Vision Transformer for Multispectral Satellite Imagery: Advancing Landcover Classification[*]

Ryan Rad
Northeastern University
Khoury College of Computer Science
r.rad@northeastern.edu

Abstract

Climate change is a global issue with significant impacts on ecosystems and human populations. Accurately classifying land cover from multi-spectral satellite imagery plays a crucial role in understanding the Earth's changing landscape and its implications for environmental processes. However, traditional methods struggle with challenges like limited data availability and capturing complex spatial-spectral relationships. Vision Transformers have emerged as a promising alternative to convolutional neural networks (CNN architectures), harnessing the power of self-attention mechanisms to capture global and long-range dependencies. However, their application to multi-spectral images is still limited. In this paper, we propose a novel Vision Transformer designed for multi-spectral satellite image datasets of limited size to perform reliable land cover identification with forty-four classes. We conduct extensive experiments on a curated dataset, simulating scenarios with limited data availability, and compare our approach to alternative architectures. The results demonstrate the potential of our Vision Transformer-based method in achieving accurate land cover classification, contributing to improving climate change modeling and environmental understanding.

1. Introduction

Climate change has become a pressing global issue, impacting ecosystems and human populations worldwide. In the context of climate change modeling, accurate landcover classification plays a critical role in understanding the Earth's changing landscape and its implications for various environmental processes. Multispectral satellite imagery has proven to be a valuable resource for obtaining comprehensive and diverse information about the Earth's surface. Specifically, it provides detailed data across multiple spectral bands, allowing for a deeper understanding of land surface characteristics.

However, landcover classification from multispectral satellite images poses significant challenges, especially cloud coverage, cloud shadow, resolution, accuracy, etc [17]. Traditional methods often struggle to effectively capture the complex spatial relationships and meaningful representations within multispectral data, hindering accurate landcover classification for climate change modeling.

Recent years have witnessed remarkable breakthroughs in deep learning, particularly with the advent of vision transformers. Vision transformers have demonstrated exceptional success in various computer vision tasks by effectively capturing long-range dependencies and global context within images, surpassing the capabilities of traditional convolutional neural networks (CNNs). These models have shown promise in tasks such as object recognition and image classification, leveraging their strengths in learning meaningful representations and spatial dependencies.

In this paper, we propose a novel vision transformer-based approach designed specifically for learning from multispectral satellite images with limited data to perform accurate landcover classification in the context of climate change modeling. Our objective is to address the challenges posed by limited data availability while harnessing the capabilities of vision transformers to understand complex spatial patterns and spectral relationships present in multispectral imagery.

By integrating vision transformers into the landcover classification task, we aim to enhance the accuracy and efficiency of the classification process, ultimately contributing to better climate change modeling and environmental understanding. The vision transformer model will be trained on a carefully curated dataset of multispectral satellite images from all over the world.

The remainder of this paper is organized as follows. Section 2 provides an overview of related work in the field

[*] Thanks to Google Cloud Research Credit for providing us with access to Google's high-performance computing capabilities for this research project.

of deep learning for landcover classification and vision transformers. Section 4 details the methodology, including the architecture design of our novel vision transformer model tailored for multispectral satellite images. Section 5 presents the experimental setup, including the dataset used, evaluation metrics, and comprehensive results and analysis. We conclude in Section 6, summarizing the contributions of our approach and outlining future research directions in the domain of multispectral image analysis and climate change modeling.

2. Related Work

Deep Convolutional Neural Networks (DCNN): Convolutional Neural Networks (CNNs) have proven to be highly effective for image segmentation or pixel-wise classification tasks, with the U-Net architecture [12] being a popular choice due to its simplicity and impressive performance. Building upon U-Net, several variants have been proposed [6, 9–11] aiming to further enhance segmentation performance. The significance of global contextual information for semantic segmentation has been widely acknowledged [3,8,16]. For instance, *PSPNet* [16] introduced a pyramid pooling module that employs pooling operations at multiple scales, while *DeepLabv3* [3] proposed parallel Atrous convolutions with different rates to incorporate global context. However, it's worth noting that the pooling operation with striding in [16] led to information loss at object boundaries, and the use of dilated convolutions with a large dilation rate in [3] resulted in the ”grinding” problem. CNN-based methods have witnessed remarkable success in this field, primarily due to their exceptional ability to represent and learn intricate features from complex data.

