An Efficient Image Fusion Network Exploiting Unifying Language and Mask Guidance

Zi-Han Cao, Yu-Jie Liang, Liang-Jian Deng, *Senior Member, IEEE*, and Gemine Vivone, *Senior Member, IEEE*

Abstract—Image fusion aims to merge image pairs collected by different sensors over the same scene, preserving their distinct features. Recent works have often focused on designing various image fusion losses, developing different network architectures, and leveraging downstream tasks (e.g., object detection) for image fusion. However, a few studies have explored how language and semantic masks can serve as guidance to aid image fusion. In this paper, we investigate how the combination of language and masks can guide image fusion tasks, discarding the previously complex frameworks, which rely on downstream tasks, GAN-based cycle training, diffusion models, or deep image priors. Additionally, we exploit a recurrent neural network-like architecture to build a lightweight network that avoids the quadratic-cost of traditional attention mechanisms. To adapt the receptance weighted key value (RWKV) model to an image modality, we modify it into a bidirectional version using an efficient scanning strategy (ESS). To guide image fusion by language and mask features, we introduce a multi-modal fusion module (MFM) to facilitate information exchange. Comprehensive experiments show that the proposed framework achieved state-of-the-art results in various image fusion tasks (i.e., visible-infrared image fusion, multi-focus image fusion, multi-exposure image fusion, medical image fusion, hyperspectral and multispectral image fusion, and pansharpening). Code will be available at https://github.com/294coder/RWKVFusion.

Index Terms—Multi-modal guided image fusion, efficient network, attention, image fusion, pansharpening, remote sensing, deep learning.

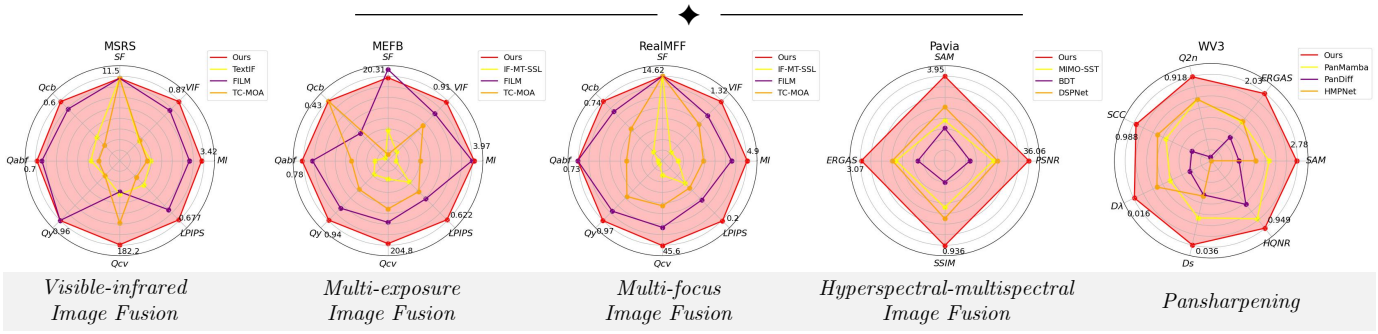


Fig. 1: Performance assessment on several image fusion tasks comparing with state-of-the-art methods.

1 Introduction

Image fusion, a low-level vision task, holds a pivotal role in various fields, such as computational imaging, military reconnaissance, and medical diagnosis [1], [2], [3], [4].

Due to some limitations of current sensors and optical imaging technologies, the information acquired by a single modality does not fully represent real-world scenarios [5]. However, we can utilize different sensors to capture images with different modalities, which are often complementary (e.g., remote sensing satellites capture both panchromatic and multispectral images [1], [6] or hyperspectral images [2], or visible and near-infrared images [3], [5], [7]). The goal of image fusion is to create a composite (synthetic) image by combining complementary information from multiple source images that are captured using various sensors or optical configurations. We aim for the fused image to represent the information contained in the input images referring to different modalities [7]. A network showing elevated fusion performance can serve as pre-processing for downstream tasks (e.g., object detection [8] and segmentation [9]).

In the field of image fusion, numerous methods have been proposed, mainly divided into: i) traditional fusion frameworks and ii) deep fusion frameworks using neural networks. Because of some limitations in traditional frameworks, such as defining handcrafted features and fusion rules, to date, state-of-the-art fusion methods mainly focus on the second class. Recent deep fusion frameworks

- *This research is supported by NSFC (Grant No. 12271083), and the Project of the Department of Science and Technology of Sichuan Province (Grant No. 2025YFNH0001).*
- *Z.-H. Cao, Y.-J. Liang are with the School of Mathematical Sciences, University of Electronic Science and Technology of China, 611731 Chengdu, China (e-mails: iamzihan666@gmail.com, yujieliang0219@gmail.com). L.-J. Deng is with the School of Mathematical Sciences/Multi-Hazard Early Warning Key Laboratory of Sichuan Province, University of Electronic Science and Technology of China (e-mail: liangjian.deng@uestc.edu.cn).*
- *G. Vivone is with the National Research Council, Institute of Methodologies for Environmental Analysis, CNR-IMAA, 85050 Tito, Italy, and with the National Biodiversity Future Center, NBFC, 90133 Palermo, Italy (e-mail: gemine.vivone@imaa.cnr.it).*
- *Z.-H. Cao and Y.-J. Liang have equal contributions.*
- *Corresponding authors: L.-J. Deng.*
- *Supported by Center for HPC, University of Electronic Science and Technology of China.*

usually put a spotlight on designing: i) fusion frameworks (e.g., combining loss functions or incorporating additional information to aid the fusion process); and ii) efficient deep networks and neural operators.

Regarding fusion frameworks, recent works have explored various strategies, such as different fusion losses [10], additional discriminators to aid the fusion process [11], diffusion priors [12], [13], deep image priors (DIPs) with iterative algorithms [14], and incorporating some high-level information [13], [15], [16]. These fusion frameworks demonstrated that solely relying on fusion losses for constraints is limiting. Instead, utilizing other relevant information is beneficial for the fusion process. Despite good outcomes, current fusion frameworks still show evident drawbacks, e.g., semantic segmentation leads to training overhead and additional priors need to train dual generator/discriminator networks or inference on diffusion processes.

A few studies explored how language and semantic masks can serve as guidance to aid image fusion tasks. Hence, in this work, we investigate the unification of language and masks to guide image fusion tasks, discarding previously developed (more complex) frameworks that rely on downstream tasks, generative adversarial networks (GANs), diffusion models, or DIPs. To the best of our knowledge, the only approaches that use language as guidance for image fusion are TextIF [17] and FILM [18]. We will review them pointing out the differences with the proposed framework in Sect. 2.1.

Designing an efficient network for image fusion is crucial, especially when processing high-resolution images that face large floating point operations per second (FLOPs) and memory issues. Many architectures have been developed for visual tasks. Early deep vision networks, such as convolutional neural networks (CNNs) with residual blocks [19] and autoencoder-based feature extraction [20], offer fast execution times but struggle with smaller receptive fields. Vision transformer (ViT) [21], with its global receptive field and less inductive bias, excels in fine-grained tasks like segmentation, especially when large datasets are considered. By patching images, ViT reduces the impact of the self-attention's quadratic spatial cost, but leads to a loss of spatial information.

However, since image fusion is a low-level vision task, it significantly differs from other high-level tasks for the following reasons: i) it involves inputs from different modalities (e.g., visible and near-infrared images) requiring tailored modules for feature handling; ii) models operate at pixel-level (not in a latent space as for diffusion models [22]) requiring more computational burden; iii) traditional CNNs fail to extract enough global (context-based) information and the quadratic memory consumption of classical attention is unacceptable for processing high-resolution images.

Most current image fusion networks aim to balance performance and complexity by adopting hybrid architectures that combine convolution and attention mechanisms (including window attention [23] or other linear attentions [24], [25]). Although this approach alleviates some issues, handling high-resolution images still requires substantial memory overhead. The recent introduction of vision Mamba (VMamba) [26] into the image fusion literature led to the reduction of memory costs to near-linear levels [27]. However, vision Mamba performance remains controversial [28].

A novel recurrent neural network (RNN)-like architecture, named receptance weighted key value (RWKV) [29], has recently demonstrated outstanding performance in language modeling. Inspired by the attention-free transformer [30], which decomposes attention into vector operations, RWKV combines the efficient parallelizable training of transformers with the efficient inference of RNNs. Vision-RWKV [31] extends this approach to vision tasks, yielding promising results. However, the currently designed RWKV networks neglect some specific requirements of image fusion tasks, in particular, overlooking language and other semantic guidance. Hence, we introduce the efficient linear attention mechanism of RWKV as a fundamental operator for image fusion tasks designing an efficient backbone network, which is multi-scale, has a global receptive field, exhibits low latency, and unifies various guidance.

In this work, we are committed to addressing the shortcomings of existing fusion frameworks by proposing RWKVFusion, a multi-modal fusion framework. This framework integrates various guidance into a single efficient network to tackle image fusion tasks, thereby enhancing fusion performance, as illustrated in Fig. 1. In the proposed fusion framework, with data examples shown in Fig. 2, we introduce both language and semantic masks to guide the fusion process overcoming the drawbacks of previous fusion frameworks.

The contribution of our paper is three-fold:

- We propose a new image fusion framework with multi-modal guidance. Our image fusion framework combines global-level language overview and object-level semantic mask guidance, thus overcoming the drawbacks of previous fusion frameworks (e.g., limited fusion guidance, complex priors, and dependence on downstream prediction heads to introduce semantic information).
- We introduce a linear-cost fusion network based on the RWKV operator. We adapted RWKV to image fusion developing a multi-scale architecture, named RWKVFusion, which integrates RWKV with an efficient 2D image scanning strategy. To incorporate diverse guidance, we designed a multi-modal fusion module (MFM) that exploits both language and mask features to semantically and globally guide image fusion processes.
- We conducted experiments on various image fusion tasks, including visible-infrared image fusion (VIF), multi-exposure image fusion (MEF), multi-focus image fusion (MFF), medical image fusion (MIF), hyperspectral and multispectral image fusion (HMIF), and pansharpening. Extensive experiments assessed the effectiveness of RWKVFusion.

The rest of this paper is organized as follows. We first review related works in Sect. 2. The proposed RWKVFusion framework is discussed in Sect. 3. The experimental analysis on several image fusion tasks is shown in Sect. 4. Sects. 5 and 6 are devoted to some ablation studies and discussions, respectively. Finally, concluding remarks are drawn in Sect. 7.

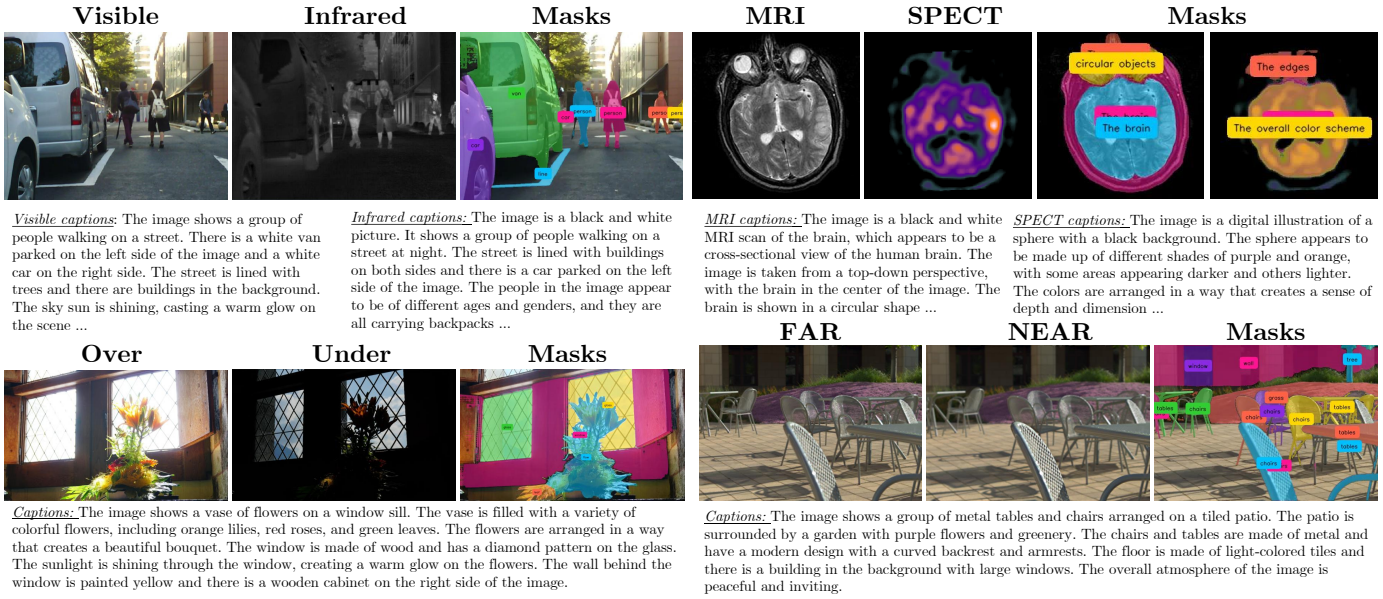


Fig. 2: Representative multi-modal data with language and mask guidance. The image fusion tasks from left to right and from top to bottom are: visible-infrared, medical image, multi-exposure, and multi-focus. To serve as fusion guidance, dense captions and masks are generated by Florence [32] and SAM [33], respectively.

2 Related Works

2.1 Image Fusion and High-level Semantic Guidance

Image fusion is an image processing task with many downstream applications such as object detection and segmentation. Its general formulation is as follows:

$$\mathbf{F} = \mathcal{F}_{\boldsymbol{\theta}}(\mathbf{S}_1, \mathbf{S}_2, \cdots, \mathbf{S}_n), \tag{1}$$

where $\mathcal{F}_{\boldsymbol{\theta}}(\cdot)$ is a deep fusion network with parameters $\boldsymbol{\theta}$ that takes different modalities $\mathcal{S} = \{\mathbf{S}_1, \mathbf{S}_2, \cdots, \mathbf{S}_n\}$ as input and generates a fused image, $\mathbf{F}$.

The optimal fused image is required to have the consistency property, i.e., retaining the information contained in the input modalities as much as possible. This property motivated many works in i) developing training/inference strategies to adapt to the task and/or ii) designing more efficient and effective networks. We start reviewing previous fusion frameworks, postponing discussions on network designs to the next subsection.

Fusion frameworks mainly rely upon end-to-end learning paradigms, as shown in Fig. 3 (left panel). Previously developed fusion frameworks usually focused on: i) designing effective loss functions [7], [34], [35]; ii) integrating fusion tasks with downstream applications such as detection and segmentation [9], [36]; and iii) incorporating generative priors (e.g., GANs [11], [37] or diffusion models [38]). In terms of loss function design, image pairs are processed through deep neural networks, computing losses starting from the output images, see "Standard fusion loss" in the left panel of Fig. 3. These losses are specifically designed to preserve essential information from each modality (consistency property). For downstream task integration, existing approaches typically adopt task-specific networks either sequentially or in parallel to the fusion network, jointly optimizing them using both fusion and task-specific losses, denoted as "Seg. or Det. loss" in the left panel of Fig. 3. However, this coupled training strategy introduces significant computational overhead and requires manual annotations. Regarding generative priors, previous methods have attempted to incorporate GAN-based cycle training [39] into image fusion processes. Moreover, diffusion models have been utilized by either injecting diffusion priors [13] or controlling diffusion trajectories [40]. However, these complex prior mechanisms often require additional network training or introduce computationally intensive diffusion inference loops.

