SWViT-RRDB: Shifted Window Vision Transformer Integrating Residual in Residual Dense Block for Remote Sensing Super-Resolution

Mohamed Ramzy Ibrahim[1,2][a], Robert Benavente[2][b], Daniel Ponsa[2][c] and Felipe Lumbreras[2][d]

[1] *Computer Engineering Department, Arab Academy for Science, Technology and Maritime Transport, Alexandria, Egypt*

[2] *Computer Vision Center & Computer Science Department, Universitat Autònoma de Barcelona, Barcelona, Spain*

Keywords: Computer Vision, Super-Resolution, Remote Sensing, Deep Learning.

Abstract: Remote sensing applications, impacted by acquisition season and sensor variety, require high-resolution images. Transformer-based models improve satellite image super-resolution but are less effective than convolutional neural networks (CNNs) at extracting local details, crucial for image clarity. This paper introduces SWViT-RRDB, a new deep learning model for satellite imagery super-resolution. The SWViT-RRDB, combining transformer with convolution and attention blocks, overcomes the limitations of existing models by better representing small objects in satellite images. In this model, a pipeline of residual fusion group (RFG) blocks is used to combine the multi-headed self-attention (MSA) with residual in residual dense block (RRDB). This combines global and local image data for better super-resolution. Additionally, an overlapping cross-attention block (OCAB) is used to enhance fusion and allow interaction between neighboring pixels to maintain long-range pixel dependencies across the image. The SWViT-RRDB model and its larger variants outperform state-of-the-art (SoTA) models on two different satellite datasets in terms of PSNR and SSIM.

1 INTRODUCTION

Satellite images are key for different tasks like agriculture and weather forecasting, with more data coming from missions like Landsat-8 and Sentinel-2 (Wang et al., 2022a). However, getting high-resolution (HR) images is expensive and slow, especially for smaller satellite missions. A solution involves capturing low-resolution (LR) images in space and applying super-resolution (SR) methods on Earth to obtain HR images. Moreover, satellite images are particularly challenging due to their bird's-eye perspective (leading to the loss of small object details and textures), atmospheric conditions (haze, clouds, rain, etc.), variations in spectral and spatial resolution from different sensors, and wide geographic coverage (Ibrahim et al., 2022). Over the past few years, research in remote sensing SR has primarily focused on convolutional neural networks (CNNs) (Dong et al., 2016; Wang et al., 2022b). However, CNN architectures face two key issues: first, their content-independent convolution kernels face challenges in recovering different image regions; second, their fixed-size convolutions limit long-range dependency modeling in images (Liang et al., 2021). Recently, vision transformers have addressed some of these concerns by capturing global interactions using self-attention mechanisms (Dosovitskiy et al., 2021). However, vision transformers also have some drawbacks that come from dividing images into fixed patch sizes. First, border pixels cannot share information with neighboring patches. Second, super-resolved patches may have border artifacts that affect the final full-sized super-resolved image (Chen et al., 2023). Different modified vision transformer approaches (Liang et al., 2021; Liu et al., 2021) attempted to address the issues. However, the proposed solutions do not utilize more input pixels than CNNs, leading to some information loss (Liang et al., 2021). The recent HAT model (Chen et al., 2023) combines Swin transformer (Liang et al., 2021; Liu et al., 2021) modules and attention techniques to enhance feature representation. However, it falls short in remote sensing SR, particularly for small details in satellite bird's-eye LR images due to the lack of dense feature representation. Conversely, residual in residual dense blocks (RRDB) (Wang et al., 2019; Ibrahim et al.,

[a] https://orcid.org/0000-0002-7483-4468
[b] https://orcid.org/0000-0001-9819-4445
[c] https://orcid.org/0000-0002-7330-6524
[d] https://orcid.org/0000-0003-2887-8053

Ibrahim, M., Benavente, R., Ponsa, D. and Lumbreras, F.
SWViT-RRDB: Shifted Window Vision Transformer Integrating Residual in Residual Dense Block for Remote Sensing Super-Resolution.
DOI: 10.5220/0012399300003660

In *Proceedings of the 19th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2024) - Volume 3: VISAPP*, pages 575-582
ISBN: 978-989-758-679-8; ISSN: 2184-4321

2022) excel in finer details representation in LR images due to their generation of extensive features.

This paper proposes a new SWViT-RRDB model for satellite imagery SR, inspired by the HAT (Chen et al., 2023). It combines a multi-headed self-attention (MSA) module for global information with a deep CNN (RRDB) to extract detailed local information and prevent gradient loss during training. This fusion addresses the information loss seen in models based solely on CNNs or transformers. The model also includes an overlapping cross-attention window (OCAB) to enable cross-window interaction. The model is evaluated on two datasets: real-captured LR-HR image pairs from different satellites and a dataset with generated HR images from pan-sharpened LR images. These diverse datasets cover variations in geographical conditions, acquisition seasons, sensor types, and resolutions.

