



UNIVERSITÀ DI PARMA

UNIVERSITÀ DEGLI STUDI DI PARMA

Dottorato di Ricerca in Tecnologie dell'Informazione

XXXVIII Ciclo

**Deep Learning for Enhanced Representation of
Multispectral Imagery in Remote Sensing**

Coordinatore:

Chiar.mo Prof. Marco Locatelli

Tutor:

Chiar.mo Prof. Andrea Prati

Dottorando: *Vittorio Bernuzzi*

Anni 2022-2025

*"Se puoi sognare, e non far diventare i sogni i tuoi padroni
se puoi pensare e non far diventare i pensieri il tuo scopo [...]
Tua sarà la Terra e tutto ciò che è in essa,
E — quel che più conta — sarai un Uomo, figlio mio!"
(Rudyard Kipling, Se)*

*Questo lavoro è dedicato in primis ai miei genitori, per
l'incondizionato supporto sotto qualunque aspetto, senza di
loro tutto questo non sarebbe stato possibile.*

*A tutte le persone che durante questo percorso di dottorato
hanno incrociato il mio cammino e preso altre strade o che
tutt'ora condividono il tragitto assieme a me, che direttamente o
indirettamente hanno saputo trasmettermi la motivazione per
portarlo a termine.*

Summary

Introduction	1
1 Background	7
1.1 Vision Transformers	7
1.1.1 The attention operation	9
1.1.2 Swin Transformer	9
1.2 Single-Image Super-Resolution	9
1.2.1 Super-Resolution for Remote Sensing Images via Local–Global Combined Network	13
1.2.2 High Quality Remote Sensing Image Super-Resolution Us- ing Deep Memory Connected Network	14
1.2.3 Transformer-based restoration and SwinIR	15
1.2.4 SwinFIR: Revisiting the SwinIR with Fast Fourier Convolu- tion and Improved Training for Image Super-Resolution . . .	17
1.2.5 Swin2SR: SwinV2 for compressed-image super-resolution and restoration	19
1.2.6 Dual Aggregation Transformer (DAT) for Image Super-Resolution	22
1.3 Masked Image Modelling	24
1.3.1 Masked Autoencoders Are Scalable Vision Learners	27
1.3.2 SatMAE: Pre-training Transformers for Temporal and Multi- Spectral Satellite Imagery	29

1.3.3	Rethinking Transformers Pre-training for Multi-Spectral Satellite Imagery	31
1.3.4	SpectralGPT: Spectral Remote Sensing Foundation Model	33
1.4	Discrete Wavelet Transform in computer vision	35
1.4.1	Wave-ViT: Unifying Wavelet and Transformers for Visual Representation Learning	36
1.4.2	Wavelet-Driven Masked Image Modeling: A Path to Efficient Visual Representation	38
2	Research findings	41
2.1	Swin2-MoSE: A New Single Image Super-Resolution Model for Remote Sensing	41
2.1.1	Introduction	41
2.1.2	Related	42
2.1.3	Architecture	43
2.1.4	Experiments	49
2.1.5	Conclusions	61
2.2	WaveMAE: Wavelet decomposition Masked Auto-Encoder for Remote Sensing	62
2.2.1	Introduction	62
2.2.2	Related	63
2.2.3	Theory priors	64
2.2.4	Architecture	66
2.2.5	Experiments	75
2.2.6	Conclusion	92
3	Conclusions and Future Work	93
3.1	Conclusions	93
3.2	Future Work	94
4	Publications	97

Summary	iii
A Memory-augmented Online Video Anomaly Detection	99
A.1 Introduction	99
A.2 Related	101
A.3 Architecture	103
A.4 Experiments	105
A.4.1 Ablation study	110
A.4.2 Comparison with the state of the art	113
A.5 Conclusions	113
Bibliography	115
Acknowledgements	127

List of Figures

1	Example an RGB image (left) and derived NDVI index (right), illustrating how multispectral data can quantify vegetation health. Green color indicates a good vegetation health, while red color suggests bad vegetation health. [1]	2
1.1	The structure of a Vision Transformer encoder. The figure on the right, illustrates the internal architecture of the deep network, highlighting the primary modules within each ViT block. Here, L denotes the architectural depth, indicating the frequency with which the transformer's blocks are consecutively reiterated.	8
1.2	The architecture of a Swin-Transformer.	10
1.3	An illustration of the windowing mechanism implemented in Swin-Transformer. The attention is calculated within the non overlapping windows, alleviating the quadratic complexity requirements for the operation.	10
1.4	Architecture of the CNN based method LGCNet for remote sensing super resolution.	13
1.5	Architecture of the CNN based method DMCN for remote sensing super resolution.	14
1.6	Detailed representation of the DMCN's building blocks.	15
1.7	Schematic overview: shallow convolutional embedding, stacked residual Swin Transformer blocks with shifted-window attention, and a compact reconstruction head.	16

1.8	In this figure the SwinFIR pipeline is shown (shallow conv $\rightarrow$ RSTB stack $\rightarrow$ reconstruction) and the SwinFIR augmentation (a compact FFC branch fused into the spatial stream).	18
1.9	Swin2SR architecture: SwinV2-based Residual Transformer Blocks (RSTB_V2) in the deep extractor, long-skip shallow path, pixel-shuffle reconstruction, and the auxiliary/high-frequency losses used during training.	19
1.10	Overview of the Dual Aggregation Transformer (DAT) architecture. Each block alternates between Spatial Self-Attention (SSA) and Channel Self-Attention (CSA). The Adaptive Interaction Module (AIM) fuses outputs from both domains, while the Spatial-Gate Feed-Forward Network (SGFN) enhances local detail. The architecture preserves a long skip connection from shallow to deep features.	23
1.11	Illustration of the MAE workflow: random masking of patches, encoder processes visible patches only, small decoder reconstructs masked patches.	28
1.12	Overview of the SatMAE framework for multispectral imagery. The encoder processes a subset of visible spatial patches, while a lightweight decoder reconstructs masked patches across all spectral channels. . .	30
1.13	SatMAE++ pipeline for multi-spectral imagery. The input image is downsampled to multiple scales, the encoder processes the coarsest scale, and the decoder reconstructs higher-resolution images through convolution-based upsampling blocks. Multi-scale reconstruction encourages the encoder to learn hierarchical spatial and cross-spectral representations.	31
1.14	Overview of the SpectralGPT architecture. The model employs a 3D tokenization strategy to process spectral RS images, capturing both spatial and spectral information through its encoder-decoder framework.	33

1.15	Wave-ViT Wavelets block. Keys/values undergo DWT, a 3×3 convolution at half resolution, attention is computed with full-resolution queries, and IDWT reconstructs a high-resolution feature that is fused with the attention output.	36
1.16	The specific designs of (a) the foundational ViT Backbone self-attention block, (b) the self-attention block that employs a down-sampling operation (DS(2, 2)) to halve the spatial scale of height and width, and (c) WaveViT's Wavelets block, which leverages wavelet transforms to accomplish lossless down-sampling.	38
1.17	WaMIM training: images are patchified and masked; the encoder provides layerwise features, the decoder reconstructs multi-level wavelet targets (detail subbands from shallow features, approximation from deep features). Loss weights w_ψ^k and w_ϕ^k prioritize the most informative scales.	39
2.1	Overview of the Swin2-MoSE architecture (solid colours highlight modifications with respect to Swin2SR). The <i>Feature Extractor</i> component is analysed later in Section 2.1.4 in connection with the semantic segmentation experiments.	43
2.2	Mixture-of-Experts design (solid colours denote modified components): (a) the Swin v2 MoSE Layer (S2ML) couples a Self-Attention block (see Figure 2.3) with an MoE-SM in place of the MLP; (b) internal layout of MoE-SM featuring the Smart Merger for fusing expert outputs.	44
2.3	Comparison of positional encodings within the self-attention module: (a) Relative Position Encoding (RPE) in Swin Transformer; (b) Log-Spaced CPB (log-CPB) in Swin Transformer v2; (c) Locally-enhanced Positional Encoding (LePE) in CSwin; (d) integrated configuration combining multiple encodings.	46
2.4	Evolution of the NCC score under different training-loss configurations.	50

2.5	Evolution of the SSIM score under different training-loss configurations.	51
2.6	Evolution of the PSNR score under different training-loss configurations.	51
2.7	Learned bias matrices for RPE and LePE visualised post-training. .	53
2.8	Comparison between 4x SISR methods on Sen2Ven μ s dataset. HR reference is Ven μ s B8 (red edge, 5 m). The crop of the HR reference image is shown in the red square, and the zoomed cropped part that was chosen is shown in the blue square.	57
2.9	Comparison between 4x SISR methods on Sen2Ven μ s dataset. HR reference is Ven μ s B9 (red edge, 5 m). The crop of the HR reference image is shown in the red square, and the zoomed cropped part that was chosen is shown in the blue square.	58
2.10	Comparison between 4x SISR methods on Sen2Ven μ s dataset. HR reference is Ven μ s B10 (red edge, 5 m). The crop of the HR reference image is shown in the red square, and the zoomed cropped part that was chosen is shown in the blue square.	58
2.11	Comparison between 4x SISR methods on Sen2Ven μ s dataset. HR reference is Ven μ s B11 (NIR, 5 m). The crop of the HR reference image is shown in the red square, and the zoomed cropped part that was chosen is shown in the blue square.	59
2.12	Comparison between 3x SISR methods on OLI2MSI dataset. HR reference is Sentinel-2 B2 (blue, 10 m). The crop of the HR reference image is shown in the red square, and the zoomed cropped part that was chosen is shown in the blue square.	59
2.13	Comparison between 3x SISR methods on OLI2MSI dataset. HR reference is Sentinel-2 B3 (green, 10 m). The crop of the HR reference image is shown in the red square, and the zoomed cropped part that was chosen is shown in the blue square.	60

- 2.14 Comparison between 3x SISR methods on OLI2MSI dataset. HR reference is Sentinel-2 B4 (red, 10 m). The crop of the HR reference image is shown in the red square, and the zoomed cropped part that was chosen is shown in the blue square. 60
- 2.15 Semantic segmentation on SeasoNet: training curves for F1-micro and F1-macro. 61
- 2.16 Three-level 2D DWT of an image. The pixel region highlighted in red gives rise to the components LL (approximation), LH (horizontal details), HL (vertical details), and HH (diagonal details) across scales. 64
- 2.17 Illustration of WaveMAE architecture, the core design follows that of the MAE [2]. The method applies a 4-level DWT decomposition to the input image x of resolution (H, W) is obtaining 4 set of wavelet components at different scales of resolutions $(H/2, W/2), (H/4, W/4), (H/8, W/8)$ and $(H/16, W/16)$. Following this operation the components in S are fed to the Multi-level patch embedding, here lighter colors stand for shallow decomposition levels and low-scale feature extraction, on the opposite darker colors mean deeper decomposition level and high-scale feature extraction. After, the full sequence is fed to the MAE framework, the decoder reconstructs the masked wavelet components $\bar{S}$ at the same resolution specified previously. A smooth L1 loss is calculated between the input and the predicted wavelet components. Additionally, the predicted components are used to calculate the inverse DWT, obtaining the reconstructed multispectral image $\bar{x}$. An MSE loss is calculated between the latter and the input image x . In both losses the method attends only to the previously masked parts of the components/reconstructed image for the calculation. 67
- 2.18 Tube Masking scheme: for a three-level DWT, spatially aligned locations are masked identically across all levels. From top to bottom, rows correspond to shallow-to-deep levels; components are colour-coded [LL, HL, LH, HH]; grey indicates masked patches. 69

- 2.19 Mechanism of the Multi-level Patch Embedding. For an input of size 64×64 and $N=4$ levels, DWT subbands have the resolutions reported in the table. Adapting the patch size to each component's spatial resolution aligns tokens at the same image location across levels. White indicates pixel grids; yellow outlines patch boundaries. 70
- 2.20 Computation of the Geo-conditioned Positional Encoding (GPE). Meta-data provide latitude/longitude for each image x ; after converting degrees to radians, Spherical Harmonics (Section 2.2.4) are evaluated and then projected with a 1-layer MLP to the encoder embedding dimension $D=768$ 71
- 2.21 Multi-level Decoder. The masked sequence is first projected to the decoder width $\hat{D}=512$, then filled with learnable mask tokens. After adding APE, the complete sequence is processed by 8 transformer blocks. Four independent prediction heads (one per level) operate on level-specific subsequences obtained by splitting the decoded tokens. Owing to the Multi-level Patch Embedding, each component yields a fixed 196-token sequence per level. Thus, the fourth (deepest) level aggregates 784 tokens (for $HF_4=\{HL_4, LH_4, HH_4\}$ plus LL_4), while the remaining levels contain 588 tokens each (high-frequency components only). Each level's head comprises separate MLPs (three, or four at the deepest level), each dedicated to one component, mapping $512 \rightarrow p_H \times p_W$ where (p_H, p_W) depend on the level (Section 2.2.4). An unpatchify step finally produces the reconstructed component set $\bar{S}$ 74

- 2.22 Qualitative reconstructions at epoch 50 for four pretraining configurations differing only by the DWT decomposition level. For each level, the *top* row shows (left-to-right) the target image, the visible (unmasked) region, and the target high-frequency component for the red channel (fMoW-S2 band 2). The *bottom* row reports the corresponding model outputs: the reconstructed image, the previously masked area recovered by the model, and the reconstructed high-frequency component for the same spectral band. At levels 1–3 the encoder exhibits limited recovery of fine high-frequency structure and instead exploits low-frequency cues for approximation. At level 4, by contrast, the lower-resolution high-frequency bands become easier to model, and these coarse HF cues assist reconstruction of inner (higher-resolution) components. Overall, the figure highlights the increasing difficulty of restoring high-frequency content at finer scales and the stabilizing role of coarse HF signals during reconstruction. 81
- 2.23 Qualitative comparison of semantic segmentation task performance on Mados dataset across different SSL methods. From left to right the figure illustrates the NIR band (B8A) of the input image, the target segmentation map, and following the predicted segmentation maps respectively produced by MAE [2], SatMAE [3], SatMAE++ [4], SpectralGPT [5] and finally **WaveMAE**. The labels on the bottom indicate the dataset’s classes and the respective color assigned in the segmentation map. (Best viewed zoomed in for fine details.) 89
- 2.24 Qualitative comparison of semantic segmentation task performance on HLSBurnScars dataset across different SSL methods. From left to right the figure illustrates the NIR band (B8A) of the input image, the target segmentation map, and following the predicted segmentation maps respectively produced by MAE [2], SatMAE [3], SatMAE++ [4], SpectralGPT [5] and finally **WaveMAE**. White color is assigned to show the presence of a burn scar in the image. 90

2.25	Qualitative comparison of change-detection task performance on Spacenet7 (MUDS) dataset across different SSL methods. From left to right the figure illustrates the RGB image at timestep t_1 , the RGB image at timestep t_2 with $t_1 < t_2$, the target segmentation map, and following the predicted segmentation maps respectively produced by MAE [2], SatMAE [3], SatMAE++ [4], SpectralGPT [5] and finally WaveMAE . The white color in segmentation maps corresponds to areas where urban development occurred between the two RGB images.	91
A.1	The online frame-level VAD architecture. $f[t]$ is the frame at time t , x the output of the Reducer, NF the number of frames in input to the VST, $s[t]$ the anomaly classification score of frame $f[t]$	103
A.2	Performance comparison, changing the number of frames (NF) in input to the STMM (from 1 to 6).	106
A.3	Performance comparison, changing the number of LSTM cells. “LSTM (2 cells)” corresponds to “ $NF = 3$ ” in Fig. A.2.	107
A.4	Performance comparison, changing the VCL. The “8 frames” configuration corresponds to “ $NF = 3$ ” in Fig. A.2.	109

List of Tables

2.1	Ablation study on loss usage. The table shows the performance of the model using different combinations of loss functions. The combination with the highest performance is shown in bold and the second-highest performance is shown in red	52
2.2	Ablation study on positional encoding. The best performance is shown in bold and the second-highest performance is shown in red	54
2.3	Ablation study on Transformer MLP and MoE layer architectures. APC / SPC: active and sparse parameters count inside a MLP and MoE layers. Latency: average amount of seconds required to elaborate a batch of 8 images on a NVIDIA Quadro RTX6000. SM: Smart Merger. Model params: amount of sparse parameters of the full model architecture. Model FLOPs: floating point operations in a forward pass of the full model. The best performance is shown in bold and the second-highest performance is shown in red	55
2.4	Quantitative comparison with SOTA models on Sen2Ven μ s and OLI2MSI datasets. GiB, IT/s: gpu memory usage and iterations per second on inference. APC / SPC: active and sparse parameters count inside a MLP and MoE layers. FLOPs: floating point operations in a forward pass of the model. The best performance is shown in bold and the second-highest performance is shown in red	56

2.5	Ablation study on WaveMAE method applying an increasing level of decomposition of the DWT. Each model has been tested across five datasets. Results reported as mean $\pm$ standard deviation over 5 runs. Best result per column is bold , second best is <u>underlined</u> . Results are mean $\pm$ std computed over 5 independent runs with different seed. .	82
2.6	Intrinsic GPE metrics (GPE-pairs) on frozen encoders. Reported are mean cosine similarity for close/far pairs and Spearman correlation between cosine similarity and Δ_{GCD} (km).	84
2.7	Intrinsic GPE metrics (Geo-conditioned embeddings) on frozen encoders. Reported are mean cosine distances between anchor (A) and positives/negatives.	85
2.8	Ablation on the effect of GPE and $\mathcal{L}_{cmp}$ on downstream tasks. Results reported as mean $\pm$ standard deviation over 5 runs. Best result per column is bold , second best is <u>underlined</u> . Results are mean $\pm$ std computed over 5 independent runs with different seed.	85
2.9	Ablation on masking ratio applied to the sequence in input to the encoder during pretraining. Each model has been tested across five datasets. Results reported as mean $\pm$ standard deviation over 5 runs. Best result per column is bold , second best is <u>underlined</u> . Results are mean $\pm$ std computed over 5 independent runs with different seed. .	86
2.10	Ablation on pretrain effectiveness against a random initialization considering the full WaveMAE architecture. Each model has been tested across five datasets. Results reported as mean $\pm$ standard deviation over 5 runs. Best result per column is bold . Results are mean $\pm$ std computed over 5 independent runs with different seed.	86
2.11	Ablation on token spatial size, results shows that increasing the granularity of the input tokens leads to better performance overall. Each model has been tested across five datasets. Results reported as mean $\pm$ standard deviation over 5 runs. Best result per column is bold . Results are mean $\pm$ std computed over 5 independent runs with different seed.	87

2.12	Comparison of WaveMAE against previous state-of-the-art models using fMoW-S2 [6] as pretraining dataset for all methods. Each model has been tested across five datasets. Results reported as mean $\pm$ standard deviation over 5 runs. Best result per column is bold , second best is <u>underlined</u> . Results are mean $\pm$ std computed over 5 independent runs with different seed.	88
A.1	Performance comparison with and w/out short and long-term memory. Short-term: with ($NF = 3$) and w/out ($NF = 1$). Long-term: with (2 cells) and w/out (0 cells).	108
A.2	Benchmarks of VAD methods on the DoTA dataset. Both MOVAD models are trained with the best configuration of: VCL of 8, 3 LSTM cells and $NF = 4$. (Note: results with (*) and (**) are taken from DoTA [7] and [8] paper, respectively)	110
A.3	Detection accuracy (AUC) for individual accident categories. “*” non-ego anomaly categories. “†” if input resolution is 640×480 instead of 320×240 . N/A=Not Available. Bold and red values are the best and second-best results.	111

Introduction

*Happiness can be found even in the darkest of times,
if only one remembers to turn on the light.*

Global food security and environmental sustainability are among the greatest challenges of the 21st century. Population growth, climate change, and resource constraints demand more efficient crop production, for this reason precision agriculture has emerged as a key strategy to address these challenges by optimizing farm management at fine spatial scales. In the context of remote sensing, meaning the use of satellite or airborne imagery to observe the Earth's surface, provides indispensable support. Satellite imagery can capture large agricultural areas repeatedly and non-invasively, offering insights into crop health, soil properties, water stress, and yield variability without requiring exhaustive field surveys.

As Phiri et al. [9] note, Sentinel-2's improved spatial, spectral and temporal resolution was designed specifically to meet the needs of the agricultural community, and remote sensing now provides crucial contributions to large-scale agriculture monitoring. In other words, multispectral satellite data can serve as a tool for laborious manual crop surveys, enabling precision management of fertilizers, irrigation, and pest control on regional or even global scales. All of this is possible because information captured by remote sensors give access to plant physiology via spectral signatures. In fact, plants strongly absorb visible light (especially blue and red wavelengths) for photosynthesis, while reflecting strongly in the near-infrared (NIR) and due to cellular structure [10]. By measuring reflectance in specific bands, one can derive veg-

etation indices that correlate with biomass, chlorophyll content, or water stress. For example, the well-known Normalized Difference Vegetation Index (NDVI) is defined as

$$\text{NDVI} = \frac{\rho_{\text{NIR}} - \rho_{\text{Red}}}{\rho_{\text{NIR}} + \rho_{\text{Red}}}, \quad (1)$$

where ρ_{NIR} and ρ_{Red} denote the surface reflectance in the near-infrared and red bands, respectively. NDVI provides a normalized measure of vegetation greenness and is often used to estimate crop vigor and predict yields. Figure 1 shows an illustrative Sentinel-2 scene and the corresponding NDVI map, demonstrating how multispectral data highlights vegetated areas. By monitoring NDVI and related indices over time, one can detect nutrient deficiencies, water stress, or disease outbreaks early, guiding targeted interventions. In summary, the synergy of precision agriculture and satellite remote sensing arises because satellites like Sentinel-2 enable repeat, high-coverage monitoring of crop fields, transforming global agricultural management.



Figure 1: Example an RGB image (left) and derived NDVI index (right), illustrating how multispectral data can quantify vegetation health. Green color indicates a good vegetation health, while red color suggests bad vegetation health. [1]

Motivated by the importance of multispectral imagery in the remote sensing panorama, the research presented in this thesis focuses on enhancing representations

processed by deep neural networks. More in detail, this work addresses two key challenges in this applicative context: the inherently low spatial resolution of multispectral imagery and the scarcity of large-scale annotated datasets. To summarize, the main contributions presented in this thesis are:

- Swin2-MoSE, a novel Single-Image Super-Resolution (SISR) [11] method is proposed to solve the issue of low spatial resolution in multispectral remote sensing imagery.
- A new self-supervised method for pretraining geospatial image-processing deep learning models, which exploits the hierarchical structure of Discrete Wavelet decomposition to build richer feature and transfer them to specific downstream tasks, namely WaveMAE.

Single-Image Super-Resolution Single-Image Super-Resolution (SISR) aims to reconstruct a high-resolution (HR) image from a single low-resolution (LR) observation. The objective is to recover high-frequency spatial details that are lost due to degradation processes such as downsampling, sensor limitations, or transmission artifacts. Formally, given an LR image $I_{LR} \in \mathbb{R}^{H \times W}$, the task is to estimate its HR counterpart $I_{HR} \in \mathbb{R}^{sH \times sW}$, where s denotes the desired upscaling factor.

The SISR problem is inherently ill-posed, since multiple HR images can correspond to the same LR observation. As a consequence, early methods relied on handcrafted priors and interpolation-based approaches (e.g., bicubic upsampling), which often produced overly smooth results. The introduction of learning-based techniques—first through sparse coding and later through deep convolutional neural networks (CNNs)—has dramatically advanced the field by enabling the model to learn powerful image priors directly from data.

In recent years, deep learning-based methods have achieved remarkable performance improvements, driven by architectures such as residual networks, attention mechanisms, and, more recently, Vision Transformers (ViTs). These models can infer high-frequency textures and edges more accurately, leading to reconstructions that are both quantitatively and perceptually superior. Despite these advances, several challenges remain open: achieving a balance between pixel-wise accuracy (e.g.,

PSNR, SSIM) and perceptual quality, improving generalization to real-world degradations, and reducing computational cost for high-resolution inference.

SISR has widespread applications, including remote sensing, medical imaging, video surveillance, and digital photography enhancement. Its ability to enhance visual detail from limited data makes it an essential component in many computer vision and image restoration pipelines.

Masked Image Modelling and Self-supervised learning

Self-supervised learning (SSL) denotes a class of learning paradigms in which supervisory signals (labels) are derived automatically from the raw data, removing the need for manual annotations. By training on proxy (pretext) tasks that require reasoning about structure, geometry, or semantics, SSL methods induce deep networks, to learn representations transferable effectively to downstream tasks, even with high parameter count and limited labeled data, a situation often leading to the inability of generalizing data. Common pretext tasks include contrastive instance discrimination, rotation prediction, and reconstruction-based objectives. These encourage deep networks to capture invariant and semantically meaningful features on large scale unlabelled dataset.

Masked Image Modelling (MIM) is a reconstruction-based SSL pretext task tailored to images. In MIM, an input image is partitioned into patches and a subset $\mathcal{M}$ of patches is retained, the model is trained to predict the content of the masked patches from the visible context. Architecturally, MIM typically employs an encoder that processes visible patches and a lightweight decoder that reconstructs masked regions. The learning objective focuses on reconstruction over masked locations, for instance:

$$\mathcal{L}_{\text{MIM}} = \frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} \ell(\hat{p}_i, p_i),$$

where p_i and $\hat{p}_i$ are the ground-truth and reconstructed patches, and $\ell(\cdot, \cdot)$ denotes a suitable regression or perceptual loss.

MIM promotes learning of both local texture and long-range contextual dependencies, producing representations well-suited for dense prediction and restoration tasks. Its simplicity, scalability to large unlabelled datasets, and strong empirical

transferability make MIM an attractive pretraining strategy for downstream problems such as single-image super-resolution.

The thesis is structured in the following way:

- Chapter 1: provides a detailed overview of state-of-the-art approaches related to this dissertation's topics.
- Chapter 2: presents an in-depth analysis of the research findings. Specifically Section 2.1 regards the proposed architecture for the Single-Image Super-Resolution task, Section 2.2 includes all the details and studies on the novel Discrete Wavelet Transform powered self-supervised Masked Image Modelling method.
- Chapter 3: explains the conclusions drawn from this research and outlines potential future work.
- Chapter 4: lists the publications produced during the doctoral research.
- Appendix: contains other relevant work by the author that is not directly related to remote sensing.

Chapter 1

Background

All we have to decide is what to do with the time that is given us.

1.1 Vision Transformers

All work presented in this dissertation builds upon the Vision Transformer architecture and its variants. The Vision Transformer is an adaptation of the Transformer architecture [12] from Natural Language Processing domain to the Image Processing domain. Vision Transformers (ViT) transplant the Transformer encoder to vision by tokenizing an image into a sequence of fixed-size patches, projecting each patch to a token embedding, adding positional encodings to let encoder discriminate the position of the patches in the original image. Successively the resulting token sequence are processed with multi-head self-attention and feed-forward blocks [13]. By removing the explicit convolutional inductive bias, ViT relies on large-scale pretraining to learn useful spatial priors and has been shown to produce strong results on image classification and transfer benchmarks when trained at sufficient scale [13]. The full architecture can be seen in Fig. 1.1.

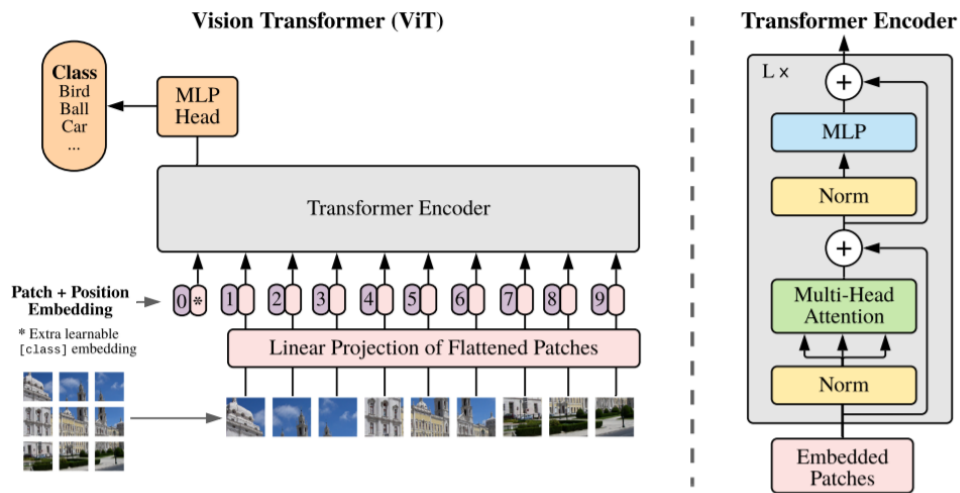


Figure 1.1: The structure of a Vision Transformer encoder. The figure on the right, illustrates the internal architecture of the deep network, highlighting the primary modules within each ViT block. Here, L denotes the architectural depth, indicating the frequency with which the transformer's blocks are consecutively reiterated.

1.1.1 The attention operation

The fundamental operation of Transformer models is scaled dot-product attention [12]. Given queries Q , keys K and values V , the attention output is

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)V,$$

where d is the dimensionality of the queries/keys. Multi-head attention splits the representation into several subspaces (heads) and concatenates independent attention outputs to allow the model to attend to different types of relationships in parallel [14]. This mechanism enables direct pairwise interactions between all tokens and is highly parallelizable, which partly explains the Transformer’s empirical success across modalities.

1.1.2 Swin Transformer

The Swin Transformer adapts the ViT design to a hierarchical feature representation and shifted window self-attention [15]. Instead of global attention over all tokens (quadratic in the number of patches), Swin computes self-attention within non-overlapping local windows and shifts the window partition between successive layers to allow cross-window connections, as shown in Fig. 1.3. The model also performs stage-wise patch merging to produce multi-scale feature maps similarly to CNN feature pyramids, the complete architecture can be seen in Fig. 1.2. These design choices reduce computational complexity with respect to image size while preserving the ability to model long-range context through successive windowing and hierarchical aggregation.

1.2 Single-Image Super-Resolution

As previously introduced, remote sensing has grown in significance due to its many uses, the most notable of which are in agriculture. Increased crop productivity and reduced resource waste are two potential outcomes of using multi-spectral satellite

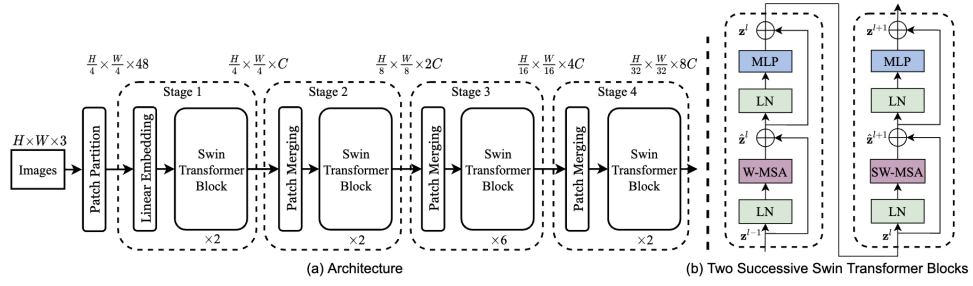


Figure 1.2: The architecture of a Swin-Transformer.