Moreover, there are a few works related to the use of language as guidance to fusion processes. In particular, TextIF [17] incorporates language information into a two-stream encoder and a unified decoder transformer network, utilizing language instructions for visible-infrared image fusion. Notably, for language features, TextIF adopts a coarse-grained guidance similar to a scale-shift mechanism to inject language information. Similarly, FILM [18] employs language generated by ChatGPT [41] for guidance, utilizing cross-attention to facilitate information exchange between language and image tokens, while neglecting semantic mask guidance. The use of cross-attention introduces a substantial computational burden. Since the language describes the entire image, the input image cannot be cropped in small patches. FILM adopts Restormer attention [42] to mitigate the high memory overhead. However, this kind of attention results in diminished representational capability. The proposed method leverages RWKV to effectively reduce memory and computational overheads, while also incorporating globally aware language modality and objective-level semantic mask information to guide image fusion processes.

2.2 Emerging Trends in Image Fusion Network Design

Network design is crucial for a given image fusion task. Recent works explored the development of fusion-oriented architectures to accomplish this task, including properties such as multi-scale [27], [38], multi-branch [43],

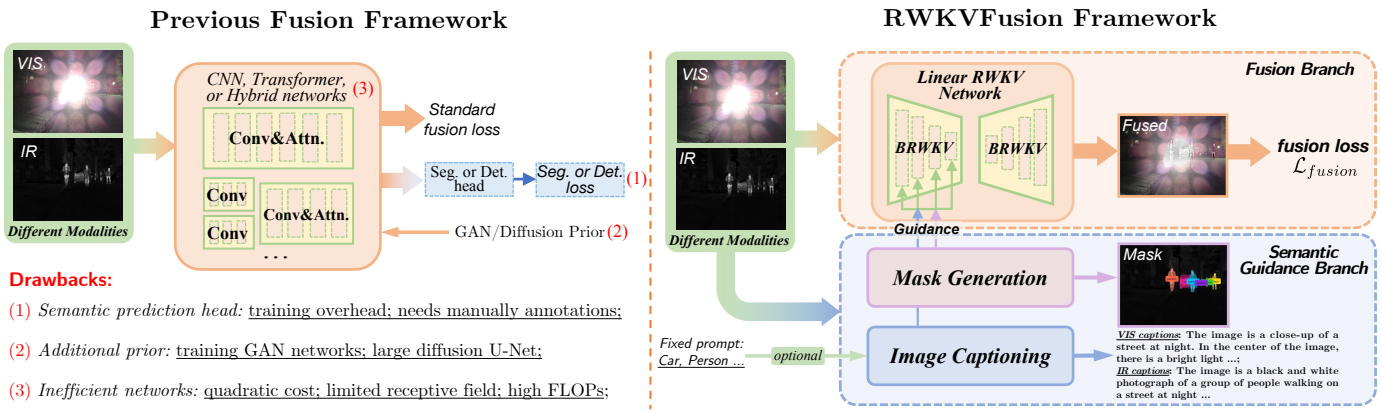


Fig. 3: Comparison between previously developed fusion frameworks and the proposed RWKVFusion. The main differences are as follows: i) some of these frameworks incorporate additional tasks, such as segmentation [9], [15] and object detection [54], requiring semantic information, thus introducing training overhead and costly manual annotations; ii) complex priors lead to the use of further networks [11] or a larger number of parameters and longer inference time [12], [13]; iii) they exploit inefficient networks characterized by high computational demands, restricted receptive fields, and some other limitations [55], [56]. Our RWKVFusion introduces automatic open-set detection and mask generation to include semantic information to guide fusion processes without expensive annotation. VIS and IR stand for visible and infrared.

large-receptive field [44], scale- or fusion size-ware [44], linear-memory or linear-runtime [27], [45], and invertible networks [46], [47]. These state-of-the-art architectures share some similar characteristics as multi-scale and large-receptive properties and task-oriented fusion modules. Some successful pioneering works, as DSPNet [48] and EMMA [49], are multi-scale architectures designed with special feature-gathering operators (e.g., attention [21], Restormer [42], and flatten attention [50]). PanMamba [51], LE-Mamba [27], and CDDFuse [35] have been designed as fusion-oriented modules.

Nevertheless, these approaches often make trade-offs for several reasons, such as the balance between receptive field and memory consumption, the high FLOPs for image fusion tasks, and the time cost during training and inference. Most architectures follow the design philosophy of using convolution at high resolutions and attention at low resolutions, but this still limits network performance. Recently, linear attention mechanisms, such as flatten attention [50] and Restormer attention [42], have emerged to address this issue and have been applied to the design of image fusion architectures [52], [53]. However, although the memory overhead has been reduced from $\mathcal{O}(L^2)$ to $\mathcal{O}(C^2)$, the computational complexity increases from $\mathcal{O}(C^2L)$ to $\mathcal{O}(CL^2)$, where L and C are the token length and the number of channels, respectively. Additionally, state space models represented by VMamba [26] offer linear overhead for global receptive fields. However, their core selective scan mechanism remains controversial [28], especially due to significant long runtime latency. Thus, designing a linear memory/FLOPs overhead, low latency backbone with large receptive fields remains challenging.

2.3 RWKV

A novel linear attention mechanism, named RWKV [29], has recently shown promising results in the field of language modeling. Before introducing RWKV, let us first review the basic form of attention. Unlike traditional RNNs [57], [58], attention utilizes query, $\mathbf{Q}$, key, $\mathbf{K}$, and value, $\mathbf{V}$, matrices to model the relationship between input and output sequences, $\mathrm{Attn}(\mathbf{Q}, \mathbf{K}, \mathbf{V})$. Thus, we have:

$$\mathrm{Attn}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathrm{softmax}(\mathbf{Q}\mathbf{K}^\top)\mathbf{V}, \tag{2}$$

where $\cdot^\top$ is the transpose operator and $\mathrm{softmax}(\cdot)$ is the softmax function. The multi-headness and the scale factor, $1/\sqrt{d_k}$, are omitted (please refer to ViT [21]). Consider two attended positions (t, i), with T token length, the attention score at t can be written in vector form:

$$\mathrm{Attn}(\mathbf{Q}, \mathbf{K}, \mathbf{V})_t = \frac{\sum_{i=1}^{T} e^{\mathbf{q}_t^\top \mathbf{k}_i} \odot \mathbf{v}_i}{\sum_{i=1}^{T} e^{\mathbf{q}_t^\top \mathbf{k}_i}}, \tag{3}$$

where query, $\mathbf{q}_t$, key, $\mathbf{k}_i$, value, $\mathbf{v}_i$, are vectors from $\mathbf{Q}$, $\mathbf{K}$, $\mathbf{V}$, and $\odot$ indicates the element-wise multiplication. By introducing the weight matrix, $\mathbf{W} = \{w_{t,i}\} \in \mathbb{R}^{T\times T}$, with its elements dependent on both i and t, we can transform attention into an RNN:

$$\mathrm{Attn}(\mathbf{W}, \mathbf{K}, \mathbf{V})_t = \frac{\sum_{i=1}^{T} e^{w_{t,i}\mathbf{k}_i} \odot \mathbf{v}_i}{\sum_{i=1}^{T} e^{w_{t,i}+\mathbf{k}_i}}. \tag{4}$$

The previously developed RWKV converts the scalar $w_{t,i}$ into a channel-wise decay vector $\mathbf{w} \in \mathbb{R}^d$ multiplied by the relative position, enhancing the language modeling capability while maintaining the RNN form. The linear RNN model is mainly designed to handle text sequences and can be viewed as a linear attention mechanism with causal masks. Despite recent efforts to extend RWKV to areas, such as image classification [31], image restoration [59], and image generation [60], its use in image fusion is still unexplored.

3 RWKVFusion: Image Fusion with RWKV Backbone Guided by Language and Masks

3.1 Language and Mask Guidance in RWKVFusion

To address the issues raised in Sects. 2.1 and 2.2, we propose a novel approach designed to transfer high-level semantic

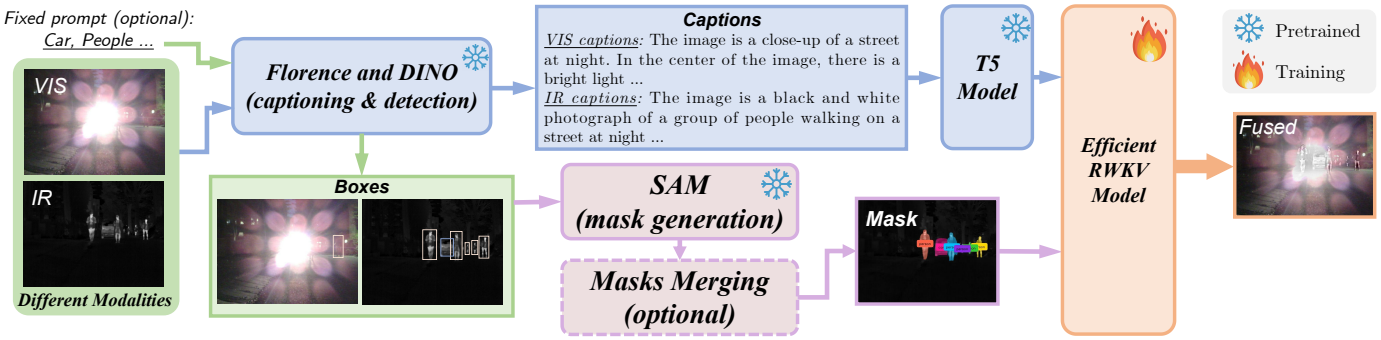


Fig. 4: An overview of the proposed semantic guidance branch. Image pairs of different modalities, optionally with fixed prompts, are input into pre-trained large multi-modal models (DINO [62]) to detect objects and generate captions (using Florence [32]). Object boxes are processed by SAM [33] to produce instance-level masks. A mask merging algorithm is used to merge objects in different masks. The encoded captions from the T5 [61] model, along with the mask, are then fed into the proposed RWKV model to guide fusion processes.

information to fusion tasks. A schematic illustration of our framework is depicted in Fig. 3 (right panel). The framework is divided into two branches: i) a fusion branch with an efficient multi-modal network, and ii) a semantic branch that provides semantic information to the fusion branch. Input images from different modalities (e.g., visible and infrared images) are fed into an efficient network guided by image captions, $\mathbf{T}$, and masks, $\mathbf{M}$, to get the fused image. Thus, the fusion process in (1) is reformulated as follows:

$$\mathbf{F} = \mathcal{F}_{\boldsymbol{\theta}}(\mathbf{S}_1, \cdots, \mathbf{S}_n, \mathbf{T}, \mathbf{M}). \tag{5}$$

For the fusion branch, we propose a pure RWKV network with a linear overhead with respect to token length, see Sect. 3.2 for more details. The semantic branch, serving as fusion guidance, includes image captioning and mask generation. To generate a language description of an image, we use the pre-trained Florence model [32]. The captions are encoded by the pre-trained T5 [61] model.

As detailed in Fig. 4, to segment the objects, a prompt provided by users or Florence is sent to DINO [62] to detect in an open-set manner. Afterward, the mask segmentation is performed starting from these boxes to obtain object-level semantic masks. It is worth to remark that, because of the different information content in the input modalities, the masks derived from the same prompt can differ and get unsatisfactory results. To address this, we introduce a mask merging algorithm to face these discrepancies, see Sect. 3.7 for details. Then, the encoded captions, $\mathbf{T}$, and masks, $\mathbf{M}$, are input into the RWKV encoder to guide fusion processes. The RWKV decoder is used to decode features to get fused outcomes. Finally, we compute the fusion loss, $\mathcal{L}_{fusion}$, using the fused images to update the parameters of the network.

3.2 RWKVFusion Backbone Overview

In this subsection, we detail the fusion backbone of the proposed framework. As shown in Fig. 5, RWKVFusion is a multi-scale encoder-decoder architecture, instead of a plain one. We conducted some experiments, see Sect. 6.1, where the multi-scale architecture is modified to have a plain backbone, similar to SwinIR [63], getting lower performance with respect to the proposed multi-scale architecture.

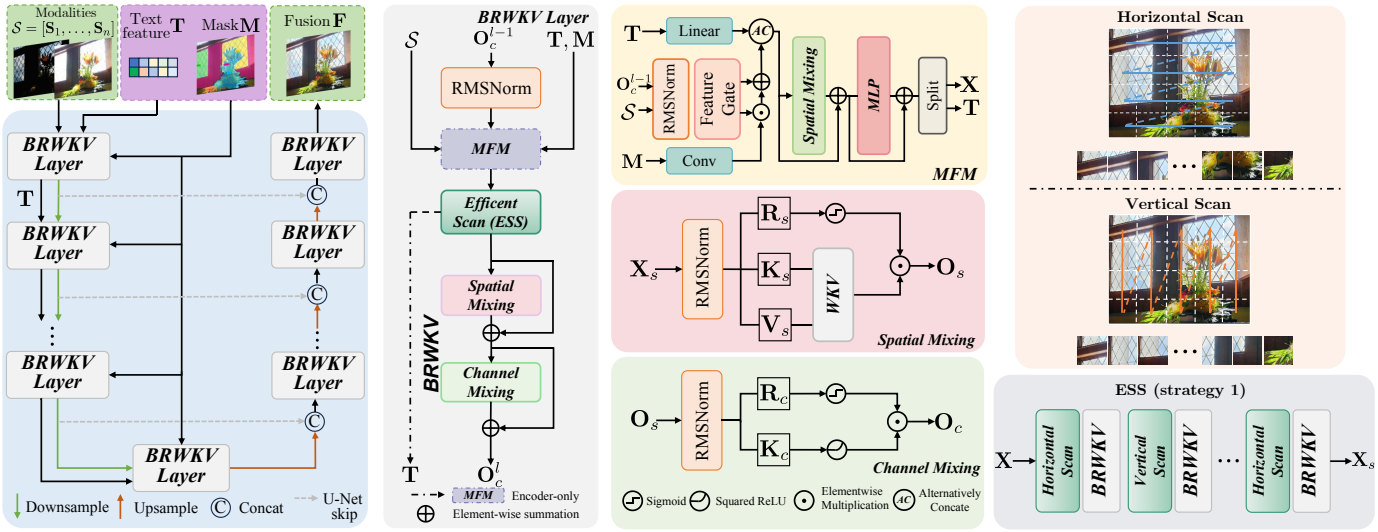
The inputs of the network are images from different modalities, encoded caption features, and semantic masks. We first concatenate all image modalities along channel dimensions before feeding them into the model. The first convolutional layer projects data into a latent space, which is then sent to the encoder to be encoded as features. As shown in Fig. 5(b), the encoder consists of bi-directional RWKV (BRWKV) layers, with each BRWKV layer consisting of a multi-modal fusion module (MFM) and spatial and channel mixing blocks.

