Image Fusion via Vision-Language Model

Zixiang Zhao [1 2]   **Lilun Deng** [1]   **Haowen Bai** [1]   **Yukun Cui** [1]   **Zhipeng Zhang** [2 3]   **Yulun Zhang** [4]
Haotong Qin [2]   **Dongdong Chen** [5]   **Jiangshe Zhang** [1]   **Peng Wang** [3]   **Luc Van Gool** [2 6 7]

Abstract

Image fusion integrates essential information from multiple images into a single composite, enhancing structures, textures, and refining imperfections. Existing methods predominantly focus on pixel-level and semantic visual features for recognition, but often overlook the deeper text-level semantic information beyond vision. Therefore, we introduce a novel fusion paradigm named *image Fusion via vIsion-Language Model* (**FILM**), for the first time, utilizing explicit textual information from source images to guide the fusion process. Specifically, FILM generates semantic prompts from images and inputs them into ChatGPT for comprehensive textual descriptions. These descriptions are fused within the textual domain and guide the visual information fusion, enhancing feature extraction and contextual understanding, directed by textual semantic information via cross-attention. FILM has shown promising results in four image fusion tasks: infrared-visible, medical, multi-exposure, and multi-focus image fusion. We also propose a vision-language dataset containing ChatGPT-generated paragraph descriptions for the eight image fusion datasets across four fusion tasks, facilitating future research in vision-language model-based image fusion. Code and dataset are available at https://github.com/Zhaozixiang1228/IF-FILM.

1. Introduction

Image fusion (Zhao et al., 2023b;c; Liu et al., 2022a; Zhang, 2021b), standing as a critical technique in computer vision,
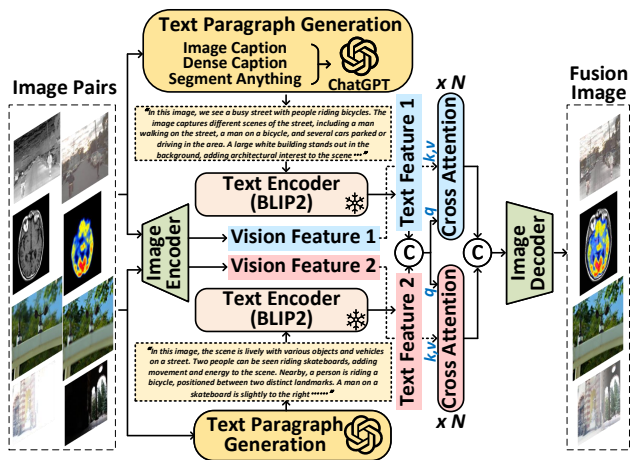

Figure 1: Workflow for our FILM. Input images are first processed to create prompts for ChatGPT (OpenAI, 2023), which then generate detailed textual descriptions. These descriptions help to get fused textual features via the frozen BLIP2 (Li et al., 2023b) model. Then, these textual features are fused and guide the extraction and fusion of visual features via cross-attention, enhancing contextual understanding with text-based semantic information. Finally, the fusion image is output by the image decoder.

combines information from multiple source images to create a single image that is more informative and suitable for human or machine perception. The realm of image fusion encompasses several sub-tasks, each addressing unique challenges and applications. Representatively, infrared-visible image fusion combines infrared and visible images, enhancing object representation under varied illumination conditions (Zhao et al., 2023b;c). Medical image fusion integrates different modalities of clinical images such as MRI and CT scans, offering a more comprehensive view for diagnosis and treatment planning (James & Dasarathy, 2014; Xu & Ma, 2021; Li et al., 2023c). Multi-exposure image fusion merges images taken with different exposure settings to capture a wider range of luminance, which is crucial in high dynamic range imaging (Ma et al., 2020b; 2017). Lastly, multi-focus image fusion merges images focused on different planes to produce a uniformly sharp image, invaluable in microscopy and macro photography (Deng & Dragotti, 2021; Zhang, 2021b).

Despite its widespread application, the current state of im-

[1]Xi'an Jiaotong University, China [2]ETH Zürich, Switzerland [3]Northwestern Polytechnical University, China [4]MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University, China [5]Heriot-Watt University, United Kingdom [6]KU Leuven, Belgium [7]INSAIT, Bulgaria. Correspondence to: Yulun Zhang <yulun100@gmail.com>, Dongdong Chen <d.chen@hw.ac.uk>.

age fusion is marred by a notable limitation: an over-reliance on visual features. The prevalent methodologies in this field predominantly center around the vision feature extraction, alignment, fusion, and reconstruction, prioritizing aspects like texture, contrast, highlight information, and pixel registration (Zhao et al., 2023b; Liang et al., 2022). This consequently neglects the deeper, semantic layers of information that images inherently possess. Approaches that integrate downstream pattern recognition tasks like semantic segmentation (Tang et al., 2022a;b; Liu et al., 2023a) and object detection (Liu et al., 2022a; Sun et al., 2022; Zhao et al., 2023a), although progressive, still fall short as they remain concentrate on the superficial semantics derived from visual pixel level cues rather than the deeper, more nuanced textual information that images can convey. Therefore, how to better utilize the deeper-semantic features that go beyond visual information in images, becomes a breakthrough point that urgently needs to be addressed.

With the development of large language models (OpenAI, 2023), some work (Zhang et al., 2023; Zhu et al., 2023; Radford et al., 2021; OpenAI, 2023) attempts to utilize Vision Language Model (VLM) for information fusion and alignment as supplementary. These models, which include notable architectures like CLIP (Radford et al., 2021) and GPT4 (OpenAI, 2023), demonstrate remarkable proficiency in understanding and generating content that synergizes visual and textual components. They not only tap into the knowledge capabilities of the large language model, but also align and fuse with visual information (Li et al., 2023b). This synergy has the potential to significantly enhance image fusion processes, offering a pathway to incorporate deeper semantic understanding guided by language, thereby enabling a more comprehensive and contextually rich fusion process. For instance, when describing two multi-modal images from the same scene, the descriptions should focus on the response characteristics unique to each modality; for descriptions of multi-focus images, the text should pay greater attention to the areas of perfect imaging that align with their focal points. Thus, we can extract textual descriptions from images based on the large vision-language model and, after integrating descriptions on the textual feature level, we then use fused text features to guide the extraction and fusion of features at the image and vision level.

Therefore, in this paper, we propose a novel fusion algorithm called *Image Fusion via VIsion-Language Model* (**FILM**). This approach integrates the capabilities of VLMs into the image fusion process, for the first time, leveraging the semantic understanding derived from textual data to guide and enhance the fusion of visual features. Our methodology comprises three components: text feature fusion, language-guided vision feature fusion, and vision feature decoding. The workflow of these components is depicted in Figure 1. Our contributions are summarized as follows:

- We propose a novel paradigm for image fusion. To our knowledge, this is the first instance of incorporating explicit (*language model derived*) textual guidance into image fusion algorithms. This approach aids in the deeper understanding of text-level semantic information, facilitating the extraction and fusion of strengths from each source image.

- Our model has achieved satisfactory results in infrared-visible, medical, multi-exposure, and multi-focus image fusion tasks, demonstrating its effectiveness across various application scenarios.

- We introduce a series of vision-language benchmark datasets for image fusion covering eight fashion datasets across four fusion tasks. These datasets comprise manually refined prompts tailored for the ChatGPT model, alongside paired textual descriptions generated by ChatGPT, which are designed to facilitate subsequent research in image fusion using vision-language models.

2. Related Work

In this section, we will review the image fusion algorithms in the era of deep learning (DL) and introduce the key technology used in our paper: the Vision-Language Model (VLM).

DL-Based Image Fusion. In the era of deep learning, neural networks are often used for source feature extraction, feature fusion, and fused image reconstruction (Zhao et al., 2023b; Zhang, 2021b). **(a)** In *multi-modal image fusion*, since there is no ground truth available, it inherently belongs to an unsupervised task. The fusion methods can be divided into generative and discriminative categories (Zhang & Demiris, 2023). Generative algorithms model the latent space manifold through the generative adversarial network (GAN) (Ma et al., 2019b; Liu et al., 2022a) or denoising diffusion model (Zhao et al., 2023c), making the distribution gap between source and fused images as close as possible. On the other hand, discriminative models, based on regression ideas, use the model-driven (Xu et al., 2021; Zhao et al., 2022a; Li et al., 2023a; Zhao et al., 2022b) or data-driven (Zhao et al., 2020; 2023b; Li & Wu, 2018) auto-encoder structures to learn the source-fusion images mapping. Additionally, downstream cross-modal pattern recognition tasks, such as object detection (Liu et al., 2022a; Sun et al., 2022; Zhao et al., 2023a) and semantic segmentation (Tang et al., 2022a;b; Liu et al., 2023a), are employed to make the fused images highlight features and regions containing vision-based semantic information (Zhao et al., 2023a). **(b)** In *digital image fusion*, supervised fusion algorithms often obtain a mapping from imperfect source images to perfect ground truth by predicting decision maps or reconstructing images (Deng & Dragotti, 2021; Liu et al., 2017; Xiao et al., 2021; 2020; Ma et al., 2020a; Yin et al., 2021b). For issues

where perfect training ground truth, like normally-exposed and all-focus images, are hard to obtain, unsupervised algorithms often reconstruct fused images based on CNN (Prabhakar et al., 2017; Han et al., 2022; Gao et al., 2022; Bouzos et al., 2023; Guan et al., 2023; Deng et al., 2021), Transformer (Qu et al., 2022; Guan et al., 2023), or GAN (Yin et al., 2021a; Cai et al., 2018; Xu et al., 2020b; Guo et al., 2019). The feature substitution or fusion, with the help of no-reference quality metrics (Xu et al., 2020c; Liu et al., 2023b), usually occurrence in image domains, frequency domains, or feature spaces (Ma et al., 2020a; Wang et al., 2022b; 2023). **(c)** Furthermore, registration-based methods focus on solving the misalignment issue in multi-source image inputs, reducing artifacts in the fused images (Wang et al., 2022a; Jiang et al., 2022; Xu et al., 2022b). Meanwhile, unified frameworks explore the meta-information in different fusion tasks, investigating the mutual promotion effects and alleviating the issues of lacking paired training data and absence of ground truth (Xu et al., 2022a; Liang et al., 2022).

