross fusion layer output are firstly enhanced by self-attention (SE-ATT), *i.e.*, $\hat{F}_f^0 = softmax(Q_f K_f / d_k)V_f$. Q_f, K_f and V_f are the Q, K, and V of F_f^0. Subsequently, it is interactively guided by semantic text features.

The semantic interaction fusion decoder is designed to interact text semantic features $F_{text} \in \mathbb{R}^{N \times L}$ and image fusion features F_f. Specifically, it is constructed by the Transformer-based decoder block and Semantic Interaction Guidance Module (SIGM) which will be introduced in Sec. 3.3. The fusion decoder block and SIGM are tightly coupled together in a multi-stage cascade to achieve the effect of dense regulation and guidance. Briefly, the semantic interaction fusion decoder can be described as:

$$F_f^{k+1} = \{\mathcal{F}_f^D(\mathcal{L}_f^s(F_f^k, F_{text}))\}_r, \quad (6)$$

where F_f^k denotes the image fusion feature at the k-th block stage. $\{\cdot\}_r$ represents the multilevel repetition. $\mathcal{F}_f^D$ and $\mathcal{L}_f^s$ denote the Transformer-based block and SIGM. Note that the upsampling is required between the levels of decoders to correspond to the downsampling at the encoder.

3.3. Text Interaction Guidance Architecture

The preset image fusion pipeline can effectively obtain the corresponding fusion features F_f. And the Text Interaction Guidance Architecture is the key part to couple the text semantic information and image fusion.

Text Semantic Encoder. Given a text T_{text} that provides the corresponding semantic feature to guide the image fusion network to obtain the specified fusion result (*e.g.*, specify the task type and the degradation type), the text semantic encoder of the text interaction guidance architecture should transfer it into the text embedding. As a large pre-trained visual language model, CLIP has a good effect on text feature extraction. We tend to freeze the good text encoder from the CLIP to maintain good linguistic consistency. With $\{\cdot\}_e$ denoting the frozen weights, this process can be expressed as:

$$F_{text} = \{\mathcal{F}_t^I\}_e(T_{text}), \quad (7)$$

where $F_{text} \in \mathbb{R}^{N \times L}$ denotes the text semantic feature. In different but semantically similar texts, the extracted features should be close in the reduced Euclidean space.

Furthermore, we design the MLP Φ_m^i to mine this connection and further map the text semantic information and the semantic parameters. Therefore, it can be obtained:

$$\gamma_m = \Phi_m^I(F_{text}), \ \beta_m = \Phi_m^{II}(F_{text}), \quad (8)$$

where Φ_m^I and Φ_m^{II} are the chunk operations of Φ_m to form the semantic parameters.

Semantic Interaction Guidance Module (SIGM). In the semantic interaction guidance module, semantic parameters interact through feature modulation and fusion features F_f^i, so as to obtain the effect of guidance. The feature modulation consists of scale scaling and bias control, which adjust the features from two perspectives, respectively. In particular, a residual connection is used to reduce the difficulty of network fitting. For simplicity, it can be described as:

$$\hat{F}_f^i = (1 + \gamma_m) \odot F_f^i + \beta_m, \quad (9)$$

where $\odot$ denotes Hadamard product. F_f^i denotes the fusion feature. $\hat{F}_f^i$ is that with textual semantic information.

3.4. Loss Functions

The loss function largely determines the type of extracted source information and the proportion relationship between source information. From the perspective of text guidance, we not only hope to solve various degradation problems through text freedom. It is also expected that the text can autonomously choose the optimal loss corresponding to the fusion task according to the needs of the users. Therefore, in the text guided image fusion task, the construction of loss function is a relation of open-set multi-point mapping.

The fusion-related losses include the intensity loss, structural similarity (*SSIM*) loss [40], maximum gradient loss, and color consistency loss. Considering degradations, we adopt manually obtained high-quality visible image I_{vis}^g and infrared image I_{ir}^g as the constraints in the loss.

