Supplementary of U-Know-DiffPAN

A. More Details of Methods

A.1. Details of Vector Extractor

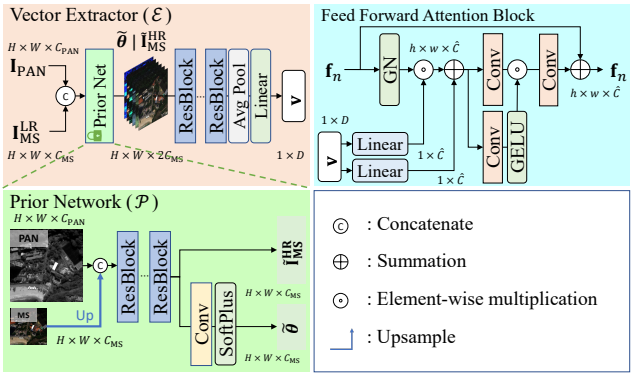

Figure 1. Detailed structure of Vector Extractor ($\mathcal{E}$). The simple representation of $\mathcal{E}$ is presented in Fig. 3 of the main paper.

The detailed structure of the Vector Extractor $\mathcal{E}$ in Fig.3 of the main paper is depicted in Fig. 1 and Eq. 1, which extracts an compact vector representation $\mathbf{v} \in \mathbb{R}^{1\times D}$ from two inputs: $\mathbf{I}_{\text{PAN}}$ and $\mathbf{I}_{\text{MS}}^{\text{LR}}$. $\mathcal{E}$ leverages a pretrained lightweight prior network $\mathcal{P}$ that takes $\mathbf{I}_{\text{PAN}}$ and $\mathbf{I}_{\text{MS}}^{\text{LR}}$ as inputs to generate a prior HRMS $\tilde{\mathbf{I}}_{\text{MS}}^{\text{HR}}$ and a prior uncertainty map $\tilde{\boldsymbol{\theta}}$. These outputs, $\tilde{\mathbf{I}}_{\text{MS}}^{\text{HR}}$ and $\tilde{\boldsymbol{\theta}}$, are then processed through multiple ResBlocks, followed by average pooling and a linear layer, to produce the final compact vector representation $\mathbf{v}$.

$$\begin{aligned} \left[\tilde{\boldsymbol{\theta}} \mid \tilde{\mathbf{I}}_{\text{MS}}^{\text{HR}}\right] &= \mathcal{P}\left(\mathbf{I}_{\text{PAN}}, \mathbf{I}_{\text{MS}}^{\text{LR}}\right), \\ \mathbf{v} &= \text{Linear}\left(\text{AvgPool}\left(\text{ResBlock}^{n}\left(\left[\tilde{\boldsymbol{\theta}} \mid \tilde{\mathbf{I}}_{\text{MS}}^{\text{HR}}\right]\right)\right)\right). \end{aligned} \tag{1}$$

where $\tilde{\boldsymbol{\theta}}$ denotes prior uncertainty map and $\tilde{\mathbf{I}}_{\text{MS}}^{\text{HR}}$ is prior HRMS obtained from prior network $\mathcal{P}$. The compact vector representation $\mathbf{v}$ not only encapsulates the combined information from $\mathbf{I}_{\text{PAN}}$ and $\mathbf{I}_{\text{MS}}^{\text{LR}}$, but also integrates the uncertainty information $\tilde{\boldsymbol{\theta}}$ derived from the prior network $\mathcal{P}$. This representation is subsequently used as a conditioning input to the Feed Forward Attention (FFA) block in the encoder blocks of FSA-T.

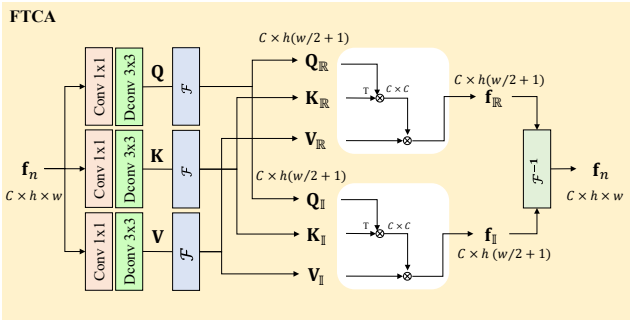

Figure 2. Detailed structure of Fourier Transform Channel Attention (FTCA) block. The simple representation of FTCA is presented in Fig. 3 of the main paper.

A.2. Fourier Transform Channel Attention (FTCA)

The detailed structure of FTCA in Fig. 3 of the main paper is depicted in Fig. 2, which is designed to enhance frequency domain features by applying the Discrete Fourier Transform (DFT) to feature maps, allowing self-attention to be performed more effectively in the frequency domain. As shown in Fig. 2, the input feature map is first transformed from the time domain to the frequency domain using 2D DFT. The real and imaginary parts of the 2D DFT feature map are then processed separately using channel attention. After applying channel attention to both real and imaginary parts, the Inverse Fourier Transform is applied to bring the features back into the time domain, producing the enhanced output. This transform-domain self attention facilitates more effective computations of frequency components to selectively emphasize. This process ensures that high-quality frequency information be effectively captured and utilized, resulting in enhanced feature representation and improved model performance.