Vision-Language Model. Recently, visual language multi-modal learning (Radford et al., 2021; Li et al., 2022; Xu et al., 2015; Brooks et al., 2023; Qin et al., 2024; Huang et al., 2024) has become a hot research topic. In particular, vision-language models (Zhu et al., 2023; Wang et al., 2021; Li et al., 2022; OpenAI, 2023; Zhang et al., 2023) such as BLIP (Li et al., 2022; 2023b), DALL-E (Ramesh et al., 2021; 2022), and GPT4 (OpenAI, 2023) have shown powerful performance in several downstream tasks. BLIP demonstrates powerful knowledge prompt capabilities in bridging between frozen visual feature pre-trained encoders and frozen large language models. GPT4 also shows strong general performance based on visual language pre-training. With the help of these large models (Touvron et al., 2023; OpenAI, 2023), a lot of studies (Zhang et al., 2023; Zhu et al., 2023) in image captioning have turned into guiding the large models to provide detailed descriptions of images in the form of natural language. These large models provide external common knowledge for image caption. The key details information from the image such as dense caption, can provide a strong explicit prompt. It allows the image to be presented in a descriptive form that covers the key information. Inspired by this, we aim to introduce a vision-language model to image fusion so that text can guide image fusion in an effective and intuitive way.

Comparison with Existing Approaches. The most closely related approaches to our method are the ones that use pattern recognition tasks to provide guidance through visual semantic information (Tang et al., 2022b; Liu et al., 2023a; 2022a; Zhao et al., 2023a). In contrast to such methods, our approach transcends the limitations of visual semantic information by utilizing deeper-level textual semantic information, guiding the feature extraction and selection in fusion

tasks through language and textual features. The integration of VLM in image fusion tasks promises a transformative shift, enabling a more holistic understanding of images through the combined perspectives of both visual perception and textual context, thereby paving the way for more sophisticated and application-specific fusion techniques.

3. Method

In this paper, we denote the input pairs of images, which may be a pair of infrared-visible, medical, multi-exposure, or multi-focus images, as I_1 and I_2. The algorithm ultimately outputs a fused image, represented as F. In this section, we will provide a comprehensive description of our FILM algorithm, denoted as $I_F = \mathsf{FILM}(I_1, I_2)$, elucidating its workflow and design details.

3.1. Workflow Overview

Brief and detailed workflows of our FILM paradigm are illustrated in Figures 1 and 2. FILM is segmented into three components: *text feature fusion*, *language-guided vision feature fusion*, and *vision feature decoding*, corresponding to the first, second, and third columns of Figure 2, and denoted as $\mathcal{T}(\cdot)$, $\mathcal{V}(\cdot)$, and $\mathcal{D}(\cdot)$, respectively. Specifically, our FILM algorithm takes two source images $\{I_1, I_2\}$ as input, which are initially processed by the *text feature fusion* component $\mathcal{T}$. This component encompasses generating prompts from image caption (Li et al., 2023b), dense caption (Nguyen et al., 2022), and Segment Anything (Kirillov et al., 2023), to produce textual descriptions via ChatGPT (OpenAI, 2023). The descriptions are encoded via the text encoder of BLIP2 (Li et al., 2023b), and the text features are subsequently fused. The *language-guided vision feature fusion* component $\mathcal{V}$ then utilizes the fused text features to guide the extraction of visual features from the source images using cross-attention. This process identifies and integrates the salient aspects and advantages to be incorporated into the fused image. Finally, the fusion result F is output by the *vision feature decoding* component $\mathcal{D}$, which decodes the fused vision features into an image. Each component's details will be elaborated upon separately.

3.2. Fusion via Vision-Language Model

Component I: Text Feature Fusion. In the text feature fusion component, paired source images $\{I_1, I_2\}$ are input, resulting in the fused text feature Φ_F^T, i.e.,

$$\Phi_F^T = \mathcal{T}(I_1, I_2). \qquad (1)$$

Initially, inspired by Li et al. (2023b); Nguyen et al. (2022); Kirillov et al. (2023); Zhao (2023), we input the images into the BLIP2 (Li et al., 2023b), GRIT (Nguyen et al., 2022), and Segment Anything (Kirillov et al., 2023) models to
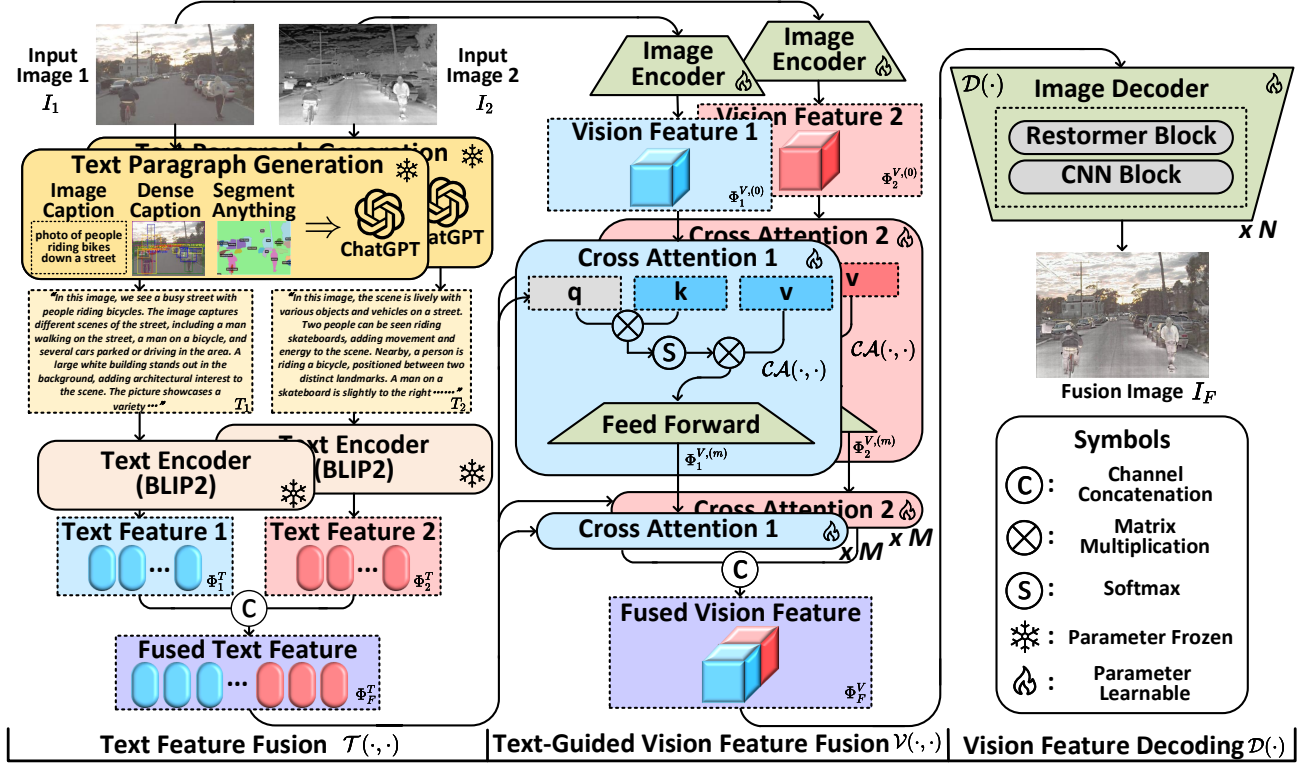
Figure 2: Workflow for our FILM, which encompasses three components: text paragraph generation and text feature fusion, language-guided vision feature fusion via cross attention and vision feature decoding, corresponding to the first, second, and third columns.

extract image semantic information from holistic to fine-grained, as *Image Caption*, *Dense Caption*, and *Semantic Mask*. Subsequently, these three prompts are fed into the ChatGPT (OpenAI, 2023) model to generate paired text descriptions $\{T_1, T_2\}$ for the source images $\{I_1, I_2\}$. We then input $\{T_1, T_2\}$ into the text encoder of parameter-frozen BLIP2 (Li et al., 2023b) model, obtaining the corresponding text features $\{\Phi_1^T, \Phi_2^T\}$. Ultimately, the fused text feature Φ_F^T is obtained by concatenating $\{\Phi_1^T, \Phi_2^T\}$. For more details on feature prompting and text generation, please refer to Section 4.

Component II: Language-Guided Vision Feature Fusion. In the language-guided vision feature fusion component, we guide the extraction of visual features from the source image through text features, resulting in the visual fusion feature Φ_F^V, i.e.

$$\Phi_F^V = \mathcal{V}(\Phi_F^T, I_1, I_2). \tag{2}$$

Firstly, source images $\{I_1, I_2\}$ are fed into the image encoders, producing shallow visual features $\{\Phi_1^{V,(0)}, \Phi_2^{V,(0)}\}$ from $\{I_1, I_2\}$, respectively. The image encoder, consisting of Restormer blocks (Zamir et al., 2022) and CNN blocks (He et al., 2016), focuses on both global and local visual representations while maintaining computational efficiency and effective feature extraction. Subsequently, these shallow

features are input into the cross-attention mechanism, where fused text features direct the visual feature extraction process, specifically emphasizing aspects of the source image that are desired to be preserved in the output fused image. That is:

$$\Phi_1^{V,(m)} = \mathcal{CA}\left(\Phi_F^T, \Phi_1^{V,(m-1)}\right), \tag{3}$$

where $m = 1, \cdots, M$. $\mathcal{CA}(\cdot)$ represents the Cross-Attention module, and $\Phi_2^{V,(m)}$ can be obtained similarly by replacing the subscripts. In $\mathcal{CA}(\cdot)$, *Key* (K) and *Value* (V) are provided by $\Phi_1^{V,(m-1)}$ or $\Phi_2^{V,(m-1)}$, while *Query* (Q) is provided by Φ_F^T. Moreover, the feed-forward operation in $\mathcal{CA}(\cdot)$ is also implemented through the Restormer block (Zamir et al., 2022).