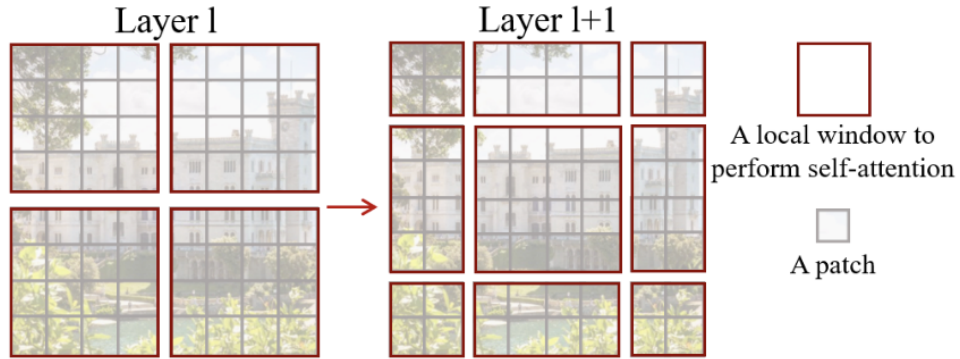


Figure 1.3: An illustration of the windowing mechanism implemented in Swin-Transformer. The attention is calculated within the non overlapping windows, alleviating the quadratic complexity requirements for the operation.

imagery for crop segmentation and classification, disease diagnosis, and autonomous irrigation control. However, satellites frequently deliver images with insufficient resolution for their intended uses, owing to optical and sensor technologies. I have been working on a framework that offers a fresh approach and surpasses the present state of the art, highlighting the critical significance of researching new and powerful remote sensing super resolution techniques.

The scientific community, have recently focused on image Super-Resolution (SR), which is the process of creating high-quality images from Low-Resolution (LR) ones [16–18]. Transforming low-resolution (LR) photographs that lack clarity or sharpness into high-resolution (HR) images that do have improved details is the main objective of sharpness restoration (SR). Classification and segmentation are only two of the many image processing and machine learning applications that could benefit from this enhanced picture detail.

The advancement of super-resolution techniques in the niche area of remote sensing imagery has received substantial research attention [19–22]. Fixing the naturally poor spatial resolution of UAV or satellite-obtained images was the initial focus of the literature. Today, this super-resolution problem is more important than ever before because of the abundance of satellites that can detect a wide range of supplementary spectral bands, including infrared and near-infrared spectra. The greater spectral range of frequencies per pixel in multi-spectral photography comes at the cost of even lower spatial resolution. As a matter of fact, due to financial and technological limitations, Earth observation satellites' spatial and spectral resolutions seldom reach the targeted high level. There are numerous applications that could gain advantages from having access to higher resolution images. For instance, land use and land cover classification, object detection, change detection, and resource management involving forests, water, and minerals. Because of this, I have been working on a framework to remedy the problem of low-resolution multispectral imagery in remote sensing by integrating the abundant low-resolution data sets with an approach that reliably recreates high-frequency details compared to earlier approaches. In addition, the lack of super-resolution methods in the field of remote sensing was addressed by developing a framework that is optimized for satellite data. This type of imagery differs signifi-

cantly in scale and semantics from the standard RGB imagery that most SR methods are trained on.

One important form of SR approach is multi-image super-resolution (MISR), and another is single-image super-resolution (SISR). According to Anger et al. [23], Lafenetre et al. [24] and Nguyen [25], MISR is used to create a high-resolution image (HR) by concatenating together multiple low-resolution (LR) images captured at various times or from different angles during satellite passes. In this thesis, the focus will be on SISR, a method that uses a single LR input to create the HR image. Several factors make SISR more attractive than MISR. The first benefit is a simplification of operations: processing a single image eliminates the requirement for the computationally costly alignment and registration of numerous images. The second advantage is that it is more versatile than MISR as it doesn't need a bunch of photos of the same location. Lastly, it's less likely to cause blurring or loss of clarity during interpolation, and it's quicker to implement as it only needs one image to be captured. MISR and SISR are both ill-posed problems due to the fact that each given LR image could correlate to an unlimited number of HR photos, for this reason SISR presents a more formidable challenge due to the fact that, by definition, the access is restricted to a single image.

In this section several important methods will be shown to highlight the state of the art in both natural imagery and remote sensing. When compared to natural imagery, remote sensing images are significantly different in terms of their acquisition method, spectral properties, and intended uses. Photos taken with regular cameras or other natural sources usually only record in the visible light spectrum (red, green, and blue) because that's what humans use for visual interpretation. Contrarily, sensors affixed to satellites or planes collect data across various spectral bands, such as the visible, infrared, and microwave ones, to provide remote sensing images. Differences in detail, scale, and data format are also caused by the fact that remote sensing images frequently have different spatial, spectral, temporal, and radiometric resolutions. These resolutions are determined by the type of sensor used and the mission objectives. Because of the unique challenges posed by sensor properties, atmospheric influences, and georeferencing, natural image analysis is not applicable to the inter-

pretation and processing of remote sensing images.

1.2.1 Super-Resolution for Remote Sensing Images via Local-Global Combined Network

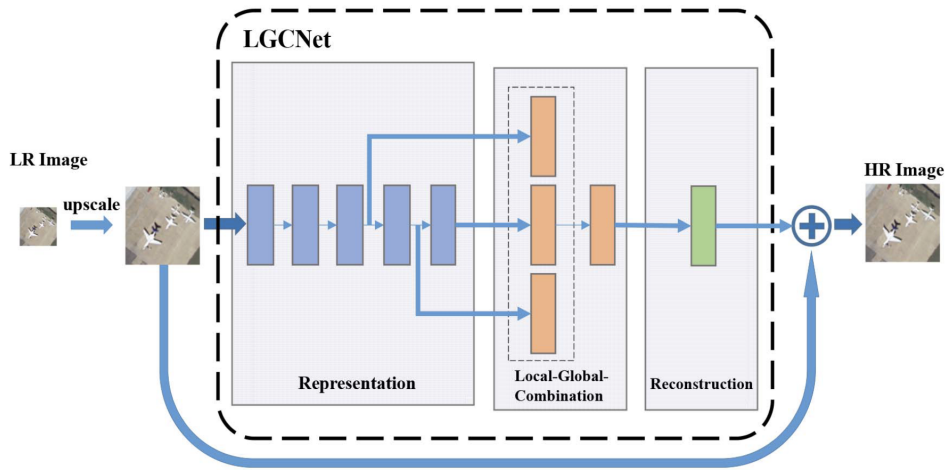


Figure 1.4: Architecture of the CNN based method LGCNet for remote sensing super resolution.

Applying traditional super-resolution methods that use IHS (Intensity, Hue, Saturation) transformations to very high resolution satellite data acquired through SPOT6/7 and Pleiades 1A/1B platforms, such as SRCNN [26] and VDSR [27], has shown encouraging results in earlier studies. Nevertheless, it is crucial to develop dedicated models for Remote Sensing Single-Image Super-Resolution (RS-SISR) [28] because of the different nature of remote sensing images. Among the first CNN approaches dedicated to RS-SISR is worth mentioning Local-Global Combined Network, introduced by Lei et al. [29], proposes a new network design to exploit both local-global feature for a better image representation. Multilevel information could be very useful for tasks that need to make images super-resolution, especially for remote sensing images. For this reason, as shown in Fig. 1.4, the first part of the network is composed

of a concatenation of different convolutional layers dedicated to learning a deep representation. At its core, multi-scale learning is performed in the second component of the network. By merging the convolutional outputs of various layers, a forked structure is typically used to execute local-global combination. Finally, for reconstruction, these merged representations are fed into one convolutional layer.

1.2.2 High Quality Remote Sensing Image Super-Resolution Using Deep Memory Connected Network

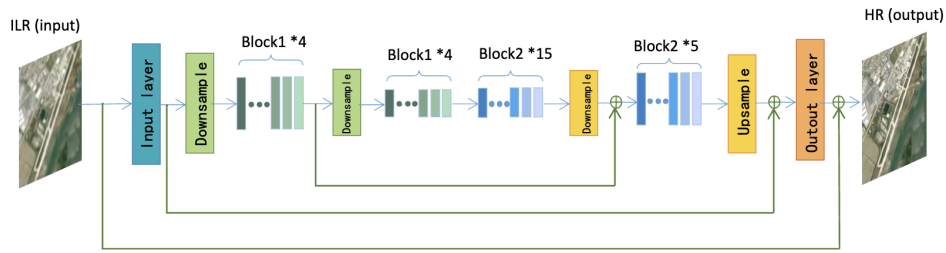


Figure 1.5: Architecture of the CNN based method DMCN for remote sensing super resolution.

In their study, Xu et al. [30] introduced Deep Memory Connected Network, a network that uses local and global memory connections to include picture information and environmental context into its architecture. As the previous work, the core idea is to mix the local and global features to build a richer representation that effectively fuses local details with global information in each image. The overall structure of DMCN is illustrated in Fig. 1.5. An input layer, a downsampling unit, an upsampling unit, and an output layer make up DMCN. The proposal for reducing the spatial size of the feature map in DMCN is an hourglass structure. A convolutional layer with stride=2 is included in every downsampling unit, which reduces the feature map by a factor of 2. Each upsampling unit is designed with an upscale factor of 2 to reconstruct the feature map. By implementing this hourglass structure, the architecture is able to drastically cut down on time complexity without sacrificing performance.

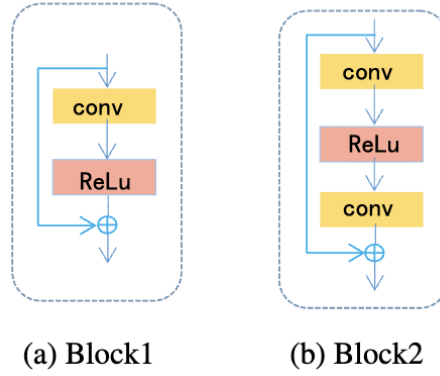


Figure 1.6: Detailed representation of the DMCN's building blocks.

The feature encoding is performed by two kind of convolutional blocks shown in Fig. 1.6, each used respectively in the downsampling unit and in the upsampling unit. By downsampling the feature maps, it's possible to access features at higher scale, in the upsampling unit these high-scale features are mixed to lower-scale features using skip connections.

1.2.3 Transformer-based restoration and SwinIR

Building upon the Swin transformer architecture, authors of SwinIR [31] introduced a novel Residual Swin Transformer Block for deep feature extraction exhibiting noteworthy performance for this task. The model couples a small convolutional front end with a stack of locally attentive transformer blocks and a concise reconstruction head, refer to Fig. 1.7 for a representation of the full architecture. This hybrid arrangement preserves the inductive bias of convolution for coarse structure while using attention to recover fine, content-dependent detail that standard convnets often blur.

Formally, given a degraded input image $I_{LQ} \in \mathbb{R}^{H \times W \times C_{in}}$, the network first computes a shallow embedding by a small convolutional operator:

$$F_0 = H_{SF}(I_{LQ}),$$

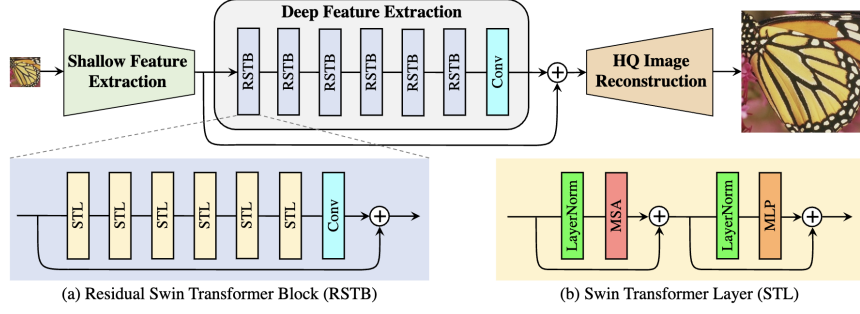


Figure 1.7: Schematic overview: shallow convolutional embedding, stacked residual Swin Transformer blocks with shifted-window attention, and a compact reconstruction head.

where H_{SF} is typically a 3×3 convolution mapping image space into a feature manifold of dimension C . A deep extractor H_{DF} , implemented as K residual Swin Transformer blocks (RSTBs), refines the shallow features:

$$F_{\text{DF}} = H_{\text{DF}}(F_0), \quad F_{\text{DF}} = H_{\text{CONV}}(F_K),$$

with $F_i = H_{\text{RSTB}_i}(F_{i-1})$ for $i = 1, \dots, K$ and H_{CONV} a final convolutional refinement. The reconstruction module combines the shallow and deep streams through a long skip connection and produces the restored image:

$$I_{\text{HQ}} = H_{\text{REC}}(F_0 + F_{\text{DF}}).$$

For super-resolution the reconstruction head uses a sub-pixel upsampling stage, for denoising or artifact removal a single convolutional output suffices. Training is commonly performed with a robust pixel-wise objective, typically the L_1 loss:

$$\mathcal{L} = \|I_{\text{HQ}} - I_{\text{GT}}\|_1.$$

Two design elements are central to SwinIR's behaviour. First, self-attention is confined to small, non-overlapping windows rather than applied globally. This choice keeps computational and memory cost near-linear in the pixel count: full self-attention

scales quadratically with the number of tokens, whereas windowed attention reduces the dominant cost to the product of image size and the square of the window side. To avoid the isolation typical of fixed-window processing, the window partition is shifted between consecutive layers: the shift allows adjacent windows to exchange information and thus extends the effective receptive field without incurring global attention cost. Second, residual grouping—stacking several transformer layers inside a residual block and following them with a convolutional refinement—preserves gradient flow and restores translation-friendly inductive bias that pure attention may weaken. Together these choices enable the model to act as a content-adaptive filter: attention weights effectively implement spatially varying convolution kernels that depend on local context.

Empirically, SwinIR attains state-of-the-art or near state-of-the-art peak-signal metrics on classical color restoration benchmarks while often using fewer parameters than some earlier convolutional baselines. The paper reports consistent gains across multiple restoration variants (standard super-resolution, lightweight SR, real-world settings and denoising), indicating that the hybrid architecture generalizes beyond a single narrow task. At the same time, applying the design directly to remote-sensing data requires several practical adaptations. Multispectral sensors produce much richer spectral content than RGB images, tile sizes are typically larger, and radiometric scales can differ markedly from consumer photography. Therefore, the initial three-channel embedding should be replaced by a band-aware projection that mixes spectral channels efficiently, and tiling strategies (with overlaps selected relative to window size) must be used to prevent border artifacts at very large image scales. Normalization and radiometric preconditioning informed by sensor calibration are also advisable.

1.2.4 SwinFIR: Revisiting the SwinIR with Fast Fourier Convolution and Improved Training for Image Super-Resolution

SwinFIR revisits the SwinIR design and addresses a central limitation of windowed attention: the slow accumulation of truly global context in early layers due to the windowing mechanism. The authors introduce a Fast Fourier Convolution (FFC) branch

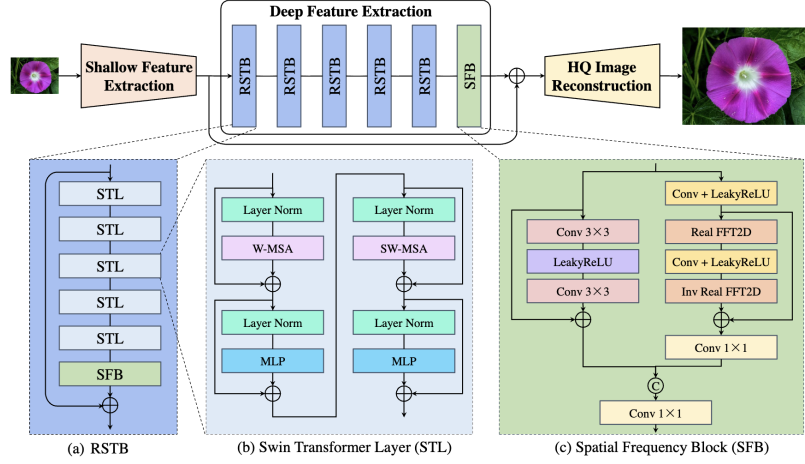


Figure 1.8: In this figure the SwinFIR pipeline is shown (shallow conv $\rightarrow$ RSTB stack $\rightarrow$ reconstruction) and the SwinFIR augmentation (a compact FFC branch fused into the spatial stream).

that provides an image-wide receptive field by performing lightweight learned operations in the frequency domain and fusing them with the local, windowed features inherited from the Swin backbone. In compact mathematical form the FFC branch may be seen as

$$\text{FFC}(X) = \text{IFFT}(M(\omega) \odot \text{FFT}(X)) + G(X),$$

where X are intermediate spatial features, FFT/IFFT denote the discrete Fourier transform pair, $M(\omega)$ is a learned modulation over frequency coordinates ω , $\odot$ is pointwise multiplication in frequency space, and G is a small local operator that preserves or refines locality. Because FFT-based mixing attains an image-wide receptive field at modest cost (FFT complexity scales like $O(N \log N)$), the FFC branch supplies immediate global cues, slow gradients, extended linear structures, repetitive patterns—that the windowed Swin path only acquires gradually through shifted-window propagation. SwinFIR fuses the FFC output back into the spatial stream inside a Spatial Frequency Block, keeping the fusion lightweight so that practical inference cost

risers little compared to the SwinIR baseline. The authors also show that modest but careful changes to the training recipe (stronger augmentation, pragmatic pretraining and simple ensemble techniques at test time) compound the architectural benefits into consistent metric gains on standard super-resolution benchmarks [32].

Viewed together, SwinIR and SwinFIR offer complementary design lessons. SwinIR supplies an efficient local-attention backbone with an explicit low-frequency skip that stabilises training and isolates high-frequency recovery, while SwinFIR demonstrates that grafting a compact, frequency-domain global mixer onto that backbone remedies the slow accrual of global context and improves recovery of large-scale structures. The full architecture of SwinFIR is shown in Fig. 1.8.

1.2.5 Swin2SR: SwinV2 for compressed-image super-resolution and restoration

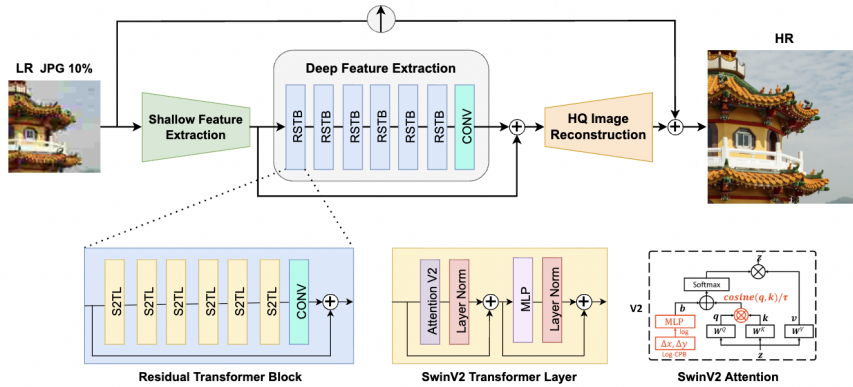


Figure 1.9: Swin2SR architecture: SwinV2-based Residual Transformer Blocks (RSTB_V2) in the deep extractor, long-skip shallow path, pixel-shuffle reconstruction, and the auxiliary/high-frequency losses used during training.

Swin2SR adapts the SwinV2 transformer to the problem of super-resolution and restoration of compressed images, with a design and training recipe that explicitly target the practical problems that arise when inputs contain heavy codec artifacts (JPEG etc.). The methodology demonstrates that modest but well-chosen changes at

the attention, normalization and loss-function level yield a model that trains more stably, generalises better to higher-resolution inference, and converges faster than the original SwinIR instantiation [33]. Like SwinIR, Swin2SR follows a three-stage decomposition: a shallow convolutional embedding, a deep feature extractor formed by a stack of residual transformer blocks, and a compact reconstruction head that produces the high-resolution output (pixel-shuffle when upscaling), refer to Fig. 1.9. Denoting the low-quality input by $I_{\text{LQ}} \in \mathbb{R}^{H \times W \times C}$, the shallow embedding is

$$F_0 = H_{\text{SF}}(I_{\text{LQ}}),$$

and the deep extractor (now composed of Residual SwinV2 Transformer Blocks, RSTB_V2) refines these features:

$$F_{\text{DF}} = H_{\text{DF}}(F_0), \quad F_{\text{DF}} = H_{\text{CONV}}(F_K),$$

where each block contains several SwinV2 Transformer layers (S2TL). The reconstruction module fuses F_0 and F_{DF} through a long skip and yields the restored image

$$I_{\text{HQ}} = H_{\text{REC}}(F_0 + F_{\text{DF}}).$$

The most important novelties introduced by Swin2SR are the post-layer normalization, a scaled cosine attention and a continuous, log-spaced relative position bias. In the first case, SwinV2 replaces the pre-norm configuration by post-norm, which empirically reduces feature variance in deep layers and improves numerical stability when scaling depth and capacity. This change helps avoid exploding/vanishing internal statistics that otherwise complicate long training runs or large models. Secondly, the dot-product similarity in conventional attention is replaced by a cosine-similarity kernel scaled by a learnable per-head scalar. Concretely, for queries q , keys k and values v it's possible to write the S2 attention weights as

$$\text{Attn}(q, k, v) = \text{softmax} \left(\tau \cdot \frac{\langle q, k \rangle}{\|q\| \|k\|} + B \right) v,$$

where $\tau > 0$ is a learnable scale (not shared across heads/layers) and B denotes the relative positional bias term. Normalising by vector norms reduces the domination of

a few large dot products and produces a more balanced attention distribution across pixels, which in turn helps stability and generalisation. Finally, the discrete table of relative biases used in the original Swin is replaced by a continuous formulation that can be evaluated at arbitrary window resolutions, this mitigates the so-called resolution gap between pre-training at one size and inference at a larger size because the model can generalise position bias to unseen spatial scales. Losses designed for compression robustness. Swin2SR augments the simple reconstruction loss with auxiliary terms crafted for compressed inputs. Two of these are particularly relevant and easy to state. The auxiliary (downsample) consistency loss enforces that the downsampled prediction matches the downsampled ground truth:

$$\mathcal{L}_{\text{aux}} = \|\mathcal{D}(I_{\text{pred}}) - \mathcal{D}(I_{\text{GT}})\|_1,$$

where $\mathcal{D}$ denotes a downsampling operator (for instance bicubic with the appropriate scale). The high-frequency (HF) loss focuses supervision on details by extracting a high-pass component $H(\cdot)$ (e.g., via Laplacian or guided high-pass filtering) and penalising its mismatch:

$$\mathcal{L}_{\text{HF}} = \|H(I_{\text{pred}}) - H(I_{\text{GT}})\|_1.$$

The final training objective is a weighted combination of the main reconstruction loss and these auxiliary terms:

$$\mathcal{L} = \mathcal{L}_{\ell_1} + \lambda_{\text{aux}}\mathcal{L}_{\text{aux}} + \lambda_{\text{HF}}\mathcal{L}_{\text{HF}}.$$

These terms encourage the network to avoid producing artifacts that survive a coarse downsampling and to prioritise faithful recovery of high-frequency content that codecs tend to obliterate. Swin2SR’s advantage over earlier methods such as SwinIR is not the result of a single radical change but rather of a careful rebalancing of algorithmic, numerical and optimisation elements that together make the model both easier to train and better suited to degraded, compressed inputs. At the numerical level, replacing dot-product attention with a scaled cosine similarity and switching to post-layer normalization markedly reduce pathological variance in deep features. Normalising the similarity scores in this way prevents a few large dot products from dominating

the softmax, which produces a more balanced distribution of attention weights and reduces sensitivity to learning-rate schedules and initialization. In practice, this stability means models can be trained deeper or with larger capacity without extensive manual tuning, and they converge more reliably across different datasets. A second, related benefit concerns generalisation across resolutions. The continuous, log-spaced formulation of relative position bias used in this work mitigates the resolution gap that commonly afflicts transformer backbones: models pre-trained or tuned at one patch size often suffer when evaluated at larger tiles because discrete, table-based biases do not extrapolate. By parameterising positional bias in a continuous manner the model can evaluate relative positions at arbitrary scales, which translates into more graceful behaviour when a pre-trained backbone is applied to higher-resolution satellite tiles or to inference regimes that differ from the training crops. From a systems perspective the combined effect of these changes is faster and more robust training. The stability improvements reduce the need for brittle learning-rate schedules, long warmups, or extensive ad-hoc normalisation, the specialised losses steer optimisation toward solutions that generalise to compressed inputs and the architecture retains the windowed attention that preserves near-linear memory scaling. Finally, Swin2SR offers a foundation that is both dependable in training and adaptable at inference, making it an attractive baseline to build upon or to combine with complementary ideas such as frequency-domain mixers or spectral-aware preconditioning [31–33].

1.2.6 Dual Aggregation Transformer (DAT) for Image Super-Resolution

The Dual Aggregation Transformer (DAT) [34] represents a significant step forward in transformer-based image restoration by explicitly integrating attention along both spatial and channel dimensions. Unlike previous architectures that rely primarily on spatial self-attention (SSA), DAT argues that strong cross-channel correlations remain underexploited. For natural and remote-sensing imagery alike, this assumption is justified: many textures or materials manifest distinctive patterns in the joint spatial–spectral domain. DAT’s design is therefore motivated by the need to aggregate global dependencies both within spatial neighborhoods and across feature channels, achieving a richer representation that enhances reconstruction quality and stability.

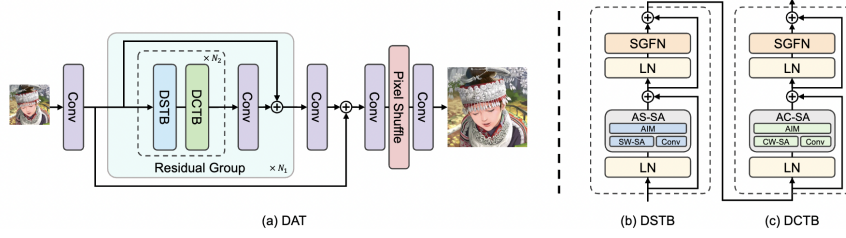


Figure 1.10: Overview of the Dual Aggregation Transformer (DAT) architecture. Each block alternates between Spatial Self-Attention (SSA) and Channel Self-Attention (CSA). The Adaptive Interaction Module (AIM) fuses outputs from both domains, while the Spatial-Gate Feed-Forward Network (SGFN) enhances local detail. The architecture preserves a long skip connection from shallow to deep features.

The core component of DAT is the Dual Aggregation Transformer Block (DATB), which alternates between two complementary forms of attention: Spatial Self-Attention (SSA) and Channel Self-Attention (CSA). Given an input feature tensor $X_t \in \mathbb{R}^{H \times W \times C}$, each block performs attention either across spatial tokens or across channels, depending on its position in the stack. Formally, the block operation can be expressed as

$$X_{t+1} = \text{SGFN}\left(\text{AIM}(\text{Attn}(X_t)) + X_t\right),$$

where $\text{Attn}(\cdot)$ denotes either SSA or CSA, $\text{AIM}(\cdot)$ is an Adaptive Interaction Module that merges and rebalances features from both attention domains, and $\text{SGFN}(\cdot)$ denotes a Spatial-Gate Feed-Forward Network (SGFN) introducing spatially adaptive filtering after the attention step. The Spatial Self-Attention (SSA) module focuses on aggregating global context across spatial positions within each feature map. It is designed to capture structural dependencies such as edges, repetitive patterns, and spatial textures that determine visual sharpness. In contrast, the Channel Self-Attention (CSA) module operates transversely across feature channels, modelling inter-channel or inter-spectral correlations. By alternating these two types of attention across the network depth, DAT progressively enriches the feature representations and allows spatial and channel cues to inform one another through residual feedback. The Adaptive Interaction Module (AIM) plays a central role in maintaining coherence between

these two attention dimensions. After each attention operation, AIM fuses the features from SSA and CSA branches through a learned reweighting mechanism:

$$Y = \alpha \cdot X_{\text{SSA}} + (1 - \alpha) \cdot X_{\text{CSA}},$$

where the coefficient α is dynamically predicted from the feature statistics. This fusion allows the model to self-adjust the relative importance of spatial versus channel cues depending on the input content, for instance, favouring channel cues in spectrally rich regions or spatial cues in highly textured areas. Finally, the Spatial-Gate Feed-Forward Network (SGFN) extends the conventional position-wise feed-forward network by introducing spatial gates that act as attention filters over the feature maps. In contrast to the pointwise MLPs in traditional transformers, SGFN explicitly modulates non-linear transformations according to spatial context, improving the ability to refine edges and local textures. The complete DAT architecture follows a hierarchical encoder–decoder design similar to SwinIR, shown in Fig. 1.10, consisting of an initial shallow convolutional embedding, a deep sequence of DAT blocks, and a reconstruction module with pixel-shuffle upsampling. A long skip connection between shallow and deep features ensures stable convergence and retention of low-frequency content.

In essence, DAT offers a unified perspective on spatial structure and spectral coherence by alternating and integrating both forms of attention: the model achieves a joint optimisation of geometric accuracy and radiometric consistency, a prerequisite for reliable downstream analysis such as classification, segmentation, or change detection.

1.3 Masked Image Modelling

Contemporary visual representation learning increasingly relies on methods that presume abundant labeled data due to larger and deeper models which require high amount of data to train. For large collections of remote sensing images, annotation is costly and often impractical: this motivates learning useful feature extractors from raw measurements alone. In this thesis I investigate the directions masked image

modelling as a Self-Supervised Learning (SSL) pretraining method to refrain from large labelled dataset and obtain better performance in several specific tasks. In this family of methods it's necessary to define surrogate tasks that forces an encoder network to capture structure in the data without human-provided labels, after which the pretrained encoder can be adapted to downstream tasks (super-resolution, classification, segmentation, regression) by finetuning in a supervised fashion.

The remainder of this introduction explains the fundamental principles behind self-supervised learning, describes the masked image modeling paradigm that has recently proven effective for vision transformers, and clarifies how these ideas apply to multispectral remote sensing.

The canonical SSL workflow splits into three components: an encoder f_θ that maps an input image x to a representation $z = f_\theta(x)$, a lightweight projection or prediction head g_ϕ optionally used during pretraining, and a surrogate loss $\mathcal{L}_{\text{pre}}$ that encourages useful invariances or predictive structure. After pretraining, the projection head is typically discarded and the encoder f_θ is transferred to downstream tasks.

Contrastive methods create positive and negative pairs and train the encoder to pull positives together while pushing negatives apart. Let x be an image and $t(\cdot)$ a stochastic data augmentation. Form two correlated views $x_i = t_i(x)$ and $x_j = t_j(x)$. Passing these through encoder and projector yields

$$v_i = g_\phi(f_\theta(x_i)), \quad v_j = g_\phi(f_\theta(x_j)).$$

A widely used contrastive loss is the InfoNCE objective: for a positive pair (i, j) and a set $\mathcal{N}$ of negatives,

$$\mathcal{L}_{\text{NCE}}(i, j) = -\log \frac{\exp(\text{sim}(v_i, v_j)/\tau)}{\exp(\text{sim}(v_i, v_j)/\tau) + \sum_{k \in \mathcal{N}} \exp(\text{sim}(v_i, v_k)/\tau)}, \quad (1.1)$$

where $\text{sim}(a, b)$ is typically cosine similarity and $\tau > 0$ a temperature parameter. Methods in this family include SimCLR [35] and momentum contrast approaches that enlarge the set of negatives via memory banks or momentum-updated encoders.

Other SSL variants avoid explicit negative sampling. Bootstrap methods train a predictor network to match representations produced by a slowly evolving target net-

work, these methods can learn meaningful features without contrastive discrimination. Architectures such as BYOL [36] and DINO [37] apply careful design choices (predictor, stop-gradient, target update rule) to prevent collapse and stabilize training. Evaluation commonly uses transfer learning protocols: linear probing (training a linear classifier on frozen features) and full fine-tuning on downstream labelled data.

