VolFormer: Explore More Comprehensive Cube Interaction for Hyperspectral Image Restoration and Beyond

Supplementary Material

6. Data Alignment

We define two same restoration tasks (image SR or image denoising) on the HSI dataset and RGBI dataset as Γ_{HSI}, Γ_{RGBI}, and desire to improve the learning of model Γ_{HSI} by using the knowledge from Γ_{RGBI}. Given an HSI dataset $\Omega_{HSI} = \left\{x^i_{HSI}, X^i_{HSI}\right\}_{i=1}^{N_{HSI}}$ and RGBI dataset $\Omega_{RGBI} = \left\{x^i_{RGBI}, X^i_{RGBI}\right\}_{i=1}^{N_{RGBI}}$, where $x_{HSI} \in \mathbb{R}^{h\times w\times D}$ represents the degraded HSI, $X_{HSI} \in \mathbb{R}^{H\times W\times D}$ represents the high-quality HSI counterpart. Similarly, $x_{RGBI} \in \mathbb{R}^{h\times w\times 3}$ is the degraded RGB image and $X_{RGBI} \in \mathbb{R}^{H\times W\times 3}$ is the high-quality counterpart. h, w, H and W denote the width and height of the degraded image and desired image, respectively, and D is the number of HSI bands. For HSI SR, we have $H = \lambda h$, $W = \lambda w$, and λ is the scaling factor. For HSI denoising, λ is set to 1, and N_{HSI} and N_{RGBI} are the numbers of HSI and RGBI samples, respectively. We attempt to exploit the advantage of the RGBI dataset since it provides numerous high-quality samples. Thus, we have $N_{RGBI} = vN_{HSI}$ and $v \geq 1$.

HSIs provide tens to hundreds of spectral bands, and it is time-consuming to learn all of the band knowledge at once. Inspired by [19], we divide each HSI input into samples with overlapping groups of bands. This strategy can retain the spectral correlation among neighboring bands and reduce the number of parameters. Another purpose of using a grouping strategy is that it offers the possibility to train our restoration task using Transformer. More specifically, we divide the D bands of the HSIs into groups of S bands. For RGBI samples, we increase the channels to S via the spectral band interpolation strategy [23]. This strategy follows a distance rule in that the correlation between neighboring HSI bands should be higher than that between distant bands. Therefore the generated RGBI dataset $\bar{\Omega}_{RGBI} = \left\{\bar{x}^i_{RGBI}, \bar{X}^i_{RGBI}\right\}_{i=1}^{N_{RGBI}}$ and HSI dataset $\bar{\Omega}_{HSI} = \left\{\bar{x}^i_{HSI}, \bar{X}^i_{HSI}\right\}_{i=1}^{N_{HSI}}$ have similar formats, where $\bar{x}_{RGBI} \in \mathbb{R}^{h\times w\times S}$, $\bar{X}_{RGBI} \in \mathbb{R}^{H\times W\times S}$ and $\bar{x}_{HSI} \in \mathbb{R}^{h\times w\times S}$, $\bar{X}_{HSI} \in \mathbb{R}^{H\times W\times S}$.

7. Loss Function

We combine the L_1 loss and the spatial-spectral total variation (SSTV) loss [19] to optimize the VolFormer parameters. The L_1 loss computes the mean absolute error (MAE) between the restored images and the ground truth. It is beneficial to penalize pixel errors and ensure better convergence throughout the training process,

$$L_1(\Theta) = \frac{1}{N}\sum_{n=1}^{N}\|X^n - I^n\|, \tag{11}$$

where I^n and X^n are the n-th reconstructed result and ground truth, respectively. N denotes the number of images in one training batch, and Θ refers to the VolFormer parameters. the SSTV loss is designed to smooth the reconstructed result in both the spatial and spectral dimensions,

$$L_{SSTV}(\Theta) = \frac{1}{N}\sum_{n=1}^{N}\left(\|\bigtriangledown_h I^n\|_1 + \|\bigtriangledown_w I^n\|_1 + \|\bigtriangledown_c I^n\|_1\right), \tag{12}$$

where $\bigtriangledown_h$, $\bigtriangledown_w$, and $\bigtriangledown_c$ represent the horizontal, vertical, and spectral direction gradients of the reconstruction result, respectively. The final objective loss function for the our VolFormer is the sum of the L_1 loss and SSTV loss.

$$L(\Theta) = L_1 + L_{SSTV}. \tag{13}$$

The overall loss of our restoration task contains the loss of the RGBI restoration task and the HSI restoration task.

$$L_{Total}(\Theta) = L^{HSI}\left(X^n_{HSI}, I^n_{HSI}\right) + L^{RGBI}\left(X^n_{RGBI}, I^n_{RGBI}\right). \tag{14}$$

8. Super-Resolution Experiments

8.1. Datasets

CAVE [54]. This dataset contains 32 images with a spatial resolution of 512 × 512 and 31 bands collected by a tunable filter and a cooled CCD camera ranging from 400 nm to 700 nm. We use 20 images for training and 12 images for testing in our experiment.

Pavia [12]. The Pavia Dataset is a remote sensing hyperspectral dataset acquired by the ROSIS sensor about Pavia, northern Italy. The images contain 102 spectral bands with a spatial resolution of 1096×1096. The hyperspectral image contains 9 land-over categories. We crop non-overlapped patches with a spatial resolution of 128×714. For each image in Pavia, three patches are used for testing and the rest for training.