A.3. Stationary Wavelet Transform Cross Attention (SWTCA)

Following the approach of ResDiff [13], WaveDiff [11], and WINet [19], the SWTCA in Fig. 3 of the main paper is detailed in Fig. 3, which utilizes wavelet components of PAN and MS images decomposed using the Stationary Wavelet Transform (SWT). The SWTCA block is designed to inject additional frequency information into the features $\mathbf{f}_n$

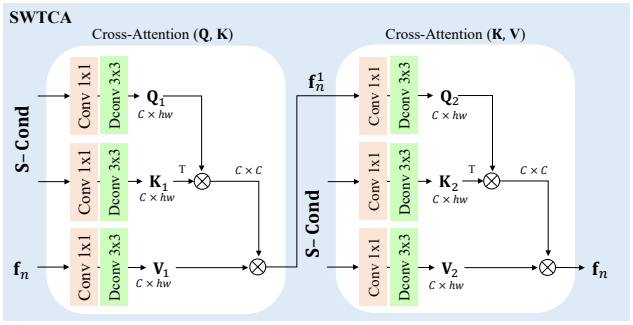

Figure 3. Detailed structure of our Stationary Wavelet Transform Cross Attention (SWTCA) block. The simple representation of SWTCA is presented in Fig. 3 of the main paper.

enhanced by the preceding FTCA. It selectively incorporates useful frequency components derived from the $\mathbf{I}_{\text{PAN}}$ and $\mathbf{I}_{\text{MS}}^{\text{LR}}$ using the SWT while maintaining shift invariance as conditions for our Diffusion Model Ψ . The outputs of SWT for $\mathbf{I}_{\text{PAN}}$ and $\mathbf{I}_{\text{MS}}^{\text{LR}}$ are constructed respectively as follows:

$$\begin{aligned}\mathcal{S}(\mathbf{I}_{\text{PAN}}) &= \left[\mathbf{L}_{\text{PAN}} \mid \mathbf{H}_{\text{PAN}} \mid \mathbf{V}_{\text{PAN}} \mid \mathbf{D}_{\text{PAN}}\right], \\ \mathcal{S}(\mathbf{I}_{\text{MS}}^{\text{LR}}) &= \left[\mathbf{L}_{\text{MS}}^{\text{LR}} \mid \mathbf{H}_{\text{MS}}^{\text{LR}} \mid \mathbf{V}_{\text{MS}}^{\text{LR}} \mid \mathbf{D}_{\text{MS}}^{\text{LR}}\right],\end{aligned} \tag{2}$$

where $\mathbf{L}$ represent the low-frequency approximation components, while $\mathbf{H}$, $\mathbf{V}$, $\mathbf{D}$ denote the high-frequency horizontal, vertical, and diagonal details, respectively. Since $\mathbf{I}_{\text{PAN}}$ provides highly detailed texture from its high spatial resolution but lacks spectral information, only its high-frequency components, which include $\mathbf{H}_{\text{PAN}}$, $\mathbf{V}_{\text{PAN}}$ and $\mathbf{D}_{\text{PAN}}$, are used. Conversely, $\mathbf{L}_{\text{MS}}^{\text{LR}}$ from $\mathbf{I}_{\text{MS}}^{\text{LR}}$ contains richer spectral information. So, these components are concatenated channel-wise to form $\mathcal{S}$-Cond that is the condition to be inputted to Ψ. $\mathcal{S}$-Cond is then constructed as follows:

$$\mathcal{S}\text{-Cond} = \left[\mathbf{L}_{\text{MS}}^{\text{LR}} \mid \mathbf{H}_{\text{PAN}} \mid \mathbf{V}_{\text{PAN}} \mid \mathbf{D}_{\text{PAN}}\right]. \tag{3}$$

Note that the construction of this kind condition is motivated from the traditional MRA(Multi-Resolution Analysis)-based [1, 10, 22] PAN-sharpening methods that have utilized high-frequency Wavelet components of $\mathbf{I}_{\text{PAN}}$ and low-frequency Wavelet components of $\mathbf{I}_{\text{MS}}^{\text{LR}}$. To inject $\mathcal{S}$-Cond into the feature $\mathbf{f}_n$ obtained from FTCA at the n-th decoder block, we employ a two-stage cross-attention process. In the first stage, $\mathbf{Q}_1$ and $\mathbf{K}_1$ are derived from $\mathcal{S}$-Cond, directing attention toward important frequency components, while $\mathbf{V}_1$ is derived from $\mathbf{f}_n$. The resulting intermediate feature $\mathbf{f}_n^1 \in \mathbb{R}^{C\times h\times w}$ is given as:

$$\begin{aligned}&[\mathbf{Q}_1 \mid \mathbf{K}_1] = \mathsf{Conv}(\mathcal{S}\text{-Cond}),\ \mathbf{V}_1 = \mathsf{Conv}(\mathbf{f}_n), \\ &\mathbf{f}_n^1 = \mathsf{SoftMax}\left(\mathbf{Q}_1\mathbf{K}_1^{\mathsf{T}}/\sqrt{C}\right)\mathbf{V}_1,\end{aligned} \tag{4}$$

where $\mathbf{Q}_1$, $\mathbf{K}_1$, $\mathbf{V}_1 \in \mathbb{R}^{C\times hw}$ are a Query, a Key and a Value, respectively. The $\mathbf{Q}_1$, $\mathbf{K}_1$ and $\mathbf{V}_1$ are reshaped to $C \times hw$ for channel attention. In the second stage, $\mathbf{Q}_2$ is derived from the intermediate feature $\mathbf{f}_n^1$, while $\mathbf{K}_2$ and $\mathbf{V}_2$ come from $\mathcal{S}$-Cond, enabling the model to refine the correlation between the reconstructed feature $\mathbf{f}_n$ and the frequency components in $\mathcal{S}$-Cond. This stage yields the final feature $\mathbf{f}_n \in \mathbb{R}^{C\times h\times w}$ after reshaping $\mathbf{Q}_2$, $\mathbf{K}_2$ and $\mathbf{V}_2$ into $C \times hw$ for channel attention, which is given as follows:

$$\begin{aligned}&\mathbf{Q}_2 = \mathsf{Conv}(\mathbf{f}_n^1),\ [\mathbf{K}_2 \mid \mathbf{V}_2] = \mathsf{Conv}(\mathcal{S}\text{-Cond}), \\ &\mathbf{f}_n \leftarrow \mathsf{SoftMax}\left(\mathbf{Q}_2\mathbf{K}_2^{\mathsf{T}}/\sqrt{C}\right)\mathbf{V}_2.\end{aligned} \tag{5}$$

The first cross-attention block in Fig. 3 uses both Query and Key mappings from $\mathcal{S}$-Cond, focusing the attention map purely on $\mathcal{S}$-Cond information, and aiding in concentrating on the significant channels of the feature $\mathbf{f}_n$. The second cross-attention block in Fig. 3 uses Query from the intermediate feature $\mathbf{f}_n^1$ and Key from $\mathcal{S}$-Cond, learning the correlation between the reconstructed feature $\mathbf{f}_n^1$ and the meaningful frequency information in $\mathbf{I}_{\text{PAN}}$ and $\mathbf{I}_{\text{MS}}^{\text{LR}}$. This block helps emphasizing significant frequency components within $\mathcal{S}$-Cond.