Intensity Loss. To highlight the salient objects in infrared and visible images, the intensity values of results are maximized to ensure the target saliency. It is defined as:

$$L_{int} = \frac{1}{HW}\|I_f - max(I_{vis}^g, I_{ir}^g)\|_1. \quad (10)$$

Structural Similarity Loss. The structural similarity loss measures the similarity between the fusion image and

the source images, so that the fusion image is similar to the source images in structure. It is expressed as:

$$L_{SSIM}(t) = \left(1 - SSIM(I_f, I^g_{vis})\right) + \delta_{ir}(t)\left(1 - SSIM(I_f, I^g_{ir})\right), \tag{11}$$

where $\delta_{ir}(t)$ denotes the ratio of infrared structural similarity loss which is a function of the text semantic.

Maximum Gradient Loss. This loss preserves the maximum edges in two source images. Then, a clearer texture representation can be obtained. It can be expressed as:

$$L_{grad} = \frac{1}{HW}\|\nabla I_f - max(\nabla I^g_{vis}, \nabla I^g_{ir})\|_1. \tag{12}$$

Color Consistency Loss. It keeps the fusion image and the visible image with consistent colors. We transpose the image to YCbCr space and constrain it with the Euclidean distance of Cb and Cr channels. It can be expressed as:

$$L_{color} = \frac{1}{HW}\|\mathcal{F}_{CbCr}(I_f) - \mathcal{F}_{CbCr}(I^g_{vis})\|_1, \tag{13}$$

where $\mathcal{F}_{CbCr}$ denotes the transfer function of RGB to CbCr.

Total Loss. The overall loss function is a combination of fusion-related losses and is regulated by semantic information. Simply, it can be expressed as:

$$\begin{aligned} L_{total} = &\alpha_{int}(t)L_{int} + \alpha_{SSIM}(t)L_{SSIM}(t) + \\ &\alpha_{grad}(t)L_{grad} + \alpha_{color}(t)L_{color}, \end{aligned} \tag{14}$$

where $\alpha_{int}(t)$, $\alpha_{SSIM}(t)$, $\alpha_{grad}(t)$, and $\alpha_{color}(t)$ are semantically regulated hyper-parameters related to the task t. The trade-off of the fusion result plays a large role.

4. Experiments

In this section, we first introduce the implementation details and relevant configuration. Then, the effectiveness and superiority of the proposed method are evaluated through qualitative and quantitative comparisons. In particular, the specific results of text guided image fusion are analyzed. Finally, ablation experiments are performed.

4.1. Implementation Details and Datasets

Implementation Details. The proposed Text-IF is trained with the text guided image fusion data. The learning rate is 0.0001 with the AdamW optimizer. And the batch size is 16. The source images are cropped to 96×96. The set of hyper-parameters $\{\alpha_{int}(t), \alpha_{SSIM}(t), \alpha_{grad}(t), \alpha_{color}(t)\}$ is essentially a discrete complex map associated with the semantic text. See the additional material for details. All the experiments are conducted on the NVIDIA GeForce RTX 3090 GPU with PyTorch framework.

Datasets. To verify generalization, the commonly used infrared and visible image fusion datasets are MSRS [29], MFNet [5], RoadScene [34] and LLVIP [8]. These original datasets come with degradations in different situations, such as low light, overexposure, *etc.*, in visible images, and low contrast, noise, *etc.*, in infrared images. We select images where the scene is different and use manual restoration to obtain the high-quality source image, and add the corresponding hundreds of description instructions to ensure that users can input text freely for interaction. Totally we use 3618 image pairs for training, and 1135 for testing.

Metric. Metrics include the sum of the correlations of differences (SCD) [2], standard deviation (SD), information entropy (EN) [18], visual information fidelity (VIF) [6], quality of gradient-based fusion ($Q^{AB/F}$) [18], CLIP-IQA [32], NIQE [22], MUSIQ [10], BRISQUE [21], and spatial frequency (SF) [4]. Higher values of SCD, SD, EN, VIF, $Q^{AB/F}$, CLIP-IQA, MUSIQ, and SF indicate higher quality of the fusion image. Besides, the lower values of NIQE, and BRISQUE indicate the higher quality.