Chikusei [55]. The Chikusei dataset was captured by a Headwall Hyperspec VNIR-C imaging sensor in an urban area in Chikusei, Ibaraki, Japan. This dataset contains a remote sensing hyperspectral image with a spatial resolution of 2517×2335 and 128 spectral bands. We first crop the center region to obtain a sub-image with $2048 \times 2048 \times 128$ pixels. Then the sub-image is further divided into training data. The remaining region is used to crop the overlap patches for training. Specifically, four nonoverlapping hyperspectral images with $512 \times 512 \times 128$ pixels are generated from the top region of the sub-image to form the testing data.

DIV2K [1]. DIV2K is adopted for the auxiliary RGBI SR task. There are 1000 high-quality images with 2k resolution. We use 800 samples for the training set and the remaining 200 for the test set.

8.2. Experimental Parameters.

During training, for upsampling factor ×4, we let the extracted patches be 64×64 pixels with 32 overlapping pixels; for upsampling factor ×8, we let the extracted patches be 128×128 pixels with 64 overlapping pixels. The corresponding LR images are generated by Bicubic downsampling with 16×16 pixels. **In addition, we have performed the super-resolution experiment based on Blur-downscale (BD) degradation mode in Appendix 8.4.** The training samples of DIV2K are approximately 24, 1, and 8 times larger than the CAVE, Pavia, and Chikusei datasets, respectively. We use the Adam optimizer and the initial learning rate is set to 10^{-4}. The batch size is set to 12 and the number of epochs is set to 20. We use the PyTorch framework to implement our models and all variants. For the SR task, the TB number, TL number, window size and attention head number are generally set to 8, 6, 8, and 6, respectively.

8.3. Super-resolution Experimental Results

In this section, we provide a more comprehensive visual analysis for hyperspectral image super-resolution. Figure 6 and Figure 7 denote the super-resolution result on CAVE with the scaling factor of 4 and 8. Figure 8 and Figure 7 represent the super-resolution result on Chikusei with the scaling factor of 4 and 8. Figure 10 shows the visual results on Pavia with the scaling factor of 4. Compared with other methods, the reconstructed results generated by the proposed VolFormer have smoother borders and sharper textures.

8.4. Results with the Blur-downscale (BD) Degradation Model

We apply our method to the blur-down (BD) degradation model for super-resolution image reconstruction, which has been commonly used recently [61, 64]. The quantitative results of the proposed VolFormer and other comparative methods on the CAVE dataset are shown in Table 7. Our VolFormer is robust and maintains significant performance with the BD degradation model.

We also conducted experiments with the BD model on the Chikusei dataset. The quantitative results are provided in Table 8. Similarly, our VolFormer obtains notable quantitative gains over eight state-of-the-art methods on the Chikusei dataset with the BD degradation model.

9. Denoise Experiments

9.1. Datasets

ICVL [3]. The ICVL dataset contains 201 images with a spatial resolution of 1392×1300 over 31 spectral bands. We randomly selected 100 images as training data and 50 images for testing.

HSIDwRD [63]. The HSIDwRD dataset is the first real-world dataset for training and testing HSI denoising model. There are 62 real-world HSIs collected by the SOC710-VP hyperspectral camera, each with a size of $696 \times 520 \times 34$. All scenes in the dataset are static. High-quality inference images are captured by adjusting the aperture, focus, and exposure time. The noisy counterparts are captured with the same aperture, focus and 1/50 exposure time as the reference images. We selected paired noisy and reference images of 45 scenes to form the training set and the remaining 17 scenes were chosen for the test set.

RENOIR [2]. RENOIR is a dataset for real noise image denoising tasks that contains 40 scenes with a spatial resolution of 3684×2760 collected by the Canon S90, 40 scenes collected at 5202×3465 spatial resolution collected by the Canon Rebel T3i, and 40 scenes collected at 4208×3120 by the Xiaomi Mi3. RENOIR is adopted for the auxiliary RGBI real-world denoise task. We selected image form 30 scenes to form the testing set and the remaining scenes were used for training. All of the experiments presented in this section are performed on an NVIDIA GeForce RTX 3090 GPU.

9.2. Denoise Experimental Settings

During training, all of the training samples were cropped into patches of 64×64 pixels. For the denoising experiment on the synthetic dataset, additive Gaussian white noise was added to each input HSI with four different noise levels, including 30, 50, 70, and random strengths ranging from 30 to 70. Besides, we add the mixture noise to the synthetic dataset. Each band is randomly corrupted by at least two kinds of Gaussian noise, impulse noise, deadlines and stripes. For synthetic experiments, the RGBI denoising task was performed on DIV2K by adding synthetic noise. The training samples of DIV2K are approximately 10 times larger than those of ICVL.

For real HSI denoising, the auxiliary RGBI denoising task was performed on RENOIR. The training samples of RENOIR are approximately 20 times larger than those of HSIDwRD. The batch size was set to 12 and the number of

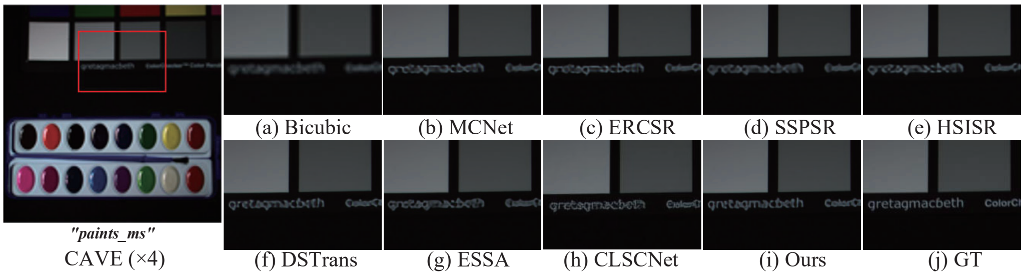


Figure 6. Visual comparison for HSI SR on the representative test image *paints_ms* from CAVE dataset with spectral bands 23-15-7 as R-G-B with the scale factor 4.

