ChromaFormer: A Scalable Multi-Spectral Transformer for Large-Scale Land Cover Classification

Anonymous authors
Paper under double-blind review

Abstract

Remote sensing satellites such as Sentinel-2 provide high-resolution, multi-spectral imagery that enables dense, large-scale land cover classification. However, most deep learning models used in this domain—frequently CNN-based architectures—are limited in their ability to process high-dimensional spectral data and scale with increasing dataset sizes. Moreover, while transformer architectures have recently been introduced for remote sensing tasks, their performance on large, densely labeled multi-spectral datasets remains underexplored.

In this paper, we present ChromaFormer, a scalable family of multi-spectral transformer models designed for large-scale land cover classification. We introduce a novel Spectral Dependency Module (SDM) that explicitly learns inter-band relationships through attention across spectral channels, enabling efficient spectral-spatial feature fusion. Our models are evaluated on the Biological Valuation Map (BVM) of Flanders, a large, densely labeled dataset spanning over 13,500 km^2 and 14 classes. ChromaFormer models achieve substantial accuracy gains over baselines: while a 23M-parameter UNet++ achieves less than 70% accuracy, a 655M-parameter ChromaFormer attains over 96% accuracy. We also analyze performance scaling trends and demonstrate generalization to standard benchmarks. Our results highlight the effectiveness of combining scalable transformer architectures with explicit spectral modeling for next-generation remote sensing tasks.

1 Introduction and Background

Remote sensing imagery provides crucial information for applications in environmental monitoring, urban planning, and disaster management, yet the high dimensionality and volume of satellite data pose significant challenges for traditional analysis techniques. Deep learning models, particularly CNNs, have greatly advanced remote sensing analysis (Maggiori et al., 2016; Hu et al., 2015; Zhu et al., 2017; Li et al., 2022; Roy et al., 2020), but challenges remain in scalability and spectral feature learning. Vision transformers are recently proven to be more effective on capturing long-range dependencies in remote sensing applications (Dosovitskiy et al., 2021). Among many transformer variations, the Swin Transformer further improved efficiency via localized self-attention windows, and subsequent studies applied transformers to multi-modal remote sensing imagery with great success (Aleissaee et al., 2023; Roy et al., 2023; Bergamasco et al., 2023), indicating careful architectural choices may be crucial for extracting meaningful features from multi/hyper-spectral data. Additionally, recent hybrid models (Yuan et al., 2023; Zhang et al., 2023) and large-scale pretraining techniques (Wang et al., 2024; 2022b) further showcase the direction of scaling vision models in remote sensing.

1.1 Scaling challenges in remote sensing

Despite the progress in model design, the scaling properties of neural networks for remote sensing remain under-explored. Intuitively, larger models and larger training datasets can yield better performance, as observed in computer vision and NLP domains (Brown et al., 2020). However, most remote sensing models to date are relatively small (often <200 million parameters) and are evaluated on limited datasets. For example, it is common to apply a 25M-parameter ResNet50 model to both a tiny hyperspectral scene (Firat

& Hanbay, 2021) and a much larger satellite image collection (Papoutsis et al., 2021). Such a one-size-fits-all approach ignores potential efficiency and accuracy gains when matching model capacity to dataset scale. With the rise of large, labeled datasets (e.g. province/nation-scale), it becomes crucial to investigate how increasing model size and complexity impacts performance. Models such as scaled Swin-transformer variants (Peng et al., 2023; Yuan et al., 2023) have demonstrated the capability of scalable remote sensing architectures. However, it must be tested on a realistic dense dataset. For this study we use the Biological Valuation Map (BVM) of Flanders, a land-cover dataset with complete annotations for an area of over 13,500 km^2 (De Saeger et al., 2017; De Saeger et al., 2020; Li et al., 2024b). A comparison of dataset sizes and corresponding model complexities is shown in Figure 1 and Table 3 (Appendix).

1.2 Leveraging spectral information

Another limitation of many existing models is the under-utilization of rich spectral information. Most CNN-based frameworks were originally designed for RGB imagery and struggle to utilize the spectral information provided by earth observation satellites such as Sentinel (VITO, 2020) and Landsat (Earth Resources Observation and Science (EROS) Center, 2020). Prior works have proposed spectral attention modules to enhance CNNs' performance (Hang et al., 2021; Roy et al., 2021; 2020). A transformer-based spectral–spatial network was even explored via neural architecture search (Zhong et al., 2022). Recent work such as ShapeFormer (Lv et al., 2023), AerialFormer (Hanyu et al., 2024), and LSKNet (Li et al., 2025) also demonstrated improvements in remote sensing classification by combining multiscale spatial and spectral modeling. Furthermore, self-supervised models like SSL4EO (Wang et al., 2023) show promise in pretraining with rich spectral bands. However, these methods typically model spectral correlations in a limited or sparse manner (e.g. separate 3D convolution blocks for spectral features) rather than a unified token-level attention across all bands. In other words, existing approaches do not fully learn inter-band relationships in an end-to-end fashion, leaving an opportunity to design architectures that more directly attend to spectral dependencies.

In this work, we address these gaps by proposing ChromaFormer, a multi-spectral transformer architecture tailored for multi-scale land cover segmentation. It is built on a Swin Transformer backbone and augmented with a novel Spectral Dependency Module that learns joint representations from all spectral bands. The selection of the Swin Transformer as a foundational architecture is a direct response to the dual challenges of global context modeling and computational scalability inherent to high-resolution remote sensing. While CNNs excel at extracting local features, they are constrained by a limited receptive field, failing to capture the long-range spatial dependencies crucial for complex land-use/land-classification tasks. Standard Vision Transformer (ViT) architectures, while capable of modeling global context , introduce a different prohibitive bottleneck: a computational complexity that scales quadratically with the input image size. This quadratic cost makes naive ViT models impractical for dense, pixel-level prediction on high-resolution satellite imagery. By integrating spectral attention into the transformer's self-attention layers, our model can weight and fuse information across different spectra – unlike prior CNN-based spectral modules that operate on fixed or handcrafted spectral groupings. We systematically study how the performance of ChromaFormer and other architectures varies with model size and training data size, leveraging a large scale Sentinel-2 dataset (Li et al., 2024b).

1.3 Dataset choice

To meaningfully evaluate the scaling properties of the chosen architectures, a correspondingly large-scale and high-quality dataset is required. The Biological Valuation Map (BVM)(Li et al., 2024b) of Flanders was selected as it is positioned at the intersection of three critical criteria: scale, density, and quality. We argue that the BVM dataset is suited and necessary for meaningfully evaluating the scaling properties of large-capacity remote sensing models.

Large Scale: Large-capacity models, such as Swin-h and ChromaFormer-h, are data-hungry. They require a massive dataset to "unlock their potential and avoid overfitting". The BVM provides this necessary scale, covering over 13,500 km^2. This is over 25 times larger than other common densely labeled benchmarks like LoveDA and is one to two orders of magnitude larger (in number of labeled pixels) than classic benchmarks. This scale permits a true big-data, big-model analysis.

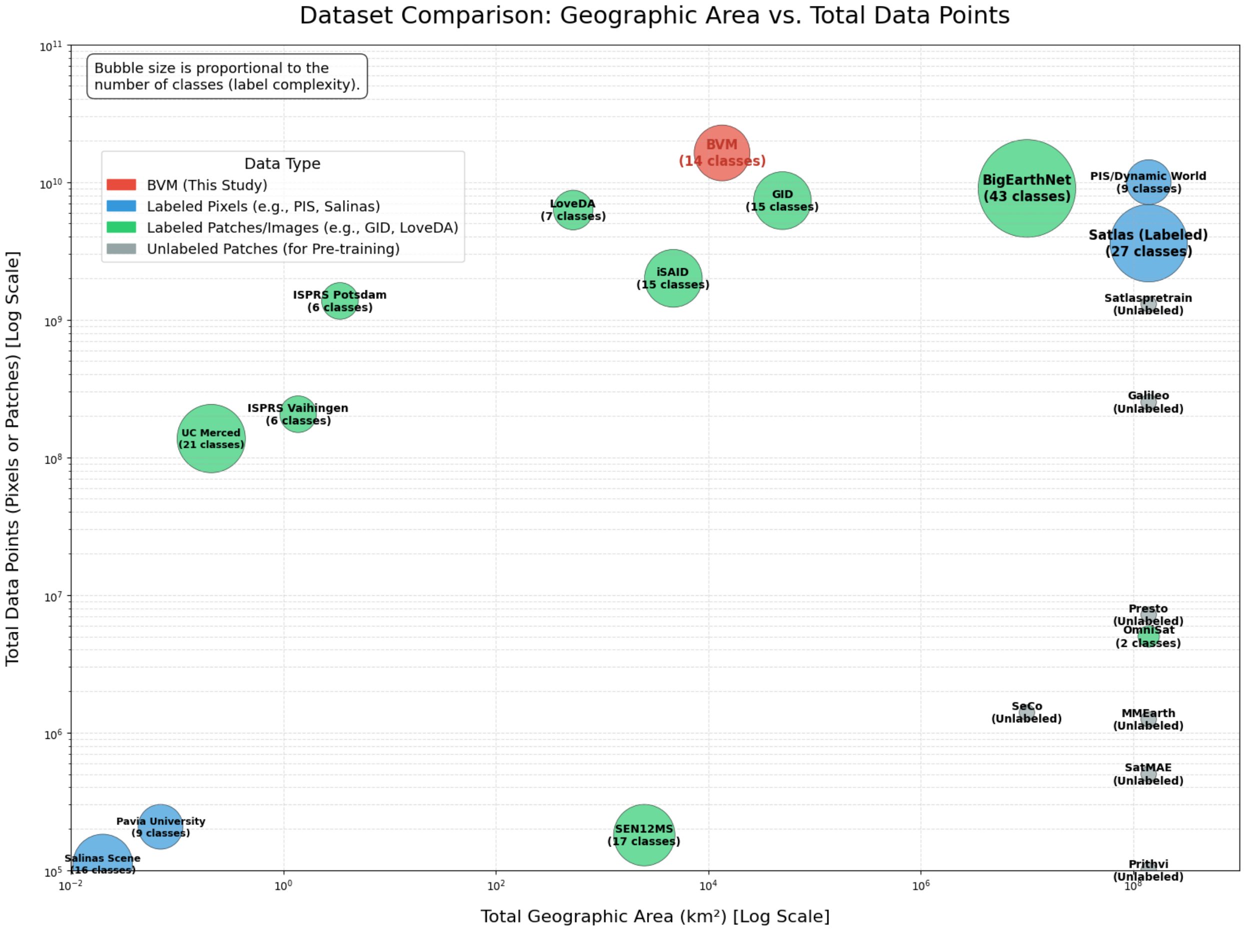


Figure 1: Dataset Comparison: Geographic Area vs. Total Data Points. Bubble size is proportional to the number of classes (label complexity). Datasets are sourced from BVM (Li et al., 2024b) or their respective publications. Labeled pixel datasets include PIS/Dynamic World, Satlas, Pavia University, and Salinas Scene (Brown et al., 2022; Bastani et al., 2023; Gamba, 2003a; Center for Hyperspectral Remote Sensing Scenes (EHU/UPV), 1998a). Labeled patch/image datasets include BigEarthNet, GID, LoveDA, iSAID, ISPRS, UC Merced, OmniSat, and SEN12MS (Sumbul et al., 2019a; Tong et al., 2020a; Wang et al., 2021a; Waqas Zamir et al., 2019; ISPRS Commission II / WG4, 2012a;b; Yang & Newsam, 2010a; Astruc et al., 2024; Schmitt et al., 2019). Unlabeled pre-training datasets are also shown (Bastani et al., 2023; Ulbricht et al., 2025; Tseng et al., 2023; Manas et al., 2021; Nedungadi et al., 2024; Cong & Zhou, 2022; Jakubik et al., 2023).