It is worth mentioning that our RWKVFusion network can be directly trained on large images without the need for window partitioning/merging [64] while maintaining a relatively low computational burden. An ablation study on using window partitioning/merging is shown in Sect. 5.1. Spatial mixing models the relationship among tokens at the token level, similarly to the attention operation. Channel mixing performs feature fusion along channels, as a feed-forward network. To preserve modality information and semantic guidance, we feed image pairs, captions, $\mathbf{T}$, and masks, $\mathbf{M}$, into each encoder layer as conditions. For this purpose, MFM, see Sect. 3.5 and Fig. 5(c), is designed to guide the fusion process. Each encoder layer concludes with a downsampling layer that has a downsampling factor of 2, achieved through a stridden convolution.

The decoder is quite similar to the encoder but without caption features and mask guidance, since the decoder should focus on decoding fused images. Its input includes not only features from the previous layer but also features from the corresponding encoder layer, weighed by a learnable factor. After being processed via the encoder and decoder, the features are sent to a final convolutional layer that projects them back into the pixel domain to get the fused image.

3.3 Spatial and Channel Mixing

In the following, we will introduce the architecture of the proposed BRWKV and the related motivations. When considering architectural design choices, we typically aim for the spatial operator in the block to have a sufficiently large receptive field. However, pursuing a large receptive field simultaneously leads to an increment of complexity. RWKV, as an alternative to attention, follows a design



(a) The proposed RWKVFusion (b) BRWKV layer (c) MFM, spatial mixing, and channel mixing (d) Horizontal and vertical scans and ESS

Fig. 5: Architecture of the proposed RWKV network. It relies upon a U-Net multi-scale framework, incorporating raw information from each modality, captions, and masks at various layers as conditions to guide the fusion process. Each BRWKV layer has four key components: *i*) multimodal fusion module (MFM), *ii*) efficient scanning strategy (ESS), *iii*) spatial mixing, and *iv*) channel mixing. Note that text feature $\mathbf{T}$ is input from the previous BRWKV encoder layer, but mask $\mathbf{M}$ is a condition injected into each encoder layer.

similar to spatial and channel mixing. The original RWKV is designed for causal language modeling, thus differing from 2D images. To extract information from pixels, we need 2D spatial operators or sequences. Based on this, we first apply an efficient scanning strategy, as introduced in Sect. 3.4, to the input, scanning the 2D feature, $\mathbf{X} \in \mathbb{R}^{H\times W\times C'}$, where H, W, and C' are the three dimensions of the input data, into a 1D sequence stacked along the channel dimension, $\mathbf{X}_s \in \mathbb{R}^{L\times C}$, where $C = n \cdot C'$, $L = H \cdot W$, and n is the scanning ratio. Then, the scanned image sequence is fed into three linear layers with weights $\mathbf{W}_R$, $\mathbf{W}_K$, and $\mathbf{W}_V$, to obtain receptance $\mathbf{R}_s$, key $\mathbf{K}_s$, and value $\mathbf{V}_s$:

$$\mathbf{R}_s = \mathbf{X}_s\mathbf{W}_R, \mathbf{K}_s = \mathbf{X}_s\mathbf{W}_K, \mathbf{V}_s = \mathbf{X}_s\mathbf{W}_V, \tag{6}$$

where the subscript "s" stands for spatial. Then, $\mathbf{K}_s$ and $\mathbf{V}_s$ are fed into the WKV operator, $\mathcal{O}_{WKV}(\cdot,\cdot)$, to obtain WKV global attention $\mathbf{A} \in \mathbb{R}^{L\times C}$, whose factorized vector at position t, $\mathbf{A}_t$, is as follows:

$$\begin{aligned}\mathbf{A}_t &= \mathcal{O}_{WKV}(\mathbf{K}_s, \mathbf{V}_s)_t \\ &= \frac{\sum_{i=1,i\neq t}^{L} e^{-(|t-i|-1)/L\cdot\mathbf{w}+\mathbf{k}_i}\mathbf{v}_i + e^{\mathbf{u}+\mathbf{k}_t}\mathbf{v}_t}{\sum_{i=1,i\neq t}^{L} e^{-(|t-i|-1)/L\cdot\mathbf{w}+\mathbf{k}_i} + e^{\mathbf{u}+\mathbf{k}_t}},\end{aligned} \tag{7}$$

where L is the sequence length, $\mathbf{k}_i$ and $\mathbf{v}_i$ are the i-th token of $\mathbf{K}_s$ and $\mathbf{V}_s$, respectively, and $\mathbf{k}_t$ and $\mathbf{v}_t$ are the t-th token of $\mathbf{K}_s$ and $\mathbf{V}_s$, respectively. $|\cdot|$ is the absolute value operator. $\mathbf{w}$ and $\mathbf{u}$ are C-dimensional learnable parameters, which control the channel-wise spatial decay and the bonus of the current token. It can be noted that the WKV operator models the global attention since it is a weighted summation of $\mathbf{v}_t$ and calculated along the L dimension by explicit decay defined by the relative positions, $(|t-i|-1)/L$, and the key, $\mathbf{k}_i$. Differently from traditional attention, which can output the explicit token-level attention map, the global attention of WKV does not directly yield the same kind of fine-grained attention. It compresses the memory of tokens into global attention, $\mathbf{A}_t$. We will discuss this aspect in Suppl. Sect. 6. After gathering global attention, receptance $\mathbf{R}_s$ is used to gate the attention, and a linear layer is employed to obtain the spatial output. Thus, we have:

$$\mathbf{O}_s = (\sigma(\mathbf{R}_s) \odot \mathbf{A})\mathbf{W}_{O_s}, \tag{8}$$

where $\mathbf{W}_{O_s}$ is the weight matrix of the linear layer, σ denotes the sigmoid function, $\odot$ is the element-wise multiplication operator, and $\mathbf{O}_s$ is the spatial output.

The channel mixing is used to interact with the channel information. We first normalize the spatial output:

$$\mathbf{X}_c = \text{RMSNorm}(\mathbf{O}_s), \tag{9}$$

where RMSNorm$(\cdot)$ stands for the root mean square normalization [65] and $\mathbf{X}_c$ is the normalized spatial output. Subsequently, three linear layers with weights $\mathbf{W}_R$, $\mathbf{W}_K$, and $\mathbf{W}_V$ are used to project features into channel receptance $\mathbf{R}_c$, key $\mathbf{K}_c$, and value $\mathbf{V}_c$:

$$\mathbf{R}_c = \mathbf{X}_c\mathbf{W}_R, \mathbf{K}_c = \mathbf{X}_c\mathbf{W}_K, \mathbf{V}_c = \text{ReLU}^2(\mathbf{K}_c)\mathbf{W}_V. \tag{10}$$

Unlike spatial mixing, the value $\mathbf{V}_c$ is derived from $\mathbf{K}_c$ activated by the squared rectified linear unit, i.e., ReLU$^2(\cdot)$ [66], used to enhance the nonlinearity. Similarly, the value $\mathbf{V}_c$ is gated by sigmoid receptance $\mathbf{R}_c$ and linearly projected into the feature space:

$$\mathbf{O}_c = (\sigma(\mathbf{R}_c) \odot \mathbf{V}_c)\mathbf{W}_{O_c}, \tag{11}$$

where $\mathbf{W}_{O_c}$ is the linear layer's weight matrix and $\mathbf{O}_c$ is the channel output.

BRWKV can be compared with popular attention mechanisms, including standard attention [21], flatten attention [50], window attention [23], and VMamba [26]. Their number of parameters, time, and space consumption are reported in Tab. 1. As can be seen, our BRWKV has a

TABLE 1: Complexity of various vision models. The input image has a sequence length L and C input channels. For window-based sequences, the length is L' of P windows. The hidden dimension is N (i.e., the output channel). We show parameter counts (Params), FLOPs, and space requirements (Space). D indicates the state size for VMamba.

	$k \times k$ Conv	Attention	Swin	Flatten Attn.	VMamba	BRWKV
Params	k^2CN	$3CN + 2N^2$	$3CN + 2N^2$	$3CN + 2N^2$	$3CN + 2N^2$	$4CN + 3N^2$
FLOPs	Lk^2CN	$L^2 + 5LCN$	$P(L'^2 + 5L'CN)$	$3LCN + L^2N + LN^2$	$\frac{1}{16}LN^2 + 2LND$	$26LC + 4LCN$
Space	LN	$L^2 + 3LN$	$P(L'^2 + L'C)$	$3LN + N^2$	$ND + \frac{1}{16}N^2L$	$5LN$

linear space complexity and also demonstrates advantages in terms of time complexity. Quantitative results comparing different attention mechanisms are reported in Sect. 5.1.

We can pass from the summation to the RNN forms of our BRWKV. When we separate the numerator and denominator in (7), we can obtain the following hidden states:

$$\begin{cases} \mathbf{a}_{t-1} = \sum_{i=0}^{t-1} e^{-(|t-i|-1)/L \cdot \mathbf{w} + \mathbf{k}_i} \mathbf{v}_i, \\ \mathbf{b}_{t-1} = \sum_{i=t+1}^{L-1} e^{-(|t-i|-1) L \cdot \mathbf{w} + \mathbf{k}_i} \mathbf{v}_i, \\ \mathbf{c}_{t-1} = \sum_{i=0}^{t-1} e^{-(|t-i|-1)/L \cdot \mathbf{w} + \mathbf{k}_i}, \\ \mathbf{d}_{t-1} = \sum_{i=t+1}^{L-1} e^{-(|t-i|-1)/L \cdot \mathbf{w} + \mathbf{k}_i}. \end{cases} \tag{12}$$

Therefore, the summation form of the WKV operator in (7) can be transformed into the recursively updated form:

$$\mathbf{A}_t = \frac{\mathbf{a}_{t-1} + \mathbf{b}_{t-1} + e^{\mathbf{k}_t + \mathbf{u}} \mathbf{v}_t}{\mathbf{c}_{t-1} + \mathbf{d}_{t-1} + e^{\mathbf{k}_t + \mathbf{u}}}, \tag{13}$$

which highlights the relationship with RNNs. It is worth to remark that an image patch is updated in each step and the whole WKV matrix $\mathbf{A}$ requires L steps.

Considering the size of inputs $\mathbf{K}$ and $\mathbf{V}$ as $L \times C$, the FLOPs of the WKV operator $\mathcal{O}_{WKV}$ are linear with respect to the length of the image patch sequence:

$$\text{FLOPs}(\mathcal{O}_{WKV}) = 2 \times 13 \times L \times C, \tag{14}$$

where factor 2 is related to the stacking along the channel dimension of ESS and factor 13 comes from the computation of the exponential for the hidden states $(\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d})$ updates.

3.4 Efficient Scanning Strategy (ESS) for 2D Images

Since RWKV can only model causal language sequences and modeling 2D images often requires bidirectional attention operations, we introduce an efficient scanning strategy (ESS) to convert 2D images into bidirectional 1D sequences, as shown in Fig. 5(d). Specifically, we scan the image horizontally and vertically, flattening it into 1D sequences, and then we concatenate them along the channel dimension. To extract image information more efficiently, we also scan the image after flipping it vertically and horizontally. Furthermore, scanning can also be performed along the diagonal of the image. This can lead to three variants of the scanning strategies in each BRWKV block:

1) Alternately (with respect to the block index) horizontal and vertical scanning before and after flipping (2 scans in total);
2) Horizontal and vertical scanning before and after flipping (4 scans in total);
3) Horizontal and vertical scanning before and after flipping plus two diagonal scans (8 scans in total).

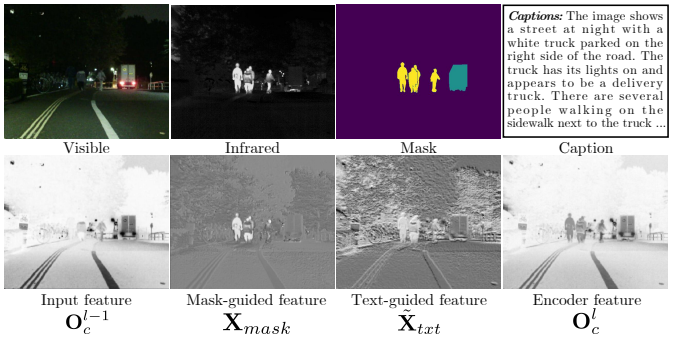


Fig. 6: Feature visualization for the MFM.

The third strategy is the most effective in terms of information aggregation. However, with many scans (i.e., 8), the number of channels grows with a factor of 8, increasing the number of parameters and FLOPs. In practice, using the first strategy, we got a slight performance degradation, while obtaining clear advantages in terms of model parameters and FLOPs. Ablation studies on different scanning strategies are reported in Sect. 5.2.

3.5 Multi-modal Fusion Module (MFM)

MFM takes image pairs from different modalities, $\mathcal{S}$, text features, $\mathbf{T}$, and masks, $\mathbf{M}$, as input, to inject semantic information and guide fusion. In MFM, as shown in Fig. 5(c), there are three paths to guide fusion: i) raw information replenishment; ii) caption information overview guidance; and iii) object-level mask guidance. In i), the input modalities, $\mathcal{S}$, and the features from the previous layer $l-1$, $\mathbf{O}_c^{l-1}$, are processed in a feature gate:

$$\begin{cases} \mathbf{X}_{feat} &= \text{RMSNorm}(\text{Conv}(\mathbf{O}_c^{l-1})), \\ \mathbf{S}_{mod} &= \text{RMSNorm}(\text{Conv}(\mathcal{S})), \\ \mathbf{X}_{act} &= \kappa(\text{AdapPool}(\mathbf{X}_{feat} + \mathbf{S}_{mod})), \\ \mathbf{X}_{feat} &= (\mathbf{X}_{feat} + \mathbf{S}_{mod}) \odot \mathbf{X}_{act}, \end{cases} \tag{15}$$

where AdapPool$(\cdot)$ adaptively pools the input with size of $C \times H \times W$ to $C \times 1$, κ is the GELU activation [67], and Conv$(\cdot)$ is a convolutional layer. In path ii), the mask $\mathbf{M}$ is processed by a convolutional layer, obtaining $\mathbf{M}_{feat}$, and then element-wise multiplied by the sum of the network feature, $\mathbf{X}_{feat}$, and the modality feature, $\mathbf{S}_{mod}$:

$$\begin{cases} \mathbf{M}_{feat} &= \text{Conv}(\mathbf{M}), \\ \mathbf{X}_{mask} &= (\mathbf{X}_{feat} + \mathbf{S}_{mod}) \odot \mathbf{M}_{feat}, \end{cases} \tag{16}$$

where $\mathbf{X}_{mask}$ is the mask-guided feature.