A.4. Student network FSA-S (ψ) architecture

The student model, denoted as FSA-S ψ, is a diffusion-based model designed with the same number of encoder and decoder blocks as FSA-T Ψ. Each block consists solely of ResBlocks, and unlike the teacher Ψ, the FSA-S operates without any additional conditional inputs. As shown in Fig. 2, the lightweight FSA-S ψ predicts denoised image $\widetilde{\mathbf{X}}_0$ from input $\mathbf{X}_t$ at timestep t as:

$$\widetilde{\mathbf{X}}_0 = \psi\left(\left[\mathbf{X}_t \mid \mathbf{I}_{\text{PAN}} \mid \mathbf{I}_{\text{MS}}^{\text{LR}}\right]; t\right). \tag{6}$$

B. Stationary Wavelet Transform compared to Discrete Wavelet Transform

Discrete Wavelet transforms have been widely used in image processing and PAN-sharpening tasks due to their ability to analyze signals across multiple resolutions. Discrete Wavelet Transform (DWT) decomposes signals into low-frequency (approximation) and high-frequency (detail) components at each level. Although DWT has proven to be effective in numerous applications, it suffers from shift variance due to its inherent downsampling operation at each decomposition level. To address this limitation, the Stationary Wavelet Transform (SWT) was introduced. Unlike DWT, SWT omits the downsampling step, ensuring that the signal size remains constant across all levels. This design enables SWT to maintain shift invariance, meaning that minor shifts in the input signal do not affect the transformation results. This property makes SWT highly advantageous in applications such as PAN-sharpening, where maintaining consistency in transformed results is crucial. Specifically, SWT ensures that the wavelet coefficients remain stable even

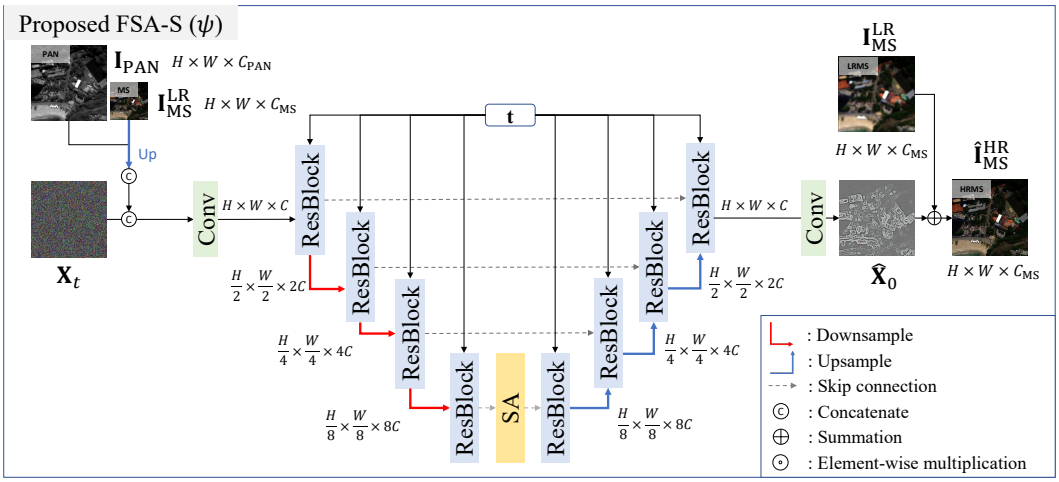

Figure 4. Detailed structure of FSA-S ψ.

when the input signal undergoes small translations, leading to more robust performance in PAN-sharpening tasks. In addition, SWT allows for higher resolution analysis across all levels, as the image or signal is not downsampled during the transformation. Compared to DWT that inherently reduces the image resolutions at each level, SWT retains the original resolution, making it more suitable for tasks that require detailed frequency information.

B.1. SWT condition versus DWT condition

	GF2 Dataset (Reduced-Resolution)			
Condition	SAM↓	ERGAS↓	SCC↑	Q4↑
DWT	0.646 ± 0.117	0.567 ± 0.095	0.993 ± 0.002	0.987 ± 0.007
SWT	**0.603 ± 0.102**	**0.537 ± 0.077**	**0.994 ± 0.001**	**0.988 ± 0.006**

Table 1. Comparison of results between DWT and SWT conditioning at the SWTCA block in FSA-T Ψ, with the best values highlighted in red.

Table. 1 represents the performance differences between DWT and SWT for the $\mathcal{S}$-Cond in Eq. 3 as the conditioning input to the SWTCA block in FSA-T Ψ. The SWT-based $\mathcal{S}$-Cond outperforms the DWT-based one across all metrics. The results in Table. 1 highlight that SWT achieves lower SAM and ERGAS values, as well as higher SCC and Q4 scores, indicating its superior performance in PAN-sharpening tasks. This improvement underscores the advantages of shift invariance and resolution preservation provided by SWT.