High Density: The BVM provides dense, pixel-wise ground truth for the entire region , making it a valid and information rich testbed. This full-coverage pixel-level density is a key feature. In contrast, other large remote sensing datasets, such as BigEarthNet or LoveDA, are noted to have labels that are often sparse or less reliable.

High-Fidelity Quality: The BVM's label quality is a critical differentiator. It is not automatically generated, which can suffer from misinformation. Instead, it is a field-driven survey curated by vegetation experts (De Saeger et al., 2017; De Saeger et al., 2020). This expert-derived, high-fidelity ground truth ensures that the high accuracy scores achieved in the study are a robust measure of model capability, not an artifact of label noise.

In summary, this unique combination of massive scale, pixel level density, and expert verified quality makes the BVM an ideal and necessary benchmark for evaluating large-scale remote sensing models. In this work, it is paired with Sentinel-2 imagery to evaluate land cover classification models. Sentinel-2, operated by the European Space Agency (ESA), is a multispectral satellite system providing 13 spectral bands across visible, near-infrared, and shortwave infrared regions at resolutions of 10–60m. Similar to Landsat, Sentinel-2 provides higher temporal resolution and richer spectral information. Li et al. (2024b) partitioned the BVM dataset into training and evaluation sets with nearly identical class distributions, ensuring that model performance is assessed on representative data. The Chi-squared distance between data splits was reported. Chi-squared distance is often used to determine the similarity of two categorical distributions, and it is formulated as follows: $D_{\chi^2}(P, Q) = \sum_i \frac{(P(i)-Q(i))^2}{P(i)+Q(i)}$, where P(i) and Q(i) are the probabilities of the i-th element in distributions P and Q, respectively, as per Table 1.

To complement our evaluation, we also consider Pavia University (Gamba, 2003b) and Indian Pines (Baumgardner et al., 2015) benchmarks. These two datasets are widely used in hyperspectral classification, especially in evaluating channel-wise modeling performance. Prior benchmarks on these datasets include CNN-based architectures (Li et al., 2022), attention-guided fusion (Hang et al., 2021), token-mixing spectral transformers (Jeong et al., 2025), and WaveMix (Jeevan & Sethi, 2024), which applies Fourier/wavelet decomposition to image patches.

Table 1: Class distribution across BVM data splits. (Li et al., 2024a)

Class	**Train (%)**	**Val. (%)**	**Test (%)**
Coastal dune habitats	0.10	0.06	0.25
Cultivated land	34.27	33.92	32.85
Grasslands	23.07	22.92	22.14
Heathland	0.54	1.08	0.95
Inland marshes	0.24	0.23	0.22
Marine habitats	0.26	0.04	0.34
Pioneer vegetation	0.69	0.65	0.56
Small landscape features - not specified	0.05	0.06	0.07
Small non-woody landscape features	0.14	0.13	0.12
Small woody landscape features	0.63	0.58	0.73
Unknown	0.01	0.007	0.006
Urban areas	26.27	27.05	28.56
Water bodies	2.08	2.05	1.74
Woodland and shrub	11.65	11.22	11.47

2 Methodologies

This section outlines the design and implementation of our proposed ChromaFormer framework on the BVM dataset. We first introduce the key component of ChromaFormer: the Spectral Dependency Module (SDM). We then emphasize ChromaFormer's architectural components, embedding and training strategies, which together enable effective modeling of both spectral dependencies and spatial structures inherent to multi/hyperspectral data. We also highlight how our design improves on scalability concerns (Hafner et al., 2024), that model performance may plateau or degrade if training regimes are not tuned to match the model depth and data volume.

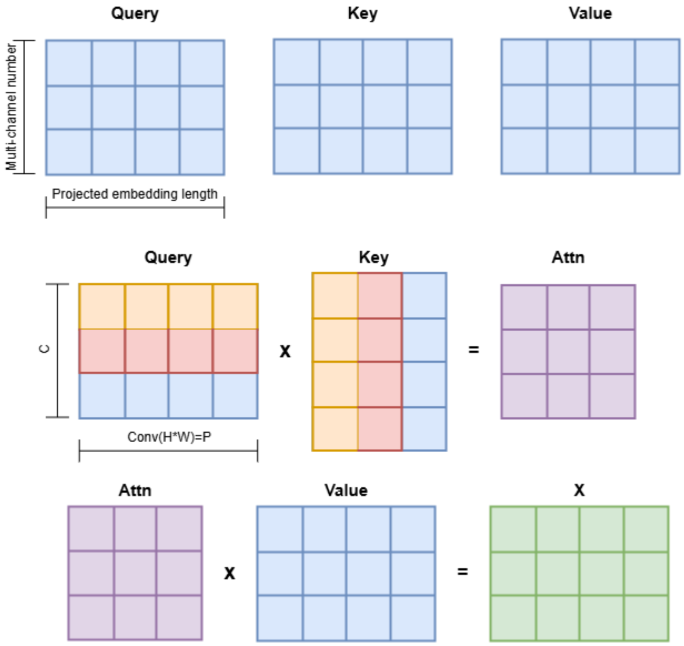


Figure 2: Spectral Dependency Module (SDM): Similar to self-attention in transformers, but attention is computed between spectral bands (channels) across the entire spatial domain.

2.1 Spectral Dependency Module (SDM)

To better utilize spectral information in multi-spectral imagery, we propose the Spectral Dependency Module (SDM), designed to explicitly model inter-band relationships.

Given an input tensor $X \in \mathbb{R}^{B \times C \times H \times W}$, we reshape the spatial dimensions to produce $Q, K, V \in \mathbb{R}^{B \times C \times N}$, where $N = H \times W$, and compute a spectral attention map of size $C \times C$. Then the channel-wise attention map is computed using scaled dot-product attention across spectral bands:

$$A = \text{softmax}\left(\frac{QK^\top}{\sqrt{N}}\right) \in \mathbb{R}^{B \times C \times C}.$$

The output is obtained by multiplying the attention weights with the value tensor:

$$O = A \cdot V \in \mathbb{R}^{B \times C \times N},$$

which is then reshaped back to $\mathbb{R}^{B \times C \times H \times W}$ to match the original feature map dimensions.

This mechanism enables the model to learn inter-band spectral dependencies across the full spatial extent of the image, thereby enhancing spectral-spatial feature representations. Unlike spatial attention, SDM emphasizes global interactions between spectral bands, and can be easily inserted into existing convolutional or transformer architectures due to its simple interface. It is worth mentioning that SDM is a lightweight, end-to-end learnable attention module that models channel-to-channel interactions, unlike fixed or non-learnable spectral priors. SDM is also resolution-independent and highly parallelizable, making it compatible with vision transformer pipelines.

2.2 Transformer backbone

Following SDM, we employ a hierarchical Swin Transformer backbone composed of four stages. Each stage consists of multiple Swin Transformer blocks with increasing receptive field and hidden dimensions. Our ChromaFormer models are defined by their specific configurations of embedding channels, the number of layers per stage (depths), and the attention head configurations across these stages. These parameters dictate the model's capacity and its ability to aggregate information across local and global spatial contexts, complementing the SDM's spectral modeling.

The configurations for the different ChromaFormer variants are as follows:

- **ChromaFormer-t (Tiny):** Uses 96 embedding channels, depths of $(2, 2, 6, 2)$ layers per stage, and attention head configurations of $(3, 6, 12, 24)$.
- **ChromaFormer-s (Small):** Similar to ChromaFormer-t in embedding channels (96) and attention heads $(3, 6, 12, 24)$, but with increased depths of $(2, 2, 18, 2)$ layers per stage.
- **ChromaFormer-b (Base):** Uses 128 embedding channels, depths of $(2, 2, 18, 2)$, and attention head configurations of $(4, 8, 16, 32)$.
- **ChromaFormer-l (Large):** Uses 192 embedding channels, depths of $(2, 2, 18, 2)$, and attention head configurations of $(6, 12, 24, 48)$.
- **ChromaFormer-h (Huge):** The largest variant, uses 352 embedding channels, depths of $(2, 2, 18, 2)$, and attention head configurations of $(11, 22, 44, 88)$.

This systematic scaling of parameters allows us to investigate the performance and efficiency of ChromaFormer across a wide range of model capacities. Code and models will be made available at the time of publication.

2.3 Training strategy and evaluation metrics

Models are trained with Adam optimizer on 4×NVIDIA A100-80G GPUs using DDP and HuggingFace Accelerate, we adopted distributed data parallelism with shared gradient synchronization across the GPUs. Batch size is 16, learning rate is 1e-4. Inputs and labels are normalized and padded. We use standard cross-entropy loss and follow recent best practices for fine-tuning.

We train the model with a learning rate of 10^{-4} and a batch size of 16. The training loop is implemented using the Accelerate library to support multi-GPU setups. Standard cross-entropy loss is used as the training objective. We normalize input tensors and apply padding where necessary to fit the fixed-resolution transformer architecture. In addition to classic griding and patching methods (Chen et al., 2014; Li et al., 2016), we follow recent recommendations for pretraining and fine-tuning from SSL4EO (Wang et al., 2023) and LSKNet (Li et al., 2025), using normalized input and label maps.

