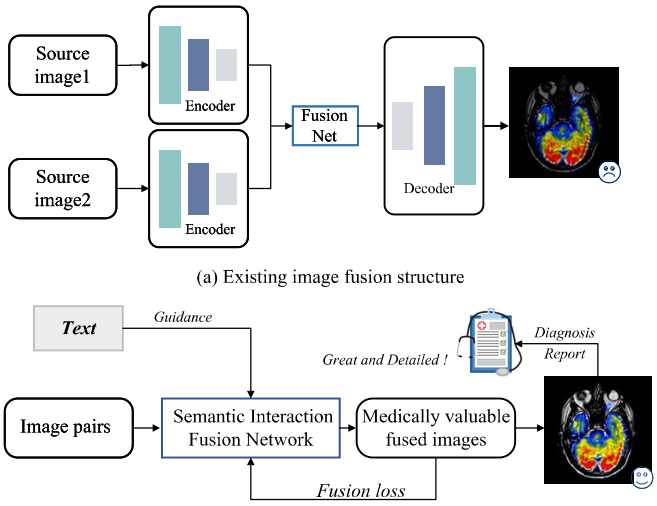
SMFusion: Semantic-Preserving Fusion of Multimodal Medical Images for Enhanced Clinical Diagnosis

Haozhe Xiang, Han Zhang, Yu Cheng, Xiongwen Quan, Wanwan Huang

arXiv:2505.12251v1 [cs.CV] 18 May 2025

***Abstract*—Multimodal medical image fusion plays a crucial role in medical diagnosis by integrating complementary information from different modalities to enhance image readability and clinical applicability. However, existing methods mainly follow computer vision standards for feature extraction and fusion strategy formulation, overlooking the rich semantic information inherent in medical images. To address this limitation, we propose a novel semantic-guided medical image fusion approach that, for the first time, incorporates medical prior knowledge into the fusion process. Specifically, we construct a publicly available multimodal medical image-text dataset, upon which text descriptions generated by BiomedGPT are encoded and semantically aligned with image features in a high-dimensional space via a semantic interaction alignment module. During this process, a cross attention based linear transformation automatically maps the relationship between textual and visual features to facilitate comprehensive learning. The aligned features are then embedded into a text-injection module for further feature-level fusion. Unlike traditional methods, we further generate diagnostic reports from the fused images to assess the preservation of medical information. Additionally, we design a medical semantic loss function to enhance the retention of textual cues from the source images. Experimental results on test datasets demonstrate that the proposed method achieves superior performance in both qualitative and quantitative evaluations while preserving more critical medical information.**

***Index Terms*—Image fusion, medical prior knowledge, semantic alignment, diagnostic reports.**



(a) Existing image fusion structure

(b) Overview of proposed method

Fig. 1. Comparison of Existing Medical Image Fusion Methods and Our Proposed Approach. (a) Existing fusion methods: These methods are primarily based on deep learning networks and integrate information from multiple source images through unsupervised learning strategies. (b) Proposed Semantic-Guided Image Fusion: Our approach integrates medical prior knowledge to explicitly guide the fusion process, establishing an interactive framework that ensures high-quality fusion results.

I. Introduction

MEDICAL imaging plays a crucial role in modern healthcare, facilitating disease diagnosis, treatment planning, and surgical navigation. Due to differences in imaging principles, various imaging modalities provide complementary information. For instance, computed tomography (CT) offers high-resolution details of bones and dense tissues, whereas magnetic resonance imaging (MRI) provides high-contrast visualization of muscles, nerves, and soft tissues. Functional imaging techniques such as positron emission tomography (PET) and single-photon emission computed tomography (SPECT) reveal metabolic activities and blood flow distribution, making them valuable for tumor detection and neurological disease diagnosis[1]. However, each imaging modality has inherent limitations. In clinical practice, the information captured in a single imaging modality is often insufficient for comprehensive diagnosis and treatment planning[2]. Each modality captures distinct but complementary features, which requires the integration of multiple sources to achieve a more comprehensive and accurate representation of the patient's condition. Medical image fusion addresses this need by combining information from various imaging modalities into a single, more informative image. By leveraging fusion techniques, clinicians can access structural, functional, and even molecular information simultaneously, facilitating improved diagnosis, treatment planning, and intraoperative guidance. The fused image must preserve critical anatomical details while retaining complementary modality-specific features, ensuring that no clinically relevant information is lost.

Multimodal Medical Image Fusion (MMIF) techniques can be broadly categorized into traditional approaches and deep learning-based methods. Traditional techniques, including spatial domain methods, transform domain approaches, sparse representation models, and hybrid strategies, have been extensively applied in multimodal image integration. Spatial domain methods, such as pixel-level weighted averaging[3] and Principal Component Analysis (PCA)[4], offer computational simplicity but often result in blurred boundaries and the loss of essential anatomical details due to their limited capacity to model spatial dependencies.Transform domain approaches,

This paper is supported by National Natural Science Foundation of China 62403194.(*Corresponding author: Wanwan Huang.*)

Haozhe Xiang and Wanwan Huang are with the College of Information and Intelligence, Hunan Agricultural University, Changsha 410128, China (e-mail: haozhexiang@163.com; hnnydx1993@163.com).

Han Zhang, Yu Cheng and Xiongwen Quan are with the College of Artificial Intelligence, Nankai University, Tianjin 300071, China (e-mail: zhanghan@nankai.edu.cn; yucheng@mail.nankai.edu.cn; quanxw@nankai.edu.cn)

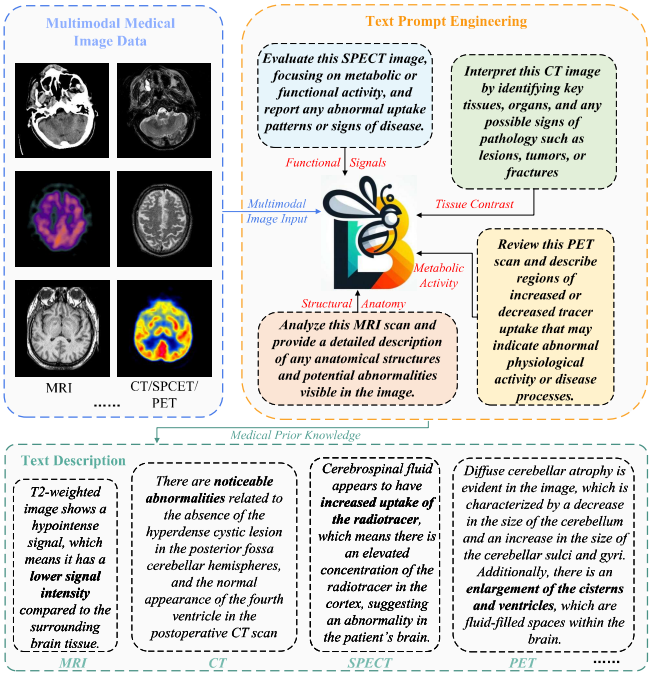


Fig. 2. Schematic illustration of multimodal medical image-text dataset construction. For each imaging modality, tailored text prompts are provided to guide BiomedGPT in generating corresponding descriptions.

including Wavelet Transform (WT)[5], curvelet transform [6], and Nonsubsampled Shearlet Transform (NSST)[7], improve fusion by enabling multi-scale and multi-directional analysis. However, their reliance on manually designed decomposition frameworks and heuristic fusion rules limits their flexibility and robustness across different clinical contexts. Sparse representation methods[8] aim to enhance detail preservation and noise robustness by learning sparse coefficients over overcomplete dictionaries. Despite their effectiveness, these methods typically involve high computational costs and are sensitive to the choice of dictionary. Hybrid strategies[9] combine multiple paradigms to capitalize on their respective advantages, but this often introduces additional complexity and parameter tuning challenges, reducing practical applicability in real-time scenarios. In contrast, deep learning-based methods offer a data-driven framework for end-to-end feature extraction and adaptive fusion. Convolutional Neural Networks (CNNs)[10] have been widely adopted to extract hierarchical representations from source images, capturing both low-level textures and high-level semantic information. Nonetheless, CNNs often face difficulties in modeling long-range dependencies and global contextual relationships. Transformer-based architectures[11] address these limitations by employing self-attention mechanisms to capture multi-scale cross-modal interactions, thereby improving anatomical alignment and semantic consistency. And hybrid architectures that integrate CNNs and Transformers[12] have been explored to balance local feature representation and global contextual modeling, offering a more comprehensive fusion strategy. Additionally, GAN-based fusion methods[13] have been proposed to improve perceptual quality through adversarial learning, but training instability remains a significant hurdle. As illustrated in fig.1(a), most current deep learning-based fusion frameworks adopt a sequential pipeline in which the registered source images are passed through an encoder, a fusion network, and a decoder to produce the final fused result. While this approach often achieves high performance on standard visual evaluation metrics due to its reliance on learned visual features, it may still suffer from problems such as brightness stacking and the loss of fine details[14], which can adversely affect the reliability of clinical diagnosis.

