Metadata, Wavelet, and Time Aware Diffusion Models for Satellite Image Super Resolution

Luigi Sigillo, Renato Giamba, Danilo Comminiello
Dept. of Information Engineering, Electronics, and Telecomm.,
Sapienza University of Rome, Italy
`luigi.sigillo@uniroma1.it`

Abstract

The acquisition of high-resolution satellite imagery is often constrained by the spatial and temporal limitations of satellite sensors, as well as the high costs associated with frequent observations. These challenges hinder applications such as environmental monitoring, disaster response, and agricultural management, which require fine-grained and high-resolution data. In this paper, we propose MWT-Diff, an innovative framework for satellite image super-resolution (SR) that combines latent diffusion models with wavelet transforms to address these challenges. At the core of the framework is a novel metadata-, wavelet-, and time-aware encoder (MWT-Encoder), which generates embeddings that capture metadata attributes, multi-scale frequency information, and temporal relationships. The embedded feature representations steer the hierarchical diffusion dynamics, through which the model progressively reconstructs high-resolution satellite imagery from low-resolution inputs. This process preserves critical spatial characteristics including textural patterns, boundary discontinuities, and high-frequency spectral components essential for detailed remote sensing analysis. The comparative analysis of MWT-Diff across multiple datasets demonstrated favorable performance compared to recent approaches, as measured by standard perceptual quality metrics including FID and LPIPS.

1 Introduction

Satellite Imaging. The need for high-resolution (HR) satellite imagery has grown exponentially with applications spanning environmental monitoring, urban planning, disaster response, and agricultural management (Hazmy et al., 2024). However, the acquisition of satellite images, particularly high-resolution imagery, can be both costly and limited by the spatial and temporal resolution constraints of satellite sensors (Khanna et al., 2024). For instance, Sentinel-2 satellites provide images of Earth's land surface every five days, with resolutions ranging from 10 to 60 meters (Cornebise et al., 2022). While valuable, this level of detail often falls short for tasks requiring fine-grained monitoring, such as assessing agricultural health or managing wildfires. Acquiring high-resolution images, such as those captured by SPOT6, for large areas and at high frequencies remains a significant challenge due to cost and technical limitations.

Multi-Image and Hyperspectral Image Super-Resolution. Image super-resolution (SR) has emerged as a pivotal area in computer vision, aiming to enhance the resolution of low-quality images to generate high-resolution outputs (Wang et al., 2020), this has been also exploited in LR satellite images (Cornebise et al., 2022). Khanna et al. (2024) introduced DiffusionSat which represents an important step towards this task which allows the use of LR image sequences as input, along with metadata, to generate high-quality HR output. These inputs often consist of multispectral satellite images captured at varying ground sampling distances (GSDs) e.g. it can utilize LR images from Sentinel-2 to produce HR images comparable in quality to those from HR datasets like fMoW (Christie et al., 2018).

Another milestone in the topic is SatDiffMoe Luo et al. (2024) which leverages the complementary information embedded in multiple LR satellite images captured at different time points. The key innovation lies in its diffusion-based fusion algorithm, which conditions the reconstruction of HR

images on each individual LR input while merging their outputs to achieve enhanced detail and accuracy.

SR exploiting prior knowledge. To enhance the quality of generated images, many SR techniques have incorporated prior knowledge to guide the reconstruction process. Reference-based SR methods leverage similar HR images as explicit priors to assist in generating the corresponding HR outputs (Zheng et al., 2018). However, aligning the features of reference images with the LR input is challenging in real-world scenarios, and such explicit priors are not always readily available, limiting their practicality. Recent SR approaches have shifted to implicit priors, which offer more flexibility and generalization (Wang et al., 2018a; Yu et al., 2018; Zhao et al., 2022; Menon et al., 2020). However, these methods often suffer from insufficient representational capacity or inaccurate feature alignment, resulting in suboptimal output quality. StableSR Wang et al. (2024) represents a significant advancement by leveraging the robust and extensive generative prior found in pre-trained diffusion models, fine-tuning a frozen pre-trained diffusion model with only a small number of trainable parameters, making it both efficient and highly effective.

In this paper we explore latent diffusion models within the problem of satellite image SR, establishing a robust framework capable of reconstructing HR images from LR input while addressing the unique characteristics and challenges of satellite imagery. The proposed methodology integrates wavelet transforms and conditional generation, to better capture fine details and contextual relationships within the data.