After obtaining both features, they are added to form the feature sequence $\mathbf{X}_{img}$:

$$\mathbf{X}_{img} = \mathbf{X}_{feat} + \mathbf{X}_{mask}. \tag{17}$$

In path iii), $\mathbf{X}_{img}$ is alternately concatenated with the text feature, $\mathbf{T}$, to generate a sequence $\mathbf{X}_{txt}$ that contains image and text features:

$$\mathbf{X}_{txt} = \begin{cases} \text{Concat}(\mathbf{T}, \mathbf{X}_{img}), & \text{if } j \text{ is even}, \\ \text{Concat}(\mathbf{X}_{img}, \mathbf{T}), & \text{if } j \text{ is odd}, \end{cases} \tag{18}$$

where j denotes the layer index and $\text{Concat}(\cdot, \cdot)$ is the concatenation operator. This allows caption $\mathbf{T}$ to be conditioned on both the start and the end of sequences. Furthermore, $\mathbf{X}_{txt}$ is sent to a spatial mixing block to exchange language and image information, and to a multilayer perceptron (MLP) in the channel dimension, finally producing a text-guided feature, $\tilde{\mathbf{X}}_{txt}$. Afterward, captions and image features are split. Captions are sent to the next MFM and image features are fed into the BRWKV block.

To illustrate the effectiveness of MFM, we visualize, in Fig. 6, the features in the first encoder. It can seen that $\mathbf{X}_{mask}$ is more focused on the objects provided by masks. Moreover, $\tilde{\mathbf{X}}_{txt}$ has a global response and objectives are all highlighted. Thus, the semantic and object-level information is injected into the encoded feature at layer l, $\mathbf{O}_c^l$.

3.6 Loss Functions

For different fusion tasks, it is necessary to adopt different supervised/unsupervised loss functions. For HMIF and pansharpening tasks, we exploit the following supervised loss function, $\mathcal{L}_{sharpening}$:

$$\mathcal{L}_{sharpening} = \|\mathbf{F} - \mathbf{GT}\|_1 + \lambda(1 - \text{SSIM}(\mathbf{F}, \mathbf{GT})), \tag{19}$$

where $\mathbf{GT}$ is the ground-truth (GT) image, λ is a weighting coefficient, $\|\cdot\|_1$ is the ℓ_1 norm, and $1 - \text{SSIM}$ is the SSIM loss, with $\text{SSIM}(\cdot, \cdot)$ being the structural similarity index measure [68].

For the VIF, MFF, MEF, and MIF tasks, which fuse two modalities, i.e., $\mathbf{S}_1$ and $\mathbf{S}_2$, we employ the following unsupervised loss function, $\mathcal{L}_{fusion}$:

$$\mathcal{L}_{fusion} = \eta_1 \mathcal{L}_{inten} + \eta_2 \mathcal{L}_{ssim} + \eta_3 \mathcal{L}_{grad}, \tag{20}$$

where η_1, η_2, and η_3 are weighting coefficients. $\mathcal{L}_{inten}$ is the intensity loss calculated as:

$$\mathcal{L}_{inten} = \|\mathbf{F} - \mathbf{S}_1\|_1 + \|\mathbf{F} - \mathbf{S}_2\|_1. \tag{21}$$

Instead, $\mathcal{L}_{ssim}$ is based on two combined SSIM losses calculated as:

$$\mathcal{L}_{ssim} = 2 - \text{SSIM}(\mathbf{F}, \mathbf{S}_1) - \text{SSIM}(\mathbf{F}, \mathbf{S}_2). \tag{22}$$

The SSIM loss is used to measure the structural similarity between the fused image and the input modalities. Finally, $\mathcal{L}_{grad}$ is the gradient loss calculated as:

$$\mathcal{L}_{grad} = \|\nabla\mathbf{F} - \max(\nabla\mathbf{S}_1, \nabla\mathbf{S}_2)\|_1, \tag{23}$$

where ∇ is the Sobel operator, which extracts the edge information of images, and $\max(\cdot, \cdot)$ is the maximum operator.

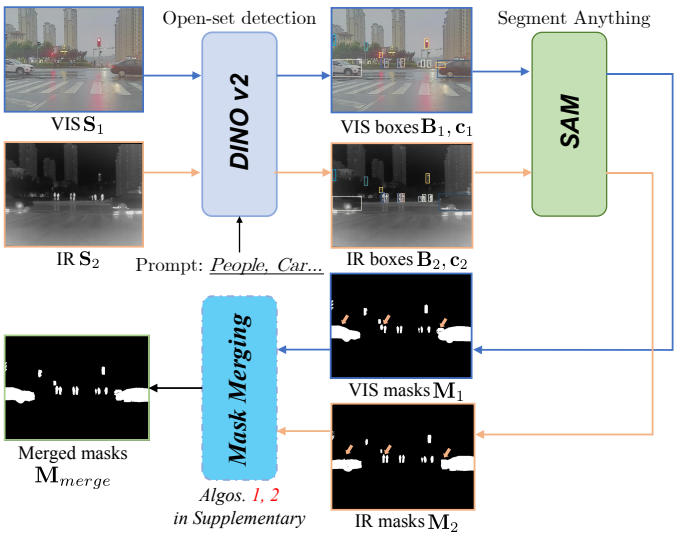


Fig. 7: An overview of mask segmentation and merging. Images from different modalities are fed into DINO and SAM to detect boxes and segment masks in an open-set manner. A novel mask merging algorithm is proposed. Technical details and symbol definitions are reported in Algos. I and II in Suppl. Sect. 3.

3.7 Mask Merging: Dealing with Unsatisfactory Masks

In Sect. 3.1, we introduced a semantic branch (see right panel of Fig. 3) that generates object masks. Because of the varying information content captured by different sensors, given a prompt (prefixed or generated by Florence), derived masks may differ from each others. Moreover, masks are sometimes unsatisfactory to serve as guidance for fusion. To address this problem, we developed an algorithm that merges masks from different image modalities, which makes the merged mask to better reflect the true number and shapes of the objects. Specifically, our approach utilizes predicted intersection over unions (IoUs) of masks to generate high-quality masks, avoiding object duplication or omission. The resulting high-quality mask generation provides a robust foundation for RWKVFusion to perform effective image fusion and combine semantic information from both modalities. The flowchart of the proposed algorithms is shown in Fig. 7. The algorithms for generating and merging image masks are detailed in Algos. I and II in Suppl. Sect. 3.

4 Experimental Analysis

We conducted experiments on six image fusion tasks to validate the effectiveness of the RWKVFusion framework. Additionally, we performed ablation studies and discussions to demonstrate the performance and design rationality of our model. For space limitations, implementation, benchmark details, and more experimental results are provided in Suppl. Sects. 2, 4 and 7.

4.1 Datasets

4.1.1 Data

To evaluate the effectiveness of the proposed framework, we included six tasks involving modalities from: 1) different

imaging sensors (i.e., VIF, MIF, pansharpening, and HMIF) and 2) the same sensor but varying imaging parameters (i.e., MEF and MFF). The datasets are listed as follows:

1) VIF: MSRS, M3FD, and TNO datasets;
2) MIF: Medical Harvard dataset;
3) MEF: SICE and MEFB datasets;
4) MFF: MFI-WHU and RealMFF datasets;
5) Pansharpening: WV3, GF2, and QB datasets;
6) HMIF: Chikusei and Pavia datasets.

Detailed information (links and citations) of each dataset is provided in Suppl. Sect. 1.

4.1.2 Mask Generation

For the VIF task, we manually provided the prompts, "People, Car, Bus, Lamp, Motorcycle, Truck" for the M3FD dataset, and applied the mask merging algorithm (see Suppl. Sect. 3 Algos. I and II and Fig. 7 for a better illustration) to deal with unsatisfactory resulting masks. Since the MSRS dataset provided human-annotated masks, we directly used them. For the MEF, MFF, and MIF datasets, due to varying objects, we utilized the pre-trained Florence model [32] for prompt extraction. When feeding data into the RWKVFusion network, we scatter different objects in the mask into separate channels, setting a maximum channel number (experimentally set to 20) to enable the fusion network to handle masks with varying object numbers. For pansharpening and HMIF tasks, due to the small spatial size of training samples (i.e., 64×64), we omitted mask guidance, thus only applying language guidance. More examples of generated masks and image captions are shown in Suppl. Sect. 8.

4.2 Benchmarks

For the VIF fusion task, we chose recent state-of-the-art fusion approaches for comparison. They can be categorized as follows: i) decomposition methods: DeFuse [55] and CDDFuse [35]; ii) task designed methods: U2Fusion [7], SegMIF [9], and TC-MOA [70]; iii) prior-based methods: DDFM [12]; iv) architectural design methods: SwinFusion [56] and MGDN [69]; v) modality guided methods: FILM [18] and TextIF [17].

For the MIF task, we added MATR [71], which has been specifically designed for this task. Moreover, we exploited: U2Fusion [7], SwinFusion [56], CDDFuse [35], DDFM [12], MGDN [69], and TC-MOA [70].

For the MEF task, we included: U2Fusion [7], DeFuse [55], TC-MOA [70], FILM [18], and IF-MT-SSL [73]. Moreover, we added two approaches specifically designed for this task, i.e., HoLoCo [72] and HSDS [10].

For the MFF task, we added: U2Fusion [7], DeFuse [55], DDFM [12], and TC-MOA [70]. Moreover, we included two approaches specifically designed for this task, i.e., ZMFF [14] and IF-MT-SSL [73].

We provided many widely used quality metrics to assess performance for the VIF, MIF, MEF, and MFF tasks: i) information theory-based metrics: MI, VIF, and SF; ii) human perception inspired metrics: Q_{cb}, Q_y, Q_{cv}, and Q_{abf}; iii) deep model perceptual metrics: LPIPS [74].

For the HMIF task, we considered comparing three traditional model-based methods and nine advanced deep learning-based methods, and the baseline approach based on the simple bicubic interpolation. The traditional model-based methods are: CSTF-FUS [75], LTMR [76], and IR-TenSR [77]. Instead, the deep learning-based methods are: HSRNet [78], MogDCN [79], Fusformer [80], DHIF [81], PSRT [82], 3DT-Net [83], DSPNet [48], BDT [84], and MIMO-SST [85]. The adopted quality metrics are: PSNR, SAM, ERGAS, and SSIM. The computational burden is evaluated by calculating network parameters and FLOPs.

For the pansharpening task, we included three traditional methods in our benchmark, i.e., MTF-GLP-FS [86],

TABLE 2: Performance of recent state-of-the-art fusion methods applied to the VIF and MIF fusion tasks. The best results are in red and the second-best results are in blue.

Methods	MSRS dataset							
	MI↑	VIF↑	SF↑	Q_{cb}↑	Q_{abf}↑	Q_y↑	Q_{cv}↓	LPIPS↓
U2Fusion [7]	1.27	0.67	5.88	0.45	0.24	0.74	611.1	0.874
DeFuse [55]	1.87	0.68	8.13	0.49	0.43	0.90	265.1	0.751
SwinFusion [56]	1.46	0.56	8.09	0.48	0.40	0.79	1270.3	0.788
CDDFuse [35]	1.74	0.57	**11.79**	0.45	0.45	0.85	268.7	0.811
DDFM [12]	1.20	0.59	6.34	0.40	0.22	0.46	535.5	0.961
SegMIF [9]	1.42	0.54	10.59	0.42	0.46	0.65	248.1	0.828
MGDN [69]	1.89	**0.85**	9.66	0.52	0.51	0.90	224.1	**0.685**
TC-MOA [70]	1.93	0.61	9.18	0.50	0.48	**0.92**	**186.9**	0.792
FILM [18]	**3.20**	0.83	11.77	**0.59**	**0.69**	**0.96**	199.0	0.692
TextIF [17]	2.10	0.60	**11.90**	0.53	0.53	**0.92**	197.6	0.754
Proposed	**3.42**	**0.87**	11.50	**0.60**	**0.70**	**0.96**	**182.2**	**0.677**
Methods	**M3FD dataset**							
U2Fusion [7]	1.89	0.55	4.15	0.42	0.14	0.79	658.8	0.835
DeFuse [55]	1.99	0.70	7.45	0.43	0.34	0.88	511.7	0.693
SwinFusion [56]	1.92	0.65	10.71	0.42	0.46	0.91	537.6	0.650
CDDFuse [35]	**2.63**	0.70	14.7	0.48	0.62	**0.95**	436.7	**0.637**
DDFM [12]	1.84	0.51	8.65	0.40	0.33	0.81	511.7	0.693
SegMIF [9]	1.87	0.59	13.46	0.41	0.58	0.72	556.5	0.753
MGDN [69]	1.97	0.66	10.22	0.43	0.47	0.91	557.7	0.661
TC-MOA [70]	1.90	0.55	10.00	0.46	0.49	0.92	**385.4**	0.850
FILM [18]	2.42	0.52	**15.00**	0.47	0.55	0.93	424.0	0.779
TextIF [17]	2.54	**0.72**	**15.53**	**0.52**	**0.68**	0.93	395.0	0.661
Proposed	**2.57**	**0.78**	14.46	**0.54**	**0.70**	**0.96**	**358.9**	**0.647**
Methods	**TNO dataset**[1]							
U2Fusion [7]	1.42	0.65	6.37	0.48	0.33	0.83	508.4	0.751
DeFuse [55]	1.78	**0.70**	5.98	0.47	0.35	0.88	382.4	0.781
SwinFusion [56]	1.26	0.57	8.05	**0.49**	0.43	0.85	503.1	0.749
CDDFuse [35]	**2.24**	0.68	10.76	0.45	0.50	0.89	366.3	0.752
DDFM [12]	1.36	0.59	5.69	0.42	0.24	0.79	531.7	0.935
SegMIF [9]	1.97	0.63	**11.97**	0.47	0.54	0.76	427.9	0.789
MGDN [69]	1.43	0.65	7.29	0.47	0.34	0.84	463.9	**0.736**
TC-MOA [70]	1.70	0.59	6.94	0.48	0.40	0.90	315.7	0.933
TextIF [17]	2.17	0.64	**11.22**	**0.52**	**0.56**	**0.91**	**278.8**	0.749
Proposed	**2.67**	**0.74**	11.07	**0.52**	**0.63**	**0.94**	**289.0**	**0.688**
Methods	**Medical Harvard dataset**							
U2Fusion [7]	1.89	0.53	11.60	0.26	0.32	0.57	523.8	0.646
SwinFusion [56]	1.78	0.49	12.0	0.57	0.29	0.86	1011.2	**0.532**
MATR [71]	1.71	0.46	14.54	0.26	0.44	0.53	416.1	0.766
CDDFuse [35]	1.88	0.43	**21.77**	0.52	0.53	0.84	335.2	0.548
DDFM [12]	1.89	0.46	12.51	0.52	0.34	0.75	513.9	0.556
MGDN [69]	**1.97**	**0.55**	17.70	**0.58**	**0.54**	**0.88**	341.8	**0.530**
TC-MOA [70]	1.95	0.53	13.94	0.55	0.52	**0.88**	**325.2**	0.577
Proposed	**2.02**	**0.57**	**22.03**	**0.60**	**0.63**	**0.90**	**300.2**	0.534

1. FILM does not provide the encoded text guidance (by ChatGPT) for the TNO and MFI-WHU datasets. Thus, we did not compare with it.

TABLE 3: Performance of recent state-of-the-art fusion methods applied to the MEF and MFF fusion tasks. The best results are in red and the second-best results are in blue.