Despite significant progress made by deep learning-based fusion methods, these approaches often overlook the intrinsic semantic information inherent in medical images. Unlike natural images, where pixel intensities typically correspond to intuitive visual features, the brightness values in different medical imaging modalities convey fundamentally distinct clinical meanings[15]. Consequently, simple channel concatenation fails to capture meaningful cross-modal correlations and may result in fused images with limited interpretability. Moreover, the evaluation of fusion performance still largely relies on conventional computer vision metrics, which inadequately reflect clinical diagnostic standards and real-world applicability. Although recent studies have attempted to incorporate textual descriptions to enhance semantic awareness in medical image fusion, their effectiveness remains limited[16]. This limitation primarily arises from the reliance on automatically generated captions by large language models such as ChatGPT. These descriptions often focus only on superficial visual attributes, such as describing "*a bright circular region*" or "*a darker surrounding area*", without interpreting the underlying pathological significance, for example, "*a hyperintense lesion indicative of possible edema*" or "*a hypodense region suggesting chronic infarction.*" In addition, they lack integration of domain-specific medical knowledge, including anatomical terminology, radiological pattern recognition, or context-aware diagnostic reasoning, which are essential for producing clinically meaningful fused outputs. As a result, the overall quality of the fusion becomes overly dependent on these shallow textual cues, significantly undermining the model's generalizability and applicability across diverse and complex clinical scenarios.

To address the limitations of existing medical image fusion methods, this paper proposes a novel semantic-guided medical image fusion framework. Specifically, we integrate biomedical artificial intelligence models to generate descriptive texts aligned with medical images. Among these models, BiomedGPT exhibits human-level performance in generating complex radiology reports, achieving an average error rate of only 8.3%, and has received expert-level preference ratings[17]. As illustrated in Fig. 2, we design modality-specific prompt engineering strategies to generate tailored textual prompts for different input modalities. These prompts are then fed into BiomedGPT to produce high-quality textual descriptions, which effectively guide the fusion network to attend to clinically relevant regions from multiple perspectives. This design significantly enhances both the semantic richness and clinical interpretability of the fused images. The SMFusion framework comprises three core components: text-image encoding, semantic interaction alignment, and a

text injection module. Each source image is paired with its corresponding expert-level textual description and jointly embedded into a high-dimensional feature space. Semantic information is integrated into the fusion process through a text injection module based on affine transformations. Finally, BiomedGPT is employed to generate diagnostic reports from the fused images, enabling assessment of improvements in lesion depiction and localization accuracy. To ensure the preservation of fine-grained anatomical details, we introduce a medical semantic loss that encourages the fused output to retain clinically important features from the source images. Experimental results demonstrate that the proposed SMFusion framework outperforms seven state-of-the-art fusion methods in terms of both fusion quality and clinical relevance. In summary, the primary contributions of this paper can be summarized as follows:

- We propose a novel semantic-guided medical image fusion framework to address the low clinical utility of existing fused medical images. To the best of our knowledge, this is the first work to explicitly focus on enhancing the practical diagnostic value of fused medical images. Our framework not only generates diagnostically valuable fusion outputs but also produces corresponding diagnostic reports, introducing a new paradigm for quantifying the clinical applicability of fused images.
- We construct a multimodal medical image–text dataset tailored for the medical image fusion task, facilitating future research in diagnostic report generation and clinical value quantification for fused images. In this study, the dataset is designed specifically for fusion tasks involving multimodal medical images; in future work, we aim to extend it to broader multimodal medical scenarios. The dataset will be publicly released on Github soon.
- We introduce a semantic interaction alignment module and a medical semantic loss function to guide the generation of semantically enriched fused medical images. Our approach not only preserves key diagnostic information from the source modalities but also achieves superior performance in terms of brightness, detail preservation, and overall image quality.
- Experimental results demonstrate that SMFusion achieves superior performance in both qualitative and quantitative evaluations. Moreover, it is capable of generating detailed diagnostic reports from fused images, offering a novel paradigm for evaluating the practical value of medical image fusion methods.

II. Related Work

A. Multimodal Medical Image Fusion Methods

The development of multimodal medical image fusion has progressed from handcrafted fusion rules to deep learning-based frameworks that emphasize semantic adaptability and data-driven representation learning. While traditional methods established the foundations for spatial and transform domain fusion, deep learning approaches have introduced greater flexibility in capturing the heterogeneous characteristics of medical modalities. Early deep fusion models predominantly utilized Convolutional Neural Networks (CNNs) to extract hierarchical features and perform pixel-level or feature-level fusion. Representative methods such as IFCNN[10] demonstrated the ability to preserve structural details, but often failed to capture long-range dependencies and modality-specific semantics. Transformer-based architectures SwinFusion[11], have been explored to model global context via self-attention mechanisms. However, these models are typically computationally intensive and may still struggle with fine-grained spatial accuracy. Hybrid frameworks combining CNNs and Transformers[12] aim to balance local texture preservation and global semantic integration, yet their deployment is often constrained by architectural complexity and interpretability concerns. Attention mechanisms and Generative Adversarial Networks (GANs) have also been introduced to refine modality-aware feature selection and enhance perceptual realism. Nevertheless, issues such as unstable training, semantic inconsistency, and inadequate cross-modal interaction persist in many existing methods.

Although existing models have achieved significant improvements in structure preservation and perceptual fidelity, most still rely solely on visual cues and lack effective integration of high-level semantic knowledge. This limitation motivates the incorporation of external vision-language pre-trained models to enhance the interpretability and clinical relevance of fused medical images. Unlike current approaches, our method introduces textual prompts, enabling the model to focus more on regions of interest. This not only improves lesion identification in specific areas but also aligns the model's reasoning process more closely with clinical needs.

B. Generative Pre-trained Models

In recent years, generative pre-trained models (GPMs) have emerged as a transformative paradigm in vision-language research[18]. By jointly modeling visual and textual modalities, these models achieve a unified semantic representation space, enabling cross-modal reasoning and generation. Unlike traditional models that rely on task-specific supervision, GPMs are typically trained on large-scale image-text pairs in a self-supervised or weakly supervised manner, demonstrating strong generalization across downstream tasks. Multimodal models such as CLIP[19] and BLIP[20] bridge vision and language through contrastive learning and vision-language alignment, effectively capturing high-level semantic correspondences. On the generative side, GPT-4[21] have demonstrated the ability to synthesize realistic images from textual prompts, extending multimodal interaction and complex scene understanding. However, despite these advances, a key limitation of large-scale models is their constrained understanding of domain-specific knowledge, particularly in specialized fields like medicine. Their generalist nature often lacks the depth required to fully grasp the intricacies of medical imagery and terminology, limiting their effectiveness in clinical applications.These challenges underscore the potential of using rich textual descriptions not only for tasks like captioning and retrieval but also for guiding fine-grained visual processes. BiomedGPT addresses this issue by integrating multimodal biomedical data

to generate diagnostic-level textual descriptions from medical images[17]. Its ability to process both visual and textual data makes it particularly well-suited for medical image fusion, where precise semantic alignment is crucial.

Despite the success of these models, the application of generative pre-trained models to medical image fusion remains underexplored. Most existing approaches lack explicit semantic guidance during the fusion process, limiting the clinical interpretability of the results. Our work leverages BiomedGPT to provide context-aware, diagnostic-level textual prompts, thereby enhancing the semantic fidelity and practical utility of fused medical images.

C. Text Prompt Learning

Text prompt learning has emerged as a pivotal technique in the field of vision-language models (VLMs)[22], leveraging textual information to guide and enhance various visual tasks. In the domain of image fusion, text prompt learning has been employed to guide the fusion process, resulting in more semantically controlled and interpretable outcomes. For example, Zhao et al. [16] introduced FILM, a novel fusion algorithm that extracts explicit textual information from source images to guide the fusion process, thereby enhancing feature extraction and contextual understanding. Additionally, Liu et al. [23] proposed PromptFusion, a semantic-guided fusion approach that employs positive and negative prompts to improve the model's ability to recognize and distinguish targets in multi-modality images. These approaches have demonstrated promising results in a range of image fusion tasks, including infrared-visible, medical, multi-exposure, and multi-focus image fusion. Integrating text prompt learning into image fusion not only improves the semantic coherence of fused images but also enhances the performance of downstream tasks such as object detection and segmentation. For instance, Cheng et al. [24] introduced TextFusion, which leverages a vision-language model to establish a coarse-to-fine association mechanism between text and image signals, enabling controllable image fusion. Experimental results indicate that this approach outperforms traditional appearance-based fusion methods in terms of both visual quality and task-specific performance. Furthermore, Yi et al. [25] developed Text-IF, a text-guided image fusion model that addresses issues of image degradation and non-interactivity in existing fusion methods. By leveraging semantic text guidance, this method facilitates interactive and high-quality fusion results. Overall, text prompt learning holds great potential for enhancing the performance of vision-language models and guiding image fusion tasks.