2 THE PROPOSED MWT-DIFF METHOD

Feature Modulation. The conditional generation approach for SR with diffusion models leverages the latent representation of degraded input images, extracted through an encoder, to guide the generation process. To enhance the guidance during generation, in StableSR Wang et al. (2024) an additional encoder is introduced to extract multi-scale features $\{F^n\}_{n=1}^N$ from the degraded LR image. These extracted features were subsequently used to modulate the intermediate feature maps, $\{F_{\text{dif}}^n\}_{n=1}^N$, within the residual blocks of Stable Diffusion (SD) (Rombach et al., 2021) via spatial feature transformations (SFT) introduced by Wang et al. (2018a).

During fine-tuning, the weights of the SD model are frozen, and only the encoder and the SFT layers are trained. This approach ensured the preservation of the generative prior inherent in SD while allowing structural information from the LR image to be incorporated effectively.

Additionally, the capacity of the model to integrate textual information is improved by utilizing CLIP-encoded text embeddings during training. This enhancement facilitates a more nuanced understanding of both the image category and the country metadata embedded in the dataset.

MWT-Encoder. We introduce a metadata- and wavelet- and time-aware encoder (MWT-Encoder) as the core of the baseline model architecture inspired by Sigillo et al. (2024) that introduced a class- and time-aware encoder, and Khanna et al. (2024) which highlighted the importance of metadata (e.g., timestamp, latitude, longitude) associated with satellite images.

In this context, each image $x \in \mathbb{R}^{3 \times H \times W}$ is paired with numerical metadata $K \in \mathbb{R}^M$, where M represents the number of metadata items. Following the methodology introduced in StableSR, we incorporate temporal information into the encoder using a time-embedding layer. We encode metadata using sinusoidal timestep embeddings, as in Khanna et al. (2024), to avoid discretization errors from CLIP's poor numerical encoding (Radford et al., 2021).

Recent works show that the use of wavelets can improve SR within diffusion models (Aloisi et al., 2024; Zhao et al., 2024). Our idea is to take advantage of the wavelets, not to perform diffusion in frequency space (Phung et al., 2023), but instead to extract features from the wavelet transform of the satellite images through the pre-training of a Wavelet Vision Transformer (WaveViT) (Yao et al., 2022).

WaveViT consists of four stages, each comprising a patch embedding layer, a series of Wavelet blocks, and subsequent feed-forward layers. In the initial phase of the architecture, discrete wavelet transforms (DWT) are applied to the input images. This operation decomposes the images into multiple frequency components, enabling the model to simultaneously capture both low-frequency (global structure) and high-frequency (fine detail) information. This is particularly critical for pre-
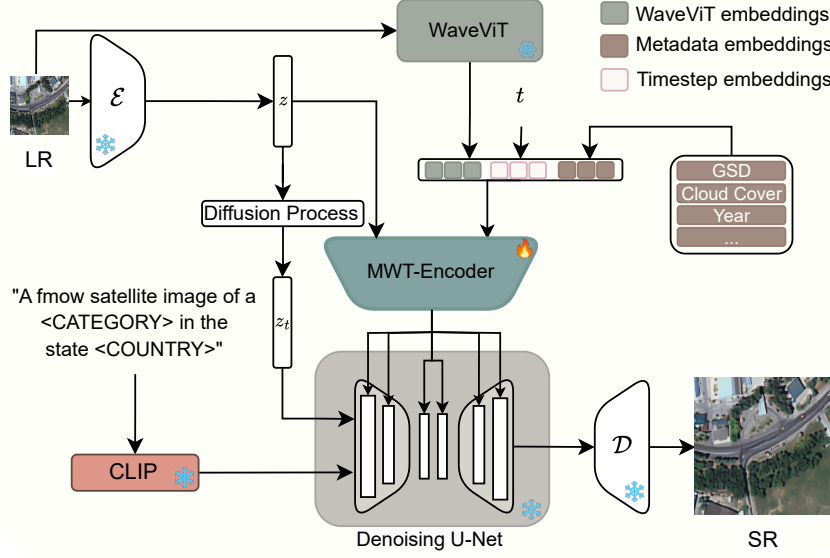
Figure 1: Overview of MWT-Diff framework with our MWT-Encoder.

serving the intricate textures and details commonly found in satellite imagery. After pre-training, which extracts robust feature representation leveraged throughout the network to perform the super-resolution task, we integrate WaveViT in our architecture, concatenating the embeddings along with the timesteps and metadata embeddings.