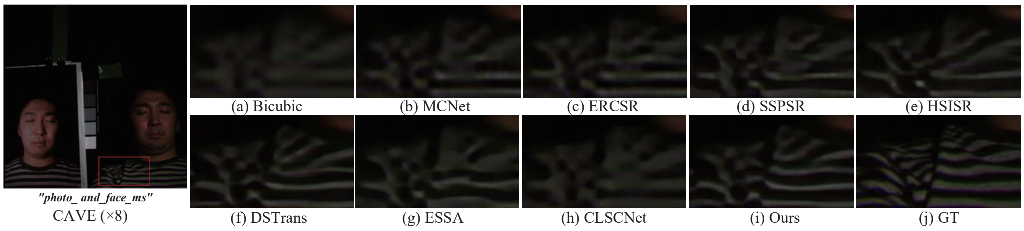


Figure 7. Visual comparison for HSI SR on the representative test image *photo_and_face_ms* from CAVE dataset with spectral bands 23-15-7 as R-G-B with the scale factor 8.

Scale	Method	SAM ↓	CC ↑	ERGAS ↓	RMSE ↓	MPSNR ↑	MSSIM ↑
×4	Blur	3.983	0.9852	5.632	0.0228	34.050	0.9251
	MCNet[26]	3.748	0.9826	4.226	0.0160	37.032	0.9488
	ERCSR[27]	3.444	0.9840	3.632	0.0153	37.484	0.9510
	SSPSR[19]	3.442	0.9929	3.680	0.0152	37.944	0.9556
	HSISR[23]	3.452	0.9946	3.474	0.0143	38.377	0.9591
	DSTrans [56]	3.161	0.9948	3.220	0.0125	39.793	0.9647
	ESSA [62]	3.191	0.9942	3.336	0.0132	39.061	0.9598
	CLSCNet [53]	3.201	0.9931	3.4126	0.0137	39.256	0.9595
	Ours	**3.117**	**0.9951**	**2.941**	**0.0120**	**39.983**	**0.9658**

Table 7. Quantitative evaluation with BD degradation model on CAVE dataset of state-of-the-art SR methods by SAM, CC, ERGAS, RMSE, MPSNR and MSSIM for scale factors 4. (Best results are shown in bold)

Scale	Method	SAM ↓	CC ↑	ERGAS ↓	RMSE ↓	MPSNR ↑	MSSIM ↑
×4	Blur	3.568	0.9139	7.128	0.0164	37.175	0.8858
	MCNet[26]	3.266	0.9250	6.421	0.0145	37.903	0.9173
	ERCSR[27]	3.414	0.9266	6.273	0.0140	38.202	0.9126
	SSPSR[19]	2.815	0.9448	5.662	0.0128	39.266	0.9303
	HSISR[23]	2.789	0.9452	5.602	0.0128	39.326	0.9302
	DSTrans [56]	2.721	0.9485	5.466	0.0125	39.472	0.9343
	ESSA [62]	2.680	0.9497	5.410	0.0124	39.614	0.9355
	CLSCNet [53]	2.850	0.9481	5.493	0.0127	39.428	0.9317
	Ours	**2.595**	**0.9508**	**5.316**	**0.0121**	**39.769**	**0.9372**

Table 8. Quantitative evaluation with BD degradation model on Chikusei dataset of state-of-the-art SR methods by SAM, CC, ERGAS, RMSE, MPSNR and MSSIM for scale factors 4. (Best results are shown in bold)

epochs was set to 50. We use the PyTorch framework to implement our models and all variants. For the denoising task, the TB number, TL number, window size and attention head number were generally set to 6, 6, 8 and 6, respectively.

9.3. Denoise Experimental Results

To visually demonstrate the superiority of our method, we output the qualitative results and performed a local magnifi-

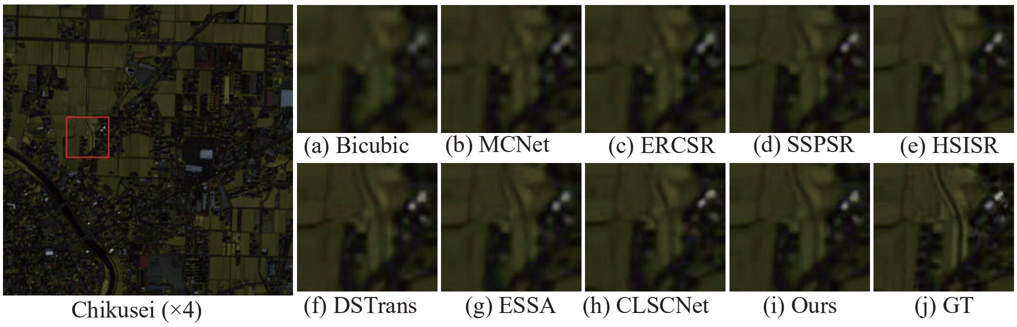


Figure 8. Visual comparison for HSI SR on the representative test image from Chikusei dataset with spectral bands 97-76-63 as R-G-B with the scale factor 4.