III. Method

A. Problem Formulation

In most image fusion networks the input consists of two source images (e.g., I_1, I_2), and the output tends to produce a fixed final fusion result (e.g.I^f)[26]. This network structure $\mathcal{T}_n$, focusing on the fusion of feature dimensions can be described as:

$$I^f = \mathcal{F}(I_1, I_2; \mathcal{T}_n). \tag{1}$$

Current fusion paradigms predominantly focus on visual-level representations in medical imaging. Although existing deep learning methodologies demonstrate superior performance metrics in quantitative evaluations, they exhibit significant limitations in comprehending high-level semantic information. This deficiency fundamentally conflicts with the clinical imperative to enhance the interpretability of medical imaging diagnostics. We propose an expert knowledge-based semantic-guided fusion mechanism that establishes hierarchical correlation mappings between diagnostic text descriptions and visual features. The diagnostic texts are generated by the advanced vision-language model BiomedGPT[17]. This strategy not only augments the multi-dimensional information representation capacity of fused images but also maintains exceptional visual fidelity metrics.The multimodal fusion task incorporating textual prompts can be formally defined as:

$$I^f = \mathcal{F}_{prompt}(I_1, I_2, T_{text}; \mathcal{T}_{n-t}). \tag{2}$$

This text-guided controllable image fusion method integrates adjustable textual semantic prompts into the original feature mapping function $\mathcal{F}(\cdot)$, thereby constructing an enhanced mapping function $\mathcal{F}_{\text{prompt}}(\cdot)$. In particular, $\mathcal{T}_{n-t}$ denotes the image fusion network incorporating medical prior knowledge, and T_{text} represents the corresponding textual descriptions used to guide the fusion process. This dynamic text-guidance mechanism enables explicit regulation of semantic spatial distribution in feature fusion, achieving multi-granularity control over the image fusion process.

B. Image Fusion Pipeline

1) Encoder: multimodal medical images contain abundant anatomical details and pathological characteristics, posing significant challenges for deep semantic extraction. Although Zhao et al.[26] attempted to employ ChatGPT for medical image captioning, the outputs were limited to superficial descriptions of apparent features. Building upon the biomedical AI model BiomedGPT[17], and leveraging the authoritative Harvard Medical Image Dataset, we construct professional-grade multimodal text-images dataset. The implementation workflow proceeds as follows: For input image pairs I_1, I_2, BiomedGPT generates precisely aligned pathological descriptions $\{T_1, T_2\}$. These textual outputs are then fed into a parameter-freezing CLIP text encoder to obtain semantic feature vectors $\left\{\phi_1^T, \phi_2^T\right\}$. Regarding image encoding, to address the spatial feature extraction requirements of multimodal medical images, we introduce dual-layer Restormer blocks to construct the feature extraction network. This architecture efficiently decouples deep semantic information from spatial topological features through the synergistic interaction of multi-head transposed attention and gated feed-forward networks[27]. This process can be stated as follows:

$$F_1 = \mathcal{F}_v^1(I_1), F_2 = \mathcal{F}_v^2(I_2), \tag{3}$$

where $I_1 \in \mathbb{R}^{H\times W\times 1}$ and $I_2 \in \mathbb{R}^{H\times W\times 3}$ denote input images from two distinct modalities. H and W represent the height and width of the images, respectively. $\mathcal{F}_v^1$ and $\mathcal{F}_v^2$ are the encoders applied to I_1 and I_2 to extract the visual features, denoted as F_1 and F_2, from each modality.

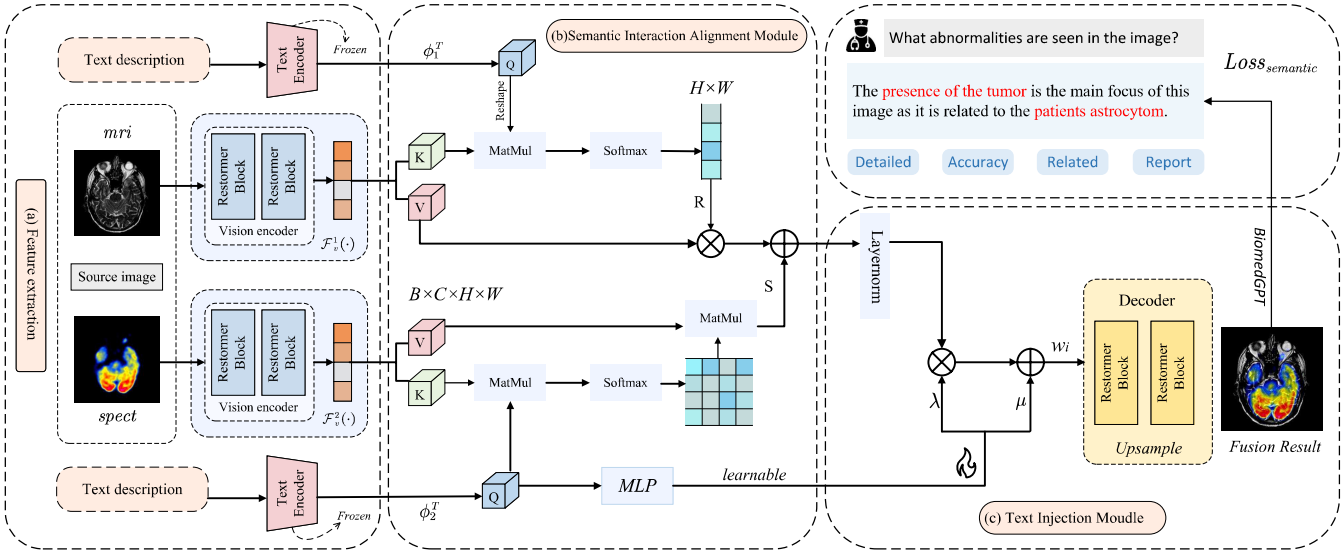


Fig. 3. The workflow of SMFusion. It consists of three key components: the feature extraction module (a), the semantic interaction alignment module (b), and the text injection module (c). The encoded text features are guided through the L-layer (b) and (c) module to facilitate image fusion.

2) Semantic Feature Modulation: . The Semantic Feature Modulation (SFM) block is designed to achieve cross-modal dynamic regulation of fused features F_f through textual feature space $\phi^T \in \mathbb{R}^{2\times128}$. This architecture comprises two pivotal components: a semantic interaction alignment module based on cross-attention mechanisms[24], [28] and a learnable semantic injection gating module which will be introduced in Sec.III-C. To facilitate deep feature interaction and efficient semantic guidance, L cascaded SFM blocks are sequentially applied through a hierarchical coupling framework, followed by the final decoding stage performed by the decoder. In simple terms, the Feature Modulation can be described as:

$$F_f^{l+1} = \{\mathcal{F}_f^D(\mathcal{A}_n^s(F_f^l, \phi_i^T))\}_m, \tag{4}$$

where F_f^l denotes the fused image features at the l-th stage. $\{\cdot\}_m$ is used to represent the hierarchical coupling architecture. $\mathcal{F}_f^D$ and $\mathcal{A}_n^s$ refer to the decoder based on Restormer Blocks and SFM, respectively. Note that the decoder includes downsampling operations corresponding to the upsampling in the encoder, ensuring consistency in the size of the feature maps.

3) Diagnostic Report and Loss Function: The current state-of-the-art fusion frameworks employing deep neural networks often result in brightness accumulation in the output images. Despite achieving high scores in certain visual evaluation metrics, these methods inevitably lead to a loss of semantic information[14]. To address this issue, we innovatively propose a novel framework that automatically generates a diagnostic report for the fusion results. This feature is designed to assist medical professionals in reducing their workload and enhancing diagnostic efficiency. Specifically, we introduce a medical semantic loss function $\mathcal{L}(\cdot)$, which will be detailed in Sec.III-D. This loss function is designed to explicitly guide the fusion process to retain more semantic information. The final output of our model comprises two components: the fused image F and the diagnostic report.

C. Semantic Guidance Fusion Architecture

1) Semantic Interaction Alignment Module: The semantic interaction alignment module enables the model to handle diverse textual prompts, thereby effectively editing the latent representations of image features and more accurately localizing lesion regions. Medical images, on one hand, typically contain rich spatial information, which is critical for precisely identifying the location of pathological areas. On the other hand, they often carry fine-grained semantic content, such as blood flow patterns that can offer higher-level contextual cues. These characteristics are complementary to the contextual information provided by textual prompts, jointly enhancing the model's holistic understanding of medical images. Inspired by FFCLIP [28], we establish a semantic alignment between text and image through an affine transformation, where all transformation parameters are learned via a cross-attention mechanism. As illustrated in Fig. 3, for the output of the i-th semantic interaction alignment module w_i (where $i \in 1, 2, \ldots, L$ and w_0 is initialized as the features encoded from multimodal images), we compute cross-attention maps between w_i and ϕ^T along both the spatial and channel dimensions, tailored to the information conveyed by each modality. For the position dimension, ϕ_1^T is passed through a convolutional block and set as the query $Q_m \in \mathbb{R}^{256}$, while the latent encoding w_i is set as both the value $V \in \mathbb{R}^{256\times256}$ and the key $K \in \mathbb{R}^{256\times256}$. Subsequently, we calculate the cross-attention map to obtain the scaling parameter R. This parameter undergoes matrix multiplication with the value on the position dimension, so as to transform and match the source image feature space with its corresponding text semantic embedding. Briefly, This process can be expressed as:

$$\begin{aligned} &Q_m = \phi_1^T W^Q, K = w_i W^K, V = w_i W^V, \\ &R = Softmax\left(\frac{Q_m K^T}{d_k}\right), \end{aligned} \tag{5}$$

where $W^Q, W^K, W^v \in \mathbb{R}^{256\times256}$, d_k is the scaling factor, we are able to dynamically adjust the contribution degree of

different dimensions via these parameters, given the semantic prompt ϕ_1^T. As the features along the positional dimension alone are insufficient to capture the complete semantic attributes, it remains necessary to compute the parameters along the channel dimension of the source image, thereby further enhancing semantic alignment.