Masked image modelling (MIM) transfers the classical idea of masked language modelling from text to images. The general recipe is to hide (mask) parts of the input and force the model to predict the missing content from the visible context. Two central design choices distinguish method variants: the representation used for prediction (pixel values, discretized visual tokens, or latent features) and the architectural partition between encoder and lightweight decoder.

Formally, let x be an image partitioned into units (patches or pixels) $x = [x_1, \dots, x_N]$. A binary mask $m \in \{0, 1\}^N$ selects visible positions with $m_p = 1$ and masked positions with $m_p = 0$. The encoder f_θ receives only the visible subset $x_{\text{vis}} = \{x_p : m_p = 1\}$ (often with positional encodings), producing latent codes $z = f_\theta(x_{\text{vis}})$. A small decoder h_ψ reconstructs targets $\hat{t} = h_\psi(z)$ defined on the masked positions. The training objective minimizes a reconstruction loss over masked tokens:

$$\mathcal{L}_{\text{MIM}} = \mathbb{E}_{x,m} \left[\sum_{p:m_p=0} \ell(t_p, \hat{t}_p) \right], \quad (1.2)$$

where t_p is the chosen target for position p and ℓ is a per-position loss (mean-squared error, L_1 , or cross-entropy against discrete codebook indices). When targets are raw pixels and ℓ is squared error, the method resembles classical inpainting. When targets are discrete visual tokens (obtained from a discrete tokenizer) the loss becomes categorical cross-entropy, this was the case in early BEiT-style [38] approaches.

A conceptual advantage of MIM is the locality of the prediction task: the model must learn correlations across space (and, for multispectral inputs, across bands) that are predictive of missing content. This pressure encourages encoders to internalize mid- and long-range dependencies, efficient spatial priors, and, when targets are spectral, cross-band relationships.

Contrastive SSL and MIM emphasise different aspects of representation learning. Contrastive objectives promote invariance to augmentations by collapsing mul-

multiple augmented views of the same scene onto similar embeddings, they thereby bias representations toward global semantic content and robustness to nuisances. MIM, in contrast, trains the model to perform conditional reconstruction, which emphasises the model’s ability to model local and structural correlations. Formally, contrastive training attempts to make the representation z invariant under a family of transformations $\mathcal{T}$:

$$z(t(x)) \approx z(x), \quad t \in \mathcal{T},$$

whereas MIM ensures that the conditional distribution $p(x_{\text{masked}} | x_{\text{vis}})$ is predictable from the encoder’s latent space. In practice, these pressures are complementary: contrastive methods supply global grouping cues, while MIM supplies dense, local structure [39].

Representations are typically assessed by a mix of intrinsic and transfer evaluations. Intrinsic evaluations examine the geometry of the representation space (clusterability, nearest-neighbour retrieval) and robustness to perturbations. Transfer evaluations measure downstream performance on tasks of interest showing the robustness and generalizability of the learned representation.

1.3.1 Masked Autoencoders Are Scalable Vision Learners

Masked Autoencoders (MAE) propose a simple, scalable method for self-supervised pretraining that pivots on three compact ideas:

1. remove a large fraction of the input tokens;
2. let a lightweight encoder see only the remaining visible tokens;
3. train a small decoder to reconstruct the removed tokens from the encoder’s latent representation.

This asymmetric design places most of the modelling burden on a decoder that is intentionally small, while keeping encoder cost low and focused on forming a concise, transferable representation [40].

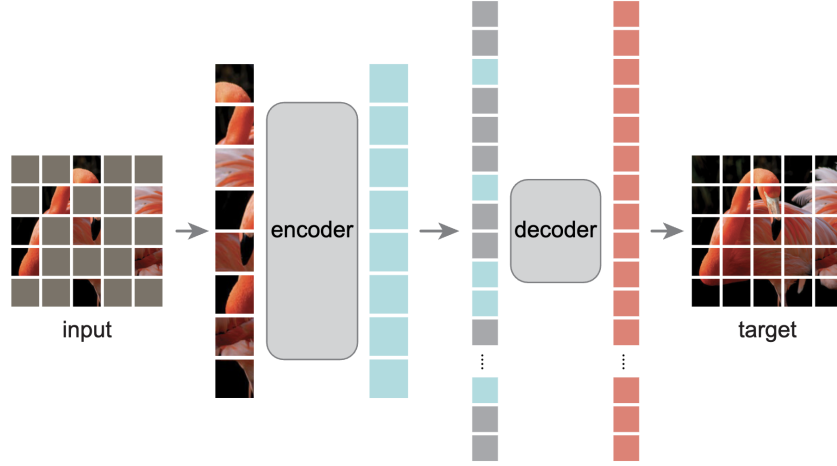


Figure 1.11: Illustration of the MAE workflow: random masking of patches, encoder processes visible patches only, small decoder reconstructs masked patches.

Formally, let an image be partitioned into N non-overlapping patches $x = [x_1, \dots, x_N]$. A random binary mask $m \in \{0, 1\}^N$ selects the visible set

$$\mathcal{V} = \{p \mid m_p = 1\}, \quad |\mathcal{V}| = (1 - r)N,$$

where $r \in (0, 1)$ is the masking ratio (the MAE paper typically uses $r \approx 0.75$). The encoder E_θ processes only the visible patches (with positional encodings) and yields latent tokens

$$z = E_\theta(\{x_p\}_{p \in \mathcal{V}}).$$

A decoder D_ψ receives z , together with simple mask tokens and positional information for the masked positions, and predicts targets for the masked patches. When targets are raw pixels and the per-patch loss is mean-squared error, the objective is

$$\mathcal{L}_{\text{MAE}} = \mathbb{E}_{x, m} \left[\sum_{p: m_p=0} \|\hat{x}_p - x_p\|_2^2 \right], \quad \hat{x}_{\text{masked}} = D_\psi(z). \quad (1.3)$$

Two implementation details are central. First, the encoder is a standard vision transformer (ViT) that never sees masked tokens, this means its self-attention cost scales

with the square of the number of visible tokens, giving a substantial computational saving when r is large. If attention cost on N tokens is $\mathcal{O}(N^2)$, then with masking the encoder cost behaves like

$$\mathcal{C}_{\text{enc}}(r) \propto ((1-r)N)^2 = (1-r)^2 N^2,$$

so a masking ratio of $r = 0.75$ reduces quadratic attention cost by a factor $(0.25)^2 = 0.0625$ relative to processing all tokens. The full architecture is illustrated in Fig. 1.11.

Empirically, MAE-style pretraining yields representations that transfer effectively across tasks: linear-probe and fine-tuning studies in the original work show improved sample efficiency and competitive downstream performance compared to other self-supervised recipes when applied to standard vision benchmarks [40]. Two intuitive reasons induce this behaviour. The reconstruction objective forces the encoder to capture both local texture and mid-range structure that are predictive of masked content, the high masking ratio prevents trivial identity shortcuts and encourages the encoder to form robust global summaries. Moreover, because the encoder processes a fraction of tokens, large-scale pretraining becomes more affordable, enabling training on larger corpora which in turn improves transfer.

In short, MAE provides a computationally efficient and conceptually clear route to dense, context-aware pretraining. Its asymmetric encoder–decoder split and the use of large masking ratios are the key design choices that permit scaling while encouraging representations that encode informative spatial structure [40].

1.3.2 SatMAE: Pre-training Transformers for Temporal and Multi-Spectral Satellite Imagery

SatMAE [3] adapts the masked autoencoding framework [40] to the unique characteristics of multispectral satellite imagery, where each observation comprises multiple spectral bands capturing different wavelengths of light. While the original SatMAE also accounts for temporal sequences, the focus will be in its ability to learn rich multispectral representations that preserve both spatial and spectral structure. The approach builds on the observation that spectral redundancy and physical correlations

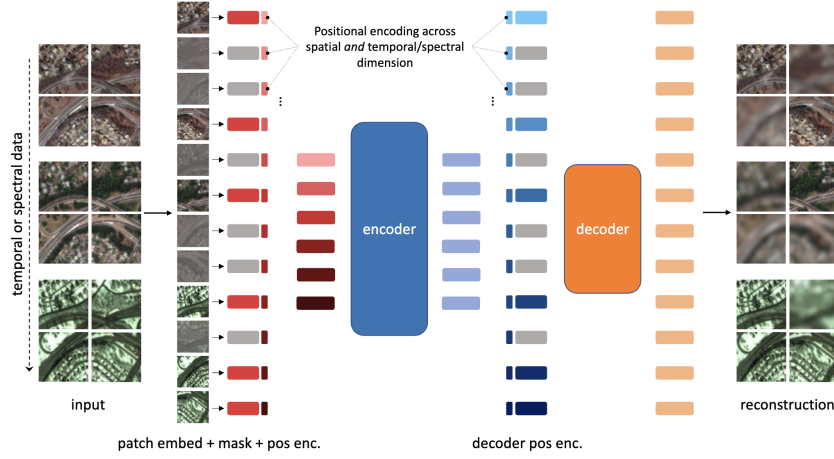


Figure 1.12: Overview of the SatMAE framework for multispectral imagery. The encoder processes a subset of visible spatial patches, while a lightweight decoder reconstructs masked patches across all spectral channels.

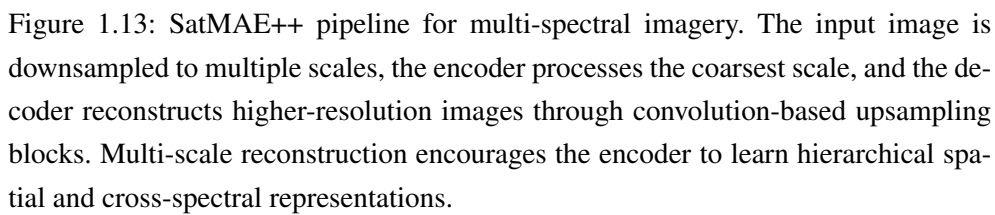
among bands, reflecting surface material properties and atmospheric conditions, can be exploited as an intrinsic supervisory signal for representation learning. Formally, let a multispectral image be $x \in \mathbb{R}^{H \times W \times B}$, where B is the number of spectral bands. The bands are partitioned into G groups, $\mathcal{B} = \{B_1, B_2, \dots, B_G\}$, based on wavelength proximity or sensor characteristics. The image is also divided into N spatial patches. During pretraining, a random mask selects a subset of patches and band groups to hide, producing a visible set $\mathcal{V}$ and masked set $\mathcal{M}$. Each visible patch and its bands are projected via a spectral-aware projection function:

$$z_i = E_\theta \left(\text{proj}(x_i^\mathcal{V}) + e_i^{\text{pos}} + e^{\text{spec}} \right), \quad i \in \mathcal{V},$$

where e_i^{pos} encodes spatial location and e^{spec} encodes spectral group identity. A lightweight decoder reconstructs the masked patches and band groups:

$$\hat{x}_\mathcal{M} = D_\psi(z, m).$$

1.3.3 Rethinking Transformers Pre-training for Multi-Spectral Satellite Imagery



In this work the architecture SatMAE++ was introduced [4]. SatMAE++ extends the SatMAE framework by explicitly leveraging multi-scale information present in multi-spectral satellite imagery. While SatMAE focuses on cross-spectral represen-

tation learning with optional temporal sequences, SatMAE++ introduces a scalable multi-resolution reconstruction approach that captures hierarchical spatial features. This design addresses a key limitation of previous methods: multi-spectral images often contain bands with different Ground Sample Distances (GSDs), and objects of interest can appear at various spatial scales within a single scene. Standard position encodings alone are insufficient to capture this variability, for this reason SatMAE++ augments masked pretraining with multi-scale supervision.

The input multi-spectral image $x \in \mathbb{R}^{H \times W \times B}$ is first downsampled to generate a pyramid of resolutions $\{x^{(s)}\}_{s=1}^S$, where $s = 1$ corresponds to the coarsest scale. Following the MAE paradigm, patches are extracted from the lowest-resolution image and a random subset is masked. Each visible patch is projected via a spectral-aware projection $\text{proj}(\cdot)$ and enriched with standard sinusoidal positional encodings:

$$z_i = E_\theta \left(\text{proj}(x_i) + e_i^{\text{pos}} \right), \quad i \in \mathcal{V}.$$

The encoder E_θ thus focuses on learning compact cross-spectral features at the coarsest scale, while a lightweight decoder reconstructs the masked patches. Crucially, SatMAE++ employs convolution-based upsampling blocks to reconstruct higher-resolution images from the encoded coarsest-scale representation. This progressive upsampling allows the model to produce feature maps at multiple scales, enforcing multi-scale consistency and encouraging the encoder to capture both global context and fine-grained spatial details. The pretraining loss is formulated as a weighted combination of reconstruction errors across scales:

$$\mathcal{L}_{\text{SatMAE++}} = \sum_{s=1}^S \alpha_s \mathbb{E}_{x,m} \left[\sum_{i \in \mathcal{M}^{(s)}} \ell(x_i^{(s)}, \hat{x}_i^{(s)}) \right], \quad (1.4)$$

where $\mathcal{M}^{(s)}$ denotes masked patches at scale s , $\ell(\cdot)$ is typically ℓ_1 or ℓ_2 loss, and α_s balances the contribution of each scale. By reconstructing multiple scales simultaneously, the model learns representations that are robust to variations in object size, sensor GSD, and spatial resolution.

Band-specific characteristics of multi-spectral imagery are preserved through spectral embeddings incorporated into each patch, allowing the model to distinguish

bands with different radiometric properties. This combination of spectral awareness and multi-scale reconstruction enables SatMAE++ to outperform prior methods on a variety of downstream tasks, including land-cover classification, semantic segmentation, and multi-label classification. For instance, on the BigEarthNet dataset, SatMAE++ achieves a 2.5% mAP gain in multi-label classification and a 3.6% absolute improvement in land-cover classification compared to SatMAE.

This approach is extensible to additional scales and modalities, making it highly effective for multi-spectral remote sensing tasks where objects vary in size and spectral composition.

1.3.4 SpectralGPT: Spectral Remote Sensing Foundation Model

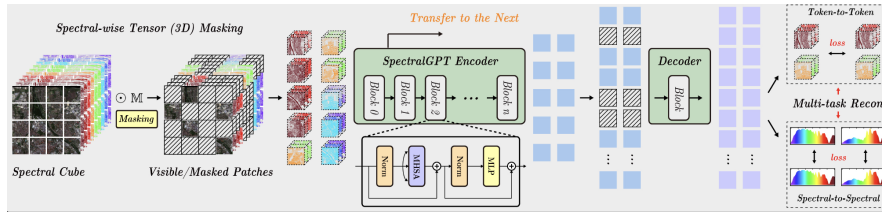


Figure 1.14: Overview of the SpectralGPT architecture. The model employs a 3D to-kenization strategy to process spectral RS images, capturing both spatial and spectral information through its encoder-decoder framework.

SpectralGPT [5] is another pioneering foundation model designed specifically for spectral remote sensing (RS) data.

The core contributions of SpectralGPT can be delineated as follows:

1. **Purpose-Built Architecture for Spectral Data:** In contrast to mere adaptation of RGB-centric models, SpectralGPT is architected from the ground up to handle the spectral modality, leveraging the full information content of the data for geoscience applications.
2. **3D Generative Pretraining:** The utilization of a 3D GPT and a 3D token generation mechanism is crucial. This mechanism ensures comprehensive spatial-

spectral coupling, enabling the model to learn representations that simultaneously capture intricate spatial features and the physical properties embedded in the spectral bands. This is a substantive departure from 2D methods that often treat spectral information as an independent channel or reduce its dimensionality prematurely.

3. **Adaptive and Progressive Training:** The model is designed to fully leverage extensive RS big data by accommodating input images with varying sizes, resolutions, time series, and regions through a progressive training strategy. This architectural flexibility addresses a persistent heterogeneity challenge in real-world RS datasets, where data collection parameters are rarely uniform.

SpectralGPT employs a 3D tokenization strategy to process spectral RS images. The input image, represented as a tensor of dimensions $H \times W \times B$ (height, width, and number of spectral bands), is divided into non-overlapping patches. Each patch is then flattened and projected into a 3D token, incorporating spatial and spectral information. This 3D tokenization allows the model to learn spatial-spectral dependencies effectively. The model architecture, similarly to previously seen approaches, consists of an encoder-decoder framework as visible in Fig. 1.14. The encoder processes the 3D tokens to extract features, while the decoder reconstructs the original input from these features. The total reconstruction loss of the previously masked tokens is composed by a token-to-token Mean Squared Error (MSE) loss and a spectral-to-spectral MSE loss on to balance the spectral coherence:

$$\begin{aligned}
 \mathcal{L} &= \mathcal{L}_{\text{token}} + \lambda \mathcal{L}_{\text{spectral}} \\
 &= \frac{1}{m} \sum_{x_i \in \text{vis}} (x_i - \hat{x}_i)^2 + \frac{1}{n} \sum_{j=1}^n (x_j - \hat{x}_j)^2 \\
 &= \frac{1}{m} \sum_{x_i \in \text{vis}} (x_i - \hat{x}_i)^2 + \frac{1}{n} \sum_{j=1}^n \left([x_{r,c,1}, \dots, x_{r,c,D_k}]_j - [\hat{x}_{r,c,1}, \dots, \hat{x}_{r,c,D_k}]_j \right)^2,
 \end{aligned}$$

where m is the number of masked spatial tokens, i.e., $H_p \times W_p$ patches, n is the number of spectral tokens, with each spectral token consisting of D_k standard tokens along the spectral dimension, (r, c) denotes the standard token in the r -th row and

c -th column of the spectral data, λ balances the contribution of spectral consistency against the token reconstruction loss.

The empirical evaluation of SpectralGPT demonstrates its profound efficacy as a robust initialization for a suite of practical RS applications. The transferability and generalization capacity of the pretrained SpectralGPT models yield significant performance improvements across multiple crucial downstream tasks. This highlights the utility of the learned spatial-spectral representations as a potent, generic feature extractor for the RS domain.

1.4 Discrete Wavelet Transform in computer vision

Having discussed transformers and super-resolution, we now turn to another tool relevant to this thesis: the Discrete Wavelet Transform (DWT). In this thesis, the application of the DWT plays a central role in part of the proposed research. For this reason, this section introduces the general concept of DWT and presents a selection of representative recent methods that employ it for image processing tasks. The DWT is a widely used mathematical tool for multi-resolution analysis and has proven valuable in remote sensing applications [41]. DWT decomposes a signal or image into a set of orthogonal basis functions, called wavelets, which capture both spatial and frequency information at multiple scales. This property makes it particularly suitable for multispectral satellite imagery, where relevant features can appear at different spatial resolutions and across spectral bands. For a 2D multispectral image $x \in \mathbb{R}^{H \times W \times B}$, the DWT can be applied independently to each spectral band $x^{(b)}$, producing a set of approximation and detail coefficients:

$$X \xrightarrow{\text{DWT}} \{X_{LL}^{(L)}, \{X_{LH}^{(l)}, X_{HL}^{(l)}, X_{HH}^{(l)}\}_{l=1}^L\},$$

where L is the number of decomposition levels, and each level l produces one low-frequency subband $X_{LL}^{(l)}$ and three high-frequency subbands $\{X_{LH}^{(l)}, X_{HL}^{(l)}, X_{HH}^{(l)}\}$. The multi-resolution decomposition allows models to capture features ranging from coarse global structures to fine spatial details, which is particularly relevant for land-cover mapping, vegetation analysis, and change detection.

1.4.1 Wave-ViT: Unifying Wavelet and Transformers for Visual Representation Learning

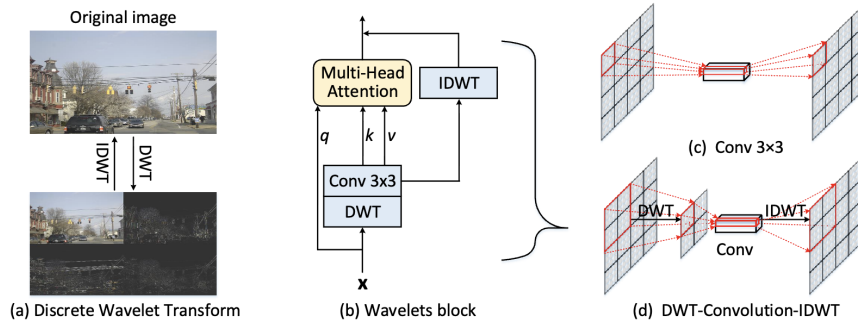


Figure 1.15: Wave-ViT Wavelets block. Keys/values undergo DWT, a 3×3 convolution at half resolution, attention is computed with full-resolution queries, and IDWT reconstructs a high-resolution feature that is fused with the attention output.

Wave-ViT [42] represents a pioneering effort in the direction of unifying wavelet transforms with vision transformers to create a hybrid architecture that uses both the inductive biases of wavelets and the representational power of self-attention mechanisms.

Wave-ViT addresses a recurrent limitation of multi-scale Vision Transformers: spatial reduction of keys/values (e.g., average pooling or learned strided conv) is irreversible and tends to discard high-frequency content that encodes edges, textures, and fine structures, precisely the details that matter for dense prediction and small-object recognition. The core idea is to replace such lossy down-sampling with an invertible wavelet transform, so that attention can operate at reduced spatial resolution without throwing away detail, and then to restore resolution by an inverse transform. The result is a self-attention block that is simultaneously computationally efficient and information-preserving, yielding stronger features at a fixed FLOP budget than prior multi-scale ViT backbones.

Concretely, consider a feature map $X \in \mathbb{R}^{H \times W \times D}$ entering a self-attention block. In a standard ViT architecture, attention over HW tokens incurs quadratic complexity

$\mathcal{O}((HW)^2)$, multi-scale ViTs reduce this by spatially down-sampling keys/values, but the reduction is lossy. Wave-ViT introduces a Wavelets block that applies a 2D DWT to the key/value feature maps, decomposing them into one low-frequency and three high-frequency components at half resolution:

$$\text{DWT}(X) = (X_{LL}, X_{LH}, X_{HL}, X_{HH}), \quad X_{LL|LH|HL|HH} \in \mathbb{R}^{\frac{H}{2} \times \frac{W}{2} \times D}.$$

Because DWT is invertible, all information is distributed across subbands without aliasing. The four subbands are stacked to form a down-sampled tensor on which a lightweight 3×3 convolution imposes locality, yielding context-aware down-sampled keys/values. Multi-head attention is then computed between the full-resolution queries and these down-sampled, wavelet-encoded keys/values. In parallel, an inverse DWT (IDWT) is applied to the subbands (after the local 3×3 conv) to reconstruct a high-resolution feature that aggregates local context with a larger receptive field. The final output of the block blends the attention output and the IDWT reconstruction, so that long-range interactions and locally contextualized details are both retained. Algebraically, by denoting K', V' the DWT-then-convolution processed keys/values at $(H/2) \times (W/2)$ resolution, attention operates as

$$\text{Attn}(Q, K', V') = \text{Softmax}\left(\frac{QK'^\top}{\sqrt{d}}\right) V',$$

while the companion branch reconstructs $\tilde{X} = \text{IDWT}(\tilde{X}_{LL}, \tilde{X}_{LH}, \tilde{X}_{HL}, \tilde{X}_{HH})$, where each $\tilde{X}_\bullet$ is the locally contextualized subband. The block output is a learned fusion of $\text{Attn}(Q, K', V')$ and $\tilde{X}$. Refer to Fig. 1.16 for the detailed design of the wavelet block.

Two properties clarify why this design improves the efficiency–accuracy trade-off. First, complexity: reducing keys/values to $(H/2) \times (W/2)$ tokens lowers the dominant attention term from $\mathcal{O}((HW)^2)$ to $\mathcal{O}(HW \cdot \frac{HW}{4})$ for the $QK^\top$ product and likewise for the subsequent multiplication with V' , while queries remain at full resolution to avoid losing detail where it matters. Second, information preservation: unlike pooling, which discards high-frequency content, DWT partitions it into explicit subbands that can be locally processed and then perfectly recombined by IDWT,

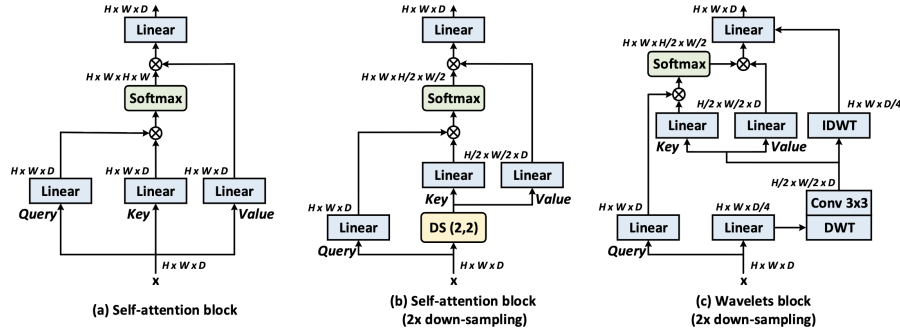


Figure 1.16: The specific designs of (a) the foundational ViT Backbone self-attention block, (b) the self-attention block that employs a down-sampling operation (DS(2, 2)) to halve the spatial scale of height and width, and (c) WaveViT’s Wavelets block, which leverages wavelet transforms to accomplish lossless down-sampling.

the DWT–Conv–IDWT path effectively acts as an invertible, large-receptive-field operator with negligible overhead relative to a single 3×3 convolution 1.16. Empirically, the authors show that integrating Wavelets blocks across stages yields consistent gains across image classification, object detection, and instance segmentation.

In summary, Wave-ViT unifies wavelet-based multi-resolution analysis with transformer attention to achieve lossless down-sampling of keys/values and large-receptive-field local context, overcoming the brittleness of pooling-based reductions. The reported gains over strong multi-scale ViTs at matched compute indicate that preserving frequency details during spatial reduction is a principled way to close the efficiency gap without sacrificing fine structure, an attractive property for high-resolution multispectral imagery.

1.4.2 Wavelet-Driven Masked Image Modeling: A Path to Efficient Visual Representation

WaMIM [43] revisits masked image modeling (MIM) through the lens of frequency analysis. Pixel-space MIM (e.g., MAE) is simple and effective but tends to over-emphasize fine textures and other redundant high-frequency details, which slows con-

vergence and can destabilize self-supervised training. Prior frequency-domain variants often rely on the Fourier transform, which captures global spectra but discards local spatial structure. WaMIM instead uses the discrete wavelet transform (DWT) to produce multi-level, spatially localized frequency targets, aligning the pretext objective with the hierarchical nature of modern encoders and yielding faster, more compute-efficient pretraining with equal or better downstream accuracy. The key idea is to replace a single pixel target with a set of wavelet targets at different decomposition levels, decoded from different encoder layers with tunable weights.

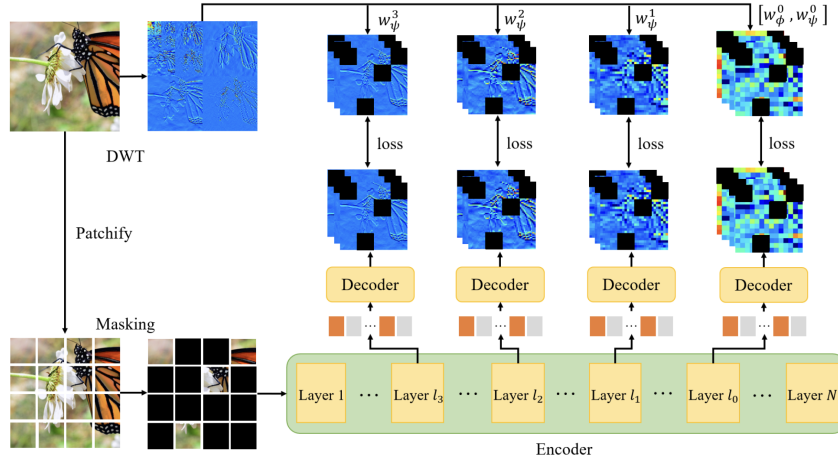


Figure 1.17: WaMIM training: images are patchified and masked; the encoder provides layerwise features, the decoder reconstructs multi-level wavelet targets (detail subbands from shallow features, approximation from deep features). Loss weights w_ψ^k and w_ϕ^k prioritize the most informative scales.

Given an input image $x \in \mathbb{R}^{H \times W \times C}$, WaMIM performs a multi-level 2D DWT to obtain level-specific approximation and detail coefficients that encode coarse structure and progressively higher-frequency content. Denote the L -level decomposition as

$$\mathcal{W}(x) = \{ A^{(L)}, \{ H^{(\ell)}, V^{(\ell)}, D^{(\ell)} \}_{\ell=1}^L \},$$

where $A^{(L)}$ is the deepest low-frequency approximation and $\{ H^{(\ell)}, V^{(\ell)}, D^{(\ell)} \}$ are

horizontal/vertical/diagonal detail subbands at level ℓ . Input patches are masked in image space as usual, the masked view $\hat{x}$ is passed to an encoder f_{EN} producing a stack of features $\{z_1, \dots, z_K\}$ from selected layers. A lightweight decoder f_{DE} uses layer- k features to predict the wavelet targets tied to a specific level, following the empirical observation that shallow layers carry more high-frequency detail while deeper layers encode lower-frequency, semantic structure:

$$\hat{T}_k = f_{\text{DE}}(z_k), \quad T_k \in \{H^{(\ell)}, V^{(\ell)}, D^{(\ell)}\} \text{ for shallow } k, \quad T_K \in \{A^{(L)}\} \text{ for deep } K.$$

WaMIM then minimizes a weighted multi-target reconstruction loss,

$$\mathcal{L}_{\text{WaMIM}} = \sum_{k=1}^K w_{\psi}^k \|T_k - \hat{T}_k\|_2^2, \quad \sum_{k=1}^K w_{\psi}^k = 1, \quad (1.5)$$

where the coefficients w_{ψ}^k control the relative importance of frequency bands and scales during pretraining. In practice, higher weights on informative low-/mid-frequency targets accelerate learning while still supervising high-frequency detail, the mapping from encoder depth to wavelet level (and the w_{ψ}^k schedule) is the central design choice. Refer to Fig. 1.17 for the full architecture depiction.

This design brings two concrete advantages over prior state of the art. First, efficiency: by supervising on compact, structured wavelet coefficients instead of raw pixels, the objective reduces redundancy and improves gradient signal quality, shortening the wall-clock needed to reach a given accuracy. Second, locality with scale: unlike Fourier targets, DWT preserves spatial support and provides intrinsic multi-scale structure, which matches the encoder’s layerwise hierarchy (shallow→high-frequency, deep→low-frequency), reducing the pretext–architecture mismatch that can damage optimization. Empirically, WaMIM consistently improves the accuracy–efficiency trade-off across ViT and Swin backbones.

Chapter 2

Research findings

*We understand how dangerous a mask can be.
We all become what we pretend to be.*

Patrick Rothfuss

2.1 Swin2-MoSE: A New Single Image Super-Resolution Model for Remote Sensing

2.1.1 Introduction

Satellite optical payloads inherently balance spectral richness against spatial detail. Operating at orbital altitudes of several hundred kilometres constrains the ground sampling distance of multispectral sensors, which in turn limits downstream tasks such as segmentation and object detection. As elaborated in Chapter 1 (Section 1.2), this trade-off is a consequence of instrument design choices. Early studies in remote-sensing SISR reported that increasing spatial resolution improves visual interpretation and also serves as an effective pre-processing step for subsequent recognition and feature-extraction pipelines. Transformer-based restoration models have recently become prominent: images are partitioned into patches, attention is computed within

local windows, and a hierarchical pathway reconstructs the output. As summarized in Chapter 1, such models typically surpass convolutional counterparts, nonetheless, fixed feed-forward MLP layers can restrict representational flexibility when facing heterogeneous multispectral content. To mitigate these limitations, *Swin2-MoSE* is presented as an end-to-end SISR framework tailored to multispectral imagery. The design extends Swin2SR and integrates three principal enhancements:

1. a sparsely gated Mixture-of-Experts (MoE) module substitutes the MLP within each transformer block, enabling experts to specialise on distinct image regimes;
2. positional information is enriched by combining a per-head relative bias (RPE) with a per-channel bias (LePE), thereby encoding both 2D spatial offsets and spectral locality;
3. training employs a hybrid objective mixing normalised cross-correlation (NCC) and structural similarity (SSIM), balancing global correlation against local perceptual structure.