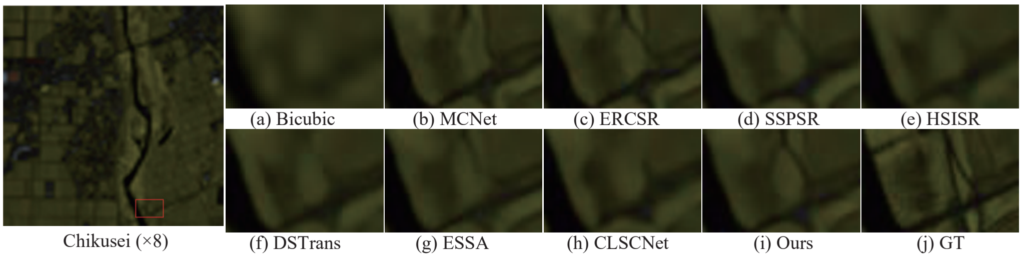


Figure 9. Visual comparison for HSI SR on the representative test image from Chikusei dataset with spectral bands 97-76-63 as R-G-B with the scale factor 8.

cation comparison. The visual results are presented in Fig. 11 and Fig. 12, "Noisy" is obtained by adding the additive Gaussian white noise with noise levels of 50. It is evident that our method is superior to the other methods since it restores more details and achieves pleasing results. The denoise result for real noise are presented in Fig. 13 and Fig. 14 . Our method eliminates real-world noise and generates the clearest text signal and boundaries.

10. Classification Experiments

In this section, we implement experiments on three benchmark datasets to illustrate the effectiveness of the proposed VolSA in classification task. The 16-class Indian Pines (IP) dataset, the 15-class University of Houston (HU2013) dataset, and the 16-class Salinas Valley (SV) dataset are utilized to validate the classification performance. We use VolSA as the key component to form a classification network. Our network is built based on the Transformer-based classification network, *i.n.,* Hybrid Former [36]. We use VolSA to replace the proposed spatial-spectral attention in HybridFormer.

The classifiers used in the comparison are eight state-of-the-art methods (SVM [42], DBMA [34], DBDA [28], SSRN [66], SSFTT [40], BS2T [39], CS2DT [51] and HybridFormer [36]. All of the experiments presented in this section are performed on RTX3090.

10.1. Dataset

The IP dataset was gathered by AVIRIS sensor over the Indian Pines test site in North-western Indiana and consisted of 145×145 pixels and 220 bands. Removing 20 bands covering the region of water absorption, 200 bands are retained for the classification task.

The SV dataset was collected by the 224-band AVIRIS sensor over Salinas Valley, California, characterized by high spatial resolution (3.7-meter pixels). Covering an area of 512 lines by 217 samples, similar to the Indian Pines scene. This image was available only as at-sensor radiance data, featuring vegetables, bare soils, and vineyard fields. The Salinas ground truth contains 16 classes.

The HU2013 dataset was published in the 2013 IEEE Geoscience and Remote Sensing Society data fusion contest, which is a relatively tricky benchmark dataset. The dataset contains 15 land-cover types with 349×1905 pixels and 144 spectral bands ranging from 380 to 1050 nm.

In the experiment, we initially preprocessed the datasets. Subsequently, we select 5.5%, 0.5% and 3% of the samples for training in IP dataset and SV dataset and HU2013 dataset, with the remaining samples reserved for performance evaluation.

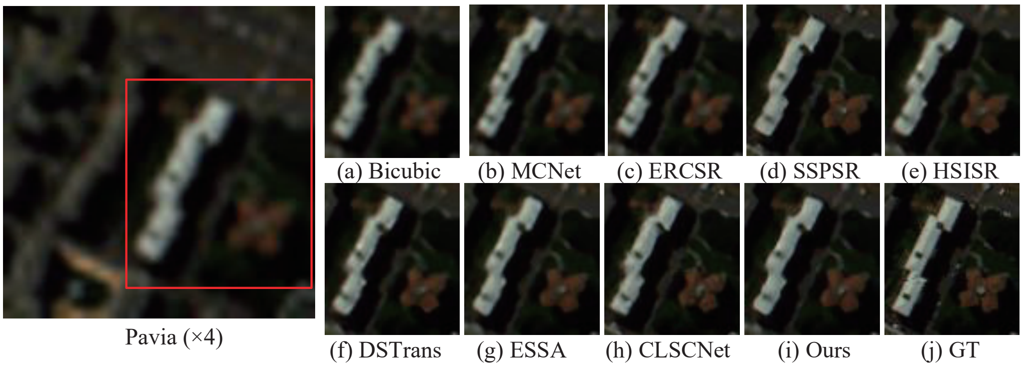


Figure 10. Visual comparison for HSI SR on the representative test image from Pavia dataset with spectral bands 60-35-10 as R-G-B with the scale factor 4.

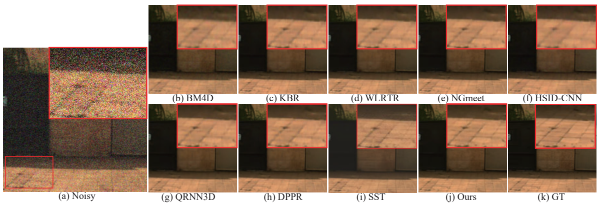


Figure 11. Denoising results and error maps of the representative image *eve_0331-1549* under Gaussian noise with spectral bands 23-15-7 as R-G-B.