Methods	SICE dataset							
	MI↑	VIF↑	SF↑	Q_{cb}↑	Q_{abf}↑	Q_y↑	Q_{cv}↓	LPIPS↓
U2Fusion [7]	**3.63**	**0.98**	16.97	0.37	0.50	0.75	260.4	0.667
DeFuse [55]	2.57	0.53	19.12	0.34	0.49	0.85	228.6	0.760
HoloCo [72]	2.38	0.48	16.88	**0.40**	0.43	0.75	265.8	0.803
HSDS [10]	2.42	0.49	24.18	**0.41**	0.52	0.74	217.8	0.762
IF-MT-SSL [73]	1.61	0.28	24.32	0.35	0.25	0.56	400.1	0.747
TC-MOA [70]	2.64	0.63	14.50	**0.41**	0.55	0.77	231.1	0.778
FILM [18]	**3.81**	0.86	**28.97**	0.38	**0.75**	**0.94**	**157.9**	**0.620**
Proposed	**3.81**	**0.93**	**29.04**	0.38	**0.78**	**0.96**	**135.8**	**0.597**
Methods	**MEFB dataset**							
U2Fusion [7]	3.93	**1.04**	11.53	0.39	0.51	0.76	356.9	0.673
DeFuse [55]	3.13	0.75	13.82	0.36	0.54	**0.91**	307.6	0.686
HoloCo [72]	2.87	0.70	12.97	**0.42**	0.53	0.82	343.2	0.716
HSDS [10]	2.84	0.68	18.30	**0.43**	0.66	0.79	277.4	0.695
IF-MT-SSL [73]	2.10	0.39	17.23	0.39	0.34	0.68	606.0	0.734
TC-MOA [70]	3.19	0.77	12.30	**0.43**	0.58	0.81	268.9	0.689
FILM [18]	**3.95**	0.85	**21.04**	**0.42**	**0.75**	0.90	**235.6**	**0.666**
Proposed	**3.97**	**0.91**	**20.31**	**0.43**	**0.78**	**0.94**	**204.8**	**0.622**
Methods	**MFI-WHU dataset**[1]							
U2Fusion [7]	**4.42**	**1.12**	15.77	**0.69**	0.60	**0.93**	**59.0**	**0.296**
DeFuse [55]	4.12	0.99	13.34	0.67	0.49	0.89	80.1	0.331
DDFM [12]	4.20	1.04	15.82	0.65	0.59	0.89	66.9	0.331
ZMFF [14]	3.58	0.87	**23.11**	0.65	**0.61**	0.92	179.0	0.428
IF-MT-SSL [73]	3.25	0.52	**24.66**	0.55	0.31	0.72	437.1	0.388
TC-MOA [70]	3.96	0.97	14.08	0.67	0.54	0.91	60.1	0.361
Proposed	**4.72**	**1.13**	22.58	**0.73**	**0.68**	**0.98**	**46.79**	**0.302**
Methods	**RealMFF dataset**							
U2Fusion [7]	4.47	**1.27**	12.72	0.68	**0.69**	0.94	70.3	**0.228**
DeFuse [55]	4.36	1.23	10.59	0.59	0.67	0.93	97.7	0.252
DDFM [12]	4.28	1.21	13.59	0.67	0.55	0.75	75.6	0.313
ZMFF [14]	4.37	1.16	15.14	**0.72**	0.67	**0.96**	60.3	0.247
IF-MT-SSL [73]	3.24	0.59	**15.66**	0.35	0.56	0.80	471.2	0.305
TC-MOA [70]	4.19	1.14	11.47	0.61	0.68	0.94	67.3	0.277
FILM [18]	**4.70**	1.22	**16.09**	0.70	**0.73**	**0.96**	**52.0**	0.235
Proposed	**4.90**	**1.32**	14.62	**0.74**	**0.73**	**0.97**	**45.60**	**0.200**

BT-H [87], and LRTCFPan [88], four CNN-based methods, i.e., DiCNN [89], FusionNet [90], LAGConv [91], and DCFNet [43], one transformer-based method (i.e., Invformer [92]), one model-driven method (i.e., HMP-Net [93]), one diffusion-based method (i.e., PanDiff [94]), and one Mamba-based approach (i.e., PanMamba [51]). As suggested in [95], we considered SAM, ERGAS, Q2n, and SCC as quality metrics at reduced resolution. Instead, we used D_λ, D_s, and HQNR for tests at full resolution.

All adopted metrics will be discussed in Suppl. Sect. 4 together with the implementation details of the methods used in our benchmarks.

4.3 Results for VIF and MIF

We evaluate the performance of our model for the VIF task using three datasets: MSRS, M3FD, and TNO. Tab. 2 reports the quantitative comparison with some state-of-the-art methods, divided into three sub-tables. Regarding the MSRS dataset, the proposed RWKVFusion achieved the best results on seven out of eight quality metrics. The same outstanding performance is obtained on the M3FD and TNO datasets, where our model ranked first or second on seven out of eight metrics, with just a slight decrease in performance with respect to the best values for SF.

Additionally, in Fig. 8, we depicted in the first two rows some results related to the VIF task (referring to the M3FD dataset). Having a look at them, we can observe two people obscured by smoke and buildings in the background. Among the fusion results of various methods, U2Fusion [7], DeFuse [55], SwinFusion [56], and MGDN [69] produce low quality outputs with blurry images. Although DDFM [12], SegMIF [9], and TC-MOA [70] effectively reveal the infrared targets in the smoke, their results suffer from either excessive brightness or darkness, along with severe color distortions when compared to the visible images. FILM [18] and TextIF [17], enhanced by language models, obtained relatively better fusion quality, but they fail to clearly preserve the buildings obscured by the smoke. In contrast, our method fully integrated complementary information from the source images, offering a more comprehensive depiction of the objects obscured by the smoke.

The experimental results related to the MIF task using the Medical Harvard dataset are reported in the fourth subtable of Tab. 2. In terms of numerical comparisons across the various adopted metrics, RWKVFusion achieved the best performance in all metrics except LPIPS. Although our LPIPS score is not the best one, it is only 0.04 and 0.02 lower than the best [69] and the second-best [56] results, respectively. Some visual analyses are provided in Suppl. Sect. 9.

4.4 Results for MEF and MFF

The quantitative assessment on the selected MEF and MFF datasets is reported in Tab. 3. The proposed method achieves the best performance in most of the metrics (i.e., six or seven out of eight quality metrics) for all four datasets. Compared with the related state-of-the-art methods, our approach demonstrates a better fusion capability.

In the middle and lower parts of Fig. 8, we present the qualitative comparisons for the MEF and MFF tasks. For the MEF task, we presented an instance of the SICE dataset. It can be observed that most of the compared methods, such as U2Fusion [7] and HoLoCo [72], achieve complementary exposure fusion but lack global exposure consistency. Compared to the most recent methods as TC-MOA [70] and FILM [18], our approach has a better balance between high-exposure and low-exposure regions, as demonstrated by the head of the sculpture in the close-up and the further sculpture. For the MFF task, we considered the RealMFF dataset in Fig. 8. Several methods, including ZMFF [14] and TC-MOA [70], lose details with respect to the source images, such as the text on the sign in the foreground. In contrast, our method demonstrates superior fidelity. The more precise fusion guidance employed by our approach makes it stand out among the compared state-of-the-art methods.

4.5 Results for Pansharpening and HMIF

Tab. 4 and Tab. 3 in Suppl. Sect. 7 report the quantitative assessment for the pansharpening task using three datasets: WV3 (8 bands), GF2 (4 bands), and QB (4 bands). For the WV3 dataset, our RWKVFusion achieved state-of-the-art

Fig. 8: Visual comparisons on the M3FD, SICE, and RealMFF datasets. Close-ups are depicted in the green and blue boxes.

performance in both the reduced resolution (RR) and full resolution (FR) assessments. It is worth to be remarked that our RWKVFusion shows outstanding performance whatever the dataset, thus pointing out its robust generalization ability. Fig. 9 (left panel) illustrates the error maps generated by the compared methods on the WV3 dataset. The comparison of these error maps clearly demonstrates that the proposed RWKVFusion method obtains residual maps that are predominantly blue, i.e., with the lowest brightness and, thus, the best results. As shown in Fig. 9, the magnified areas in red and green boxes (representing buildings) demonstrate that our RWKVFusion excels at preserving the fine structural details from the panchromatic images and the rich spectral information from the low resolution multispectral images. This is particularly evident in the transition regions with significant spectral differences.

Finally, we also evaluate the performance of the proposed RWKVFusion for the HMIF task. The results of our model compared to recent state-of-the-art models are reported in Tab. 5 calculated on both the Chikusei and the Pavia test sets. The proposed approach demonstrates significant advantages across all metrics. More specifically, on the Chikusei dataset, our model shows top performance for all quality indexes. It is worth mentioning that the proposed approach achieves these superior results with only 8.41% of the parameters and 0.67% of the FLOPs with respect to the second-ranked DHIF method [81]. Similarly, on the Pavia dataset, our approach always outperforms the second-best method, i.e., DSPNet [48], while using one-third of the parameters and half of the FLOPs. Error maps are depicted in Fig. 9 (right panel). They corroborate the numerical assessment, highlighting the superior visual performance of our model. Compared with the other methods, our model excels at fusing fine details and retains spectral features of the low resolution hyperspectral cube.

TABLE 4: The averages and standard deviations of the adopted quality metrics for the pansharpening task calculated on the WV3 test set. The best results are in red and the second-best results are in blue.

	Methods	Reduced Resolution (RR): Avg±std				Full Resolution (FR): Avg±std			#Params↓	#FLOPs↓
		SAM↓	ERGAS↓	Q2n↑	SCC↑	D_λ↓	D_s↓	HQNR↑		
WorldView-3 (WV3, 8-band)	MTF-GLP-FS [86]	5.32±1.65	4.65±1.44	0.818±0.101	0.898±0.047	0.021±0.008	0.063±0.028	0.918±0.035	—	—
	BT-H [87]	4.90±1.30	4.52±1.33	0.818±0.102	0.924±0.024	0.057±0.023	0.081±0.037	0.867±0.054	—	—
	LRTCFPan [88]	4.74±1.41	4.32±1.44	0.846±0.091	0.927±0.023	**0.018±0.007**	0.053±0.026	0.931±0.031	—	—
	DiCNN [89]	3.59±0.76	2.67±0.66	0.900±0.087	0.976±0.007	0.036±0.011	0.046±0.018	0.920±0.026	0.23M	0.19G
	FusionNet [90]	3.33±0.70	2.47±0.64	0.904±0.090	0.981±0.007	0.024±0.009	**0.036±0.014**	**0.941±0.020**	0.047M	0.32G
	LAGConv [91]	3.10±0.56	2.30±0.61	0.910±0.091	0.984±0.007	0.037±0.015	0.042±0.015	0.923±0.025	0.15M	0.54G
	Invformer [92]	3.25±0.64	2.39±0.52	0.906±0.084	0.983±0.005	0.055±0.029	0.068±0.031	0.882±0.049	0.14M	3.89G
	DCFNet [43]	3.03±0.74	**2.16±0.46**	0.905±0.088	**0.986±0.004**	0.078±0.081	0.051±0.034	0.877±0.101	2.77M	3.46G
	HMPNet [93]	3.06±0.58	2.23±0.55	**0.916±0.087**	0.986±0.005	**0.018±0.007**	0.053±0.006	0.929±0.011	1.09M	2.00G
	PanDiff [94]	3.30±0.60	2.47±0.58	0.898±0.088	0.980±0.006	0.027±0.012	0.054±0.026	0.920±0.036	45.33M	14.83G
	PanMamba [51]	**2.94±0.54**	2.24±0.51	0.916±0.090	0.985±0.006	0.020±0.007	0.042±0.014	0.939±0.020	0.48M	1.31G
	Proposed	**2.78±0.52**	**2.03±0.43**	**0.918±0.083**	**0.988±0.003**	**0.016±0.006**	**0.036±0.005**	**0.949±0.009**	1.21M	2.34G

TABLE 5: The averages and standard deviations of the adopted quality metrics for the HMIF task calculated on the Chikusei and the Pavia Centre test sets. The best results are in red and the second-best results are in blue.

Methods	**Chikusei ×4 HMIF Dataset**						**Pavia ×4 HMIF Dataset**					
	PSNR↑	SAM↓	ERGAS↓	SSIM↑	#Params↓	#FLOPs↓	PSNR↑	SAM↓	ERGAS↓	SSIM↑	#Params↓	#FLOPs↓
Bicubic	33.35±2.14	4.00±0.37	7.65±0.48	0.815±0.044	–	–	26.65±0.06	7.07±0.20	8.46±0.09	0.614±0.004	–	–
CSTF-FUS [75]	35.40±2.48	5.40±0.60	7.88±0.71	0.844±0.049	–	–	30.93±0.01	11.08±0.14	5.74±0.09	0.791±0.001	–	–
LTMR [76]	41.21±3.66	2.98±0.86	4.84±1.23	0.950±0.031	–	–	32.33±0.15	6.35±0.23	5.10±0.05	0.820±0.003	–	–
IR-TenSR [77]	36.00±0.42	5.12±0.48	7.86±0.05	0.868±0.045	–	–	30.87±0.11	6.81±0.25	5.82±0.01	0.783±0.003	–	–
HSRNet [78]	42.01±0.95	2.33±0.24	3.95±0.29	0.947±0.009	0.633M	3.041G	32.17±0.17	5.60±0.18	4.60±0.05	0.867±0.003	2.061M	2.677G
MogDCN [79]	42.21±1.00	2.27±0.23	3.76±0.29	0.936±0.009	6.840M	53.507G	33.84±0.25	4.61±0.20	4.07±0.08	0.889±0.003	7.202M	51.743G
Fusformer [80]	43.37±1.02	2.03±0.19	3.49±0.28	0.959±0.006	0.504M	10.315G	35.31±0.32	4.33±0.18	3.37±0.07	0.924±0.003	0.539M	10.263G
DHIF [81]	**43.69±1.05**	**1.94±0.19**	**3.33±0.28**	0.960±0.007	22.462M	466.313G	35.30±0.38	4.36±0.20	3.35±0.10	0.924±0.002	38.785M	311.194G
PSRT [82]	43.48±0.96	2.01±0.19	3.47±0.25	**0.961±0.006**	0.303M	1.367G	34.86±0.44	4.47±0.20	3.54±0.14	0.916±0.001	0.288M	1.304G
3DT-Net [83]	43.53±1.01	2.03±0.19	3.46±0.28	**0.963±0.006**	3.464M	75.352G	35.10±0.38	4.44±0.19	3.35±0.09	**0.927±0.002**	3.482M	73.299G
DSPNet [48]	43.55±0.98	2.03±0.20	3.44±0.25	0.960±0.007	6.138M	7.125G	**35.47±0.43**	**4.26±0.21**	**3.30±0.12**	**0.927±0.002**	6.115M	7.031G
BDT [84]	43.25±1.06	2.09±0.21	3.44±0.29	0.955±0.008	3.263M	4.372G	34.55±0.35	4.66±0.22	3.70±0.11	0.904±0.002	3.056M	3.569G
MIMO-SST [85]	43.36±1.02	2.09±0.23	3.48±0.27	0.958±0.007	4.983M	2.505G	35.37±0.39	4.48±0.17	3.34±0.10	0.922±0.002	5.227M	2.248G
Proposed	**43.89±1.10**	**1.93±0.18**	**3.33±0.28**	**0.963±0.006**	1.888M	3.134G	**36.06±0.41**	**3.95±0.17**	**3.07±0.08**	**0.936±0.003**	1.888M	3.134G

5 Ablation Study

5.1 BRWKV and Attention Mechanisms

In this section, we compare the proposed BRWKV with other related mechanisms as flash attention [97], flatten attention [50], window attention [64], and VMamba [26]. To ensure a fair comparison, we replaced BRWKV with the aforementioned attention operations while maintaining all other configurations. We maintained consistent hidden channel dimensions across all variants. We evaluated these model variants on the Pavia $\times 4$ and MSRS datasets, and the quantitative results are reported in Tab. 6(a)-(d). The results indicate that flash attention, benefiting from its global receptive field and optimized CUDA operators, achieves fusion performance close to that of our RWKVFusion. However, it has a quadratic cost with respect to token length. Flatten attention, while reducing memory consumption, limits the ability of attention in the $C \times C$ attention map, resulting in poorer fusion performance. VMamba, on the other hand, stores information in states, balancing expressive power and cost, leading to moderate performance. Our RWKVFusion outperforms these attention variants in almost all metrics.