The concatenated metadata tensors $m \in \mathbb{R}^{1024}$, wavelet embeddings $w \in \mathbb{R}^{1024}$, and timestep embeddings $t \in \mathbb{R}^{1024}$ yield a vector $b \in \mathbb{R}^{3072}$, which is combined with the latent representation of the image $z \in \mathbb{R}^{h \times w \times c}$. The latent z is obtained from the encoder $z = \mathcal{E}(x)$, where $x \in \mathbb{R}^{3 \times H \times W}$ is the low-resolution image. Both the encoder $\mathcal{E}$ and the decoder $\mathcal{D}$ are frozen components of the VAE used in SD Rombach et al. (2021), alongside the time-conditional U-Net responsible for performing the latent diffusion process.

The MWT-Encoder δ_θ generates conditioning embeddings that guide the denoising process of the conditional auto-encoder ϵ_θ at multiple scales. These embeddings are incorporated following the architecture of the U-Net. The complete objective function for training the model is defined as:

$$L = \mathbb{E}_{E(x),y,\epsilon \sim \mathcal{N}(0,1),m,w,t} \left[\left\| \epsilon - \epsilon_\theta(z_t, \delta_\theta(m, w, t, z), \tau(y)) \right\|_2^2 \right] \qquad (1)$$

In equation 1, $\tau(y)$ represents the domain-specific encoder, implemented using a frozen CLIP model, which processes the conditioning caption y. The variable m corresponds to the metadata embeddings, w is the wavelet embedding, t is the timestep embedding, z is the latent representation of the low-resolution image extracted by the VAE encoder, and z_t is the latent representation after the addition of noise during the diffusion process.

The objective is to optimize the MWT-Encoder $\delta_\theta(m, w, t, z)$ to provide effective conditioning embeddings. These embeddings ensure that the denoising process generates HR images that faithfully capture the information encoded in both the input metadata and the conditioning captions. The final MWT-Diff model used for the super-resolution task is illustrated in Figure 1 within our MWT-Encoder.

3 EXPERIMENTAL RESULTS

We evaluate our model on two datasets. The first dataset is a subset of the fMoW dataset Christie et al. (2018), used for the super-resolution task of 4x scaling, where 128x128 pixel low-resolution

Figure 2: Qualitative results on 128x128 $\rightarrow$ 512x512 with fMoW (Christie et al., 2018).

	Model	FID ↓	LPIPS ↓
fMoW	Low Resolution	114.38	0.756 +/- 0.004
	StableSR	53.85	0.345 +/- 0.002
	MWT-Diff	**53.07** (-1.44%)	**0.336** +/- 0.002 (-2.61%)
Sentinel2-fMoW	WorldStrat Cornebise et al. (2022)	426.7	0.736 ± 0.092
	MSRResNet Wang et al. (2018b)	286.5	0.783 ± 0.081
	DBPN Haris et al. (2018)	278.2	0.750 ± 0.052
	Pix2Pix Isola et al. (2017)	196.3	0.643 ± 0.045
	SatDiffMoE Luo et al. (2024)	115.6	0.606 ± 0.044
	DiffusionSat Khanna et al. (2024)	102.9	0.638 ± 0.034
	ControlNet Zhang et al. (2023)	102.3	0.644 ± 0.034
	MWT-Diff	**98.1** (-4.11%)	**0.555** ± 0.003 (-8.41%)

Table 1: Evaluation metrics comparison for the $128 \times 128 \rightarrow 512 \times 512$ super resolution task on fMoW dataset Christie et al. (2018), and on Sentinel2-fMoW dataset Cong et al. (2022).

images are upscaled to 512x512 pixel high-resolution images. The second dataset is a subset of the Sentinel2-fMoW dataset Cong et al. (2022), employed for the super-resolution task on hyperspectral images. For inference we adopt the DDPM Ho et al. (2020) sampling strategy utilizing 200 timesteps.

The results underscore the superior performance of our model across both the fMoW and hyperspectral image super-resolution tasks, as presented in Table 3. By integrating our MWT-Encoder, the model significantly enhances its ability to reconstruct intricate features and capture fine-grained details within the images. MWT-Diff consistently outperforms its counterparts, including traditional CNN-based methods and other diffusion models, as evidenced by its superior performance in both FID and LPIPS metrics.