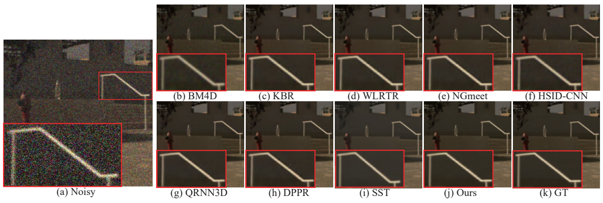


Figure 12. Denoising results and error maps of the representative image *eve_0331-1606* under Gaussian noise with spectral bands 23-15-7 as R-G-B.

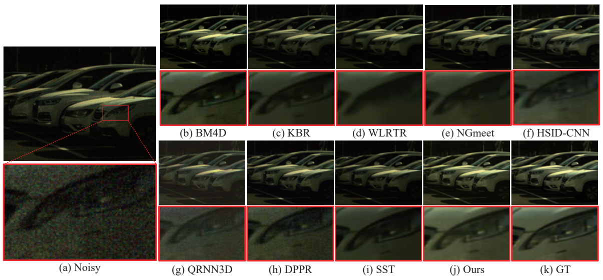


Figure 13. Denoising results of *image 46* under real-world noise with spectral bands 23-15-7 as R-G-B.

10.2. Metrics

The classification performances of the proposed method and the state-of-the-art methods are evaluated by the overall accuracy (OA), average accuracy (AA), kappa coefficient (Kappa) and the accuracy per class. All the quantitative results are presened in the form of mean value ± standard deviation by choosing random samples and calculating the result of ten times experiments.

10.3. Classification Experiment Results

Table 9 - Table 11 demonstrate the quantitative evaluation results for IP, SV and HU2013 datasets. The best results are highlighted in bold. The accuracy per class, OA, AA and Kappa are respectively listed. It can be seen that our method keeps the highest classification accuracy among all the methods.

11. Limitation

Following the preview work [56], our VolFormer is also trained based on HSI samples and RGBI samples together. Numerous training samples ensure the stable learning process of Transformer parameter distribution and achieve significant performance. Especially for the natural HSI dataset (*i.e.,* CAVE), adding the processed RGB samples brought significant gains. For example, HSISR obtains 0.7dB gains in terms of PSNR than SSPSR on CAVE dataset. Compared to those methods without using RGB samples, DSTrans and our VolFormer also achieve more than 1dB gains. Moreover, as shown in Fig. 15(a), the quantitative results gradually improved as the sample number increased. However, in the remote sensing HSI dataset (i.e., Chikusei and Pavia), the collaborative training strategy barely replicates the success of CAVE. There is a more distinctive domain gap between the RGBI samples from DIV2K and the HSI samples from the remote sensing HSI dataset. For the Chikusei dataset, the numerous RGBI samples bring limited perfor-

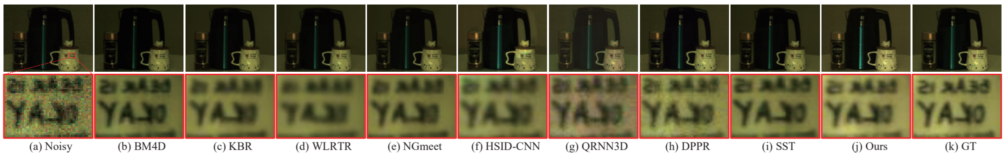


Figure 14. Denoising results of *image 48* under real-world noise with spectral bands 23-15-7 as R-G-B.