The paper contribution can be summarized as: 1) A new SWViT-RRDB model for remote sensing satellite imagery SR is presented that focuses on exploiting more information from LR images for finer detail enhancement. 2) Fusing global MSA module features with extensive local RRDB features and preventing gradient loss. 3) Model adaptation to diverse datasets, including real-captured or synthesized HR pan-sharpened with varying scaling factors.

2 RELATED WORK

2.1 CNN-Based Models

Dong et al. (Dong et al., 2014) introduced SRCNN, the first SR network with three convolutional layers. Also, they proposed VDSR (Dong et al., 2016), a very deep super-resolution approach with 20 convolutional layers for enhanced image reconstruction. Tuna et al. (Tuna et al., 2018) compared SRCNN and VDSR models on satellite images, favoring VDSR for both panchromatic and multi-spectral images. Some studies use residual blocks (Zhang and et al., 2018; Zhang et al., 2021) to overcome the vanishing gradient problem. Other studies focus on dense blocks (Jiang et al., 2018; Zhang and et al., 2018) to stack a large number of features for improved image representation and reconstruction. More recent studies proposed attention mechanisms (Zhang et al., 2018; Woo et al., 2018) to enhance image reconstruction. Wang et al. (Wang et al., 2022b) introduced a lightweight lattice block with channel separation, attention module, and feature enhancement block to improve texture extraction capability for better satellite imagery reconstruction. After the development of GANs in SR, some studies employed Generative Adversarial Networks (GANs) to reduce perceptual loss and enhance super-resolved images (Ledig and et al., 2017; Wang et al., 2019). Lanaras et al. (Lanaras et al., 2018) proposed an improved ESRGAN, based on (Wang et al., 2019), trained with multi-spectral satellite image pairs.

2.2 Transformer-Based Models

Recently, after the success of transformers in natural language processing (Aleissaee et al., 2022). Transformers have been employed in various high-level computer vision tasks like classification, segmentation, and object detection (Dosovitskiy et al., 2021; Liu et al., 2021; Zhang et al., 2022). They have also been applied to low-level tasks like image restoration and SR. Chen et al. (Chen et al., 2021) introduced the image processing transformer (IPT) for restoration, requiring extensive parameters, large datasets, and multi-task learning, but its focus on small image patches could potentially limit their effectiveness in image restoration. Liang et al. (Liang et al., 2021) introduced SwinIR for single-image super-resolution (SISR), incorporating a shifted window mechanism for efficient self-attention computation and enabling information sharing between border pixels of neighboring patches. While transformers focus on global interactions, studies suggest that adding convolutions to transformer-based networks enhances visual representation (Xiao et al., 2021). Chen et al. (Chen et al., 2023) introduced a hybrid attention transformer (HAT) for feature attention with CAB and OCAB to enhance pixel information exploration and feature fusion. He et al. (He et al., 2022) proposed a dense spectral transformer with ResNet for multispectral remote sensing images, facilitating long-range relationship learning within the data.

3 METHODS

The proposed SWViT-RRDB model pipeline, inspired by HAT (Chen et al., 2023), combines the Swin transformer MSA (Liang et al., 2021) modules, RRDB (Ibrahim et al., 2022) (a dense CNN with multiple residuals), and OCAB (Chen et al., 2023) modules to complement each other.

The SWViT-RRDB introduces a new residual in residual transformer fusion block (RRTFB) that fuses local (RRDB) and global (MSA) features, enhancing the representation of small and detailed objects. The RRDB module enhances local details, particularly in satellite images, by generating dense local information, and its mix of local and global residuals pre-