4 CONCLUSION

In conclusion, this paper presents a novel latent diffusion model to address the challenge of satellite image super-resolution. By incorporating the MWT-Encoder, the proposed framework effectively utilizes contextual information, leading to improved image fidelity and detail. Our MWT-Diff achieved superior performance with lower FID and LPIPS scores, demonstrating its ability to generate high-quality, ground-truth-like images for practical applications in satellite imaging, including urban planning, environmental monitoring, and disaster management.

REFERENCES

Lorenzo Aloisi, Luigi Sigillo, Aurelio Uncini, and Danilo Comminiello. A wavelet diffusion gan for image super-resolution, 2024. URL `https://arxiv.org/abs/2410.17966`.

Gordon Christie, Neil Fendley, James Wilson, and Ryan Mukherjee. Functional map of the world. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6172–6180, 2018.

Yezhen Cong, Samar Khanna, Chenlin Meng, Patrick Liu, Erik Rozi, Yutong He, Marshall Burke, David Lobell, and Stefano Ermon. Satmae: Pre-training transformers for temporal and multi-spectral satellite imagery. *Advances in Neural Information Processing Systems*, 35:197–211, 2022.

Julien Cornebise, Ivan Oršolić, and Freddie Kalaitzis. Open high-resolution satellite imagery: The worldstrat dataset–with application to super-resolution. *Advances in Neural Information Processing Systems*, 35:25979–25991, 2022.

Muhammad Haris, Gregory Shakhnarovich, and Norimichi Ukita. Deep back-projection networks for super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1664–1673, 2018.

Asa Ibnu Hazmy, Ammar Hawbani, Xingfu Wang, Ahmed Al-Dubai, Aiman Ghannami, Ali Abdullah Yahya, Liang Zhao, and Saeed Hamood Alsamhi. Potential of satellite-airborne sensing technologies for agriculture 4.0 and climate-resilient: A review. *IEEE Sensors Journal*, 24(4): 4161–4180, 2024. doi: 10.1109/JSEN.2023.3343428.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1125–1134, 2017.

Samar Khanna, Patrick Liu, Linqi Zhou, Chenlin Meng, Robin Rombach, Marshall Burke, David B. Lobell, and Stefano Ermon. Diffusionsat: A generative foundation model for satellite imagery. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=I5webNFDgQ`.

Zhaoxu Luo, Bowen Song, and Liyue Shen. Satdiffmoe: A mixture of estimation method for satellite image super-resolution with latent diffusion models. In *2nd Workshop on Advancing Neural Network Training: Computational Efficiency, Scalability, and Resource Optimization (WANT@ICML 2024)*, 2024. URL `https://openreview.net/forum?id=dpK5a5xnel`.

Sachit Menon, Alexandru Damian, Shijia Hu, Nikhil Ravi, and Cynthia Rudin. Pulse: Self-supervised photo upsampling via latent space exploration of generative models. *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 2434–2442, 2020.

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*, pp. 10199–10208, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Int. Conf. on Machine Learning*, 2021.

Robin Rombach, A. Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 10674–10685, 2021.

Luigi Sigillo, Riccardo Fosco Gramaccioni, Alessandro Nicolosi, and Danilo Comminiello. Ship in sight: Diffusion models for ship-image super resolution. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, 2024. doi: 10.1109/IJCNN60899.2024.10650042.

Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin CK Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. *International Journal of Computer Vision*, pp. 1–21, 2024.

Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Recovering realistic texture in image super-resolution by deep spatial feature transform. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 606–615, 2018a.

Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European conference on computer vision (ECCV) workshops*, pp. 0–0, 2018b.

Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1905–1914, 2021.

Zhihao Wang, Jian Chen, and Steven CH Hoi. Deep learning for image super-resolution: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3365–3387, 2020.

Ting Yao, Yingwei Pan, Yehao Li, Chong-Wah Ngo, and Tao Mei. Wave-vit: Unifying wavelet and transformers for visual representation learning. In *Proceedings of the European conference on computer vision (ECCV)*, 2022.

Ke Yu, Chao Dong, Liang Lin, and Chen Change Loy. Crafting a toolchain for image restoration by deep reinforcement learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2443–2452, 2018.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3836–3847, 2023.