C. Uncertainty Estimation

Uncertainty estimation has been an important topic of research in deep learning. Several works [2, 7, 12, 14] have incorporated uncertainty into regression problems. A Bayesian deep learning framework was proposed to enable its application to per-pixel computer vision tasks. Similarly, some other works [3, 6, 12] explored the role of data uncertainty by modeling both the mean and variance of the predictions. Uncertainty-based loss function of those works can be represented as:

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^{N} \frac{\|\mathbf{x}_i - f(\mathbf{y}_i)\|_2}{2\boldsymbol{\sigma}_i^2} + \frac{1}{2} \ln \boldsymbol{\sigma}_i^2 \tag{7}$$

where N denotes total number of input samples, $\mathbf{x}_i$ is a target label, $\mathbf{y}_i$ is an input, $f(\mathbf{y}_i)$ and $\boldsymbol{\sigma}_i^2$ denote the learned mean and variance, respectively. Recent studies [9, 12] have continued to explore uncertainty estimation in deep learning, particularly in a task requiring high accuracy such as image super-resolution [9]. These approaches have demonstrated that such uncertainty-based losses that utilize uncertainty terms can achieve better results than mean square error (MSE) and mean absolute error (MAE) losses. Inspired by this work, we formulated our $\mathcal{L}_{\text{U-Diff}}$ as:

$$\mathcal{L}_{\text{U-Diff}} = \left\| \frac{1}{2\hat{\boldsymbol{\theta}}} \odot \left| \hat{\mathbf{X}}_0 - \mathbf{X}_0 \right| + \frac{1}{2} \log \hat{\boldsymbol{\theta}} \right\|_1, \tag{8}$$

where $\hat{\mathbf{X}}_0$ is a predicted residual, $\mathbf{X}_0$ is a target, and $\hat{\boldsymbol{\theta}}$ serves as the estimated variance term and is regarded as the uncertainty map in our framework. In PAN-sharpening tasks, the regions with high uncertainty, such as complex textures and edges, are visually more significant than smooth areas. By prioritizing these regions, uncertainty-aware models can handle complex image details more effectively, potentially leading to improved PAN-sharpening performance in our U-Know-DiffPAN framework.

D. Additional Results

Table 2 presents the results of our proposed models, FSA-T and FSA-S, compared with all baseline models on the WV3,