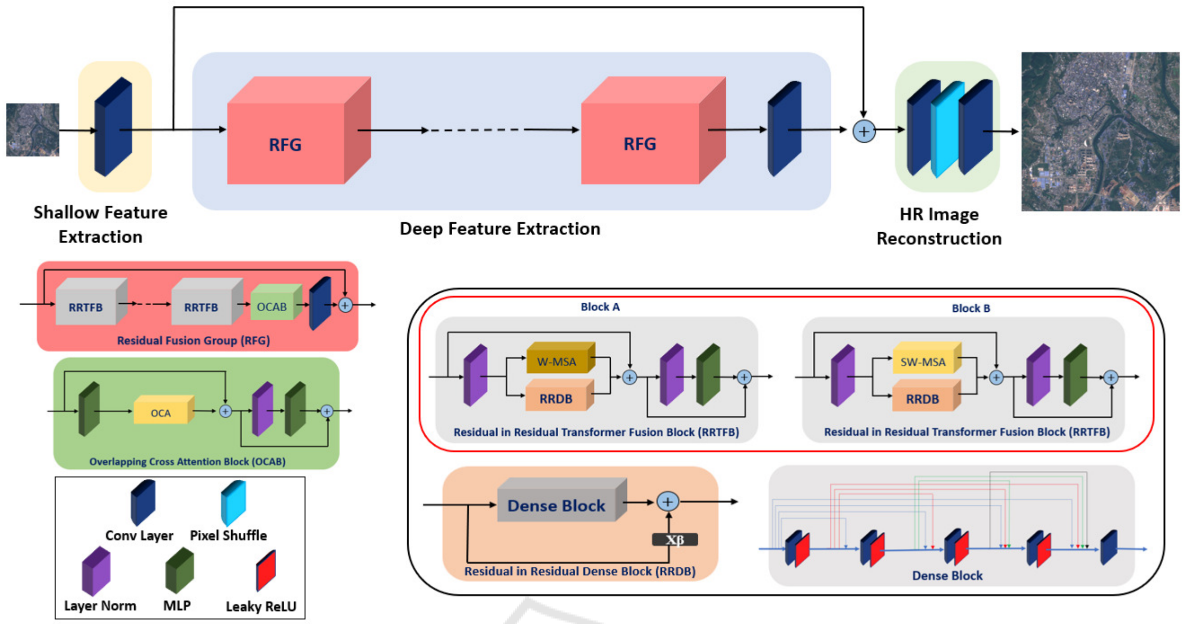
Shallow Feature Extraction
RFG
RFG
Deep Feature Extraction
HR Image Reconstruction
RRTFB
RRTFB
OCAB
Residual Fusion Group (RFG)
OCA
Overlapping Cross Attention Block (OCAB)
Conv Layer
Pixel Shuffle
Layer Norm
MLP
Leaky ReLU
Block A
W-MSA
RRDB
Residual in Residual Transformer Fusion Block (RRTFB)
Block B
SW-MSA
RRDB
Residual in Residual Transformer Fusion Block (RRTFB)
Dense Block
Xβ
Residual in Residual Dense Block (RRDB)
Dense Block

Figure 1: SWViT-RRDB model architecture, with key enhancements highlighted in black and red squares.

vents vanishing gradients. The MSA modules, alternating between a regular window multi-head self-attention (W-MSA) and a shifted window multi-head self-attention (SW-MSA), extract global information to address convolution kernel locality issues and support long-range dependencies.

3.1 Overall Model Architecture

The SWViT-RRDB model contains three phases: shallow feature extraction, deep feature extraction, and HR picture reconstruction (see Fig 1).

In the shallow feature extraction phase, shallow features F_S are extracted from each LR image of size $H \times W \times C_{in}$ where H, W, and C_{in} are the height, the width, and the number of channels in the image, using a 3×3 convolution layer $(CONV)$. These shallow features are then fed to a deep feature extraction phase P_{DFE}. The P_{DFE} phase consists of N residual fusion groups (RFG) followed by a $CONV$. RFGs are connected sequentially, with each subsequent block taking the output of the previous one as its input. The output, F_D, is obtained by fusing the initial feature F_S with the output of the $CONV$ after the last RFG block using a skip residual connection, preserving image information and preventing any loss during training. In the HR image reconstruction phase (P_{REC}), F_D features are upscaled with a pixel-shuffle module using sub-pixel convolution, trained for each feature map to enhance LR image upscaling, instead of a traditional bicubic filter (Shi et al., 2016). The upscaled features are then convolved to reconstruct the HR image.

The RFG serves as the primary component in our proposed model, emphasizing deep feature extraction. As illustrated in Fig. 1, the RFG comprises K instances of the RRTFB. The RRTFBs are arranged sequentially, with each subsequent block taking the output of the previous one as its input. The last RRTFB is succeeded by OCAB (Chen et al., 2023) and $CONV$. Additionally, the input to each RFG is combined with the output of $CONV$ using a skip residual connection, facilitating the aggregation of different feature levels for stable training and to prevent any gradient loss.

3.2 Residual in Residual Transformer Fusion Block (RRTFB)

A newly proposed RRTFB combines the global features inside image patches that are extracted from MSA and the local features inside image patches that are extracted from RRDB to enhance the image representation for the SR task. Moreover, the alternating MSA modules support long-range dependencies across the image. As shown in Fig. 1, RRTFB is composed of two parts. The first part consists of an input fed to a LayerNorm (LN) followed by two parallel branches having RRDB, and either a W-MSA or a SW-MSA (Liang et al., 2021) as presented in Block A and Block B in Fig. 1, respectively. The SW-MSA is used because in W-MSA the local window is fixed across different layers introducing a no connection across local windows of MSA. To address this, both MSA blocks (W-MSA and SW-MSA) are used interchangeably to enable cross-window connections (Liu et al., 2021). In SW-MSA, features are shifted by $[\frac{M}{2}, \frac{M}{2}]$ pixels before partitioning (M is the window size). Additionally, a skip residual connection combines the output with the input of the RRTFB. The second part consists of the output of the first part, which is fed to a LN and followed by a multi-layer perceptron (MLP) with four fully connected layers with a non-linear GeLU activating function to add more feature transformation. The output of the second part is added to the output of the first part by a skip residual connection.