By applying window partition and window merge operations to the feature map before and after the BRWKV module, we transform the global BRWKV into a windowed BRWKV, thereby reducing the computational cost. We replaced each global BRWKV with its windowed counterpart, and the fusion results are reported in Tab. 6(e). Although the performance of the windowed BRWKV is lower than that of the global BRWKV, it still outperforms the window attention. Hence, we recommend using this variant in resource-limited scenarios.

5.2 Scanning Types

In Sect. 3.4, we proposed three scanning strategies to address the limitation of the original RWKV [29] in modeling relationships between 2D image tokens. It is worth noting that scanning along different directions introduces inductive bias, which can enhance the model's learning capabilities. However, this approach has a clear drawback, i.e., it increases the number of channels, potentially compromising efficiency. The results are reported in Tab. 6(f) and (g). Compared to using the default scanning configuration, the performance of Scan Config 2) is slightly inferior. Although it offers more scanning directions, the excessive number of scanning directions can disrupt the network's spatial awareness, leading to a decline in performance. As for Scan Config 3), its performance is slightly better. We believe this is because the additional scanning directions further increase the network's channel capacity, and the benefits of over-parameterization offset the negative impact of too

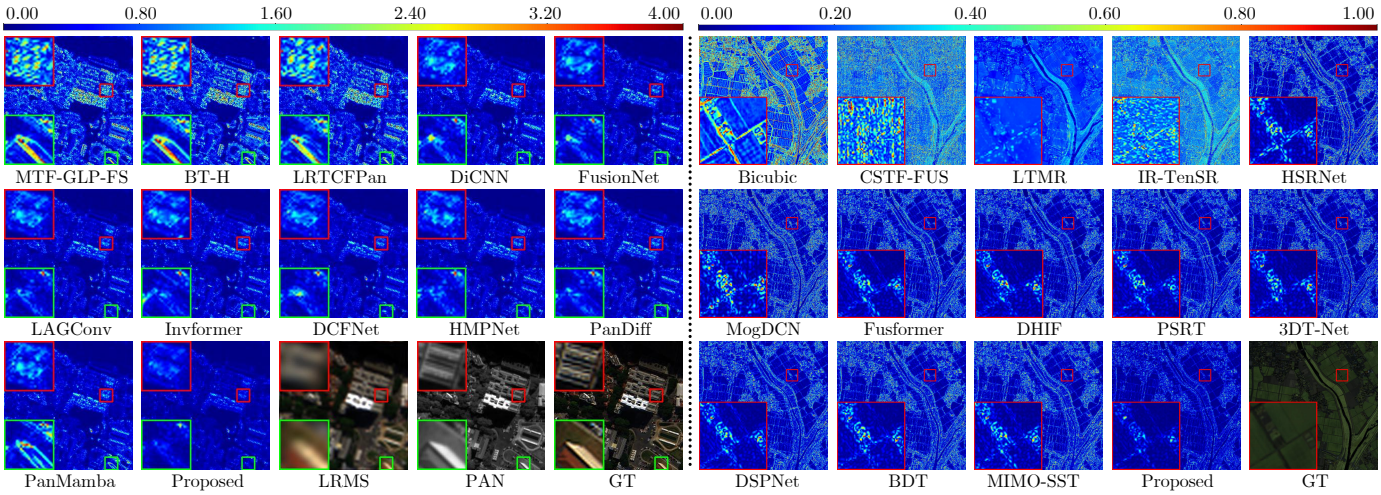


Fig. 9: Error maps with respect to the GT image for the compared approaches considering the pansharpening and the HMIF tasks. The left and right panels respectively showcase the error maps of "*area 5*" from the WV3 test set and "*area 1*" from the Chikusei ($\times$ 4) test set using a pseudo-color representation. Close-ups are depicted in the red and green boxes. PAN and LRMS stand for panchromatic and low resolution multispectral images, respectively.

TABLE 6: Ablation studies about attention operators, scanning strategies, different MFM designs, and plain backbone, for the VIF and HMIF tasks. The best results are in red, and the second-best results are in blue.

Ablations	Pavia $\times 4$ HMIF Dataset				MSRS VIF Dataset							
	PSNR↑	SAM↓	ERGAS↓	SSIM↑	MI↑	VIF↑	SF↑	Q_{cb}↑	Q_{abf}↑	Q_y↑	Q_{cv}↓	LPIPS↓
(a) Flash Attn. [97]	36.04	**3.95**	**3.06**	**0.936**	**3.44**	0.81	11.33	0.55	0.62	**0.96**	185.8	0.685
(b) Flatten Attn. [50]	35.70	4.13	3.17	0.932	3.21	0.77	10.18	0.54	0.66	0.86	198.4	0.699
(c) Window Attn. [64]	35.86	4.13	3.20	0.929	3.40	0.82	11.12	**0.60**	0.69	0.88	207.3	0.682
(d) VMamba [26]	35.61	4.17	3.18	0.930	3.32	0.79	11.03	0.53	0.64	0.89	199.7	0.701
(e) Window BRWKV	35.89	4.11	3.15	0.932	3.41	0.82	11.36	0.59	0.66	0.93	195.0	0.679
(f) Scan Config 2)	35.90	4.09	3.14	**0.933**	3.36	0.80	10.97	0.57	0.67	**0.94**	200.7	0.680
(g) Scan Config 3)	**36.09**	**3.92**	**3.05**	**0.936**	3.41	**0.88**	11.49	**0.61**	**0.71**	**0.96**	**178.1**	**0.675**
(h) Simple MLPs	35.68	4.16	3.19	0.931	3.25	0.77	10.36	0.57	0.68	0.88	195.7	0.689
(i) Cross attentions	36.00	3.97	3.07	**0.936**	3.40	**0.88**	**11.53**	0.58	0.68	0.91	192.3	0.680
(j) Plain backbone	35.29	4.30	3.29	0.929	3.12	0.71	9.81	0.49	0.62	0.81	231.9	0.710
(k) Default	**36.06**	**3.95**	3.07	**0.936**	**3.42**	**0.87**	**11.50**	**0.60**	**0.70**	**0.96**	**182.2**	**0.677**

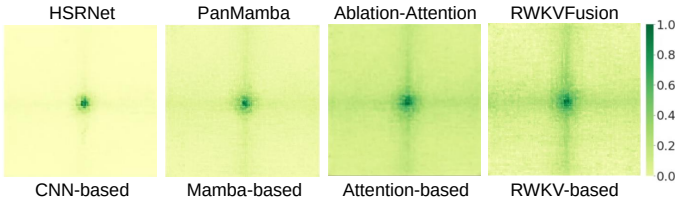


Fig. 10: Effective receptive fields (ERF) for architectures based on different operators (i.e., convolution, Mamba, attention, and RWKV) are visualized.

many scanning directions. However, this comes with greater computational costs. Therefore, the default configuration is selected as the optimal one.

5.3 MFM Design

In Sect. 3.5, we proposed an MFM to fuse language and semantic mask guidance. To investigate the effectiveness of this module, we performed ablation studies as follows:

1) Simple MLPs. We replaced the MFM module with two simple three-layer MLPs, each consisting of linear+RMSNorm+ReLU layers. The feature $\mathbf{O}_c^{l-1}$ and modalities $\mathbf{S}$ are concatenated and fed into one MLP, while the caption $\mathbf{T}$ is input to the other MLP. The outputs of both MLPs are concatenated and passed through a linear layer to produce the output of the MFM variant.
2) Cross-attentions. We substituted the RWKV's spatial mixing mechanism with two cross-attentions following Flux.1-dev [98]. In the first cross-attention, the caption features, $\mathbf{T}$, act as queries, while the image features, $\mathbf{X}_{img}$, defined in (17), serve as keys and values. Instead, the second cross-attention inverts the relationship, i.e., image features query caption features.

We provide numerical results in Tab. 6(h) and (i). The default MFM outperforms the compared fusion modules.

5.4 Semantic Guidance and Mask Merging

In Sect. 3.1, we proposed to enhance image fusion by incorporating global caption information and semantic mask guidance into our model. To evaluate the relationship between these two inputs, we conducted experiments under

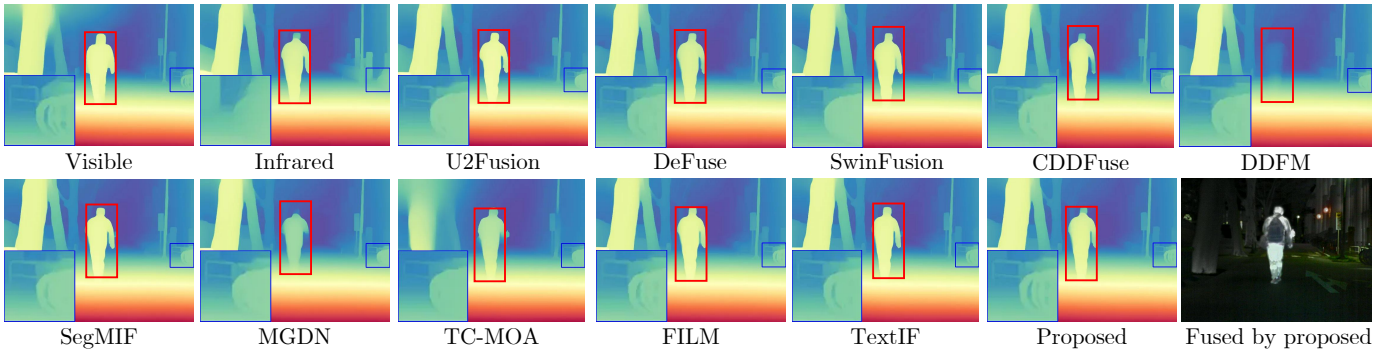


Fig. 11: Monocular depth estimation results by using Depth Anything v2 [96]. Close-ups are depicted in the blue boxes.

TABLE 7: Ablation studies on the use of language and semantic mask guidance. The best results are in red.

Ablation	MSRS VIF Dataset							
	MI↑	VIF↑	SF↑	Q_{cb}↑	Q_{abf}↑	Q_y↑	Q_{cv}↓	LPIPS↓
i) Only caption	3.20	0.76	9.93	0.54	0.64	0.91	204.8	0.685
ii) Only merged mask	3.36	0.78	11.01	0.56	0.67	0.93	196.5	0.680
iii) None of both	3.10	0.69	9.72	0.48	0.62	0.79	234.7	0.711
iv) Mask concat	3.38	0.84	11.33	0.57	0.67	0.95	189.7	0.679
v) Default	**3.42**	**0.87**	**11.50**	**0.60**	**0.70**	**0.96**	**182.2**	**0.677**

the following settings: *i*) using only the caption; *ii*) using only the merged mask; *iii*) using no guidance; *iv*) using *unmerged* mask and concatenating them along the channel dimensions (while still using caption); and *v*) default setting, using both inputs.

The validation is performed on the MSRS dataset, see Tab. 7. It can be observed that the best performance is obtained by *v*, i.e., the default setting. The fusion performance was reduced in both (*i*) and (*ii*) because they only utilized partial semantic information. This demonstrates that language and semantic mask guidance can mutually improve the fusion performance. For configuration (*iv*), the mask-merging technique was not used, but it still introduces caption information. While this approach retains the raw mask features, it *may introduce redundancy and potential misalignment* by comparing with merged masks, particularly when masks overlap or are erroneously segmented in spatial regions, which is the initial motivation of the mask merging technique. As a result, its performance was also lower compared to the default setting (i.e., using both caption and merged-mask guidance).

5.5 The Settings of Prompting for Masking

To clarify the impact of using an optional fixed prompt for masking, we conducted an ablation study comparing two different prompting strategies: auto-prompt and fixed-prompt. The auto-prompt (default for MEF, MFF, and MIF tasks) uses the text descriptions generated by the Florence model to guide the SAM segmentor in generating masks, enabling open-set segmentation. The fixed-prompt (default for VIF tasks) uses a pre-defined prompt relevant to common objects within the evaluation dataset for more targeted segmentation guidance.

Ablation results on the M3FD dataset are provided in Tab. 8. Fixed-prompt slightly outperforms the auto-prompt.

TABLE 8: Ablation studies on the settings of mask prompting in Sect. 3.1. The best results are in red.

Ablation	M3FD VIF Dataset							
	MI↑	VIF↑	SF↑	Q_{cb}↑	Q_{abf}↑	Q_y↑	Q_{cv}↓	LPIPS↓
Auto-prompt	2.55	0.75	**14.50**	0.53	0.69	**0.96**	360.3	**0.647**
Fixed-prompt	**2.57**	**0.78**	14.46	**0.54**	**0.70**	**0.96**	**358.9**	**0.647**

We attribute this to the fact that the fixed-prompt targets human-interested and easy-to-find objects, enabling more precise guidance for the fusion model, while the auto-prompt can benefit from the open-set detection but may suffer from inaccurate detection, affecting subsequent segmentation and mask injection, leading to a slight performance decrease. Nevertheless, under both prompt settings, our method still outperforms previous methods, demonstrating that injecting semantic information from masks can effectively enhance fusion performance.

6 Discussion

6.1 Plain Against Multi-scale Backbones

One of the main contributions of this paper is the proposed multi-scale design. However, non-multi-scale architectures, which we refer to as plain backbones, are also quite popular. We believe that multi-scale architectures are of crucial importance for low-level tasks, such as image fusion. Therefore, similarly to SwinIR [63], we first concatenate the inputs from different modalities and pass them through a linear layer to map them to a higher dimension. We then feed the features into multiple BRWKV layers without changing the feature size. Finally, we map the features back to the data space with the SwinIR light-weight projection head. More details about this plain backbone can be found in Suppl. Sect. 5. We compared the two architectures on the Pavia and MSRS datasets. The results are reported in Tab. 6(j). It can be observed that the multi-scale architecture performs better with a comparable parameter count.