WV3	Reduced-Resolution						Full-Resolution		
Model	**PSNR↑**	**SSIM↑**	**SAM↓**	**ERGAS↓**	**SCC↑**	**Q8↑**	**D_λ↓**	**D_s↓**	**HQNR↑**
PanNet[17]	36.148 ± 1.958	0.966 ± 0.011	3.402 ± 0.672	2.538 ± 0.597	0.979 ± 0.006	0.913 ± 0.087	0.035 ± 0.014	0.049 ± 0.019	0.918 ± 0.031
MSDCNN[18]	36.329 ± 1.748	0.967 ± 0.010	3.300 ± 0.654	2.489 ± 0.620	0.979 ± 0.007	0.914 ± 0.087	0.028 ± 0.013	0.050 ± 0.020	0.924 ± 0.030
FusionNet[15]	36.569 ± 1.666	0.968 ± 0.009	3.188 ± 0.628	2.428 ± 0.621	0.981 ± 0.007	0.916 ± 0.087	0.029 ± 0.011	0.053 ± 0.021	0.920 ± 0.030
LAGNet[5]	36.732 ± 1.723	0.970 ± 0.009	3.153 ± 0.608	2.380 ± 0.617	0.981 ± 0.007	0.916 ± 0.087	0.033 ± 0.012	0.055 ± 0.023	0.915 ± 0.033
S2DBPN[20]	37.216 ± 1.888	0.972 ± 0.009	3.019 ± 0.588	2.245 ± 0.541	0.985 ± 0.005	0.917 ± 0.091	0.025 ± 0.010	0.030 ± 0.010	0.946 ± 0.018
DCPNet[21]	37.009 ± 1.735	0.972 ± 0.008	3.083 ± 0.537	2.301 ± 0.569	0.984 ± 0.005	0.915 ± 0.092	0.043 ± 0.018	0.036 ± 0.012	0.923 ± 0.027
CANConv[4]	37.441 ± 1.788	0.973 ± 0.008	2.927 ± 0.536	2.163 ± 0.481	0.985 ± 0.005	0.918 ± 0.082	0.020 ± 0.008	0.030 ± 0.008	0.951 ± 0.013
PanDiff[8]	37.029 ± 1.796	0.971 ± 0.008	3.058 ± 0.567	2.276 ± 0.545	0.984 ± 0.004	0.913 ± 0.084	**0.014 ± 0.005**	0.034 ± 0.005	0.952 ± 0.009
TMDiff[16]	37.477 ± 1.923	0.973 ± 0.008	2.885 ± 0.549	2.151 ± 0.458	0.986 ± 0.004	0.915 ± 0.086	0.018 ± 0.007	0.059 ± 0.009	0.924 ± 0.015
FSA-T	37.894 ± 1.820	**0.976 ± 0.007**	2.801 ± 0.517	2.055 ± 0.463	0.987 ± 0.003	0.921 ± 0.083	**0.014 ± 0.005**	0.032 ± 0.003	0.954 ± 0.006
FSA-S	**37.930 ± 1.824**	**0.976 ± 0.007**	**2.797 ± 0.526**	**2.046 ± 0.454**	**0.988 ± 0.003**	**0.922 ± 0.083**	0.016 ± 0.006	**0.029 ± 0.003**	**0.955 ± 0.008**
QB	Reduced-Resolution						Full-Resolution		
Model	**PSNR↑**	**SSIM↑**	**SAM↓**	**ERGAS↓**	**SCC↑**	**Q4↑**	**D_λ↓**	**D_s↓**	**HQNR↑**
PanNet[17]	35.563 ± 1.930	0.939 ± 0.012	5.273 ± 0.946	4.856 ± 0.590	0.966 ± 0.015	0.911 ± 0.094	0.063 ± 0.019	0.092 ± 0.021	0.851 ± 0.035
MSDCNN[18]	37.040 ± 1.778	0.954 ± 0.007	4.828 ± 0.824	4.074 ± 0.244	0.977 ± 0.010	0.925 ± 0.098	0.058 ± 0.014	0.058 ± 0.027	0.888 ± 0.037
FusionNet[15]	36.821 ± 1.765	0.952 ± 0.007	4.892 ± 0.822	4.183 ± 0.266	0.975 ± 0.011	0.923 ± 0.100	0.074 ± 0.022	0.079 ± 0.025	0.853 ± 0.041
LAGNet[5]	37.565 ± 1.721	0.958 ± 0.006	4.682 ± 0.785	3.845 ± 0.323	0.980 ± 0.009	0.930 ± 0.095	0.075 ± 0.019	0.035 ± 0.009	0.892 ± 0.024
S2DBPN[20]	37.314 ± 1.782	0.956 ± 0.006	4.849 ± 0.822	3.956 ± 0.291	0.980 ± 0.008	0.928 ± 0.093	0.059 ± 0.026	0.036 ± 0.023	0.908 ± 0.044
DCPNet[21]	38.079 ± 1.454	0.963 ± 0.004	4.420 ± 0.710	3.618 ± 0.313	0.983 ± 0.010	0.935 ± 0.095	0.051 ± 0.017	0.073 ± 0.013	0.880 ± 0.013
CANConv[4]	37.795 ± 1.801	0.960 ± 0.006	4.554 ± 0.788	3.740 ± 0.304	0.982 ± 0.007	0.935 ± 0.087	0.039 ± 0.012	0.070 ± 0.017	0.893 ± 0.010
PanDiff[8]	37.842 ± 1.721	0.959 ± 0.006	4.611 ± 0.768	3.723 ± 0.280	0.982 ± 0.007	0.935 ± 0.084	**0.028 ± 0.011**	0.055 ± 0.012	0.919 ± 0.010
TMDiff[16]	37.642 ± 1.831	0.958 ± 0.006	4.627 ± 0.814	3.804 ± 0.279	0.981 ± 0.008	0.930 ± 0.096	0.034 ± 0.016	0.068 ± 0.012	0.901 ± 0.011
FSA-T	38.343 ± 1.718	**0.964 ± 0.005**	4.349 ± 0.723	3.502 ± 0.272	**0.985 ± 0.007**	**0.938 ± 0.089**	0.036 ± 0.018	**0.031 ± 0.014**	**0.934 ± 0.029**
FSA-S	**38.361 ± 1.709**	**0.964 ± 0.005**	**4.337 ± 0.733**	**3.500 ± 0.272**	0.984 ± 0.007	0.938 ± 0.090	0.035 ± 0.011	0.035 ± 0.021	0.931 ± 0.029
GF2	Reduced-Resolution						Full-Resolution		
Model	**PSNR↑**	**SSIM↑**	**SAM↓**	**ERGAS↓**	**SCC↑**	**Q4↑**	**D_λ↓**	**D_s↓**	**HQNR↑**
PanNet[17]	39.197 ± 2.009	0.959 ± 0.011	1.050 ± 0.209	1.038 ± 0.214	0.975 ± 0.006	0.963 ± 0.009	0.020 ± 0.012	0.052 ± 0.009	0.929 ± 0.013
MSDCNN[18]	40.730 ± 1.564	0.971 ± 0.006	0.946 ± 0.166	0.862 ± 0.141	0.983 ± 0.003	0.972 ± 0.009	0.026 ± 0.014	0.079 ± 0.011	0.898 ± 0.016
FusionNet[15]	39.866 ± 1.955	0.966 ± 0.009	0.971 ± 0.195	0.960 ± 0.193	0.980 ± 0.005	0.967 ± 0.008	0.034 ± 0.013	0.105 ± 0.013	0.865 ± 0.018
LAGNet[5]	41.147 ± 1.384	0.974 ± 0.005	0.886 ± 0.140	0.816 ± 0.121	0.985 ± 0.003	0.974 ± 0.009	0.030 ± 0.014	0.078 ± 0.013	0.895 ± 0.021
S2DBPN[20]	42.686 ± 1.676	0.980 ± 0.005	0.772 ± 0.149	0.686 ± 0.125	0.990 ± 0.002	0.981 ± 0.007	0.020 ± 0.012	0.046 ± 0.007	0.935 ± 0.011
DCPNet[21]	42.312 ± 1.682	0.979 ± 0.005	0.806 ± 0.153	0.724 ± 0.138	0.988 ± 0.003	0.980 ± 0.007	0.024 ± 0.022	**0.024 ± 0.008**	**0.953 ± 0.019**
CANConv[4]	43.166 ± 1.705	0.982 ± 0.004	0.722 ± 0.138	0.653 ± 0.124	0.991 ± 0.002	0.983 ± 0.006	0.019 ± 0.010	0.063 ± 0.009	0.919 ± 0.011
PanDiff[8]	42.827 ± 1.462	0.980 ± 0.005	0.767 ± 0.134	0.674 ± 0.110	0.990 ± 0.002	0.981 ± 0.007	0.020 ± 0.014	0.045 ± 0.009	0.936 ± 0.011
TMDiff[16]	41.896 ± 1.765	0.977 ± 0.005	0.764 ± 0.155	0.754 ± 0.143	0.988 ± 0.003	0.979 ± 0.007	0.029 ± 0.011	0.030 ± 0.010	0.942 ± 0.016
FSA-T	**44.757 ± 1.359**	**0.988 ± 0.003**	**0.603 ± 0.102**	**0.537 ± 0.077**	**0.994 ± 0.001**	**0.988 ± 0.006**	**0.017 ± 0.010**	0.030 ± 0.008	**0.953 ± 0.013**
FSA-S	44.585 ± 1.521	0.986 ± 0.003	0.624 ± 0.109	0.548 ± 0.091	0.993 ± 0.001	0.987 ± 0.007	0.018 ± 0.011	0.037 ± 0.007	0.944 ± 0.012

Table 2. Additional PAN-sharpening results by our U-Know-DiffPAN and other SOTA methods for the WV3, QB, and GF2 dataset. The best (second best) performance in each block is highlighted in bold **red** (underlined in blue).