Class	SVM	DBMA	DBDA	SSRN	SSFTT	BS2T	CS2DT	HybridFormer	Ours
1	99.42 ± 0.75	100.0 ± 0.00	100.0 ± 0.00	99.36 ± 0.68	99.99 ± 0.01	95.53 ± 5.53	100.0 ± 0.00	98.60 ± 1.27	93.02 ± 5.45
2	98.79 ± 0.37	99.07 ± 0.44	99.53 ± 0.42	99.83 ± 0.11	100.0 ± 0.00	97.53 ± 4.65	99.65 ± 0.43	95.76±1.67	97.57 ± 1.38
3	87.98 ± 3.76	97.41 ± 1.28	97.79 ± 1.57	95.01 ± 4.98	99.99 ± 0.01	99.06 ± 0.84	99.89 ± 0.14	97.28±1.33	97.30± 2.23
4	97.54 ± 0.59	92.39 ± 1.28	96.23 ± 1.07	96.91 ± 2.17	97.33 ± 1.10	93.93 ± 1.85	93.50 ± 3.16	96.17 ± 1.95	99.38± 0.75
5	95.09 ± 3.14	99.51 ± 0.28	99.23 ± 0.80	98.18 ± 1.93	99.96 ± 0.05	90.51 ± 8.84	98.58 ± 0.63	97.11 ± 1.68	96.45± 3.54
6	99.89 ± 0.08	99.83 ± 0.23	100.0 ± 0.00	99.92 ± 0.10	99.75 ± 0.20	100.0 ± 0.00	99.21 ± 1.05	98.50 ± 0.62	99.30± 0.59
7	95.59 ± 2.67	93.89 ± 4.56	98.63 ± 1.14	99.51 ± 0.86	99.46 ± 0.38	94.51 ± 3.12	99.96 ± 0.10	95.53 ± 6.08	100.00± 0.00
8	71.66 ± 2.54	94.00 ± 2.26	87.92 ± 1.03	88.52 ± 4.54	95.87 ± 0.63	90.58 ± 3.38	90.98 ± 2.83	100.00± 0.00	100.00± 0.00
9	98.08 ± 1.17	99.71 ± 0.33	99.26 ± 0.22	95.63 ± 7.56	99.99 ± 0.01	94.01 ± 5.68	99.53 ± 0.35	83.16±15.07	94.74± 7.44
10	85.39 ± 3.46	85.46 ± 8.47	96.67 ± 2.69	97.93 ± 1.30	98.81 ± 0.43	86.39 ± 7.50	98.20 ± 1.43	96.61 ± 1.51	97.02± 2.44
11	86.97 ± 6.81	89.98 ± 8.01	96.87 ± 2.21	83.51 ± 25.72	99.36 ± 0.66	75.53 ± 38.22	98.69 ± 1.61	97.27 ± 0.98	97.41± 1.28
12	94.20 ± 4.00	99.75 ± 0.27	98.94 ± 1.48	97.13 ± 3.15	99.23 ± 0.83	98.65 ± 0.99	99.83 ± 0.12	94.30 ± 3.55	96.57± 2.22
13	93.43 ± 3.31	96.56 ± 1.56	99.88 ± 0.09	98.19 ± 2.47	92.94 ± 3.29	99.86 ± 0.12	98.92 ± 1.55	99.28 ± 0.86	99.69 ± 0.46
14	92.03 ± 5.43	99.52 ± 0.28	96.67 ± 0.60	93.85 ± 6.14	91.15 ± 4.26	87.12 ± 5.78	92.76 ± 6.33	99.30 ± 0.90	99.40 ± 0.19
15	71.02 ± 5.59	75.77 ± 16.40	94.24 ± 0.71	80.35 ± 7.61	89.61 ± 0.94	76.27 ± 1.56	87.32 ± 5.00	98.04 ± 2.36	98.63± 1.19
16	97.81 ± 1.19	97.47 ± 1.88	100.0 ± 0.00	99.42 ± 0.58	99.68 ± 0.22	98.78 ± 2.40	99.99 ± 0.01	96.36±4.91	95.23± 4.28
OA	86.97 ± 0.86	91.43 ± 4.81	95.78 ± 0.03	92.43 ± 3.86	97.20 ± 0.25	90.85 ± 1.30	95.61 ± 0.78	97.29 ± 0.60	**97.89± 0.36**
AA	91.55 ± 0.62	95.02 ± 1.23	97.62 ± 0.12	95.20 ± 2.79	97.69 ± 0.36	92.39 ± 3.58	97.31 ± 0.48	96.45 ± 1.34	**97.61± 0.30**
Kappa	85.45 ± 0.97	90.52 ± 5.28	95.29 ± 0.04	91.58 ± 4.30	96.88 ± 0.28	89.82 ± 1.45	95.11 ± 0.86	96.92 ± 0.69	**97.60± 0.41**

Table 9. Classification results on the IP dataset. (Best results are shown in bold)

Class	SVM	DBMA	DBDA	SSRN	SSFTT	BS2T	CS2DT	HybridFormer	Ours
1	99.42 ± 0.75	100.0 ± 0.00	100.0 ± 0.00	99.36 ± 0.68	99.99 ± 0.01	95.53 ± 5.53	100.00 ± 0.000	99.78 ± 0.34	99.90 ± 0.17
2	98.79 ± 0.37	99.07 ± 0.44	99.53 ± 0.42	99.83 ± 0.11	100.0 ± 0.00	97.53 ± 4.65	99.65 ± 0.43	99.99 ± 0.01	100.00.0 ± 0
3	87.98 ± 3.76	97.41 ± 1.28	97.79 ± 1.57	95.01 ± 4.98	99.99 ± 0.01	99.06 ± 0.84	99.89 ± 0.14	95.02 ± 11.12	97.45 ± 3.57
4	97.54 ± 0.59	92.39 ± 1.28	96.23 ± 1.07	96.91 ± 2.17	97.33 ± 1.10	93.93 ± 1.85	93.50 ± 3.16	99.77 ± 0.18	98.75 ± 2.57
5	95.09 ± 3.14	99.51 ± 0.28	99.23 ± 0.80	98.18 ± 1.93	99.96 ± 0.05	90.51 ± 8.84	98.58 ± 0.63	97.67 ± 1.12	97.74 ± 1.41
6	99.89 ± 0.08	99.83 ± 0.23	100.0 ± 0.00	99.92 ± 0.10	99.75 ± 0.20	100.0 ± 0.00	99.21 ± 1.05	99.43 ± 1.02	99.55 ± 0.91
7	95.59 ± 2.67	93.89 ± 4.56	98.63 ± 1.14	99.51 ± 0.86	99.46 ± 0.38	94.51 ± 3.12	99.96 ± 0.10	99.95 ± 0.11	99.89 ± 0.18
8	71.66 ± 2.54	94.00 ± 2.26	87.92 ± 1.03	88.52 ± 4.54	95.87 ± 0.63	90.58 ± 3.38	90.98 ± 2.83	94.87 ± 2.62	95.24 ± 2.38
9	98.08 ± 1.17	99.71 ± 0.33	99.26 ± 0.22	95.63 ± 7.56	99.99 ± 0.01	94.01 ± 5.68	99.53 ± 0.35	100.00 ± 0	100.0 ± 0.00
10	85.39 ± 3.46	85.46 ± 8.47	96.67 ± 2.69	97.93 ± 1.30	98.81 ± 0.43	86.39 ± 7.50	98.20 ± 1.43	97.61 ± 1.25	98.18 ± 1.56
11	86.97 ± 6.81	89.98 ± 8.01	96.87 ± 2.21	83.51 ± 25.72	99.36 ± 0.66	75.53 ± 38.22	98.69 ± 1.61	99.79 ± 0.41	99.71 ± 0.63
12	94.20 ± 4.00	99.75 ± 0.27	98.94 ± 1.48	97.13 ± 3.15	99.23 ± 0.83	98.65 ± 0.99	99.83 ± 0.12	99.43 ± 0.76	99.98 ± 0.02
13	93.43 ± 3.31	96.56 ± 1.56	99.88 ± 0.09	98.19 ± 2.47	92.94 ± 3.29	99.86 ± 0.12	98.92 ± 1.55	99.98 ± 0.05	99.86 ± 0.29
14	92.03 ± 5.43	99.52 ± 0.28	96.67 ± 0.60	93.85 ± 6.14	91.15 ± 4.26	87.12 ± 5.78	92.76 ± 6.33	99.23 ± 0.74	99.45 ± 0.62
15	71.02 ± 5.59	75.77 ± 16.40	94.24 ± 0.71	80.35 ± 7.61	89.61 ± 0.94	76.27 ± 1.56	87.32 ± 5.00	93.06 ± 3.80	98.36 ± 1.55
16	97.81 ± 1.19	97.47 ± 1.88	100.0 ± 0.00	99.42 ± 0.58	99.68 ± 0.22	98.78 ± 2.40	99.99 ± 0.01	99.76 ± 0.43	99.06 ± 1.27
OA	86.97 ± 0.86	91.43 ± 4.81	95.78 ± 0.03	92.43 ± 3.86	97.20 ± 0.25	90.85 ± 1.30	95.61 ± 0.78	97.45 ± 0.66	**98.35 ± 0.47**
AA	91.55 ± 0.62	95.02 ± 1.23	97.62 ± 0.12	95.20 ± 2.79	97.69 ± 0.36	92.39 ± 3.58	97.31 ± 0.48	98.46 ± 0.72	**98.95 ± 0.36**
Kappa	85.45 ± 0.97	90.52 ± 5.28	95.29 ± 0.04	91.58 ± 4.30	96.88 ± 0.28	89.82 ± 1.45	95.11 ± 0.86	97.89 ± 0.55	**98.63 ± 0.39**