We report overall accuracy (OA) and per-class accuracy as our main evaluation metrics. In line with standard benchmarks, we evaluate the model on both accuracy and robustness against underrepresented classes (Wang et al., 2024). We also evaluate the scalability trends of our models following observations from the foundation model literature (Hafner et al., 2024; Brown et al., 2020), highlighting where models reach saturation and where additional parameter growth no longer yields benefits.

3 Results and Discussion

3.1 Per-class performance

The empirical analysis of per-class accuracy curves in Figure 3 reveals that ChromaFormer models consistently outperform their Swin Transformer counterparts in land cover classification, demonstrating higher peak accuracies and faster convergence rates across most land-cover classes. This performance edge is expressed even more in classes that inherently rely on spectral information for accurate discrimination. For instance, in "Coastal dune habitats," where floristic diversity correlates with spectral heterogeneity, and "Cultivated land," where dynamic spectral signatures indicate plant health and growth stages over the year, ChromaFormer models show substantial gains. In "Inland marshes" and "Marine habitats," where mixed

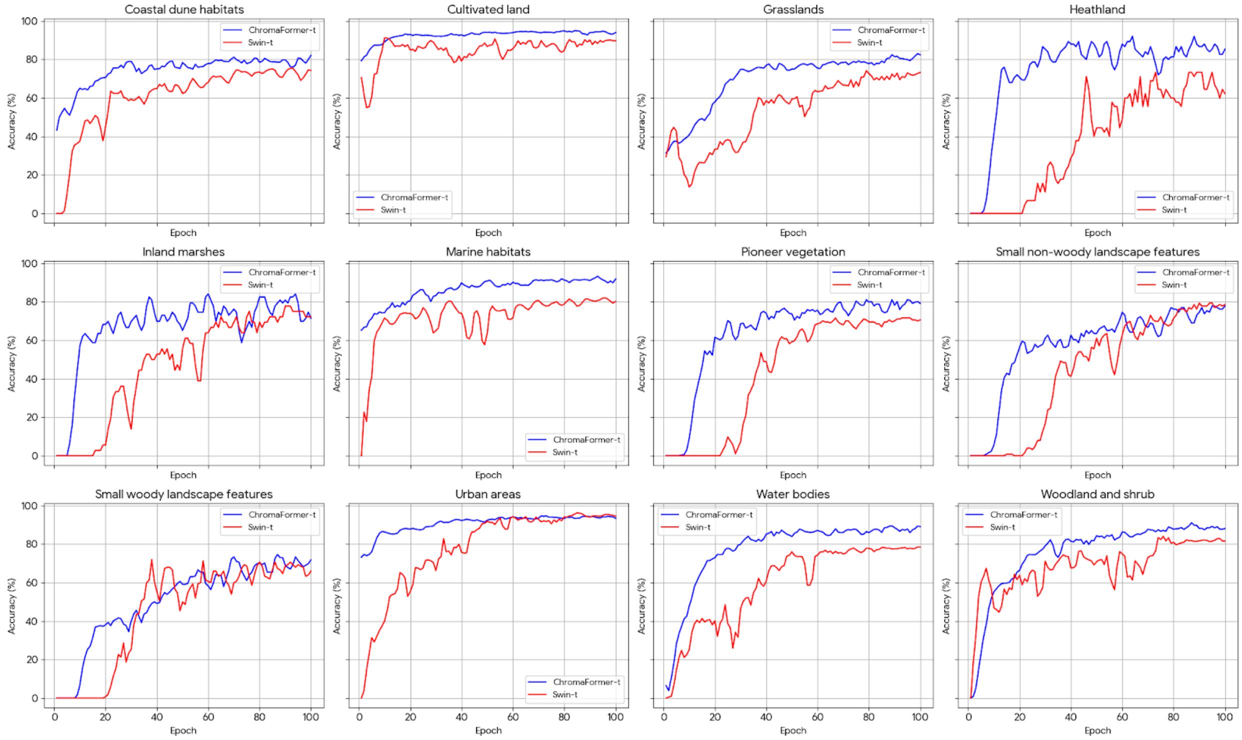


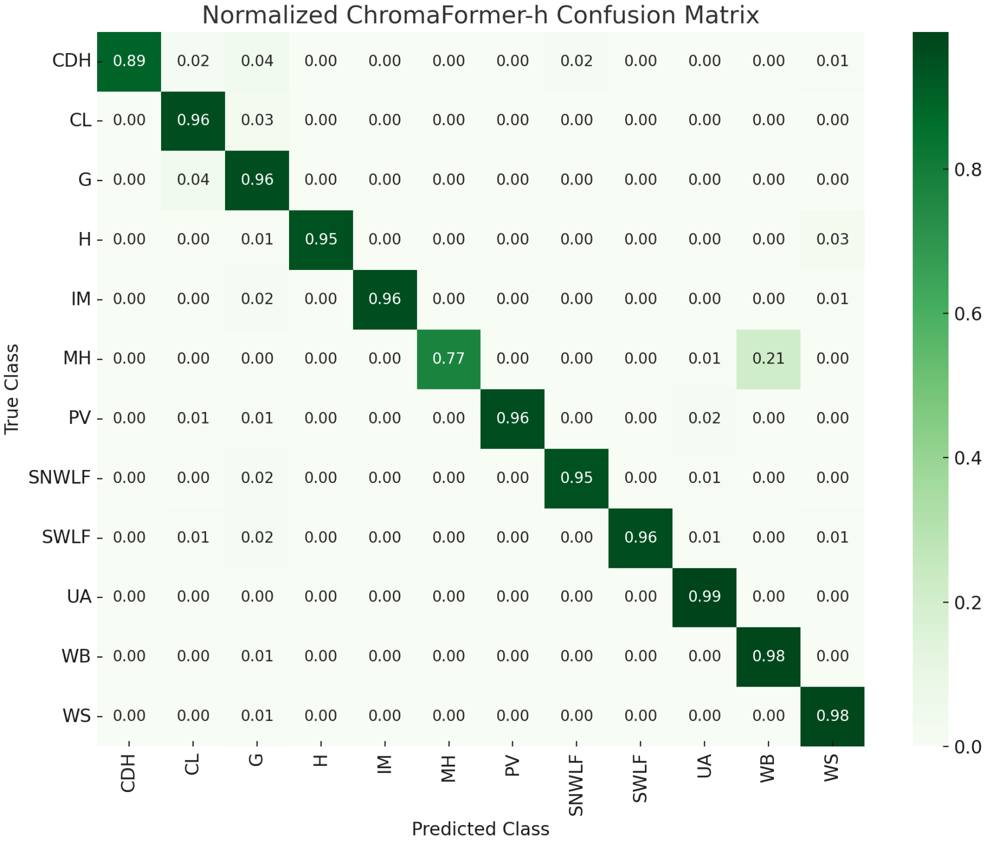
Figure 3: Per-class accuracy curves of Chromaformer-t (blue) and Swin-t (red), "Unknown" and "Small landscape features-not specified" removed as they provide less meaningful information.

water–vegetation signals and water column effects challenge classification, ChromaFormer's superior spectral processing capabilities are also evident. The confusion matrix in Figure 4 further confirms these trends, showing strong diagonal dominance for spectrally complex classes, particularly in minority categories such as CDH, MH, and IM, where misclassification rates remain low despite severe class imbalance.

3.2 Quantitative model comparison

As shown in Table 2, the ChromaFormer models exhibit superior performance compared to both the ResNet family and the UNet++ model across all sizes, based on the provided data. In the small model category, ChromaFormer-t (27 million parameters) achieves an accuracy of 90.87%, significantly outperforming ResNet-20M (20 million parameters) with 80.95% accuracy and UNet++ (23 million parameters) with 68.81% accuracy. This substantial accuracy gap highlights the efficiency of ChromaFormer models in handling complex tasks, indicating that ChromaFormer could be more parameter-efficient than its competitors. Additionally, ChromaFormer models maintain higher scaling coefficients than ResNet models, indicating more efficient scaling as model size increases.

Between the Swin Transformer and ChromaFormer models, the ChromaFormer models offer additional advantages, making them a better choice for multi-spectral segmentation tasks. The integration of the Spectral Dependency Module (SDM) into the Swin Transformer architecture allows ChromaFormer models to capture spectral dependencies more effectively, leading to higher accuracies without a significant increase in parameters or loss of scaling efficiency. For example, ChromaFormer consistently achieves higher accuracy than Swin Transformers in almost all model size ranges, with almost identical scaling coefficients. This demonstrates that ChromaFormer models enhance spectral feature learning while maintaining efficient scalability. Therefore, the ChromaFormer models are optimal for multi-spectral tasks running at different scales, as they provide superior accuracy and maintain scaling efficiency compared to both their Swin Transformer coun-



	CDH	CL	G	H	IM	MH	PV	SNWLF	SWLF	UA	WB	WS
CDH	0.89	0.02	0.04	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.01
CL	0.00	0.96	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
G	0.00	0.04	0.96	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
H	0.00	0.00	0.01	0.95	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03
IM	0.00	0.00	0.02	0.00	0.96	0.00	0.00	0.00	0.00	0.00	0.00	0.01
MH	0.00	0.00	0.00	0.00	0.00	0.77	0.00	0.00	0.00	0.01	0.21	0.00
PV	0.00	0.01	0.01	0.00	0.00	0.00	0.96	0.00	0.00	0.02	0.00	0.00
SNWLF	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.95	0.00	0.01	0.00	0.00
SWLF	0.00	0.01	0.02	0.00	0.00	0.00	0.00	0.00	0.96	0.01	0.00	0.01
UA	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.99	0.00	0.00
WB	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.98	0.00
WS	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.98