Collectively, these elements deliver state-of-the-art performance across two multispectral super-resolution benchmarks.

2.1.2 Related

Remote Sensing Single Image Super-Resolution

Traditional convolutional networks, e.g., SRCNN and VDSR, were among the first to be adapted for multispectral remote-sensing data. As detailed in Chapter 1 (Section 1.2), subsequent specialised architectures such as the Local–Global Combined Network and the Deep Memory Connected Network introduced multi-scale and hourglass-style designs to merge global context with fine spatial detail. Later transformer backbones, including SwinIR [31] and SwinFIR [32], adopted windowed self-attention with hierarchical feature pyramids, while Swin2SR [33] improved stability using cosine-similarity attention and post-norm configurations. The Dual Aggregation Transformer [34] unified spatial and spectral attention in a single module.

Despite these advances, most approaches still rely on a fixed feed-forward component and apply uniform processing to all tokens. Swin2-MoSE diverges from this pattern through a sparsely gated MoE [44], enabling conditional computation and expert specialization. In contrast to token-level routing commonly used in prior work, the proposed routing policy operates at the example level, and expert outputs are fused by a learnable Smart Merger.

2.1.3 Architecture

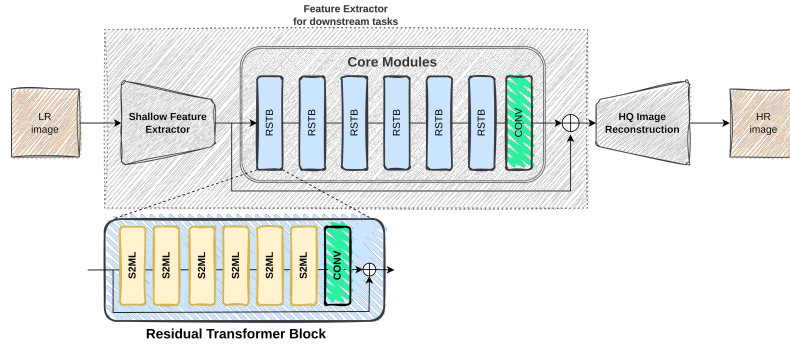


Figure 2.1: Overview of the Swin2-MoSE architecture (solid colours highlight modifications with respect to Swin2SR). The *Feature Extractor* component is analysed later in Section 2.1.4 in connection with the semantic segmentation experiments.

Figure 2.1 summarises the overall organisation of Swin2-MoSE. Building upon the limitations identified for Swin2SR [33], the next subsections describe the architectural modifications introduced to strengthen super-resolution capability. Subsection 2.1.3 presents a Mixture-of-Experts module (MoE-SM) inserted within transformer blocks, referred to as Swin v2 MoSE Layers (S2ML). Subsection 2.1.3 examines the interaction between positional encoding schemes and distils practical guidance for super-resolution transformers. Finally, Subsection 2.1.3 investigates alternative loss formulations and motivates the chosen training objective.

Mixture of Experts

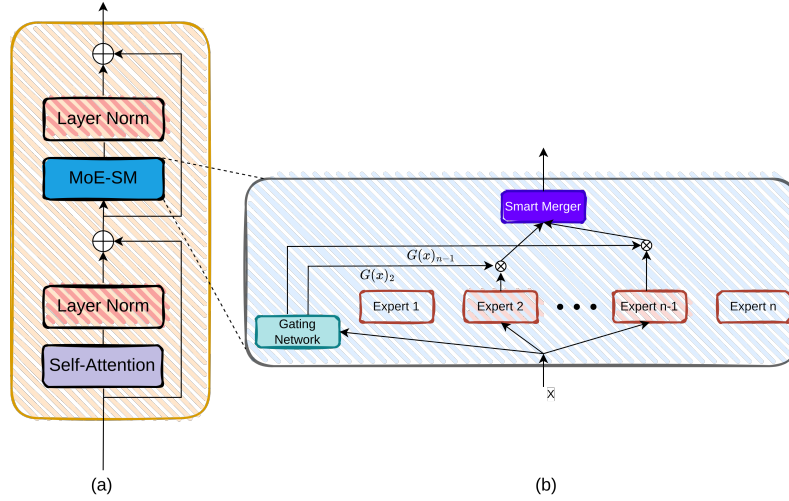


Figure 2.2: Mixture-of-Experts design (solid colours denote modified components): (a) the Swin v2 MoSE Layer (S2ML) couples a Self-Attention block (see Figure 2.3) with an MoE-SM in place of the MLP; (b) internal layout of MoE-SM featuring the Smart Merger for fusing expert outputs.

Conditional computation has emerged as an effective strategy for scaling model capacity while maintaining a bounded compute budget, particularly when training on very large datasets. Mixture-of-Experts (MoE) embodies this paradigm by activating only a subset of model parameters for a given input, thereby introducing sparsity that improves efficiency without sacrificing capacity. Here, sparsity reflects the fraction of active blocks used to produce an output relative to the total available blocks. In practice, this separates the *sparse parameter count* (SPC), the full parameter budget, from the *active parameter count* (APC), the parameters engaged for a specific example, following the convention in Mixtral [45]. Selective activation has been investigated extensively in NLP, from early LSTM-based systems with 137B parameters [46] to more recent architectures scaling to 1.6T parameters [47].

An MoE typically comprises a set of Experts, a Gating Network (GN) that selects

a subset of those Experts, and a merging mechanism that aggregates their outputs into a final representation. This setup encourages experts to specialise on distinct input characteristics, which is advantageous for complex restoration tasks. The present work exploits this property within SISR to elevate reconstruction quality while enabling future scaling via additional experts without a prohibitive increase in compute.

Swin2SR [33] is adopted as the base transformer block due to its strong performance in super-resolution. As depicted in Figure 2.2(a), and analogous to designs such as Switch Transformer [47], SwinV2-MoE [48] and V-MoE [44], the standard MLP is replaced by the MoE-SM. All MLP layers across blocks are substituted to allow MoE-SM contributions at each processing depth while keeping overhead manageable.

The MoE-SM (Figure 2.2(b)) follows a sparsely gated MoE formulation [46] but differs from the original design by employing a learnable Smart Merger (SM) to combine expert outputs. The SM is implemented as a 3×3 2D convolution with padding 1, using k input channels (equal to the number of active experts for a batch element) and a single output channel, thus realizing a learnable weighted fusion instead of a fixed linear combination.

Expert selection is controlled by a trainable Gating Network. A per-example routing strategy is used: each input instance in the batch is dispatched to a subset of experts, instead of distributing individual tokens. Unlike [46], no learned noise term is added to the gating logits: a Top-K gating function is applied instead, as the noise term was observed to grow during training and destabilize the gate. The gating operation is defined as:

$$G(x) = \text{Softmax}(\text{KeepTopK}(x \cdot W_g), k)$$

$$\text{KeepTopK}(v, k) = \begin{cases} v_i & \text{if } v_i \text{ is in the top } k \text{ elements of } v \\ -\infty & \text{otherwise} \end{cases} \quad (2.1)$$

where $G(x)$ denotes the softmax-normalized Top-K scores, W_g are the GN weights, x the input features, and k the number of active experts per example.

Following [46], an auxiliary objective is included to promote balanced expert utilization. Let $\text{Imp}(X)$ quantify the cumulative importance accrued by each expert

across a set of examples X :

$$\begin{aligned} Imp(X) &= \sum_{x \in X} G(x) \\ \mathcal{L}_{MoE} &= CV(Imp(X))^2 \end{aligned} \quad (2.2)$$

where CV is the coefficient of variation (variance divided by squared mean). The term $\mathcal{L}_{MoE}$ is the balancing loss applied to MoE-SM.

Positional Encoding

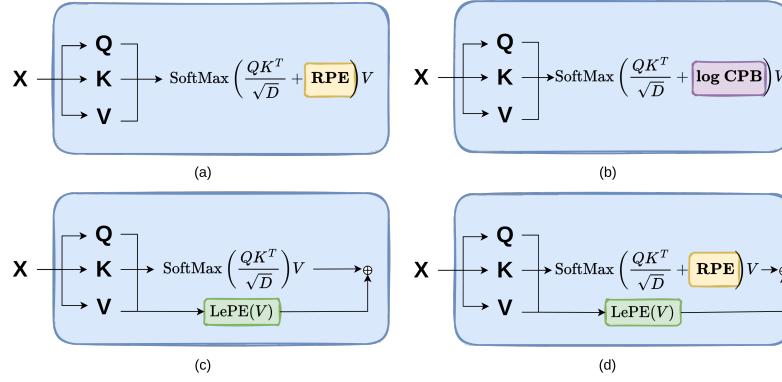


Figure 2.3: Comparison of positional encodings within the self-attention module: (a) Relative Position Encoding (RPE) in Swin Transformer; (b) Log-Spaced CPB (log-CPB) in Swin Transformer v2; (c) Locally-enhanced Positional Encoding (LePE) in CSwin; (d) integrated configuration combining multiple encodings.

Self-attention in Vision Transformers is permutation-invariant, which necessitates explicit spatial signals to preserve 2D structure. Several mechanisms address this requirement. Among them, per-head schemes such as RPE [15] and log-spaced CPB [49], and per-channel schemes such as LePE [50], have emerged as particularly effective. Beyond APE and RPE in Swin [15], there appears to be little prior analysis of how popular positional encodings interact when deployed jointly.

Two empirical observations are worth mentioning: (i) combining a per-head bias with a per-channel bias improves performance relative to using either in isolation,

as the former encodes 2D positional offsets (width/height) while the latter injects 1D locality along the channel dimension; (ii) pairing two per-head biases generally offers limited returns relative to the added overhead and is usually unnecessary.

Guided by these findings, the configuration in Figure 2.3(d) adopts RPE plus LePE, which consistently proved the most beneficial for super-resolution in the reported experiments. This pairing provides relative position information while reinforcing local inductive biases.

Losses

Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity (SSIM) are widely used to assess super-resolved imagery. PSNR, however, may favor overly smooth solutions and exhibits sensitivity to small pixel-level deviations that do not necessarily correlate with perceptual quality [51]. SSIM, while accounting for luminance, contrast and structure, can likewise struggle to reflect perceived fidelity [52]. When these quantities serve as training objectives, their limitations can propagate to the reconstructions.

To reduce such artifacts, several reconstruction losses were evaluated as training objectives. The most effective choice combined the Normalized Cross-Correlation (NCC) loss with the SSIM loss, thereby constraining distortions introduced during super-resolution. Although both are similarity measures, NCC leverages global statistics, whereas SSIM focuses on local structure: using both was observed to yield higher-quality outputs.

The overall objective is a weighted sum:

$$\mathcal{L} = \alpha \cdot \mathcal{L}_{NCC} + \beta \cdot \mathcal{L}_{SSIM} + \gamma \cdot \mathcal{L}_{MoE} \quad (2.3)$$

where α , β , and γ weight the NCC, SSIM, and MoE terms, respectively.

NCC Loss. The Normalized Cross-Correlation (NCC) or Pearson correlation coefficient quantifies similarity between two signals or images by comparing centered inner products. Scores lie in $[-1, 1]$, with 1 indicating perfect correlation, 0 no correlation, and -1 perfect inverse correlation.

For the s -th spectral channel, the NCC value (NCC_s) is:

$$NCC_s = \frac{\sum (\hat{y} - \mu_{\hat{y}}) \cdot (y - \mu_y)}{\sqrt{\sum (\hat{y} - \mu_{\hat{y}})^2 \cdot \sum_{i=1}^n (y - \mu_y)^2}} \quad (2.4)$$

where $\hat{y}$ and y denote network output and ground truth, and $\mu_{\hat{y}}, \mu_y$ their corresponding means. A per-channel score is computed and then averaged across S channels. To map the correlation range to a positive loss, the following definition is used:

$$\begin{aligned} NCC &= \frac{1}{S} \sum_{s=1}^S NCC_s \in [-1, 1] \\ \mathcal{L}_{NCC} &= 1 - \frac{1}{2} (NCC + 1) \in [0, 1] \end{aligned} \quad (2.5)$$

where $\mathcal{L}_{NCC}$ is the loss derived from NCC.

SSIM Loss. Structural Similarity (SSIM) [53] assesses luminance, contrast and structural agreement between two images. For the s -th spectral channel:

$$\begin{aligned} SSIM_s(\hat{y}, y) &= \mathbb{L}(\hat{y}, y)^\delta \cdot \mathbb{C}(\hat{y}, y)^\varepsilon \cdot \mathbb{S}(\hat{y}, y)^\eta \\ \mathbb{L}(\hat{y}, y) &= \frac{2\mu_{\hat{y}}\mu_y + C_1}{\mu_{\hat{y}}^2 + \mu_y^2 + C_1} \\ \mathbb{C}(\hat{y}, y) &= \frac{2\sigma_{\hat{y}y} + C_2}{\sigma_{\hat{y}}^2 + \sigma_y^2 + C_2} \\ \mathbb{S}(\hat{y}, y) &= \frac{\sigma_{\hat{y}y} + C_3}{\sigma_{\hat{y}}\sigma_y + C_3} \end{aligned} \quad (2.6)$$

with $\sigma_{\hat{y}}^2, \sigma_y^2$ and $\sigma_{\hat{y}y}$ denoting variances and covariance, and C_n ($n \in [1, 3]$) small constants for numerical stability. A circular-symmetric Gaussian weighting with window 11×11 is applied; local statistics and the local $SSIM_s$ are computed per window while sliding across the image. The luminance term $\mathbb{L}$ compares mean intensities, the contrast term $\mathbb{C}$ compares standard deviations, and the structural term $\mathbb{S}$ compares correlation. Typical settings use $\delta = \varepsilon = \eta = 1$.

The SSIM-based loss is the channel-wise complement averaged over S channels:

$$\mathcal{L}_{SSIM}(\hat{y}, y) = \frac{1}{S} \sum_{s=1}^S (1 - SSIM_s) \quad (2.7)$$

2.1.4 Experiments

Tasks and Datasets. A comprehensive experimental study was conducted across multiple datasets. For ablation studies, the $4\times$ setting was evaluated on the Sen2Ven μ s benchmark. Direct assessment of SISR performance employed both Sen2Ven μ s [54] and OLI2MSI [55], each providing paired low- and high-resolution observations sourced from distinct satellite sensors with differing native resolutions. In addition, the downstream implications of super-resolved quality were examined through a semantic segmentation task on SeasoNet [56]. Further details for each dataset are provided below.

Sen2Ven μ s dataset. This large-scale SISR dataset, publicly released for the remote-sensing community, pairs Sentinel-2 inputs with Ven μ s references. Specifically, Sentinel-2 patches at 10m and 20m (128×128 and 64×64 pixels, respectively) are matched to Ven μ s patches at 5 m (256×256 pixels), co-acquired on the same day with at most a 30-minute interval. The dataset spans 29 distinct locations over 2 years, totaling 132,955 patch pairs.

OLI2MSI dataset. OLI2MSI offers real paired LR–HR imagery curated under a cloud-coverage threshold within appropriate temporal proximity. Landsat-8 OLI provides the LR observations at 30m (160×160 pixels), whereas Sentinel-2 MSI contributes the HR counterparts at 10m (480×480 pixels). The collection comprises 5,325 pairs, partitioned into 5,225 for training and 100 for testing.

SeasoNet dataset. SeasoNet is a large-scale multi-label land-cover/use resource derived from Sentinel-2 tiles with 12 spectral channels, totalling 1,759,830 scenes. Each patch comprises 120×120 , 60×60 , and 20×20 pixels corresponding to the 10m, 20m, and 60m bands, respectively. Each image is associated with a segmentation tensor of size $33 \times 120 \times 120$, where 33 denotes the class set cardinality. Altogether, 519,547 unique spatial locations are sampled. For each location, the data are stratified by season into five subsets, plus an additional snowfall subset. To evaluate the quality of super-resolved outputs, the Spring (grid 1) subset is selected for training and testing, using the 10 m bands B2, B3, B4, and B8 as inputs.

Implementation details. Unless stated otherwise, all training runs use a single A100 GPU, the Adam optimiser, a learning rate of 0.0001, 70 epochs, and a batch size of 8.

For the final SeasoNet experiment, training uses Adam with the same learning rate, 40 epochs, and a batch size of 128.

Evaluation Metrics. Quantitative comparisons rely on Peak Signal-to-Noise Ratio (PSNR, dB) and Structural Similarity Index Measure (SSIM). Only for the first ablation concerning losses (Section 2.1.4) is the Normalized Cross-Correlation (NCC) also reported, reflecting the design of that study. For the SeasoNet segmentation task, F1-macro and F1-micro are employed: the former averages per-class F1 uniformly, while the latter aggregates over samples, weighting majority classes more heavily. Across all metrics, larger values indicate better performance.

Training Losses

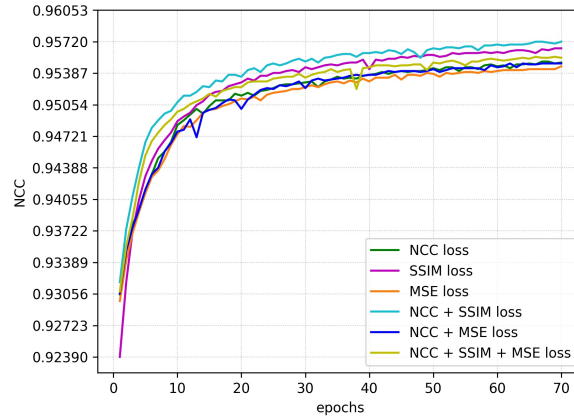


Figure 2.4: Evolution of the NCC score under different training-loss configurations.

This ablation investigates how different training losses affect reconstruction quality. Figures 2.4–2.6 and Table 2.1 report results for the NCC, SSIM, and MSE losses, evaluated with NCC, SSIM, and PSNR metrics, the backbone is Swin2SR.

Consider first rows 1–3. Using NCC as the sole loss yields strong NCC scores (row 1) but underperforms on SSIM and PSNR (see Figures 2.5 and 2.6), indicating limited suitability for perceptual fidelity. Because SSIM explicitly models luminance, contrast, and structure, it tends to be more robust to local variations, which explains

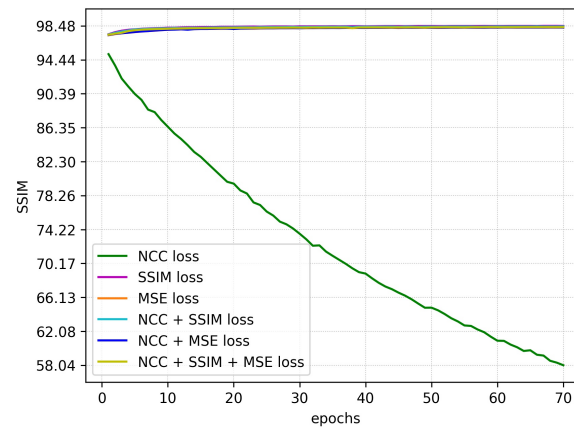


Figure 2.5: Evolution of the SSIM score under different training-loss configurations.

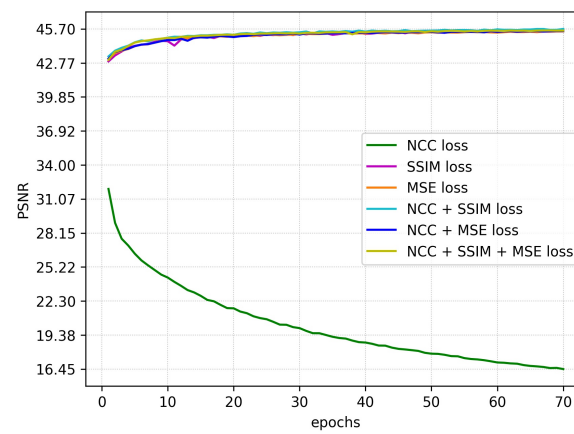


Figure 2.6: Evolution of the PSNR score under different training-loss configurations.

Table 2.1: Ablation study on loss usage. The table shows the performance of the model using different combinations of loss functions. The combination with the highest performance is shown in **bold** and the second-highest performance is shown in **red**.

	Losses			Performance		
	NCC	SSIM	MSE	NCC $\uparrow$	SSIM $\uparrow$	PSNR $\uparrow$
1	✓			0.9550	0.5804	16.4503
2		✓		0.9565	0.9847	45.5427
3			✓	0.9546	0.9828	45.4759
4	✓	✓		0.9572	0.9841	45.6986
5	✓		✓	0.9549	0.9828	45.5163
6	✓	✓	✓	0.9555	0.9833	45.5542

why training with SSIM alone secures competitive NCC values (row 2), whereas the converse does not hold (row 1). Similarly, employing MSE alone attains competitive NCC (row 3), whereas NCC as loss underperforms on PSNR, reflecting the global-versus-local distinction: NCC is driven by global statistics, while SSIM and MSE/PSNR emphasise local windows or pixelwise residuals.

Despite these caveats, NCC remains valuable when used in combination. Rows 4 and 5 show that augmenting SSIM or MSE with NCC generally improves performance relative to using SSIM or MSE alone (rows 2–3), as NCC constrains global correlations in a manner complementary to local similarity terms. Consequently, the configuration combining NCC and SSIM was selected, providing the best overall trade-off (see row 4).

Positional Encoding

This ablation analyses the joint behaviour of positional encodings. The baseline (row 2 in Table 2.2) corresponds to the best setting from the preceding experiment: Swin2SR with log-CPB and training under NCC+SSIM (row 4 in Table 2.1).

Table 2.2 indicates that the most effective strategy is to combine a per-head bias

(RPE or log-CPB) with a per-channel bias (LePE), rather than using either in isolation (row 5 versus rows 1–3). Employing both per-head encodings together (RPE and log-CPB) produces only marginal differences (rows 1-2 versus row 6). Moreover, enabling all three simultaneously does not yield gains sufficient to justify additional complexity (row 7 versus rows 4–6).

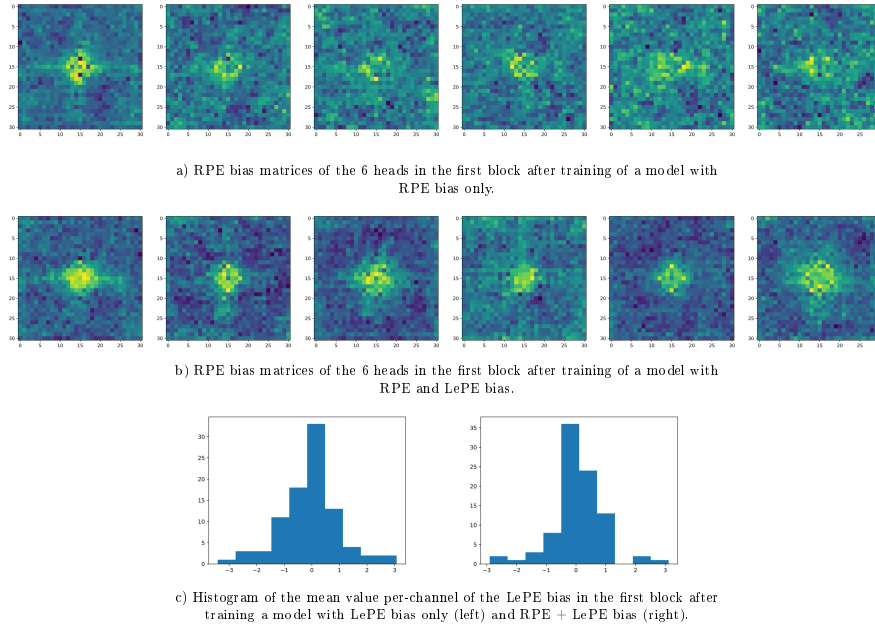


Figure 2.7: Learned bias matrices for RPE and LePE visualised post-training.

To further substantiate these choices, Figure 2.7 provides a qualitative analysis across different encoding combinations. RPE matrices are shown as $\hat{B} \in \mathbb{R}^{(2M-1) \times (2M-1)}$ (with B the batch size and M^2 the window patch count), and LePE as tensors of shape $B \times C \times 256$ (with C channels). Subfigures 2.7 (a) and 2.7 (b) depict the six attention heads in the first block using RPE alone and RPE+LePE, respectively. LePE does not degrade RPE patterns and, instead, appears to consolidate them. Subfigure 2.7 (c) plots the per-channel mean of LePE in the first block without (left) and with (right) RPE; distributions remain broadly similar, indicating that LePE behaves consistently

whether or not RPE is present.

Table 2.2: Ablation study on positional encoding. The best performance is shown in **bold** and the second-highest performance is shown in **red**.

	Positional Encoding			Performance	
	RPE	log CPB	LePE	SSIM $\uparrow$	PSNR $\uparrow$
1	✓			0.9841	45.5855
2		✓		0.9841	45.6986
3			✓	0.9843	45.7278
4		✓	✓	0.9845	45.8046
5	✓		✓	0.9847	45.8539
6	✓	✓		0.9843	45.6945
7	✓	✓	✓	0.9846	45.8185

Mixture of Experts

This ablation evaluates whether MoE-SM can surpass the baseline MLP (Table 2.3, row 3 versus row 1). Active and total expert counts follow the Mixtral choice [45], eight experts with two active per input—while keeping each expert at roughly half the MLP parameter count to equalize active computation. A subsequent study varies both the total experts and the number of active experts.

Results are reported in Table 2.3. The baseline (row 1) corresponds to the best positional-encoding configuration (row 5 in Table 2.2) using MLP. Without the Smart Merger (row 2), MoE exhibits modest fluctuations relative to MLP; introducing SM (row 3) yields consistent gains of +0.0002 SSIM and +0.0735 PSNR. A slight latency increase is observed, attributable to sequential expert execution on a single GPU.

The second block of rows (4–7) shows that reducing active experts degrades performance (row 4 vs. row 6), underscoring the role of SM and parallel expert contributions. Altering the total number of experts alone (rows 5 and 7 vs. row 6) does

2.1. Swin2-MoSE: A New Single Image Super-Resolution Model for Remote Sensing

55

Table 2.3: Ablation study on Transformer MLP and MoE layer architectures. APC / SPC: active and sparse parameters count inside a MLP and MoE layers. Latency: average amount of seconds required to elaborate a batch of 8 images on a NVIDIA Quadro RTX6000. SM: Smart Merger. Model params: amount of sparse parameters of the full model architecture. Model FLOPs: floating point operations in a forward pass of the full model. The best performance is shown in **bold** and the second-highest performance is shown in **red**

	Arch	SM	MLP # params		Performance		Latency	Model	Model
			APC	SPC	SSIM $\uparrow$	PSNR $\uparrow$	(s)	params	FLOPs
1	MLP		32'670	32'670	0.9847	45.8539	0.194	2.20M	135.54G
2	MoE 8/2		33'480	131'760	0.9845	45.8647	0.223	4.59M	135.68G
3	MoE 8/2	✓	33'499	131'779	0.9849	45.9272	0.229	4.59M	138.30G
4	MoE 8/1	✓	34'120	262'810	0.9837	45.4847	0.230	7.72M	136.96G
5	MoE 6/2	✓	33'319	98'839	0.9844	45.8328	0.230	3.80M	138.30G
6	MoE 8/2	✓	33'499	131'779	0.9849	45.9272	0.229	4.59M	138.30G
7	MoE 10/2	✓	33'679	164'719	0.9846	45.8913	0.230	5.39M	138.30G

not help. It is plausible that the optimal total-expert count depends on the training-set scale, larger datasets could benefit from further increases.

SISR Experiments

Quantitative comparison. A three-scale comparison against strong SR baselines is provided. For $2\times$ upscaling, Sen2Ven μ s is used with Sentinel-2 bands B2, B3, B4, B8 (10m) and Ven μ s bands B3, B4, B7, B11 (5m). For $3\times$, OLI2MSI is used with Landsat-8 blue/green/red (30 m) and Sentinel-2 B2/B3/B4 (10 m). For $4\times$, Sen2Ven μ s is used with Sentinel-2 B5, B6, B7, B8A (20m) and Ven μ s B8, B9, B10, B11 (5m). Unless otherwise specified for competing methods, standard MSE training is assumed, whereas the described approach employs SSIM+NCC as motivated earlier.

The results in Table 2.4 indicate consistent gains over strong Swin-based base-

Table 2.4: Quantitative comparison with SOTA models on Sen2Ven μ s and OLI2MSI datasets. GiB, IT/s: gpu memory usage and iterations per second on inference. APC / SPC: active and sparse parameters count inside a MLP and MoE layers. FLOPs: floating point operations in a forward pass of the model. The best performance is shown in **bold** and the second-highest performance is shown in **red**.

Model	Sen2Ven μ s 2 $\times$						
	SSIM $\uparrow$	PSNR $\uparrow$	GiB	IT/s	Params		FLOPs
Bicubic	0.9883	45.5588					
SwinIR [31]	0.9938	48.7064	2.37	1.32	2.10M	2.10M	515.90G
Swinfir [32]	0.9940	48.8532	3.12	1.28	2.52M	2.52M	622.42G
Swin2SR [33]	0.9942	49.0467	3.12	1.18	2.11M	2.11M	526.78G
DAT [34]	0.9946	49.4699	1.46	1.95	2.14M	2.14M	547.22G
Swin2-MoSE	0.9948	49.4883	2.41	1.16	2.20M	4.56M	543.00G
	OLI2MSI 3 $\times$						
	SSIM $\uparrow$	PSNR $\uparrow$	GiB	IT/s	Params		FLOPs
Bicubic	0.9768	42.1835					
SwinIR [31]	0.9863	44.4806	3.55	0.77	2.11M	2.11M	809.42G
Swinfir [32]	0.9863	44.4829	4.73	0.76	2.53M	2.53M	975.84G
Swin2SR [33]	0.9881	44.9614	4.80	0.70	2.12M	2.12M	826.30G
DAT [34]	0.9886	45.0404	2.10	1.10	2.15M	2.15M	858.35G
Swin2-MoSE	0.9912	45.9414	4.88	0.52	2.20M	4.56M	851.76G
	Sen2Ven μ s 4 $\times$						
	SSIM $\uparrow$	PSNR $\uparrow$	GiB	IT/s	Params		FLOPs
Bicubic	0.9674	42.0499					
SwinIR [31]	0.9825	45.3460	0.84	5.37	2.14M	2.14M	131.53G
Swinfir [32]	0.9830	45.5500	0.84	5.20	2.56M	2.56M	158.17G
Swin2SR [33]	0.9828	45.4759	1.03	4.74	2.15M	2.15M	134.40G
DAT [34]	0.9837	45.8053	0.69	6.73	2.18M	2.18M	139.36G
Swin2-MoSE	0.9849	45.9272	0.86	3.91	2.23M	4.59M	138.30G

lines. Relative to related transformer architectures, PSNR improvements up to $0.377 \sim 0.958$ dB are observed, with the largest margins appearing at the more challenging $3\times$ and $4\times$ scales, while $2\times$ remains slightly but consistently better than SOTA. Model-complexity measures are reported alongside accuracy: despite a higher sparse parameter count due to MoE, the active parameter count and FLOPs remain comparable to competing systems.