SOTA Competitors. We compare the proposed method with several state-of-the-art methods on multiple datasets. The methods for comparison include UMF-CMGR [31], TarDAL [14], ReCoNet [7], MURF [36], U2Fusion [34], MetaFusion [41], and DDFM [42].

4.2. Comparison without Text Guidance

Existing image fusion methods do not have semantic guidance. For comparison fairness, we first merely compare the fusion performances where no semantic guidance is provided. At this point, Text-IF uses default text. It means that no additional semantic information is introduced.

Qualitative Comparison. The results on three datasets are shown in Fig. 3. Text-IF shows three distinctive advantages thanks to the Transformer-based pipeline with high expressive power and the implicit embedding image restoration prior. First, our results can highlight the thermal targets. As shown in the first three groups of results, the pixel intensity of thermal targets in our results are the highest. It indicates that the thermal targets in our results are the most prominent. Second, our results exhibit more appropriate brightness and provide more details. In the second and third groups, most regions of our results show higher pixel intensity than the results of competitors. In this case, more scene content can be presented clearly. Last, our results can present more vibrant and natural colors. As shown in the last example, in our result, the colors of cars and trees are more similar to those of visible images. By reducing the interference of infrared images on the color information in visible images, our fusion results are more conducive to visual perception from the perspective of colors.

Quantitative Comparison. The quantitative results tested with five metrics on three datasets are reported in Tab. 1. On the MSRS and LLVIP datasets, our method performs best on all the five metrics, especially showing significant advantages in SCD and VIF. On the RoadScene dataset, our method also performs optimally on three met-

Figure 3. Qualitative comparison of our Text-IF without text guidance (without additional semantic information) and existing image fusion methods. From top to bottom: data from MSRS, two groups of data from LLVIP, and data from RoadScene datasets, respectively.

Table 1. Quantitative comparison of our Text-IF without text guidance (without introducing additional semantic information) and existing image fusion methods on the MSRS, LLVIP, and RoadScene datasets (**Bold**: optimal 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	$Q^{AB/F}$
UMF-CMGR	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
TarDAL	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
ReCoNet	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
MURF	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
U2Fusion	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	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	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 (ours)	**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**

rics. The results on EN, VIF, and $Q^{AB/F}$ reflect that even without text guidance, our method can also generate fusion results with the most information, cause the least distortion between the fusion and source images, and transfer the most edges into the fusion image. The optimal or comparable results on SCD and SD reflect that our results show little fusion distortion and high contrast (good visual effect). From the perspective of metrics, the superiority on multiple metrics indicates the comprehensiveness of the proposed method in terms of fusion performance. From the perspective of datasets, the superiority of the proposed method on multiple datasets reflects its generalization in multiple data distributions and multiple types of scenarios.

4.3. Comparison with Text Guidance

In real scenarios, source images may usually suffer from various degradations, *e.g.*, poor illumination, noise, and low contrast. Existing image fusion methods cannot handle these degradations, resulting in unsatisfactory fusion results while our method can handle them through simple text guidance. Thus, for fairness, we combine existing image fusion methods with image restoration methods for comparison. SOTA image restoration models for different degradations include URetinex [33] for low-light image enhancement, AirNet [12] for contrast enhancement, GDID [3] for denoising, and LMPEC [1] for overexposure correction. It is also worth noting that our approach uses the same model parameters in all scenarios, *i.e.*, for all degradations.

Qualitative Comparison. The qualitative results of Text-IF and the results of combining SOTA image restoration and image fusion competitors on degraded source images are shown in Fig. 4. In general, unlike existing methods that require manual priors for adding restoration preprocessing to fusion, Text-IF only needs to provide a simple request/description of scene and can then handle degraded source images. It avoids the tedious task of finding and switching between different restoration methods in the process of combating degradation.

