Supplementary Material: 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—Detailed information regarding the proposed approach is included in this supplementary material. We first cover the datasets utilized, followed by explanations of caption generation, mask creation, and network configurations specific to each dataset. The proposed mask merging algorithm, developed to mitigate issues with suboptimal mask generation, is then discussed in detail. For fair comparison, we provide implementation specifics for the compared methods and the plain RWKV backbone. Differences between global attention and traditional attention are subsequently discussed. Additional experimental visualization results conclude the material.

1 Datasets

We provide some basic information about various image fusion datasets used in our experimental analysis:

***i*) Visible-infrared image fusion (VIF):** We used MSRS[1] [1] and M3FD[2] [2] training sets for training, and MSRS, M3FD, and TNO[3] test sets for testing. The MSRS training set involves 1083 visible/infrared image pairs for model training, the MSRS and TNO test set has 361 and 41 pairs, respectively. The M3FD training set has 3900 visible/infrared image pairs for training, and the test set consists of 300 pairs. Since the M3FD dataset does not provide mask labels, we generated and merged both modality masks using the proposed semantic branch. When training, the image pair is randomly cropped and resized into patches of size of 224 × 280. Moreover, 80 MSRS detection image pairs are used to validate the downstream detection and monocular depth estimation tasks.

***ii*) Medical image fusion (MIF):** We chose the medical Harvard[4] dataset to conduct the MRI-SPECT image fusion task. We did not crop images even for training. 285 pairs are used for training and 71 pairs are used for testing.

***iii*) Multi-exposure image fusion (MEF):** For the MEF task, we selected the SICE[5] [3] and the MEFB[6] datasets [4], using 288 images from the SICE dataset and 60 from the MEFB dataset for training. Instead, 72 and 40 images are used for the SICE and the MEFB datasets, respectively, for testing. We did not use normally exposed images as ground-truth (GT), i.e., we considered an unsupervised training. During training, each image was randomly cropped and resized into patches of size of 256 × 256, with a crop scale range of [0.8, 1] and an aspect ratio range of [0.8, 1.4]. Following the approach in SwinFusion [5], the model fuses only the Y channel of color images. The fused Y channel is then concatenated with the weighed Cb and Cr channels of overexposed and underexposed images, before being mapped back to the RGB space.

***iv*) Multi-focus image fusion (MFF):** For the MFF task, the RealMFF[7] [6] and the MFI-WHU[8] [7] datasets have been used for both training and testing. More specifically, 639 images from the RealMFF dataset have been used for training and 71 for testing, while 92 images from the MFI-WHU dataset have been used for training and 30 for testing. The remaining processing steps are the same as in the MEF task.

***v*) Pansharpening:** The Pancollection[9] dataset for pansharpening is derived from three satellites: WorldView-3 (8 bands), GaoFen-2 (4 bands), and QuickBird (4 bands). The WorldView-3 (WV3) satellite captures panchromatic (PAN) and low resolution multispectral images (LRMS) at 0.31 m and 1.24 m spatial resolution, respectively. The WV3 dataset contains 9714 samples for training and 1080 samples for validation. Each sample consists of PAN/LRMS/GT image pairs with size of $64 \times 64 \times 1$, $16 \times 16 \times 8$, and $64 \times 64 \times 8$. The GaoFen-2 (GF2) satellite captures PAN and LRMS images at 0.8 m and 3.2 m spatial resolution, respectively. The GF2 dataset consists of 19809 training samples and 2201 validation samples. Each sample includes PAN/LRMS/GT image pairs with size of $64 \times 64 \times 1$, $16 \times 16 \times 4$, and $64 \times 64 \times 4$. The QuickBird (QB) satellite captures PAN and LRMS images at 0.7 m and 2.8 m spatial resolution, respectively. The QB dataset contains 20685 training samples and 48 test samples, with each sample consisting of PAN/LRMS/GT image pairs of the size of $64 \times 64 \times 1$, $16 \times 16 \times 4$, and $64 \times 64 \times 4$. Additionally, we followed Wald's protocol [8] to simulate reduced resolution datasets from the original (full resolution) ones. Experiments for the three test cases (i.e.,

[1] https://github.com/Linfeng-Tang/MSRS
[2] https://github.com/JinyuanLiu-CV/TarDAL
[3] https://figshare.com/articles/dataset/TNOImage_Fusion_Dataset/1008029
[4] https://www.med.harvard.edu/AANLIB/home.html
[5] https://github.com/csjcai/SICE
[6] https://github.com/xingchenzhang/MEFB
[7] https://github.com/Zancelot/Real-MFF
[8] https://github.com/HaoZhang1018/MFI-WHU
[9] https://liangjiandeng.github.io/PanCollection.html.

Algorithm 1: `maskMerging` function

```
Input: $\mathbf{M}_1, \mathbf{c}_1$: masks and classes from modality 1,
       $\mathbf{M}_2, \mathbf{c}_2$: masks and classes from modality 2, $\tilde{d}$:
       IoU threshold
Output: $\mathbf{M}_{merge}$, $\mathbf{c}_{merge}$: merged masks and classes
$\mathbf{M}_{merge}, \mathbf{c}_{merge} \leftarrow$ emptyList, emptyList;
for $M_1, c_1 \leftarrow \mathbf{M}_1, \mathbf{c}_1$ do
    $\mathbf{d} \leftarrow$ IoU$(M_1, \mathbf{M}_2)$;          ▷ IoU with all masks
    if $length(\mathbf{d} \leq \tilde{d}) = 0$ then
        ▷ Append in the list
        $\mathbf{M}_{merge} \leftarrow M_1$, $\mathbf{c}_{merge} \leftarrow c_1$;
        continue;
    end
    for $i, d \leftarrow enumerate(\mathbf{d})$ do
        $M_2, c_2 \leftarrow \mathbf{M}_2^{(i)}, \mathbf{c}_2^{(i)}$;          ▷ $i$-th mask and class
        if $d \leq \tilde{d}$ then
            $\mathbf{c}_{merge} \leftarrow c_1$;          ▷ Append in the list
            if $c_1 = c_2$ then
                $\mathbf{M}_{merge} \leftarrow M_1 \cup M_2$;
            else
                $\mathbf{M}_{merge} \leftarrow M_1$;
            end
        end
    end
end
return $\mathbf{M}_{merge}$, $\mathbf{c}_{merge}$
```

Algorithm 2: Masks using SAM and merging

```
Input: Seg$(\cdot)$: SAM model, DINO$(\cdot)$: DINO model,
       $\mathbf{S}_1$, $\mathbf{S}_2$: input modalities, $\tilde{d}$: IoU threshold,
       and $\mathbf{C}$: grounding prompt
Output: $\mathbf{M}_{merge,2}, \mathbf{c}_{merge,2}$: merged masks and
        classes
Seg$(\cdot)$, DINO$(\cdot) \leftarrow$ load pre-trained weights();
▷ Predict bounding boxes $\mathbf{B}$ and box class $\mathbf{c}$.
$\mathbf{B}_1,\ \mathbf{c}_1 \leftarrow$DINO$(\mathbf{S}_1, \mathbf{C})$;
$\mathbf{B}_2,\ \mathbf{c}_2 \leftarrow$DINO$(\mathbf{S}_2, \mathbf{C})$;
▷ Segment the first and second modality masks in
  the boxes.
$\mathbf{M}_1 \leftarrow$Seg$(\mathbf{B}_1)$, $\mathbf{M}_2 \leftarrow$Seg$(\mathbf{B}_2)$;
▷ Merge the first and second modality masks using
  Algo. 1.
$\mathbf{M}_{merge,1}, \mathbf{c}_{merge,1} \leftarrow$
  maskMerging$(\mathbf{M}_1, \mathbf{M}_2, \mathbf{c}_1, \mathbf{c}_2, \tilde{d})$;
$\mathbf{M}_{merge,2}, \mathbf{c}_{merge,2} \leftarrow$
  maskMerging$(\mathbf{M}_2, \mathbf{M}_1, \mathbf{c}_2, \mathbf{c}_1, \tilde{d})$;
▷ Remove duplicated masks
for $i \leftarrow range(length(\mathbf{M}_{merge,1}))$ do
    for $j \leftarrow range(i+1, length(\mathbf{M}_{merge,2}))$ do
        if IoU$(\mathbf{M}^{(i)}_{merge,1}, \mathbf{M}^{(j)}_{merge,2}) \leq \tilde{d}$ then
            ▷ Remove the $j$-th mask and class
            $pop(\mathbf{M}_{merge,2}, j)$, $pop(\mathbf{c}_{merge,2}, j)$;
        end
    end
end
return $\mathbf{M}_{merge,2}, \mathbf{c}_{merge,2}$
```

	VIF	MIF	MEF	MFF	HMIF	Pansharpening
Basic chan.	32					
Chan. upscale	(1, 1)	(1, 1)	(1, 1)	(1, 1)	(2, 1, 1)	(2, 1, 1)
Enc. layers	(1, 1)	(1, 1)	(2, 1)	(1, 1)	(2, 1, 1)	(2, 1, 1)
Mid. layers	1	1	1	1	1	1
Dec. layers	(1, 1)	(1, 1)	(2, 1)	(1, 1)	(2, 1, 1)	(2, 1, 1)
FFN hidden	(2, 2)	(2, 2)	(2, 2)	(2, 2)	(2, 2, 2)	(2, 2, 2)
Drop path	0.2	0.3	0.1	0.1	0.2	0.2
λ in Eq. (19)	—	—	—	—	0.1	
(η_1, η_2, η_3) in Eq. (20)	(10, 2, 20)				—	—
Batch size	10	12	12	12	64	64
Optimizer	AdamW(LR= $1e^{-3}$, $\beta_1 = 0.95$, $\beta_2 = 0.99$)					
LR scheduler	Cosine Annealing(LR: $1e^{-3} \rightarrow 1e^{-5}$)					
Weight decay	1×10^{-6}					
Epochs	200	50	50	50	800	800

TABLE 1: Model configurations for the different tasks. Chan., Enc., Mid., Dec., and LR stand for channel, encoder, middle, decoder, and learning rate, respectively.