Residual in Residual Dense Block (RRDB). RRDB extracts extensive, detailed local features from LR images using dense blocks, merging them with shallow frequencies via a global residual connection, which is important for representing small objects. This representation is needed, especially in satellite images SR for better earth observation. As shown in Fig. 1, the input X_{LN} is fed to RRDB, which consists of a dense block with a global residual connection. The global residual connection is multiplied by a residual scaling factor (β) to stabilize the results during training. After that, it is added to the output of the dense block to avoid gradient losses (Ibrahim et al., 2022).

The dense block consists of four $CONV$ with a ReLU activation function and ends with a fifth $CONV$. It consists of multiple local residual connections, and all the preceding generated feature maps, $x_0, x_1, ..., x_{l-1}$, are sent to the l^{th} layer. The residual scaling factor (β) is chosen to be equal to 0.2 as different values were tested in (Ibrahim et al., 2022) and the best result was achieved with this value. Its fusion with global and local residual connections prevents gradient losses in dense blocks and stabilizes training.

3.3 Overlapping Cross-Attention Block (OCAB)

The OCAB module (Chen et al., 2023) (presented in Fig. 1) is used to establish cross-window connections. The OCAB module is similar to the basic Swin transformer module (Liang et al., 2021), except that it replaces MSA with an overlapping cross-attention layer (OCA), which uses different window sizes with overlapping from MSA to partition the generated features. It addresses convolution challenges by enhancing long dependency representation in window self-attention.

4 EXPERIMENTAL SETTINGS

In this section, we detail the datasets, the training settings, and the performance metrics used.

4.1 Datasets Description

The evaluation of the proposed model is done on two satellite image datasets, namely OLI2MSI (Wang et al., 2021) and Alsat2B (Djerida et al., 2021).

OLI2MSI Dataset. It is a real-world remote sensing satellite imagery dataset containing LR-HR pairs of images from two different satellites (Landsat8-OLI and Sentinel2-MSI). It consists of 5225 image pairs for training and 100 pairs for testing. The LR input comprises Landsat8-OLI images with a 30-meter spatial resolution, and the HR ground truth consists of Sentinel2-MSI images with a 10-meter spatial resolution which results in a scale factor of 3 in SR.

Alsat Dataset. It is generated from Alsat-2B satellite. The LR images are captured from a satellite with a 10-meter spatial resolution, and the HR images are generated by pan-sharpening the LR images. Pan-sharpening is a widely recognized technique that enhances the spatial resolution of satellite imagery by combining information from both the PAN (2.5-meter spatial resolution) and multi-spectral bands (10-meter spatial resolution). The dataset contains 2182 training samples and three subsets for testing: ”Agriculture”, ”Urban”, and ”Special structures[1]”, with 56, 282, and 239 image pairs, respectively, and a scale factor of 4.

4.2 Training Settings

Our SWViT-RRDB, developed in Python and PyTorch, is tested on 3.80 GHz Core i7, 32 GB RAM, and 24 GB RTX 3090 systems. Training involves cropping LR and HR images into 64×64 and 192×192 patches, respectively, with augmentation via flips and rotations. We train models using 16-image batches, Adam optimizer, and cosine annealing learning rate for 800K iterations, ensuring identical training conditions across datasets for fairness.

SWViT-RRDB (baseline) comprises 6 RFG blocks, 6 RRTFB, 6 MSA modules, 1 OCAB module with 6 OCA blocks, and W-MSA/SW-MSA modules with a window size of 8. We tested its variants to evaluate the impact of different RFG numbers: SWViT-RRDB-S (3 RFGs), SWViT-RRDB-M (9 RFGs), and SWViT-RRDB-L (12 RFGs).

[1] Contain constructions such as airports, stadiums, etc.

Loss Function. SWViT-RRDB is optimized using the L_1 loss function ($L_1 = \| I_{SR} - I_{HR} \|$), minimizing the difference between the super-resolved image (I_{SR}) and the ground truth HR image (I_{HR}).

Performance Metrics. We use peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM) for quantitative model evaluation (Ibrahim et al., 2024). These measures compare super-resolved images to ground-truth HR images for fidelity. For qualitative evaluation, we generate quality maps by measuring the Euclidean distance between corresponding pixels in high-resolution and super-resolved images, normalizing the results to a 0-255 scale.