Table 10. Classification results on the SV dataset. (Best results are shown in bold)

mance gains (as shown in Fig. 15(b)). Even more, for the Pavia dataset, the numerous RGBI samples result in a negative impact on super-resolution performance, as shown in Fig. 15(c). The size of the Pavia dataset is smaller than

Class	SVM	DBMA	DBDA	SSRN	SSFTT	BS2T	CS2DT	HybridFormer	Ours
1	94.74 ± 2.32	96.53 ± 1.22	95.57 ± 1.46	98.58 ± 0.16	96.57 ± 0.45	96.00 ± 1.02	95.81 ± 2.35	96.51 ± 2.66	98.87 ± 1.91
2	96.44 ± 1.49	99.44 ± 0.13	99.47 ± 0.14	99.81 ± 0.18	98.45 ± 0.68	98.61 ± 1.14	98.46 ± 1.00	99.13 ± 1.14	97.92 ± 0.59
3	98.16 ± 1.40	99.45 ± 0.77	100.0 ± 0.00	99.87 ± 0.24	99.77 ± 0.13	99.96 ± 0.06	99.95 ± 0.09	99.44 ± 0.56	99.79 ± 0.16
4	97.65 ± 1.40	99.01 ± 0.72	97.55 ± 0.97	98.59 ± 1.10	99.06 ± 0.75	99.50 ± 0.51	94.73 ± 1.90	99.44 ± 0.57	99.46 ± 0.49
5	93.00 ± 2.04	98.35 ± 1.55	96.61 ± 2.71	98.10 ± 1.45	100.0 ± 0.00	99.86 ± 0.27	97.07 ± 2.60	100.0 ± 0.00	99.95 ± 0.11
6	98.57 ± 1.82	99.13 ± 1.22	100.0 ± 0.00	98.73 ± 2.52	98.17 ± 1.37	100.0 ± 0.00	99.02 ± 1.68	88.70 ± 9.11	97.07 ± 2.64
7	86.08 ± 2.97	90.77 ± 1.95	92.75 ± 1.34	93.64 ± 3.65	93.78 ± 1.34	95.49 ± 1.09	94.89 ± 2.37	97.80 ± 0.90	98.01 ± 0.76
8	80.90 ± 2.47	97.16 ± 0.42	97.75 ± 1.53	98.40 ± 1.75	91.82 ± 0.75	99.78 ± 0.26	95.73 ± 2.91	87.49 ± 5.16	89.62 ± 3.01
9	74.64 ± 3.56	87.18 ± 4.20	93.70 ± 2.37	93.35 ± 3.19	91.35 ± 1.53	93.17 ± 2.18	93.81 ± 3.02	93.25 ± 2.46	94.51 ± 4.54
10	83.11 ± 3.94	91.73 ± 3.09	90.78 ± 2.91	91.70 ± 5.16	97.91 ± 1.49	93.45 ± 1.84	88.84 ± 2.88	98.37 ± 1.67	97.88 ± 2.63
11	81.27 ± 2.51	92.41 ± 0.60	92.07 ± 4.55	91.47 ± 1.16	98.05 ± 1.06	97.25 ± 1.09	92.59 ± 4.06	97.83 ± 2.35	99.16 ± 1.21
12	76.29 ± 3.79	95.45 ± 0.41	88.43 ± 1.98	91.17 ± 2.41	90.70 ± 1.26	94.61 ± 1.15	92.13 ± 5.07	94.16 ± 3.83	95.66 ± 3.71
13	63.57 ± 10.27	93.64 ± 1.57	90.03 ± 3.65	88.86 ± 4.58	96.54 ± 2.17	90.60 ± 1.70	95.83 ± 5.11	89.85 ± 6.89	94.32 ± 6.51
14	93.77 ± 4.56	98.39 ± 1.13	98.50 ± 1.62	91.78 ± 5.35	99.68 ± 0.18	98.15 ± 1.91	100.0 ± 0.00	100.0 ± 0.00	99.95 ± 0.10
15	99.21 ± 0.51	98.24 ± 0.25	97.92 ± 1.25	99.46 ± 0.66	100.0 ± 0.00	96.90 ± 1.55	97.51 ± 1.40	99.28 ± 0.79	99.87 ± 0.27
OA	87.47 ± 0.66	95.17 ± 0.29	94.83 ± 0.54	95.44 ± 0.57	96.32 ± 0.28	96.76 ± 0.37	94.95 ± 0.71	96.40 ± 0.30	**97.35** ± **0.30**
AA	87.83 ± 0.94	95.79 ± 0.28	95.41 ± 0.54	95.57 ± 0.55	96.79 ± 0.27	96.89 ± 0.35	95.76 ± 0.62	96.08 ± 0.79	**97.47** ± **0.47**
Kappa	86.44 ± 0.71	94.78 ± 0.31	94.41 ± 0.58	95.07 ± 0.62	96.03 ± 0.31	96.50 ± 0.41	94.54 ± 0.77	96.11 ± 0.33	**97.13** ± **0.33**