Then, we compare the fusion results in various degradation scenarios in detail. First, in the first two examples, the

Figure 4. Comparison of our Text-IF with semantic text guidance and the combination of existing image restoration and fusion methods on degraded source images. The semantic text is reported above each group of images. Degradations from top to bottom: low-light visible (MSRS), low-light visible (LLVIP), low-contrast infrared (MFNet), noised infrared (DN-MSRS), over-exposed visible (RoadScene).

visible images suffer from low illumination. The competitors can brighten the visible images with URetinex to some extent. However, after fusion, the low pixel intensity of infrared images still reduces the brightness of their results, and also reduces the color saturation. In comparison, our method results in more suitable brightness and brighter colors. In the third and last examples, the infrared image is of low contrast or the visible image is overexposed. In these conditions, our method can expand the dynamic range of fusion results and obtain fusion results with higher contrast and ensure the correctness of its color information at the same time. Then, the results can exhibit more clear details. In the fourth example, the infrared image suffers from obvious noise. GDID fails to remove all the noise, resulting in residual noise in fusion results. By comparison, our result shows less noise pollution, presenting higher image quality. Moreover, the prominence of thermal targets in our result is also advantageous.

Quantitative Comparison. The results on datasets in different types of degradations are reported in Tab. 2. Text-IF still achieves the overall optimal performance on all the metrics of MSRS, LLVIP, MFNet, DN-MSRS, and RoadScene datasets. The results on SD, EN and SF indicate that our method can effectively transfer the information in the fusion image. The results on CLIP-IQA, NIQE, MUSIQ and BRISQUE show that our method can produce high quality fusion results facing the degradations.

4.4. Performance on High-level Task

To verify the performance of the fusion performance in downstream high-level vision tasks, we conduct object detection experiments on the fusion results on LLVIP dataset. We adopted YOLOv8[1] as the object detection backbone and fine-tuned it on the infrared visible light source images of LLVIP. Qualitative and quantitative experimental results are shown in Fig. 5 and Tab. 3.

Comparison with SOTA Competitors. In terms of qualitative comparison, our proposed method Text-IF de-

[1] https://github.com/ultralytics/ultralytics

Table 2. Quantitative comparison of our Text-IF with text guidance and the combination of existing image restoration (*eir.*) and fusion methods on source images with various types of degradations (MSRS and LLVIP datasets: low-light visible images; MFNet: low-contrast infrared images; DN-MSRS: noised infrared images; RoadScene: over-exposed visible images). (**Bold**: optimal performance)

Methods	MSRS Dataset			LLVIP Dataset			MFNet Dataset			DN-MSRS Dataset			RoadScene Dataset		
	CLIP-IQA	EN	NIQE	EN	NIQE	MUSIQ	SD	EN	MUSIQ	SD	EN	NIQE	SF	NIQE	BRISQUE
eir.+UMF-CMGR	0.101	6.316	3.738	7.087	3.891	47.543	23.684	5.414	34.113	21.047	5.645	6.279	11.047	3.792	32.485
eir.+TarDAL	0.082	5.855	4.750	7.042	3.659	41.735	33.454	6.142	25.120	23.316	5.399	7.353	11.789	3.667	32.436
eir.+ReCoNet	0.117	7.216	5.769	7.109	4.695	44.187	41.654	5.161	29.299	41.525	4.463	8.631	10.312	4.785	37.775
eir.+MURF	0.111	5.872	4.199	6.757	4.177	**50.589**	23.741	5.601	35.626	20.456	5.280	6.549	15.605	3.779	30.594
eir.+U2Fusion	0.127	6.724	3.997	7.439	3.969	48.481	33.940	5.740	34.255	28.812	4.609	7.185	18.006	4.215	34.577
eir.+MetaFusion	0.106	**7.302**	3.584	**7.495**	3.722	49.620	42.026	6.665	34.762	39.956	6.398	4.337	**26.653**	3.473	29.500
eir.+DDFM	0.094	6.723	**3.465**	7.150	5.184	35.933	30.465	6.480	26.902	27.362	6.120	4.644	10.493	3.717	32.334
Text-IF (ours)	**0.132**	7.172	3.708	7.391	**3.502**	48.625	**43.933**	**6.683**	**35.650**	**43.448**	**6.669**	**4.012**	17.766	**3.342**	**29.021**