Vision Transformers: Transformers [14] initially gained acclaim in the domain of Natural Language Processing (NLP) by achieving state-of-the-art performance. As the success of Transformers became evident, their application expanded to the realm of computer vision. Vision transformers possess a unique capability to capture global dependencies and long-range interactions, making them highly valuable for tasks such as image recognition, object detection, and semantic segmentation. Vision Transformers (ViT) [5] were introduced to handle image recognition tasks, but they require pre-training on large datasets. To address this, approaches like Deit [13] have been proposed to improve the training of ViT. A notable example is Swin TransformerSwin Transformer [7], an efficient hierarchical vision Transformer, for various vision tasks, including image classification, object detection, and semantic segmentation. One notable example of this family of models is [1]. Transformers possess much lighter inductive biases, when compared with CNNs, enabling them to benefit from extensive pre-training using vast datasets to achieve cutting-edge performance.

Self-attention/Transformer in conjunction with DCNNs: In pursuit of improved network performance, researchers have delved into the amalgamation of self-attention mechanisms, commonly employed in Transformers, with Convolutional Neural Networks (CNNs). Several approaches have emerged that integrate self-attention with CNN-based U-shaped architectures, particularly for medical image segmentation tasks. Moreover, studies have focused on synergizing Transformers and CNNs to enhance segmentation capabilities, particularly in the domains of multi-modal brain tumor segmentation and 3D medical image segmentation. Among the notable examples of this model fusion are [2] and [15], both of which demonstrate the potential of such combined architectures in medical image analysis. However, these models lie beyond the primary focus of this paper, which centers on the development of a pure Vision Transformer for multispectral satellite imagery for advanced landcover classification in the context of climate change modeling.

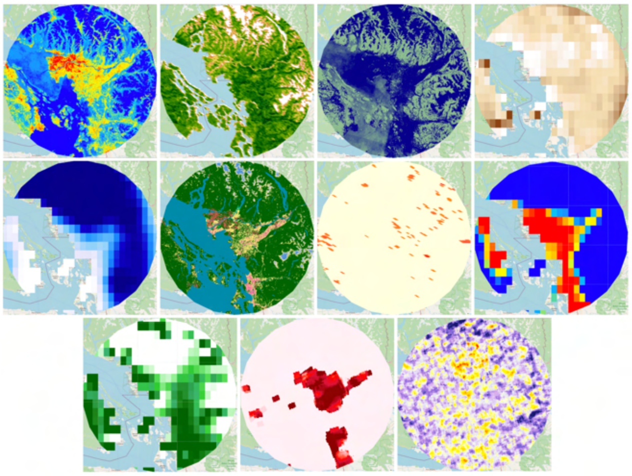

Figure 1. Sample satellite band visualization above Vancouver, BC. From left to right, Row 1: Temperature $2m$, Coastal Aerosols, EVI, and Evaporation. Row 2: Precipitation, True colour (432), T21, Soil Temperature. Row 3: LAI, CH4, and HCHO

3. Data

In this paper, we explore the *landcover identification* as a pixel-level classification problem generating landcover maps at a high resolution of 30 meters. A total number of 1m patches of size 256 × 256, each covering an area of 7680 × 7680 meters from selected countries were extracted for training. The following 15 bands were obtained from Google Earth Engine (GEE), selected bands are visualized in Fig. 1.

- From *ERA5* climate reanalysis data with 11.132 kilometers (0.1 × 0.1 decimal degree) spatial resolution:

Temperature $2m$*:* Temperature of the soil in layer 1 $(0 - 7cm)$.
Soil Temperature $L1$*:* Temperature at $2m$ height above the surface of land, sea or in-land waters.
Total Evaporation: Accumulated amount of water that has evaporated from the Earth's surface, including a simplified representation of transpiration (from vegetation), into vapor in the air above.
Total Precipitation: Accumulated liquid and frozen water, including rain and snow, that falls to the Earth's surface.
High/Low Leaf Area Index (LAI): One-half of the total green leaf area per unit horizontal ground surface area.

- From *Landsat 8* with 30 meters spatial resolution:
  B1: Atmospheric aerosols near coastal regions.
  B2-B4: Blue, green, and red.
  B10/B11: Thermal infrared 1 and 2, resampled from 100m to 30m.