Qualitative comparison. Figures 2.8–2.11 present visual comparisons for $4\times$ SISR on Sen2Ven μ s, and Figures 2.12–2.14 for $3\times$ on OLI2MSI. Highlighted red boxes indicate the regions of interest for inspection. The reconstructions produced by the model display fewer artefacts and sharper details than the considered baselines.

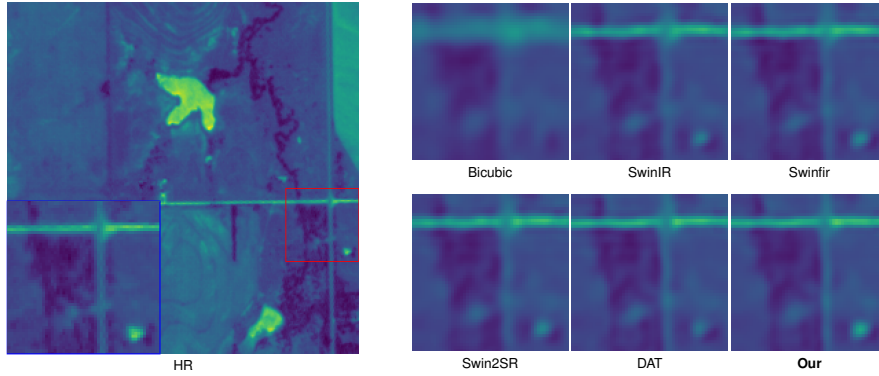


Figure 2.8: Comparison between $4\times$ SISR methods on Sen2Ven μ s dataset. HR reference is Ven μ s B8 (red edge, 5 m). The crop of the HR reference image is shown in the red square, and the zoomed cropped part that was chosen is shown in the blue square.

Semantic Segmentation Experiments

The final study examines whether super-resolution pretraining confers benefits to a different downstream task, in this case semantic segmentation task is studied. For simplicity, only the 10m Sentinel-2 bands (B2, B3, B4, B8) are used as input. A Swin2-MoSE model trained on the $2\times$ Sen2Ven μ s SISR task is employed as a frozen

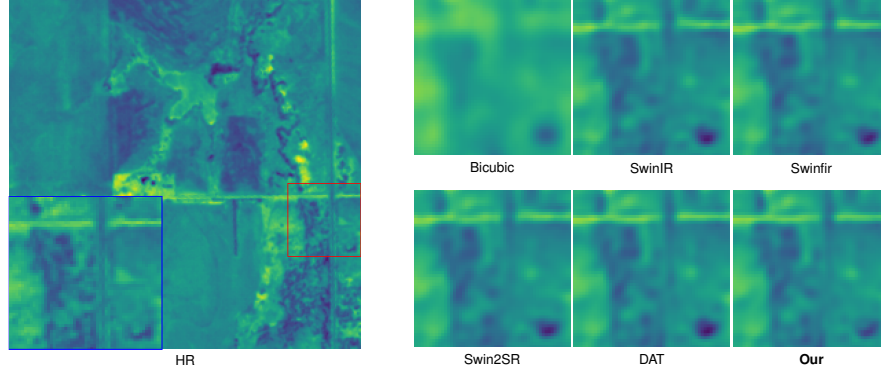


Figure 2.9: Comparison between 4x SISR methods on Sen2Ven μ s dataset. HR reference is Ven μ s B9 (red edge, 5 m). The crop of the HR reference image is shown in the red square, and the zoomed cropped part that was chosen is shown in the blue square.

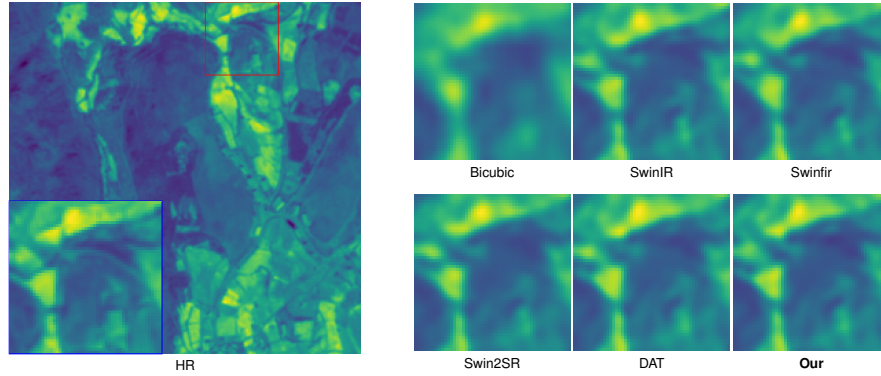


Figure 2.10: Comparison between 4x SISR methods on Sen2Ven μ s dataset. HR reference is Ven μ s B10 (red edge, 5 m). The crop of the HR reference image is shown in the red square, and the zoomed cropped part that was chosen is shown in the blue square.

feature extractor for the Foreground-Aware Relation Network (FarSeg) [57]. The objective is to determine whether the features produced in this manner enhance segmentation performance.

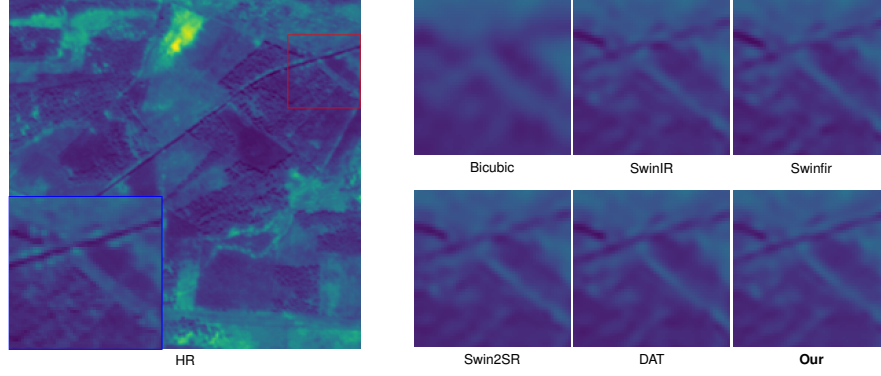


Figure 2.11: Comparison between 4x SISR methods on Sen2Ven μ s dataset. HR reference is Ven μ s B11 (NIR, 5 m). The crop of the HR reference image is shown in the red square, and the zoomed cropped part that was chosen is shown in the blue square.

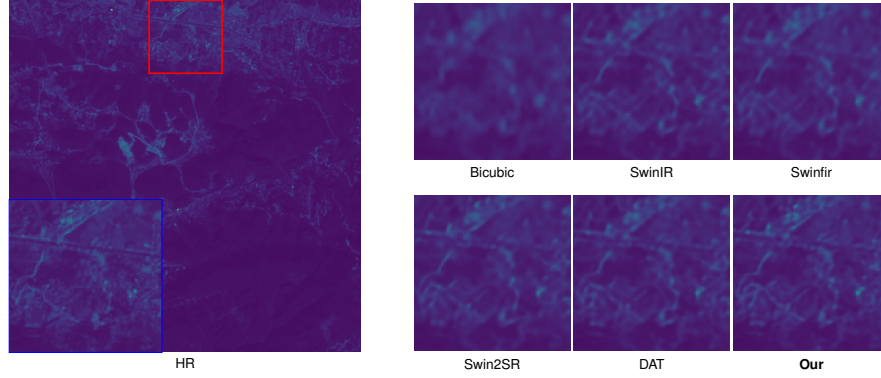


Figure 2.12: Comparison between 3x SISR methods on OLI2MSI dataset. HR reference is Sentinel-2 B2 (blue, 10 m). The crop of the HR reference image is shown in the red square, and the zoomed cropped part that was chosen is shown in the blue square.

Three configurations are compared. (1) *FarSeg*: the FarSeg architecture is trained directly on 4-channel SeasoNet inputs. (2) *FarSeg-S2MFE*: the pretrained, frozen Swin2-MoSE (without the HQ reconstruction head) serves as feature extractor (see the *Feature Extractor* in Figure 2.1), the extracted tensors of size $T \times 128 \times 128$

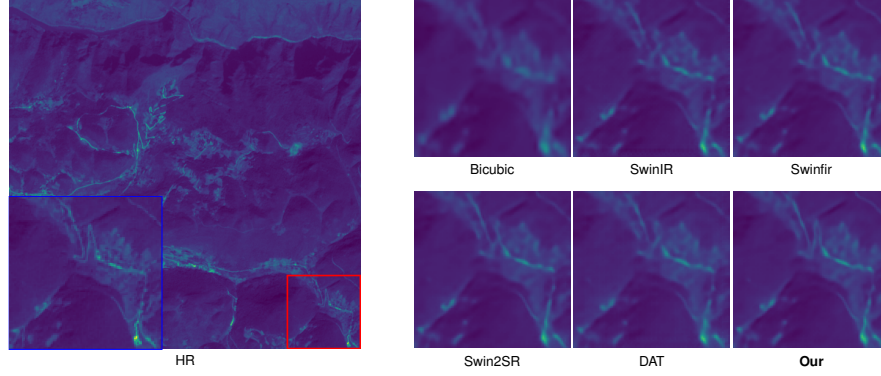


Figure 2.13: Comparison between 3x SISR methods on OLI2MSI dataset. HR reference is Sentinel-2 B3 (green, 10 m). The crop of the HR reference image is shown in the red square, and the zoomed cropped part that was chosen is shown in the blue square.

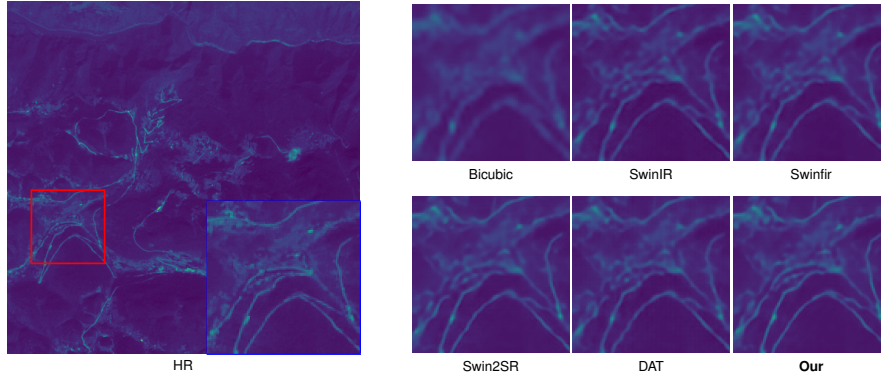


Figure 2.14: Comparison between 3x SISR methods on OLI2MSI dataset. HR reference is Sentinel-2 B4 (red, 10 m). The crop of the HR reference image is shown in the red square, and the zoomed cropped part that was chosen is shown in the blue square.

($T=90$) are then fed to FarSeg. (3) *FarSeg++*: to balance input dimensionality with (2), the 4-channel 120×120 inputs are first projected via a 1×1 convolution (no padding) to $T \times 128 \times 128$, and then passed to FarSeg—note that this variant intro-

duces a learnable projection absent in configurations (1) and (2).

As depicted in Figure 2.15, FarSeg-S2MFE surpasses both baselines across F1-macro and F1-micro throughout training, indicating that the extracted features are beneficial for segmentation. These findings support the claim that super-resolution pretraining can positively influence downstream tasks.

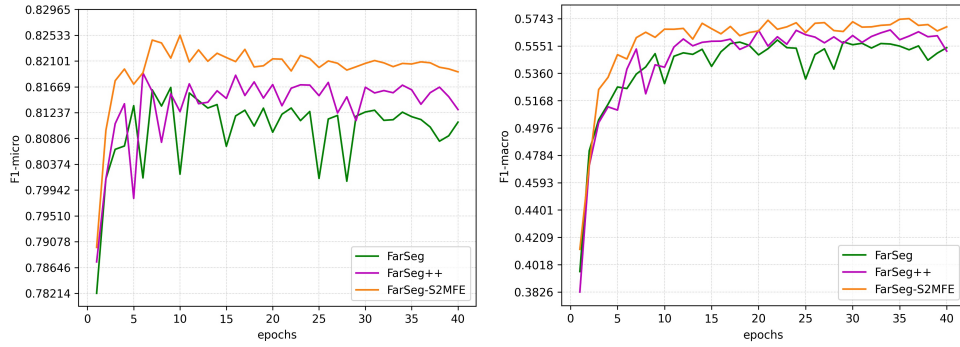


Figure 2.15: Semantic segmentation on Seasonet: training curves for F1-micro and F1-macro.

2.1.5 Conclusions

This section established Swin V2 Mixture of Super-resolution Experts (Swin2-MoSE), an end-to-end SISR framework for remote sensing that advances the state of the art. First, an enhanced Sparsely-Gated Mixture-of-Experts (MoE-SM) was integrated within transformer blocks to raise reconstruction quality, the MoE-SM features a learnable Smart Merger and adopts per-example routing rather than per-token gating. Second, following an analysis of positional encodings, a per-channel bias (LePE [50]) and a per-head bias (RPE [15]) were combined within the architecture. Third, training employed a composite objective mixing Normalized Cross-Correlation (NCC) and Structural Similarity (SSIM) to address the limitations of Mean Squared Error (MSE). Extensive experiments across multiple satellite benchmarks demonstrated improvements over strong baselines in both SSIM and PSNR. Swin2-MoSE exceeds

Swin-based competitors by as much as $0.377 \sim 0.958$ dB in PSNR and by $0.0006 \sim 0.0031$ in SSIM for the $2\times$ (Sen2Ven μ s), $3\times$ (OLI2MSI), and $4\times$ (Sen2Ven μ s) settings. The largest gains appear at higher scale factors, with consistent improvements also at $2\times$. A downstream experiment on semantic segmentation further indicated that features derived from the super-resolution model can enhance performance beyond SR alone.

2.2 WaveMAE: Wavelet decomposition Masked Auto-Encoder for Remote Sensing

2.2.1 Introduction

Labelled datasets for multispectral satellite imagery are scarce and costly to acquire, yet unlabelled archives are vast. As noted in Chapter 1 (Section 1.3), self-supervised learning provides a way to learn from unlabelled data by defining surrogate tasks that encourage structure discovery. Among these, masked image modelling has emerged as a powerful pretext: by predicting the content of masked patches from their visible neighbours, a model learns to capture spatial and spectral dependencies without manual labels. In remote sensing, MAE-style methods extend this idea to multispectral inputs by grouping bands or adopting multi-scale decoders. However, these approaches operate purely in the pixel domain, which may not fully exploit the frequency structure of multispectral images. This observation motivates incorporating frequency-domain information into the pretext task.

Wavelet decomposition offers an alternative for multi-scale representation. The DWT decomposes an image into approximation and detail coefficients, capturing both coarse structure and high-frequency variations. Wavelet analysis has long been applied to remotely sensed images because it handles non-stationary signals and provides multiresolution analysis. Motivated by this, this work proposes WaveMAE, a masked auto-encoder that reconstructs the wavelet components of an image rather than its pixels. By predicting the low-frequency approximation and a hierarchy of high-frequency components, the encoder learns representations that respect both spec-

tral and scale-based correlations. Furthermore, this work introduces a Geo-conditioned Positional Encoding (GPE) based on spherical harmonics that encodes the latitude and longitude of each scene, encouraging geographically proximate images to occupy similar positions in the embedding space. In experiments on the PANGAEA benchmark, WaveMAE consistently outperforms the previous state of the art on tasks ranging from semantic segmentation to regression and multi-label classification.

2.2.2 Related

MAE demonstrated that reconstructing masked pixels is a scalable self-supervised objective for natural images. As summarised in Chapter 1, SatMAE [3] adapts this framework to multispectral imagery by reconstructing randomly masked spectral groups and leveraging redundancy across bands, while SatMAE++ [4] employs multi-scale decoders to handle sensors with different ground sampling distances and Spectral-GPT [5] uses 3D tokenisation with a spectral consistency loss. These methods primarily focus on reconstructing the input at the pixel level. WaveMAE instead reconstructs discrete wavelet components, which encourages the encoder to learn multi-scale frequency features and provides a more compact target space.

Wavelets in deep vision. Wavelet transforms have been integrated with transformers for classification and detection. Wave-ViT [42] uses the DWT to downsample keys and values in self-attention while preserving high-frequency content, and WaMIM [43] reconstructs wavelet coefficients instead of pixels to reduce redundancy in masked image modeling. These approaches focus on natural images and do not address the unique properties of multispectral remote sensing data. WaveMAE instead applies a multi-level DWT to multispectral inputs and combines a multi-level patch embedding that aligns patches across decomposition levels, processing data not from the input image pixel domain but that of the wavelet components. As noted in the previous section, wavelet analysis is well-suited for analyzing non-stationary signals such as remote sensing images and provides a valid framework for multi-resolution analysis with an explicit frequency separation.

2.2.3 Theory priors

Discrete Wavelet Transform

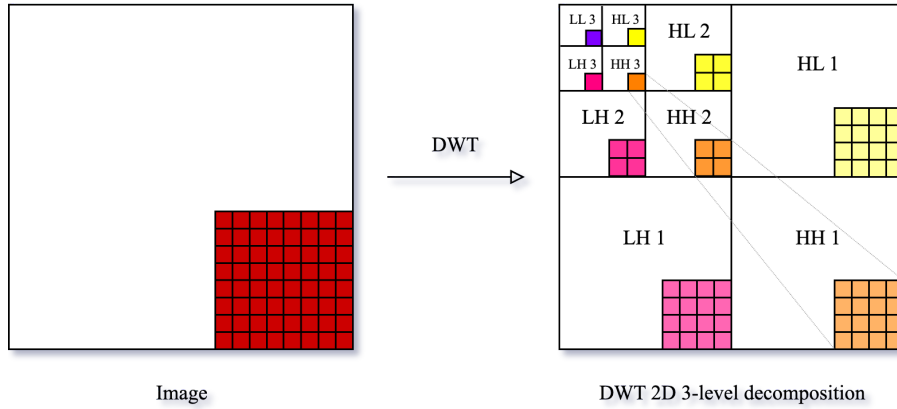


Figure 2.16: Three-level 2D DWT of an image. The pixel region highlighted in red gives rise to the components **LL** (approximation), **LH** (horizontal details), **HL** (vertical details), and **HH** (diagonal details) across scales.

Recalling the previous chapter, the DWT provides a joint time–frequency analysis framework for discrete signals. For a one-dimensional sequence x of length N , the DWT separates the signal into low-frequency approximations A_j and high-frequency details D_j via cascaded low-/high-pass filtering followed by downsampling:

$$\text{Approximation: } A_j[k] = \sum_n x[n] \cdot \eta_{j,k}[n] \quad (2.8)$$

$$\text{Detail: } D_j[k] = \sum_n x[n] \cdot \psi_{j,k}[n] \quad (2.9)$$

where $\eta_{j,k}[n]$ and $\psi_{j,k}[n]$ denote, respectively, the scaling and wavelet (high-pass) functions at scale j and translation k . In image processing, the two-dimensional extension applies the same principle across spatial axes, yielding a multi-scale representation of the content, as illustrated in Figure 2.16. A single 2D DWT stage filters the image $I(x, y)$ with a pair of complementary kernels η (scaling) and ψ (wavelet)

and downsamples by a factor of two along both dimensions:

$$\begin{aligned} \text{LL}_{j+1}(x, y) &= \sum_{m,n} \eta(m)\eta(n) \cdot \text{LL}_j(2x - m, 2y - n) \\ \text{LH}_{j+1}(x, y) &= \sum_{m,n} \eta(m)\psi(n) \cdot \text{LL}_j(2x - m, 2y - n) \\ \text{HL}_{j+1}(x, y) &= \sum_{m,n} \psi(m)\eta(n) \cdot \text{LL}_j(2x - m, 2y - n) \\ \text{HH}_{j+1}(x, y) &= \sum_{m,n} \psi(m)\psi(n) \cdot \text{LL}_j(2x - m, 2y - n) \end{aligned}$$

with $\text{LL}_0 = I(x, y)$ the original image, $j \geq 0$ the level index, and m, n filter taps. The procedure outputs four subbands: LL (approximation), LH (horizontal details), HL (vertical details), and HH (diagonal details). The L/H letters indicate whether low- or high-pass filtering is applied horizontally and vertically during the 2D decomposition.

Spherical Harmonics

Spherical harmonics (SH) describe functions on the surface of a sphere through orthogonal basis functions. SH are complex, continuous and limited functions defined as the angular part of the solution to Laplace's equation in spherical coordinates $Y_\ell^m(\theta, \phi)$, where ℓ is the degree ($\ell = 0, 1, 2, \dots$), m is the order ($m = -\ell, -\ell + 1, \dots, \ell - 1, \ell$).

The equation can be explicated as:

$$Y_\ell^m(\theta, \phi) = \sqrt{\frac{2\ell + 1}{4\pi} \frac{(\ell - |m|)!}{(\ell + |m|)!}} P_\ell^m(\cos \theta) e^{im\phi} \quad (2.10)$$

where P_ℓ^m are the associated Legendre polynomials. Intuitively, SH can be viewed as the spherical analogue of the Fourier series expansion. For a periodic function defined on the circle (i.e., the interval $[0, 2\pi]$), the standard approach is to decompose it into sines and cosines, which represent the natural vibration modes of the circle, each mode corresponding to an integer number of oscillations along the angular coordinate. Analogously, for a function defined on a sphere S^2 , the aim is to decompose it into the natural vibration modes of the sphere. These modes are given by the

spherical harmonics $Y_\ell^m(\theta, \phi)$, where the degree ℓ determines the overall number of oscillations on the sphere, while the order m determines the azimuthal oscillation pattern. Lower-degree harmonics ($\ell = 0, 1, 2$) represent smooth, global variations. Higher degrees capture increasingly fine-grained details and rapid variations on the sphere's surface. Any square-integrable function on the unit sphere can be expressed as follows:

$$f(\theta, \phi) = \sum_{\ell=0}^{\infty} \sum_{m=-\ell}^{\ell} c_\ell^m Y_\ell^m(\theta, \phi) \quad (2.11)$$

with $(lat, lon) = (\theta, \phi)$ and c_ℓ^m is the weight associated to the orthogonal spherical harmonic basis functions. The orthogonality property ensures that different basis functions are independent, enabling efficient computation and analysis. Similarly to [58], only the real form of the spherical harmonics is selected for this application, and they are defined as:

$$Y_\ell^m(\theta, \phi) = \begin{cases} (-1)^m \sqrt{2} \bar{P}_\ell^m(\cos \theta) \sin(|m|\phi) & \text{if } m < 0 \\ \bar{P}_\ell^m(\cos \theta) & \text{if } m = 0 \\ (-1)^m \sqrt{2} \bar{P}_\ell^m(\cos \theta) \cos(m\phi) & \text{if } m > 0 \end{cases} \quad (2.12)$$

where $\bar{P}_\ell^m(\cos \theta)$ is the normalized associated Legendre polynomial. In this way, the orthogonality and completeness properties are preserved while removing the complex term, making them more suitable for applications dealing with real-valued data on geographic coordinates, where θ represents latitude and ϕ represents longitude.

2.2.4 Architecture

The method is designed around the premise that the 2D DWT provides a natural method to expose multi-scale frequency structure in optical RS imagery, enabling an encoder to form representations that couple fine details with broader spatial patterns. Building on this idea, a foundation-style vision encoder is constructed for optical RS, with utility across multiple downstream tasks. In addition, a Geo-conditioned Positional Encoding (GPE) is introduced to inject geolocation priors into the embedding space by means of Spherical Harmonics, producing geography-aware features with

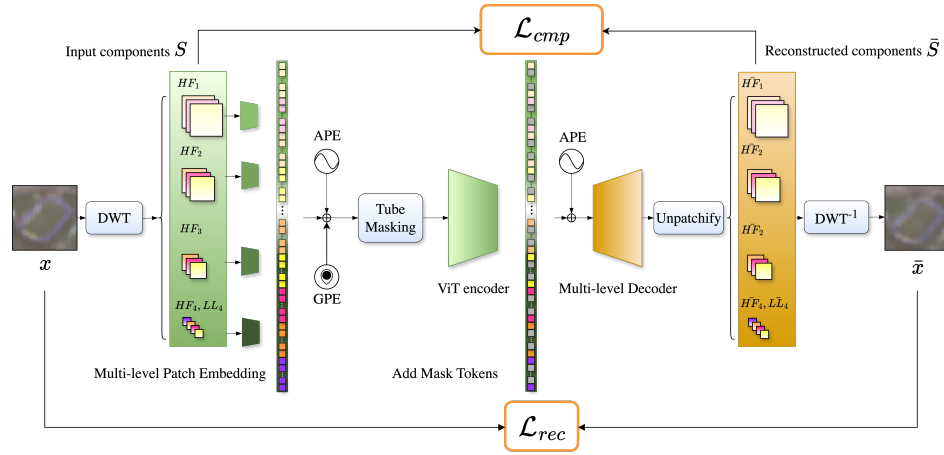


Figure 2.17: Illustration of WaveMAE architecture, the core design follows that of the MAE [2]. The method applies a 4-level DWT decomposition to the input image x of resolution (H, W) is obtaining 4 set of wavelet components at different scales of resolutions $(H/2, W/2), (H/4, W/4), (H/8, W/8)$ and $(H/16, W/16)$. Following this operation the components in S are fed to the Multi-level patch embedding, here lighter colors stand for shallow decomposition levels and low-scale feature extraction, on the opposite darker colors mean deeper decomposition level and high-scale feature extraction. After, the full sequence is fed to the MAE framework, the decoder reconstructs the masked wavelet components $\hat{S}$ at the same resolution specified previously. A smooth L1 loss is calculated between the input and the predicted wavelet components. Additionally, the predicted components are used to calculate the inverse DWT, obtaining the reconstructed multispectral image $\hat{x}$. An MSE loss is calculated between the latter and the input image x . In both losses the method attends only to the previously masked parts of the components/reconstructed image for the calculation.

observed gains on several tasks. Figure 2.17 outlines the overall design. Given a multispectral input x of size $224 \times 224 \times C$ (with C channels defined by the pretraining dataset), an N -level DWT is first applied.

Let $\text{DWT}_N(x)$ denote the N -level 2D wavelet decomposition of x . At each level $j = 1, 2, \dots, N$, the high-frequency sets HL_j, LH_j, HH_j are produced, and at the last level the approximation LL_N is obtained. Collect all components in

$$S = \bigcup_{j=1}^N S_j = S_1 \cup \dots \cup S_N = \{HL_1, LH_1, HH_1\} \cup \dots \cup \{HL_N, LH_N, HH_N, LL_N\}.$$

Empirically, $N=4$ offered the most favourable trade-off (see the experiments). Thus, $\text{DWT}_4(x)$ yields 13 components: one low-frequency term and 12 high-frequency subbands across the four levels. Intuitively, LL_4 corresponds to a $16\times$ downsampled version of I , while $\{HL_i, LH_i, HH_i\}$ are produced at scale factors 2^i :

$$\begin{aligned} HL_1, LH_1, HH_1 &\rightarrow 2 \times \text{downscale} \\ HL_2, LH_2, HH_2 &\rightarrow 4 \times \text{downscale} \\ HL_3, LH_3, HH_3 &\rightarrow 8 \times \text{downscale} \\ HL_4, LH_4, HH_4 &\rightarrow 16 \times \text{downscale} \end{aligned}$$

This hierarchy preserves frequency content at multiple resolutions, granting access to both local textures and broader structures. Consequently, the spatial sizes of the groups are $S_1 : 112 \times 112$, $S_2 : 56 \times 56$, $S_3 : 28 \times 28$, and $S_4 : 14 \times 14$.

Multi-level Patch Embedding

WaveMAE's central design goal is to make tokens that refer to the same spatial location interact across decomposition levels, thereby coupling information from frequency components at different scales. To realize this, the Multi-level Patch Embedding comprises level-specific convolutional embeddings whose kernel (patch) sizes scale with the corresponding downsampling factor (see Figure 2.19). With $N=4$, the patch sizes per level, from shallowest to deepest, are 8×8 , 4×4 , 2×2 , and 1×1 , producing the same token count per component at every level. The kernel size of each

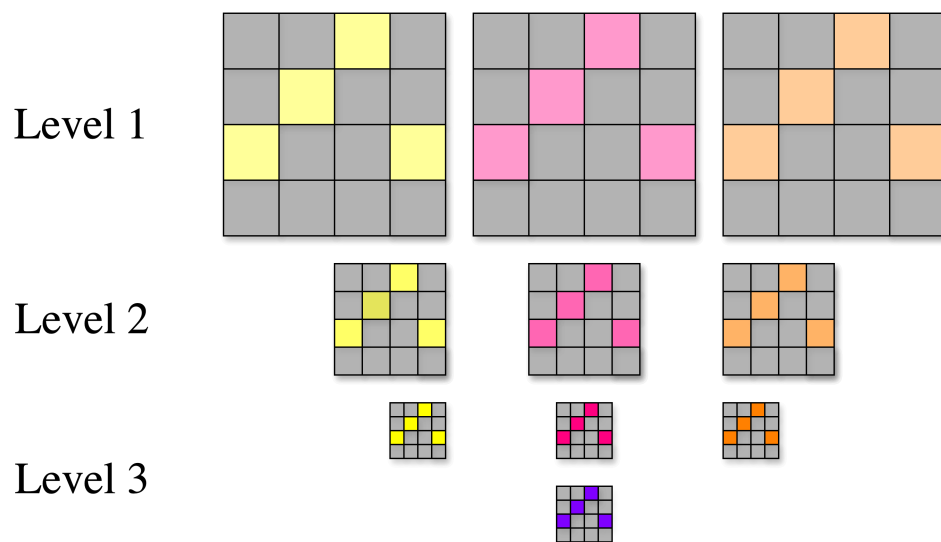


Figure 2.18: Tube Masking scheme: for a three-level DWT, spatially aligned locations are masked identically across all levels. From top to bottom, rows correspond to shallow-to-deep levels; components are colour-coded [LL, HL, LH, HH]; grey indicates masked patches.

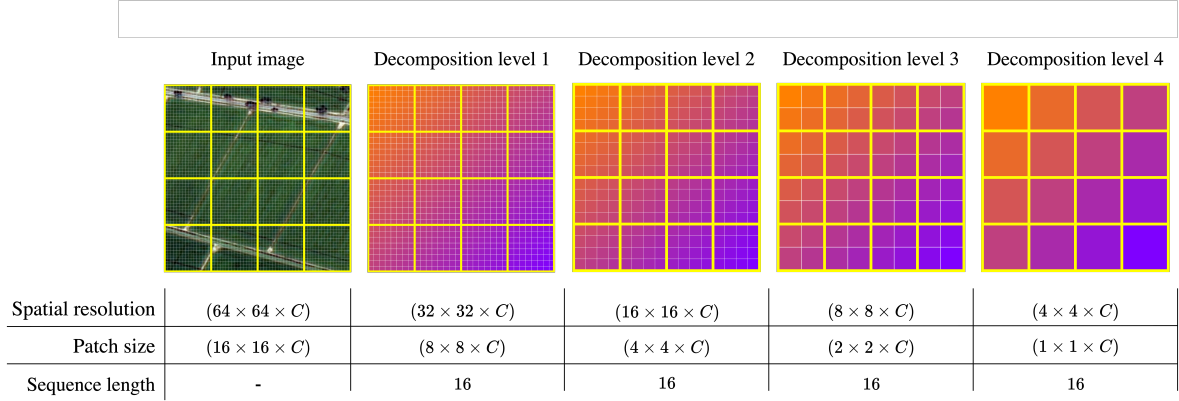


Figure 2.19: Mechanism of the Multi-level Patch Embedding. For an input of size 64×64 and $N=4$ levels, DWT subbands have the resolutions reported in the table. Adapting the patch size to each component’s spatial resolution aligns tokens at the same image location across levels. White indicates pixel grids; yellow outlines patch boundaries.

patch-embedding layer matches its patch size. This construction ensures that a token at any location aggregates the same 16×16 region in the original image x , regardless of level, guaranteeing location-wise correspondence across scales.