QB, and GF2 datasets. The evaluation includes Reduced-Resolution (RR), Full-Resolution (FR), and standard deviation. Fig. 5 to 10 illustrate the qualitative results for the WV3, QB, and GF2 datasets in both Reduced-Resolution and Full-Resolution settings. For the RR results, we visualize the RGB outputs, along with the difference between the output HRMS $\hat{\mathbf{I}}_{\text{MS}}^{\text{HR}}$ and the ground truth $\mathbf{I}_{\text{MS}}^{\text{HR}}$ using error maps and their corresponding mean absolute error (MAE) values. For the FR results, we showcase visual comparisons with the latest state-of-the-art methods. From the visual comparisons, we observe that our U-Know-DiffPAN framework significantly enhances restoration qualities, particularly in the regions with high-frequency contents, high uncertainty, and complex textures, such as edges and small objects. In RR scenarios, these regions exhibit a closer resemblance to their ground truths, compared to the previous methods. Even in FR scenarios where ground truth is unavailable, our U-Know-DiffPAN framework produces more detailed and robust results, demonstrating superior structural and spectral fidelity in high-uncertainty regions. This highlights the capability of our U-Know-DiffPAN to outperform state-of-the-art models in both qualitative and quantitative aspects.

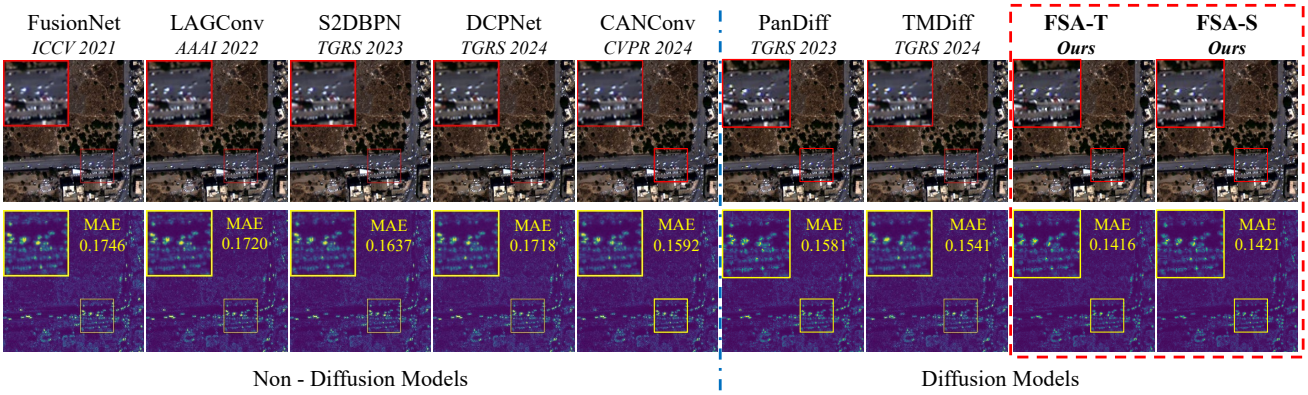

Figure 5. PAN-sharpening results for the WV3 dataset under reduced resolution (RR) scenarios. The first row depicts the output HRMS images, while the second row highlights the Error Map between the output HRMS and the corresponding ground truth images.The Mean Absolute Error (MAE) values are presented alongside the Error Map. Zoom in for better visualization.

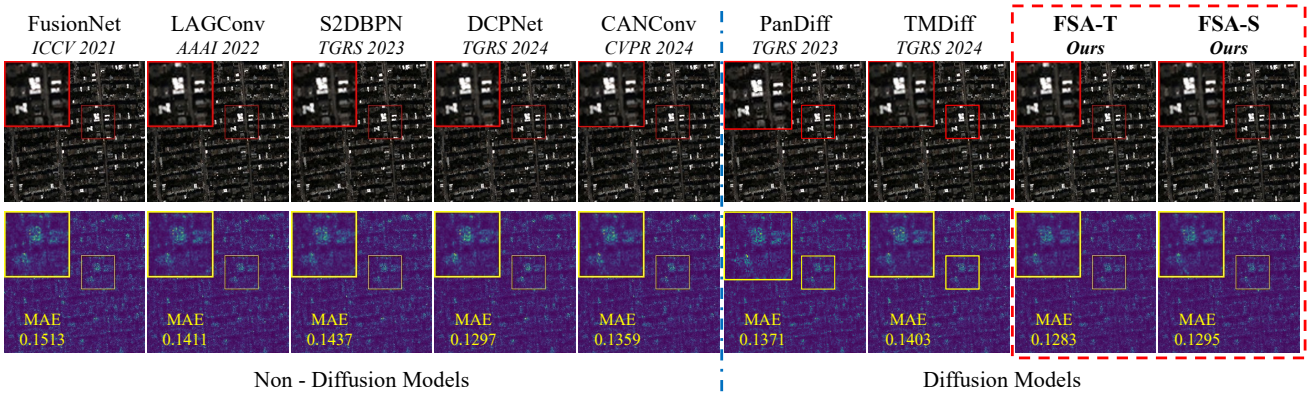

Figure 6. PAN-sharpening results for the QB dataset under reduced resolution (RR) scenarios. The first row depicts the output HRMS images, while the second row highlights the Error Map between the output HRMS and the corresponding ground truth images. The Mean Absolute Error (MAE) values are presented alongside the Error Map. Zoom in for better visualization.