For the channel dimension, similar to the position dimension, we utilize the same K and V as those for the position dimension, along with a new query to compute the feature map within the channel dimension, then, the V is reconstructed based on this feature map. Therefore, it can be described as:

$$\begin{aligned} &Q_n = \phi_2^T W_n^Q, K = w_i W^K, V = w_i W^V, \\ &S = Softmax\left(\frac{Q_n K}{d_k}\right), \end{aligned} \tag{6}$$

where $W_n^Q \in \mathbb{R}^{1\times 256}$, After obtaining the scale and translation parameters, we can obtain the semantically aligned features[28] according to the following predefined linear transformation rules:

$$\hat{F}_f^i = R \times V + S, \tag{7}$$

where $\hat{F}_f^i$ represents the output of the entire semantic feature alignment module. Subsequently, by passing through the text injection module, we are able to acquire the fused image features that incorporate semantic information. This feature space has the potential to enhance the representational capacity of the image.

2) Text Injection Module: In the text injection module, the aligned fused feature space $\hat{F}_f^i$ is combined using the semantic learning parameters λ and μ, which guide feature fusion from the perspectives of scale adjustment and bias control, respectively, so as to obtain the final fused feature F_f. Therefore, it is defined as:

$$F_f = (1+\lambda)\frac{\hat{F}_f^i - \mu_{\hat{F}_f^i}}{\sigma_{\hat{F}_f^i}} + \mu, \tag{8}$$

where $\mu_{\hat{F}f^i}$ and $\sigma_{\hat{F}_f^i}$ represent the mean and variance of the feature that incorporates text semantic information, respectively, which can help achieve more stable and efficient convergence during model training. The semantically learned parameters λ and μ are obtained via a multi-layer perceptron (MLP), which consists of two fully connected layers with the ReLU activation function.

D. Loss Functions

The loss function plays a crucial role in guiding the network optimization process, directing it to optimize in the correct direction and largely determining the quality of the fused image. From the perspective of semantic-guided image fusion, we aim for the fused image to not only preserve clear texture details and luminance information but also exhibit a high degree of similarity to the given text prompts. In this section, we introduce two main components of the loss function: Visual Quality Loss, which includes gradient loss and reconstruction loss, and Semantic Consistency Loss, which ensures that the fused image aligns well with the semantic content provided by the text prompts.

1) Medical semantic loss: We designed a medical semantic loss function based on [29], which quantifies the semantic correlation between the fused image and the text prompt. This loss function operates at the semantic level, ensuring that the generated image aligns closely with the textual description. By measuring the discrepancy between the image and the text in a shared semantic space, this loss function directly enhances the model's ability to preserve the intended meaning conveyed by the text. It is defined as follows:

$$\mathcal{L}_{\text{semantic}} = \begin{cases} 0\,, & if \quad \cos\left(\mathcal{F}_v^i\left(I^f\right), \phi^T\right) \geq \theta \\ 1 - \cos\left(\mathcal{F}_v^i\left(I^f\right), \phi^T\right), & otherwise \end{cases}, \tag{9}$$

where F_v^i denotes the encoder for the input source images, ϕ^T represents the encoded textual features. θ is a hyperparameter. Through extensive experiments, we set it to 0.85 as the text serves merely as a prompt, which only needs to guide the fused image to optimize towards semantic similarity at the initial stage.

2) Gradient consistency: Gradient Consistency: Once the text similarity loss exceeds the threshold θ, the gradient will no longer be updated. Therefore, to maximize the preservation of the texture and useful information of the source images and further enhance the quality of the fused image, we introduce the gradient loss[30]. It can be expressed as follows:

$$\mathcal{L}_{grad} = \frac{1}{HW}\left\|\nabla F_f - \max\left(\nabla F_1, \nabla F_2\right)\right\|_1, \tag{10}$$

where ∇ denotes the Sobel gradient operator, and $\max(\cdot)$ represents the operation of selecting the maximum value for each element, with H and W denoting the height and width of the image, respectively.

3) Reconstruction loss: The reconstruction loss is primarily employed to guarantee the information integrity between the source image and the fused image throughout the entire encoding process[23], ensuring that the regions of interest in the semantic prompts can be effectively preserved. It can be specifically expressed as follows:

$$\mathcal{L}_1 = ||F_1 - F_f|| + \omega\mathcal{L}_{SSIM}(\mathbf{F}_1, \mathbf{F}_f), \tag{11}$$

where $L_{ssim} = 1 - SSIM$, $SSIM(\cdot)$ represents the structural similarity index[31], and ω is the adjustment factor. Similarly, $\mathcal{L}_2$ can also be obtained using the same method.

4) Total loss: The total loss is a linear combination of the relevant losses mentioned above, and it is formulated as follows:

$$\mathcal{L}_{total} = \alpha\mathcal{L}_{semantic} + \beta\mathcal{L}_{grad} + \gamma\mathcal{L}_{rec}, \tag{12}$$

where α, β, and γ are adjustable hyper-parameters associated with the fusion task, which play a decisive role in determining the quality of the final fused image.

IV. Experiments

A. Implementation Details and Datasets

1) Implementation Details: The training and testing datasets for proposed SMFusion were both obtained from *Harvard* medical school and can be publicly available at http://www.med.harvard.edu/AANLIB/home.html. The text

descriptions were generated using BiomedGPT fine-tuned with instruction, and the images had a size of 256×256. Specifically, the dataset comprised 357 MRI-SPECT pairs, 269 MRI-PET pairs, and 184 MRI-CT pairs. During the training process, the learning rate was set to 0.0001, and the Adam optimizer was employed to train the network for 150 epochs with a batch-size of 2. Furthermore, the hyperparameter α was cautiously set to 0.1, while β and γ were determined to be 1.5 respectively. In the Retormer block, the dimension of each attention head was precisely set to 64. All the experiments mentioned above were completed on a Tesla T4 GPU and an Intel Core i5-12500H CPU with the PyTorch framework.

2) Comparison Approaches: We compared the proposed fusion method SMFusion with seven state-of-the-art methods, namely DIFF-IF[32], MetaFusion[33], EMMA[34], DDFM[35], IFCNN[10], U2Fusion[36], and SDNet[37]. The codes of all these comparison methods are publicly available. Moreover, we strictly set these methods according to the parameter configurations given in the corresponding papers to ensure the scientificity and rigor of the comparison.

3) Evaluation Metrics: To quantitatively evaluate the fusion results, we employ five widely-used metrics: spatial frequency (SF)[38], average gradient (AG)[39], standard deviation (SD)[40], edge information transfer ($Q_{ab/f}$)[14], and multi-scale structural similarity index measure (MS-SSIM)[41]. These metrics capture various critical aspects of fusion quality, including texture structure, edge sharpness, and overall visual fidelity, providing a comprehensive assessment of the method's performance in medical image fusion tasks.

$Q_{ab/f}$ can quantify the edge information transfer between the original images and the fused image, thereby reflecting the quality of visual information in the fused image. The higher the value of $Q_{ab/f}$, the better the quality of the fused image. It is written as follows:

$$Q_{ab/f} = \frac{\sum_{i=1}^{H}\sum_{j=1}^{W}\left(Q_{af}(i,j)w_1(i,j) + Q_{bf}(i,j)w_2(i,j)\right)}{\sum_{i=1}^{H}\sum_{j=1}^{W}\left(w_1(i,j) + w_2(i,j)\right)}, \tag{13}$$

where H and W denote the height and width of the source images respectively, and $Q_{af}(i,j)$ is define as:

$$Q_{fm}(i,j) = Q^e_{fm}(i,j)Q^g_{fm}(i,j), \tag{14}$$

where Q^e_{fm} and Q^g_{fm} indicate the edge intensity and orientation retention measures at position (i,j), $Q_{bf}(i,j)$ follows the same definition logic. ω_1 and ω_2 denote the weights derived from the gradient intensity of the source image, respectively.

SD is a measure that captures the distribution and contrast of the fused image. It indicates the variability of pixel values relative to the mean, and images with higher contrast exhibit superior visual representation effects. It can be expressed as follows:

$$SD = \sqrt{\frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}(x_{i,j} - \mu)^2}, \tag{15}$$

where $x_{i,j}$ denote the pixel value and μ signifies the average pixel value.

SF represents the overall activity within an image in the spatial domain. It is defined by the following equation:

$$SF = \sqrt{RF^2 + CF^2}, \tag{16}$$

where RF and CF denote the row frequency and column frequency. AG evaluates an image's ability to preserve details and textures and is commonly used to assess image sharpness. For a given image of size M×N, the average gradient is formulated as:

$$\bar{G} = \frac{\sum_{i=1}^{M-1}\sum_{j=1}^{N-1}\frac{1}{4}\sqrt{\left(\frac{\partial g(i,j)}{\partial i}\right)^2 + \left(\frac{\partial g(i,j)}{\partial j}\right)^2}}{(M-1)(N-1)},$$

where (m,n) represents the coordinates of the image, and $\frac{\partial g}{\partial i}$ and $\frac{\partial g}{\partial n}$ denote the horizontal and vertical gradient components, respectively. A higher average gradient value indicates a richer representation of image details, signifying a better-fused result.