WV3, GF2, and QB) are conducted at both reduced and full resolution.

***vi*) Hyperspectral-multispectral image fusion (HMIF):** We evaluated the performance of our RWKVFusion on two remote sensing datasets: the Chikusei [9] and the Pavia [10] datasets. The Chikusei dataset consists of 2517×2335 pixels with 128 spectral bands spanning from 363 nm to 1018 nm. We selected the upper-left region of size of 1000×2200 pixels to train the network, extracting 64×64 overlapping patches from this region as ground-truth. The high resolution multispectral image patch size is $64 \times 64 \times 3$ and the low resolution hyperspectral image patch size is $16 \times 16 \times 128$. The Pavia dataset consists of 102 spectral bands (reduced from the original 115 bands by removing water absorption and noisy bands) and has a spatial size of 1096×1096 pixels. We selected a region from the top of the captured area as training set, cropping patches of size of 64×64, with the remaining area used as test set. The test set consists of two non-overlapping hyperspectral patches of size of 400×400 pixels.

2 Implementation Details

We first present the building of training and test datasets followed by the architectural designs and training configurations for the different tasks and datasets.

For all datasets, we generated captions with a maximum sentence length of 512, encoded them within the pre-trained T5 model [11]. Masks have been generated for all explored tasks except for HMIF and pansharpening, where they have not been generated because of the small patch size (i.e., 64×64). More specifically, for the VIF task, we employed an additional mask merging process to combine masks from visible and infrared images, while for the other tasks, open-vocabulary segmentation has been performed using Florence. These procedures enabled us to build image/caption/mask pairs for model training and testing.

The configurations of our RWKVFusion for the different image fusion tasks are reported in Tab. 1. In the following, we will provide some additional details about RWKVFusion. In particular, the proposed network is a U-net like architecture, which has encoder, middle layers, and decoder. Encoder and decoder have multiple stages, and each stage has several layers. For instance, in the VIF configuration, the encoder layers (Enc. Layers in Tab. 1) are set to (1, 1), which means the encoder has 2 stages and the first stage has 1 BRWKV layer. Channel upscale means that the number of hidden channels will be multiplied by a factor after a stage. Drop path indicates the use of the drop path technique to prevent overfitting. Cosine annealing [12] is used to change the learning rate from the initial value, $1e^{-3}$, to the minimum, $1e^{-5}$.

We implemented the proposed RWKVFusion using Pytorch on a workstation with an Intel 13700k CPU and two NVIDIA RTX 3090 GPUs. All training procedures can be performed within three 3090 GPU days. We implemented the ESS proposed in Sect. 3.4 of the main paper by Triton [13] and compiled it as CUDA kernel for fast training.

3 Mask Merging Algorithm

In Sect. 3.7 of the main paper, we introduced a mask generation pipeline to automatically segment objects for image pairs. In cases where there are significant differences between modalities, such as visible/infrared images, we found that the generated masks often result in missing objects and incomplete segmentation. Therefore, we proposed a mask merging algorithm to deal with these unsatisfactory masks. As shown in Algos. 1 and 2, the core of this algorithm checks the object class and the intersection of union (IoU). If two objects have the same class and the IoU is less than a threshold, then we merge the two masks.

4 Benchmark Details

This section is devoted to an introduction to the quality metrics used, with some implementation details for the compared methods. We provided many widely used quality metrics to assess performance for the VIF, MIF, MEF, and MFF tasks: i) information theory-based metrics: mutual information (MI), visual information fidelity (VIF), and spatial frequency (SF); ii) human perception inspired metrics: Q_{cb}, Q_y, Q_{cv}, and Q_{abf}; iii) deep model perceptual metrics: LPIPS [14]. We refer readers to previous image fusion literature [4], [15] to gain a comprehensive understanding of the metrics employed. For the MI, VIF, SF, Q_{cb}, Q_y, and Q_{abf} metrics, a higher value means better fusion performance. Conversely, for the Q_{cv} and LPIPS metrics, a lower value indicates better performance.

As suggested in [16], [17], for the pansharpening task, we considered the spectral angle mapper (SAM) [18], the erreur relative globale adimensionnelle de synthèse (ERGAS) [19], universal image quality index for multiband images (Q2n) [20], and the spatial correlation coefficient (SCC) [21] as quality metrics at reduced resolution. SAM is primarily used to measure spectral similarity in hyperspectral images. ERGAS is used to assess the global error of spectral images. It takes into account both spectral and spatial information errors and is commonly used for the quality evaluation of multispectral and hyperspectral images. Q2n is an extension of the universal image quality index (UQI) designed specifically for multispectral images. It evaluates the quality of a fused image by considering both spatial and spectral distortions. SCC measures the spatial correlation between the reference and the fused images. Since there is no reference at full resolution, we used D_λ, D_s, and the (overall) hybrid quality with no reference (HQNR) index for quality assessment at full resolution. D_λ and D_s are the spectral and spatial distortion indexes, respectively. HQNR combines D_λ and D_s to simultaneously represent both spectral and spatial quality.

As for the HMIF task, we employ the SAM, the ERGAS, the peak signal-to-noise ratio (PSNR) [22], and the Structural similarity index measure (SSIM) [21] as quality assessment metrics. The higher the values of PSNR and SSIM, the better the fusion results.

For most of the methods, we employed pre-trained weights provided in their codebases. If pre-trained weights were unavailable, we retrained models using the default configurations specified in their original code implementations/papers. In particular, we have:

1) For the VIF task: U2Fusion [23], MGDN [24], and TC-MOA [25] have been retrained;
2) For the MIF task: all methods have been trained from scratch;
3) For the MEF task: U2Fusion and TC-MOA have been retrained;
4) For the MFF task: TC-MOA has been trained using the original paper configuration;
5) For the HMIF and pansharpening tasks: all models have been trained from scratch.

We did not compare our method with FILM on the WHI-WFU and TNO datasets, as FILM neither released the descriptions generated by ChatGPT [26] nor the feature maps extracted by BLIP2 [27], making a direct comparison infeasible.

5 Implementation Details of Plain RWKV Backbone

In Sect. 6.1 of the main paper, we introduced a straightforward RWKV architecture similar to SwinIR [28]. This architecture is depicted in Fig. 1. The network consists of: i) shallow feature extraction; ii) several RWKV layers; and iii) high-quality image reconstruction. The shallow feature extraction employs a 3 × 3 convolutional layer to handle early visual processing. For the RWKV layers, we set the number of layers to 4 for the MSRS dataset in the ablation study, and to 8 for the Pavia dataset, to ensure the parameter count matches each default setting. For high-quality image reconstruction, we directly applied sub-pixel convolution [29] to map the fused features to pixel space. We empirically omit the long skip-connection between shallow feature extraction and high-quality image reconstruction in SwinIR since the tasks are different.

6 Differences on Attention Masks

Standard cross-attention mechanisms indeed produce sequence-length-squared ($L \times L$) attention weight maps,

$$\mathbf{A} = Softmax\left(\frac{\mathbf{Q}\cdot\mathbf{K}^\top}{\sqrt{d}}\right) \in \mathbb{R}^{L\times L}, \quad (1)$$

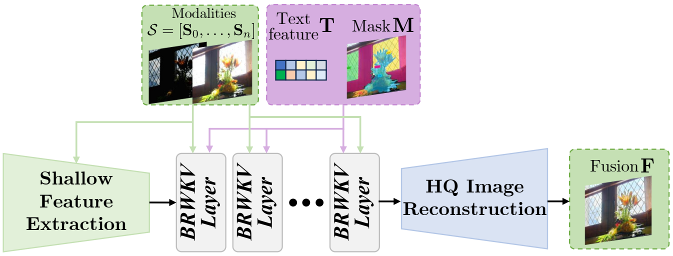

Modalities
$\mathcal{S} = [\mathbf{S}_0, \ldots, \mathbf{S}_n]$
Text feature $\mathbf{T}$
Mask $\mathbf{M}$
Shallow Feature Extraction
BRWKV Layer
BRWKV Layer
BRWKV Layer
HQ Image Reconstruction
Fusion $\mathbf{F}$

Fig. 1: An overview of the compared plain RWKV architecture. Guidance including language and masks, is injected in every RWKV layer.

that can be visualized as fine-grained masks correlating queries and keys (*e.g.*, text and image regions), the WKV (Weighted Key Value) operator used within the spatial mixing blocks of our RWKV backbone operates differently. As detailed in the WKV operator,

$$\mathbf{A}_t = \mathcal{O}_{WKV}(\mathbf{K}_s, \mathbf{V}_s)_t = \frac{\sum_{i=1, i\neq t}^{L} e^{-(|t-i|-1)/L \cdot \mathbf{w} + \mathbf{k}_i} \mathbf{v}_i + e^{\mathbf{u}+\mathbf{k}_t} \mathbf{v}_t}{\sum_{i=1, i\neq t}^{L} e^{-(|t-i|-1)/L \cdot \mathbf{w} + \mathbf{k}_i} + e^{\mathbf{u}+\mathbf{k}_t}}, \quad (2)$$

WKV relies on channel-wise learnable decay parameters ($\mathbf{w}$) and relative positional encoding, calculated efficiently, often via a recurrent formulation. This mechanism models long-range spatial dependencies with linear complexity but does not inherently compute or store an explicit matrix of pairwise token attention scores comparable to those in standard Transformers. Therefore, it does not directly yield the same kind of fine-grained attention masks.