Positional Encoding

Before encoding, positional information is added to the token sequence. For absolute ordering, sinusoidal Absolute Positional Encoding (APE) is adopted due to its demonstrated effectiveness [12]. To introduce geospatial context, a Geo-Conditioned Positional Encoding (GPE) is further constructed using Spherical Harmonics, as depicted in Figure 2.20. This encourages samples from similar regions to occupy nearby locations in the representation space. For a pretraining image x with metadata (lat, lon),

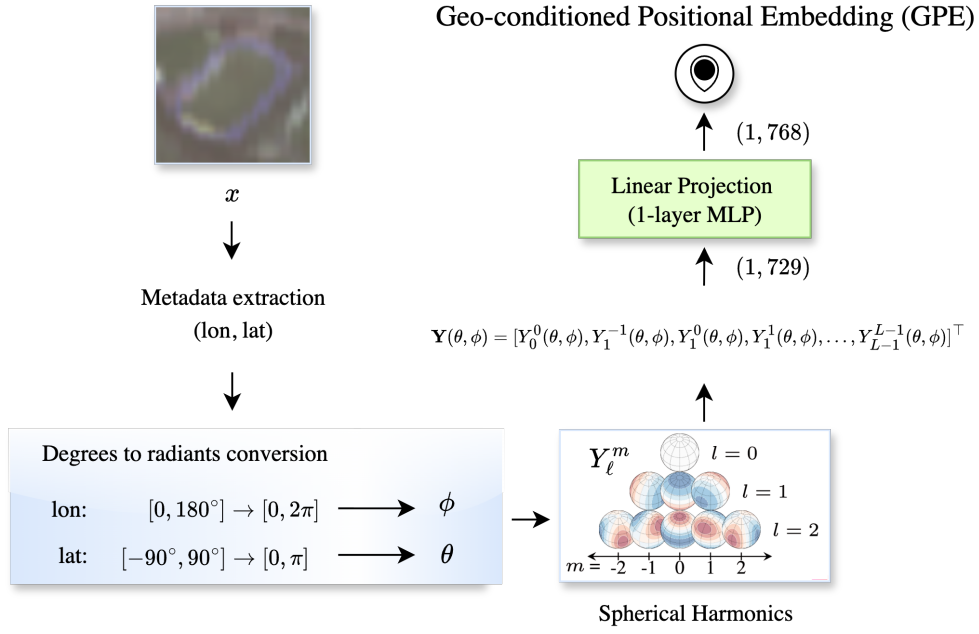


Figure 2.20: Computation of the Geo-conditioned Positional Encoding (GPE). Meta-data provide latitude/longitude for each image x ; after converting degrees to radians, Spherical Harmonics (Section 2.2.4) are evaluated and then projected with a 1-layer MLP to the encoder embedding dimension $D=768$.

angles are first mapped to radians:

$$\begin{aligned}\theta &= \text{deg2rad}(\text{lat} + 90^\circ) \in [0, \pi] \\ \phi &= \text{deg2rad}(\text{lon} + 180^\circ) \in [0, 2\pi]\end{aligned}$$

where θ is the polar angle and ϕ the azimuth. Spherical harmonics $Y_\ell^m(\theta, \phi)$ are then computed for degrees $\ell \in \{0, \dots, L-1\}$ and orders $m \in \{-\ell, \dots, \ell\}$. Increasing L augments angular resolution by adding higher degrees and orders [58], low degrees capture global trends, whereas high degrees model fine variations. A setting of $L=27$ is adopted, producing $\sum_{\ell=0}^{L-1} (2\ell+1) = L^2 = 729$ coefficients, concatenated as

$$\mathbf{Y}(\theta, \phi) = [Y_0^0(\theta, \phi), Y_1^{-1}(\theta, \phi), Y_1^0(\theta, \phi), Y_1^1(\theta, \phi), \dots, Y_{L-1}^{L-1}(\theta, \phi)]^\top \in \mathbb{R}^{729}.$$

A linear layer maps this vector to the transformer dimension $D=768$. The GPE is added to all tokens, complementing APE so that the embedding encodes both local spatial position and global geolocation cues.

Masking Strategy

The masking policy exploits the spatial alignment of tokens across all levels/components induced by the Multi-level Patch Embedding, enabling a Tube Masking protocol [59]. This approach reduces spatial redundancy while preserving cross-level frequency correspondence by applying identical spatial masks to every component (Figure 2.18). A masking ratio of 75% is adopted, in line with prior self-supervised work [3] [4]. Formally, define the token set

$$\mathcal{T} = \{t_{i,j}^l : i = 1, \dots, n_H, j = 1, \dots, n_W, l = 1, \dots, L\}$$

where (i, j) indexes spatial location, l the level, and n_H, n_W the token grid dimensions for x . Draw a random spatial mask

$$M \subseteq \{1, \dots, H\} \times \{1, \dots, W\} \text{ with } |M| = \rho HW \text{ and } \rho = 0.75.$$

Retained tokens are

$$T = \bigcup_{(i,j) \notin M} \{t_{i,j}^l : l = 1, \dots, L\}.$$

Compared with fully random masking, tube-wise masking led to faster convergence in preliminary trials: enforcing shared spatial locations across levels encourages the model to exploit the hierarchical wavelet structure and stabilizes training. The working hypothesis is that, for any unmasked location, attention can integrate information from all levels/components at that same site, making this association easier facilitates the convergence during training.

Multi-level Reconstruction

After encoding, embeddings are linearly projected to the decoder dimension $\hat{D}=512$ and learnable mask tokens are inserted at previously removed positions, following the MAE pipeline [2]. Absolute positional embeddings are then added to the full sequence. The Geo-conditioned Positional Encoding is applied only on the encoder side and is not provided to the decoder, preventing direct access to geolocation during reconstruction and compelling the encoder to internalize geographic cues. Figure 2.21 depicts the decoder layout. Unlike standard MAE, the target is not the pixel image $\bar{x}$, instead, the decoder estimates the set of wavelet components $\bar{S}$, from which $\bar{x}$ is obtained via the inverse transform:

$$\bar{x} = DWT_N^{-1}(\bar{S})$$

To this end, the final linear prediction stage comprises L level-specific heads (Pred modules), mirroring the Multi-level Patch Embedding, so that each head reconstructs all components at its level. Concatenating all predicted components yields $\bar{S}$.

Losses

Reconstructing $\bar{S}$ alone was found insufficient to guarantee high-fidelity images. Accordingly, two complementary losses are employed:

$$\mathcal{L}_{tot} = \mathcal{L}_{rec} + \mathcal{L}_{cmp} \quad (2.13)$$

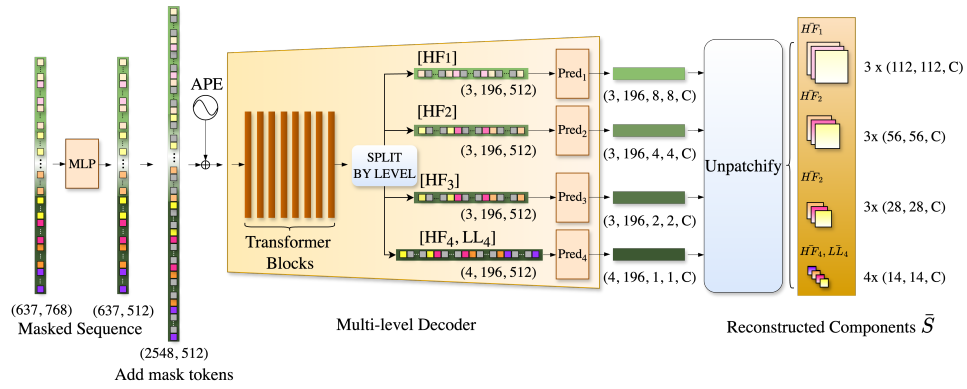


Figure 2.21: Multi-level Decoder. The masked sequence is first projected to the decoder width $\hat{D}=512$, then filled with learnable mask tokens. After adding APE, the complete sequence is processed by 8 transformer blocks. Four independent prediction heads (one per level) operate on level-specific subsequences obtained by splitting the decoded tokens. Owing to the Multi-level Patch Embedding, each component yields a fixed 196-token sequence per level. Thus, the fourth (deepest) level aggregates 784 tokens (for $HF_4=\{HL_4, LH_4, HH_4\}$ plus LL_4), while the remaining levels contain 588 tokens each (high-frequency components only). Each level's head comprises separate MLPs (three, or four at the deepest level), each dedicated to one component, mapping $512 \rightarrow p_H \times p_W$ where (p_H, p_W) depend on the level (Section 2.2.4). An unpatchify step finally produces the reconstructed component set $\tilde{S}$.

Both terms are computed only over masked spatial locations (defined by M). The reconstruction term $\mathcal{L}_{rec}$ compares the original image x with the inverse-DWT reconstruction $\bar{x} = DWT_4^{-1}(\bar{S})$:

$$\mathcal{L}_{rec} = \frac{1}{P} \sum_{i=1}^P (x_i - \bar{x}_i)^2$$

where P is the number of masked pixels. This term steers the encoder toward representations that are informative for the inverse DWT, encouraging high-frequency information to propagate even into shallower levels. The complementary loss $\mathcal{L}_{cmp}$ regularizes at the component level using a smooth L1 criterion between true and predicted components:

$$\mathcal{L}_{cmp} = \frac{1}{|M|} \sum_{p \in M} \text{SmoothL1}(S_p, \bar{S}_p),$$

which, in conjunction with GPE, empirically benefits several downstream tasks.

2.2.5 Experiments

A fair evaluation requires disentangling gains due purely to pretraining strategy from confounders such as dataset choice and scale. Accordingly, all compared methods were retrained on a single, shared pretraining corpus so that improvements can be attributed to the pretraining strategy rather than to data discrepancies. To rigorously test generalization, downstream evaluation relies on the PANGAEA benchmark [60], which offers a broad and standardized suite of tasks. Five datasets spanning four objectives—semantic segmentation, change detection, regression, and multilabel classification were selected. All methods used for the comparison are in the optical masked-autoencoder families for remote sensing imagery. Within this setting, the question under examination is whether explicitly injecting a wavelet decomposition benefits optical encoders by structuring high-frequency, scale-aware content that conventional MAE-style pipelines may underexploit. For this reason, the comparison is restricted to foundation models that (i) adopt masked autoencoding and (ii) are applicable to optical inputs only, thereby avoiding confounds from extra modalities such as SAR. The roster of methods includes MAE [2], the canonical formulation for natural images, together with SatMAE [3], which adapts MAE to multispectral data, and SatMAE++ [4], which augments reconstruction via multi-scale constraints. To broaden

the scope, SpectralGPT [5] is also considered as a large-scale transformer explicitly designed for multispectral structure through spectral priors. Collectively, these baselines span general-purpose and domain-specialized masked autoencoders, providing a balanced reference frame to assess the contribution of WaveMAE.

Datasets The empirical study focuses on optical—and specifically multispectral—imagery, which has demonstrated advantages beyond RGB alone. Consequently, fMoW-S2 [6] serves as the common pretraining dataset for every method assessed in this section.

fMoW-S2 (Functional Map of the World – Sentinel-2) constitutes the Sentinel-2 variant of the IARPA fMoW collection. It comprises 882,779 image patches sampled across hundreds of countries and provides the full Sentinel-2 band set (B1–B12 plus B8A; 13 bands) over 62 categories. These include four 10 m bands (B02, B03, B04, B08), six 20 m bands (B05, B06, B07, B8A, B11, B12), and three 60 m bands (B01, B09, B10).

Transfer capability is measured under the PANGAEA protocol [60], finetuning the pretrained encoders on five downstream datasets covering four tasks and multiple sensors/resolutions:

MADOS. MADOS [61] targets semantic segmentation for marine pollution (oil spills, marine debris) using Sentinel-2 data. It supplies the multispectral imagery (B1, B2, B3, B4, B5, B6, B7, B8, B8A, B11, B12; 11 bands), enabling evaluation of spectral sensitivity and fine-grained delineation in maritime scenes.

HLSBurnScars. HLSBurnScars [62] contains burn-scar segmentation scenes derived from Harmonized Landsat–Sentinel products. A typical subset (B2, B3, B4, B8A, B11, B12; six bands) is provided at HLS harmonized resolutions. The public release includes 804 scenes at 512×512 pixels (2018–2021, CONUS), supporting tests of robustness under cross-sensor harmonization.

SpaceNet7 (MUDS). SpaceNet7 [63]—the Multi-temporal Urban Development SpaceNet challenge—addresses urban change detection with PlanetScope RGB stacks ($\sim 3\text{--}4\text{ m}$ GSD) and building labels. Its standardized monthly stacks and area-of-interest (AOI) patches facilitate temporal change analysis and medium-resolution localization.

BioMassters. BioMassters [64] is a regression benchmark for above-ground biomass that combines Sentinel-1 (VV, VH) and Sentinel-2 time series with LiDAR-derived references in Finland. The optical stream includes B2, B3, B4, B5, B6, B7, B8, B8a, B11, B12 (12 bands). The dataset comprises 310,000 patches (each covering roughly $2,560 \times 2,560$ m) with per-patch biomass targets.

BigEarthNet. BigEarthNet [65] is a large-scale multilabel scene classification dataset built from Sentinel-2 Level-2A (B1–B12, B8A; 13 bands). Version 2 contains 549,488 patches, typical pre-processing maps 10m bands to $\sim 120 \times 120$ pixels (with matched sizes for 20m, 60m). Labels cover 19 classes in a multilabel setting.

Together, these datasets span heterogeneous resolutions, band sets, and objectives, yielding a comprehensive testbed for transfer evaluation. Since the study targets optical scenarios exclusively, Sentinel-1 patches are discarded during finetuning, Sentinel-2 multispectral or RGB imagery is retained as available.

Pretraining Details

All contenders are pretrained on fMoW-S2 [6]. Following [5], pre-processing normalizes reflectances to $[0, 1]$, applies a random crop with scale sampled uniformly in $[0.2, 1.0]$, resizes to $I = (224, 224)$, and uses horizontal flip with probability 0.5. The spectral dimensionality is fixed at $C = 13$ for every method except Spectral-GPT [5], which adopts $C = 12$ (excluding B1). All models are trained for 50 epochs with AdamW with learning rate set to 10^{-4} . WaveMAE includes an additional step: a level- $N=4$ Discrete Wavelet Transform is computed (Haar basis), producing 13 components that are stacked as the input sequence. For baseline MAE-style encoders, the patch size is $P = (16, 16)$, yielding $(I_h/p_h) \times (I_w/p_w) = 196$ tokens. WaveMAE employs a multi-level patch embedding (Section 2.2.3), that means components at each level are downsampled by a factor of $2 \times$ w.r.t. the original resolution. The base patch size is scaled to $P = (8, 8)$ so that each token corresponds to a $(16, 16)$ region in the original image (refer to Fig. 2.19 for a visual explanation). Geo-coordinates for the Geo-conditioned Positional Encoding are parsed from the fMoW-S2 GeoTIFF meta-data and processed as in Section 2.2.3. Unless otherwise noted, the masking ratio is 75% across all methods, in line with [2]. Pretraining runs are executed on 2 NVIDIA L40S GPUs under matched conditions.

Downstream Tasks

Transfer is assessed on the PANGAEA suite [60]. Only encoder weights are reused; task-specific heads are trained from scratch with aligned hyperparameters to remove optimization confounds. For dense prediction, intermediate features from encoder layers $\{3, 5, 7, 11\}$ are aggregated as decoder inputs following established protocols. In all methods except WaveMAE and MAE, each output layer has shape (B, N, L, D) , features are summed over the N (spectral-group) axis, layer-normalized over D , then permuted and reshaped to (B, D, f_h, f_w) with $f_h = I_h/p_h$ and $f_w = I_w/p_w$. WaveMAE preserves (B, N, L, D) where N denotes wavelet components. Low-frequency (LL) and high-frequency (HF) subsets are normalized independently to respect their distinct statistics, HF components are then summed over N and normalized. The LL and aggregated-HF streams are concatenated to $(B, 2, L, D)$, reduced by summation over the component axis, and layer-normalized to obtain (B, L, D) for 2D map reconstruction. Geo-conditioned Positional Encoding is applied only when downstream metadata provide geolocation, thus, for MADOS and BioMassters no GPE is injected during finetuning, limiting its effect to pretraining. Standard MAE retains $N=1$ and requires no special handling. All downstream experiments are run on a single NVIDIA L40S GPU for consistency.

Semantic Segmentation. UperNet [66] is adopted following PANGAEA. Training uses cross-entropy and AdamW (learning rate 1×10^{-4} , betas $(0.9, 0.999)$, weight decay 0.05), batch size 32, for 80 epochs with validation every 20 epochs; the best checkpoint is selected on validation performance. Sliding-window inference is used for inputs larger than the model window. A multi-step scheduler triggers at $[0.6, 0.9]$ of the total epochs. Mean intersection over union (mIoU) is used as a metric to assess model’s performance. This protocol is applied to MADOS and HLSBurnScars datasets.

Regression. A regression UperNet head is employed with the same optimizer, scheduler, and feature aggregation used for segmentation. Training spans 50 epochs with validation every 10, batch size 32, and MSE loss. Performance is reported as RMSE on the test split using the best validation checkpoint. This setup is used for BioMassters dataset.

Multilabel scene classification. For multilabel classification, the 2D decoder is replaced by a compact MLP (Linear $\rightarrow$ BatchNorm1d $\rightarrow$ ReLU $\rightarrow$ Linear) operating on the highest-level encoder features. Training proceeds for 20 epochs with validation every 5, batch size 256, and Binary Cross Entropy loss; optimizer and scheduler mirror the dense tasks. In this case the metric used to assess performance is mean average precision (mAP). This configuration is applied to BigEarthNet dataset.

Protocols, optimizers, and feature usage are standardized across tasks, with only the task head varying by objective. This alignment permits attribution of differences to the pretrained representations of MAE, SatMAE, SatMAE++, SpectralGPT, and WaveMAE rather than to training heuristics.

Ablation Study

Ablations are designed to quantify the contribution of each architectural component. For every configuration, a single pretrained model is evaluated across the five downstream datasets listed in Section 2.2.5. Results are reported as the average over five independent runs with different seeds, limiting sensitivity to initialization and training noise.

To keep the ablation phase computationally tractable while preserving representational capacity, a base patch size of (16,16) is adopted at the first decomposition level (except where token size is explicitly studied in Section 2.2.5), scaling with component resolution as in Section 2.2.4. For inputs of 224×224 , first-level components are 112×112 ; with 16×16 patches, this yields $7 \times 7 = 49$ spatial tokens per level (versus 196 tokens in the final configuration).

The analysis proceeds through five questions in sequence:

1. the effect of Discrete Wavelet Transform depth, probing hierarchical multi-resolution representations;
2. the impact of Geo-conditioned Positional Encoding (GPE) relative to standard positional schemes;
3. sensitivity to the masking ratio, trading reconstruction difficulty against representation quality;

4. benefits of pretraining compared to random initialization;
5. the role of token size in multi-level patch embedding and the resulting optimal spatial resolution.

Decomposition level

To determine an effective depth for the wavelet hierarchy in the WaveMAE pipeline, this study varies the number of decomposition levels and measures transfer on the full benchmark suite. This choice is very important because the decomposition depth governs both the granularity of frequency separation and the reconstruction difficulty posed during pretraining. With fewer levels, the representation exposes fewer frequency components and thus provides a weaker basis for capturing the spectral–spatial variability of multispectral optical data, which in turn increases reconstruction ambiguity. Conversely, deeper decompositions yield a richer, explicitly hierarchical set of components that supports learning of multi-scale semantics, coupling fine high-frequency details with broader low-frequency structures. The qualitative behavior in Figure 2.22 mirrors this intuition: moving from shallow to deeper decompositions progressively strengthens the pathway by which high-frequency evidence guides the recovery of finer-scale content. Empirically, the results in Table 2.5 indicate a marked improvement when adopting 3–4 levels over 1–2 levels, underscoring the importance of hierarchical factorization for robust representation learning in remote sensing. Improvements are consistent with an enhanced capacity to encode spatial patterns across scales and to model frequency-domain relations characteristic of Earth observation scenes, where semantics span from local texture to regional structure. A fifth level was not pursued: memory overhead would rise over the hardware availability in the configuration used in this comparison, moreover, the gains beyond the 3-to-4 level transition appear marginal, making four levels a practical optimum under the resource envelope considered.



Figure 2.22: Qualitative reconstructions at epoch 50 for four pretraining configurations differing only by the DWT decomposition level. For each level, the *top* row shows (left-to-right) the target image, the visible (unmasked) region, and the target high-frequency component for the red channel (fMoW-S2 band 2). The *bottom* row reports the corresponding model outputs: the reconstructed image, the previously masked area recovered by the model, and the reconstructed high-frequency component for the same spectral band. At levels 1–3 the encoder exhibits limited recovery of fine high-frequency structure and instead exploits low-frequency cues for approximation. At level 4, by contrast, the lower-resolution high-frequency bands become easier to model, and these coarse HF cues assist reconstruction of inner (higher-resolution) components. Overall, the figure highlights the increasing difficulty of restoring high-frequency content at finer scales and the stabilizing role of coarse HF signals during reconstruction.

Table 2.5: Ablation study on WaveMAE method applying an increasing level of decomposition of the DWT. Each model has been tested across five datasets. Results reported as mean $\pm$ standard deviation over 5 runs. Best result per column is **bold**, second best is underlined. Results are mean $\pm$ std computed over 5 independent runs with different seed.

Decomposition level	#Params (Encoder only)	Mados mIoU $\uparrow \pm$ std	HLSBurnScars mIoU $\uparrow \pm$ std	BioMassters mRMSE $\downarrow \pm$ std	Spacenet7 mIoU $\uparrow \pm$ std	BigEarthNet mAP $\uparrow \pm$ std
1 Level	85.21M	43.421 $\pm$ 1.543	83.123 $\pm$ 0.726	125.107 $\pm$ 0.077	42.633 $\pm$ 3.787	42.710 $\pm$ 0.423
2 Levels	85.32M	43.323 $\pm$ 0.849	83.272 $\pm$ 0.664	124.167 $\pm$ 0.334	42.222 $\pm$ 1.662	43.206 $\pm$ 0.211
3 Levels	85.43M	<u>46.840 $\pm$ 1.980</u>	84.703 $\pm$ 0.291	<u>123.741 $\pm$ 0.151</u>	<u>44.582 $\pm$ 1.981</u>	44.224 $\pm$ 0.468
4 Levels	85.55M	50.401 $\pm$ 2.963	<u>83.294 $\pm$ 0.918</u>	123.244 $\pm$ 0.122	48.280 $\pm$ 2.292	<u>43.760 $\pm$ 0.657</u>

Geo-conditioned Positional Encoding

The effect of Geo-conditioned Positional Encoding (GPE) is examined both intrinsically via alignment of representations with geodesic distance and extrinsically on downstream transfer. GPE injects latitude/longitude priors through Spherical Harmonics, encouraging embeddings from nearby locations to concentrate in feature space while preserving semantic coherence. Accordingly, a stronger monotonic relation is expected between pairwise GPE encodings and great-circle distance, together with clearer separation between geographically close and distant samples. In conjunction with the component loss $\mathcal{L}_{cmp}$, such structure is hypothesized to translate into consistent downstream improvements.

GPE-pairs. This test verifies that GPE encodes location information by correlating great-circle distances with similarities of the learned GPE vectors. Approximately 4k coordinate pairs $pair_j(p_1, p_2)$ are sampled from fMoW-S2, balanced between *close* (< 200 km) and *far* (> 200 km) regimes, with $p_i = (\phi_i, \theta_i)$. Great-circle distance is computed as

$$\Delta_{GCD}(p_1, p_2) = R \cdot \arccos(\sin \phi_1 \sin \phi_2 + \cos \phi_1 \cos \phi_2 \cos(\theta_2 - \theta_1))$$

with Earth radius $R=6371$ km. For each point, a GPE vector is evaluated, and the co-

sine similarity $\cosSim(GPE(p_1), GPE(p_2))$ is correlated (Spearman correlation) with $\Delta_{GCD}(p_1, p_2)$. Two settings are considered: GPE alone, and GPE plus $\mathcal{L}_{cmp}$. Table 2.6 shows a pronounced gap in cosine similarity between *close* and *far* pairs in both settings, and a strong negative correlation with distance, confirming that GPE captures geospatial structure.

GPE-embeddings. A second probe constructs $\sim 3k$ balanced tuples over the 62 fMoW-S2 categories. For each Anchor (A), four samples are paired: a Positive (P) within 200 km but from a different class, an Easy Positive (EP) within 200 km and same class, a Negative (N) beyond 200 km and same class, and an Easy Negative (EN) beyond 200 km and different class. Embeddings are extracted with frozen encoders under four configurations: no GPE / no $\mathcal{L}_{cmp}$, GPE only, $\mathcal{L}_{cmp}$ only, and GPE+ $\mathcal{L}_{cmp}$. Average cosine distances between A and {P, EP, N, EN} are reported in Table 2.7. This evaluates whether geolocation cues learned during pretraining modulate representation geometry beyond visual similarity alone.

Insights. As previously demonstrated, GPE induces a strong negative correlation between cosine similarity and geographical distance, with consistently high similarity for geographically close pairs of locations and substantially reduced similarity for distant ones. Comparing the two configurations without GPE in Table 2.8, one might observe that the best downstream performance corresponds to smaller margins between Positive and Negative embeddings, as highlighted in Table 2.7. This suggests that a more homogeneous representation space may be advantageous for transfer learning. Conversely, the strong bias introduced by GPE induces the largest separation between Positives and Negatives (Table 2.7), this separation however does not translate into improved downstream performance, rather the opposite, as shown in Table 2.8. The hypothesis is that the unregulated shift of tokens in the embedding space caused by GPE is excessively strong, impairing the model’s reconstruction ability and thus degrading the learned representations. This hint is the reason behind the introduction of a regularization loss to enforce reconstruction consistency. By introducing the loss $\mathcal{L}_{cmp}$ the overly strong bias introduced by GPE was mitigated while retaining its semantics, thereby allowing the injection of geolocation priors without excessively disrupting the latent space. Although the resulting margin remains

slightly larger than that of $\mathcal{L}_{cmp}$ only configuration, the embedded geographic priors yield significant improvements in downstream tasks across segmentation, regression, and change-detection tasks, as summarized in Table 2.8.

Table 2.6: Intrinsic GPE metrics (GPE-pairs) on frozen encoders. Reported are mean cosine similarity for close/far pairs and Spearman correlation between cosine similarity and Δ_{GCD} (km).

GPE	$\mathcal{L}_{cmp}$	Close pairs mean sim. $\uparrow$	Far pairs mean sim. $\downarrow$	Spearman corr. ($\cos Sim$ vs Δ_{GCD})
✓		0.992	0.401	-0.81
✓	✓	0.995	0.592	-0.80

Masking ratio

The masking ratio sets the proportion of tokens removed during pretraining and therefore controls reconstruction difficulty. Table 2.9 compares three settings—60%, 75%, and 90%—across all downstream tasks.

Across datasets, a 75% mask emerges as the most reliable operating point, typically delivering the best or second-best scores. Notably, this configuration yields strong gains on Mados, BioMassters, and Spacenet7, while remaining competitive on HLSBurnScars and BigEarthNet. A 60% mask performs well on HLSBurnScars and shows competitive behavior on BioMassters but lags elsewhere, pushing to 90% produces results close to 75% with slightly reduced variance yet generally lower means.

These trends emphasize careful tuning of the masking ratio: too little masking makes the task trivial and weakens representation quality, whereas overly aggressive masking diminishes usable context. Under the present setup, 75% strikes a balanced compromise, encouraging learning of robust, transferable features.

Table 2.7: Intrinsic GPE metrics (Geo-conditioned embeddings) on frozen encoders. Reported are mean cosine distances between anchor (A) and positives/negatives.

GPE	$\mathcal{L}_{cmp}$	(A,P)	(A,EP)	(A,N)	(A,EN)
		$dist_{cos} \downarrow$	$dist_{cos} \downarrow$	$dist_{cos} \uparrow$	$dist_{cos} \uparrow$
		0.119	0.084	0.138	0.169
	✓	0.101	0.072	0.106	0.123
✓		0.111	0.079	0.137	0.163
✓	✓	0.115	0.082	0.128	0.148

Table 2.8: Ablation on the effect of GPE and $\mathcal{L}_{cmp}$ on downstream tasks. Results reported as mean $\pm$ standard deviation over 5 runs. Best result per column is **bold**, second best is underlined. Results are mean $\pm$ std computed over 5 independent runs with different seed.

GPE	$\mathcal{L}_{cmp}$	#Params (Encoder only)	Mados mIoU $\uparrow \pm$ std	HLSBurnScars mIoU $\uparrow \pm$ std	BioMassters mRMSE $\downarrow \pm$ std	Spacenet7 mIoU $\uparrow \pm$ std	BigEarthNet mAP $\uparrow \pm$ std
		85.55M	<u>50.401 $\pm$ 2.963</u>	<u>83.294 $\pm$ 0.918</u>	123.244 $\pm$ 0.122	48.280 $\pm$ 2.292	43.760 $\pm$ 0.657
	✓	85.55M	47.694 $\pm$ 1.934	83.288 $\pm$ 0.650	<u>123.071 $\pm$ 0.168</u>	<u>48.574 $\pm$ 2.875</u>	43.838 $\pm$ 0.334
✓		85.55M	48.997 $\pm$ 2.272	82.705 $\pm$ 1.056	123.516 $\pm$ 0.120	47.586 $\pm$ 1.779	43.364 $\pm$ 0.521
✓	✓	85.55M	52.258 $\pm$ 2.549	83.565 $\pm$ 1.353	122.843 $\pm$ 0.173	50.703 $\pm$ 2.216	<u>43.830 $\pm$ 0.483</u>

Pretrain effectiveness

The contribution of pretraining is quantified against a randomly initialized encoder under identical finetuning protocols. Table 2.10 reports downstream results for the full WaveMAE configuration when initialized from fMoW-S2 versus training from scratch.

Across benchmarks, pretraining confers a systematic advantage. Notably, the gains are substantial on *Mados* (+11.4 mIoU), *Spacenet7* (+8.5 mIoU), and *BioMassters* (-1.46 mRMSE), indicating that the features learned during self-supervision carry over effectively to dense prediction and regression. The improvement on *HLSBurnScars* is modest yet positive, whereas *BigEarthNet* exhibits parity with random ini-

Table 2.9: Ablation on masking ratio applied to the sequence in input to the encoder during pretraining. Each model has been tested across five datasets. Results reported as mean $\pm$ standard deviation over 5 runs. Best result per column is **bold**, second best is underlined. Results are mean $\pm$ std computed over 5 independent runs with different seed.