Chen Zhao, Weiling Cai, Chenyu Dong, and Chengwei Hu. Wavelet-based fourier information interaction with frequency diffusion adjustment for underwater image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8281–8291, 2024.

Yang Zhao, Yu-Chuan Su, Chun-Te Chu, Yandong Li, Marius Renn, Yukun Zhu, Changyou Chen, and Xuhui Jia. Rethinking deep face restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7652–7661, 2022.

Haitian Zheng, Mengqi Ji, Haoqian Wang, Yebin Liu, and Lu Fang. Crossnet: An end-to-end reference-based super resolution network using cross-scale warping. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 88–104, 2018.

A APPENDIX

Figure 3: Additional results on 128x128 → 512x512 with fMoW Christie et al. (2018). Comparison between MWT-Diff and StableSR.

A.1 DATASET DETAILS

A.1.1 FUNCTIONAL MAP OF THE WORLD (FMOW)

The metadata accompanying each image of the fMoW dataset Christie et al. (2018) provides contextual information, including location, time, sun angles, physical sizes, and other attributes, facilitating a more informed analysis of the objects within the images.

The original fMoW dataset consists of over 1 million satellite images collected from more than 200 countries, each annotated with at least one bounding box corresponding to one of 63 categories. For the purpose of conducting a 4x scaling super-resolution task—transforming 128×128 pixel LR images into 512×512 pixel HR images we select a balanced subset of 100k samples from the original dataset, ensuring proportional representation of all 63 classes.

To generate the HR images, a resize operation to 512×512 pixels followed by a center crop was applied, preserving the critical details of the images. The LR images were derived from the HR counterparts by degrading them using various techniques outlined Real-ESRGAN Wang et al. (2021), effectively reducing their resolution while simulating realistic degradations.

Additionally, to establish a baseline for comparison with the super-resolution model outputs, a standard upscaling technique was applied to the LR images, creating reference images for evaluation. The resulting refined dataset is optimized and prepared for training and benchmarking super-resolution models.

A.1.2 Functional Map of the World - Sentinel-2 (fMoW-Sentinel)

The dataset Cong et al. (2022) comprises temporal image series collected by the Sentinel-2 satellite, corresponding to the geographic locations of the original fMoW dataset. These images were captured at multiple time points and are categorized into 62 distinct types of building and land use, following the classification scheme of the fMoW dataset.

The fMoW-Sentinel dataset consists of Sentinel-2 images with a 10-meter spatial resolution, created from cloud composites aggregated over 90-day intervals. Each image includes 13 spectral bands from the Sentinel-2 surface reflectance dataset, offering a comprehensive multispectral representation. Building on previous works Cong et al. (2022); Luo et al. (2024), we use a derived dataset named fMoW-Sentinel-fMoW-RGB. This dataset pairs Sentinel-2 images (10m–60m ground sampling distance, GSD) with high-resolution fMoW-RGB images (0.3m–1.5m GSD) at each of the original fMoW locations.

A key distinction from SatdiffMoE is the handling of high-resolution images. While it uses bounding boxes provided in the metadata to extract relevant areas and crop them into 512×512 patches, the fMoW high-resolution images in this work were resized in their entirety to 512×512 pixels.

For each high-resolution fMoW image, we select one corresponding Sentinel-2 image from the same location. Similar to SattdiffMoE, only the RGB bands of the Sentinel-2 images were retained. These RGB images were then resized to match the dimensions of the fMoW-RGB images, ensuring consistency in resolution for downstream tasks.

A.2 Training details

The MWT-Diff model is trained for eight and four epochs, respectively. Further training beyond these limits resulted in worsening FID scores. The Adam optimizer was employed, with a learning rate set to 5e-5. An instance of the WaveViT-B model was pre-trained and evaluated directly on the fMoW dataset. This evaluation was conducted on 1,000 images from the fMoW validation set.

A.3 Evaluation details

In line with the methodology presented by Luo et al. (2024), the evaluation of the WaveViT is further refined by focusing on the first 100 images from the validation set across six specific classes: Airport, Amusement Park, Car Dealership, Crop Field, Educational Institution, and Electric Substation. We generate a synthetic set of captions for each class of the fMoW dataset using the pattern: "A fMoW satellite image of a class_type". Each image is paired with these synthetic captions and input into the model under evaluation. The model then predicted the most relevant caption associated with each image.