However, our RWKVFusion framework is designed to incorporate fine-grained spatial and semantic guidance **explicitly** through the Multi-modal Fusion Module (MFM), as detailed in Sect. 3.5 and illustrated in Fig. 5 (c). The MFM directly integrates user-provided or automatically generated object masks ($\mathbf{M}$) for precise spatial guidance and encoded image captions ($\mathbf{T}$) for global semantic context into each encoder layer. As visualized in Fig. 6 (of the main text), this explicit guidance effectively modulates the network's features (*e.g.*, $\mathbf{X}_{mask}$ highlighting masked objects), achieving the desired semantic control over the fusion process.

Our design choice prioritizes the efficiency benefits of the RWKV architecture (linear FLOPs and space complexity *w.r.t.* sequence length, see Tab. 1) while ensuring rich semantic and spatial guidance through the dedicated MFM, rather than relying on implicit attention maps derived from the backbone's internal workings.

7 Additional Experimental Results

In this section, we provide additional experimental results for the Pansharpening, VIF, and MFF tasks, which have been conducted on datasets mentioned in the main paper, i.e., the GF2, QB, RoadScene, and Lytro datasets.

The Lytro[10] [30] dataset has 20 images, just used for testing MFF approaches. Instead, the RoadScene[11] [31] dataset has 22 pairs of visible and infrared images, just used for testing VIF methods. Details of the GF2 and QB datasets are provided in Sect. 1.

TABLE 2: Performance of recent state-of-the-art fusion methods on the Lytro MFF and RoadScene VIF datasets. The best results are in red, the second-best results are in blue.

Methods	Lytro MFF Dataset							
	MI↑	VIF↑	SF↑	Q_{cb}↑	Q_{abf}↑	Q_y↑	Q_{cv}↓	LPIPS↓
U2Fusion [23]	4.24	**1.05**	12.71	0.60	0.58	0.92	190.2	**0.366**
DeFuse [32]	4.23	1.03	11.42	0.58	0.53	0.91	228.1	0.375
DDFM [33]	4.07	0.98	13.41	0.60	0.60	0.85	159.0	0.398
ZMFF [34]	**4.35**	0.98	18.88	**0.71**	**0.68**	**0.97**	**52.5**	0.381
IF-MT-SSL [35]	2.19	0.18	**19.06**	0.42	0.18	0.53	292.3	0.650
TC-MOA [25]	4.07	0.96	13.33	0.64	0.59	0.93	112.8	0.383
Proposed	**4.70**	**1.09**	**19.51**	**0.72**	**0.73**	**0.98**	**48.3**	**0.356**
Methods	**RoadScene VIF Dataset**							
U2Fusion [23]	1.92	0.46	9.41	0.51	0.29	0.76	606.0	0.660
DeFuse [32]	**2.29**	**0.72**	9.57	**0.57**	0.43	0.86	356.1	0.637
SwinFusion [5]	2.04	0.61	14.81	**0.57**	0.61	0.91	369.9	**0.601**
CDDFuse [36]	2.13	0.57	15.05	0.47	0.52	0.90	328.7	0.619
DDFM [33]	2.13	0.58	10.34	0.52	0.39	0.84	356.2	0.700
SegMIF [37]	2.00	0.55	17.12	0.55	0.59	0.91	328.9	0.647
MGDN [24]	2.13	0.66	10.56	0.54	0.38	0.83	383.5	**0.591**
TC-MOA [25]	2.09	0.59	11.01	0.55	0.49	0.88	**295.9**	0.761
FILM [38]	1.70	0.25	15.84	0.47	0.30	0.69	873.7	0.772
TextIF [39]	2.09	0.61	**17.25**	**0.57**	**0.64**	**0.93**	303.6	0.632
Proposed	**2.58**	**0.69**	**17.48**	**0.61**	**0.68**	**0.95**	**300.0**	0.605

The results of the RoadScene and Lytro datasets are reported in Tab. 2. It can be seen that our RWKVFusion can largely outperform previous state-of-the-art methods, keeping at least the second-best results for six out of eight quality metrics. The results of the GF2 and QB datasets are reported in Tab. 3. Moreover, the related visual comparisons are provided in Sect. 9.

8 More Caption and Mask Examples

More examples of image captions and segmented masks are depicted in Fig. 2.

9 More Visual Results

In this section, we include more visual results related to the MSRS, TNO, RoadScene, WFI-WHU, Lytro, GF2, QB, Pavia, and medical Harvard datasets in Figs. 3 , 4, 5, 6, 7 and 8, respectively.

As shown in Fig. 3, it is evident that U2Fusion [23], SwinFusion [5], and DDFM [33] struggle to clearly present road landmarks due to underexposure in the visible images. While other methods can show the arrow landmarks in a clearer way, they inevitably compromise salient targets. For example, pedestrians within the red and yellow circles are blurred in DeFuse [32] and color distortion issues appear in the MGDN [24] and SegMIF [37] results. Instead, our method fully integrates complementary information from the source images, providing a more comprehensive description of scenes with insufficient lighting.

Fig. 4 presents an example from the TNO dataset, containing two targets as highlighted by the related mask. Observing the overall results, many methods exhibit clear

[10] https://mansournejati.ece.iut.ac.ir/content/lytro-multi-focus-dataset

[11] https://github.com/hanna-xu/RoadScene

TABLE 3: The averages and standard deviations of the adopted quality metrics for the pansharpening task calculated on the GF2 and QB test sets. 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↑		
GaoFen2 (GF2, 4-band)	MTF-GLP-FS [40]	1.68±0.35	1.60±0.35	0.891±0.026	0.939±0.020	0.035±0.014	0.143±0.028	0.823±0.035	—	—
	BT-H [41]	1.68±0.32	1.55±0.36	0.909±0.029	0.951±0.015	0.060±0.025	0.131±0.019	0.817±0.031	—	—
	LRTCFPan [42]	1.30±0.31	1.27±0.34	0.935±0.030	0.964±0.012	0.033±0.027	0.090±0.014	0.881±0.023	—	—
	DiCNN [43]	1.05±0.23	1.08±0.25	0.959±0.010	0.977±0.006	0.041±0.012	0.099±0.013	0.864±0.017	0.23M	0.19G
	FusionNet [44]	0.97±0.21	0.99±0.22	0.964±0.009	0.981±0.005	0.040±0.013	0.101±0.013	0.863±0.018	0.047M	0.32G
	LAGConv [45]	0.78±0.15	0.69±0.11	0.980±0.009	0.991±0.002	0.032±0.013	0.079±0.014	0.891±0.020	0.15M	0.54G
	Invformer [46]	0.83±0.14	0.70±0.11	0.977±0.012	0.980±0.002	0.059±0.026	0.110±0.015	0.838±0.024	2.77M	3.46G
	DCFNet [47]	0.89±0.16	0.81±0.14	0.973±0.010	0.985±0.002	0.023±0.012	0.066±0.010	0.912±0.012	2.77M	3.46G
	HMPNet [48]	0.80±0.14	**0.56±0.10**	0.981±0.030	**0.993±0.003**	0.080±0.050	0.115±0.012	0.815±0.049	1.09M	2.00G
	PanDiff [49]	0.89±0.12	0.75±0.10	0.979±0.010	0.989±0.002	0.027±0.020	0.073±0.010	0.903±0.021	45.33M	14.83G
	PanMamba [50]	**0.68±0.12**	0.64±0.10	**0.982±0.008**	0.985±0.006	**0.016±0.008**	**0.045±0.009**	**0.939±0.010**	0.48M	1.31G
	Proposed	**0.62±0.12**	**0.55±0.11**	**0.986±0.007**	**0.993±0.002**	**0.019±0.009**	**0.045±0.010**	**0.936±0.012**	1.21M	2.34G
QuickBird (QB, 4-band)	MTF-GLP-FS [40]	8.11±1.96	7.51±0.79	0.830±0.091	0.900±0.020	0.049±0.015	0.138±0.024	0.820±0.034	—	—
	BT-H [41]	7.19±1.55	7.40±0.84	0.833±0.088	0.916±0.015	0.230±0.072	0.165±0.017	0.643±0.065	—	—
	LRTCFPan [42]	7.19±1.71	6.93±0.81	0.855±0.087	0.917±0.013	**0.023±0.012**	0.071±0.035	**0.909±0.044**	—	—
	DiCNN [43]	5.38±1.03	5.14±0.49	0.904±0.094	0.962±0.013	0.092±0.014	0.107±0.021	0.811±0.031	0.23M	0.19G
	FusionNet [44]	4.92±0.91	4.16±0.32	0.925±0.090	0.976±0.010	0.059±0.019	**0.052±0.009**	0.892±0.022	0.047M	0.32G
	LAGConv [45]	4.55±0.83	3.83±0.42	0.934±0.088	0.981±0.009	0.084±0.024	0.068±0.014	0.854±0.018	0.15M	0.54G
	Invformer [46]	4.66±0.78	3.70±0.29	0.932±0.007	**0.983±0.007**	0.174±0.033	0.073±0.024	0.766±0.043	2.77M	3.46G
	DCFNet [47]	**4.54±0.74**	3.83±0.29	0.933±0.090	0.974±0.010	0.045±0.015	0.124±0.027	0.836±0.016	2.77M	3.46G
	HMPNet [48]	4.72±0.38	**3.66±0.40**	0.930±0.110	0.980±0.009	0.183±0.054	0.079±0.025	0.754±0.065	1.09M	2.00G
	PanDiff [49]	4.58±0.74	3.74±0.31	**0.935±0.090**	0.982±0.090	0.059±0.022	0.064±0.025	0.881±0.042	45.33M	14.83G
	PanMamba [50]	5.14±0.90	4.95±0.42	0.921±0.086	0.976±0.009	**0.036±0.012**	0.067±0.015	0.900±0.010	0.48M	1.31G
	Proposed	**4.36±0.73**	**3.53±0.27**	**0.938±0.090**	**0.984±0.007**	0.037±0.016	**0.065±0.017**	**0.900±0.015**	1.21M	2.34G