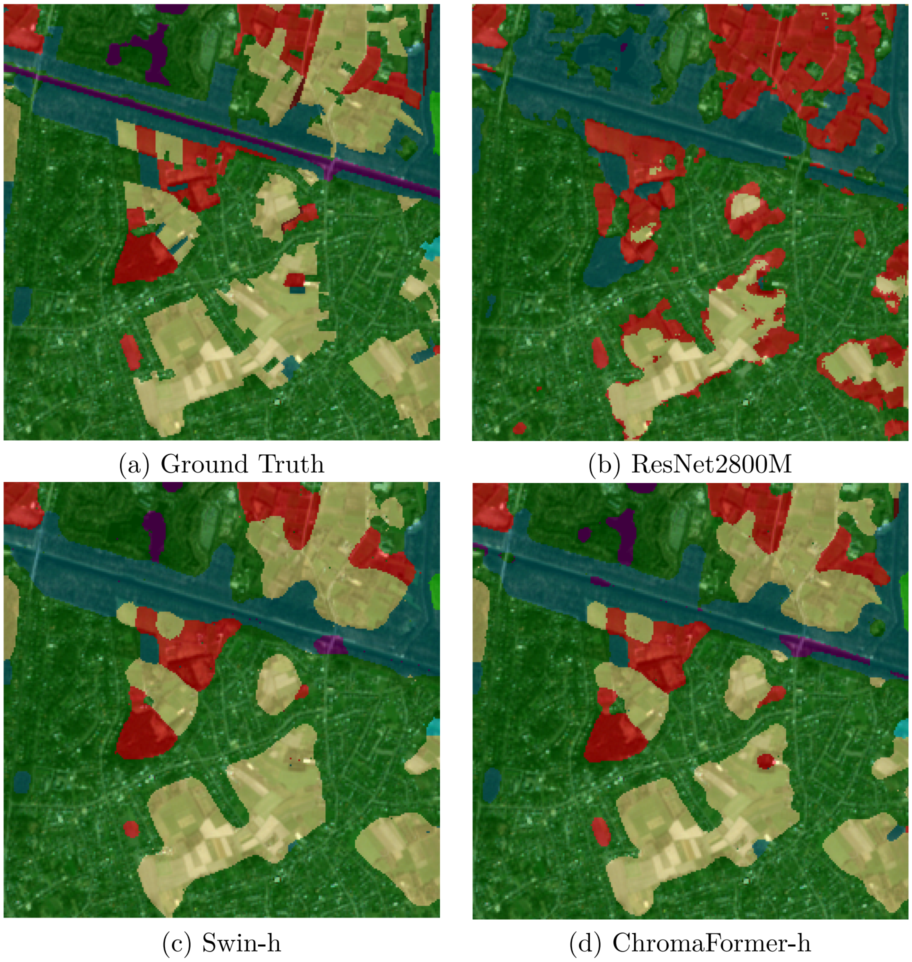
Figure 4: Normalized ChromaFormer-h confusion matrix on the test dataset. CDH = Coastal dune habitats, CL = Cultivated land, G = Grasslands, H = Heathland, IM = Inland marshes, MH = Marine habitats, PV = Pioneer vegetation, SNWLF = Small non-woody landscape features, SWLF = Small woody landscape features, UA = Urban areas, WB = Water bodies, WS = Woodland and shrub.

terparts and conventional models like ResNet and UNet++. Figure 5 shows output results from different models; as we can see, ChromaFormer is better than ResNet and Swin Transformers in predicting minor classes.

3.3 Accuracy and loss analysis

Figure 6 compares the accuracy curves of ResNet-1M, ResNet-20M, ChromaFormer-t, Swin-t, and U-Net++ over 100 training epochs. ChromaFormer-t achieves the highest accuracy, steadily improving and reaching 90.87% by the end of training. Swin-t follows closely, reaching 88.98%. ResNet-20M converges and stabilizes earlier, indicating moderate capacity. ResNet-1M performs worse, plateauing at around 75.92%. U-Net++ shows the lowest performance, with accuracy saturating below 70% and minimal improvement after the initial epochs.

Figure 10 (Appendix) shows the loss curve comparison between ChromaFormer and ResNet. The analysis of loss curves for various neural network models reveals distinct behaviors based on model size and architecture. Smaller models like ResNet-1M and ResNet-20M show rapid early loss descent but quickly reach saturation, indicating a capacity limitation in handling complex patterns as training progresses. In contrast, larger ResNet models and all ChromaFormer models demonstrate a more gradual loss reduction, suggesting that they can continue improving with extended training due to their greater capacity to learn complex data representations.


(a) Ground Truth

(b) ResNet2800M

(c) Swin-h

(d) ChromaFormer-h

Figure 5: Qualitative results: (a) Ground truth, (b) ResNet2800M, (c) Swin-h, and (d) ChromaFormer-h.

ChromaFormer models exhibit a slower but more persistent decline in loss, suggesting that they benefit from extended training epochs before showing signs of saturation, unlike conventional models like ResNet. This prolonged effectiveness in learning indicates that SDM modules are capable of exploiting their architectural efficiency to handle complex data relationships over longer periods. While larger ResNet models tend to plateau earlier, larger ChromaFormer models (like ChromaFormer-l and ChromaFormer-h) continue to show potential for improvement beyond the typical saturation points of conventional architectures, highlighting the distinct advantage of transformer-based models in sustained learning capability. We also observe that scaling beyond a certain point stops being rewarding for ResNet (e.g., the loss gap between ResNet-1550M and ResNet-2800M becomes negligible around 1×10^8 sample passes), whereas the losses for ChromaFormer-l and ChromaFormer-h only start to converge around 2×10^8 sample passes. This underscores the fact that transformer models, despite their size, are inherently designed to scale more effectively with increasing data and training duration before experiencing diminishing returns.

3.4 Scaling efficiency comparison

Figure 7 illustrates the relationship between model accuracy and scaling coefficient across a range of architectures, including ResNets, Swin Transformers, U-Net++, and ChromaFormers. While most models exhibit a general trend where higher scaling coefficients correlate with increased accuracy, ChromaFormer models consistently achieve superior accuracy even at moderate or low scaling coefficients. Notably, ChromaFormer-t, ChromaFormer-s, and ChromaFormer-b match or exceed the performance of their Swin counterparts at

Table 2: Comparison of models with parameters, training time, accuracy (with 95% error bars computed using $N = 21,500,000$ samples), and scaling coefficients. Scaling efficiency coefficient S quantifies how effectively a neural network scales its performance relative to the increase in parameters and computational resources. It is mathematically defined as: $S = -\log\left(\frac{G}{P \times C}\right)^{-1}$ where: G is the Performance Gain Factor, P is the Parameter Count Scaling Factor, C is the Computation Increase Factor.

Model	**Parameters (M)**	**Time/Epoch (h)**	**Accuracy (%)**	**Scaling Coefficient**
Small Models (∼1M to 30M Parameters)				
ResNet-1M	1	0.8	75.92 ± 0.02	Baseline
ResNet-2M	2	1.0	76.03 ± 0.02	1.09
UNet++	23	1.0	68.81 ± 0.02	N/A
ResNet-20M	20	1.0	80.95 ± 0.02	0.32
Swin-t	27	2.2	88.98 ± 0.01	Baseline
ChromaFormer-t	27	2.2	90.87 ± 0.01	Baseline
Medium Models (∼50M to 100M Parameters)				
ResNet-230M	230	1.8	84.10 ± 0.02	0.16
Swin-s	49	3.1	92.19 ± 0.01	1.10
ChromaFormer-s	49	3.1	93.47 ± 0.01	1.09
Swin-b	86	3.7	93.08 ± 0.01	0.61
ChromaFormer-b	86	3.7	94.02 ± 0.01	0.61
Large Models (∼150M to 300M Parameters)				
ResNet-1550M	1550	6.3	87.32 ± 0.02	0.11
Swin-l	195	4.7	94.57 ± 0.01	0.37
ChromaFormer-l	195	4.7	95.98 ± 0.01	0.37
Extra-Large Models (∼650M to 2800M Parameters)				
ResNet-2800M	2800	10.0	89.19 ± 0.01	0.10
Swin-h	655	6.0	96.64 ± 0.01	0.24
ChromaFormer-h	656	6.0	96.67 ± 0.01	0.24

equivalent parameter scales. This suggests that ChromaFormer architectures offer improved scaling efficiency, achieving better accuracy without a proportional increase in computational cost.

3.5 Performance on benchmark datasets

To further validate the generalizability and effectiveness of our proposed ChromaFormer architecture, we conduct experiments on two widely used hyperspectral image classification benchmarks: Pavia University and Indian Pines. These datasets are standard testbeds for remote sensing models and have been extensively used to benchmark prior state-of-the-art methods.

Our lightweight variant, ChromaFormer-t, achieves a satisfying classification accuracy of 91.03% on the Pavia University dataset and 99.70% on the Indian Pines dataset, surpassing the 92.37% achieved by HyperspectralMAE (Jeong et al., 2025). Qualitative results are shown in Figure 8 and Figure 9. These results exceed previously reported accuracies on public leaderboards, including CNN-based and hybrid models. We emphasize that our model achieves these results without task-specific tuning or architectural overfitting to the dataset characteristics. The same model configuration was used across both datasets, underscoring its robustness in learning spatial–spectral representations. These results demonstrate that ChromaFormer is not only effective on a large-scale, complex dataset like BVM, but also achieves state-of-the-art performance on standard benchmarks, making it a compelling general-purpose solution for hyperspectral image classification tasks.

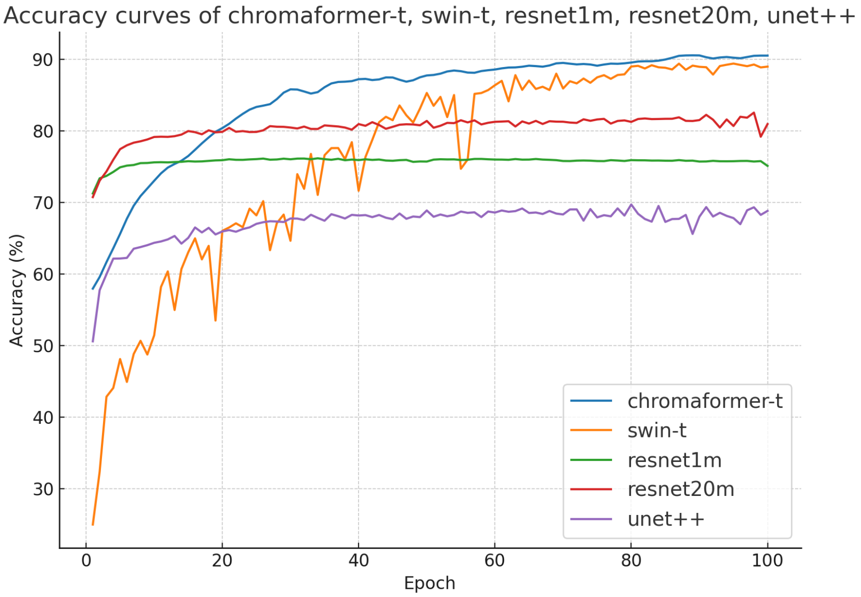


Figure 6: Accuracy curves of different models, "m" in legend stands for million parameters. We only plot small models for better visual clarity.