A.4 Additional Qualitative Comparisons

Figure 4 illustrates qualitative results on the Sentinel2-fMoW dataset Cong et al. (2022), showcasing the low-resolution Sentinel-2 input, the super-resolved output generated by MWT-Diff, and the corresponding high-resolution fMoW reference image. Figure 3 depicts an additional comparison on the standard fMoW dataset Christie et al. (2018), where MWT-Diff is directly compared against the StableSR baseline.

The results demonstrate consistent qualitative improvements across datasets. In the Sentinel2-fMoW case, MWT-Diff successfully reconstructs fine-grained structures despite the challenging resolution gap between Sentinel-2 inputs and fMoW ground truth. Meanwhile, in the fMoW comparison, MWT-Diff outputs exhibit clearer boundaries and more coherent textures than those produced by StableSR, indicating the effectiveness of our metadata- and frequency-aware conditioning mechanisms.

These examples highlight the model's generalization capability and the robustness of its generative priors across different spatial resolutions and sensor characteristics.
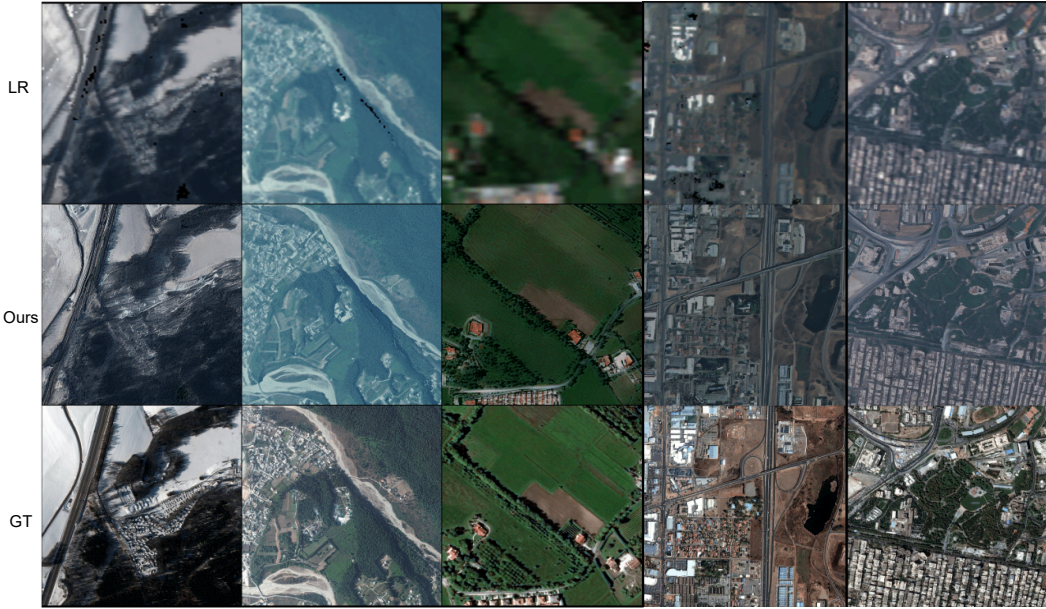
Figure 4: Qualitative results on Sentinel2-fMoW dataset Cong et al. (2022). The image includes a comparison of the LR Sentinel-2 input, the output of the model, and the corresponding HR fMoW images.

A.5 ABLATION STUDY

We conduct a series of ablation experiments to evaluate the individual contributions and trade-offs of key components in the MWT-Diff framework. This analysis explores the impact of removing conditioning modules and examines design choices that affect the model's complexity, efficiency, and applicability.

A.5.1 CLIP-BASED TEXT ENCODER

To assess the necessity of the CLIP-based caption encoder, we removed this component from the architecture and retrained the model. The goal was to investigate whether this simplification would reduce computational overhead or accelerate inference. Empirical results indicated no measurable improvement in sampling speed. On the contrary, a slight degradation in performance was observed across FID and LPIPS metrics. These findings suggest that while the CLIP encoder introduces additional complexity, it provides marginal gains in perceptual quality and semantic alignment, which justify its inclusion in the final model.