color bias. For instance, U2Fusion [23] and CDDFuse [36] closely resemble the visible image, while SwinFusion [5], DDFM [33], and MGDN [24] are overly influenced by the infrared counterpart. These methods fail to achieve effective fusion. Both DeFuse [32] and TC-MOA [25] also suffer from significant color deviations. Our method stands out by better highlighting targets with respect to the compared approaches, clearly distinguishing the sky and the building, and preserving more details.

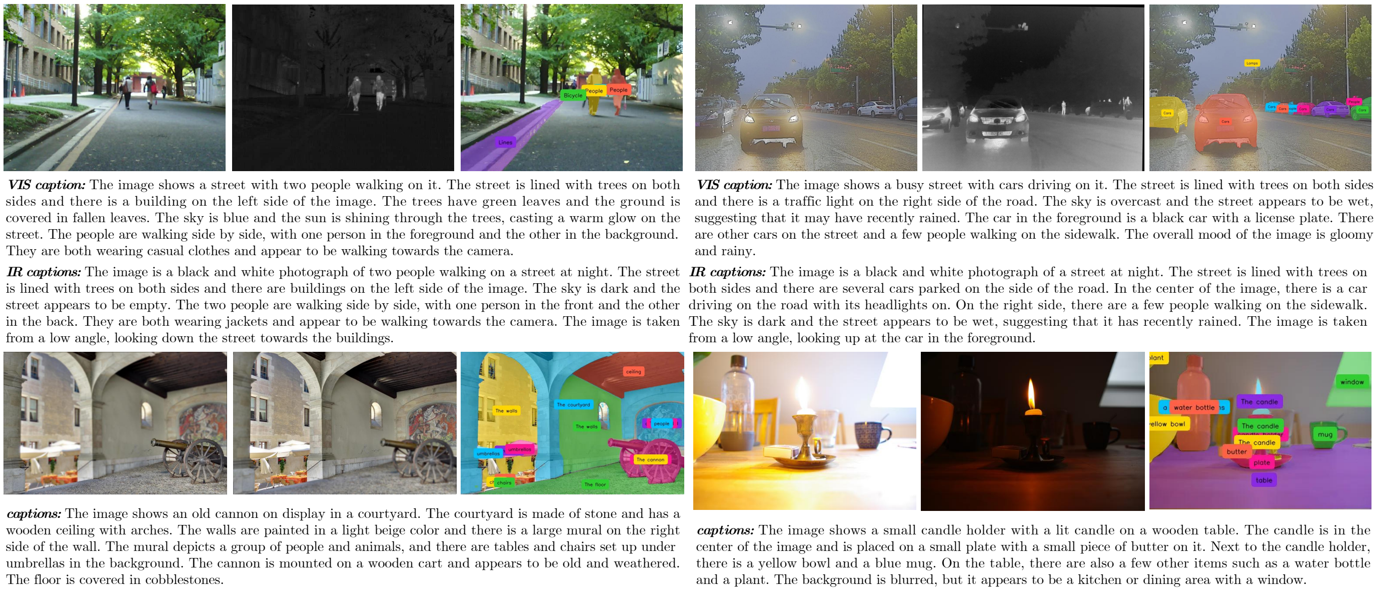
Fig. 5 displays a nighttime image pair from the RoadScene dataset. We focus on two specific details: the cars on the road and the slogans posted on the wall. Our method achieves the highest clarity for vehicles, closely resembling the infrared image. The slogans on the wall are also enhanced after fusion. From a global perspective, our fusion results effectively retain the road's zebra crossings from the visible image and the tree details from the infrared image. Compared to the other recent methods, our approach clearly demonstrates benefits.

Fig. 6 shows some visual results for the MFI-WHU and Lytro datasets. Fig. 6 (left panel) shows a golf course with a red flag in the foreground and houses and trees in the background. Comparing our approach with the other state-of-the-art methods, the proposed solution clearly preserves the cables in the blue box, while the distant houses remain unaffected by the foreground focus. The green box highlights the integration of the foreground and background objects, where our method achieves high clarity for both the red flag and the background houses. Instead, other methods exhibit lower clarity and introduce artifacts, such as the additional sky gap in ZMFF [34]. Fig. 6 (right panel) shows a golfer in a yellow polo shirt in the foreground and a yellow checkered flag in the background. Key details, such as the golfer's hand and the far flag, have been chosen for analysis. While most of the methods achieve successful fusion, DeFuse [32] and DDFM [33] show poor fusion quality. Observing the hand details (blue box) and the green lawn in the background, our method achieves higher fidelity. The checkered flag in the green box also demonstrates successful color reconstruction by incorporating details from the near image.

Fig. 7 provides visual comparisons for the pansharpening task on the GF2 and QB datasets. Error maps are characterized by significant spectral transformations among buildings. Fig. 7 (left panel) shows the white and blue buildings in the red and green boxes representing challenges for fusion. Our approach gets high fusion quality in these areas, while the other methods exhibit poor performance. In Fig. 7 (right panel), we focus on cars on the harbor road (green box) and isolated buildings near the coastline. Our method mainly obtains deep blue pixels in the related error map, indicating small errors. In the close-ups (red and green boxes), our approach avoids fusion defects, commonly seen in the other compared methods.

Fig. 8 illustrates fusion results for various methods on the Pavia dataset. The selected image showcases dense building clusters, where retaining the structural distribution of the building clusters poses a major challenge for this task. Both traditional and deep-learning methods generate error maps with large high-error regions. However, our method obtains overall low-error areas and preserves the structural distribution of the building clusters. This demonstrates that our RWKVFusion performs well even for the HMIF task.

Fig. 9 depicts the fusion results on two pairs of MRI-SPECT images. In the top two rows of Fig. 9, one can observe that the fused images produced by SwinFusion [5],



VIS caption: The image shows a street with two people walking on it. The street is lined with trees on both sides and there is a building on the left side of the image. The trees have green leaves and the ground is covered in fallen leaves. The sky is blue and the sun is shining through the trees, casting a warm glow on the street. The people are walking side by side, with one person in the foreground and the other in the background. They are both wearing casual clothes and appear to be walking towards the camera.

IR captions: The image is a black and white photograph of two people walking on a street at night. The street is lined with trees on both sides and there are buildings on the left side of the image. The sky is dark and the street appears to be empty. The two people are walking side by side, with one person in the front and the other in the back. They are both wearing jackets and appear to be walking towards the camera. The image is taken from a low angle, looking down the street towards the buildings.

VIS caption: The image shows a busy street with cars driving on it. The street is lined with trees on both sides and there is a traffic light on the right side of the road. The sky is overcast and the street appears to be wet, suggesting that it may have recently rained. The car in the foreground is a black car with a license plate. There are other cars on the street and a few people walking on the sidewalk. The overall mood of the image is gloomy and rainy.