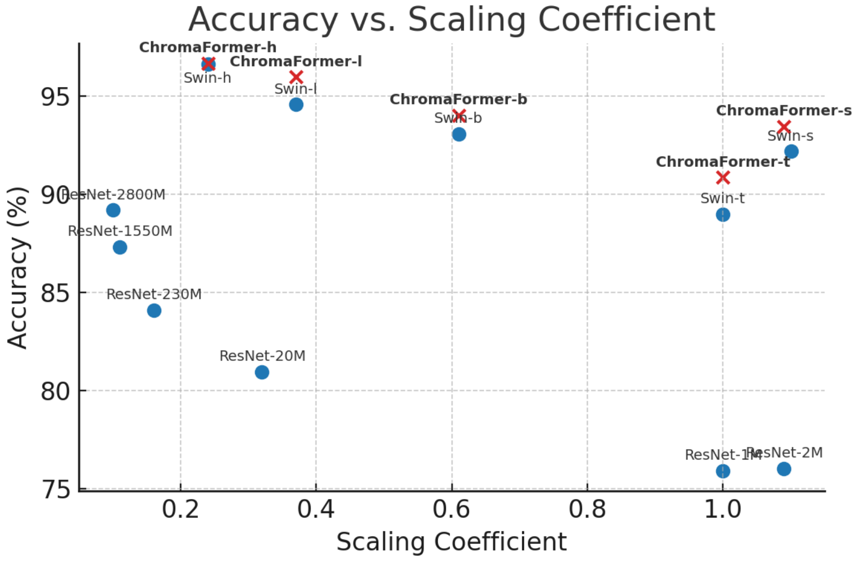


Figure 7: Accuracy versus scaling coefficient for various model architectures, including ResNet, Swin Transformer, U-Net++, and ChromaFormers. ChromaFormer models are highlighted with red crosses and bold labels. Notably, ChromaFormers achieve superior performance with equal or lower scaling cost compared to their counterparts.

4 Conclusions and limitations

In this work, through extensive experiments on the large-scale Biological Valuation Map (BVM) of Flanders, Indian Pines and Pavia University dataset, we demonstrated that ChromaFormer outperforms conventional CNN-based models and pure vision transformers in terms of accuracy, scaling efficiency, and robustness to class imbalance. The novel SDM module improves spectral fusion in a lightweight, learnable manner. Our key

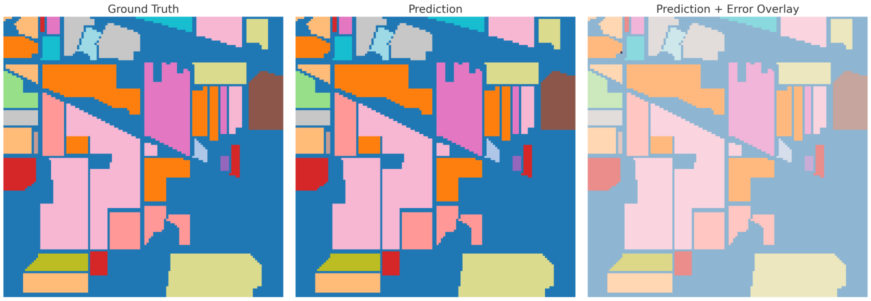


Figure 8: Visual comparison between ground truth and ChromaFormer-t prediction results on the Indian Pines dataset.(Left) Ground truth map with 16 semantic classes; (Middle) Predicted map using 5-fold cross-validated ChromaFormer-t ensemble with majority voting with 99.70% accuracy; (Right) Prediction map overlaid with error regions: correctly predicted regions are shown with opacity.

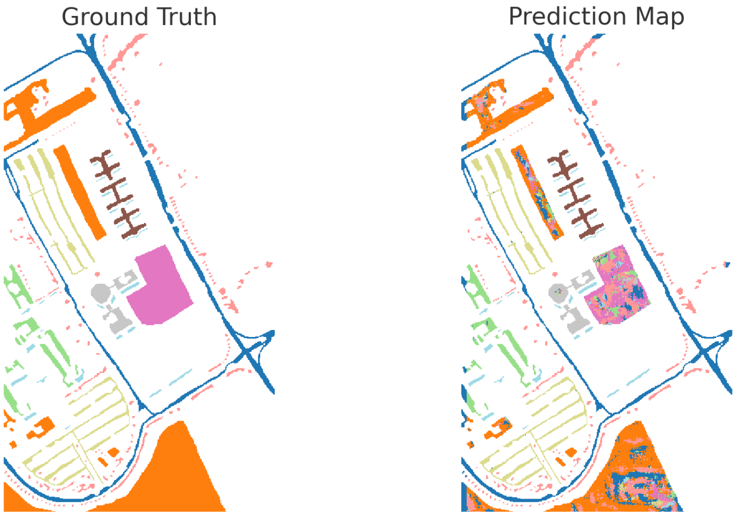


Figure 9: Visual comparison between ground truth and ChromaFormer-t prediction results on the Pavia University dataset. (Left) Ground truth map; (Right) Prediction map generated using 5-fold cross-validated ChromaFormer-t ensemble with majority voting. Visually, most class boundaries are well preserved.

finding is that aligning model complexity with dataset scale—both in terms of spatial coverage and spectral dimensionality—is crucial for effective land cover classification. The SDM module enables the model to leverage inter-band spectral correlations in a learnable, scalable manner, resulting in superior generalization and performance across different classes and resolutions. SDM's capability of utilizing temporal data is further detailed in Table 4 (Appendix). Additionally, ChromaFormer models maintain high scaling efficiency even at hundreds of millions of parameters, suggesting their suitability for processing increasingly large remote sensing datasets being released worldwide.

Nevertheless, our study has some limitations. First, BVM dataset is geographically constrained to the Flemish region of Belgium, evaluation on more diversed geographic zones is necessary to validate global generalization. Second, computational constraints limited our ability to run extensive ablation tests with ChromaFormer-b/l/h variations and explore even larger model variants or longer training regimes, which could uncover further scaling benefits. Finally, while we benchmarked against several popular baselines, additional comparisons with emerging foundation models or hybrid transformer–CNNs would help further position ChromaFormer within the broader model landscape. Future work may also investigate integrating SDM into other model families or extending the architecture to handle temporal sequences in multi-temporal satellite imagery.

References

Abdulaziz Amer Aleissaee, Amandeep Kumar, Rao Muhammad Anwer, Salman Khan, Hisham Cholakkal, Gui-Song Xia, and Fahad Shahbaz Khan. Transformers in remote sensing: A survey. *Remote Sensing*, 15(7), 2023. ISSN 2072-4292. doi: 10.3390/rs15071860. URL https://www.mdpi.com/2072-4292/15/7/1860.

Guillaume Astruc, Erlend Mørk, Conrad Ulbricht, Manuel Foks, Zhaoyang He, Selma Zaïri, Céleste Robin, Yonatan Lencz, Elad David, and Natan Malkin. Omnisat: Self-supervised modality fusion for earth observation. In *European Conference on Computer Vision (ECCV)*, 2024. arXiv:2404.08351.

Favyen Bastani, Piper Wolters, Ritwik Gupta, Joseph Ferdinando, and Ani Kembhavi. Satlaspretrain: A large-scale dataset for remote sensing image understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 21003–21014, 2023.

Marion F. Baumgardner, Larry L. Biehl, and David A. Landgrebe. 220 band aviris hyperspectral image data set: June 12, 1992 indian pine test site 3, Sep 2015. URL https://purr.purdue.edu/publications/1947/1.

Luca Bergamasco, Francesca Bovolo, and Lorenzo Bruzzone. A dual-branch deep learning architecture for multisensor and multitemporal remote sensing semantic segmentation. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 16:2147–2162, 2023.

Qi Bi, Kun Qin, Han Zhang, and Gui-Song Xia. Local semantic enhanced convnet for aerial scene recognition. *IEEE Transactions on Image Processing*, 30:6498–6511, 2021a.

Qi Bi, Han Zhang, and Kun Qin. Multi-scale stacking attention pooling for remote sensing scene classification. *Neurocomputing*, 436:147–161, 2021b.

Qi Bi, Beichen Zhou, Kun Qin, Qinghao Ye, and Gui-Song Xia. All grains, one scheme (agos): Learning multigrain instance representation for aerial scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–17, 2022.

Christopher F Brown, Steven P Brumby, Brookie Guzder-Williams, Tanya Birch, Samantha Brooks Hyde, Joseph Mazzariello, Wanda Czerwinski, Valerie J Pasquarella, Robert Haertel, Simon Ilyushchenko, et al. Dynamic world, near real-time global 10 m land use land cover mapping. *Scientific data*, 9(1):251, 2022.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Center for Hyperspectral Remote Sensing Scenes (EHU/UPV). Salinas scene dataset. https://www.ehu.eus/ccwintco/index.php/Hyperspectral_Remote_Sensing_Scenes, 1998a.

Center for Hyperspectral Remote Sensing Scenes (EHU/UPV). Salinas scene dataset. https://www.ehu.eus/ccwintco/index.php/Hyperspectral_Remote_Sensing_Scenes, 1998b.

Keumgang Cha, Junghoon Seo, and Taekyung Lee. A billion-scale foundation model for remote sensing images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, pp. 1–17, 2024. doi: 10.1109/JSTARS.2024.3401772.

Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 801–818, 2018.

Yushi Chen, Zhouhan Lin, Xing Zhao, Gang Wang, and Yanfeng Gu. Deep learning-based classification of hyperspectral data. *IEEE Journal of Selected topics in applied earth observations and remote sensing*, 7 (6):2094–2107, 2014.

Yijun Cong and Tat-Jun Zhou. SatMAE: Pre-training Transformers for Temporal and Multi-Spectral Satellite Imagery. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 2552–2567, 2022.

S. De Saeger, P. Oosterlynck, and D. Paelinckx. The biological valuation map (bvm): a fielddriven survey of land cover and vegetation in the flemish region of belgium. *Documents phytosociologiques - Actes du colloque de Saint-Mandé 2012 - Prodrome et cartographie des végétations de France*, 6:372–382, 2017.

Steven De Saeger, Robin Guelinckx, Patrik Oosterlynck, Adinda De Bruyn, Klaas Debusschere, Pieter Dhaluin, Rémar Erens, Pieter Hendrickx, Dirk Hennebel, Indra Jacobs, Myriam Kumpen, Jorgen Opdebeeck, Toon Spanhove, Ward Tamsyn, Frank Van Oost, Guy Van Dam, Martine Van Hove, Carine Wils, and Desiré Paelinckx. *Biologische Waarderingskaart en Natura 2000 Habitatkaart, uitgave 2020.* Number 35 in Rapporten van het Instituut voor Natuur- en Bosonderzoek. Instituut voor Natuur- en Bosonderzoek, België, 2020. doi: 10.21436/inbor.18840851.