After passing through M Cross-Attention modules, the visual features from text-guided extraction are represented as $\{\Phi_1^{V,(M)}, \Phi_2^{V,(M)}\}$. Subsequently, after the concatenation through channel dimension, $\{\Phi_1^{V,(M)}, \Phi_2^{V,(M)}\}$ yield the fused visual feature Φ_F^V, as shown in Figure 2.

Component III: Vision Feature Decoding. Finally, the fused visual feature Φ_F^V is input into the image decoder $\mathcal{D}$, comprising N layers of Restormer (Zamir et al., 2022) and CNN (He et al., 2016) blocks, from which the fused image is output, denoted as $I_F = \mathcal{D}(\Phi_F^V)$. I_F refers to the final output fusion image of FILM.
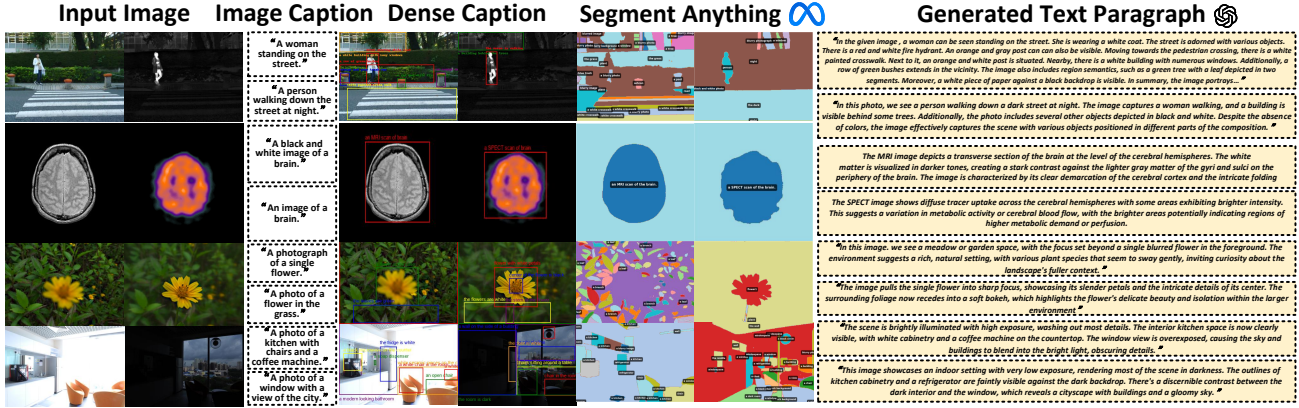
Figure 3: Visualization of the VLF dataset creation process and representative data displays.

4. Vision-Language Fusion Dataset

In this section, we will introduce details of the proposed *Vision-Language Fusion* (**VLF**) Dataset, including prompts generation, paragraph descriptions output, and representative visualization displays.

Overview. Considering the high computational cost of invoking various vision-language components, and to facilitate subsequent research on image fusion based on vision-language models, we propose the VLF Dataset. This dataset encompasses paired paragraph descriptions generated by ChatGPT, covering all image pairs from the training and test sets of the eight widely-used fusion datasets. These include the MSRS (Tang et al., 2022c), M³FD (Liu et al., 2022a) and RoadScene (Xu et al., 2020a) datasets for *infrared-visible image fusion* (IVF) task, the Harvard medical dataset (Johnson & Becker) for *medical image fusion* (MIF) task, the RealMFF (Zhang et al., 2020a) and Lytro (Nejati et al., 2015) datasets for *multi-focus image fusion* (MFF) task, and the SICE (Cai et al., 2018) and MEFB (Zhang, 2021a) datasets for *multi-exposure image fusion* (MEF) task.

Prompt Generation. The output of each component from the Text Paragraph Generation module in FILM is shown in Figure 3. Firstly, inspired by Zhao (2023), BLIP2 (Li et al., 2023b), GRIT (Nguyen et al., 2022) and Segment Anything (Kirillov et al., 2023) models output Image Caption, Dense Caption, and Semantic Mask, respectively. They provide the one-sentence caption, object-level information, and semantic mask for the input and representing semantic information ranging from coarse-grained to fine-grained.

Generated Paragraph Descriptions. Subsequently, the generated semantic prompts from paired images are input into ChatGPT (OpenAI, 2023) to generate paragraph descriptions, which are used to guide subsequent fusion tasks.

Statistical Information. This dataset contains 7040 paragraph descriptions, with each description consisting of seven sentences and 186 words on average. We present examples

of representative infrared-visible, medical, multi-exposure, multi-focus image pairs in Figure 3. More dataset details can be found in Appendix A.

5. Experiment

In this section, we will demonstrate the performance of FILM on various image fusion tasks, showcasing its superiority. Due to space constraints, more visual results are presented in the supplementary material (Appendix C).

Loss Function. For the total training loss, we set it as:

$$\mathcal{L}_{total} = \mathcal{L}_{int} + \alpha_1 \mathcal{L}_{grad} + \alpha_2 \mathcal{L}_{\text{SSIM}}, \qquad (4)$$

where α_1, α_2 are tuning parameters. In the IVF task, following the setting in Zhao et al. (2023b), $\mathcal{L}_{int} = \frac{1}{HW}\|I_F - \max(I_1, I_2)\|_1$, and $\mathcal{L}_{grad} = \frac{1}{HW}\| |\nabla I_F| - \max(|\nabla I_1|, |\nabla I_2|)\|_1$. ∇ indicates the Sobel gradient operator. α_1 and α_2 are set to 20 and 0, respectively. MIF task does not need fine-tuning training, therefore it has no loss function. For MFF and MEF tasks, inspired by Liu et al. (2023b), we set $\mathcal{L}_{int} = \frac{1}{HW}\|I_F - mean(I_1, I_2)\|_1$, $\mathcal{L}_{grad} = \frac{1}{HW}\| |\nabla I_F| - \max(|\nabla I_1|, |\nabla I_2|)\|_1$, and $\mathcal{L}_{\text{SSIM}} = 2 - \text{SSIM}(I_1, I_F) - \text{SSIM}(I_2, I_F)$. $\{\alpha_1, \alpha_2\}$ are set to $\{300, 1\}$ and $\{500, 1\}$ in MFF and MEF tasks respectively, in order to ensure the magnitude comparable in each term.

Training Details. A machine with eight NVIDIA GeForce RTX 3090 GPUs is utilized for our experiments. We train the network for 300 epochs using the Adam optimizer, with an initial learning rate of 1e-4 and decreasing by a factor of 0.5 every 50 epochs. The Adam optimization strategy is employed with the batchsize set as 16. We incorporate Restormer blocks (Zamir et al., 2022) in both language-guided vision encoder $\mathcal{V}(\cdot)$ and vision feature decoder $\mathcal{D}(\cdot)$, with each block having 8 attention heads and a dimensionality of 64. M and N, representing the number of blocks in $\mathcal{V}(\cdot)$ and $\mathcal{D}(\cdot)$, are set to 2 and 3, respectively.

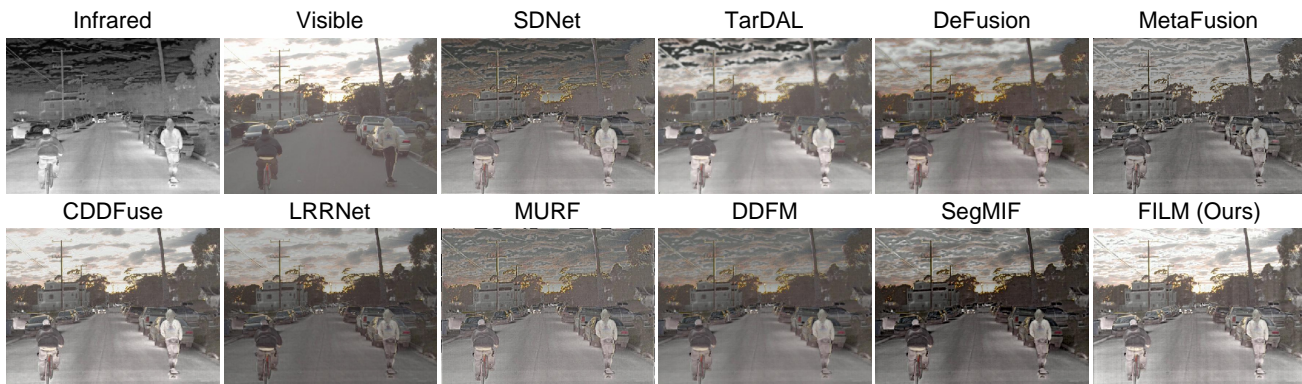**Metrics.** We employ six quantitative metrics to assess the

Figure 4: Visualization comparison of the fusion results in the infrared-visible image fusion task.

Table 1: Quantitative results of IVF. The red and blue markers represent the best and second-best values, respectively.