Mask ratio	#Params (Encoder only)	Mados mIoU $\uparrow \pm$ std	HLSBurnScars mIoU $\uparrow \pm$ std	BioMassters mRMSE $\downarrow \pm$ std	Spacenet7 mIoU $\uparrow \pm$ std	BigEarthNet mAP $\uparrow \pm$ std
60%	85.55M	44.571 $\pm$ 1.161	83.586 $\pm$ 0.617	<u>123.108 $\pm$ 0.100</u>	49.061 $\pm$ 2.025	43.602 $\pm$ 0.763
75%	85.55M	52.258 $\pm$ 2.549	<u>83.565 $\pm$ 1.353</u>	122.843 $\pm$ 0.173	50.703 $\pm$ 2.216	43.830 $\pm$ 0.483
90%	85.55M	<u>51.266 $\pm$ 0.660</u>	83.540 $\pm$ 0.249	123.162 $\pm$ 0.149	<u>50.108 $\pm$ 2.168</u>	<u>43.818 $\pm$ 0.602</u>

Table 2.10: Ablation on pretrain effectiveness against a random initialization considering the full WaveMAE architecture. Each model has been tested across five datasets. Results reported as mean $\pm$ standard deviation over 5 runs. Best result per column is **bold**. Results are mean $\pm$ std computed over 5 independent runs with different seed.

Pretrain Dataset	#Params (Encoder only)	Mados mIoU $\uparrow \pm$ std	HLSBurnScars mIoU $\uparrow \pm$ std	BioMassters mRMSE $\downarrow \pm$ std	Spacenet7 mIoU $\uparrow \pm$ std	BigEarthNet mAP $\uparrow \pm$ std
Random Init.	85.55M	40.816 $\pm$ 1.560	81.932 $\pm$ 0.838	124.303 $\pm$ 0.097	42.215 $\pm$ 1.503	43.842 $\pm$ 0.429
fMoW-S2	85.55M	52.258 $\pm$ 2.549	83.565 $\pm$ 1.353	122.843 $\pm$ 0.173	50.703 $\pm$ 2.216	43.830 $\pm$ 0.483

tialization, consistent with the intuition that large, label-rich classification corpora benefit less from pretraining when finetuned with abundant supervision.

Taken together, these observations support the conclusion that pretraining on a large and diverse optical corpus such as *fMoW-S2* enhances transfer, with the strongest benefits emerging on tasks where supervision is limited or signal structure is complex. Random initialization, by contrast, struggles to attain comparable generalization, underscoring pretraining as a key factor in representation quality and robustness.

2.2. WaveMAE: Wavelet decomposition Masked Auto-Encoder for Remote Sensing

87

Table 2.11: Ablation on token spatial size, results shows that increasing the granularity of the input tokens leads to better performance overall. Each model has been tested across five datasets. Results reported as mean $\pm$ standard deviation over 5 runs. Best result per column is **bold**. Results are mean $\pm$ std computed over 5 independent runs with different seed.

Patch size	#Params (Encoder only)	Mados mIoU $\uparrow \pm$ std	HLSBurnScars mIoU $\uparrow \pm$ std	BioMassters mRMSE $\downarrow \pm$ std	Spacenet7 mIoU $\uparrow \pm$ std	BigEarthNet mAP $\uparrow \pm$ std
16	85.55M	52.258 $\pm$ 2.549	83.565 $\pm$ 1.353	122.843 $\pm$ 0.173	50.703 $\pm$ 2.216	43.830 $\pm$ 0.483
8	85.91M	64.550 $\pm$ 3.307	85.595 $\pm$ 0.198	106.383 $\pm$ 0.170	52.303 $\pm$ 1.987	45.278 $\pm$ 0.376

Token size

The spatial granularity of input tokens is ablated by comparing encoders pretrained with patches of 16×16 versus 8×8 , while preserving the rest of the architecture. Table 2.11 summarizes the results.

Reducing the patch size consistently improves transfer across all tasks. The switch from 16 to 8 yields sizable gains on *Mados* (+12.3 mIoU), *HLSBurnScars* (+2.0 mIoU), and *BioMassters* (-16.5 mRMSE). Even for *Spacenet7* and *BigEarthNet*, where sensitivity to very fine details is generally lower, the finer tokenization still confers measurable benefits.

These outcomes indicate that preserving higher spatial resolution at the token level facilitates the capture of high-frequency structure and subtle local variations, which benefits both dense and global objectives. The computational overhead associated with a larger token count is moderate relative to the observed improvements, suggesting that 8×8 patching represents a favorable accuracy–efficiency compromise for optical remote sensing encoders.

Comparison with State-of-the-Art

A head-to-head evaluation is conducted against representative optical SSL baselines—MAE [2], SatMAE [3], SatMAE++ [4], and SpectralGPT [5], all pretrained on fMoW-S2 [6] with harmonized settings. Table 2.12 reports performance over five downstream bench-

Table 2.12: Comparison of WaveMAE against previous state-of-the-art models using fMoW-S2 [6] as pretraining dataset for all methods. Each model has been tested across five datasets. Results reported as mean $\pm$ standard deviation over 5 runs. Best result per column is **bold**, second best is underlined. Results are mean $\pm$ std computed over 5 independent runs with different seed.

Encoder type	#Params (Encoder only)	Mados mIoU $\uparrow \pm$ std	HLSBurnScars mIoU $\uparrow \pm$ std	BioMassters mRMSE $\downarrow \pm$ std	Spacenet7 mIoU $\uparrow \pm$ std	BigEarthNet mAP $\uparrow \pm$ std
MAE [2]	87.76M	25.414 $\pm$ 2.083	83.163 $\pm$ 0.329	123.870 $\pm$ 3.887	41.416 $\pm$ 1.515	<u>45.000 $\pm$ 0.361</u>
SatMAE [3]	87.03M	47.091 $\pm$ 3.094	<u>83.626 $\pm$ 0.274</u>	<u>109.040 $\pm$ 0.212</u>	45.115 $\pm$ 1.231	44.874 $\pm$ 0.401
SatMAE++ [4]	87.9M	39.842 $\pm$ 2.839	81.744 $\pm$ 0.472	110.358 $\pm$ 0.273	55.228 $\pm$ 1.126	43.096 $\pm$ 0.290
SpectralGPT [5]	85.8M	<u>51.143 $\pm$ 0.926</u>	81.734 $\pm$ 0.822	109.201 $\pm$ 0.295	49.100 $\pm$ 2.893	44.256 $\pm$ 0.145
WaveMAE-Small	22.7M	61.665 $\pm$ 3.422	84.063 $\pm$ 1.009	108.008 $\pm$ 0.146	51.227 $\pm$ 3.770	- $\pm$ -
WaveMAE-Base	85.91M	64.550 $\pm$ 3.307	85.595 $\pm$ 0.198	106.383 $\pm$ 0.170	<u>52.303 $\pm$ 1.987</u>	45.278 $\pm$ 0.376
		(+26.21%)	(+2.35%)	(+2.5%)	(-5.29%)	(+0.61%)

marks. To avoid confounding by input granularity, the comparison employs an input size of $224 \times 224 \times C$ with a patch size of 16, in WaveMAE this matches an effective ratio of $112 \times 112 \times C$ with base patch size 8 at level 1 of the decomposition.

In the base configuration, WaveMAE attains the top results on four of five tasks. The improvements are most pronounced on *Mados* (+26.2% mIoU) and *HLSBurnScars* (+2.35% mIoU), and a new state-of-the-art is set on *BioMassters* (-2.5% mRMSE). On *BigEarthNet*, a consistent albeit smaller margin is observed (+0.61% mAP). *Spacenet7* remains an outlier where SatMAE++ leads, plausibly due to its stronger temporal modeling, which is particularly beneficial for change detection. A compact WaveMAE-Small variant is also reported, maintaining competitive accuracy with a parameter budget of only 26.4% of the base model. Overall, the evidence supports the conclusion that explicitly modeling scale-aware high-frequency content via wavelet decomposition yields more transferable optical representations than standard MAE-style pretraining. A qualitative comparison is provided in Figures 2.23, 2.24 and 2.25.

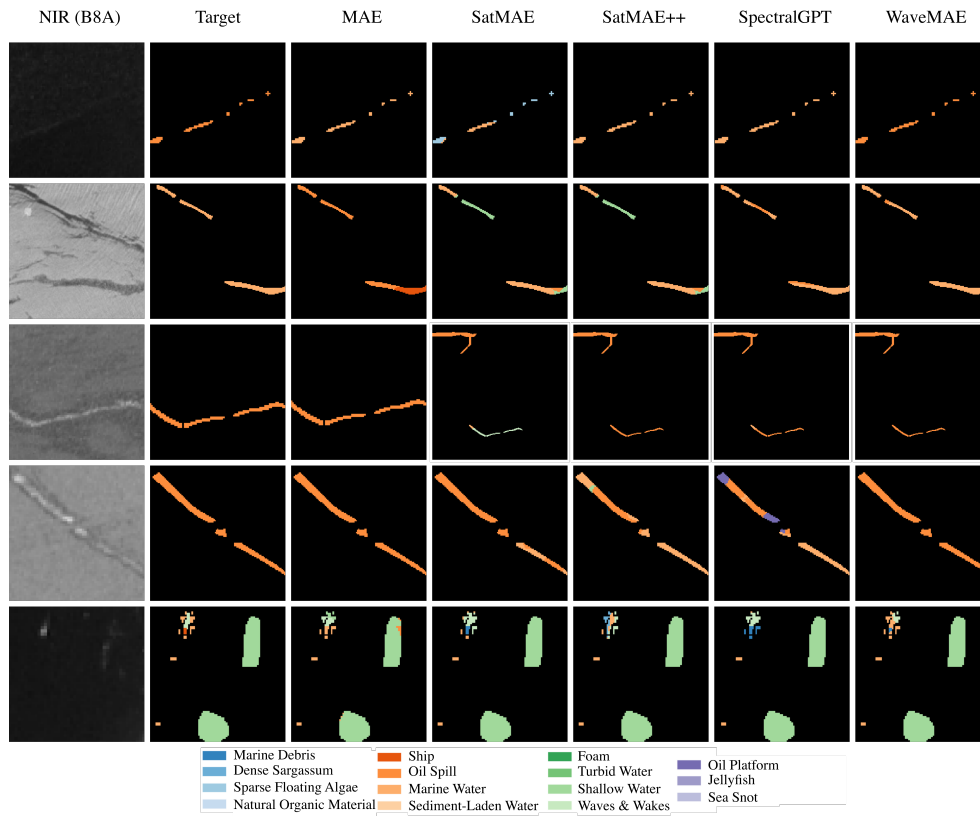


Figure 2.23: Qualitative comparison of semantic segmentation task performance on Mados dataset across different SSL methods. From left to right the figure illustrates the NIR band (B8A) of the input image, the target segmentation map, and following the predicted segmentation maps respectively produced by MAE [2], SatMAE [3], SatMAE++ [4], SpectralGPT [5] and finally **WaveMAE**. The labels on the bottom indicate the dataset’s classes and the respective color assigned in the segmentation map. (Best viewed zoomed in for fine details.)

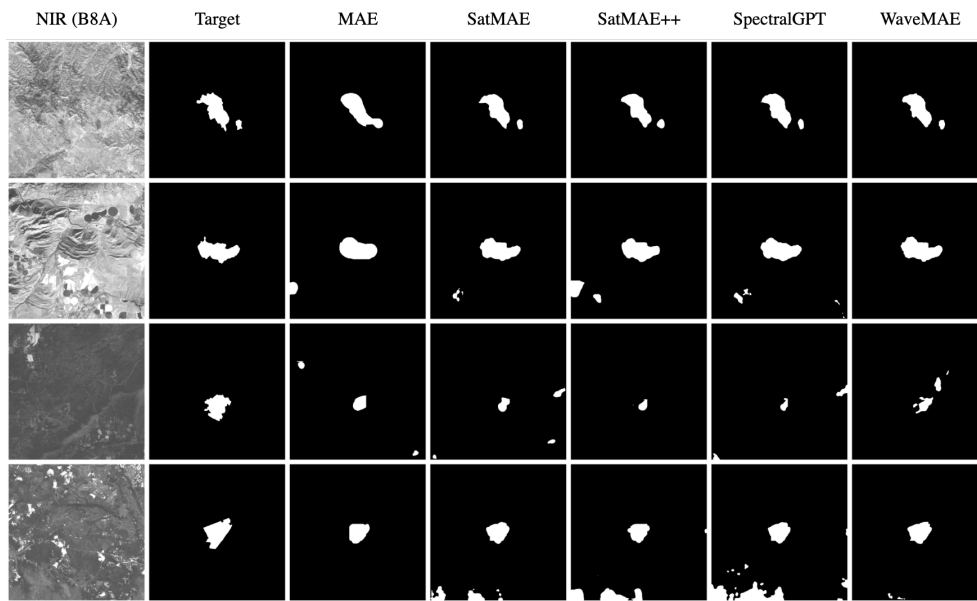


Figure 2.24: Qualitative comparison of semantic segmentation task performance on HLSBurnScars dataset across different SSL methods. From left to right the figure illustrates the NIR band (B8A) of the input image, the target segmentation map, and following the predicted segmentation maps respectively produced by MAE [2], SatMAE [3], SatMAE++ [4], SpectralGPT [5] and finally **WaveMAE**. White color is assigned to show the presence of a burn scar in the image.

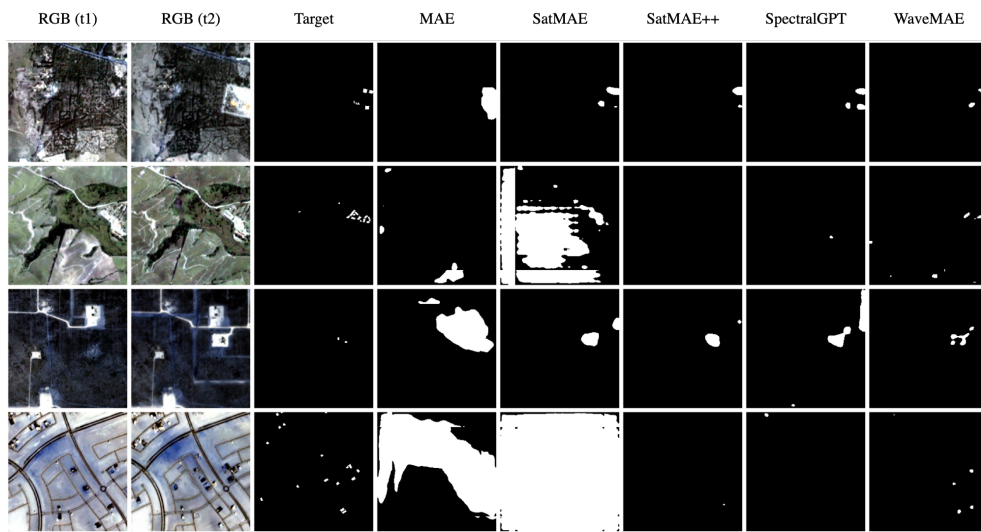


Figure 2.25: Qualitative comparison of change-detection task performance on Spacenet7 (MUDS) dataset across different SSL methods. From left to right the figure illustrates the RGB image at timestep t_1 , the RGB image at timestep t_2 with $t_1 < t_2$, the target segmentation map, and following the predicted segmentation maps respectively produced by MAE [2], SatMAE [3], SatMAE++ [4], SpectralGPT [5] and finally **WaveMAE**. The white color in segmentation maps corresponds to areas where urban development occurred between the two RGB images.

2.2.6 Conclusion

This study presents WaveMAE, a self-supervised framework for multispectral optical imagery that augments masked autoencoding with Discrete Wavelet Transform decomposition and Geo-conditioned Positional Encoding (GPE). By factorizing content into multi-resolution frequency components exploiting DWT and injecting geospatial priors through Spherical Harmonics, the approach enables more faithful reconstruction targets and leads to richer, geographically structured representations.

Extensive evaluation on the PANGAEA-bench shows consistent improvements over prior foundation models across semantic segmentation, change detection, regression, and multilabel classification. Ablations studies the role of each novelty in the architecture and motivates the design choices: wavelet decomposition enhances high-frequency capture and reconstruction fidelity, GPE organizes the embedding space with respect to geography and, when coupled with the component loss, improves transfer. Analysis of masking ratio, token size, and pretraining effectiveness further demonstrate robustness to hyperparameter choices and clear gains over random initialization.

Overall, the evidence indicates that frequency-aware decomposition and geographically informed encodings are effective drivers of self-supervised learning in remote sensing. The unified evaluation protocol also contributes to more standardized comparisons in the field, facilitating objective progress measurement across optical SSL methods.

Chapter 3

Conclusions and Future Work

3.1 Conclusions

This thesis contributes to two complementary areas: satellite image supervised single image super-resolution and self-supervised masked image modeling for remote sensing.

In the supervised setting, Swin2-MoSE shows that sparsely gated transformer backbones, coupled with example-level routing and a loss term based on Normalized Cross-Correlation (NCC), can preserve fine structures while maintaining cross-band consistency. The design encourages experts to specialize without fragmenting the feature space, yielding sharper edges and more faithful spectra than conventional CNN or dense-transformer baselines. Swin2-MoSE relies on a balanced expert utilization regime: excessive sparsity may underfit rare patterns, whereas insufficient sparsity can dilute specialization.

In the self-supervised regime, WaveMAE demonstrates that frequency-aware masked

modeling—implemented via discrete wavelet decomposition and complemented by a light geospatial positional encoding—produces transferable representations. The learned features generalize across segmentation, regression, and multilabel classification, and often outperform larger models trained to reconstruct pixels directly. The evidence indicates that aligning self-supervision with the signal’s natural multi-scale structure is an effective inductive bias for multispectral data. WaveMAE’s gains depend on appropriate choices of wavelet depth and token granularity, overly shallow decompositions miss fine structure, while deeper ones increase memory and may offer diminishing returns. Geo-conditioned encodings provide consistent benefits when geotags are reliable, and can have benefits also in cases when metadata are not available.

3.2 Future Work

Several possibilities emerge for extending this research. While the current Swin2-MoSE uses a modest number of experts, scaling the MoE architecture could further increase model capacity without a proportional increase in FLOPs. Future work could investigate dynamic expert assignment across spectral bands, automatic expert specialisation through evolutionary algorithms, or hierarchical MoE designs that operate at different spatial scales. An intriguing direction is to couple the MoE with perceptual or task-aware losses that better align super-resolved outputs with downstream tasks such as crop classification or land-cover mapping.

The WaveMAE framework could be extended to multi-modal data, for example by integrating synthetic aperture radar (SAR) or hyperspectral imagery. Combining wavelet decompositions across modalities may enable richer cross-channel correlations to be learned. Additionally, exploring joint reconstruction and contrastive pre-text tasks may yield more robust representations. The Geo-conditioned positional encoding could be augmented with auxiliary metadata, such as acquisition time or sun angles, to capture temporal and illumination dynamics.

As large-scale multi-modal datasets become available, there is an opportunity to train universal foundation models that unify super-resolution, denoising and cloud

removal within a single architecture. Such models might incorporate both masked autoencoding and diffusion-based generative modelling, offering controllable outputs and improved generalisation. Transfer to downstream tasks could be enhanced via adapter modules or task-specific heads, reducing the need for full fine-tuning.

Chapter 4

Publications

Journal Papers

- Rossi, L., **Bernuzzi, V.**, Fontanini, T., Bertozzi, M., and Prati, A. (2025). Swin2-MoSE: A new single image supersolution model for remote sensing. IET Image Processing, 19(1), e13303.
- **Bernuzzi, V.**, Rossi, L., Fontanini, T., Bertozzi, M., and Prati, A. (2025). Wavelet decomposition Masked Auto-Encoder for Remote Sensing. IEEE. (Under-submission)

International Conference Papers

- Rossi, L., **Bernuzzi, V.**, Fontanini, T., Bertozzi, M., and Prati, A. (2024, April). Memory-augmented online video anomaly detection. In ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 6590-6594). IEEE.

Appendix A

Memory-augmented Online Video Anomaly Detection

“Roads? Where we’re going, we don’t need roads.”

A.1 Introduction

Autonomous Vehicles (AVs) are increasingly transitioning from laboratory prototypes to operational systems as a result of sustained scientific and engineering progress across perception, planning, and control. Yet, safety remains a determining factor for large-scale deployment and public acceptance. One concrete pathway to improving safety is the rapid identification of anomalous situations during driving, enabling active interventions such as collision avoidance, pedestrian protection, and route re-planning [67]. Among the sensing modalities typically integrated in AV stacks (e.g., LiDAR, radar, GNSS/IMU, ultrasonics), camera systems offer dense, semantically rich, and cost-effective observations of the environment at high frame rates. Despite these advantages, detecting anomalous situations from video in unconstrained traffic remains difficult due to appearance variability, diverse motion patterns, occlusions,

weather and illumination changes, and the non-stationarity induced by ego-motion.

From an operational standpoint, anomaly detection for AVs imposes strict real-time requirements. Effective deployment entails not only computationally efficient models but also predictable latency and bounded memory usage to guarantee timely actuation. In online settings, the detector must operate causally—relying on the current frame and, at most, a limited history—without access to future observations. This constraint excludes batch-mode algorithms that aggregate information symmetrically in time and necessitates architectures explicitly designed for streaming inputs, deterministic execution time, and graceful handling of frame-rate jitter.

A foundational challenge concerns the absence of a universally agreed definition of “anomaly” in road scenes. Risk perception is highly context-dependent: for example, a brief lane encroachment might be benign in sparse traffic yet hazardous in congested conditions. Ambiguity also extends to temporal localization, where the precise onset and termination of an anomalous event are often subjective and may vary across annotators. The scarcity and long-tailed nature of incident categories further complicate supervised learning, since many classes are under-represented compared with normal driving behaviour. Several datasets introduce operational criteria to standardize labeling.

Within this context, a memory-augmented online detector namely MOVAD (Memory-augmented Online Video Anomaly Detection) is introduced to address anomaly localization across diverse road scenarios under strict online constraints. The design is deliberately minimalist with respect to input assumptions, consuming only RGB frames and avoiding dependencies on auxiliary modalities such as optical flow, object detections, or vehicle odometry. Two architectural elements are central. First, a Video Swin Transformer [68] is adapted to extract short-term memory in a streaming regime, allowing the backbone to harvest recent spatiotemporal correlations from the current frame and a compact buffer of preceding frames. Second, a lightweight LSTM module is incorporated in the classification head to encode long-term memory, enabling the decision layer to integrate contextual trends accumulated over the observed history. This division of labour—short-term dynamics in the backbone, long-term context in the head—yields complementary signals that improve robust-

ness to transient, non-anomalous perturbations and enhance sensitivity to genuinely hazardous deviations.

A comprehensive ablation on the DoTA benchmark [69] quantifies the contribution of each component and evaluates performance relative to prior art. Under identical online constraints and without auxiliary inputs, the system achieves superior Area Under the Curve (AUC), indicating effective discrimination between normal and anomalous segments. It is also notable that the inference pipeline is designed for constant execution time, a property desirable for real-time deployment on embedded platforms where compute budgets and latency guarantees are tightly controlled.

Contributions. The study offers the following contributions:

- An end-to-end, online architecture for Video Anomaly Detection (MOVAD) [70] that operates exclusively on RGB frames with minimal assumptions about scene content or actor availability, thereby simplifying integration in camera-only AV stacks.
- A Video Swin Transformer [68] backbone configured for short-term memory extraction in streaming conditions, capturing recent spatiotemporal correlations from the current and few preceding frames while maintaining efficiency.
- An LSTM-based classification head to encode long-term memory, leveraging context distributed over the entire observed sequence to stabilize decisions and reduce susceptibility to transient artifacts.
- An extensive ablation on DoTA [69] demonstrating competitive or superior AUC compared with methods that require auxiliary inputs, together with analysis of the effect of online constraints, temporal depth, and memory design on detection quality.

A.2 Related

Fang et al. [71] define the anomaly onset as the moment when at least half of the object involved becomes visible in the field of view, while Yao et al. [69] (DoTA)

annotate the start as the first instant after which the accident becomes unavoidable. The DoTA dataset [69] is commonly adopted as a benchmark for road-traffic anomaly detection due to its scale, carefully designed temporal annotations, and widespread use in the community.

Research on video anomaly detection (VAD) spans reconstruction-based, prediction-based, and hybrid approaches. Early work trained convolutional autoencoders to reconstruct normal frames and flag deviations at test time [72]. Subsequent methods integrated temporal recurrence via convolutional LSTM autoencoders to encode appearance and motion jointly [73, 74]. As highlighted in [75], reconstruction-style systems can be sensitive to contamination by anomalous content during training and often require ad-hoc post-processing (e.g., temporal smoothing, adaptive thresholds) to stabilize decisions. Prediction-driven techniques estimate short-term futures and compare predictions against observed frames; AnoPred [7] formulates a multi-task objective over intensity, optical flow, gradients, and an adversarial component. However, as discussed in [69], surveillance-oriented predictors may degrade on dash-cam footage, where rapid ego-motion, perspective changes, and complex interactions produce challenging dynamics. Two-stage pipelines such as STFE [8] first compute coarse temporal descriptors—e.g., Histogram of Optical Flow (HOF) [76] and CNN frame features—and subsequently refine decisions by modeling spatial relations among objects. Actor-centric formulations avoid full-frame prediction by tracking road users and forecasting their trajectories; FOL [69] follows this strategy to reduce prediction complexity. While effective in scenes with reliable detections, such methods depend on the presence and quality of object tracks within the field of view and frequently assume access to auxiliary inputs (optical flow, bounding boxes, or ego-motion), which can limit applicability in adverse conditions.

An alternative perspective emphasizes representation learning with explicit temporal reasoning. The TRN model [77] combines action detection with recurrent modeling of temporal dependencies and a decoding branch for anticipation, encouraging features that encode predictive structure. In online driving, a useful compromise is to couple short-term evidence—capturing abrupt changes and recent motion—with long-term context that reflects scene layout, intent cues, and evolving interactions,

without requiring explicit forecasting of future frames. Modern video transformers offer a suitable backbone for this objective: localized spatiotemporal attention and hierarchical feature extraction enable efficient processing of streaming video while remaining sensitive to motion and appearance cues.

A.3 Architecture

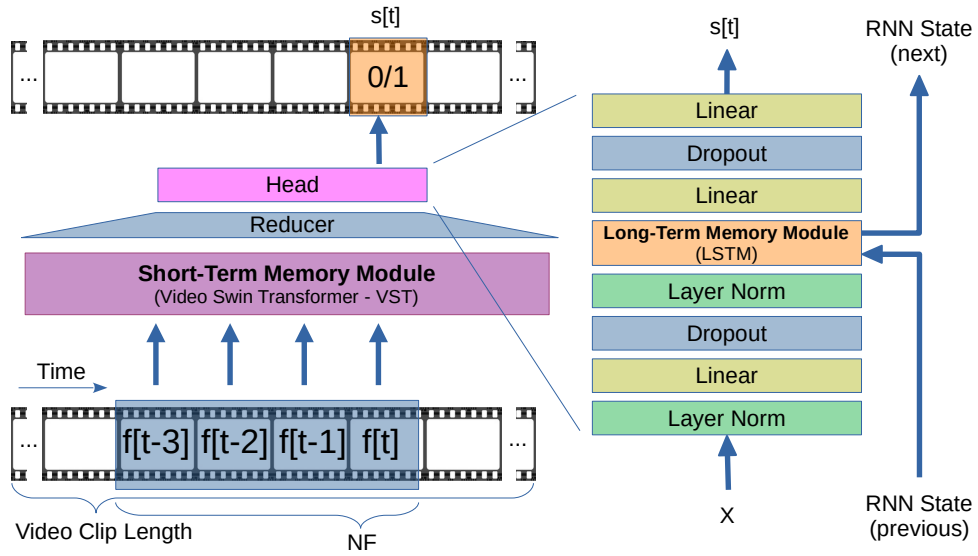


Figure A.1: The online frame-level VAD architecture. $f[t]$ is the frame at time t , x the output of the Reducer, NF the number of frames in input to the VST, $s[t]$ the anomaly classification score of frame $f[t]$.

The design treats recently observed frames as primary evidence of the ongoing activity while also leveraging more distant frames to supply contextual cues about scene layout and evolving interactions. To meet the constraints of Online VAD, the architecture decomposes temporal reasoning into two complementary components: a Short-Term Memory Module (STMM), responsible for extracting compact representations from the current frame and a short window of near-past inputs, and a

Long-Term Memory Module (LTMM) embedded in the classification head, tasked with accumulating and integrating longer-horizon context (see Fig. A.1).

Short-Term Memory Module. For pre-processing the incoming frame $f[t]$ while jointly retaining information from the immediate past $f[t-1] \dots f[t-(NF-1)]$, a transformer-based video backbone is selected in place of a recurrent encoder. This choice enables parallel computation over spatiotemporal tokens and exploits the ability of self-attention to capture long-range dependencies in both space and time [78]. Among transformer variants, the Video Swin Transformer (VST) [68] is adopted as the STMM backbone due to its favourable accuracy-throughput trade-off and efficient windowed attention relative to alternatives such as ViViT [79]. While VST was originally introduced for offline action classification—processing a full clip at once—the network is adapted here to operate in a causal, frame-level setting by consuming a sliding window of a few frames and producing a per-frame decision.

Because self-attention scales quadratically with token count, unrestricted temporal context is impractical for streaming scenarios. To bound latency and memory, the input is restricted to a compact window of $NF = 4$ frames, spanning the current time t to $t - (NF - 1)$. This configuration preserves salient short-term motion cues (e.g., abrupt decelerations, trajectory deviations) while keeping compute predictable. Formally, VST ingests a tensor of shape $NF \times H \times W \times 3$, where NF , H , W , and 3 denote the number of frames, spatial height, spatial width, and RGB channels, respectively. Frames are partitioned into non-overlapping 3D patches, yielding $\frac{NF}{2} \times \frac{H}{4} \times \frac{W}{4}$ spatiotemporal tokens. A 3D shifted-window attention mechanism is then applied to enable cross-window interactions and to fuse appearance with motion across the short horizon. The resulting short-term descriptor serves as the evidence stream subsequently consumed by the LTMM.

Long-Term Memory Module. Empirical analysis (Section A.4.1) indicates that simply enlarging the STMM temporal window ($NF > 4$) can degrade performance by giving equal salience to frames irrespective of their temporal recency, thereby diluting cues from the present with stale context. To address this limitation, longer-horizon information is injected through a distinct, lightweight memory pathway within the classification head. At each time step, the short-term representation emitted by the

STMM is forwarded to the LTMM, which updates a persistent state summarizing the distant past. This separation allows the decision process to remain anchored to current evidence while gradually accumulating background knowledge about the scene (e.g., road topology, typical motion flows), improving temporal stability and disambiguating transient artifacts.

Concretely, the VST output is first passed through an Adaptive Average Pool 3D layer (the *Reducer* in Fig. A.1) to condense spatiotemporal activations into a compact feature vector. The classification head comprises alternating normalization, linear, and dropout layers; a three-layer stacked LSTM is inserted after the final normalization stage to realize the LTMM. The recurrent state is composed of hidden $h[t]$ and cell $c[t]$ vectors for each of the three layers, updated once per incoming frame. The LSTM consumes features of shape $[B, 1024]$ (with B the batch size) and returns an equally sized transformed representation along with the updated states. Because the recurrent state is modest in size and updated incrementally, the added computational burden remains fixed and small, preserving constant-time inference. For each frame $f[t]$, the head predicts an anomaly score $s[t] \in [0, 1]$, where 0 denotes normal content and 1 indicates anomalous content. To mitigate class imbalance at training time, a weighted cross-entropy objective is employed, assigning larger weight to the anomaly class according to $w_i = e/e_i$, where e is the total number of training examples and e_i the number for class i .