Ivica Dimitrovski, Vlatko Spasev, Suzana Loshkovska, and Ivan Kitanovski. U-net ensemble for enhanced semantic segmentation in remote sensing imagery. *Remote Sensing*, 16(12):2077, 2024.

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 *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=YicbFdNTTy`.

Earth Resources Observation and Science (EROS) Center. Landsat 8-9 operational land imager / thermal infrared sensor level-2, collection 2 [dataset], 2020. URL `https://doi.org/10.5066/P9OGBGM6`. Public domain dataset.

Hüseyin Firat and Davut Hanbay. Classification of hyperspectral images using 3d cnn based resnet50. In *2021 29th Signal Processing and Communications Applications Conference (SIU)*, pp. 1–4. IEEE, 2021.

Paolo Gamba. Hyperspectral remote sensing scenes: grupo de inteligencia computacional (gic). `https://www.ehu.eus/ccwintco/index.php/Hyperspectral_Remote_Sensing_Scenes#Pavia_Centre_and_University`, 2003a.

Paolo Gamba. Hyperspectral remote sensing scenes - grupo de inteligencia computacional (gic), 2003b. URL `https://www.ehu.eus/ccwintco/index.php/Hyperspectral_Remote_Sensing_Scenes#Pavia_Centre_and_University`.

Meng-Hao Guo, Cheng-Ze Lu, Qibin Hou, Zhengning Liu, Ming-Ming Cheng, and Shi-Min Hu. Segnext: Rethinking convolutional attention design for semantic segmentation. *Advances in Neural Information Processing Systems*, 35:1140–1156, 2022.

Sebastian Hafner, Heng Fang, Hossein Azizpour, and Yifang Ban. Continuous urban change detection from satellite image time series with temporal feature refinement and multi-task integration. *arXiv preprint arXiv:2406.17458*, 2024.

Renlong Hang, Zhu Li, Qingshan Liu, Pedram Ghamisi, and Shuvra S. Bhattacharyya. Hyperspectral image classification with attention-aided cnns. *IEEE Transactions on Geoscience and Remote Sensing*, 59(3): 2281–2293, 2021. doi: 10.1109/TGRS.2020.3007921.

Taisei Hanyu, Kashu Yamazaki, Minh Tran, Roy A. McCann, Haitao Liao, Chase Rainwater, Meredith Adkins, Jackson Cothren, and Ngan Le. Aerialformer: Multi-resolution transformer for aerial image segmentation. *Remote Sensing*, 16(16), 2024. ISSN 2072-4292. doi: 10.3390/rs16162930. URL `https://www.mdpi.com/2072-4292/16/16/2930`.

Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7132–7141, 2018.

Wei Hu, Yangyu Huang, Li Wei, Fan Zhang, and Hengchao Li. Deep convolutional neural networks for hyperspectral image classification. *Journal of Sensors*, 2015(1):258619, 2015. doi: https://doi.org/10.1155/2015/258619. URL https://onlinelibrary.wiley.com/doi/abs/10.1155/2015/258619.

Gyutae Hwang, Jiwoo Jeong, and Sang Jun Lee. Sfa-net: Semantic feature adjustment network for remote sensing image segmentation. *Remote Sensing*, 16(17):3278, 2024.

ISPRS Commission II / WG4. Isprs potsdam 2d semantic labeling benchmark dataset. https://www.isprs.org/resources/datasets/benchmarks/UrbanSemLab/2d-sem-label-potsdam.aspx, 2012a.

ISPRS Commission II / WG4. Isprs vaihingen 2d semantic labeling benchmark dataset. https://www.isprs.org/resources/datasets/benchmarks/UrbanSemLab/2d-sem-label-vaihingen.aspx, 2012b.

Johannes Jakubik, L. Chu, A. Fraccaro, L. Godwin, S. K. Mukkavilli, S. Roy, C. Phillips, et al. Foundation models for generalist geospatial artificial intelligence. *arXiv preprint arXiv:2310.18660*, 2023. URL https://arxiv.org/abs/2310.18660.

Pranav Jeevan and Amit Sethi. Which backbone to use: A resource-efficient domain specific comparison for computer vision. *arXiv preprint arXiv:2406.05612*, 2024.

Wooyoung Jeong, Hyun Jae Park, Seonghun Jeong, Jong Wook Jang, Tae Hoon Lim, and Dae Seoung Kim. Hyperspectralmae: The hyperspectral imagery classification model using fourier-encoded dual-branch masked autoencoder. *arXiv preprint arXiv:2505.05710*, 2025.

Jiaju Li, Hefeng Wang, Anbing Zhang, and Yuliang Liu. Semantic segmentation of hyperspectral remote sensing images based on pse-unet model. *Sensors*, 22(24), 2022. ISSN 1424-8220. doi: 10.3390/s22249678. URL https://www.mdpi.com/1424-8220/22/24/9678.

Mingshi Li, Dusan Grujicic, Steven De Saeger, Stien Heremans, Ben Somers, and Matthew B. Blaschko. Biological valuation map of flanders: A sentinel-2 imagery analysis. In *IGARSS 2024 - 2024 IEEE International Geoscience and Remote Sensing Symposium*, pp. 9539–9543, 2024a. doi: 10.1109/IGARSS53475.2024.10640788.

Mingshi Li, Dusan Grujicic, Steven De Saeger, Stien Heremans, Ben Somers, and Matthew B Blaschko. Biological valuation map of Flanders: A Sentinel-2 imagery analysis. In *IEEE International Geoscience and Remote Sensing Symposium*, 2024b.

Rui Li, Shunyi Zheng, Ce Zhang, Chenxi Duan, Jianlin Su, Libo Wang, and Peter M Atkinson. Multiattention network for semantic segmentation of fine-resolution remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–13, 2021a.

Wei Li, Guodong Wu, Fan Zhang, and Qian Du. Hyperspectral image classification using deep pixel-pair features. *IEEE Transactions on Geoscience and Remote Sensing*, 55(2):844–853, 2016.

Xiaolong Li, Yuyin Li, Jinquan Ai, Zhaohan Shu, Jing Xia, and Yuanping Xia. Semantic segmentation of uav remote sensing images based on edge feature fusing and multi-level upsampling integrated with deeplabv3+. *Plos one*, 18(1):e0279097, 2023a.

Xin Li, Feng Xu, Runliang Xia, Xin Lyu, Hongmin Gao, and Yao Tong. Hybridizing cross-level contextual and attentive representations for remote sensing imagery semantic segmentation. *Remote Sensing*, 13(15): 2986, 2021b.

Xin Li, Feng Xu, Xi Yong, Deqing Chen, Runliang Xia, Baoliu Ye, Hongmin Gao, Ziqi Chen, and Xin Lyu. Sscnet: A spectrum-space collaborative network for semantic segmentation of remote sensing images. *Remote Sensing*, 15(23), 2023b. ISSN 2072-4292. doi: 10.3390/rs15235610. URL https://www.mdpi.com/2072-4292/15/23/5610.

Yuxuan Li, Xiang Li, Yimain Dai, Qibin Hou, Li Liu, Yongxiang Liu, Ming-Ming Cheng, and Jian Yang. Lsknet: A foundation lightweight backbone for remote sensing. *International Journal of Computer Vision*, 133(3):1410–1431, Mar 2025. ISSN 1573-1405. doi: 10.1007/s11263-024-02247-9. URL `https://doi.org/10.1007/s11263-024-02247-9`.

Nannan Liao, Jianglei Gong, Wenxing Li, Cheng Li, Chaoyan Zhang, and Baolong Guo. Smale: Hyperspectral image classification via superpixels and manifold learning. *Remote Sensing*, 16(18), 2024. ISSN 2072-4292. doi: 10.3390/rs16183442. URL `https://www.mdpi.com/2072-4292/16/18/3442`.

Bo Liu, Jinwu Hu, Xiuli Bi, Weisheng Li, and Xinbo Gao. Pgnet: Positioning guidance network for semantic segmentation of very-high-resolution remote sensing images. *Remote Sensing*, 14(17), 2022. ISSN 2072-4292. doi: 10.3390/rs14174219. URL `https://www.mdpi.com/2072-4292/14/17/4219`.

Li Liu, Emad Mahrous Awwad, Yasser A. Ali, Muna Al-Razgan, Ali Maarouf, Laith Abualigah, and Azadeh Noori Hoshyar. Multi-dataset hyper-cnn for hyperspectral image segmentation of remote sensing images. *Processes*, 11(2), 2023. ISSN 2227-9717. doi: 10.3390/pr11020435. URL `https://www.mdpi.com/2227-9717/11/2/435`.

Mushui Liu, Jun Dan, Ziqian Lu, Yunlong Yu, Yingming Li, and Xi Li. Cm-unet: Hybrid cnn-mamba unet for remote sensing image semantic segmentation, 2024. URL `https://arxiv.org/abs/2405.10530`.

Pengyuan Lv, Lusha Ma, Qiaomin Li, and Fang Du. Shapeformer: A shape-enhanced vision transformer model for optical remote sensing image landslide detection. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 16:2681–2689, 2023.

Emmanuel Maggiori, Yuliya Tarabalka, Guillaume Charpiat, and Pierre Alliez. Convolutional neural networks for large-scale remote-sensing image classification. *IEEE Transactions on geoscience and remote sensing*, 55(2):645–657, 2016.

Oscar Manas, Eteni Gbodjo, Alexandre Lacoste, Marco Pedersoli, Marc-Antoine Drouin, Salvatore Gaetano, Lê V. K., and David Co. Seasonal contrast: A self-supervised framework for satellite image enhancement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pp. 2978–2987, 2021.

Vishal Nedungadi, Ankit Kariryaa, Stefan Oehmcke, Serge Belongie, Christian Igel, and Nico Lang. Mmearth: Exploring multi-modal pretext tasks for geospatial representation learning. In *European Conference on Computer Vision*, pp. 164–182. Springer, 2024.

Tareque Bashar Ovi, Shakil Mosharrof, Nomaiya Bashree, Muhammad Nazrul Islam, and Md Shofiqul Islam. Deeptrinet: A tri-level attention-based deeplabv3+ architecture for semantic segmentation of satellite images. In *International Conference on Human-Centric Smart Computing*, pp. 373–384. Springer, 2023.

Fatih Özyurt, Engin Ava, and Eser Sert. Uc-merced image classification with cnn feature reduction using wavelet entropy optimized with genetic algorithm. *Traitement du Signal*, 2020.