	MSRS Infrared-Visible Fusion Dataset							**M³FD Infrared-Visible Fusion Dataset**							**RoadScene Infrared-Visible Fusion Dataset**					
	EN	SD	SF	AG	VIF	Qabf		EN	SD	SF	AG	VIF	Qabf		EN	SD	SF	AG	VIF	Qabf
SDN	5.25	17.35	8.67	2.67	0.50	0.38	SDN	6.79	34.63	14.86	5.16	0.56	0.54	SDN	7.34	44.74	14.99	5.94	0.62	0.55
TarD	5.28	25.22	5.98	1.83	0.42	0.18	TarD	6.79	40.75	8.18	2.92	0.53	0.30	TarD	7.25	47.57	11.46	4.23	0.56	0.43
DeF	6.46	37.63	8.60	2.80	0.77	0.54	DeF	6.84	35.09	9.65	3.37	0.59	0.42	DeF	7.39	47.60	11.26	4.47	0.63	0.50
Meta	5.65	24.97	9.99	3.40	0.63	0.48	Meta	6.68	29.62	16.22	5.68	0.68	0.57	Meta	6.87	31.95	14.40	5.55	0.55	0.46
CDDF	6.70	43.39	11.56	3.74	1.05	0.69	CDDF	7.08	41.29	16.49	5.42	0.78	0.63	CDDF	7.41	54.59	17.04	6.07	0.63	0.51
LRR	6.19	31.78	8.46	2.63	0.54	0.46	LRR	6.60	30.19	11.69	3.95	0.57	0.51	LRR	7.09	38.77	11.50	4.36	0.43	0.33
MURF	5.04	20.63	10.49	3.38	0.44	0.36	MURF	6.52	27.90	11.43	4.51	0.39	0.30	MURF	6.91	33.46	13.74	5.31	0.53	0.47
DDFM	6.19	29.26	7.44	2.51	0.73	0.48	DDFM	6.72	31.15	9.84	3.42	0.63	0.47	DDFM	7.27	42.94	10.89	4.20	0.63	0.50
SegM	5.95	37.28	11.10	3.47	0.88	0.63	SegM	6.89	35.64	16.11	5.52	0.78	0.65	SegM	7.29	47.10	15.07	5.78	0.65	0.56
Ours	6.72	43.17	11.70	3.84	1.06	0.73	Ours	7.09	41.53	16.77	5.55	0.83	0.67	Ours	7.43	49.25	17.34	6.60	0.69	0.62

Figure 5: Visualization comparison of the fusion results in the medical image fusion task.

Table 2: Quantitative results of MIF. The red and blue markers represent the best and second-best values, respectively.

	Harvard Medical Image Fusion Dataset					
	EN	SD	SF	AG	VIF	Qabf
DIFNet	4.58	49.99	14.93	4.09	0.61	0.59
DeFusion	3.90	54.77	16.87	4.30	0.62	0.57
CDDFuse	4.00	70.58	22.84	5.75	0.71	0.69
DDFM	3.82	56.47	16.17	4.16	0.68	0.65
SDNet	3.53	48.85	23.15	5.53	0.54	0.63
U2Fusion	3.56	49.95	19.70	4.98	0.47	0.53
MATR	4.09	48.63	17.87	4.70	0.75	0.72
GeSeNet	4.31	62.47	22.72	5.85	0.76	0.76
MsgFusion	4.06	75.01	20.34	5.09	0.49	0.50
Ours	4.74	65.26	23.36	6.19	0.78	0.76

fusion outcomes: entropy (EN), standard deviation (SD), spatial frequency (SF), average gradient (AG), visual information fidelity (VIF) and $Q^{AB/F}$. Higher metric values indicate superior quality in the fused image. Further information is available in Ma et al. (2019a).

5.1. Infrared and Visible Image Fusion

Setup. Following Zhao et al. (2023b;c), infrared-visible fusion experiments are conducted on the MSRS (Tang et al.,

2022c), M³FD (Liu et al., 2022a) and RoadScene (Xu et al., 2020a) datasets. 1083 image pairs in MSRS are for training and 361 pairs are for testing. The generalizability of FILM is further assessed by M³FD and RoadScene without fine-tuning. We evaluated FILM against various state-of-the-art (SOTA) methods including SDNet (Zhang & Ma, 2021), TarDAL (Liu et al., 2022a), DeFusion (Liang et al., 2022), MetaFusion (Zhao et al., 2023a), CDDFuse (Zhao et al., 2023b), LRRNet (Li et al., 2023a), MURF (Xu et al., 2023), DDFM (Zhao et al., 2023c), and SegMIF (Liu et al., 2023a).

Comparison with SOTA Methods. In Figure 4, FILM successfully integrated the thermal radiation information with the detailed texture features. Leveraging textual features
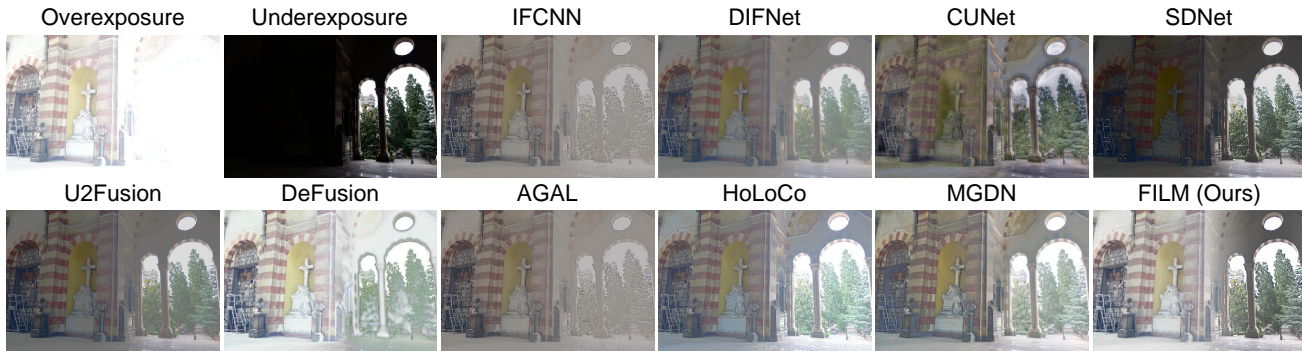
Figure 6: Visualization comparison of the fusion results in the multi-exposure image fusion task.

Table 3: Quantitative results of MEF. The red and blue markers represent the best and second-best values, respectively.

	SICE Multi-exposure Image Fusion Dataset							MEFB Multi-exposure Image Fusion Dataset					
	EN	SD	SF	AG	VIF	Qabf		EN	SD	SF	AG	VIF	Qabf
IFCNN	6.67	39.43	16.93	4.59	0.73	0.71	IFCNN	6.99	52.49	18.16	5.34	0.71	0.69
DIFNet	6.56	35.76	11.86	3.09	0.46	0.50	DIFNet	6.99	50.23	11.79	3.47	0.51	0.53
CUNet	6.90	34.18	11.87	3.80	0.69	0.50	CUNet	7.18	45.37	12.78	4.28	0.71	0.50
SDNet	6.47	38.25	19.34	4.80	0.48	0.45	SDNet	6.59	51.77	20.53	5.27	0.55	0.42
U2Fusion	6.43	34.77	10.71	3.17	0.48	0.57	U2Fusion	6.67	46.73	12.54	3.82	0.51	0.56
DeFusion	6.87	44.73	14.28	4.04	0.87	0.57	DeFusion	7.10	56.46	14.86	4.48	0.70	0.59
AGAL	7.06	46.03	16.64	4.91	0.72	0.53	AGAL	7.14	60.63	17.77	5.33	0.79	0.65
HoLoCo	7.04	42.73	9.33	3.47	0.74	0.37	HoLoCo	7.20	53.88	12.80	4.34	0.73	0.54
MGDN	6.94	43.69	15.04	4.59	0.88	0.64	MGDN	7.25	55.97	18.09	5.76	0.96	0.65
Ours	7.07	54.21	19.42	5.15	1.05	0.79	Ours	7.31	69.02	20.98	6.15	0.98	0.77

and knowledge, the fusion process enhanced the visibility of objects in low-light environments, making textures and contours clearer, and reducing artifacts. For the quantitative results in Table 1, our method showcases exceptional performance in almost all metrics, confirming its adaptability for various environmental scenarios and object categories. Hence, FILM is proven to well maintain the completeness and richness of the information from source images, and generate results that conform to human visual perception.

5.2. Medical Image Fusion

Setup. Following Zhao et al. (2023c), we engage the Harvard Medical dataset (Johnson & Becker), which consisted of 50 pairs of MRI-CT, MRI-PET, and MRI-SPECT images, to evaluate the generalizability of our model. Notably, we employ the model trained on the IVF experiments and conducted a generalization test on the Harvard Medical dataset without any fine-tuning. The competitors include DIFNet (Jung et al., 2020), SDNet (Zhang & Ma, 2021), U2Fusion (Xu et al., 2022a), DeFusion (Liang et al., 2022), MATR (Tang et al., 2022d), CDDFuse (Zhao et al., 2023b), DDFM (Zhao et al., 2023c), GeSeNet (Li et al., 2023d) and Msg-Fusion (Wen et al., 2023). Results from DIFNet, DeFusion, CDDFuse, and DDFM are the generalized outcomes of IVF models without fine-tuning, whereas the other results are

from models specialized training using the MIF datasets.

Comparison with SOTA Methods. In terms of visual perception and quantitative analysis (Figure 5 and Table 2), FILM has shown outstanding accuracy in extracting cross-modal structural highlights and detailed texture features, effectively integrating source information into the fused images. These achievements surpass even those of fusion models specifically fine-tuned via medical image pairs.

5.3. Multi-exposure Image Fusion

Setup. We conduct MEF experiments on the SICE (Cai et al., 2018) and MEFB (Zhang, 2021a) dataset. We utilized 499 pairs from SICE dataset for training, while 90 pairs from SICE and 40 pairs from MEFB for testing. Our comparison methods encompass IFCNN (Zhang et al., 2020b), DIFNet (Jung et al., 2020), CUNet (Deng & Dragotti, 2021), SDNet (Zhang & Ma, 2021), U2Fusion (Xu et al., 2022a), DeFusion (Liang et al., 2022), AGAL (Liu et al., 2022b), HoLoCo (Liu et al., 2023b) and MGDN (Guan et al., 2023).