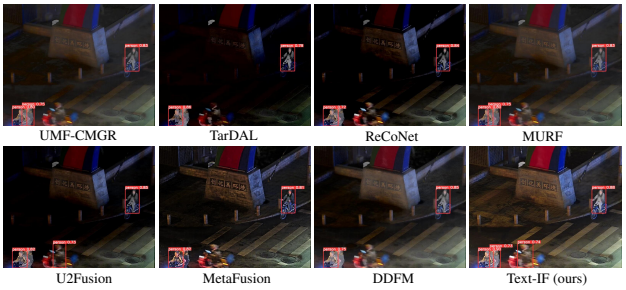

Figure 5. Qualitative comparison of object detection performance on LLVIP (without introducing additional semantic information).

Table 3. Quantitative comparison of object detection on the LLVIP dataset. Text-IF uses the default Text (without introducing additional semantic information). (**Bold**: optimal performance)

Method	UMF-CMGR	TarDAL	ReCoNet	MURF
mAP@0.50	0.925	0.922	0.916	0.926
mAP@0.75	0.659	0.646	0.617	0.675
mAP@0.50:0.95	0.599	0.582	0.568	0.599
Method	U2Fusion	MetaFusion	DDFM	**Text-IF (ours)**
mAP@0.50	0.921	0.916	0.921	**0.941**
mAP@0.75	0.655	**0.690**	0.655	0.676
mAP@0.50:0.95	0.591	0.590	0.592	**0.602**

tects all the objects in the scene, while other methods have the miss detection. In terms of the quantitative comparison, Text-IF obtains the best detection performance.

4.5. Ablation Experiment

To verify the effectiveness of the proposed method, we conduct a series of ablation experiments on LLVIP dataset. It mainly includes the ablation of image fusion loss, including the intensity loss, the structural similarity (*SSIM*) loss, the maximum gradient loss, and the color consistency loss. As shown in Fig. 6 and Tab. 4, qualitative and quantitative results are presented.

In terms of qualitative results, the intensity loss preserves the significant thermal radiation of targets. The color loss keeps consistent color. The maximum gradient loss provides clear texture information. In terms of the quantitative

Figure 6. Qualitative comparison of ablation experiment of the loss function on LLVIP.

Table 4. Quantitative comparison of the ablation experiment of the loss function on LLVIP. (**Bold**: optimal performance)

L_{int}	L_{SSIM}	L_{grad}	L_{color}	SCD	SD	EN	VIF	$Q^{AB/F}$
✓				1.389	46.147	7.205	0.794	0.552
	✓	✓	✓	1.481	42.530	7.063	**1.020**	0.674
✓	✓		✓	1.485	47.559	7.182	0.831	0.594
✓	✓	✓		1.547	46.798	7.274	0.987	**0.688**
✓	✓	✓	✓	**1.591**	**48.834**	**7.325**	1.011	0.616

results, each loss has a corresponding contribution to the final quantitative evaluation result. Our method achieves the best qualitative and quantitative evaluation among all ablation methods, which proves the effectiveness of the method.

5. Conclusion

In this paper, we extend the image fusion task and propose a novel text guided image fusion framework to address the problem that existing methods have difficulty in solving the complex scenes fusion with the degradations and getting the user-required fusion image with interactivity. Through the image fusion pipeline, the text semantic feature extraction and the semantic interaction guidance module, the goal of image fusion guided by text semantics is realized. Extensive experimental results demonstrate the obvious advantages of the proposed method in both the fusion performance and degradations treatment. It makes it possible to generate the corresponding fusion image according to the interactive user text input freely, which plays a promotive role in practice and subsequent theoretical research.