6.2 RWKV Effective Receptive Field

The effective receptive field (ERF) plays a crucial role in determining whether a model possesses a large receptive field. We adopted the RepLKNet's [100] method for visualizing ERF to illustrate the ERFs of various models, including CNNs, transformers, Mambas, and RWKVFusion.

TABLE 9: Semantic segmentation metrics, including per-class IoU, mean IoU, and mean accuracy, are reported. Segformer is trained on visible infrared images and different fusion methods on the MSRS VIF dataset. The best results are in red and the second-best results are in blue.

Methods	MSRS VIF Dataset Segmented by Segformer										
	Background	Car	Person	Bike	Curve	Car Stop	Guardrail	Color Cone	Bump	mIoU	mAcc
Visible	98.06	88.13	61.84	69.45	52.46	64.59	74.13	56.68	71.45	70.75	80.55
Infrared	97.92	86.46	71.49	66.62	48.16	55.63	42.90	50.97	65.26	65.04	74.73
DeFuse [55]	98.64	91.53	74.73	76.67	59.62	76.88	85.07	64.90	79.92	78.66	87.22
U2Fusion [7]	98.38	89.92	73.48	69.76	52.11	71.71	81.28	60.72	73.67	74.56	82.83
SwinFusion [56]	98.37	89.89	74.08	69.70	50.24	73.05	72.80	59.47	68.39	72.89	80.60
CDDFuse [35]	98.58	91.19	74.03	75.51	55.42	76.79	85.53	64.31	79.25	77.85	86.39
DDFM [12]	97.86	85.90	69.68	60.43	36.64	6393	78.11	50.88	57.27	66.74	73.29
SegMIF [9]	98.54	90.62	75.08	73.81	56.74	75.64	82.14	62.83	78.55	77.11	86.36
MGDN [69]	98.59	91.08	75.03	75.86	56.95	76.17	84.23	64.48	78.84	77.92	86.39
TC-MOA [70]	98.59	90.98	73.82	76.44	59.38	76.12	85.41	60.66	78.39	77.75	85.90
FILM [18]	**98.71**	**92.13**	**76.42**	**77.40**	**62.32**	76.79	**85.86**	64.27	**80.94**	**79.43**	**88.64**
TextIF [17]	98.70	92.04	75.24	77.25	61.99	**77.39**	85.79	**64.56**	80.47	79.27	88.30
Proposed	**98.72**	**92.25**	**76.21**	**77.49**	**62.63**	**77.68**	**85.99**	**64.61**	**80.89**	**79.61**	**88.72**

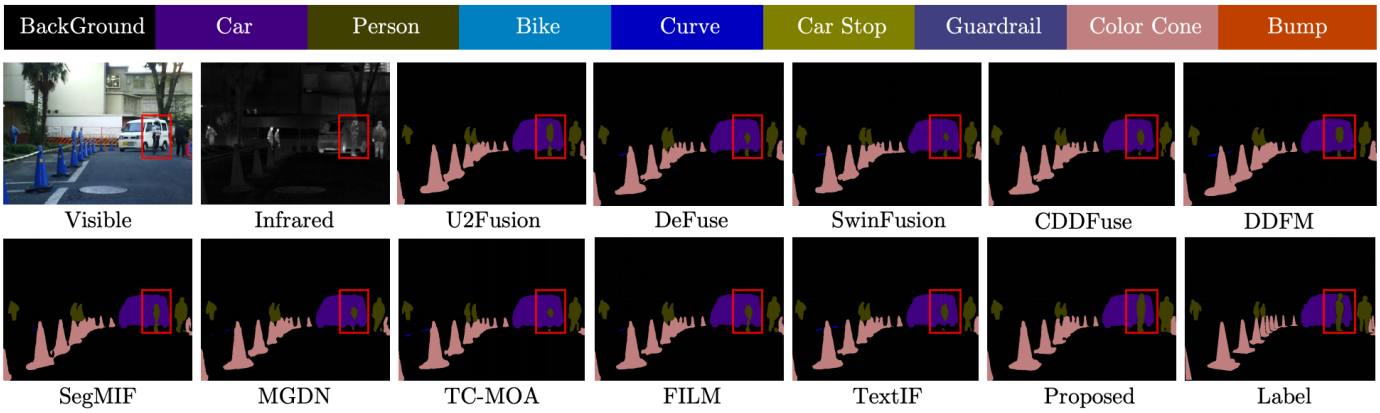


Fig. 12: Semantic segmentation results using Segformer [99] on images fused by the proposed RWKVFusion and compared with previous methods. Segformer was trained separately on the corresponding fused/GT image pairs, with all configurations held constant across experiments. Red boxes show segmented objects that cause mIoU differences.

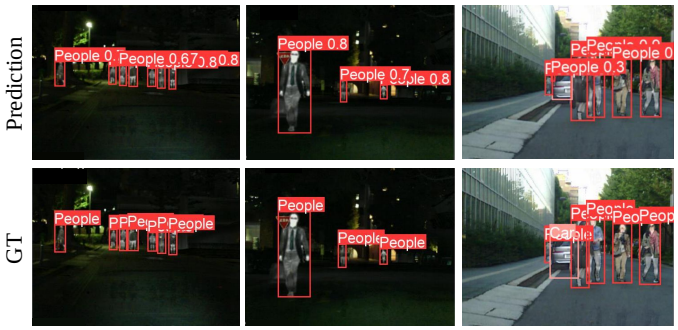


Fig. 13: Object detection by using YOLO v5 [8] on the RWKVFusion's fused product and the related GT.

The qualitative results are presented in Fig. 10. It is evident that RWKVFusion exhibits a larger and more concentrated ERF, closely resembling the global receptive field of global attention, but with a clear advantage with respect to the latter, i.e., a less-than-quadratic increase in memory consumption with respect to token length.

TABLE 10: Object detection performance by applying recent state-of-the-art fusion methods on the MSRS VIF dataset. The best results are in red and the second-best results are in blue.

Method	People↑	Car↑	mAP@0.5↑	mAP0.5:0.9↑
DeFuse [55]	0.922	0.792	0.857	0.646
U2Fusion [7]	0.941	0.650	0.875	0.654
SwinFusion [56]	0.950	0.784	0.874	0.668
CDDFuse [35]	0.935	0.750	0.843	0.615
DDFM [12]	0.909	0.760	0.834	0.579
SegMIF [9]	**0.961**	**0.840**	**0.900**	**0.703**
MGDN [69]	0.942	0.830	0.886	0.656
TC-MOA [70]	0.923	0.811	0.867	0.671
TextIF [17]	0.954	0.838	0.896	0.690
Proposed	**0.966**	**0.847**	**0.907**	**0.697**

6.3 Downstream Tasks for VIF

Image fusion often serves as pre-processing for downstream tasks. In this section, we will focus on VIF considering three kinds of downstream applications: monocular depth estimation, object detection, and segmentation.

Monocular depth estimation: Due to the lack of GT, we simply present visual results. We employed Depth Anything v2 [96] for zero-shot inference. From Fig. 11, it is

clear that the depth maps estimated by our RWKVFusion exhibit clearer contours, effective estimation of distant backgrounds, and align well with human visual perception of the depth.

Object detection: We tested our approach on the MSRS dataset using the YOLO v5 detector [8]. The detection performance for several VIF fusion methods is reported in Tab. 10. Some detection examples are depicted in Fig. 13 and Suppl. Sect. 11. It is worth to be remarked that SegMIF simultaneously optimizes both downstream and fusion tasks, and its performance overcomes that of the proposed RWKVFusion just for the mAP0.5:0.9 metric. Instead, our RWKVFusion obtains the best performance for all other metrics, clearly demonstrating its effectiveness.

Semantic segmentation: We adopt Segformer [99] with a pretrained MiT-B3 backbone, trained for 100 epochs at a fixed learning rate of $1e^{-4}$. As shown in Tab. 9, our method achieves the best mIoU and the highest segmentation accuracy. Fig. 12 reveals that fused images preserve target boundary continuity, reducing the segmentation error of small-scale objects. More segmentation results can be found in Suppl. Sect. 11.

7 Conclusion

In this paper, we proposed a novel image fusion framework that simultaneously addresses the limitations of conventional fusion frameworks and existing neural architectures. To address the semantic deficiencies in the previous fusion frameworks, we integrated multi-modal guidance, such as language information and semantic masks, into the fusion process. To overcome the limitations of existing network designs, we introduced an efficient RWKV-based architecture featuring linear computational complexity, multi-scale processing capabilities, and a global receptive field, while maintaining low latency and seamlessly integrating various guidance. Extensive experiments and ablation studies on different fusion tasks (i.e., VIF, MIF, MEF, MFF, HMIF, and pansharpening) demonstrated the superior performance and versatility of the proposed framework.

8 Acknowledgement

This research is supported by NSFC (Grant No. 12271083) and Sichuan Province's Science and Technology Empowerment for Disaster Prevention, Mitigation, and Relief Project (2025YFNH0001). We are deeply grateful to Jocelyn Chanussot for his assistance during the first round of review. Additionally, we thank Yingying Wang at Xiamen University for providing the medical Harvard dataset.

References

[1] L.-J. Deng, G. Vivone, M. E. Paoletti, G. Scarpa, J. He, Y. Zhang, J. Chanussot, and A. Plaza, "Machine learning in pansharpening: A benchmark, from shallow to deep networks," *IEEE Geosci. Remote Sens. Mag.*, vol. 10, no. 3, pp. 279–315, 2022.

[2] G. Vivone, "Multispectral and hyperspectral image fusion in remote sensing: A survey," *Inf. Fus.*, vol. 89, pp. 405–417, 2023.

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

[4] T. Zhou, S. Ruan, and S. Canu, "A review: Deep learning for medical image segmentation using multi-modality fusion," *Array*, vol. 3, p. 100004, 2019.

[5] X. Zhang and Y. Demiris, "Visible and infrared image fusion using deep learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 8, pp. 10 535–10 554, 2023.

[6] R. Dian, A. Guo, and S. Li, "Zero-shot hyperspectral sharpening," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 10, pp. 12 650–12 666, 2023.

[7] H. Xu, J. Ma, J. Jiang, X. Guo, and H. Ling, "U2Fusion: A unified unsupervised image fusion network," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 1, pp. 502–518, 2020.

[8] G. Jocher, "Ultralytics yolov5," 2020. [Online]. Available: https://github.com/ultralytics/yolov5

[9] J. Liu, Z. Liu, G. Wu, L. Ma, R. Liu, W. Zhong, Z. Luo, and X. Fan, "Multi-interactive feature learning and a full-time multi-modality benchmark for image fusion and segmentation," in *ICCV*, 2023.

[10] G. Wu, H. Fu, J. Liu, L. Ma, X. Fan, and R. Liu, "Hybrid-supervised dual-search: Leveraging automatic learning for loss-free multi-exposure image fusion," in *AAAI*, vol. 38, no. 6, 2024, pp. 5985–5993.

[11] 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 Trans. Image Process.* , vol. 29, pp. 4980–4995, 2020.

[12] Z. Zhao, H. Bai, Y. Zhu, J. Zhang, S. Xu, Y. Zhang, K. Zhang, D. Meng, R. Timofte, and L. Van Gool, "DDFM: denoising diffusion model for multi-modality image fusion," in *CVPR*, 2023, pp. 8082–8093.

[13] J. Yue, L. Fang, S. Xia, Y. Deng, and J. Ma, "Dif-fusion: Towards high color fidelity in infrared and visible image fusion with diffusion models," *IEEE Trans. Image Process.*, 2023.

[14] X. Hu, J. Jiang, X. Liu, and J. Ma, "ZMFF: Zero-shot multi-focus image fusion," *Inf. Fus.*, vol. 92, pp. 127–138, 2023.

[15] L. Tang, J. Yuan, and J. Ma, "Image fusion in the loop of high-level vision tasks: A semantic-aware real-time infrared and visible image fusion network," *Inf. Fus.*, vol. 82, pp. 28–42, 2022.

[16] L. Tang, J. Yuan, H. Zhang, X. Jiang, and J. Ma, "PIAFusion: A progressive infrared and visible image fusion network based on illumination aware," *Inf. Fus.*, vol. 83, pp. 79–92, 2022.

[17] 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 *CVPR*, 2024, pp. 27 026–27 035.

[18] 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.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[20] D. P. Kingma, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[21] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *ICLR*, 2021.

[22] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *CVPR*, 2022, pp. 10 684–10 695.

[23] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *ICCV*, 2021, pp. 10 012–10 022.

[24] J. Ho, N. Kalchbrenner, D. Weissenborn, and T. Salimans, "Axial attention in multidimensional transformers," *arXiv preprint arXiv:1912.12180*, 2019.

[25] A. Vaswani, P. Ramachandran, A. Srinivas, N. Parmar, B. Hechtman, and J. Shlens, "Scaling local self-attention for parameter efficient visual backbones," in *CVPR*, 2021, pp. 12 894–12 904.

[26] Y. Liu, Y. Tian, Y. Zhao, H. Yu, L. Xie, Y. Wang, Q. Ye, J. Jiao, and Y. Liu, "VMamba: Visual state space model," in *NeurIPS*, 2024.

[27] Z. Cao, X. Wu, L.-J. Deng, and Y. Zhong, "A novel state space model with local enhancement and state sharing for image fusion," in *ACM MM*, 2024.

[28] W. Yu and X. Wang, "Mambaout: Do we really need mamba for vision?" *arXiv preprint arXiv:2405.07992*, 2024.

[29] B. Peng, E. Alcaide, Q. Anthony, A. Albalak, S. Arcadinho, S. Biderman, H. Cao, X. Cheng, M. Chung, M. Grella *et al.*,

"Rwkv: Reinventing rnns for the transformer era," *arXiv preprint arXiv:2305.13048*, 2023.

[30] S. Zhai, W. Talbott, N. Srivastava, C. Huang, H. Goh, R. Zhang, and J. Susskind, "An attention free transformer," *arXiv preprint arXiv:2105.14103*, 2021.

[31] Y. Duan, W. Wang, Z. Chen, X. Zhu, L. Lu, T. Lu, Y. Qiao, H. Li, J. Dai, and W. Wang, "Vision-rwkv: Efficient and scalable visual perception with rwkv-like architectures," *arXiv preprint arXiv:2403.02308*, 2024.

[32] B. Xiao, H. Wu, W. Xu, X. Dai, H. Hu, Y. Lu, M. Zeng, C. Liu, and L. Yuan, "Florence-2: Advancing a unified representation for a variety of vision tasks," in *CVPR*, 2024, pp. 4818–4829.

[33] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," in *ICCV*, 2023, pp. 4015–4026.

[34] X. Zhang, "Deep learning-based multi-focus image fusion: A survey and a comparative study," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 9, pp. 4819–4838, 2021.

[35] Z. Zhao, H. Bai, J. Zhang, Y. Zhang, S. Xu, Z. Lin, R. Timofte, and L. Van Gool, "CDDFuse: Correlation-driven dual-branch feature decomposition for multi-modality image fusion," in *CVPR*, 2023, pp. 5906–5916.

[36] H. Zhang, X. Zuo, J. Jiang, C. Guo, and J. Ma, "MRFS: Mutually reinforcing image fusion and segmentation," in *CVPR*, 2024, pp. 26 974–26 983.