Comparison with SOTA Methods. Both quantitative and qualitative results in Table 3 and Figure 6 demonstrate the effectiveness of FILM, which adeptly handles multiple images with varying exposures, expanding the dynamic range while simultaneously improving image quality and enhanc-
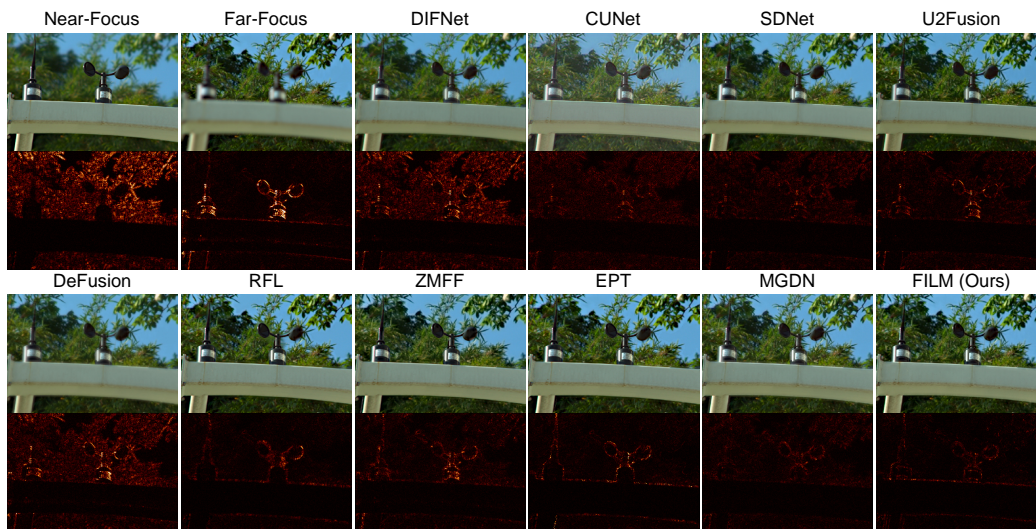
Figure 7: Visualization comparison of the fusion results and error maps in the multi-focus image fusion task.

Table 4: Quantitative results of MFF. The red and blue markers represent the best and second-best values, respectively.

	RealMFF Multi-focus Image Fusion Dataset							Lytro Multi-focus Image Fusion on Dataset					
	EN	SD	SF	AG	VIF	Qabf		EN	SD	SF	AG	VIF	Qabf
DIFNet	7.01	51.17	10.78	3.96	0.89	0.69	DIFNet	7.43	52.52	11.47	4.30	0.73	0.54
CUNet	6.72	38.97	13.59	4.81	0.77	0.65	CUNet	7.25	45.78	15.54	5.58	0.71	0.65
SDNet	6.95	50.96	15.22	5.02	0.93	0.73	SDNet	7.47	55.25	16.88	5.84	0.84	0.69
U2Fusion	6.77	48.49	14.07	5.09	0.95	0.70	U2Fusion	7.30	51.95	14.83	5.60	0.83	0.65
DeFusion	7.09	54.42	11.24	4.08	0.98	0.69	DeFusion	7.52	56.65	11.55	4.35	0.80	0.55
RFL	7.00	51.62	14.93	5.03	0.96	0.75	RFL	7.53	57.53	18.43	6.84	0.94	0.73
ZMFF	6.99	51.15	13.93	4.95	0.94	0.70	ZMFF	7.53	56.96	18.84	6.76	0.93	0.69
EPT	7.00	51.64	14.97	5.04	0.96	0.75	EPT	7.53	57.55	18.44	6.84	0.94	0.74
MGDN	7.09	54.24	15.15	5.24	1.07	0.75	MGDN	7.54	57.50	18.81	6.67	0.93	0.74
Ours	7.11	54.93	15.62	5.43	1.10	0.76	Ours	7.56	59.15	19.57	6.97	0.98	0.74

ing contrast.

5.4. Multi-focus Image Fusion

Setup. MFF experiments are conducted using RealMFF (Zhang et al., 2020a) and Lytro (Nejati et al., 2015). 639 image pairs from RealMFF are employed for training, while 71 pairs from it are reserved for testing and 20 image pairs in Lytro are utilized for generalizability test. Comparative methods encompass DIFNet (Jung et al., 2020), CUNet (Deng & Dragotti, 2021), SDNet (Zhang & Ma, 2021), U2Fusion (Xu et al., 2022a), DeFusion (Liang et al., 2022), RFL (Wang et al., 2022b), ZMFF (Hu et al., 2023), EPT (Wang et al., 2023), and MGDN (Guan et al., 2023).

Comparison with SOTA Methods. As illustrated in Figure 7, benefiting from textual descriptions, FILM excels in identifying clear regions within multi-focus image pairs, ensuring sharp foreground and background elements. The quantitative results in Table 4 further underscore the excellence of our methodology.

5.5. Ablation Studies

To explore the effectiveness of each module in our proposed method, using the infrared-visible fusion task as an example, we conduct ablation studies on the test dataset of RoadScene (Xu et al., 2020a). The results are presented in Table 5.

Textual Guidance. In Exp. I, we remove the guidance through textual information and only use image features for fusion, i.e., the cross-attention layers between text and image features are eliminated, aiming to demonstrate the effect of text-guided feature extraction and fusion in FILM. By increasing the number of Restormer blocks, we maintain the total number of parameters close to the original model.

Semantic Prompts. Then, in Exp. II-IV, we test the guiding role of text semantic prompts from holistic to fine-grained, including image caption (IC), dense caption (DC), and segment mask (SM). In Exp. II, we directly feed the source images into ChatGPT. By manually providing prompts, GPT generates overall descriptions of the images, which are used as text inputs for image fusion. This study bypassed the

Table 5: Ablation experiments results, with red resent best values.

Descriptions	Configurations				Metrics					
	Image caption	Dense caption	Segment mask	GPT	EN	SD	SF	AG	VIF	Qabf
Exp. I: w/o text					7.16	43.45	11.58	5.63	0.51	0.48
Exp. II: w/o caption				✓	7.25	45.66	11.94	5.99	0.54	0.51
Exp. III: w/o DC and SM	✓			✓	7.26	47.71	11.87	6.05	0.55	0.53
Exp. IV: w/o SM	✓	✓		✓	7.33	49.09	16.36	6.94	0.62	0.55
Exp. V: w/o GPT	✓	✓	✓		7.29	50.38	14.39	6.55	0.58	0.53
FILM (Ours)	✓	✓	✓	✓	7.43	49.25	17.34	6.60	0.69	0.62

steps involving prompts from IC, DC and SM. In Exp. III, only IC is input into GPT, whereas in Exp. IV, both IC and DC are together input into GPT, revealing the importance of different aspects of the captions from coarse-grained to fine-grained.

ChatGPT. Finally, in Exp. V, after extracting IC, DC and SM from images, we directly concatenate these three captions as the text description without inputting them into GPT. This is to demonstrate GPT's capability in integrating textual information and its effort for fusion performance.

In conclusion, ablation experiments demonstrate that relying on the comprehensive information from different grains of captions and the powerful summarization capability of GPT, our experimental setup achieved optimal fusion performance, validating the rationality of our FILM setting.

6. Conclusion

This study addresses a significant shortcoming of existing image fusion techniques: their insufficient exploitation of deeper semantic information beyond visual features. To this end, we present, for the first time, a novel paradigm called *Image Fusion via VIsion-Language Model* (**FILM**), which employs explicit textural descriptions of source images from large language models to guide and enhance the fusion process, enabling a more comprehensive understanding of image content. Furthermore, we explore the feasibility of integrating the vision-language model framework into the image fusion process. Notably, in FILM, any component within the model, such as BLIP2 or ChatGPT, is replaceable.

FILM has shown promising results on various image fusion tasks, including infrared-visible, medical, multi-exposure and multi-focus scenarios. In addition, we present a novel benchmark vision-language dataset, including ChatGPT-generated descriptions for eight image fusion datasets. We hope that our study will open up new opportunities for large-scale vision-language models in the realm of image fusion.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Bouzos, O., Andreadis, I., and Mitianoudis, N. A convolutional neural network-based conditional random field model for structured multi-focus image fusion robust to noise. *IEEE Transactions on Image Processing*, 2023.

Brooks, T., Holynski, A., and Efros, A. A. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18392–18402, 2023.

Cai, J., Gu, S., and Zhang, L. Learning a deep single image contrast enhancer from multi-exposure images. *IEEE Transactions on Image Processing*, 27(4):2049–2062, 2018.

Deng, X. and Dragotti, P. L. Deep convolutional neural network for multi-modal image restoration and fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(10):3333–3348, 2021.

Deng, X., Zhang, Y., Xu, M., Gu, S., and Duan, Y. Deep coupled feedback network for joint exposure fusion and image super-resolution. *IEEE Transactions on Image Processing*, 30:3098–3112, 2021.

Gao, F., Deng, X., Xu, M., Xu, J., and Dragotti, P. L. Multimodal convolutional dictionary learning. *IEEE Transactions on Image Processing*, 31:1325–1339, 2022.

Guan, Y., Xu, R., Yao, M., Wang, L., and Xiong, Z. Mutual-guided dynamic network for image fusion. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, pp. 1779–1788, 2023.

Guo, X., Nie, R., Cao, J., Zhou, D., Mei, L., and He, K. Fusegan: Learning to fuse multi-focus image via conditional generative adversarial network. *IEEE Transactions on Multimedia*, 21(8):1982–1996, 2019.

Han, D., Li, L., Guo, X., and Ma, J. Multi-exposure image fusion via deep perceptual enhancement. *Information Fusion*, 79:248–262, 2022.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

Hu, X., Jiang, J., Liu, X., and Ma, J. Zmff: Zero-shot multifocus image fusion. *Information Fusion*, 92:127–138, 2023.

Huang, W., Liu, Y., Qin, H., Li, Y., Zhang, S., Liu, X., Magno, M., and Qi, X. Billm: Pushing the limit of post-training quantization for llms. *arXiv preprint arXiv:2402.04291*, 2024.

James, A. P. and Dasarathy, B. V. Medical image fusion: A survey of the state of the art. *Information Fusion*, 19: 4–19, 2014.

Jiang, Z., Zhang, Z., Fan, X., and Liu, R. Towards all weather and unobstructed multi-spectral image stitching: Algorithm and benchmark. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, pp. 3783–3791, 2022.

Johnson, B. A. and Becker, J. A. Harvard medical website. http://www.med.harvard.edu/AANLIB/home.html.

Jung, H., Kim, Y., Jang, H., Ha, N., and Sohn, K. Unsupervised deep image fusion with structure tensor representations. *IEEE Transactions on Image Processing*, 29: 3845–3858, 2020.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollar, P., and Girshick, R. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4015–4026, October 2023.

Li, H. and Wu, X.-J. Densefuse: A fusion approach to infrared and visible images. *IEEE Transactions on Image Processing*, 28(5):2614–2623, 2018.