IR captions: The image is a black and white photograph of a street at night. The street is lined with trees on both sides and there are several cars parked on the side of the road. In the center of the image, there is a car driving on the road with its headlights on. On the right side, there are a few people walking on the sidewalk. The sky is dark and the street appears to be wet, suggesting that it has recently rained. The image is taken from a low angle, looking up at the car in the foreground.

captions: The image shows an old cannon on display in a courtyard. The courtyard is made of stone and has a wooden ceiling with arches. The walls are painted in a light beige color and there is a large mural on the right side of the wall. The mural depicts a group of people and animals, and there are tables and chairs set up under umbrellas in the background. The cannon is mounted on a wooden cart and appears to be old and weathered. The floor is covered in cobblestones.

captions: The image shows a small candle holder with a lit candle on a wooden table. The candle is in the center of the image and is placed on a small plate with a small piece of butter on it. Next to the candle holder, there is a yellow bowl and a blue mug. On the table, there are also a few other items such as a water bottle and a plant. The background is blurred, but it appears to be a kitchen or dining area with a window.

Fig. 2: Additional examples of image captions and segmented masks for the VIF, MEF, and MFF tasks. The two image modalities are shown on the left. Instead, the mask is on the right and the captions are shown in the bottom panel.

Captions: The image shows a group of people walking on a street at night. The street is lined with trees and buildings on both sides, and there are street lamps on the right side of the image. The sky is dark and the street is lit up with street lamps, creating a warm glow. The people are walking in a line, and they appear to be walking towards the camera. The image is taken from a low angle, looking down the street towards the buildings.

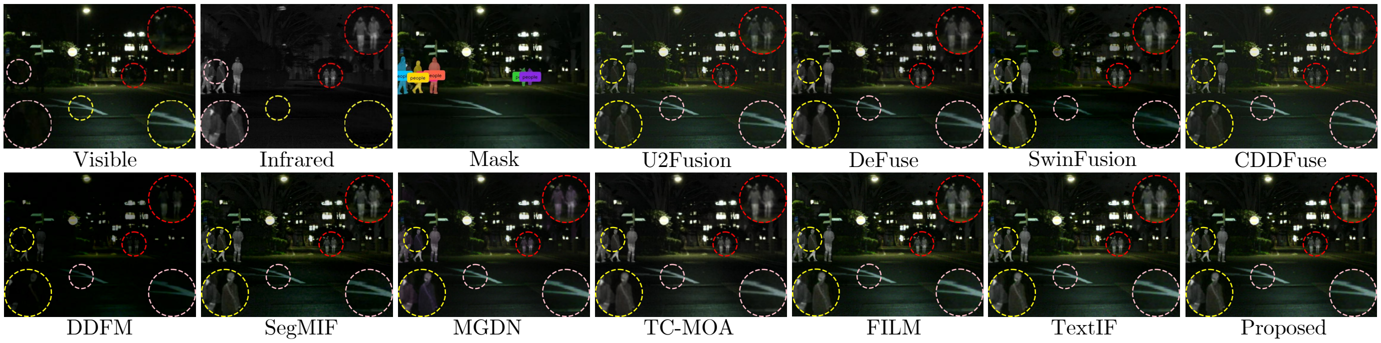


Fig. 3: Visual results for the compared approaches on the MSRS dataset (VIF task). Several close-ups are shown in the dashed circles.

Captions: The image is a black and white photograph of a small house in the woods. The house is a two-story structure with a sloping roof and a chimney. It appears to be old and dilapidated, with peeling paint and broken windows. There is a small porch on the front of the house with a door and a window on the second floor. A man and a woman are standing outside the house, looking at the camera. The ground is covered in fallen leaves and there are trees in the background. The image is taken from a low angle, looking up at the house.

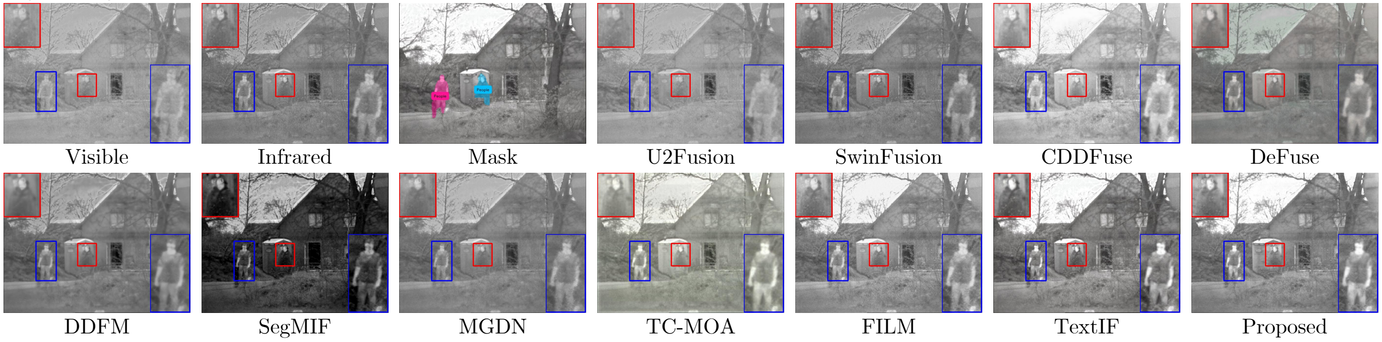
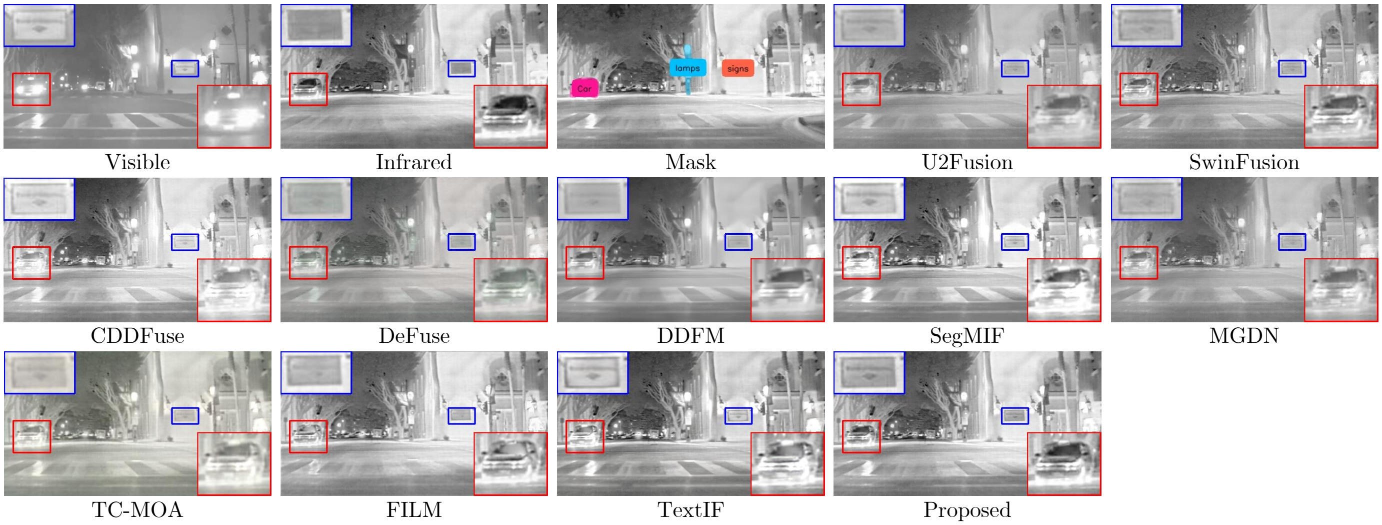


Fig. 4: Visual results for the compared approaches on the TNO dataset (VIF task). Close-ups are shown in the blue and red boxes.

Captions: The image is a black and white photograph of a street at night. The street is lined with trees and buildings on both sides, and there are street lamps on the right side of the image. In the center of the street, there is a white building with a sign that reads "Cafe". The building appears to be a restaurant or a bar, and the street is wet, suggesting that it has recently rained. There are a few cars driving on the street and a few people walking on the sidewalk. The sky is dark and the overall mood of the photograph is gloomy.