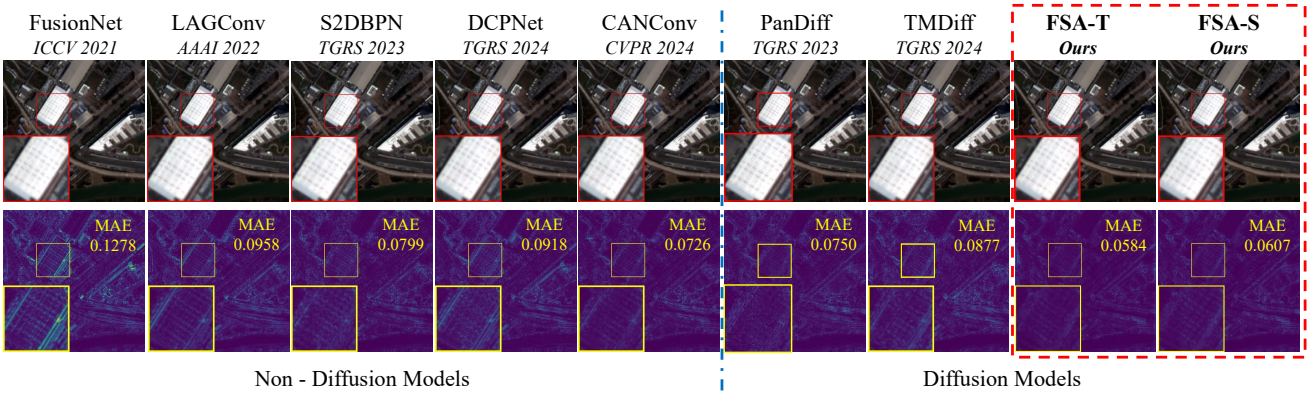

Figure 7. PAN-sharpening results for the GF2 dataset under reduced resolution (RR) scenarios. The first row depicts the output HRMS images, while the second row highlights the Error Map between the output HRMS and the corresponding ground truth images. The Mean Absolute Error (MAE) values are presented alongside the Error Map. Zoom in for better visualization.

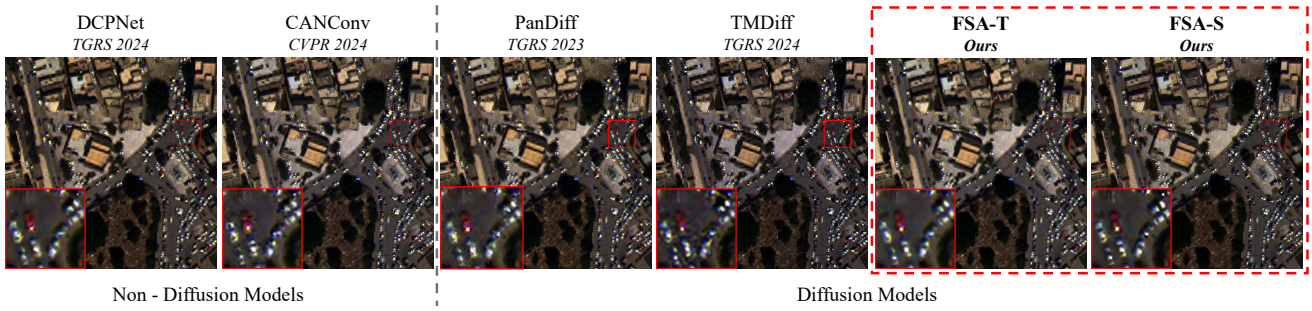

Figure 8. PAN-sharpening results for the WV3 dataset under full resolution (FR) scenarios. The first row depicts the output HRMS images. Zoom in for better visualization.

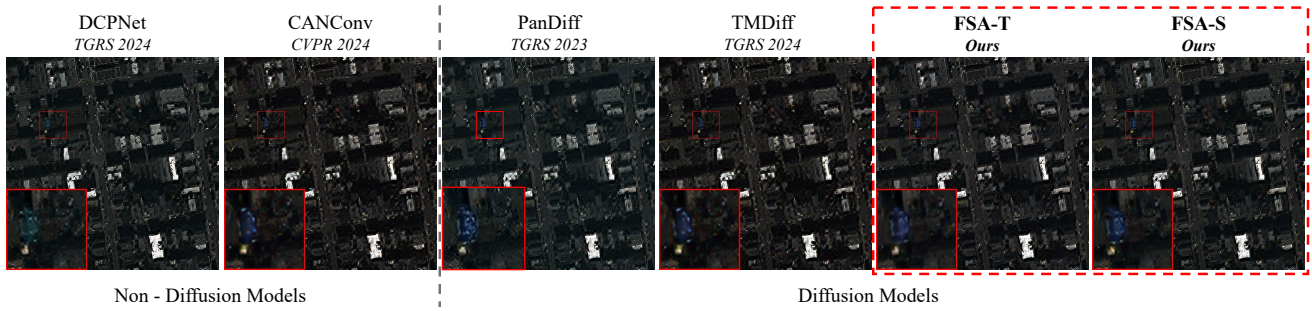

Figure 9. PAN-sharpening results for the QB dataset under full resolution (FR) scenarios. The first row depicts the output HRMS images. Zoom in for better visualization.