Li, H., Xu, T., Wu, X., Lu, J., and Kittler, J. Lrrnet: A novel representation learning guided fusion network for infrared and visible images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):11040–11052, 2023a.

Li, J., Li, D., Xiong, C., and Hoi, S. C. H. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 12888–12900, 2022.

Li, J., Li, D., Savarese, S., and Hoi, S. C. H. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 19730–19742, 2023b.

Li, J., Liu, J., Zhou, S., Zhang, Q., and Kasabov, N. K. Gesenet: A general semantic-guided network with couple mask ensemble for medical image fusion. *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–14, 2023c.

Li, J., Liu, J., Zhou, S., Zhang, Q., and Kasabov, N. K. Gesenet: A general semantic-guided network with couple mask ensemble for medical image fusion. *IEEE Transactions on Neural Networks and Learning Systems*, 2023d.

Liang, P., Jiang, J., Liu, X., and Ma, J. Fusion from decomposition: A self-supervised decomposition approach for image fusion. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 719–735, 2022.

Liu, J., Fan, X., Huang, Z., Wu, G., Liu, R., Zhong, W., and Luo, Z. 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)*, pp. 5792–5801, 2022a.

Liu, J., Shang, J., Liu, R., and Fan, X. Attention-guided global-local adversarial learning for detail-preserving multi-exposure image fusion. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(8):5026–5040, 2022b.

Liu, J., Liu, Z., Wu, G., Ma, L., Liu, R., Zhong, W., Luo, Z., and Fan, X. 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)*, pp. 8115–8124, October 2023a.

Liu, J., Wu, G., Luan, J., Jiang, Z., Liu, R., and Fan, X. Holoco: Holistic and local contrastive learning network for multi-exposure image fusion. *Information Fusion*, 95: 237–249, 2023b.

Liu, Y., Chen, X., Peng, H., and Wang, Z. Multi-focus image fusion with a deep convolutional neural network. *Information Fusion*, 36:191–207, 2017.

Ma, H., Liao, Q., Zhang, J., Liu, S., and Xue, J.-H. An α-matte boundary defocus model-based cascaded network for multi-focus image fusion. *IEEE Transactions on Image Processing*, 29:8668–8679, 2020a.

Ma, J., Ma, Y., and Li, C. Infrared and visible image fusion methods and applications: A survey. *Information Fusion*, 45:153–178, 2019a.

Ma, J., Yu, W., Liang, P., Li, C., and Jiang, J. Fusiongan: A generative adversarial network for infrared and visible image fusion. *Information Fusion*, 48:11–26, 2019b.

Ma, K., Li, H., Yong, H., Wang, Z., Meng, D., and Zhang, L. Robust multi-exposure image fusion: A structural patch decomposition approach. *IEEE Transactions on Image Processing*, 26(5):2519–2532, 2017.

Ma, K., Duanmu, Z., Zhu, H., Fang, Y., and Wang, Z. Deep guided learning for fast multi-exposure image fusion. *IEEE Transactions on Image Processing*, 2020b.

Nejati, M., Samavi, S., and Shirani, S. Multi-focus image fusion using dictionary-based sparse representation. *Information Fusion*, 25:72–84, 2015.

Nguyen, V.-Q., Suganuma, M., and Okatani, T. Grit: Faster and better image captioning transformer using dual visual features. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 167–184. Springer, 2022.

OpenAI. Gpt-4 technical report. *ArXiv*, abs/2303.08774, 2023. URL https://arxiv.org/abs/2303.08774.

OpenAI. *ChatGPT*, 2023. URL https://www.openai.com/chatgpt.

Prabhakar, K. R., Srikar, V. S., and Babu, R. V. Deepfuse: A deep unsupervised approach for exposure fusion with extreme exposure image pairs. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 4724–4732, 2017.

Qin, H., Ma, X., Zheng, X., Li, X., Zhang, Y., Liu, S., Luo, J., Liu, X., and Magno, M. Accurate lora-finetuning quantization of llms via information retention. *arXiv preprint arXiv:2402.05445*, 2024.

Qu, L., Liu, S., Wang, M., and Song, Z. Transmef: A transformer-based multi-exposure image fusion framework using self-supervised multi-task learning. In *Proceedings of the AAAI conference on artificial intelligence (AAAI)*, volume 36, pp. 2126–2134, 2022.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *Proceedings of the International conference on machine learning (ICML)*, pp. 8748–8763, 2021.

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., and Sutskever, I. Zero-shot text-to-image generation. In *Proceedings of the International conference on machine learning (ICML)*, pp. 8821–8831, 2021.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

Sun, Y., Cao, B., Zhu, P., and Hu, Q. Detfusion: A detection-driven infrared and visible image fusion network. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, pp. 4003–4011, 2022.

Tang, L., Deng, Y., Ma, Y., Huang, J., and Ma, J. Superfusion: A versatile image registration and fusion network with semantic awareness. *IEEE/CAA Journal of Automatica Sinica*, 9(12):2121–2137, 2022a.

Tang, L., Yuan, J., and Ma, J. 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, 2022b.

Tang, L., Yuan, J., Zhang, H., Jiang, X., and Ma, J. Piafusion: A progressive infrared and visible image fusion network based on illumination aware. *Information Fusion*, 83-84:79–92, 2022c.

Tang, W., He, F., Liu, Y., and Duan, Y. Matr: Multimodal medical image fusion via multiscale adaptive transformer. *IEEE Transactions on Image Processing*, 31:5134–5149, 2022d.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Wang, D., Liu, J., Fan, X., and Liu, R. Unsupervised misaligned infrared and visible image fusion via cross-modality image generation and registration. In *Proceedings of the International Joint Conferences on Artificial Intelligence (IJCAI)*, pp. 3508–3515, 2022a.

Wang, Z., Yu, J., Yu, A. W., Dai, Z., Tsvetkov, Y., and Cao, Y. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*, 2021.

Wang, Z., Li, X., Duan, H., and Zhang, X. A self-supervised residual feature learning model for multifocus image fusion. *IEEE Transactions on Image Processing*, 31:4527–4542, 2022b.

Wang, Z., Li, X., Zhao, L., Duan, H., Wang, S., Liu, H., and Zhang, X. When multi-focus image fusion networks meet traditional edge-preservation technology. *International Journal of Computer Vision*, pp. 1–24, 2023.

Wen, J., Qin, F., Du, J., Fang, M., Wei, X., Chen, C. P., and Li, P. Msgfusion: Medical semantic guided two-branch network for multimodal brain image fusion. *IEEE Transactions on Multimedia*, 2023.

Xiao, B., Xu, B., Bi, X., and Li, W. Global-feature encoding u-net (geu-net) for multi-focus image fusion. *IEEE Transactions on Image Processing*, 30:163–175, 2020.

Xiao, B., Wu, H., and Bi, X. Dtmnet: a discrete tchebichef moments-based deep neural network for multi-focus image fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 43–51, 2021.

Xu, H. and Ma, J. Emfusion: An unsupervised enhanced medical image fusion network. *Information Fusion*, 76:177–186, 2021.

Xu, H., Ma, J., Le, Z., Jiang, J., and Guo, X. Fusiondn: A unified densely connected network for image fusion. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pp. 12484–12491, 2020a.

Xu, H., Ma, J., and Zhang, X.-P. Mef-gan: Multi-exposure image fusion via generative adversarial networks. *IEEE Transactions on Image Processing*, 29:7203–7216, 2020b.

Xu, H., Ma, J., Jiang, J., Guo, X., and Ling, H. U2fusion: A unified unsupervised image fusion network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1):502–518, 2022a.

Xu, H., Ma, J., Yuan, J., Le, Z., and Liu, W. 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)*, pp. 19647–19656, 2022b.

Xu, H., Yuan, J., and Ma, J. MURF: mutually reinforcing multi-modal image registration and fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):12148–12166, 2023.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., and Bengio, Y. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the International conference on machine learning (ICML)*, pp. 2048–2057, 2015.

Xu, S., Ji, L., Wang, Z., Li, P., Sun, K., Zhang, C., and Zhang, J. Towards reducing severe defocus spread effects for multi-focus image fusion via an optimization based strategy. *IEEE Transactions on Computational Imaging*, 6:1561–1570, 2020c.

Xu, S., Zhang, J., Zhao, Z., Sun, K., Liu, J., and Zhang, C. Deep gradient projection networks for pan-sharpening. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1366–1375, 2021.

Yin, J.-L., Chen, B.-H., and Peng, Y.-T. Two exposure fusion using prior-aware generative adversarial network. *IEEE Transactions on Multimedia*, 24:2841–2851, 2021a.

Yin, J.-L., Chen, B.-H., Peng, Y.-T., and Hwang, H. Automatic intermediate generation with deep reinforcement learning for robust two-exposure image fusion. *IEEE Transactions on Neural Networks and Learning Systems*, 33(12):7853–7862, 2021b.

Zamir, S. W., Arora, A., Khan, S., Hayat, M., Khan, F. S., and Yang, M. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5718–5729, 2022.

Zhang, H. and Ma, J. Sdnet: A versatile squeeze-and-decomposition network for real-time image fusion. *International Journal of Computer Vision*, 129(10):2761–2785, 2021.

Zhang, J., Liao, Q., Liu, S., Ma, H., Yang, W., and Xue, J.-H. Real-mff: A large realistic multi-focus image dataset with ground truth. *Pattern Recognition Letters*, 138:370–377, 2020a.

Zhang, S., Sun, P., Chen, S., Xiao, M., Shao, W., Zhang, W., Chen, K., and Luo, P. Gpt4roi: Instruction tuning large language model on region-of-interest. *arXiv preprint arXiv:2307.03601*, 2023.

Zhang, X. Benchmarking and comparing multi-exposure image fusion algorithms. *Information Fusion*, 74:111–131, 2021a.

Zhang, X. Deep learning-based multi-focus image fusion: A survey and a comparative study. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):4819–4838, 2021b.

Zhang, X. and Demiris, Y. Visible and infrared image fusion using deep learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8):10535–10554, 2023.