MS-SSIM can represent the multi-scale assessment of image quality, considering the human visual system's sensitivity to structural information at different scales. The overall MS-SSIM index is calculated by combining the luminance, contrast, and structure comparisons at each scale, weighted by their respective importance. It can be formulated as:

$$MS(x,y) = \left[l_M(x,y)\prod_{j=1}^{M}[c_j(x,y)]^{\alpha_j}[s_j(x,y)]^{\beta_j}\right]^{\gamma_j}, \tag{17}$$

where M is the number of scales, and α_j, β_j, and γ_j are parameters that define the relative importance of each scale.

B. Comparison with SOTA methods

1) Qualitative analysis: To further validate the effectiveness of SMFusion as a semantic-guided image fusion paradigm, we selected brain images from three representative patients for fusion experiments. As shown in Fig.4, compared with seven state-of-the-art fusion methods, SMFusion exhibits three distinct advantages. SMFusion accurately delineates the location and size of the lesion, providing more reliable diagnostic information. It maintains a high level of color fidelity and hierarchical information during the fusion process, ensuring effective integration of multimodal information. Finally, high-density structures such as bones in CT images do not completely obscure the soft tissue details in MRI images, thereby achieving a balanced information representation. Specifically, the fusion results generated by DDFM, SDNet, and U2Fusion exhibit low brightness, leading to significant structural detail loss and reducing visual quality. While MetaFusion produces images with higher brightness, it introduces substantial artifacts, limiting its fusion performance. EMMA excessively incorporates information from CT images, resulting in an inaccurate depiction of the lesion region. In contrast, DIFF-IF and IFCNN effectively capture complementary information from the original images; however, they exhibit limitations in preserving fine soft tissue details from MRI images. Overall, SMFusion outperforms

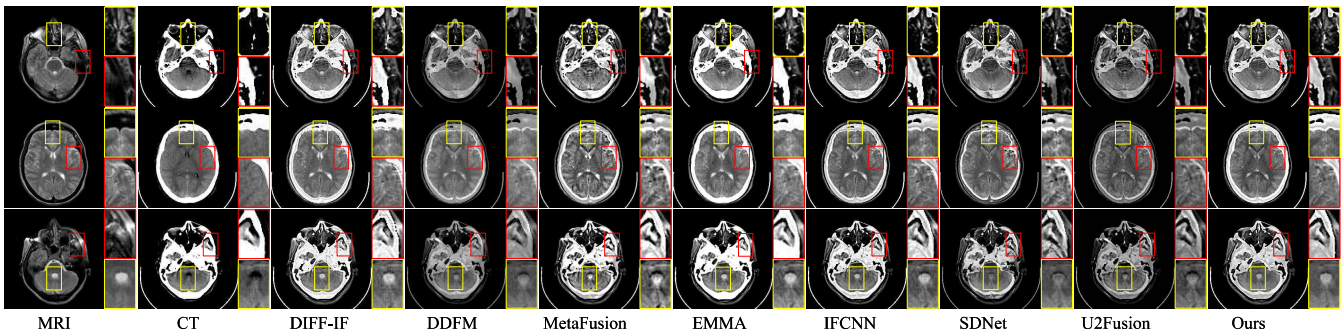


Fig. 4. Qualitative comparison of our SMFusion with seven state-of-the-art methods on three MRI and CT image pairs. Enlarged detail patches are highlighted with red and yellow boxes for visualization.

existing methods in lesion representation, color fidelity, and balanced information preservation.

As shown in Fig.5, we conducted a subjective qualitative analysis of MRI-PET fusion images. DDFM and EMMA exhibit strong color extraction capabilities for PET images. However, their unbalanced information extraction strategy leads to brightness accumulation, which negatively impacts the preservation of cortical folds in MRI images. In contrast, SDNet and U2Fusion show lower performance in information extraction and color consistency, failing to fully capture the essential features from the original images. MetaFusion suffers from noticeable color distortion, particularly evident in the first set of fused images. Additionally, while DIFF-IF and IFCNN perform well in preserving edge and texture details, they fall short in integrating soft tissue structures and functional information compared to SMFusion. This limitation is particularly observable in the first set of fusion results. The MRI-SPECT fusion results are shown in Fig.6. Similar to the MRI-PET fusion results, SMFusion effectively extracts true color information while preserving rich texture details. In the first row of fusion results, MetaFusion and U2Fusion fail to effectively extract soft tissue information from MRI images and exhibit significant color deviations from the source images, which degrades the overall fusion quality. In contrast, SMFusion excels in integrating structural information from MRI and functional information from SPECT, ensuring that essential features are preserved while maintaining high visual quality. Compared to other methods, SMFusion achieves a well-balanced fusion of multimodal information, avoiding information loss or excessive enhancement, and delivering superior visual coherence. In general, SMFusion outperforms the other seven state-of-the-art methods in both visual quality and structural detail preservation, further validating its effectiveness as a semantic-guided image fusion paradigm.

2) Quantitative analysis: To objectively evaluate the fusion performance of SMFusion, we randomly selected 50 brain fusion images from different patients and quantitatively compared them with seven state-of-the-art fusion algorithms. The quantitative results are presented in Table.I. Experimental results demonstrate that SMFusion achieves the highest performance in SF, AG and MS-SSIM, indicating that our method effectively preserves more details and texture information from the original images while avoiding brightness accumulation issues, resulting in the clearest fusion images. These findings are highly consistent with the qualitative evaluation results. For SD and Qabf metrics, SMFusion achieves the second-best performance. This is primarily due to the introduction of a semantic interaction alignment module in our method, which enhances the fused image's ability to understand textual descriptions, thereby extracting richer semantic features from the source images. Unlike other methods, our approach prioritizes the retention of high-level semantic information, which may result in a slight compromise in these specific quantitative metrics. However, this trade-off ensures that the fused images effectively preserve critical medical details, such as lesion regions, making them more clinically valuable. Overall, these results confirm that SMFusion achieves a well-balanced fusion outcome, maintaining both high perceptual quality and essential diagnostic information.

C. Performance in Generating Diagnostic Reports

1) Comparative Analysis of Different Methods: Medical image fusion aims to provide physicians with more accurate and comprehensive diagnostic information. However, generating diagnostic reports from fused images remains labor-intensive and time-consuming. Existing research primarily focuses on improving the visual quality of fused images, with limited attention given to their effectiveness in downstream clinical applications. To address this gap, we randomly selected 27 brain images from the test set and leveraged the medical reasoning capabilities of BiomedGPT, which has demonstrated human-level performance in generating complex radiology reports, to generate diagnostic reports based on different fusion methods. This model's ability to synthesize detailed, context-aware reports makes it particularly suited for clinical use. We evaluated eight fusion approaches quantitatively across three dimensions: average text length, information entropy, and keyword count. These dimensions provide a balanced assessment of the reports, capturing aspects such as depth, informativeness, and relevance, while reflecting both the richness and clarity of the diagnostic information. The evaluation results are summarized in Table III and Table IV, where we present a comparison with SDnet due to space constraints. Additional comparison results can be found in the appendix.

Experimental results demonstrate that our proposed method achieves the highest performance in terms of information

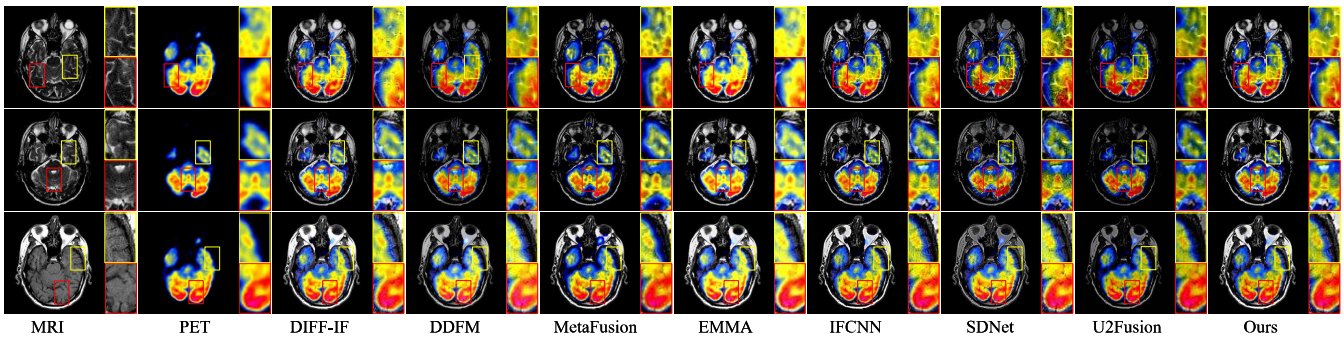


Fig. 5. Qualitative comparison of SMFusion with seven state-of-the-art methods on three MRI and PET image pairs. Notably, our method demonstrates superior performance in blood flow representation and texture preservation (e.g., the first row).