A.4 Experiments

Dataset. Experimental evaluation is carried out on Task 1 of the DoTA benchmark [69] (frame-level VAD), using only the anomaly label and its temporal boundaries under a strictly online protocol. The corpus consists of 4,677 YouTube videos collected across multiple countries and conditions (illumination, weather), with a spatial resolution of 1280×720 . The original split is retained, yielding approximately 70% for training and 30% for validation.

Evaluation Metrics. Frame-level Area Under the ROC Curve (AUC) is used to quantify the ability to temporally localize anomalous frames within each video. This met-

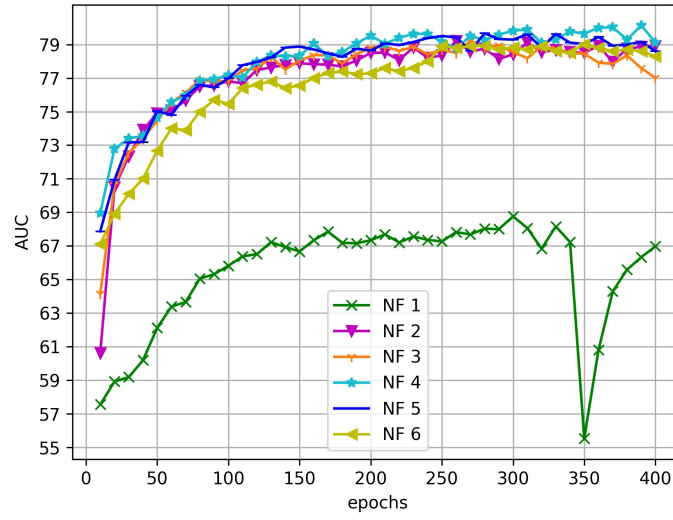


Figure A.2: Performance comparison, changing the number of frames (NF) in input to the STMM (from 1 to 6).

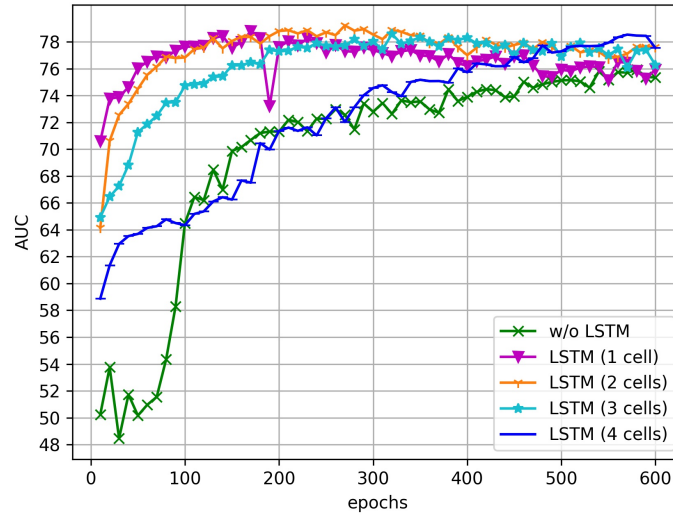


Figure A.3: Performance comparison, changing the number of LSTM cells. “LSTM (2 cells)” corresponds to “ $NF = 3$ ” in Fig. A.2.

ric reflects threshold-free discrimination quality and is standard in VAD comparisons. **Training details.** Reported baselines are taken from the corresponding publications. Training is performed on a single A100 GPU using Stochastic Gradient Descent (SGD) with learning rate 1×10^{-4} and momentum 0.9. SGD is preferred over Adam after observing that Adam led to instability and occasional divergence in preliminary runs on this setting. Unless otherwise specified: batch size is 8; input frames are re-sized to 320×240 ; the Video Clip Length (VCL)—the number of frames per video within each batch—is 8; the LSTM stack uses 2 cells; $NF = 3$ frames are supplied to the STMM; linear layers are initialized from a uniform distribution; recurrent weights use (semi-)orthogonal initialization; biases are zero-initialized; and VST is initialized from a model pretrained on Something-Something V2 [80]. Class imbalance in DoTA is handled with the weighted cross-entropy described at the end of Section A.3, using $w_n = 0.3$ for the normal class and $w_a = 0.7$ for the anomaly class. This configuration provides stable optimization, predictable per-frame latency, and a consistent memory footprint compatible with online deployment.

#	Short-term	Long-term	<i>AUC</i>
1			66.53
2	✓		74.46
3		✓	68.76
4	✓	✓	79.21

Table A.1: Performance comparison with and w/out short and long-term memory. Short-term: with ($NF = 3$) and w/out ($NF = 1$). Long-term: with (2 cells) and w/out (0 cells).

Since the videos contain a non-equal number of frames, we fixed the Video Clip Length (VCL), that is the number of frames inside the batch for each video, in order to be able to train the model with a batch size larger than one. To make the training as diverse as possible and reduce the effect of overfitting, at each iteration, we randomly choose the starting frame for each video, adjusting the ground-truth accordingly. Batch shape is $[B, V, H, W]$, where B is batch size, V is VCL, H and

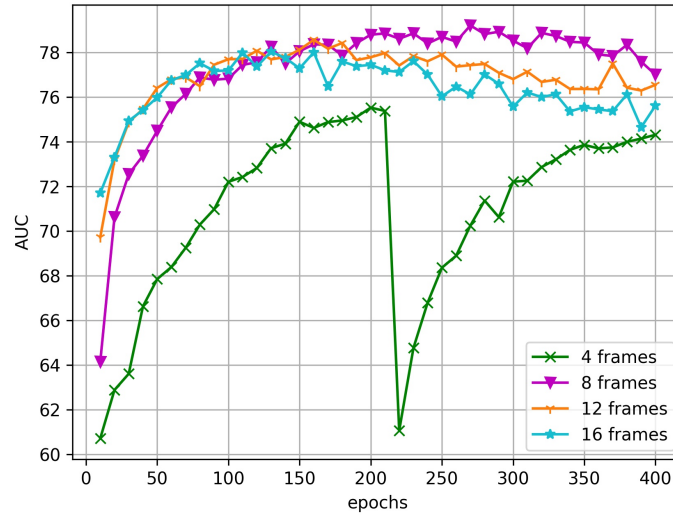


Figure A.4: Performance comparison, changing the VCL. The “8 frames” configuration corresponds to “ $NF = 3$ ” in Fig. A.2.

#	Method	Type	Input	<i>AUC</i>
1	ConvAE [72] (*)	Unsupervised	Gray	64.3
2	ConvAE [72] (*)		Flow	66.3
3	ConvLSTMAE [81] (*)		Gray	53.8
4	ConvLSTMAE [81] (*)		Flow	62.5
5	AnoPred [7] (*)		RGB	67.5
6	AnoPred [7] (*)		Masked RGB	64.8
7	FOL-Ensemble [69] (*)		RGB + Box + Flow + Ego	73.0
8	TRN [77] (**)	Supervised	RGB	78.0
9	STFE [8] (**)		RGB + Flow	79.3
10	Our (MOVAD)	Supervised	RGB (320×240)	80.09
11	Our (MOVAD)		RGB (640×480)	82.17

Table A.2: Benchmarks of VAD methods on the DoTA dataset. Both MOVAD models are trained with the best configuration of: VCL of 8, 3 LSTM cells and $NF = 4$. (Note: results with (*) and (**) are taken from DoTA [7] and [8] paper, respectively)

V are height and width, respectively. Unless otherwise specified, the following configuration is adopted: input video size is 320×240 , VCL is 8, LSTM cell number is 2, NF is 3, linear weights are initialized using a uniform distribution, the LSTM cells with a (semi-)orthogonal matrix, bias parameters are set to zero and the VST is initialized with a model pretrained on Something-Something v2 [80].

A.4.1 Ablation study

This section systematically explores alternative architectural and training settings, quantifies their effect on performance, and identifies the operating point used for the comparative evaluation in Section A.4.2.

Memory modules effectiveness. This experiment isolates the contribution of short- and long-horizon temporal reasoning; full configuration details are provided in the caption of Table A.1. As summarized in Table A.1, enabling either the Short-Term Memory Module (STMM) or the Long-Term Memory Module (LTMM) yields gains

Model	<i>ST</i>	<i>AH</i>	<i>LA</i>	<i>OC</i>	<i>TC</i>	<i>VP</i>	<i>VO</i>	<i>OO</i>	<i>UK</i>
AnoPred [7]	69.9	73.6	75.2	69.7	73.5	66.3	N/A	N/A	N/A
AnoPred [7] + Mask	66.3	72.2	64.2	65.4	65.6	66.6	N/A	N/A	N/A
FOL-STD [69]	67.3	77.4	71.1	68.6	69.2	65.1	N/A	N/A	N/A
FOL-Ensemble [69]	73.3	81.2	74.0	73.4	75.1	70.1	N/A	N/A	N/A
STFE [8]	75.2	84.5	72.1	77.3	72.8	71.9	N/A	N/A	N/A
Our (MOVAD)	85.6	85.1	83.9	82.2	85.3	86.2	79.3	86.7	77.1
Our (MOVAD) †	86.6	86.3	84.9	83.7	85.5	81.6	77.4	87.9	73.8
Model	<i>ST*</i>	<i>AH*</i>	<i>LA*</i>	<i>OC*</i>	<i>TC*</i>	<i>VP*</i>	<i>VO*</i>	<i>OO*</i>	<i>UK*</i>
AnoPred [7]	70.9	62.6	60.1	65.6	65.4	64.9	64.2	57.8	N/A
AnoPred [7] + Mask	72.9	63.7	60.6	66.9	65.7	64.0	58.8	59.9	N/A
FOL-STD [69]	75.1	66.2	66.8	74.1	72.0	69.7	63.8	69.2	N/A
FOL-Ensemble [69]	77.5	69.8	68.1	76.7	73.9	71.2	65.2	69.6	N/A
STFE [8]	80.6	65.6	69.9	76.5	74.2	N/A	75.6	70.5	N/A
Our (MOVAD)	72.1	71.6	72.3	76.5	75.7	74.1	77.9	71.7	69.1
Our (MOVAD) †	72.2	74.0	74.8	80.2	79.6	76.8	82.2	78.3	72.9

Table A.3: Detection accuracy (AUC) for individual accident categories. “*” non-ego anomaly categories. “†” if input resolution is 640×480 instead of 320×240 . N/A=Not Available. Bold and red values are the best and second-best results.

over a purely per-frame classifier, indicating that explicit temporal modeling is beneficial. Activating both modules concurrently produces the highest Area Under the Curve (AUC), highlighting their complementarity: exploiting near-past dynamics is advantageous, and augmenting this with longer-range context further improves discrimination. In general, incorporating information from the past outperforms single-frame inference.

Short-Term Memory Module. The STMM is instantiated with a Video Swin-B backbone. After linear patch projection, the embedding dimension is $C = 128$, and the stage depths are $d_i = \{2, 2, 18, 2\}$ for the i -th VST block. In practice, a higher-capacity variant (VST-L) exhibits a tendency to overfit under the available supervi-

sion, whereas lighter variants (VST-S/T) compress features too aggressively and fail to preserve sufficient scene detail. The influence of the short-term window is assessed by varying NF , the number of consecutive frames processed by the STMM at each step (see Fig. A.2). Using only the current frame consistently produces the weakest results and unstable optimization, as the model lacks temporal cues and tends to present nearly identical successive inputs to downstream layers, encouraging overfitting. Increasing NF generally improves AUC until a point at which the benefit reverses: for $NF \geq 5$ the transformer is burdened with temporally more diverse evidence without an internal mechanism to discount stale information, which degrades performance. Overall, a window of $NF = 4$ provides the most favorable trade-off between temporal richness and robustness.

Long-Term Memory Module. The contribution of long-range context is evaluated by varying the number of stacked recurrent cells from 0 (no LSTM) to 4; the resulting learning curves are reported in Fig. A.3. A GRU-based alternative exhibited qualitatively similar trends and is omitted for readability. Removing recurrent memory slows convergence and yields the lowest terminal AUC. Increasing the number of cells raises the attainable AUC but delays saturation: a single cell converges quickly and then drifts slightly downward, whereas four cells converge slowly and do not reach the best performance within the same training budget. The global maximum is observed with two cells, only +0.02 AUC above three cells; however, the three-cell configuration promotes steadier optimization and, when paired with the preferred short-term window ($NF = 4$), attains a higher final AUC. A plausible explanation is that both settings (three LSTM cells in Fig. A.3 and $NF = 4$ in Fig. A.2) reach their respective optima at approximately the same epoch (around 400), leading to more synchronized learning dynamics.

Video clip length (VCL). The temporal extent per training sample is varied in $\{4, 8, 12, 16\}$ to study its effect (Fig. A.4). Very short clips (VCL= 4) yield the most unstable optimization and the lowest AUC, likely because there is insufficient temporal evidence for the LTMM to exploit. Increasing VCL accelerates performance saturation; however, overly long clips (VCL= 12 or 16) reduce the final AUC, consistent with heightened redundancy and greater overfitting risk arising from high inter-frame sim-

ilarity. A VCL of 8 offers a balanced compromise, providing adequate context for the recurrent pathway while constraining redundancy to maintain generalization.

A.4.2 Comparison with the state of the art

Table A.2 presents a head-to-head comparison against representative state-of-the-art (SOTA) approaches, evaluating MOVAD at input resolutions of 320×240 and 640×480 . The configuration follows the ablation-derived setting: VCL of 8, three LSTM cells in the LTMM, and $NF = 4$ for the STMM. Both evaluated resolutions surpass the competing methods. The strongest configuration achieves an AUC of 82.17%, exceeding the previous best by +2.87 absolute AUC. Class-wise results, grouped into ego- and non-ego-centric anomaly categories (see [69] for category definitions), are reported in Table A.3. Unlike pipelines that rely on auxiliary detectors, the system operates directly on RGB inputs even when traffic participants are scarce or not visible, which is reflected in ego-involved categories such as VO (vehicle–obstacle collision) and OO (oncoming out-of-control).

A.5 Conclusions

This study presents MOVAD, an end-to-end architecture for online, frame-level video anomaly detection that enforces causal inference (past and present only) and maintains fixed, scene-agnostic execution time while requiring only RGB frames. Temporal reasoning is decoupled into a Short-Term Memory Module—implemented via a Video Swin Transformer to encode near-past dynamics—and a Long-Term Memory Module—implemented via stacked LSTMs to accumulate compact summaries of remote history within the classification head. On the DoTA benchmark, which comprises dash-mounted recordings of real-world traffic incidents, the system attains an AUC of 82.17%, outperforming prior state-of-the-art by +2.87 AUC. These results indicate that combining windowed transformer-based short-term encoding with lightweight recurrent long-term aggregation is an effective strategy for anomaly detection under strict online constraints.

Future research could explore unsupervised or self-supervised training regimes that reduce reliance on labelled anomalies. Extending the model to fuse LiDAR or radar could provide complementary motion cues and improve robustness to visual occlusion or adverse weather. Finally, applying MOVAD to broader surveillance contexts, such as industrial safety or wildlife monitoring, would further demonstrate its versatility.

Bibliography

- [1] coptrz.com. <https://coptrz.com/wp-content/uploads/2023/06/phantom-4-multispectral-5.jpg>. [Accessed 26-10-2025].
- [2] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.
- [3] 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.
- [4] Mubashir Noman, Muzammal Naseer, Hisham Cholakkal, Rao Muhammad Anwer, Salman Khan, and Fahad Shahbaz Khan. Rethinking transformers pre-training for multi-spectral satellite imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27811–27819, 2024.
- [5] Danfeng Hong, Bing Zhang, Xuyang Li, Yuxuan Li, Chenyu Li, Jing Yao, Naoto Yokoya, Hao Li, Pedram Ghamisi, Xiuping Jia, et al. Spectralgpt: Spectral remote sensing foundation model. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 46(08):5227–5244, 2024.

- [6] 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.
- [7] Wen Liu, Weixin Luo, Dongze Lian, and Shenghua Gao. Future frame prediction for anomaly detection—a new baseline. In *Proc. of Conf. on Computer Vision and Pattern Recognition*, pages 6536–6545, 2018.
- [8] Zhili Zhou, Xiaohua Dong, Zhetao Li, Keping Yu, Chun Ding, and Yimin Yang. Spatio-Temporal Feature Encoding for Traffic Accident Detection in VANET Environment. *IEEE Tran. on Intelligent Transportation Systems*, 23(10):19772–19781, October 2022.
- [9] Darius Phiri, Matamyo Simwanda, Serajis Salekin, Vincent R Nyirenda, Yuji Murayama, and Manjula Ranagalage. Sentinel-2 data for land cover/use mapping: A review. *Remote sensing*, 12(14):2291, 2020.
- [10] Josep Peñuelas and Iolanda Filella. Visible and near-infrared reflectance techniques for diagnosing plant physiological status. *Trends in plant science*, 3(4):151–156, 1998.
- [11] Leonardo Rossi, Vittorio Bernuzzi, Tomaso Fontanini, Massimo Bertozzi, and Andrea Prati. Swin2-mose: A new single image supersolution model for remote sensing. *IET Image Processing*, 19(1):e13303, 2025.
- [12] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- [13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In

- 9th Intl. Conf. on Learning Representations, ICLR 2021, May 3-7, 2021.* OpenReview.net, 2021. URL: <https://openreview.net/forum?id=YicbFdNTTy>.
- [14] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.
- [15] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proc. of the IEEE/CVF Intl. Conf. on Computer Vision (ICCV)*, 2021.
- [16] Honggang Chen, Xiaohai He, Linbo Qing, Yuanyuan Wu, Chao Ren, Ray E Sheriff, and Ce Zhu. Real-world single image super-resolution: A brief review. *Information Fusion*, 79:124–145, 2022.
- [17] Karansingh Chauhan, Shail Patel, Malaram Kumhar, Jitendra Bhatia, Sudeep Tanwar, Innocent E Davidson, Thokozile F Mazibuko, and Ravi Sharma. Deep learning-based single-image super-resolution: A comprehensive review. *IEEE Access*, 2023.
- [18] Hanadi Al-Mekhlafi and Shiguang Liu. Single image super-resolution: a comprehensive review and recent insight. *Frontiers of Computer Science*, 18(1):181702, 2024.
- [19] Hui Liu, Yurong Qian, Xiwu Zhong, Long Chen, and Guangqi Yang. Research on super-resolution reconstruction of remote sensing images: A comprehensive review. *Optical Engineering*, 60(10):100901–100901, 2021.
- [20] Ruben Fernandez-Beltran, Pedro Latorre-Carmona, and Filiberto Pla. Single-frame super-resolution in remote sensing: A practical overview. *International journal of remote sensing*, 38(1):314–354, 2017.

- [21] Xuan Wang, Jinglei Yi, Jian Guo, Yongchao Song, Jun Lyu, Jindong Xu, Weiqing Yan, Jindong Zhao, Qing Cai, and Haigen Min. A review of image super-resolution approaches based on deep learning and applications in remote sensing. *Remote Sensing*, 14(21):5423, 2022.
- [22] Peijuan Wang, Bulent Bayram, and Elif Sertel. A comprehensive review on deep learning based remote sensing image super-resolution methods. *Earth-Science Reviews*, page 104110, 2022.
- [23] J  r  my Anger, Thibaud Ehret, Carlo de Franchis, and Gabriele Facciolo. Fast and accurate multi-frame super-resolution of satellite images. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 1:57–64, 2020.
- [24] Jamy Lafenetre, Ngoc Long Nguyen, Gabriele Facciolo, and Thomas Eboli. Handheld burst super-resolution meets multi-exposure satellite imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2055–2063. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023.
- [25] Ngoc Long Nguyen, J  r  my Anger, Axel Davy, Pablo Arias, and Gabriele Facciolo. L1bsr: Exploiting detector overlap for self-supervised single-image super-resolution of sentinel-2 11b imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2012–2022. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023.
- [26] Gozde Unal Caglayan Tuna and Elif Sertel. Single-frame super resolution of remote-sensing images by convolutional neural networks. *International Journal of Remote Sensing*, 39(8):2463–2479, 2018. arXiv:<https://doi.org/10.1080/01431161.2018.1425561>, doi:10.1080/01431161.2018.1425561.

- [27] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(2):295–307, 2016. doi:10.1109/TPAMI.2015.2439281.
- [28] Kinga Karwowska and Damian Wierzbicki. Using super-resolution algorithms for small satellite imagery: A systematic review. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15:3292–3312, 2022. doi:10.1109/JSTARS.2022.3167646.
- [29] Sen Lei, Zhenwei Shi, and Zhengxia Zou. Super-resolution for remote sensing images via local–global combined network. *IEEE Geoscience and Remote Sensing Letters*, 14(8):1243–1247, 2017. doi:10.1109/LGRS.2017.2704122.
- [30] Wenjia Xu, Guangluan XU, Yang Wang, Xian Sun, Daoyu Lin, and Yirong WU. High quality remote sensing image super-resolution using deep memory connected network. In *Proceedings of the IEEE International Geoscience and Remote Sensing Symposium IGARSS*, pages 8889–8892. Proceedings of the IEEE International Geoscience and Remote Sensing Symposium IGARSS, 2018. doi:10.1109/IGARSS.2018.8518855.
- [31] Jingyun Liang, Jiezhong Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1833–1844. Proceedings of the IEEE/CVF international conference on computer vision, 2021.
- [32] Dafeng Zhang, Feiyu Huang, Shizhuo Liu, Xiaobing Wang, and Zhezhu Jin. Swinfr: Revisiting the swinir with fast fourier convolution and improved training for image super-resolution. *arXiv preprint arXiv:2208.11247*, 2022.
- [33] Marcos V Conde, Ui-Jin Choi, Maxime Burchi, and Radu Timofte. Swin2sr: Swin2 transformer for compressed image super-resolution and restoration. In

- European Conference on Computer Vision*, pages 669–687. European Conference on Computer Vision, 2022.
- [34] Zheng Chen, Yulun Zhang, Jinjin Gu, Linghe Kong, Xiaokang Yang, and Fisher Yu. Dual aggregation transformer for image super-resolution. In *ICCV. Proceedings of the Computer Vision – ICCV*, 2023.
- [35] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607, 2020. URL: <http://proceedings.mlr.press/v119/chen20j/chen20j.pdf>.
- [36] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.
- [37] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022.
- [38] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.
- [39] Qi Zhang, Yifei Wang, and Yisen Wang. How mask matters: Towards theoretical understandings of masked autoencoders. *Advances in Neural Information Processing Systems*, 35:27127–27139, 2022.
- [40] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*,

- pages 16000–16009, 2022. URL: https://openaccess.thecvf.com/content/CVPR2022/papers/He_Masked_Autoencoders_Are_Scalable_Vision_Learners_CVPR_2022_paper.pdf.
- [41] I. Daubechies. The wavelet transform, time-frequency localization and signal analysis. *IEEE Transactions on Information Theory*, 36(5):961–1005, 1990. doi:10.1109/18.57199.
- [42] Ting Yao, Yingwei Pan, Yehao Li, Chong-Wah Ngo, and Tao Mei. Wave-vit: Unifying wavelet and transformers for visual representation learning. In *European conference on computer vision*, pages 328–345. Springer, 2022.
- [43] Wenzhao Xiang, Chang Liu, Hongyang Yu, and Xilin Chen. Wavelet-driven masked image modeling: A path to efficient visual representation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 8611–8619, 2025.
- [44] Carlos Riquelme, Joan Puigcerver, Basil Mustafa, Maxim Neumann, Rodolphe Jenatton, André Susano Pinto, Daniel Keysers, and Neil Houlsby. Scaling vision with sparse mixture of experts. *Advances in Neural Information Processing Systems*, 34:8583–8595, 2021.
- [45] Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024.
- [46] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarczyk, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017.
- [47] William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *The Journal of Machine Learning Research*, 23(1):5232–5270, 2022.

- [48] Changho Hwang, Wei Cui, Yifan Xiong, Ziyue Yang, Ze Liu, Han Hu, Zilong Wang, Rafael Salas, Jithin Jose, Prabhat Ram, et al. Tutel: Adaptive mixture-of-experts at scale. *Proceedings of Machine Learning and Systems*, 5, 2023.
- [49] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12009–12019. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022.
- [50] Xiaoyi Dong, Jianmin Bao, Dongdong Chen, Weiming Zhang, Nenghai Yu, Lu Yuan, Dong Chen, and Baining Guo. Cswin transformer: A general vision transformer backbone with cross-shaped windows. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12124–12134. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022.
- [51] Mehdi SM Sajjadi, Bernhard Scholkopf, and Michael Hirsch. Enhancenet: Single image super-resolution through automated texture synthesis. In *Proceedings of the IEEE international conference on computer vision*, pages 4491–4500. Proceedings of the IEEE international conference on computer vision, 2017.
- [52] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690. Proceedings of the IEEE conference on computer vision and pattern recognition, 2017.
- [53] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

- [54] Julien Michel, Juan Vinasco-Salinas, Jordi Inglada, and Olivier Hagolle. Sen2ven μ s, a dataset for the training of sentinel-2 super-resolution algorithms. *Data*, 7(7):96, 2022.
- [55] Junwei Wang, Kun Gao, Zhenzhou Zhang, Chong Ni, Zibo Hu, Dayu Chen, and Qiong Wu. Multisensor remote sensing imagery super-resolution with conditional gan. *Journal of Remote Sensing*, 2021.
- [56] Dominik Koßmann, Viktor Brack, and Thorsten Wilhelm. Seasonet: A seasonal scene classification, segmentation and retrieval dataset for satellite imagery over germany. In *Proceedings of the IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium*, pages 243–246. Proceedings of the IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium, 2022.
- [57] Zhuo Zheng, Yanfei Zhong, Junjue Wang, and Ailong Ma. Foreground-aware relation network for geospatial object segmentation in high spatial resolution remote sensing imagery. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4096–4105. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020.
- [58] Marc Rußwurm, Konstantin Klemmer, Esther Rolf, Robin Zbinden, and Devis Tuia. Geographic location encoding with spherical harmonics and sinusoidal representation networks. *arXiv preprint arXiv:2310.06743*, 2023.
- [59] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093, 2022.
- [60] Valerio Marsocci, Yuru Jia, Georges Le Bellier, David Kerekes, Liang Zeng, Sebastian Hafner, Sebastian Gerard, Eric Brune, Ritu Yadav, Ali Shibli, Heng Fang, Yifang Ban, Maarten Vergauwen, Nicolas Audebert, and Andrea Nascetti. Pangaea: A global and inclusive benchmark for geospatial foundation models,

2024. URL: <https://arxiv.org/abs/2412.04204>, arXiv:2412.04204.
- [61] Katerina Kikaki, Ioannis Kakogeorgiou, Ibrahim Hoteit, and Konstantinos Karantzas. Detecting marine pollutants and sea surface features with deep learning in sentinel-2 imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2024.
- [62] Christopher Phillips, Sujit Roy, Kumar Ankur, and Rahul Ramachandran. HLS Foundation Burnscars Dataset, August 2023. URL: https://huggingface.co/ibm-nasa-geospatial/hls_burn_scars, doi:10.57967/hf/0956.
- [63] Adam Van Etten, Daniel Hogan, Jesus Martinez Manso, Jacob Shermeyer, Nicholas Weir, and Ryan Lewis. The multi-temporal urban development spacenet dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6398–6407, 2021.
- [64] Andrea Nascetti, RITU YADAV, Kirill Brodt, Qixun Qu, Hongwei Fan, Yuri Shendryk, Isha Shah, and Christine Chung. Biomasssters: A benchmark dataset for forest biomass estimation using multi-modal satellite time-series. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL: <https://openreview.net/forum?id=hrWsIC4Cmz>.
- [65] Kai Norman Clasen, Leonard Hackel, Tom Burgert, Gencer Sumbul, Begüm Demir, and Volker Markl. reben: Refined bigearthnet dataset for remote sensing image analysis, 2025. URL: <https://arxiv.org/abs/2407.03653>, arXiv:2407.03653.
- [66] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European conference on computer vision (ECCV)*, pages 418–434, 2018.

- [67] Tarak Gandhi and Mohan Manubhai Trivedi. Pedestrian protection systems: Issues, survey, and challenges. *IEEE Tran. on Intelligent Transportation Systems*, 8(3):413–430, 2007.
- [68] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video Swin Transformer. In *2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 3192–3201, New Orleans, LA, USA, June 2022. IEEE.
- [69] Yu Yao, Xizi Wang, Mingze Xu, Zelin Pu, Yuchen Wang, Ella Atkins, and David Crandall. Dota: Unsupervised detection of traffic anomaly in driving videos. *IEEE Tran. on Pattern Analysis and Machine Intelligence*, pages 1–1, 2022.
- [70] Leonardo Rossi, Vittorio Bernuzzi, Tomaso Fontanini, Massimo Bertozzi, and Andrea Prati. Memory-augmented online video anomaly detection. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6590–6594. IEEE, 2024.
- [71] Jianwu Fang, Dingxin Yan, Jiahuan Qiao, and Jianru Xue. Dada: A large-scale benchmark and model for driver attention prediction in accidental scenarios, 2019.
- [72] Mahmudul Hasan, Jonghyun Choi, Jan Neumann, Amit K Roy-Chowdhury, and Larry S Davis. Learning temporal regularity in video sequences. In *Procs. of the IEEE Conf. on computer vision and pattern recognition*, pages 733–742, 2016.
- [73] Weixin Luo, Wen Liu, and Shenghua Gao. Remembering history with convolutional lstm for anomaly detection. In *2017 IEEE Intl. Conf. on Multimedia and Expo*, pages 439–444, 2017.
- [74] Lin Wang, Fuqiang Zhou, Zuoxin Li, Wangxia Zuo, and Haishu Tan. Abnormal event detection in videos using hybrid spatio-temporal autoencoder. In *2018 25th IEEE Intl. Conf. on Image Processing (ICIP)*, pages 2276–2280. IEEE, 2018.

- [75] Bharathkumar Ramachandra, Michael J Jones, and Ranga Raju Vatsavai. A survey of single-scene video anomaly detection. *IEEE transactions on pattern analysis and machine intelligence*, 44(5):2293–2312, 2020.
- [76] Heng Wang and Cordelia Schmid. Action recognition with improved trajectories. In *Procs. of the IEEE international Conf. on computer vision*, pages 3551–3558, 2013.
- [77] Mingze Xu, Mingfei Gao, Yi-Ting Chen, Larry S Davis, and David J Crandall. Temporal recurrent networks for online action detection. In *Procs. of the IEEE/CVF Intl. Conf. on Computer Vision*, pages 5532–5541, 2019.
- [78] Oumaima Moutik, Hiba Sekkat, Smail Tigani, Abdellah Chehri, Rachid Saadane, Taha Ait Tchakoucht, and Anand Paul. Convolutional neural networks or vision transformers: Who will win the race for action recognitions in visual data? *Sensors*, 23(2):734, 2023.
- [79] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Procs. of the IEEE/CVF Intl. Conf. on Computer Vision (ICCV)*, pages 6836–6846, October 2021.
- [80] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The "something something" video database for learning and evaluating visual common sense. In *Procs. of the IEEE international Conf. on computer vision*, pages 5842–5850, 2017.
- [81] Yong Shean Chong and Yong Haur Tay. Abnormal event detection in videos using spatiotemporal autoencoder. In *Advances in Neural Networks-ISNN 2017: 14th Intl. Symposium, ISNN 2017, Sapporo, Hakodate, and Muroran, Hokkaido, Japan, June 21–26, 2017, Procs., Part II 14*, pages 189–196. Springer, 2017.

Acknowledgement

Project ECS_00000033_ECOSISTER funded under the National Recovery and Resilience Plan (NRRP), Mission 4 Component 2 Investment 1.5 - funded by the European Union – NextGenerationEU.

This research benefits from the High Performance Computing facility of the University of Parma, Italy (HPC.unipr.it).