5 RESULTS AND DISCUSSION

We conduct ablation studies on SWViT-RRDB modules and compare SWViT-RRDB and its variants against SoTA.

5.1 Ablation Study

Effect of OCAB and RRDB on SWViT-RRDB. We studied the impact of OCAB and RRDB on SWViT-RRDB using both datasets, OLI2MSI and Alsat. The best results, shown in Table 1, came from combining OCAB and RRDB, increasing PSNR by 0.38 dB for OLI2MSI and 0.21 dB for Alsat, compared to SWViT-RRDB without these modules. The OCAB module enables cross-window connections, complementing the fusion of extensive local RRDB features and global MSA features. This could potentially explain the improved representation of small objects in satellite images needed for better image SR. The results show that adding either OCAB or RRDB to the model improves it compared to just using MSA modules. The OCAB module is especially effective, slightly better than just using RRDB. This emphasizes the need for OCAB and MSA modules to work with RRDB for better image reconstruction.

Table 1: Study on adding (✓) or removing (✗) OCAB and RRDB modules in SWViT-RRDB.

Dataset	**Module**	**SWViT-RRDB Model**			
	OCAB **RRDB**	✗ ✗	✗ ✓	✓ ✗	✓ ✓
OLI2MSI	PSNR	37.98	38.23	38.27	38.36
	SSIM	0.9492	0.9523	0.9529	0.9533
Alsat	PSNR	17.33	17.43	17.50	17.54
	SSIM	0.2934	0.3358	0.3566	0.3582

Table 2: Study on modifying the number of RFGs (N) in our SWViT-RRDB.

Dataset	**Metric**	**No. of RFG blocks in SWViT-RRDB**			
		N=3	**N=6**	**N=9**	**N=12**
OLI2MSI	PSNR	38.11	38.36	38.41	38.48
	SSIM	0.9504	0.9533	0.9542	0.9557
Alsat	PSNR	17.46	17.54	17.58	17.63
	SSIM	0.3562	0.3582	0.3620	0.3649

Effect of Changing the Number of RFGs. Experiments were performed to study how changing the number of RFG blocks affects the SWViT-RRDB model(6 RFGs) and its variants: SWViT-RRDB-S (3 RFGs), SWViT-RRDB-M (9 RFGs), and SWViT-RRDB-L (12 RFGs). Table 2 shows the quantitative results (PSNR and SSIM) obtained. As a general trend, we can see that the results improve as the number of RFG blocks increases. Increasing RFGs from 3 to 6 significantly improves PSNR by 0.25 dB in OLI2MSI and 0.08 dB in Alsat datasets. Increasing RFG blocks from 6 to 9 and then to 12 leads to near-linear PSNR improvements: 0.05 dB and 0.07 dB in OLI2MSI, and 0.04 dB and 0.05 dB in Alsat, respectively. From these results, we conclude that increasing RFG blocks from 3 to 6 leads to a PSNR improvement more than twice the sum of PSNR improvements resulting from increasing RFG blocks from 6 to 9 and 9 to 12 in OLI2MSI (real-captured dataset). While in Alsat (simulated dataset), increasing RFG blocks from 3 to 6 leads to a PSNR improvement that nearly equals the sum of PSNR improvements resulting from increasing RFG blocks from 6 to 9 and 9 to 12.

5.2 Comparison with SoTA

Quantitative Results. Our proposed model (SWViT-RRDB) and its larger variants are evaluated against various DL SoTA models. The models included in experiments corresponds to the best SoTA results in the last decade. It includes methods from several approaches such as: bicubic (baseline method), CNN-based models (SRCNN (Dong et al., 2016), VDSR (Kim et al., 2016), SRResNet (Ledig and et al., 2017), EDSR (Lim et al., 2017), and 2DRRDB (Ibrahim et al., 2022)), GAN-based models (ESRGAN (Wang et al., 2019), and cDRSRGAN (Wang et al., 2021)), models utilizing channel attention blocks (CAB) such as RCAN (Zhang et al., 2018), transformers such as SwinIR and SwinIR+ (Liang et al., 2021), and a fusion of transformers and

Table 3: Quantitative results (PSNR(dB) / SSIM) comparison with SoTA models.