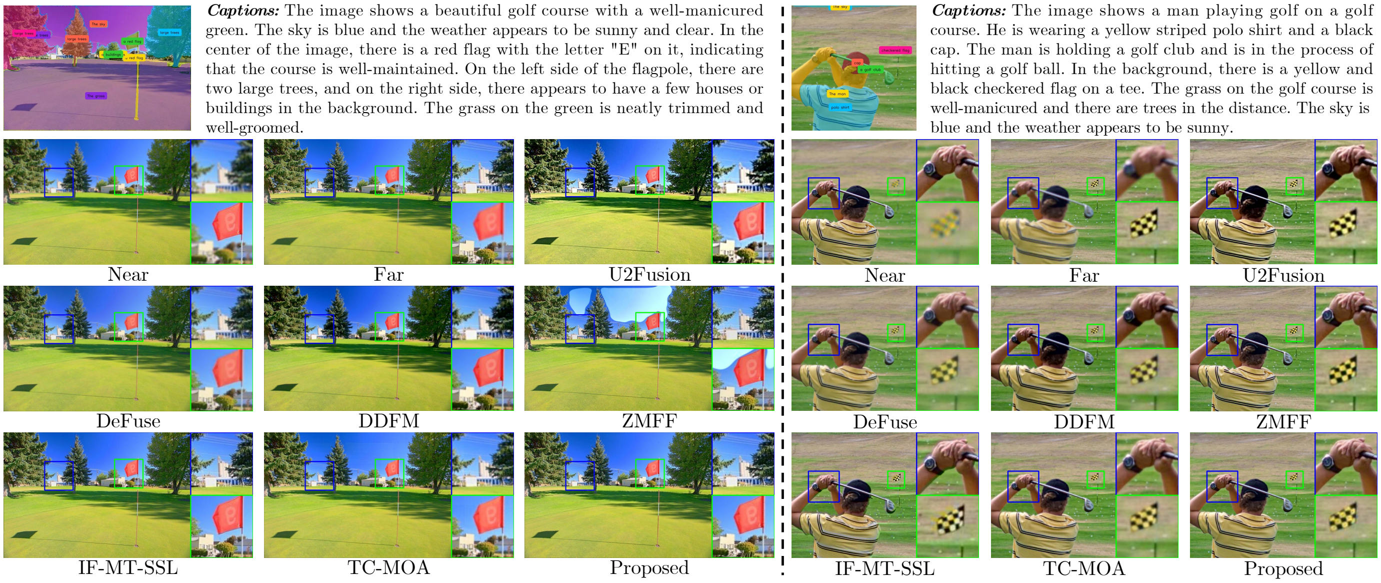


Visible Infrared Mask U2Fusion SwinFusion

CDDFuse DeFuse DDFM SegMIF MGDN

TC-MOA FILM TextIF Proposed

Fig. 5: Visual results for the compared approaches on the RoadScene dataset (VIF task). Close-ups are shown in the blue and red boxes.



Captions: The image shows a beautiful golf course with a well-manicured green. The sky is blue and the weather appears to be sunny and clear. In the center of the image, there is a red flag with the letter "E" on it, indicating that the course is well-maintained. On the left side of the flagpole, there are two large trees, and on the right side, there appears to have a few houses or buildings in the background. The grass on the green is neatly trimmed and well-groomed.

Captions: The image shows a man playing golf on a golf course. He is wearing a yellow striped polo shirt and a black cap. The man is holding a golf club and is in the process of hitting a golf ball. In the background, there is a yellow and black checkered flag on a tee. The grass on the golf course is well-manicured and there are trees in the distance. The sky is blue and the weather appears to be sunny.

Near Far U2Fusion

DeFuse DDFM ZMFF

IF-MT-SSL TC-MOA Proposed

Near Far U2Fusion

DeFuse DDFM ZMFF

IF-MT-SSL TC-MOA Proposed

Fig. 6: Visual results for the compared approaches on *"44"* from the MFI-WHU test set (MFF task; left panel) and *"1"* from the Lytro test set (MFF task; right panel). Close-ups are depicted in the green and blue boxes.

CDDFuse [36], DDFM [33], and MGDN [24] (see the close-ups in the green boxes) all exhibit varying degrees of artifacts at the interface between the brain's internal structures and the exoskeleton, caused by SPECT pixel blocks. TC-MOA [25], due to its excessively blurred fused images, is likely to hinder medical experts from making further accurate judgments. This issue is also evident in the last two rows (second example) in Fig. 9. In contrast, our method not only eliminates these artifacts but also preserves the color information from SPECT while retaining the rich texture details from MRI.

10 Computational Burden

To assess the computational burden, we comprehensively evaluate the number of parameters, computational complexity (FLOPs), and throughput. Taking the visible-infrared fusion (VIF) task as an example, Tab. 4 compares these metrics across multiple methods. The proposed RWKVFusion, leveraging its global receptive field and efficient CUDA implementation, demonstrates differentiated advantages under three configurations:

$*$ configuration (fusion network only) achieves a high throughput of 136.6 images/s with an ultra-low parameter count of 0.35M, delivering near-SOTA performance (see ablation study in Tab. 7 (*iii*) of the main paper.);

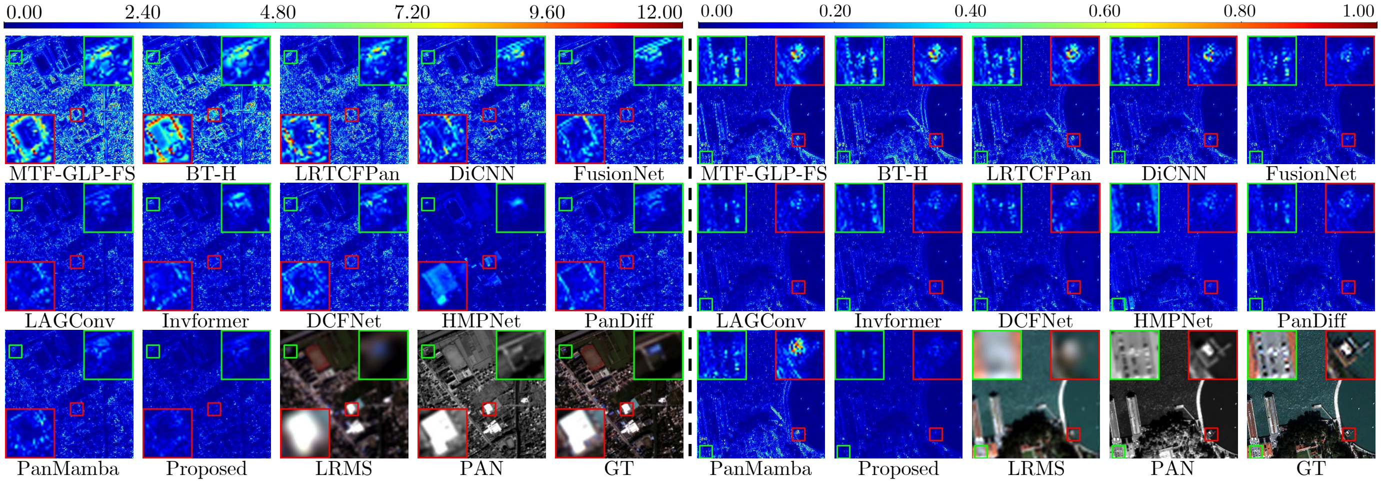


Fig. 7: Error maps for the compared approaches on *"area 19"* of the GF2 test set (left panel) and on *"area 2"* of the QB test set (right panel). Close-ups are depicted in the green and red boxes.

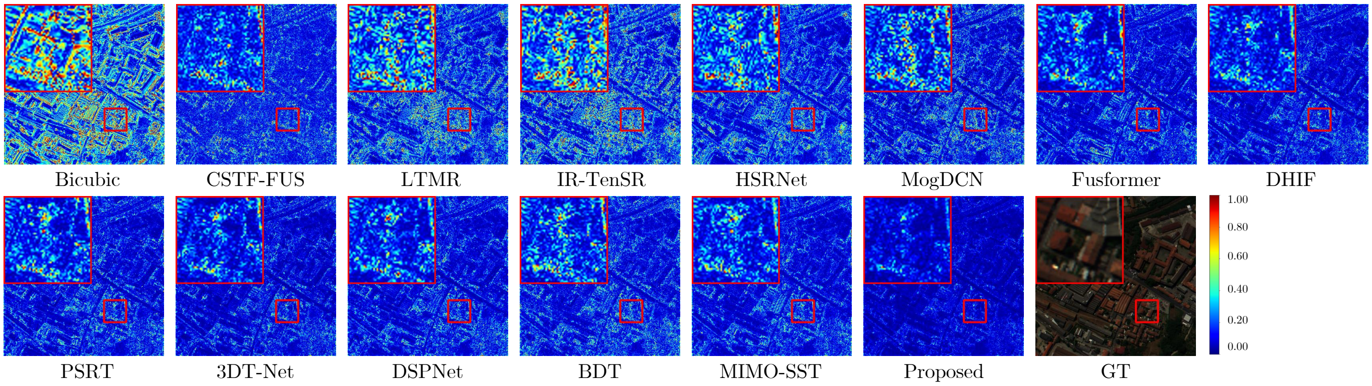


Fig. 8: Error maps for the compared approaches on *"area 2"* of the Pavia test set. Close-ups are depicted in the red boxes.

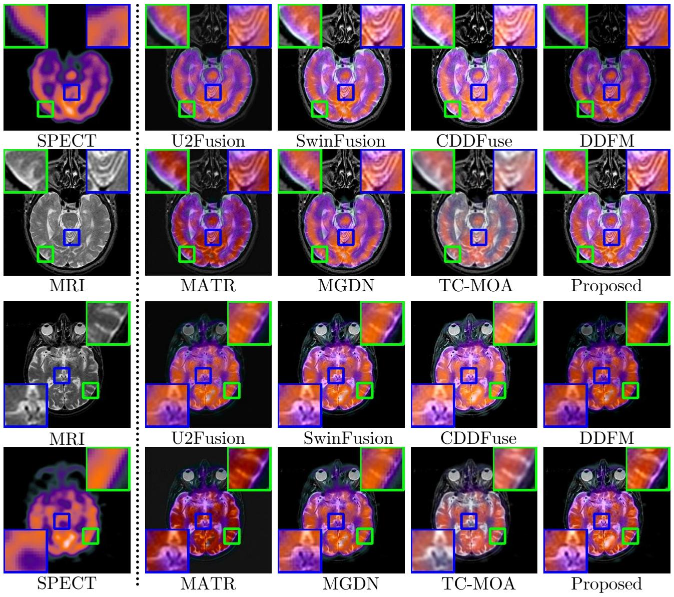


Fig. 9: Visual comparisons on the medical Harvard dataset. Close-ups are depicted in the green and blue boxes.