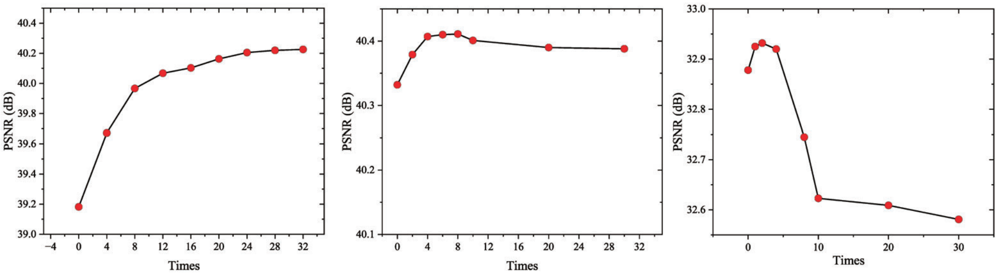
Table 11. Classification results on the HU2013 dataset. (Best results are shown in bold)

the CAVE and Chikusei datasets. In the beginning, a small amount of RGBI samples are introduced to enrich the training set and lead to positive feedback. With the increase of RGBI samples, the influence of the domain gap is amplified, resulting in performance drops. Therefore, for different scales and types of HSI datasets, the effectiveness of RGBI samples is varied. Especially for the small-scale (far smaller than Pavia) remote sensing HSI dataset, the limited samples would lead to the limited performance of the Transformer network.

12. Ablation Studies

Discussion of the TB Number, TL Number and Attention Head Number. We show the effects of the TB number, TL number and attention head number on the model's performance in Fig. 16(a), Fig. 16(b) and Fig. 16(c). The MPSNR is positively correlated with the TB number and TL number. As the TB number and TL number increase, the performance gain gradually becomes saturated. Meanwhile, the MPSNR is positively correlated with a small number of attention heads. When the attention head number does not exceed 6, the performance gain gradually. While the attention head number exceeds 12, the performance of VolFormer drops. To balance the performance and model size, the TB , TL and attention head number were set to 8, 6 and 6 to obtain a relatively effective and small model.

Discussion of the Computational Cost. We have conducted a comparison of the parameters and computational cost for our model with SOTA methods across various super-resolution tasks in detail shown in Table 12. The input HSI is 1×128×16×16, and 4× and 8× super-resolution tasks are employed.



(a) CAVE (b) Chikusei (c) Pavia

Figure 15. The performance with different ratios of the RGBI samples among different datasets.

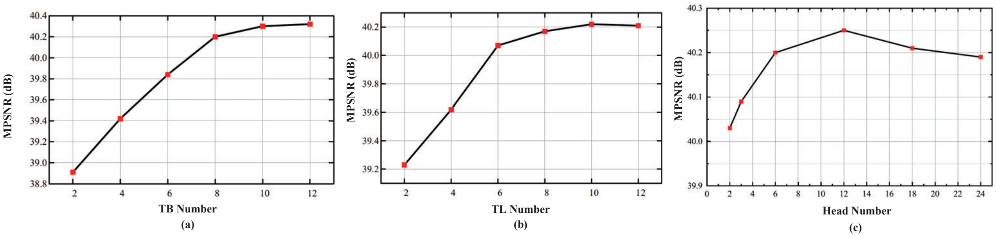


Figure 16. Ablation study on different settings of TB number, TL number and attention head number.

	Scale	MCNet	ERCSR	HSISR	DSTrans	ESSA	VolFormer
FLOPs	×4	289.26	287.74	203.23	137.63	50.45	90.15
Params		2.17	1.59	18.11	25.20	11.64	5.10
FLOPs	×8	2637	2691	761.90	166.05	199.07	120.29
Params		2.96	2.38	20.47	25.34	14.14	5.74

Table 12. Comparisons of computational cost and parameters.