Method	**OLI2MSI** (×3)		**Alsat** (×4)							
	ALL		**ALL**		**Agriculture**		**Urban**		**Special**	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Bicubic	33.24	0.879	14.30	0.260	14.68	0.342	13.41	0.213	14.98	0.283
ESRGAN	33.82	0.888	15.97	0.268	17.61	0.372	14.89	0.213	16.57	0.294
SRCNN	35.26	0.907	15.99	0.270	17.68	0.377	14.88	0.217	16.59	0.293
VDSR	35.87	0.919	16.02	0.272	17.79	0.383	14.85	0.218	16.67	0.296
SRResNet	37.13	0.939	16.06	0.272	17.82	0.387	14.90	0.219	16.71	0.295
EDSR	37.68	0.946	16.12	0.274	17.83	0.387	14.93	0.218	16.79	0.298
cDRSRGAN	37.71	0.947	16.31	0.269	17.94	0.388	15.23	0.217	16.91	0.290
RCAN	37.85	0.947	16.78	0.276	18.62	0.391	15.82	0.221	17.23	0.299
2DRRDB	37.89	0.948	16.86	0.282	18.78	0.390	15.89	0.223	17.31	0.311
SwinIR	37.98	0.949	17.33	0.293	18.96	0.392	16.29	0.226	17.88	0.330
SwinIR+	38.08	0.951	17.36	0.310	19.01	0.416	16.33	0.249	17.91	0.341
HAT	38.21	0.952	17.43	0.340	19.03	0.427	16.42	0.274	17.97	0.378
HAT-L	38.29	0.953	17.48	0.352	19.03	0.432	16.44	0.299	18.06	0.382
SWViT-RRDB (ours)	**38.36**	**0.953**	**17.54**	**0.358**	**19.12**	**0.440**	**16.52**	**0.309**	**18.10**	**0.384**
SWViT-RRDB-M (ours)	**38.41**	**0.954**	**17.58**	**0.362**	**19.16**	**0.447**	**16.56**	**0.316**	**18.12**	**0.384**
SWViT-RRDB-L (ours)	**38.48**	**0.956**	**17.63**	**0.365**	**19.17**	**0.451**	**16.61**	**0.321**	**18.17**	**0.385**

CABs such as HAT and HAT-L (Chen et al., 2023). As seen in Table 3, our proposed SWViT-RRDB variants (SWViT-RRDB, SWViT-RRDB-M, and SWViT-RRDB-L), combining transformers and CNNs, achieve superior results compared to the latest SoTA models using CNNs, CABs, GANs, transformers, or a fusion of transformers and CABs. They scored the highest results in terms of PSNR and SSIM in both datasets, OLI2MSI and Alsat. Similary, for Alsat dataset classes, the baseline SWViT-RRDB achieves the highest PSNR and SSIM in each class (Agriculture, Urban, and Special structure) compared to SoTA models. Among the SWViT-RRDB variants, SWViT-RRDB-L, with 12 RFGs, achieves the highest PSNR and SSIM on OLI2MSI and Alsat datasets. From Table 3, first, our model (SWViT-RRDB) and its larger variants outperform CNN-based, GANs-based, and CAB-based methods due to their ability to capture long dependencies and process information more comprehensively. Second, they do better than modern transformer-based methods, which cannot generate local information due to the absence of convolutions or share information between neighboring pixels in different fixed-sized patches. Additionally, the transformer-based methods address the issue of border artifacts that can affect the generated super-resolved full image. Finally, they surpass the fusion of transformers and CABs (HAT and HAT-L) due to the potential lack of extensive local features generated by deep CNNs. Our model (SWViT-RRDB) and its larger variants propose using the RRDB CNN module, which generates extensive local features and fuses local and global residual connections inside it to avoid the vanishing gradient problem, which enables better representation of small objects in satellite images.

Qualitative Results. Fig. 3(A) visually compares our models (SWViT-RRDB-L,SWViT-RRDB-M, and baseline) with SoTA (bicubic, SwinIR, SwinIR+, HAT, HAT-L) on OLI2MSI dataset. Our models show sharper, clearer edges and lines with less haze than the others. Moreover, Fig. 3(B) reveals that SWViT-RRDB-L and SWViT-RRDB have darker quality maps than other models (bicubic, SwinIR+, HAT-L), indicating fewer errors, particularly compared to SwinIR+, which shows lighter, error-prone areas.

6 CONCLUSIONS

In conclusion, our proposed SWViT-RRDB model addresses the limitations of prior SR models in satellite imagery. Building on the previous work that combines CNN and transformers, our pipeline involves shallow feature extraction, deep feature extraction via RFG blocks, and HR image reconstruction. Experiments and ablation studies reveal that our SWViT-RRDB model, along with its larger variants, SWViT-RRDB-M and SWViT-RRDB-L, surpass SoTA models in PSNR and SSIM. The fusion of global MSA features, local RRDB features, and the OCAB module shows improvement in image reconstruction compared to SoTA models. Extensive local features of the RRDB module and combined global-local residual connections enhance small object representation in satellite images and prevent vanishing gradients