Zhang, Y., Liu, Y., Sun, P., Yan, H., Zhao, X., and Zhang, L. IFCNN: A general image fusion framework based on convolutional neural network. *Information Fusion*, 54: 99–118, 2020b.

Zhao, H. Image.txt: Transform image into unique paragraph. https://zhaohengyuan1.github.io/image2paragraph.github.io/, 2023.

Zhao, W., Xie, S., Zhao, F., He, Y., and 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)*, pp. 13955–13965. IEEE, 2023a.

Zhao, Z., Xu, S., Zhang, C., Liu, J., Zhang, J., and Li, P. DIDFuse: Deep image decomposition for infrared and visible image fusion. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 970–976, 2020.

Zhao, Z., Xu, S., Zhang, J., Liang, C., Zhang, C., and Liu, J. Efficient and model-based infrared and visible image fusion via algorithm unrolling. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(3):1186–1196, 2022a.

Zhao, Z., Zhang, J., Xu, S., Lin, Z., and Pfister, H. Discrete cosine transform network for guided depth map super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5687–5697, 2022b.

Zhao, Z., Bai, H., Zhang, J., Zhang, Y., Xu, S., Lin, Z., Timofte, R., and Van Gool, L. Cddfuse: Correlation-driven dual-branch feature decomposition for multi-modality image fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5906–5916, June 2023b.

Zhao, Z., Bai, H., Zhu, Y., Zhang, J., Xu, S., Zhang, Y., Zhang, K., Meng, D., Timofte, R., and Van Gool, L. Ddfm: Denoising diffusion model for multi-modality image fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8082–8093, October 2023c.

Zhu, D., Chen, J., Shen, X., Li, X., and Elhoseiny, M. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

A. More Visualization Results for VLF dataset

More visualization results for the VLF dataset are displayed in Figures 8 and 9.

B. Detailed Illustration to Datasets

We adopt widely-used benchmarks MSRS (Tang et al., 2022c), M^{3}FD (Liu et al., 2022a) and RoadScene (Xu et al., 2020a) for *Infrared-Visible image Fusion* (IVF), Harvard Medical dataset (Johnson & Becker) for *Medical Image Fusion* (MIF), SICE (Cai et al., 2018) and MEFB (Zhang, 2021a) dataset for *Multi-exposure Image Fusion* (MEF), as well as RealMFF (Zhang et al., 2020a) and Lytro (Nejati et al., 2015) dataset for *Multi-focus Image Fusion* (MFF), respectively.

- MSRS dataset[1]: 1083 pairs for IVF training and 361 pairs for IVF testing.

- M^{3}FD dataset[2]: 100 pairs for IVF testing.

- RoadScene dataset[3]: 70 pairs for IVF validation, 70 pairs for IVF testing.

- Harvard Medical Image dataset[4]: 50 pairs for MIF testing.

- SICE dataset[5]: 499 pairs for MEF training and 90 pairs MEF testing.

- MEFB dataset[6]: 40 pairs for MEF testing.

- RealMFF dataset[7]: 639 pairs for MFF training and 71 pairs for MFF testing.

- Lytro dataset[8]: 20 pairs for testing.

C. More Qualitative Comparison Fusion Results

- More qualitative comparisons for *Infrared-Visible image Fusion* are shown in Figure 10.

- More qualitative comparisons for *Medical Image Fusion* are shown in Figure 11.

- More qualitative comparisons for *Multi-exposure Image Fusion* are shown in Figure 12.

- More qualitative comparisons for *Multi-focus Image Fusion* are shown in Figure 13.

---

[1]https://github.com/Linfeng-Tang/MSRS
[2]https://github.com/JinyuanLiu-CV/TarDAL
[3]https://github.com/hanna-xu/RoadScene
[4]http://www.med.harvard.edu/AANLIB/home.html
[5]https://github.com/csjcai/SICE
[6]https://github.com/xingchenzhang/MEFB
[7]https://github.com/Zancelot/Real-MFF
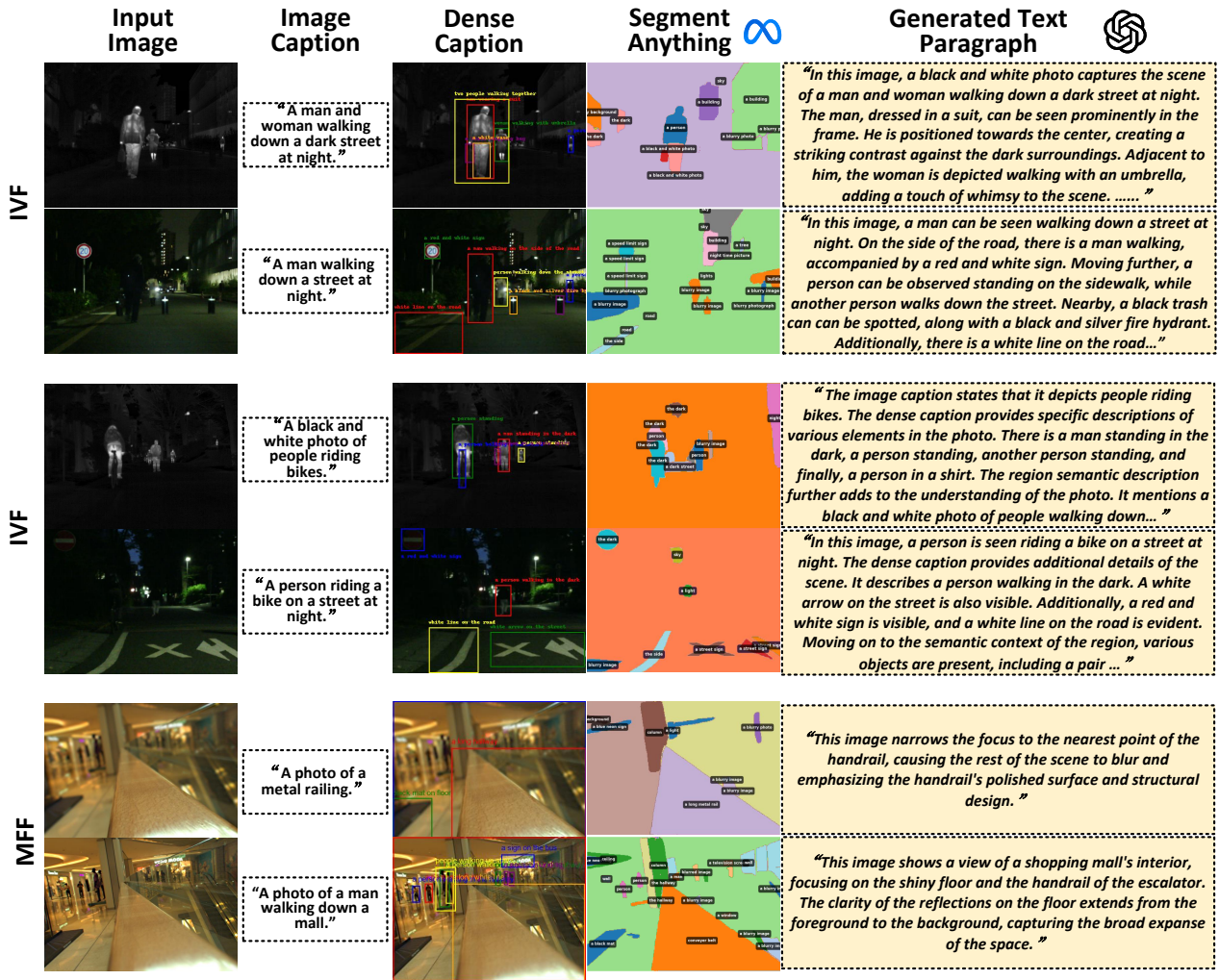[8]http://mansournejati.ece.iut.ac.ir/content/lytro-multi-focus-dataset