TABLE I
QUANTITATIVE COMPARISON OF OUR SMFUSION WITH SEVEN STATE-OF-THE-ART METHODS ON THE TEST DATASET IN TERMS OF MEAN AND STANDARD DEVIATION. RED AND BLUE FONTS INDICATE THE BEST AND SECOND-BEST RESULTS, RESPECTIVELY.

Method	**SF**$_{\uparrow}$	**AG**$_{\uparrow}$	**MS_SSIM**$_{\uparrow}$	**SD**$_{\uparrow}$	$\mathbf{Q}_{ab/f}$ $\uparrow$
MetaFusion	17.742 ± 7.839	4.734 ± 2.347	1.160 ± 0.337	46.457 ± 2.347	0.453 ± 3.302
DDFM	16.317 ± 4.636	4.189 ± 1.328	1.298 ± 0.185	56.287 ± 10.827	0.635 ± 0.101
IFCNN	23.135 ± 7.74	5.852 ± 2.048	1.297 ± 0.201	55.066 ± 10.641	0.702 ± 0.058
U2Fusion	12.738 ± 3.944	3.346 ± 1.216	1.129 ± 0.241	33.733 ± 10.58	0.378 ± 0.062
SDNet	21.958 ± 6.241	5.501 ± 1.779	1.235 ± 0.199	49.137 ± 8.759	0.647 ± 0.071
EMMA	21.809 ± 6.576	5.601 ± 1.951	1.263 ± 0.204	69.128 ± 10.701	0.592 ± 0.089
DIFF-IF	23.852 ± 7.428	6.065 ± 1.978	1.276 ± 0.207	59.410 ± 10.246	0.756 ± 0.058
Ours	25.770 ± 8.034	6.333 ± 2.055	1.298 ± 0.197	60.230 ± 10.674	0.707 ± 0.064

entropy, indicating that the generated diagnostic reports contain more semantically rich medical information. As shown in Table IV, a direct comparison between our method and SDNet, the best-performing baseline, further validates this advantage. In the first set of fusion results, both methods successfully identified a defect in the posterior fossa. However, our method additionally detected an ependymoma and indicated potential pathological abnormalities. In the second set, when BiomedGPT was applied to SDNet's fusion output, it merely highlighted regions of increased metabolic activity without providing meaningful medical insights.

In addition, we also show the diagnostic reports of single-modal and multi-modal multi-channel (non-fusion) medical images in Table VIII. It is evident that these pre-fusion images exhibit certain limitations in conveying comprehensive diagnostic information. Specifically, both single-modality and non-fused multi-modality images are only capable of capturing partial brain features, such as cerebral edema in the right cerebellum, morphological changes in the sella turcica, and enlargement of the pituitary gland, or merely referencing brain metabolic activity. Such information has some diagnostic value; however, it does not offer a systematic or comprehensive view. For example, in the second case presented, neither the single-modality image nor the combined information from two separate modalities was sufficient to accurately identify the lesion. In contrast, the fused image clearly reveals the presence of an adenoma and provides interpretable lesion characteristics, so as to assist the doctor to further confirm the accuracy of the results.

2) Mean Opinion Score testing: To ensure the reliability of the diagnostic reports generated by large language models, we conducted a questionnaire-based subjective assessment. Specifically, we recruited 13 medical students with medical backgrounds to participate in the study. Each participant was asked to rate the diagnostic reports using one of three levels: Moderate, Satisfactory, or Excellent, based on how well the content aligned with clinical expectations. We randomly selected 10 diagnostic reports based on fused images from the test set, covering a total of 8 experimental groups representing 8 different fusion methods. To eliminate potential bias, all method-related information was blinded to the participants. The results of the survey are summarized in Table VII. As shown, over 80% of the reports across all fusion methods received a rating of Satisfactory or above. Notably, the reports generated based on our fusion results received the highest proportion of Excellent ratings. These findings suggest that the majority of diagnostic reports generated by BiomedGPT based on fused images are considered accurate.

D. Ablation studies

To validate the effectiveness of the proposed method, we conducted a systematic analysis of the key innovations in SMFusion, focusing on the impact of the semantic interaction alignment module, loss function and the influence of different hyperparameter settings in the loss function on model performance.

1) Ablation analysis of different modules: In this paper, we propose a deep network incorporating a semantic interaction alignment module and apply it to medical image fusion. To simplify the analysis, we decompose the network into two main components, *i.e.*, $\mathcal{K}$ and $\mathcal{M}$, which represent the semantic alignment modules for different modalities. Furthermore,

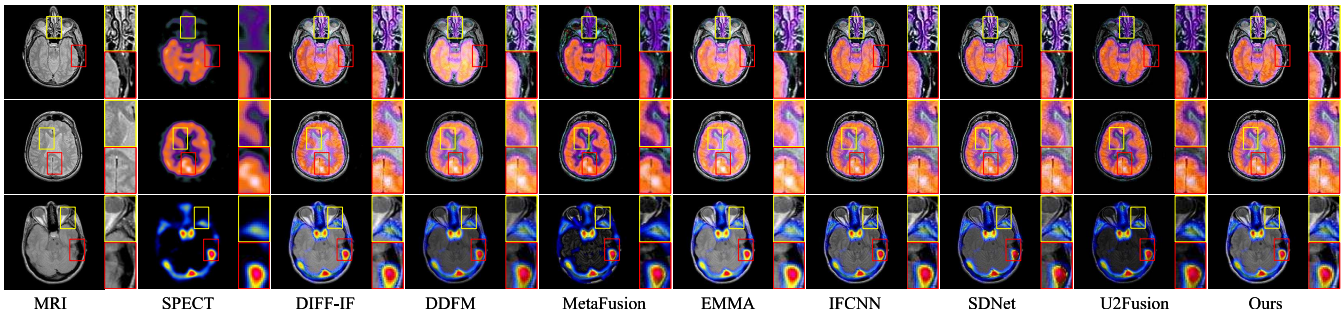


Fig. 6. Qualitative comparison of ours with seven state-of-the-art methods on three MRI and SPECT image pairs. Enlarged detail patches are highlighted with red and yellow boxes and displayed on the right side of the original images for visualization.

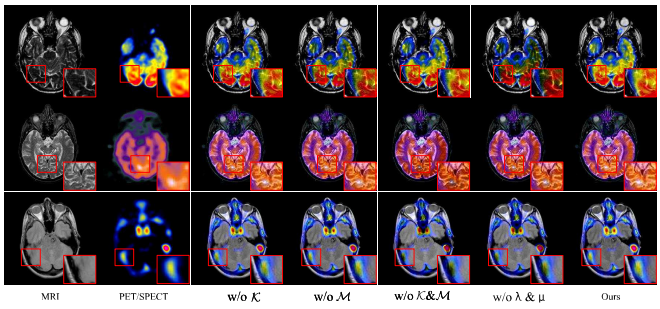


Fig. 7. Ablation qualitative experiments were performed on three different image pairs. Each module shows a positive contribution to the fusion performance of SMFusion.

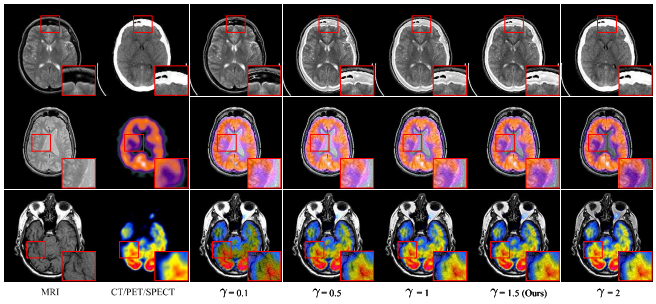


Fig. 8. Ablation experiments between different hyperparameter settings. The enlarged details are displayed in the bottom-right corner of the source image.

the semantic learning parameters λ and μ in the text injection module facilitate the fusion of aligned features in a high-dimensional space. As shown in Fig.7, qualitative results demonstrate that each module contributes positively to the final fusion image. Specifically, the absence of $\mathcal{K}$ leads to suboptimal preservation of fine details and edge information, resulting in less effective enhancement of soft tissue structures in MRI images. Meanwhile, omitting $\mathcal{M}$ results in a decline in color processing capability. Furthermore, when both $\mathcal{K}$ and $\mathcal{M}$ are removed, the network fails to learn sufficient medical semantic information, making it unable to focus on abnormal regions in the source images, as illustrated in the second set of fusion results. Further analysis reveals that removing the text injection module prevents the high-dimensional text-image feature space from being fully disentangled, introducing additional noise during decoding and causing noticeable color distortions in the fused images. This indicates that the text injection module plays a crucial role in extracting effective

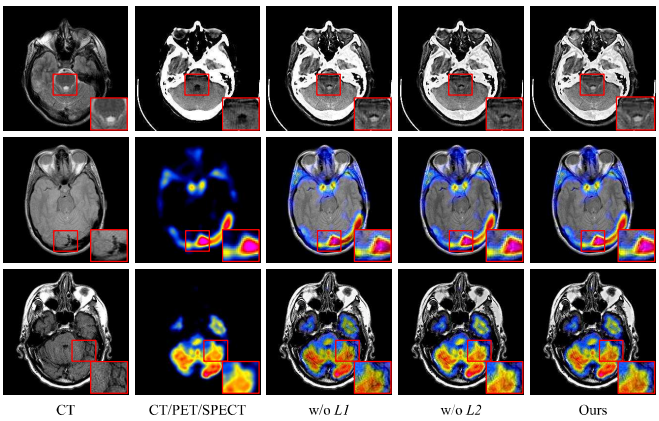


Fig. 9. Qualitative analysis of different loss functions. $L1$ and $L2$ represent the proposed medical semantic loss and gradient loss, respectively.