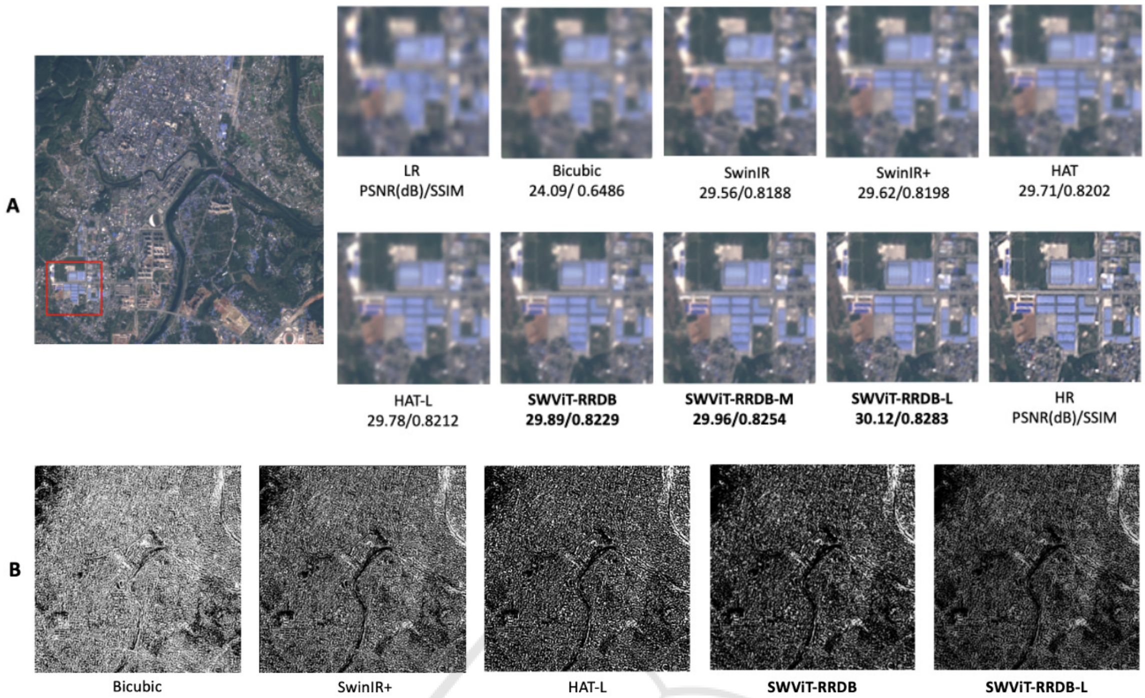

Figure 2: Qualitative results on OLI2MSI dataset (A) Visual comparison between bicubic, SwinIR, SwinIR+, HAT, HAT-L, SWViT-RRDB, SWViT-RRDB-M, and SWViT-RRDB-L on a selected part (red square) inside the image. (B) Quality map comparsion between bicubic, SwinIR+, HAT-L, SWViT-RRDB, SWViT-RRDB-L on the full image.

during training. Moreover, the alternating MSA modules (W-MSA and SW-MSA) and OCAB module highlight the value of cross-window connections for long-range dependencies, complementing the RRDB for improved image reconstruction. Our SWViT-RRDB model surpasses SoTA on two diverse satellite datasets: OLI2MSI (real, ×3 scale) and Alsat (synthetic HR from pan-sharpened LR, ×4 scale), proving effectiveness across different scales and satellites. Future work will extend SWViT-RRDB to different domains and smaller scales (×8 and ×16).

ACKNOWLEDGEMENTS

This work is partially supported by Grant PID2021-128945NB-I00 funded by MCIN/AEI/10.13039/501100011033 and by "ERDF A way of making Europe". The authors acknowledge the support of the Generalitat de Catalunya CERCA Program to CVC's general activities, and the Departament de Recerca i Universitats from the Generalitat de Catalunya with reference 2021SGR01499.

REFERENCES

Aleissaee, A. A., Kumar, A., Anwer, R. M., Khan, S., Cholakkal, H., Xia, G.-S., and khan, F. S. (2022). Transformers in Remote Sensing: A Survey. *arXiv preprint arXiv:2209.01206*.

Chen, H., Wang, Y., Guo, T., Xu, C., Deng, Y., Liu, Z., Ma, S., Xu, C., Xu, C., and Gao, W. (2021). Pre-Trained Image Processing Transformer. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12294–12305.

Chen, X., Wang, X., Zhou, J., Qiao, Y., and Dong, C. (2023). Activating More Pixels in Image Super-Resolution Transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22367–22377.

Djerida, A., Djerriri, K., Karoui, M. S., and El Amin larabi, M. (2021). A New Public Alsat-2B Dataset for Single-Image Super-Resolution. In *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, pages 8095–8098. ISSN: 2153-7003.

Dong, C., Loy, C. C., He, K., and Tang, X. (2014). Learning a Deep Convolutional Network for Image Super-Resolution. In Fleet, D., Pajdla, T., Schiele, B., and Tuytelaars, T., editors, *Computer Vision – ECCV 2014*, volume 8692, pages 184–199. Springer International Publishing, Cham.