† configuration (offline semantic generation) maintains competitiveness at 112.7 images/s, exhibiting 139% higher efficiency than offline methods like FILM (47.1 images/s);

‡ configuration (full online processing, including caption and masks), while requiring higher parameters/FLOPs, concurrently provides an image caption and a unified segmentation mask subsequent to mask merging, validating the enhancement of fusion quality by high-level semantics. While in this setting, the view shows our pipeline might not be superior solely based on parameter count and FLOPs, it is essential to consider the unique value proposition of our method. Beyond achieving SOTA fusion performance, our approach **concurrently provides an image caption and a unified segmentation mask subsequent to mask merging**. These constitute key advantages. This outcome validates our foundational premise that high-level semantic information from captions and masks can significantly improve fusion quality, a research avenue rarely explored in image fusion.

11 More Results of Downstream Tasks

In this section, we provide more results of RWKVFusion on downstream tasks and comparisons with other models.

Fig. 10 presents monocular depth estimation results using Depth Anything v2 [52] across multiple image fusion tasks. The first row displays fused images and their

TABLE 4: Comparisons of the model parameters, FLOPs, and throughput for fusing a 256 × 256 image. "A+B" (of previous FILM, TextIF, and proposed†) indicates that A represents the fusion network's parameters (or FLOPs), while B corresponds to the language model's parameters (or FLOPs). "A+B+C+D" for the proposed‡ method means parameters and FLOPs of RWKVFusion model, T5 model, Florence and DINO model, and SAM model. N/A, M, G, and s stand for not available, million, giga, and second, respectively. Gray color indicates methods of image processing-based. Proposed* indicates that do not use semantic information (i.e., without caption and masks), the performance of this model is shown in Tab. 7 (*iii*) of the main paper.

Architecture	Params (M)	FLOPs (G)	Throughput (images/s)
DeFuse [32]	7.87	15.17	139.2
U2Fusion [23]	0.66	43.31	315.8
SwinFusion [5]	0.97	59.41	16.4
CDDFuse [36]	1.19	118.2	25.4
DDFM [33]	553.1	1112	20.1
MATR [51]	0.013	3.36	51.4
MGDN [24]	0.91	65.0	12.0
TC-MOA [25]	340.9	524.3	26.6
Proposed*	0.352	15.08	136.6
FILM [38]	2.1+N/A	209.1+N/A	47.1
TextIF [39]	64.8+151.2	336.1+2.91	35.8
Proposed†	0.352+35.3	15.08+11.27	112.7
Proposed‡	0.352+35.3 +765.2+481.8	15.08+11.27 +934.3+179.9	0.742

depth maps generated by our RWKVFusion, which exhibit precise geometric structures and coherent depth hierarchy for distant backgrounds. The last two rows compare input modalities (visible, infrared, etc.) and corresponding depth estimations for MEF, MFF, and visible-VIF tasks. As can be seen, the fused images generated by the proposed RWKVFusion under challenging illumination (e.g., overexposed regions) and cross-modal scenarios (e.g., low-visibility infrared images) yield depth maps that align with human spatial perception.

The object detection outcomes obtained using YOLOv5 [53] on our fusion results for the MSRS dataset are depicted in Fig. 11. The YOLOv5 model, pretrained on the COCO dataset, has been directly applied to detect objects in the fused images. It can be observed that some objects remain undetected, and this issue can be attributed to the distribution shift between the COCO and MSRS datasets. To address this limitation, the fine-tuning of the detector on the fused images can be considered a promising solution.

As for the semantic segmentation task, we employ the Segformer [54] architecture with a pretrained MiT-B3 backbone. Models are trained and evaluated on fused image/mask pairs generated by different VIF methods from the MSRS VIF dataset. Fig. 12 demonstrates an additional visual segmentation result on fused images generated by different fusion methods. As shown in the figure, the segmentation results of our proposed method align most closely with the ground truth labels. The primary limitations of other methods include edge fragmentation in complex scenes (e.g., bicycles) and incomplete segmentation of small-scale objects (e.g., distant pedestrians and road curves).

Comprehensive downstream experiments demonstrate RWKVFusion's superiority in fusion performance and downstream adaptability.

12 Code and Data

We will release the code and data, including image modalities, captions, and masks, at https://github.com/294coder/RWKVFusion.

13 Contributions

Zi-Han Cao: Developed the main idea, conducted experiments, and wrote the manuscript.
Yu-Jie Liang: Developed the main idea, conducted experiments, and wrote the manuscript.
Liang-Jian Deng: Provided main supervision and revised the manuscript.
Gemine Vivone: Revised the manuscript.
Jocelyn Chanussot: Revised the manuscript in the first round review.

We are deeply grateful to Jocelyn Chanussot for his assistance during the first round of reviews. However, due to time constraints, he is no longer involved in the revision of the manuscript.

References

[1] 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.

[2] 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.

[3] J. Cai, S. Gu, and L. Zhang, "Learning a deep single image contrast enhancer from multi-exposure images," *IEEE Trans. Image Process.*, vol. 27, no. 4, pp. 2049–2062, 2018.

[4] X. Zhang, "Benchmarking and comparing multi-exposure image fusion algorithms," *Inf. Fus.*, pp. 111–131, 2021.

[5] 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.

[6] J. Zhang, Q. Liao, S. Liu, H. Ma, W. Yang, and J.-H. Xue, "Real-mff: A large realistic multi-focus image dataset with ground truth," *Pattern Recognit. Lett.*, vol. 138, pp. 370–377, 2020.

[7] H. Zhang, Z. Le, Z. Shao, H. Xu, and J. Ma, "Mff-gan: An unsupervised generative adversarial network with adaptive and gradient joint constraints for multi-focus image fusion," *Inf. Fus.*, vol. 66, pp. 40–53, 2021.

[8] L. Wald, T. Ranchin, and M. Mangolini, "Fusion of satellite images of different spatial resolutions: Assessing the quality of resulting images," *Photogramm. Eng. Remote Sens.*, vol. 63, no. 6, pp. 691–699, 1997.

[9] N. Yokoya and A. Iwasaki, "Airborne hyperspectral data over chikusei," Space Application Laboratory, University of Tokyo, Japan, Tech. Rep. SAL-2016-05-27, May 2016.

[10] Y.-W. Zhuo, T.-J. Zhang, J.-F. Hu, H.-X. Dou, T.-Z. Huang, and L.-J. Deng, "A deep-shallow fusion network with multi-detail extractor and spectral attention for hyperspectral pansharpening," *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 7539–7555, 2022.

[11] 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.

[12] I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts," in *ICLR*, 2017.

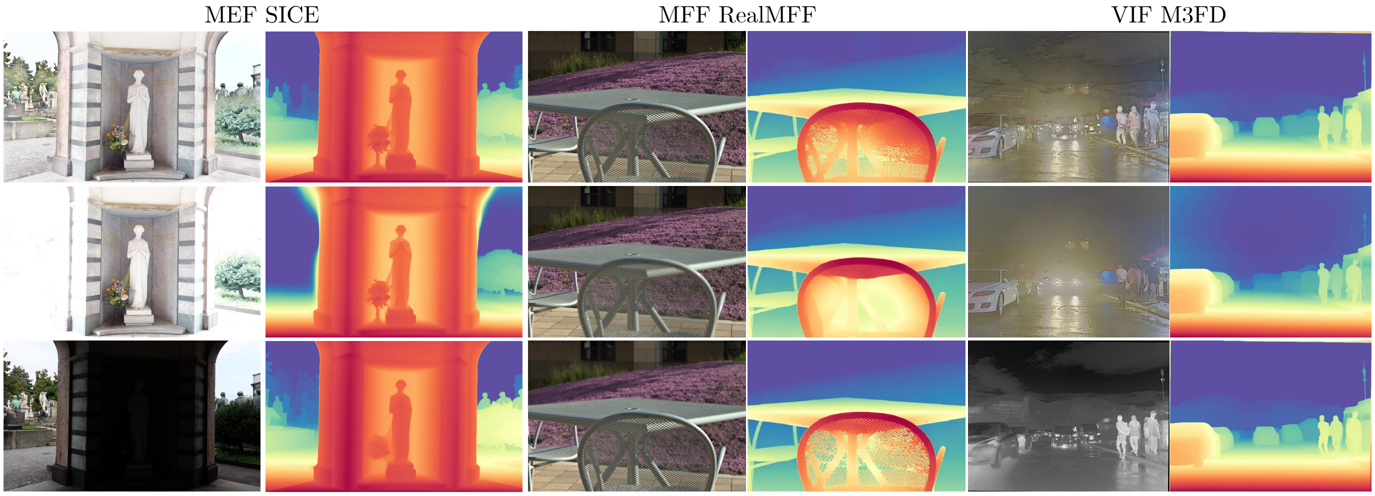


Fig. 10: Monocular depth estimation results by DepthAnything v2 [52] on different image fusion tasks including MEF, MFF, and VIF. The first row includes our fused images and corresponding estimated depth maps. The last two rows are the image modalities and their estimated depth maps.