[37] H. Xu, P. Liang, W. Yu, J. Jiang, and J. Ma, "Learning a generative model for fusing infrared and visible images via conditional generative adversarial network with dual discriminators," in *IJCAI*, 2019, pp. 3954–3960.

[38] Z. Cao, S. Cao, L.-J. Deng, X. Wu, J. Hou, and G. Vivone, "Diffusion model with disentangled modulations for sharpening multispectral and hyperspectral images," *Inf. Fus.*, vol. 104, p. 102158, 2024.

[39] Q. Xu, Y. Li, J. Nie, Q. Liu, and M. Guo, "Upangan: Unsupervised pansharpening based on the spectral and spatial loss constrained generative adversarial network," *Inf. Fus.*, vol. 91, pp. 31–46, 2023.

[40] L. Tang, Y. Deng, X. Yi, Q. Yan, Y. Yuan, and J. Ma, "DRMF: Degradation-robust multi-modal image fusion via composable diffusion prior," in *ACM MM*, 2024, pp. 8546–8555.

[41] OpenAI, "Chatgpt (gpt-4)," 2023. [Online]. Available: https://openai.com/chatgpt

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

[43] X. Wu, T.-Z. Huang, L.-J. Deng, and T.-J. Zhang, "Dynamic cross feature fusion for remote sensing pansharpening," in *ICCV*, 2021, pp. 14 687–14 696.

[44] Y.-J. Liang, Z. Cao, L.-J. Deng, and X. Wu, "Fourier-enhanced implicit neural fusion network for multispectral and hyperspectral image fusion," in *NeurIPS*, 2024.

[45] Y. Duan, X. Wu, H. Deng, and L.-J. Deng, "Content-adaptive non-local convolution for remote sensing pansharpening," in *CVPR*, 2024.

[46] M. Zhou, J. Huang, Y. Fang, X. Fu, and A. Liu, "Pan-sharpening with customized transformer and invertible neural network," in *AAAI*, 2022.

[47] G. Yang, X. Cao, W. Xiao, M. Zhou, A. Liu, X. Chen, and D. Meng, "Panflownet: A flow-based deep network for pan-sharpening," in *ICCV*, 2023, pp. 16 857–16 867.

[48] Y. Sun, H. Xu, Y. Ma, M. Wu, X. Mei, J. Huang, and J. Ma, "Dual spatial-spectral pyramid network with transformer for hyperspectral image fusion," *IEEE Trans. Geosci. Remote Sens.*, 2023.

[49] 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 *CVPR*, 2024, pp. 25 912–25 921.

[50] D. Han, X. Pan, Y. Han, S. Song, and G. Huang, "Flatten transformer: Vision transformer using focused linear attention," in *CVPR*, 2023, pp. 5961–5971.

[51] X. He, K. Cao, K. Yan, R. Li, C. Xie, J. Zhang, and M. Zhou, "Pan-mamba: Effective pan-sharpening with state space model," *arXiv preprint arXiv:2402.12192*, 2024.

[52] W. G. C. Bandara and V. M. Patel, "Hypertransformer: A textural and spectral feature fusion transformer for pansharpening," in *CVPR*, 2022, pp. 1767–1777.

[53] J. Hou, Z. Cao, N. Zheng, X. Li, X. Chen, X. Liu, X. Cong, M. Zhou, and D. Hong, "Linearly-evolved transformer for pan-sharpening," *arXiv preprint arXiv:2404.12804*, 2024.

[54] J. Liu, X. Fan, Z. Huang, G. Wu, R. Liu, W. Zhong, and Z. Luo, "Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection," in *CVPR*, 2022, pp. 5802–5811.

[55] P. Liang, J. Jiang, X. Liu, and J. Ma, "Fusion from decomposition: A self-supervised decomposition approach for image fusion," in *ECCV*. Springer, 2022, pp. 719–735.

[56] 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 J. Automatica Sin.*, vol. 9, no. 7, pp. 1200–1217, 2022.

[57] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[58] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," in *NIPS 2014 Workshop on Deep Learning*, 2014.

[59] Z. Yang, H. Zhang, D. Zhao, B. Wei, and Y. Xu, "Restore-rwkv: Efficient and effective medical image restoration with rwkv," *arXiv preprint arXiv:2407.11087*, 2024.

[60] Z. Fei, M. Fan, C. Yu, D. Li, and J. Huang, "Diffusion-rwkv: Scaling rwkv-like architectures for diffusion models," *arXiv preprint arXiv:2404.04478*, 2024.

[61] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *J. Mach. Learn. Res.*, vol. 21, no. 140, pp. 1–67, 2020.

[62] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, C. Li, J. Yang, H. Su, J. Zhu *et al.*, "Grounding dino: Marrying dino with grounded pre-training for open-set object detection," *arXiv preprint arXiv:2303.05499*, 2023.

[63] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "Swinir: Image restoration using swin transformer," in *ICCV*, 2021, pp. 1833–1844.

[64] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *CVPR*, 2021, pp. 10 012–10 022.

[65] B. Zhang and R. Sennrich, "Root mean square layer normalization," in *NeurIPS*, vol. 32, 2019.

[66] D. So, W. Mańke, H. Liu, Z. Dai, N. Shazeer, and Q. V. Le, "Searching for efficient transformers for language modeling," *NeurIPS*, vol. 34, pp. 6010–6022, 2021.

[67] D. Hendrycks and K. Gimpel, "Gaussian error linear units (gelus)," *arXiv preprint arXiv:1606.08415*, 2016.

[68] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, 2004.

[69] Y. Guan, R. Xu, M. Yao, L. Wang, and Z. Xiong, "Mutual-guided dynamic network for image fusion," in *ACM MM*, 2023, pp. 1779–1788.

[70] P. Zhu, Y. Sun, B. Cao, and Q. Hu, "Task-customized mixture of adapters for general image fusion," in *CVPR*, 2024, pp. 7099–7108.

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

[72] J. Liu, G. Wu, J. Luan, Z. Jiang, R. Liu, and X. Fan, "Holoco: Holistic and local contrastive learning network for multi-exposure image fusion," *Inf. Fus.*, vol. 95, pp. 237–249, 2023.

[73] W. Wang, L.-J. Deng, and G. Vivone, "A general image fusion framework using multi-task semi-supervised learning," *Inf. Fus.*, vol. 108, p. 102414, 2024.

[74] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *CVPR*, 2018.

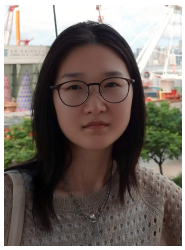
[75] S. Li, R. Dian, L. Fang, and J. M. Bioucas-Dias, "Fusing hyperspectral and multispectral images via coupled sparse tensor factorization," *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 4118–4130, 2018.

[76] R. Dian and S. Li, "Hyperspectral image super-resolution via subspace-based low tensor multi-rank regularization," *IEEE Trans. Image Process.*, vol. 28, no. 10, pp. 5135–5146, 2019.

[77] T. Xu, T. Huang, L. Deng, and N. Yokoya, "An iterative regularization method based on tensor subspace representation for

hyperspectral image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–16, 2022.
[78] J. Hu, T. Huang, L. Deng, T. Jiang, G. Vivone, and J. Chanussot, "Hyperspectral image super-resolution via deep spatiospectral attention convolutional neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, 2022.
[79] W. Dong, C. Zhou, F. Wu, J. Wu, G. Shi, and X. Li, "Model-guided deep hyperspectral image super-resolution," *IEEE Trans. Image Process.*, vol. 30, pp. 5754–5768, 2021.
[80] J. Hu, T. Huang, L. Deng, H. Dou, D. Hong, and G. Vivone, "Fusformer: A transformer-based fusion network for hyperspectral image super-resolution," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.
[81] T. Huang, W. Dong, J. Wu, L. Li, X. Li, and G. Shi, "Deep hyperspectral image fusion network with iterative spatio-spectral regularization," *IEEE Trans. Comput. Imaging*, vol. 8, pp. 201–214, 2022.
[82] S.-Q. Deng, L.-J. Deng, X. Wu, R. Ran, D. Hong, and G. Vivone, "PSRT: Pyramid shuffle-and-reshuffle transformer for multispectral and hyperspectral image fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–15, 2023.
[83] Q. Ma, J. Jiang, X. Liu, and J. Ma, "Learning a 3d-cnn and transformer prior for hyperspectral image super-resolution," *Inf. Fus.*, vol. 100, p. 101907, 2023.
[84] S. Deng, L.-J. Deng, X. Wu, R. Ran, and R. Wen, "Bidirectional dilation transformer for multispectral and hyperspectral image fusion," in *IJCAI*, 2023.
[85] J. Fang, J. Yang, A. Khader, and L. Xiao, "Mimo-sst: Multi-input multi-output spatial-spectral transformer for hyperspectral and multispectral image fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–20, 2024.
[86] G. Vivone, R. Restaino, and J. Chanussot, "Full scale regression-based injection coefficients for panchromatic sharpening," *IEEE Trans. Image Process.*, vol. 27, no. 7, pp. 3418–3431, 2018.
[87] S. Lolli, L. Alparone, A. Garzelli, and G. Vivone, "Haze correction for contrast-based multispectral pansharpening," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 12, pp. 2255–2259, 2017.
[88] Z.-C. Wu, T.-Z. Huang, L.-J. Deng, J. Huang, J. Chanussot, and G. Vivone, "LRTCFPan: Low-rank tensor completion based framework for pansharpening," *IEEE Trans. Image Process.*, vol. 32, pp. 1640–1655, 2023.
[89] L. He, Y. Rao, J. Li, J. Chanussot, A. Plaza, J. Zhu, and B. Li, "Pansharpening via detail injection based convolutional neural networks," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 12, no. 4, pp. 1188–1204, 2019.
[90] L.-J. Deng, G. Vivone, C. Jin, and J. Chanussot, "Detail injection-based deep convolutional neural networks for pansharpening," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 8, pp. 6995–7010, 2020.
[91] Z.-R. Jin, T.-J. Zhang, T.-X. Jiang, G. Vivone, and L.-J. Deng, "LAGConv: Local-context adaptive convolution kernels with global harmonic bias for pansharpening," in *AAAI*, vol. 36, no. 1, Jun. 2022, pp. 1113–1121.
[92] M. Zhou, X. Fu, J. Huang, F. Zhao, A. Liu, and R. Wang, "Effective pan-sharpening with transformer and invertible neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–15, 2022.
[93] X. Tian, K. Li, W. Zhang, Z. Wang, and J. Ma, "Interpretable model-driven deep network for hyperspectral, multispectral, and panchromatic image fusion," *IEEE Trans. Neural Netw. Learn. Syst.*, pp. 1–14, 2023.
[94] Q. Meng, W. Shi, S. Li, and L. Zhang, "Pandiff: A novel pansharpening method based on denoising diffusion probabilistic model," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–17, 2023.
[95] L.-j. Deng, G. Vivone, M. E. Paoletti, G. Scarpa, J. He, Y. Zhang, J. Chanussot, and A. Plaza, "Machine learning in pansharpening: A benchmark, from shallow to deep networks," *IEEE Geosci. Remote Sens. Mag.*, vol. 10, no. 3, pp. 279–315, 2022.
[96] L. Yang, B. Kang, Z. Huang, Z. Zhao, X. Xu, J. Feng, and H. Zhao, "Depth anything v2," *arXiv:2406.09414*, 2024.
[97] T. Dao, "FlashAttention-2: Faster attention with better parallelism and work partitioning," in *ICLR*, 2024.
[98] Black Forest Labs, "Flux: A powerful tool for text generation," https://huggingface.co/black-forest-labs/FLUX.1-dev, 2024, accessed: 2024-09-26.
[99] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," *Advances in neural information processing systems*, vol. 34, pp. 12 077–12 090, 2021.
[100] X. Ding, X. Zhang, J. Han, and G. Ding, "Scaling up your kernels to 31x31: Revisiting large kernel design in cnns," in *CVPR*, 2022, pp. 11 963–11 975.

Zi-Han Cao was born in Zhongxiang, Hubei province, China. He received his B.S. degree from the School of Information and Communication Engineering, University of Electronic Science and Technology of China (UESTC), Chengdu, China, in 2023. He is currently pursuing his M.S. degree under the supervision of Prof. Liang-Jian Deng in the School of Mathematics at UESTC. His research interests include computer vision, machine learning, and applications in low-level vision tasks, such as image fusion, inverse problems, and radar signal processing.



Yu-Jie Liang is currently a third-year master's student. She received her B.S. degree in Information and Computational Science from the School of Science, Yanshan University, Qinhuangdao, China, in 2022. She is now pursuing her M.S. degree under the supervision of Prof. Liang-Jian Deng at the School of Mathematical Sciences, University of Electronic Science and Technology of China (UESTC) in Chengdu, China. Her research interests focus on computer vision and image processing, including image fusion and image super-resolution.

Liang-Jian Deng (Senior Member, IEEE) received the B.S. and Ph.D. degrees in applied mathematics from the School of Mathematical Sciences, University of Electronic Science and Technology of China (UESTC), Chengdu, China, in 2010 and 2016, respectively. He is currently a Research Fellow with the School of Mathematical Sciences, UESTC. From 2013 to 2014, he was a Joint-Training Ph.D. student with the Case Western Reserve University, Cleveland, OH, USA. In 2017, he was a Postdoc at Hong Kong Baptist University (HKBU). In addition, he also stayed at Isaac Newton Institute for Mathematical Sciences, Cambridge University and HKBU for short visits. His research interests include the use of partial differential equations (PDE), optimization modeling, and deep learning to address several tasks in image processing, and computer vision, e.g., resolution enhancement and restoration.

Gemine Vivone (Senior Member, IEEE) received the B.Sc. and M.Sc. degrees (summa cum laude), and the Ph.D. degree in information engineering from the University of Salerno, Fisciano, Italy, in 2008, 2011, and 2014, respectively. He is a senior researcher at the National Research Council (Italy). His main research interests focus on image fusion, statistical signal processing, deep learning, and classification and tracking of remotely sensed images. Dr. Vivone is an ex-officio member of the IEEE Geoscience and Remote Sensing Society (GRSS) Administrative Committee, a Co-chair of the IEEE GRSS Image Analysis and Data Fusion Technical Committee, a member of the IEEE Task Force on "Deep Vision in Space". Dr. Vivone is currently the Editor in Chief for IEEE Geoscience and Remote Sensing eNewsletter, an Area Editor for Elsevier Information Fusion, and Associate Editor for IEEE Transactions on Geoscience and Remote Sensing (TGRS), IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing (JSTARS), and IEEE Geoscience and Remote Sensing Letters (GRSL). Moreover, he is an Advisory Board Member for ISPRS Journal of Photogrammetry and Remote Sensing, and an Editorial Board Member for MDPI Remote Sensing. Dr. Vivone received the IEEE GRSS Early Career Award in 2021, the Symposium Best Paper Award at IEEE International Geoscience and Remote Sensing Symposium (IGARSS) in 2015 and the Best Reviewer Award of the IEEE Transactions on Geoscience and Remote Sensing in 2017. Moreover, he is listed in the World's Top 2% Scientists by Stanford University.