Acknowledgments

This work was supported by NSFC (62276192).

References

[1] Mahmoud Afifi, Konstantinos G Derpanis, Bjorn Ommer, and Michael S Brown. Learning multi-scale photo exposure correction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9157–9167, 2021. 6

[2] V Aslantas and Emre Bendes. A new image quality metric for image fusion: The sum of the correlations of differences. *Aeu-International Journal of Electronics and Communications*, 69(12):1890–1896, 2015. 5

[3] Haoyu Chen, Jinjin Gu, Yihao Liu, Salma Abdel Magid, Chao Dong, Qiong Wang, Hanspeter Pfister, and Lei Zhu. Masked image training for generalizable deep image denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1692–1703, 2023. 6

[4] Ahmet M Eskicioglu and Paul S Fisher. Image quality measures and their performance. *IEEE Transactions on Communications*, 43(12):2959–2965, 1995. 5

[5] Qishen Ha, Kohei Watanabe, Takumi Karasawa, Yoshitaka Ushiku, and Tatsuya Harada. Mfnet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5108–5115, 2017. 5

[6] Yu Han, Yunze Cai, Yin Cao, and Xiaoming Xu. A new image fusion performance metric based on visual information fidelity. *Information Fusion*, 14(2):127–135, 2013. 5

[7] Zhanbo Huang, Jinyuan Liu, Xin Fan, Risheng Liu, Wei Zhong, and Zhongxuan Luo. Reconet: Recurrent correction network for fast and efficient multi-modality image fusion. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 539–555, 2022. 5

[8] Xinyu Jia, Chuang Zhu, Minzhen Li, Wenqi Tang, and Wenli Zhou. Llvip: A visible-infrared paired dataset for low-light vision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3496–3504, 2021. 5

[9] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8110–8119, 2020. 2

[10] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5148–5157, 2021. 5

[11] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2426–2435, 2022. 2

[12] Boyun Li, Xiao Liu, Peng Hu, Zhongqin Wu, Jiancheng Lv, and Xi Peng. All-in-one image restoration for unknown corruption. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17452–17462, 2022. 6

[13] Wentong Liao, Kai Hu, Michael Ying Yang, and Bodo Rosenhahn. Text to image generation with semantic-spatial aware gan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18187–18196, 2022. 2

[14] Jinyuan Liu, Xin Fan, Zhanbo Huang, Guanyao Wu, Risheng Liu, Wei Zhong, and Zhongxuan Luo. Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5802–5811, 2022. 2, 5

[15] Jinyuan Liu, Zhu Liu, Guanyao Wu, Long Ma, Risheng Liu, Wei Zhong, Zhongxuan Luo, and Xin Fan. Multi-interactive feature learning and a full-time multi-modality benchmark for image fusion and segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8115–8124, 2023. 1

[16] Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. Compositional visual generation with composable diffusion models. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 423–439, 2022. 2

[17] Yu Liu, Xun Chen, Rabab K Ward, and Z Jane Wang. Image fusion with convolutional sparse representation. *IEEE Signal Processing Letters*, 23(12):1882–1886, 2016. 2

[18] Jiayi Ma, Yong Ma, and Chang Li. Infrared and visible image fusion methods and applications: A survey. *Information Fusion*, 45:153–178, 2019. 1, 5

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

[20] Jiayi Ma, Hao Zhang, Zhenfeng Shao, Pengwei Liang, and Han Xu. Ganmcc: A generative adversarial network with multiclassification constraints for infrared and visible image fusion. *IEEE Transactions on Instrumentation and Measurement*, 70:1–14, 2021. 1

[21] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, 21 (12):4695–4708, 2012. 5