A.5.2 COMPARISON WITH STABLESR AND IMPLICATIONS FOR REAL-WORLD DEPLOYMENT

We further examined the computational cost of MWT-Diff relative to StableSR. Despite the integration of additional embeddings (metadata, wavelet, and temporal encodings), the inference time remained comparable. This parity is attributed to the lightweight nature of the MWT-Encoder and the use of a single-image input pipeline. Notably, when compared to methods that requires multi-image fusion (Khanna et al., 2024; Luo et al., 2024; Cornebise et al., 2022) from different timestamps, MWT-Diff demonstrates superior practicality. The absence of complex temporal alignment and fusion procedures enables more straightforward deployment in real-world scenarios where temporally-aligned sequences may not be available.
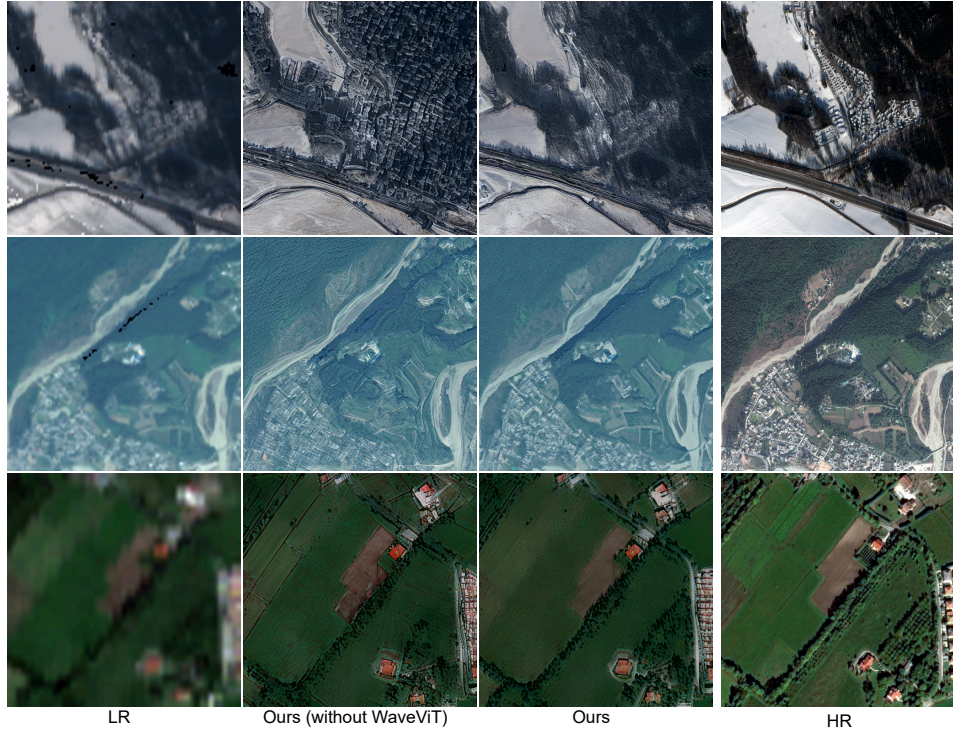
Figure 5: Qualitative comparison of super-resolution outputs with and without WaveViT embeddings. From left to right: Low-resolution input, MWT-Diff without WaveViT, MWT-Diff with WaveViT, and Ground Truth.

A.5.3 IMPACT OF RGB-ONLY INPUT

Our current implementation utilizes only the RGB bands of the Sentinel-2 imagery. This choice was made to maintain a lean architecture, facilitating efficient training and inference while enhancing the usability of the model in operational environments with limited multispectral data. Although RGB data suffice for a broad range of super-resolution tasks, we recognize the potential advantages of incorporating additional spectral channels. Future work will investigate the integration of multispectral and hyperspectral inputs to assess their impact on reconstruction accuracy and downstream utility, particularly in applications such as vegetation monitoring and land use classification.

A.5.4 EFFECT OF WAVELET-BASED EMBEDDINGS (WAVEVIT)

To evaluate the contribution of frequency-aware representations, we conducted an experiment disabling the WaveViT component. Figure 5 shows a qualitative comparison of super-resolved samples with and without WaveViT-based embeddings.

From visual inspection, it is evident that the inclusion of WaveViT significantly enhances the perceptual sharpness and texture fidelity of the generated images. Samples generated with WaveViT appear less blurry and less noisy, exhibiting structural characteristics that more closely resemble the ground truth. These results confirm that multi-scale frequency features, as captured by WaveViT, play a critical role in recovering fine details and reducing generative artifacts.