Dong, C., Loy, C. C., He, K., and Tang, X. (2016). Image Super-Resolution Using Deep Convolutional Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(2):295–307.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

He, J., Yuan, Q., Li, J., Xiao, Y., Liu, X., and Zou, Y. (2022). DsTer: A dense spectral transformer for remote sensing spectral super-resolution. *International Journal of Applied Earth Observation and Geoinformation*, 109:102773.

Ibrahim, M. R., Benavente, R., Lumbreras, F., and Ponsa, D. (2022). 3DRRDB: Super Resolution of Multiple Remote Sensing Images using 3D Residual in Residual Dense Blocks. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 322–331.

Ibrahim, M. R., Benavente, R., Ponsa, D., and Lumbreras, F. (2024). Unveiling the Influence of Image Super-Resolution on Aerial Scene Classification. In Vasconcelos, V., Domingues, I., and Paredes, S., editors, *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, Lecture Notes in Computer Science, pages 214–228.

Jiang, K., Wang, Z., Yi, P., Jiang, J., Xiao, J., and Yao, Y. (2018). Deep Distillation Recursive Network for Remote Sensing Imagery Super-Resolution. *Remote Sensing*, 10(11):1700.

Kim, J., Lee, J. K., and Lee, K. M. (2016). Accurate Image Super-Resolution Using Very Deep Convolutional Networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1646–1654.

Lanaras, C., Bioucas-Dias, J., Galliani, S., Baltsavias, E., and Schindler, K. (2018). Super-resolution of Sentinel-2 images: Learning a globally applicable deep neural network. *ISPRS Journal of Photogrammetry and Remote Sensing*, 146:305–319.

Ledig, C. and et al. (2017). Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 105–114.

Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., and Timofte, R. (2021). SwinIR: Image Restoration Using Swin Transformer. In *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 1833–1844.

Lim, B., Son, S., Kim, H., Nah, S., and Lee, K. M. (2017). Enhanced Deep Residual Networks for Single Image Super-Resolution. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1132–1140.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. (2021). Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9992–10002.

Shi, W., Caballero, J., Huszar, F., Totz, J., Aitken, A. P., Bishop, R., Rueckert, D., and Wang, Z. (2016). Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1874–1883.

Tuna, C., Unal, G., and Sertel, E. (2018). Single-frame super resolution of remote-sensing images by convolutional neural networks. *International Journal of Remote Sensing*, 39(8):2463–2479.

Wang, J., Gao, K., Zhang, Z., Ni, C., Hu, Z., Chen, D., and Wu, Q. (2021). Multisensor Remote Sensing Imagery Super-Resolution with Conditional GAN. *Journal of Remote Sensing*, 2021:2021/9829706.

Wang, P., Bayram, B., and Sertel, E. (2022a). A comprehensive review on deep learning based remote sensing image super-resolution methods. *Earth-Science Reviews*, 232:104110.

Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Qiao, Y., and Loy, C. C. (2019). ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks. In Leal-Taixé, L. and Roth, S., editors, *Computer Vision – ECCV 2018 Workshops*, volume 11133, pages 63–79. Springer International Publishing, Cham.

Wang, Z., Li, L., Xue, Y., Jiang, C., Wang, J., Sun, K., and Ma, H. (2022b). FeNet: Feature Enhancement Network for Lightweight Remote-Sensing Image Super-Resolution. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–12.

Woo, S., Park, J., Lee, J.-Y., and Kweon, I. S. (2018). CBAM: Convolutional Block Attention Module. In Ferrari, V., Hebert, M., Sminchisescu, C., and Weiss, Y., editors, *Computer Vision – ECCV 2018*, Lecture Notes in Computer Science, pages 3–19.

Xiao, T., Singh, M., Mintun, E., Darrell, T., Dollar, P., and Girshick, R. (2021). Early Convolutions Help Transformers See Better. In *Advances in Neural Information Processing Systems*, volume 34, pages 30392–30400.

Zhang, C., Jiang, W., Zhang, Y., Wang, W., Zhao, Q., and Wang, C. (2022). Transformer and CNN Hybrid Deep Neural Network for Semantic Segmentation of Very-High-Resolution Remote Sensing Imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–20.

Zhang, K., Li, Y., Zuo, W., Zhang, L., Van Gool, L., and Timofte, R. (2021). Plug-and-Play Image Restoration with Deep Denoiser Prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1.

Zhang, Y. and et al. (2018). Residual Dense Network for Image Super-Resolution. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2472–2481.

Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., and Fu, Y. (2018). Image Super-Resolution Using Very Deep Residual Channel Attention Networks. In Ferrari, V., Hebert, M., Sminchisescu, C., and Weiss, Y., editors, *Computer Vision – ECCV 2018*, volume 11211, pages 294–310. Springer International Publishing, Cham.