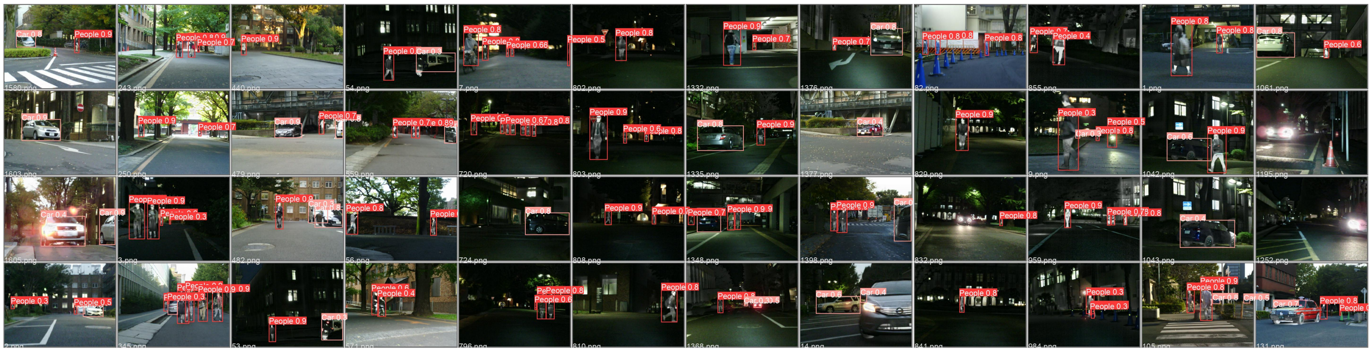


Fig. 11: Detection outcomes on our fusion results for the MSRS dataset by using the pre-trained YOLOv5 [53]. The class and the related confidence value for each object are shown on the related bounding box.

[13] P. Tillet, H.-T. Kung, and D. Cox, “Triton: an intermediate language and compiler for tiled neural network computations,” in *Proceedings of the 3rd ACM SIGPLAN International Workshop on Machine Learning and Programming Languages*, 2019, pp. 10–19.

[14] 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.

[15] 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.

[16] G. Vivone, L.-J. Deng, S. Deng, D. Hong, M. Jiang, C. Li, W. Li, H. Shen, X. Wu, J.-L. Xiao *et al.*, “Deep learning in remote sensing image fusion: Methods, protocols, data, and future perspectives,” *IEEE Geosci. Remote Sens. Mag.*, 2024.

[17] 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.

[18] R. H. Yuhas, A. F. Goetz, and J. W. Boardman, “Discrimination among semi-arid landscape endmembers using the spectral angle mapper (sam) algorithm,” in *JPL AGW-3 Vol. 1: AVIRIS Workshop.*, 1992.

[19] L. Wald, *Data fusion: definitions and architectures: fusion of images of different spatial resolutions*. Presses des MINES, 2002.

[20] A. Garzelli and F. Nencini, “Hypercomplex quality assessment of multi/hyperspectral images,” *IEEE Geoscience and Remote Sensing Letters*, vol. 6, no. 4, pp. 662–665, 2009.

[21] 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.

[22] A. Horé and D. Ziou, “Image quality metrics: Psnr vs. ssim,” in *International Conference on Pattern Recognition (ICIP)*, 2010, pp. 2366–2369.

[23] 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.

[24] 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.

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

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

[27] J. Li, D. Li, S. Savarese, and S. Hoi, “BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models,” in *ICML*. PMLR, 2023, pp. 19 730–19 742.

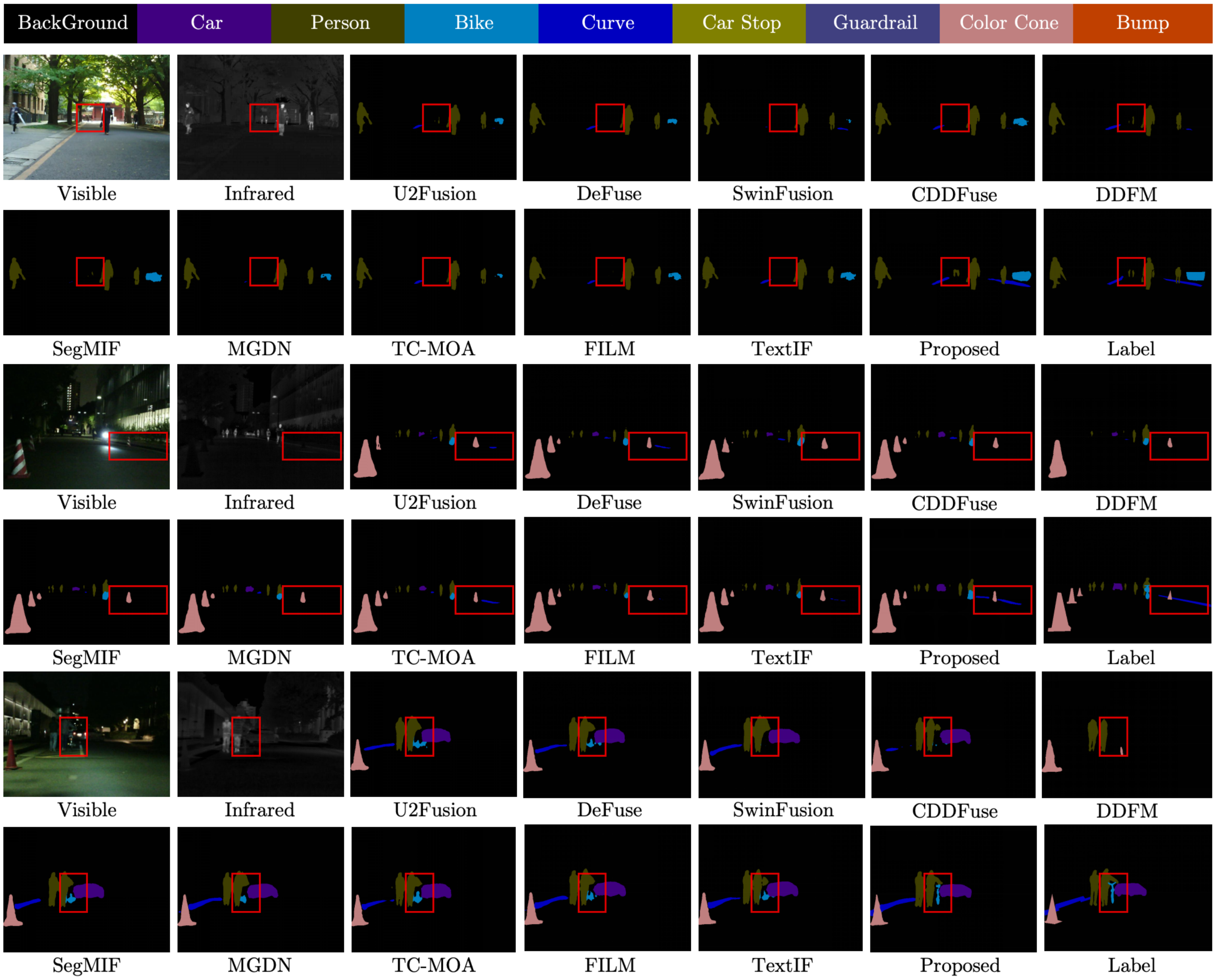
[28] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, “Swin transformer: Hierarchical vision transformer using shifted windows,” in *CVPR*, 2021, pp. 10 012–10 022.

[29] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, “Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network,” in *CVPR*, 2016, pp. 1874–1883.

[30] J. Cai, S. Gu, and L. Zhang, “Learning a deep single image contrast enhancer from multi-exposure images,” *IEEE Trans. Image Process.*, vol. 27, no. 4, pp. 2049–2062, 2018.

[31] H. Xu, J. Ma, Z. Le, J. Jiang, and X. Guo, “FusionDN: A unified densely connected network for image fusion,” in *AAAI*, 2020.

[32] 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.



Visible Infrared U2Fusion DeFuse SwinFusion CDDFuse DDFM

SegMIF MGDN TC-MOA FILM TextIF Proposed Label

Visible Infrared U2Fusion DeFuse SwinFusion CDDFuse DDFM

SegMIF MGDN TC-MOA FILM TextIF Proposed Label

Visible Infrared U2Fusion DeFuse SwinFusion CDDFuse DDFM

SegMIF MGDN TC-MOA FILM TextIF Proposed Label

Fig. 12: Semantic segmentation results using Segformer [54] 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.

[33] 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.

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

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] 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.

[40] 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.

[41] 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.

[42] 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.

[43] 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.

[44] 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.

[45] 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.

[46] 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.

[47] 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.

[48] 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.

[49] 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.

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

[51] 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.

[52] L. Yang, B. Kang, Z. Huang, Z. Zhao, X. Xu, J. Feng, and H. Zhao, "Depth anything v2," *arXiv:2406.09414*, 2024.

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

[54] 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.