[22] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212, 2012. 5

[23] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2085–2094, 2021. 2

[24] Vaishnav Potlapalli, Syed Waqas Zamir, Salman Khan, and Fahad Shahbaz Khan. Promptir: Prompting for all-in-one blind image restoration. *arXiv preprint arXiv:2306.13090*, 2023. 2

[25] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry,

Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*, pages 8748–8763, 2021. 2

[26] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022. 2

[27] Yiming Sun, Bing Cao, Pengfei Zhu, and Qinghua Hu. Detfusion: A detection-driven infrared and visible image fusion network. In *Proceedings of the ACM International Conference on Multimedia*, pages 4003–4011, 2022. 1

[28] Linfeng Tang, Jiteng Yuan, and Jiayi Ma. Image fusion in the loop of high-level vision tasks: A semantic-aware real-time infrared and visible image fusion network. *Information Fusion*, 82:28–42, 2022. 1

[29] Linfeng Tang, Jiteng Yuan, Hao Zhang, Xingyu Jiang, and Jiayi Ma. Piafusion: A progressive infrared and visible image fusion network based on illumination aware. *Information Fusion*, 83:79–92, 2022. 1, 5

[30] Ming Tao, Hao Tang, Fei Wu, Xiao-Yuan Jing, Bing-Kun Bao, and Changsheng Xu. Df-gan: A simple and effective baseline for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16515–16525, 2022. 2

[31] Di Wang, Jinyuan Liu, Xin Fan, and Risheng Liu. Unsupervised misaligned infrared and visible image fusion via cross-modality image generation and registration. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2022. 5

[32] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2555–2563, 2023. 5

[33] Wenhui Wu, Jian Weng, Pingping Zhang, Xu Wang, Wenhan Yang, and Jianmin Jiang. 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)*, pages 5901–5910, 2022. 6

[34] Han Xu, Jiayi Ma, Junjun Jiang, Xiaojie Guo, and Haibin Ling. U2fusion: A unified unsupervised image fusion network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1):502–518, 2022. 2, 5

[35] Han Xu, Jiayi Ma, Jiteng Yuan, Zhuliang Le, and Wei Liu. Rfnet: Unsupervised network for mutually reinforcing multi-modal image registration and fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19679–19688, 2022. 1

[36] Han Xu, Jiteng Yuan, and Jiayi Ma. Murf: Mutually reinforcing multi-modal image registration and fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):12148–12166, 2023. 5

[37] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5728–5739, 2022. 3

[38] Hao Zhang and Jiayi Ma. Sdnet: A versatile squeeze-and-decomposition network for real-time image fusion. *International Journal of Computer Vision*, 129:2761–2785, 2021. 1

[39] Hao Zhang, Han Xu, Xin Tian, Junjun Jiang, and Jiayi Ma. Image fusion meets deep learning: A survey and perspective. *Information Fusion*, 76:323–336, 2021. 1

[40] Hang Zhao, Orazio Gallo, Iuri Frosio, and Jan Kautz. Loss functions for image restoration with neural networks. *IEEE Transactions on Computational Imaging*, 3(1):47–57, 2016. 4

[41] Wenda Zhao, Shigeng Xie, Fan Zhao, You He, and Huchuan Lu. Metafusion: Infrared and visible image fusion via meta-feature embedding from object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13955–13965, 2023. 5

[42] Zixiang Zhao, Haowen Bai, Yuanzhi Zhu, Jiangshe Zhang, Shuang Xu, Yulun Zhang, Kai Zhang, Deyu Meng, Radu Timofte, and Luc Van Gool. Ddfm: denoising diffusion model for multi-modality image fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8082–8093, 2023. 2, 5

[43] Zhengjie Zhu, Xiaogang Yang, Ruitao Lu, Tong Shen, Xueli Xie, and Tao Zhang. Clf-net: Contrastive learning for infrared and visible image fusion network. *IEEE Transactions on Instrumentation and Measurement*, 71:1–15, 2022. 1