Ioannis Papoutsis, Nikolaos-Ioannis Bountos, Angelos Zavras, Dimitrios Michail, and Christos Tryfonopoulos. Efficient deep learning models for land cover image classification. *CoRR*, abs/2111.09451, 2021. URL `https://arxiv.org/abs/2111.09451`.

Cheng Peng, Yangyang Li, Ronghua Shang, and Licheng Jiao. Rsbnet: One-shot neural architecture search for a backbone network in remote sensing image recognition. *Neurocomputing*, 537:110–127, 2023.

Swalpa Kumar Roy, Gopal Krishna, Shiv Ram Dubey, and Bidyut B. Chaudhuri. Hybridsn: Exploring 3-d–2-d cnn feature hierarchy for hyperspectral image classification. *IEEE Geoscience and Remote Sensing Letters*, 17(2):277–281, 2020. doi: 10.1109/LGRS.2019.2918719.

Swalpa Kumar Roy, Suvojit Manna, Tiecheng Song, and Lorenzo Bruzzone. Attention-based adaptive spectral–spatial kernel resnet for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 59(9):7831–7843, 2021. doi: 10.1109/TGRS.2020.3043267.

Swalpa Kumar Roy, Ankur Deria, Danfeng Hong, Behnood Rasti, Antonio Plaza, and Jocelyn Chanussot. Multimodal fusion transformer for remote sensing image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–20, 2023.

Michael Schmitt, Lloyd H. Hughes, and Caisi Qiu. SEN12MS – A CURATED DATASET of GEOREFERENCED MULTI-SPECTRAL SENTINEL-1/2 IMAGERY for DEEP LEARNING and DATA FUSION. In *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, volume IV-2/W7, pp. 153–160, 2019. doi: 10.5194/isprs-annals-IV-2-W7-153-2019.

Gencer Sumbul, Marcela Charfuelan, Begüm Demir, and Volker Markl. Bigearthnet: A large-scale benchmark archive for remote sensing image understanding. In *IGARSS 2019-2019 IEEE international geoscience and remote sensing symposium*, pp. 5901–5904. IEEE, 2019a.

Gencer Sumbul, Marcela Charfuelan, Begüm Demir, and Volker Markl. Bigearthnet: A large-scale benchmark archive for remote sensing image understanding. In *IGARSS 2019-2019 IEEE international geoscience and remote sensing symposium*, pp. 5901–5904. IEEE, 2019b.

Xin-Yi Tong, Gui-Song Xia, Qikai Lu, Huanfeng Shen, Shengyang Li, Shucheng You, and Liangpei Zhang. Land-cover classification with high-resolution remote sensing images using transferable deep models. *Remote Sensing of Environment*, 237:111322, 2020a.

Xin-Yi Tong, Gui-Song Xia, Qikai Lu, Huanfeng Shen, Shengyang Li, Shucheng You, and Liangpei Zhang. Land-cover classification with high-resolution remote sensing images using transferable deep models. *Remote Sensing of Environment*, 237:111322, 2020b.

Gabriel Tseng, Ruben Cartuyvels, Ivan Zvonkov, Mirali Purohit, David Rolnick, and Hannah Kerner. Lightweight, pre-trained transformers for remote sensing timeseries. *arXiv preprint arXiv:2304.14065*, 2023.

Conrad Ulbricht, Erlend Mork, Manuel Foks, Guillaume Astruc, Yonatan Lencz, Selma Zaïri, Zhaoyang He, Elad David, and Natan Malkin. Galileo: Learning global & local features of many remote sensing modalities. *arXiv preprint arXiv:2502.09356*, 2025.

VITO. Sentinel-2 top of canopy (toc) products (tiles) - v2. `https://docs.terrascope.be/DataProducts/Sentinel-2/ProductsOverview.html`, 2020.

Di Wang, Jing Zhang, Minqiang Xu, Lin Liu, Dongsheng Wang, Erzhong Gao, Chengxi Han, Haonan Guo, Bo Du, Dacheng Tao, et al. Mtp: Advancing remote sensing foundation model via multi-task pretraining. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2024.

Junjue Wang, Zhuo Zheng, Ailong Ma, Xiaoyan Lu, and Yanfei Zhong. Loveda: A remote sensing land-cover dataset for domain adaptive semantic segmentation. *arXiv preprint arXiv:2110.08733*, 2021a.

Junjue Wang, Zhuo Zheng, Ailong Ma, Xiaoyan Lu, and Yanfei Zhong. Loveda: A remote sensing land-cover dataset for domain adaptive semantic segmentation. *arXiv preprint arXiv:2110.08733*, 2021b.

Libo Wang, Rui Li, Chenxi Duan, Ce Zhang, Xiaoliang Meng, and Shenghui Fang. A novel transformer based semantic segmentation scheme for fine-resolution remote sensing images. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2022a. doi: 10.1109/LGRS.2022.3143368.

Yi Wang, Conrad M Albrecht, Nassim Ait Ali Braham, Lichao Mou, and Xiao Xiang Zhu. Self-supervised learning in remote sensing: A review. *IEEE Geoscience and Remote Sensing Magazine*, 10(4):213–247, 2022b.

Yi Wang, Nassim Ait Ali Braham, Zhitong Xiong, Chenying Liu, Conrad M Albrecht, and Xiao Xiang Zhu. Ssl4eo-s12: A large-scale multimodal, multitemporal dataset for self-supervised learning in earth observation [software and data sets]. *IEEE Geoscience and Remote Sensing Magazine*, 11(3):98–106, 2023.

Syed Waqas Zamir, Aditya Arora, Akshita Gupta, Salman Khan, Guolei Sun, Fahad Shahbaz Khan, Fan Zhu, Ling Shao, Gui-Song Xia, and Xiang Bai. isaid: A large-scale dataset for instance segmentation in aerial images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 28–37, 2019.

ISPRS Commission II / WG4. Isprs potsdam 2d semantic labeling benchmark dataset. `https://www.isprs.org/resources/datasets/benchmarks/UrbanSemLab/2d-sem-label-potsdam.aspx`, 2012a.

ISPRS Commission II / WG4. Isprs vaihingen 2d semantic labeling benchmark dataset. `https://www.isprs.org/resources/datasets/benchmarks/UrbanSemLab/2d-sem-label-vaihingen.aspx`, 2012b.

Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 3–19, 2018.

Yi Yang and Shawn Newsam. Ucmerced land use dataset. `https://weegee.vision.ucmerced.edu/datasets/landuse.html`, 2010a.

Yi Yang and Shawn Newsam. UCMerced land use dataset. `https://weegee.vision.ucmerced.edu/datasets/landuse.html`, 2010b.

Wei Yuan, Xiaobo Zhang, Jibao Shi, and Jin Wang. Litest-net: a hybrid model of lite swin transformer and convolution for building extraction from remote sensing image. *Remote Sensing*, 15(8):1996, 2023.

Syed Waqas Zamir, Aditya Arora, Akshita Gupta, Salman Khan, Guolei Sun, Fahad Shahbaz Khan, Fan Zhu, Ling Shao, Gui-Song Xia, and Xiang Bai. isaid: Instance segmentation in aerial images dataset. `https://captain-whu.github.io/iSAID/`, 2019.

Cong Zhang, Jingran Su, Yakun Ju, Kin-Man Lam, and Qi Wang. Efficient inductive vision transformer for oriented object detection in remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 2023.

Zilong Zhong, Ying Li, Lingfei Ma, Jonathan Li, and Wei-Shi Zheng. Spectral–spatial transformer network for hyperspectral image classification: A factorized architecture search framework. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–15, 2022. doi: 10.1109/TGRS.2021.3115699.

Xiao Xiang Zhu, Devis Tuia, Lichao Mou, Gui-Song Xia, Liangpei Zhang, Feng Xu, and Friedrich Fraundorfer. Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE Geoscience and Remote Sensing Magazine*, 5(4):8–36, 2017. doi: 10.1109/MGRS.2017.2762307.

A Supplementary material

In this appendix, we include Table 3, Table 4, Figure 10, and Figure 11, giving insights into comparative dataset sizes, multi-season composite analysis, model convergence, and SDM module placement ablation study, respectively.

The comparative data in Table 3 is compiled from research utilizing a variety of mainstream datasets and models. Key sources include work on Salinas Scene (Center for Hyperspectral Remote Sensing Scenes (EHU/UPV), 1998b; Li et al., 2022; Liu et al., 2023; Roy et al., 2020; Liao et al., 2024), UC Merced (Yang & Newsam, 2010b; Bi et al., 2022; 2021b;a; Özyurt et al., 2020), ISPRS Vaihingen and Potsdam (WG4, 2012b;a; Liu et al., 2022; Li et al., 2021a; 2023a; Chen et al., 2018; Liu et al., 2024; Li et al., 2023b; 2021b; Wang et al., 2022a; Hanyu et al., 2024), iSAID (Zamir et al., 2019; Guo et al., 2022), LoveDA (Wang et al., 2021b; Dimitrovski et al., 2024; Hwang et al., 2024; Cha et al., 2024), GID (Tong et al., 2020b; Li et al., 2025; Ovi et al., 2023), and BigEarthNet (Sumbul et al., 2019b; Wang et al., 2023; 2022b), along with the BVM dataset used in this study (Li et al., 2024b).

Table 3: A comparison of mainstream datasets and the BVM dataset, and collection of models applied to them. Note that the pixel count (B: billion, M: million) is an estimation based on dataset specifications.

Dataset	Pixels	Model	Accuracy
Salinas Scene	0.11M	PSE-UNet	91.01% (OA)
		3D-CNN	97.55% (OA)
		HybridSN	99.84% (OA)
		SMALE	99.28% (OA)
UC Merced	137M	DenseNet-121	99.88% (OA)
		MS2AP	99.01% (OA)
		LSENet	98.69% (OA)
		VGG-VD16	95.21% (OA)
ISPRS Vaihingen	206M	PGNet	86.32% (OA)
		MANet	86.51% (OA)
		EMNet	95.42% (OA)
		DeepLabv3+	86.07% (OA)
ISPRS Potsdam	1.37B	CM-UNet	91.86% (OA)
		SSCNet	91.03% (OA)
		HCANet	90.15% (OA)
		AerialFormer-B	91.4% (OA)
		DC-Swin	92% (OA)
iSAID	2B	SegNeXt-L	70.3% (IoU)
		SegNeXt-B	69.9% (IoU)
		AerialFormer-B	69.3% (IoU)
LoveDA	6.27B	UNet-Ensemble	56.16% (IoU)
		SFA-Net	54.9% (IoU)
		ViT-G12X4	54.4% (IoU)
GID	7.34B	LSKNet-S	82.3% (OA)
		DeepTriNet	77% (OA)
BigEarthNet	9B	ResNet50	91.8% (OA)
		ViT-S	89.9% (OA)
		ResNet18	89.3% (OA)
BVM	**10.57B**	ChromaFormer-h	96.67% (OA)

Table 4: Performance of single-season vs. multi-season temporal composite models.