TABLE II
QUANTITATIVE ANALYSIS UNDER DIFFERENT COMBINATIONS OF LOSS FUNCTIONS. RED AND BLUE INDICATE THE BEST AND SECOND-BEST RESULTS, RESPECTIVELY. THE FUSION RESULTS ARE OPTIMAL WHEN ALL PROPOSED LOSS FUNCTIONS ARE INCORPORATED.

Metric	**SF**	**AG**	**MS-SSIM**	**SD**	$\mathbf{Q}_{ab/f}$
w/o $l1$	25.355	6.333	1.286	59.258	0.701
w/o $l2$	25.829	6.316	1.293	59.646	0.700
ours	25.770	6.335	1.298	60.230	0.707

image information and significantly enhances the network's robustness. Moreover, as shown in table V, quantitative experimental results further validate the proposed method. Our approach achieves state-of-the-art performance across six evaluation metrics. Each module contributes to different aspects of fusion quality improvement, demonstrating the positive impact of incorporating both the semantic interaction alignment module and the text injection module on model performance.

2) Ablation analysis of different hyperparameter: Extensive experiments have demonstrated that the choice of hyperparameters plays a crucial role in determining the quality of the fusion results. The loss function updates gradients through backpropagation, guiding the network toward minimizing the gradient and ultimately influencing the quality of the fused images. In our experiments, we observed that the hyperparameter γ in the reconstruction loss significantly affects the model's fusion performance. The qualitative analysis results, as shown in Fig.8, indicate that we set γ to 0.1, 0.5, 1, 1.5, and 2 to examine the variation in fusion images. It was observed that

TABLE III
QUANTITATIVE COMPARISON OF THE FUSION RESULTS OF DIFFERENT METHODS IN GENERATING DIAGNOSTIC REPORTS. RED AND BLUE FONTS INDICATE THE OPTIMAL AND SECOND-BEST RESULTS, RESPECTIVELY.

Method	DDFM	Diff-IF	EMMA	IFCNN	MetaFusion	U2Fusion	SDNet	Ours
Text Length	40.2593	41.8148	37.6296	43.0370	39.5185	38.9259	40.0370	33.8889
Entropy	4.0490	4.0633	4.0250	4.0604	4.0561	4.0490	4.0707	4.0762

TABLE IV
DIAGNOSTIC REPORT GENERATION FROM FUSION IMAGES PRODUCED BY SDNET AND SMFUSION. RED TEXT INDICATES KEY DIAGNOSTIC INFORMATION. CLEARLY, SMFUSION GENERATES MORE DETAILED AND SPECIFIC DESCRIPTIONS.

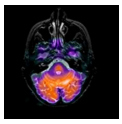
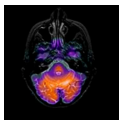
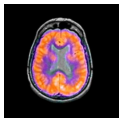
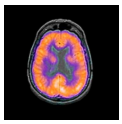
Method	Image	Fusion Result Description	Text Length	Entropy	Key Words
Ours		Defect in the posterior fossa which is an abnormality in the optic nerve sheath, a rare congenital condition where there are ependymomas found, as well as an area of increased radiotracer uptake on the surface of the brain, indicating a possible abnormality or pathology.	44	4.0130	posterior fossa, ependymomas, abnormality or pathology
SDNet		Defect in the posterior fossa which is an abnormality in the brain tissue as seen on the other side of the cerebellum, midbrain, and pons, as well as the presence of a lacunar infarct, a type of brain tissue that stands out from the surrounding healthy brain tissue.	48	3.9535	posterior fossa, lacunar infarct
Ours		Brain structures such as the basal ganglia and lateral ventricles are also visible in the image, but the main focus is on the adenoma, a type of brain tumor with increased uptake of the radiotracer due to the presence of a lacunar infarct.	43	3.9739	adenoma, presence of a lacunar infarct
SDNet		Positron emission tomography is a technique used to assess the metabolic activity of the brain. In this image, you can see areas of increased metabolic activity, represented by brighter colors compared to the surrounding brain tissue.	36	4.0725	increased metabolic activity

TABLE V
QUANTITATIVE ABLATION OF DIFFERENT MODULES, WITH RED AND BLUE MARKING THE BEST AND SECOND-BEST RESULTS.

$\mathcal{K}$	$\mathcal{M}$	λ	μ	SF	AG	MS_SSIM	SD	$\mathbf{Q}_{ab/f}$
✓	✓			25.679 ± 8.025	6.308 ± 2.111	1.210 ± 0.213	60.018 ± 10.523	0.675 ± 0.075
		✓	✓	25.691 ± 8.019	6.304 ± 2.060	1.276 ± 0.202	57.711 ± 11.001	0.706 ± 0.064
	✓	✓	✓	25.671 ± 8.128	6.323 ± 2.088	1.278 ± 0.203	59.783 ± 10.955	0.697 ± 0.062
✓		✓	✓	26.182 ± 8.327	6.315 ± 2.149	1.266 ± 0.204	60.210 ± 11.282	0.703 ± 0.063
✓	✓	✓	✓	25.770 ± 8.034	6.335 ± 2.055	1.298 ± 0.197	60.230 ± 10.674	0.707 ± 0.064

as the value of γ increased, more CT information was retained in the fused images, which led to the suppression of soft tissue structures from the MRI images, as shown in the first group of example. The quantitative analysis results, presented in Table VI, further confirm this trend, demonstrating that our method achieves superior performance across most metrics and remains in a leading position. Considering both qualitative and quantitative evaluations, the optimal fusion results are obtained when γ is set to 1.5, as it ensures a balanced extraction of information from the source images.

3) Ablation analysis of loss function: In Section III-D, we introduce a medical semantic loss function specifically designed to enhance fusion performance in medical imaging tasks. The qualitative fusion results under different loss combinations are shown in Fig. 9. As evident from the visual comparisons, the absence of the $L1$ loss leads to undesirable artifacts and noticeable color distortions in the fused images. When the $L2$ loss is excluded during training, the fusion results tend to miss critical structural details, such as the continuity of soft tissue in MRI images. The quantitative results summarized in Table II. Our method achieves the best performance across all evaluated metrics, demonstrating the effectiveness and necessity of the proposed loss functions in improving fusion quality.

V. CONCLUSION

In this study, we propose a semantic-guided medical image fusion method that incorporates expert-level prior knowledge, allowing the model to focus more precisely on pathological regions while preserving critical medical information. This approach enhances the visibility of lesions and ensures a well-balanced integration of fine details and high-level semantic information. To further improve fusion performance, we introduce a medical semantic loss, aligning the fused images more closely with real-world clinical needs in terms of content

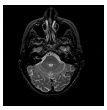
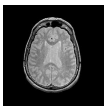
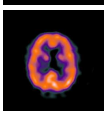
TABLE VI
QUANTITATIVE ANALYSIS OF THE ABLATION EXPERIMENTS WITH DIFFERENT HYPERPARAMETERS, WHERE THE FUSION PERFORMANCE ACHIEVES THE BEST RESULTS ACROSS ALL METRICS WHEN γ IS SET TO 1.5.

The value of γ	**SF**	**AG**	**MS_SSIM**	**SD**	$\mathbf{Q}_{ab/f}$
$\gamma = 0.1$	24.699 ± 6.721	6.066 ± 2.118	1.159 ± 0.172	53.196 ± 14.843	0.668 ± 0.083
$\gamma = 0.5$	24.525 ± 7.661	6.281 ± 2.017	1.265 ± 0.202	60.025 ± 10.599	0.734 ± 0.055
$\gamma = 1$	24.356 ± 7.970	6.478 ± 2.104	1.272 ± 0.206	60.122 ± 11.133	0.700 ± 0.066
$\gamma = 2$	25.149 ± 7.982	6.136 ± 1.959	1.280 ± 0.207	55.606 ± 12.473	0.680 ± 0.061
$\gamma = 1.5$ (Ours)	25.770 ± 8.034	6.335 ± 2.055	1.298 ± 0.197	60.230 ± 10.674	0.707 ± 0.064

TABLE VII
RESULTS OF THE SUBJECTIVE ASSESSMENT, WHERE THE PERCENTAGES REPRESENT THE PROPORTION OF PARTICIPANTS' RATINGS FOR DIAGNOSTIC REPORTS. NOTABLY, FEWER MODERATE RATINGS AND MORE OTHER RATINGS ARE BETTER.

Method	**DDFM**	**Diff-IF**	**EMMA**	**IFCNN**	**MetaFusion**	**U2Fusion**	**SDNet**	**Ours**
Moderate	7%	16%	19%	12%	19%	16%	18%	9%
Satisfactory	66%	59%	52%	62%	47%	51%	54%	56%
Excellent	27%	25%	29%	26%	34%	33%	28%	35%

TABLE VIII
DIAGNOSTIC REPORTS FOR SINGLE-MODALITY AND MULTI-MODALITY MULTI-CHANNEL (NON-FUSION) SCANNED IMAGES.