Figure 8: More visualization results for the VLF dataset on IVF and MFF.
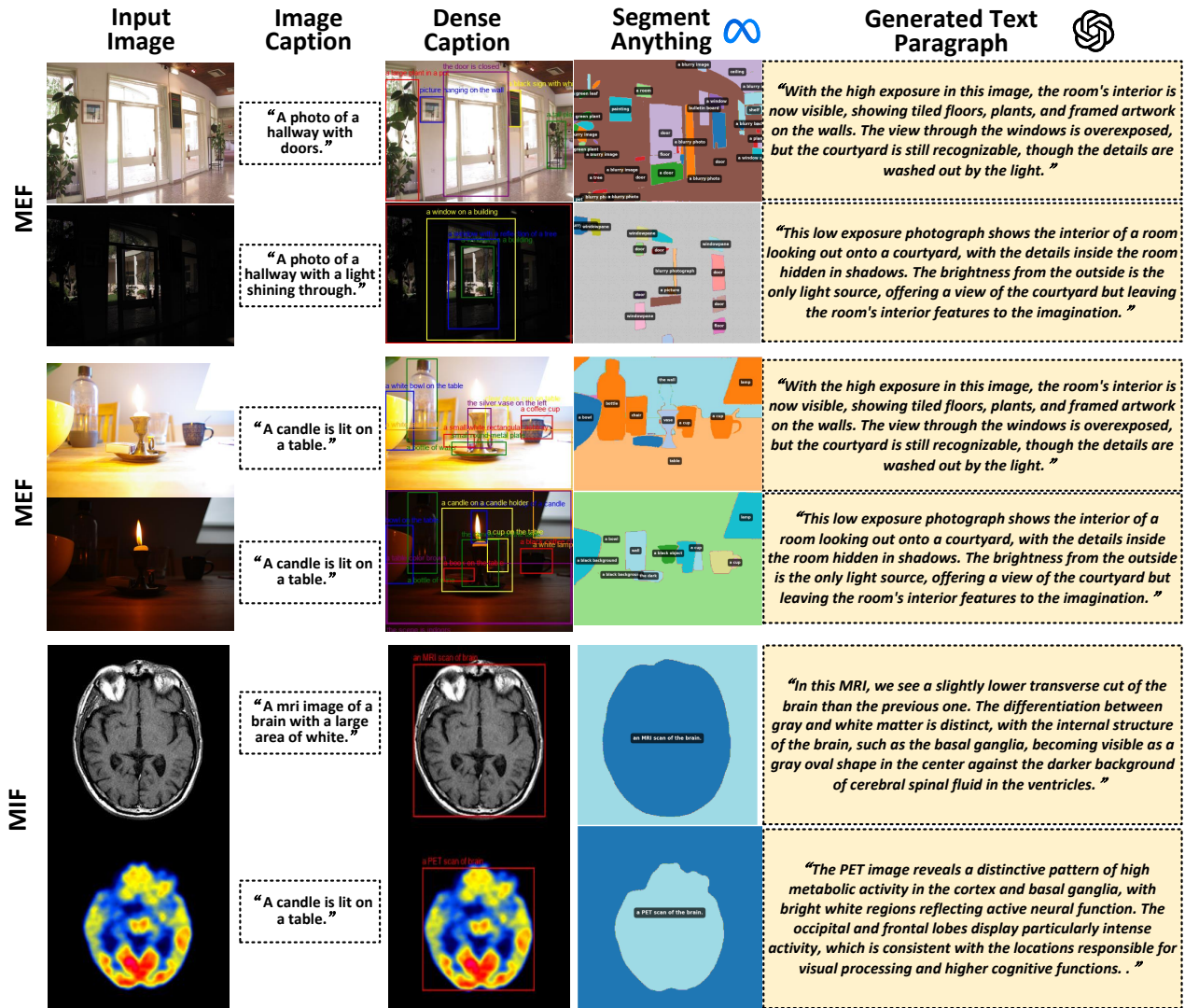
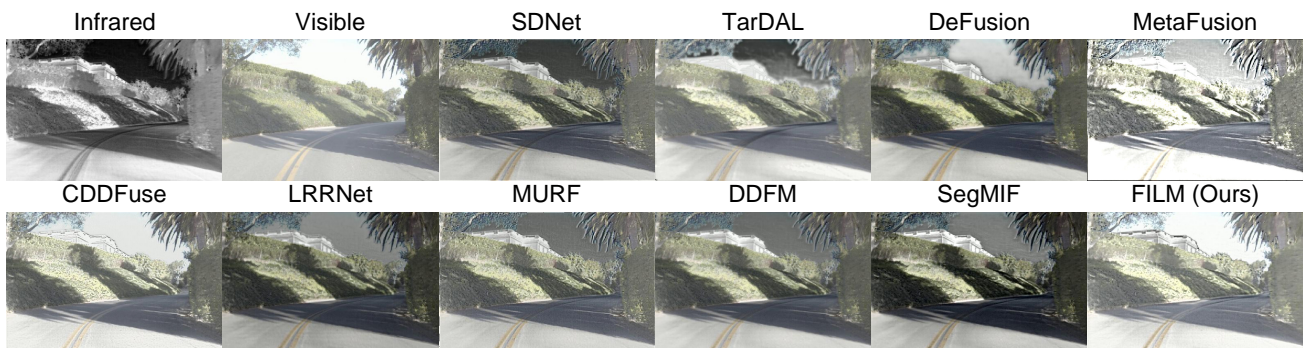Figure 9: More visualization results for the VLF dataset on MEF and MIF.

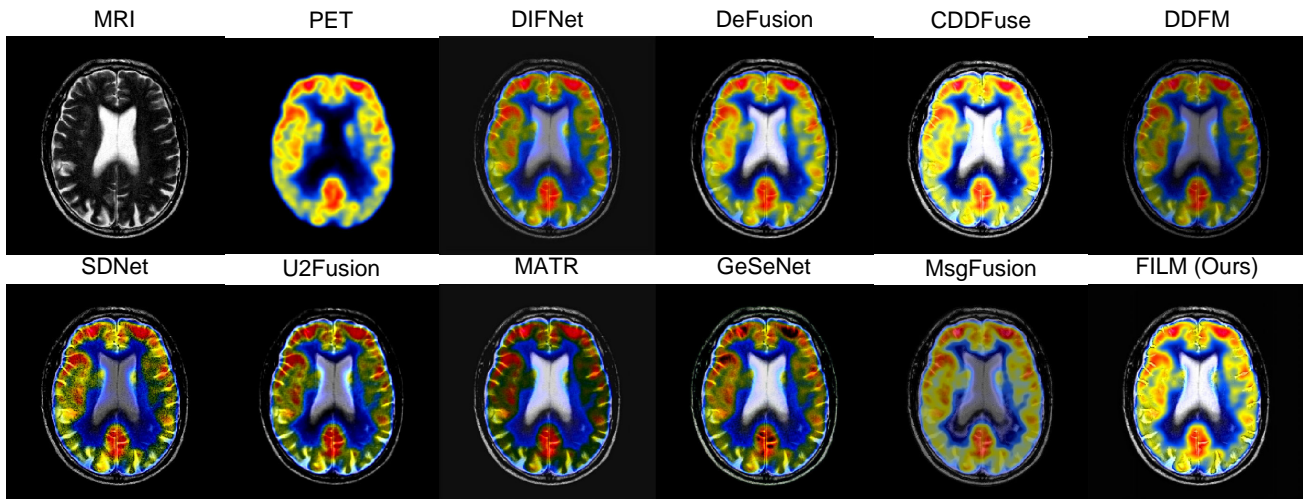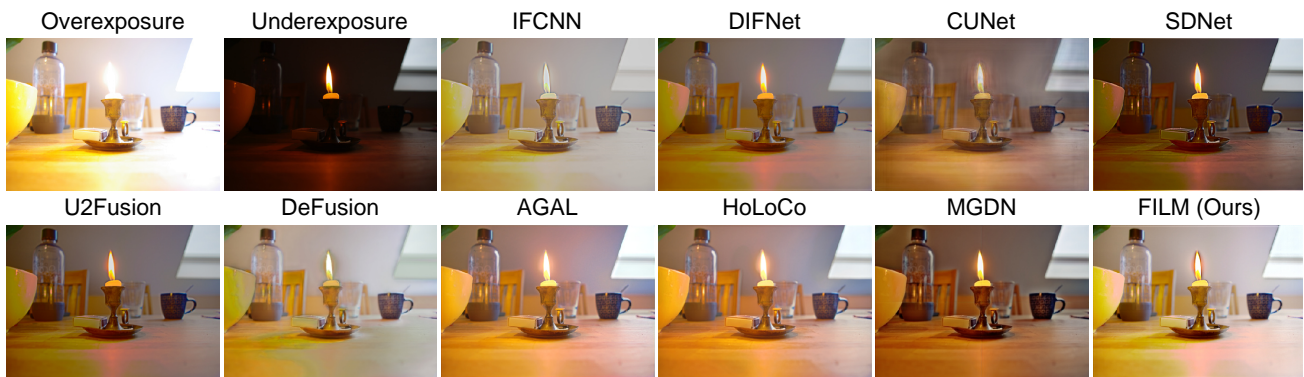Figure 10: Visualization comparison of the fusion results in the infrared-visible image fusion task.

Figure 11: Visualization comparison of the fusion results in the medical image fusion task.

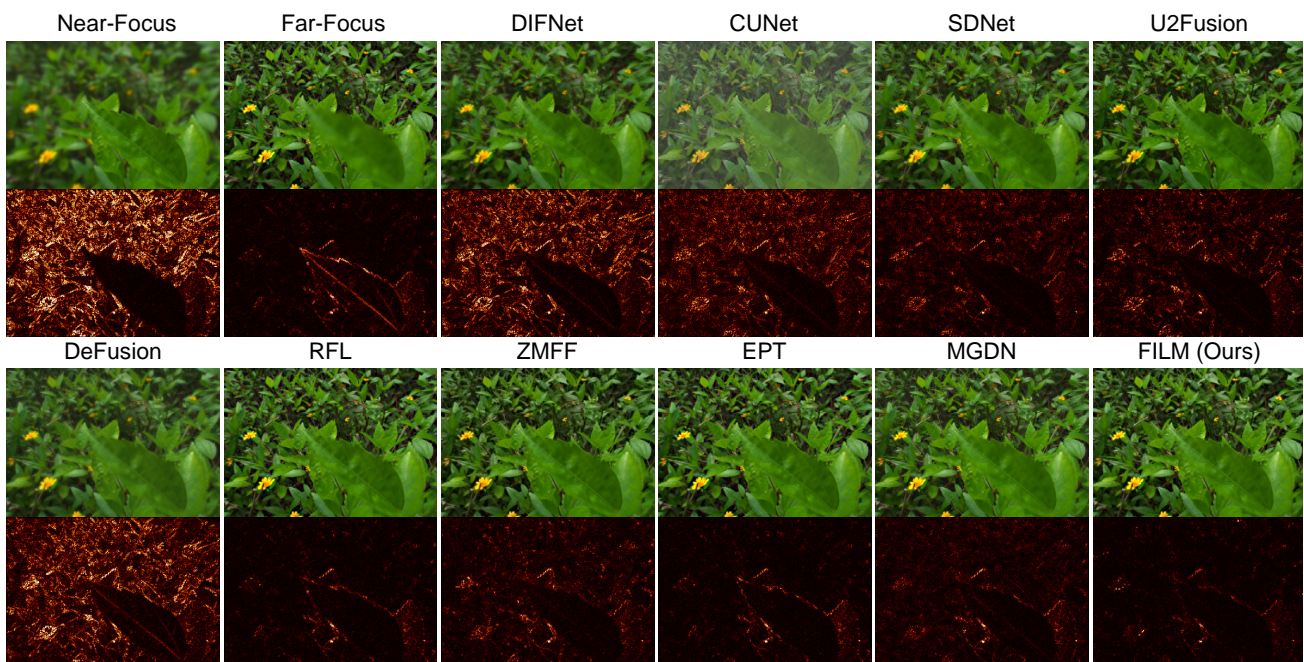Figure 12: Visualization comparison of the fusion results in the multi-exposure image fusion task.

Figure 13: Visualization comparison of the fusion results and error maps in the multi-focus image fusion task.