Model Configuration	**Overall Acc. (OA)**	**Cultivated Land Acc.**	**Woodland/Shrub Acc.**
ChromaFormer-t Models			
Spring Season Only	85.3%	82.1%	88.4%
Summer Season Only	88.1%	90.5%	89.1%
Autumn Season Only	86.5%	84.3%	87.2%
Winter Season Only	82.7%	79.8%	83.5%
4-Season Composite	**90.87%**	**92.01%**	**91.53%**
ChromaFormer-s Models			
Spring Season Only	88.0%	85.9%	89.5%
Summer Season Only	90.5%	91.8%	91.3%
Autumn Season Only	88.9%	87.5%	89.1%
Winter Season Only	85.2%	82.3%	86.0%
4-Season Composite	**93.47%**	**94.55%**	**94.02%**

A.1 Seasonal Modeling Experiments

The improved performance demonstrated by the four-season composite models is grounded in the principle of phenology. Many land cover classes that are spectrally similar in a single season, such as "Cultivated land"

and "Grasslands", exhibit unique temporal signatures throughout the year. Quantitatively, this is evidenced by the 4-season composite boosting the ChromaFormer-s model to 93.47% OA, a significant +2.97% gain over even the strongest single season (Summer, 90.5%). This gain is also seen in the challenging "Cultivated Land" class, which jumps from 91.8% to 94.55% accuracy.

Summer season yields the best results among single-season analyses due to peak vegetation health and spectral separability, while Winter performs the poorest (e.g., 85.2% for ChromaFormer-s, over 5% lower than Summer), likely due to dormant vegetation and lower light conditions masking class distinctions.

Across all scenarios, the larger ChromaFormer-s model consistently outperforms the ChromaFormer-t model, maintaining a stable 2.5% OA advantage, which demonstrates the model's robust scalability. The Spectral Dependency Module (SDM) is particularly effective here. It treats the seasonal stack as a single high-dimensional spectral-temporal input, allowing it to explicitly learn the phenological curves by modeling the crucial inter-band relationships not just within a season, but across the entire temporal stack.

A.2 Ablation on the placement of the Spectral Dependency Module (SDM)

To validate the architectural design of the ChromaFormer, we conducted an ablation study regarding the placement of the Spectral Dependency Module (SDM). Specifically, we compared our proposed early spectral fusion strategy against a late spectral fusion alternative.

Early Fusion: The SDM is inserted immediately after the patch embedding and before the first Transformer stage. This allows the model to recalibrate channel weights based on global spectral correlations before any spatial mixing occurs.

Late Fusion: The SDM is inserted within the Transformer stages, specifically after the Multi-Head Self-Attention (MSA) module in each block. This mimics the design of standard channel-attention mechanisms like SE-Blocks Hu et al. (2018) or CBAM Woo et al. (2018), where channel re-weighting is performed on spatially processed features.

As illustrated in Figure 11 the Early Fusion architecture consistently outperforms the Late Fusion variant in overall accuracy (OA). Specifically, we observe the following improvements when moving from Late Fusion to Early Fusion:

- ChromaFormer-t: Accuracy increased from 88.76% to 90.87% (+2.11%).
- ChromaFormer-b: Accuracy increased from 93.38% to 94.02% (+0.64%).
- ChromaFormer-l: Accuracy increased from 94.80% to 95.98% (+1.18%).
- ChromaFormer-h: Accuracy increased from 95.28% to 96.67% (+1.39%).

Most notably, the Early Fusion architecture resolves the performance bottleneck seen in the largest models. While the Late Fusion ChromaFormer-h (95.28%) lagged significantly behind the baseline Swin-h (96.64%), the proposed Early Fusion ChromaFormer-h (96.67%) successfully surpasses the Swin-h baseline. This suggests that as model capacity increases, decoupling spectral correlation (via Early Fusion) from spatial mixing becomes increasingly critical for maximizing accuracy.

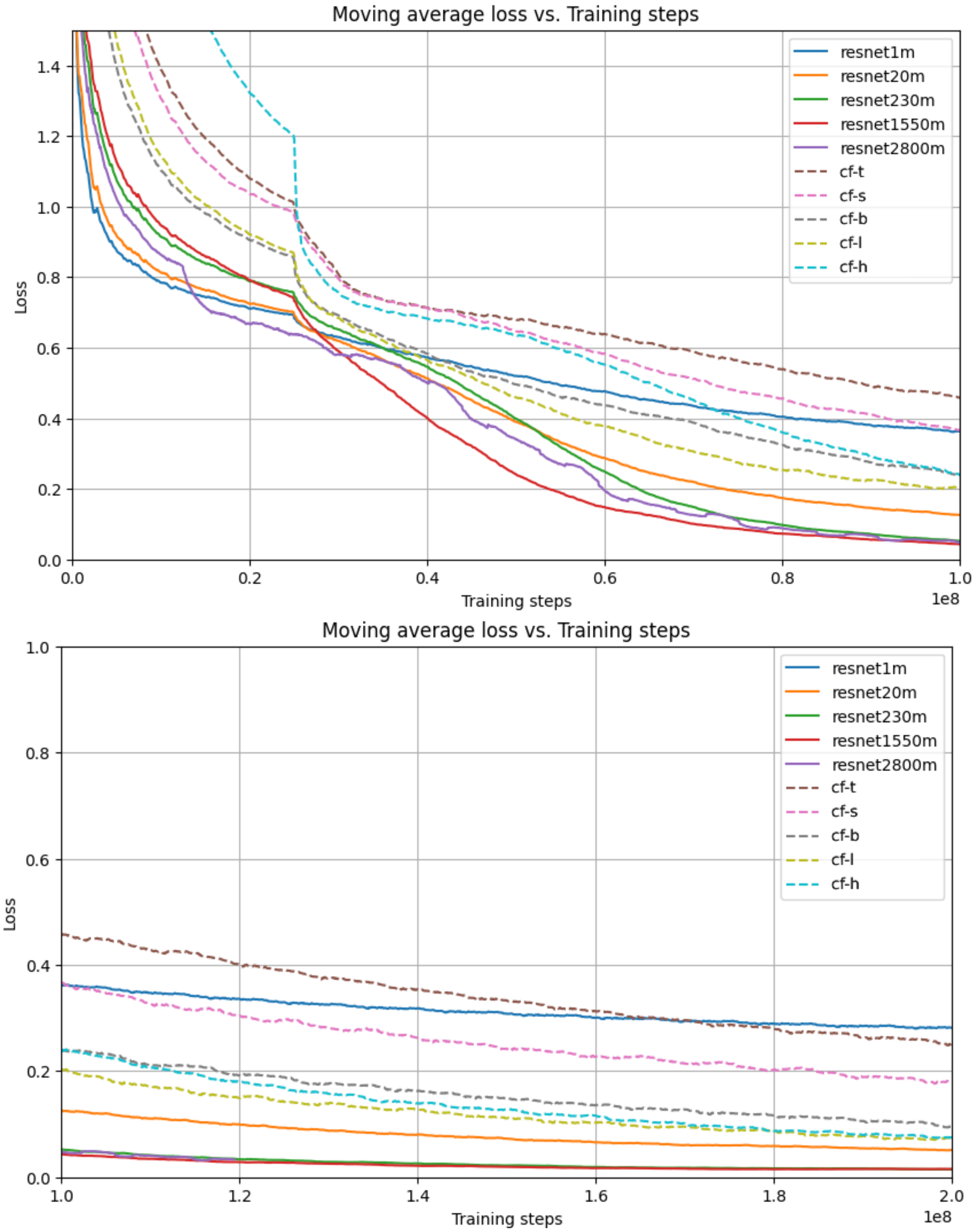


Figure 10: Loss curves of scaled networks, up: initial region of training, bottom: stable descending region. “cf” stands for ChromaFormer.

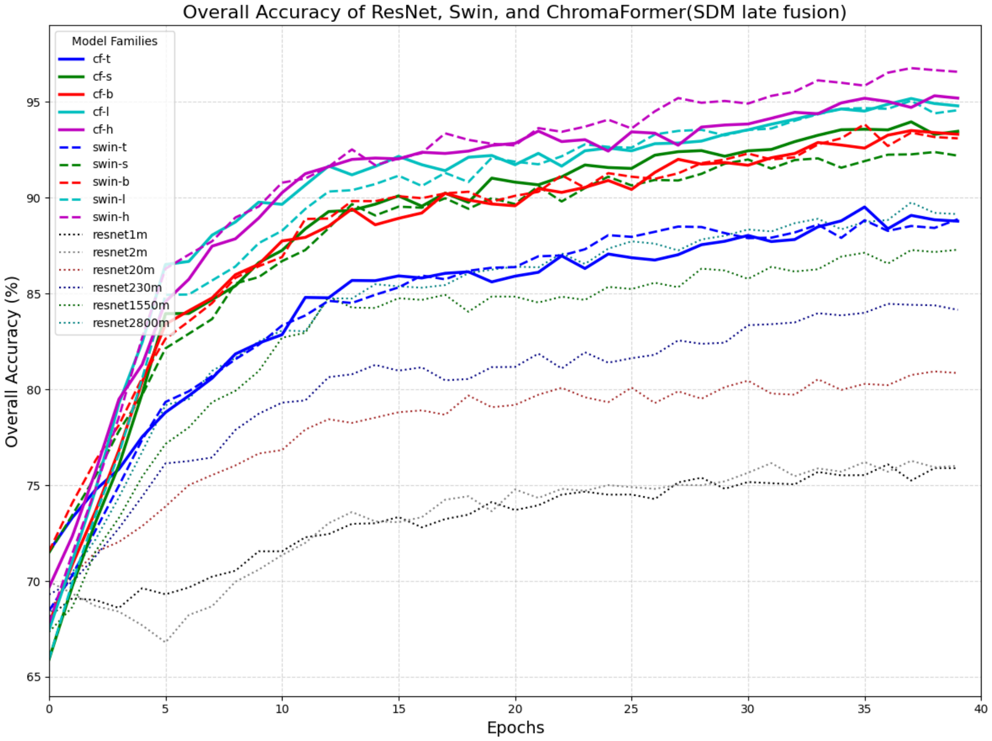


Figure 11: The overall accuracy curves of tested models, ChromaFormers in this figure are late fusion version