Image	**Diagnostic Report**
	Brain edema can be seen in the image as an area of increased signal intensity within the right cerebellum, which is located between the brainstem and the cerebrum.
	There is a noticeable reduction in the size of the sella turcica and an enlargement of the pituitary gland, which is the main focus of the scan.
	Remnant Astrocytoma, Brain Edema, Brain Nonenhancing Tumor.
	Emission computed tomography scans provide detailed information about the brain's metabolic activity, such as inflammation, dysplastic changes, or pathological changes.

consistency and medical relevance. In addition, we develop a novel evaluation mechanism that assesses the effectiveness of fused images through the diagnostic reports they generate, providing a clinically oriented perspective on fusion quality. Experimental results demonstrate that our method outperforms current approaches in preserving fine details, enhancing texture representation, and integrating semantic information, offering valuable support for clinical decision-making. In the future, we will further investigate the mechanisms of text-guided multi-modal medical image fusion and conduct extensive validation on large-scale clinical datasets to improve model robustness and generalization.

REFERENCES

[1] J. Fu, W. Li, J. Du, and B. Xiao, "Multimodal medical image fusion via laplacian pyramid and convolutional neural network reconstruction with local gradient energy strategy," *Computers in Biology and Medicine*, vol. 126, p. 104048, 2020.

[2] M. Yin, X. Liu, Y. Liu, and X. Chen, "Medical image fusion with parameter-adaptive pulse coupled neural network in nonsubsampled shearlet transform domain," *IEEE Transactions on Instrumentation and Measurement*, vol. 68, no. 1, pp. 49–64, 2018.

[3] J. Nunez, X. Otazu, O. Fors, A. Prades, V. Pala, and R. Arbiol, "Multiresolution-based image fusion with additive wavelet decomposition," *IEEE Transactions on Geoscience and Remote sensing*, vol. 37, no. 3, pp. 1204–1211, 1999.

[4] A. Maćkiewicz and W. Ratajczak, "Principal components analysis (pca)," *Computers & Geosciences*, vol. 19, no. 3, pp. 303–342, 1993.

[5] P. Chai, X. Luo, and Z. Zhang, "Image fusion using quaternion wavelet transform and multiple features," *IEEE access*, vol. 5, pp. 6724–6734, 2017.

[6] J. Ma and G. Plonka, "The curvelet transform," *IEEE signal processing magazine*, vol. 27, no. 2, pp. 118–133, 2010.

[7] H.-Y. Yang, X.-Y. Wang, P.-P. Niu, and Y.-C. Liu, "Image denoising using nonsubsampled shearlet transform and twin support vector machines," *Neural networks*, vol. 57, pp. 152–165, 2014.

[8] L. Li, M. Lv, Z. Jia, and H. Ma, "Sparse representation-based multi-focus image fusion method via local energy in shearlet domain," *Sensors*, vol. 23, no. 6, p. 2888, 2023.

[9] Z. Wang, Z. Cui, and Y. Zhu, "Multi-modal medical image fusion by laplacian pyramid and adaptive sparse representation," *Computers in Biology and Medicine*, vol. 123, p. 103823, 2020.

[10] Y. Zhang, Y. Liu, P. Sun, H. Yan, X. Zhao, and L. Zhang, "Ifcnn: A general image fusion framework based on convolutional neural network," *Information Fusion*, vol. 54, pp. 99–118, 2020.

[11] J. Ma, L. Tang, F. Fan, J. Huang, X. Mei, and Y. Ma, "Swinfusion: Cross-domain long-range learning for general image fusion via swin transformer," *IEEE/CAA Journal of Automatica Sinica*, vol. 9, no. 7, pp. 1200–1217, 2022.

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

[13] J. Ma, H. Xu, J. Jiang, X. Mei, and X.-P. Zhang, "Ddcgan: A dual-discriminator conditional generative adversarial network for multi-resolution image fusion," *IEEE Transactions on Image Processing*, vol. 29, pp. 4980–4995, 2020.

[14] W. Huang, H. Zhang, H. Guo, W. Li, X. Quan, and Y. Zhang, "Addns: An asymmetric dual deep network with sharing mechanism for medical image fusion of ct and mr-t2," *Computers in Biology and Medicine*, vol. 166, p. 107531, 2023.

[15] F. Fan, Y. Huang, L. Wang, X. Xiong, Z. Jiang, Z. Zhang, and J. Zhan, "A semantic-based medical image fusion approach," *arXiv preprint arXiv:1906.00225*, 2019.

[16] Z. Zhao, L. Deng, H. Bai, Y. Cui, Z. Zhang, Y. Zhang, H. Qin, D. Chen, J. Zhang, P. Wang *et al.*, "Image fusion via vision-language model," *arXiv preprint arXiv:2402.02235*, 2024.

[17] K. Zhang, R. Zhou, E. Adhikarla, Z. Yan, Y. Liu, J. Yu, Z. Liu, X. Chen, B. D. Davison, and H. Ren, "Biomedgpt: A generalist vision-language foundation model for diverse biomedical tasks," *Nature Medicine*, 2023.

[18] M. Zhou, L. Yu, A. Singh, M. Wang, Z. Yu, and N. Zhang, "Unsupervised vision-and-language pre-training via retrieval-based multi-granular

alignment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16 485–16 494.

[19] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PmLR, 2021, pp. 8748–8763.

[20] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *International conference on machine learning*. PMLR, 2022, pp. 12 888–12 900.

[21] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[22] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," *International Journal of Computer Vision*, vol. 130, no. 9, pp. 2337–2348, 2022.

[23] J. Liu, X. Li, Z. Wang, Z. Jiang, W. Zhong, W. Fan, and B. Xu, "Promptfusion: Harmonized semantic prompt learning for infrared and visible image fusion," *IEEE/CAA Journal of Automatica Sinica*, 2024.

[24] C. Cheng, T. Xu, X.-J. Wu, H. Li, X. Li, Z. Tang, and J. Kittler, "Textfusion: Unveiling the power of textual semantics for controllable image fusion," *Information Fusion*, vol. 117, p. 102790, 2025.

[25] X. Yi, H. Xu, H. Zhang, L. Tang, and J. Ma, "Text-if: Leveraging semantic text guidance for degradation-aware and interactive image fusion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 27 026–27 035.

[26] ——, "Text-if: Leveraging semantic text guidance for degradation-aware and interactive image fusion," *IEEE*, 2024.

[27] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and M.-H. Yang, "Restormer: Efficient transformer for high-resolution image restoration," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 5728–5739.

[28] Y. Zhu, H. Liu, Y. Song, Z. Yuan, X. Han, C. Yuan, Q. Chen, and J. Wang, "One model to edit them all: Free-form text-driven image manipulation with semantic modulations," *Advances in Neural Information Processing Systems*, vol. 35, pp. 25 146–25 159, 2022.

[29] T. Wei, D. Chen, W. Zhou, J. Liao, Z. Tan, L. Yuan, W. Zhang, and N. Yu, "Hairclip: Design your hair by text and reference image," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 072–18 081.

[30] X. Liu, H. Huo, J. Li, S. Pang, and B. Zheng, "A semantic-driven coupled network for infrared and visible image fusion," *Information Fusion*, vol. 108, p. 102352, 2024.

[31] H. Zhao, O. Gallo, I. Frosio, and J. Kautz, "Loss functions for image restoration with neural networks," *IEEE Transactions on computational imaging*, vol. 3, no. 1, pp. 47–57, 2016.

[32] X. Yi, L. Tang, H. Zhang, H. Xu, and J. Ma, "Diff-if: Multi-modality image fusion via diffusion model with fusion knowledge prior," *Information Fusion*, vol. 110, p. 102450, 2024.

[33] W. Zhao, S. Xie, F. Zhao, Y. He, and H. 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*, 2023, pp. 13 955–13 965.

[34] Z. Zhao, H. Bai, J. Zhang, Y. Zhang, K. Zhang, S. Xu, D. Chen, R. Timofte, and L. Van Gool, "Equivariant multi-modality image fusion," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 25 912–25 921.

[35] Z. Zhao, H. Bai, Y. Zhu, J. Zhang, S. Xu, Y. Zhang, K. Zhang, D. Meng, R. Timofte, and L. Van Gool, "Ddfm: denoising diffusion model for multi-modality image fusion," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 8082–8093.

[36] H. Xu, J. Ma, J. Jiang, X. Guo, and H. Ling, "U2fusion: A unified unsupervised image fusion network," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 1, pp. 502–518, 2020.

[37] H. Zhang and J. Ma, "Sdnet: A versatile squeeze-and-decomposition network for real-time image fusion," *International Journal of Computer Vision*, vol. 129, no. 10, pp. 2761–2785, 2021.

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

[39] G. Cui, H. Feng, Z. Xu, Q. Li, and Y. Chen, "Detail preserved fusion of visible and infrared images using regional saliency extraction and multi-scale image decomposition," *Optics Communications*, vol. 341, pp. 199–209, 2015.

[40] C. S. Xydeas and V. Petrovic, "Objective image fusion performance measure," *Electronics letters*, vol. 36, no. 4, pp. 308–309, 2000.

[41] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, vol. 2. Ieee, 2003, pp. 1398–1402.