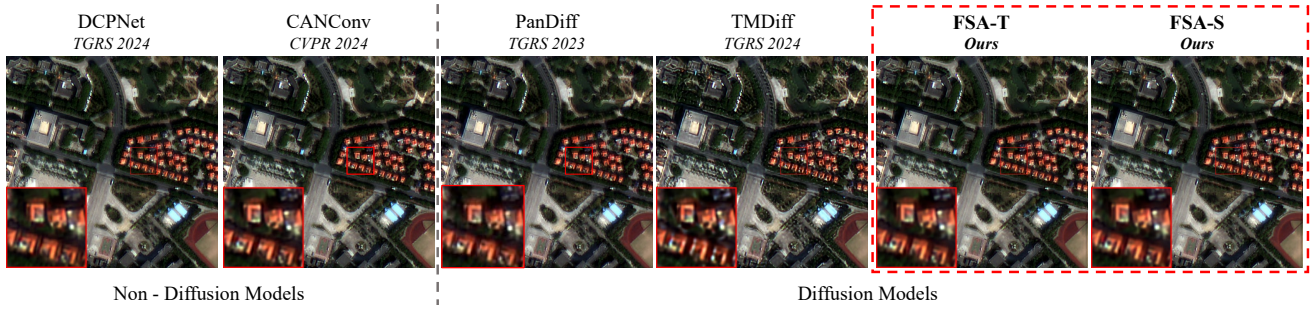

Figure 10. PAN-sharpening results for the GF2 dataset under full resolution (FR) scenarios. The first row depicts the output HRMS images. Zoom in for better visualization.

References

[1] Bruno Aiazzi, Luciano Alparone, Stefano Baronti, Andrea Garzelli, and Massimo Selva. Mtf-tailored multiscale fusion of high-resolution ms and pan imagery. *Photogrammetric Engineering & Remote Sensing*, 72(5):591–596, 2006. 2

[2] Arsenii Ashukha, Alexander Lyzhov, Dmitry Molchanov, and Dmitry Vetrov. Pitfalls of in-domain uncertainty estimation and ensembling in deep learning. *arXiv preprint arXiv:2002.06470*, 2020. 3

[3] Jie Chang, Zhonghao Lan, Changmao Cheng, and Yichen Wei. Data uncertainty learning in face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5710–5719, 2020. 3

[4] Yule Duan, Xiao Wu, Haoyu Deng, and Liang-Jian Deng. Content-adaptive non-local convolution for remote sensing pansharpening. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27738–27747, 2024. 4

[5] Zi-Rong Jin, Tian-Jing Zhang, Tai-Xiang Jiang, Gemine Vivone, and Liang-Jian Deng. Lagconv: Local-context adaptive convolution kernels with global harmonic bias for pansharpening. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1113–1121, 2022. 4

[6] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017. 3

[7] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017. 3

[8] Qingyan Meng, Wenxu Shi, Sijia Li, and Linlin Zhang. Pandiff: A novel pansharpening method based on denoising diffusion probabilistic model. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–17, 2023. 4

[9] Qian Ning, Weisheng Dong, Xin Li, Jinjian Wu, and Guangming Shi. Uncertainty-driven loss for single image super-resolution. *Advances in Neural Information Processing Systems*, 34:16398–16409, 2021. 3

[10] Xavier Otazu, María González-Audícana, Octavi Fors, and Jorge Núñez. Introduction of sensor spectral response into image fusion methods. application to wavelet-based methods. *IEEE Transactions on Geoscience and Remote Sensing*, 43(10):2376–2385, 2005. 2

[11] Hao Phung, Quan Dao, and Anh Tran. Wavelet diffusion models are fast and scalable image generators. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10199–10208, 2023. 1

[12] Maximilian Seitzer, Arash Tavakoli, Dimitrije Antic, and Georg Martius. On the pitfalls of heteroscedastic uncertainty estimation with probabilistic neural networks. *arXiv preprint arXiv:2203.09168*, 2022. 3

[13] Shuyao Shang, Zhengyang Shan, Guangxing Liu, LunQian Wang, XingHua Wang, Zekai Zhang, and Jinglin Zhang. Resdiff: Combining cnn and diffusion model for image super-resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8975–8983, 2024. 1

[14] Joost Van Amersfoort, Lewis Smith, Yee Whye Teh, and Yarin Gal. Uncertainty estimation using a single deep deterministic neural network. In *International conference on machine learning*, pages 9690–9700. PMLR, 2020. 3

[15] Xiao Wu, Ting-Zhu Huang, Liang-Jian Deng, and Tian-Jing Zhang. Dynamic cross feature fusion for remote sensing pansharpening. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14687–14696, 2021. 4

[16] Yinghui Xing, Litao Qu, Shizhou Zhang, Jiapeng Feng, Xiuwei Zhang, and Yanning Zhang. Empower generalizability for pansharpening through text-modulated diffusion model. *IEEE Transactions on Geoscience and Remote Sensing*, 2024. 4

[17] Junfeng Yang, Xueyang Fu, Yuwen Hu, Yue Huang, Xinghao Ding, and John Paisley. Pannet: A deep network architecture for pan-sharpening. In *Proceedings of the IEEE international conference on computer vision*, pages 5449–5457, 2017. 4

[18] Qiangqiang Yuan, Yancong Wei, Xiangchao Meng, Huanfeng Shen, and Liangpei Zhang. A multiscale and multidepth convolutional neural network for remote sensing imagery pan-sharpening. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 11(3):978–989, 2018. 4

[19] Jie Zhang, Xuanhua He, Keyu Yan, Ke Cao, Rui Li, Chengjun Xie, Man Zhou, and Danfeng Hong. Pansharpening with wavelet-enhanced high-frequency information. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–14, 2024. 1

[20] Kai Zhang, Anfei Wang, Feng Zhang, Wenbo Wan, Jiande Sun, and Lorenzo Bruzzone. Spatial-spectral dual back-projection network for pansharpening. *IEEE Transactions on Geoscience and Remote Sensing*, 2023. 4

[21] Yafei Zhang, Xuji Yang, Huafeng Li, Minghong Xie, and Zhengtao Yu. Dcpnet: A dual-task collaborative promotion network for pansharpening. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–16, 2024. 4

[22] Jie Zhou, Daniel L Civco, and John A Silander. A wavelet transform method to merge landsat tm and spot panchromatic data. *International journal of remote sensing*, 19(4):743–757, 1998. 2