- From *Copernicus S5P* climate reanalysis data with 11.132 kilometers spatial resolution:
  Tropospheric HCHO: The amount of formaldehyde in a vertical column of the atmosphere above a given location, expressed in molecules per square centimeter.
  CH4: The amount of methane (CH4) in a vertical column of the atmosphere above a given location, expressed as the volume mixing ratio of CH4 in dry air.

- From *MODIS/FIRMS* with 1 kilometer spatial resolution:
  T21: The brightness temperature of a fire pixel using MODIS channels 21/22.

4. Methodology

In recent years, the encoder-decoder family of Fully Convolutional Networks (FCN) achieved remarkable success in several application area including medical image analysis and remote sensing [4]. Inspired by this success, we focus on this type of architecture to address the challenging task of modelling formaldehyde concentration. The *UNet* architecture consists of two contracting and expansive paths to capture high-frequency details with low-frequency structures of the image. The contracting path encodes the input image into higher level representations and the expansive path decodes the compact representations into regression probability maps. We investigated several state-of-the-art models for the task of *HCHO* concentration prediction, particularly, we re-implemented and optimized *BLAST-Net* and *Swin UNet*.

4.1. Swin UNet

Swin Transformer [7] presented a vision transformer leveraging the advantages of the U-shaped architectural design where both the encoder and decoder components are constructed using Swin Transformer blocks introduced in [7]. The patch partitioning module splits an input multispectral image into non-overlapping patches. The individual patches are treated as "tokens" with their features constructed by concatenating the raw values from five multispectral bands. Assuming a patch size of 4×4 and X number of bands, each patch's feature dimension is calculated as $4 \times 4 \times X = 16X$. To project these raw-valued features into an arbitrary dimension represented as C, a linear embedding layer is applied. The Swin UNet Transformer architecture applies several Transformer blocks with shifted window self-attention computation to patch tokens. To create a hierarchical representation, patch merging layers are used to reduce the number of tokens as the network goes deeper.

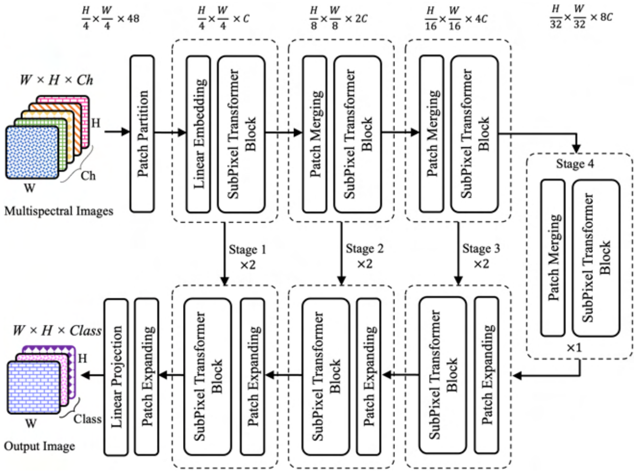
$\frac{H}{4} \times \frac{W}{4} \times 48$
$\frac{H}{4} \times \frac{W}{4} \times C$
$\frac{H}{8} \times \frac{W}{8} \times 2C$
$\frac{H}{16} \times \frac{W}{16} \times 4C$
$\frac{H}{32} \times \frac{W}{32} \times 8C$
$W \times H \times Ch$
H
Ch
W
Multispectral Images
Patch Partition
Linear Embedding
SubPixel Transformer Block
Patch Merging
SubPixel Transformer Block
Patch Merging
SubPixel Transformer Block
Stage 4
Patch Merging
SubPixel Transformer Block
×1
Stage 1 ×2
Stage 2 ×2
Stage 3 ×2
$W \times H \times Class$
H
Class
W
Output Image
Linear Projection
Patch Expanding
SubPixel Transformer Block
Patch Expanding
SubPixel Transformer Block
Patch Expanding
SubPixel Transformer Block
Patch Expanding

Figure 2. Sub-Pixel Vision Transformer Model

4.2. Overall Architecture

Building upon the remarkable success of the Swin Transformer [7], we present a novel U-shaped Encoder-Decoder architecture called *Sub-Pixel (SP) vision Transformer*, specifically designed for multispectral images. Our approach leverages the advantages of the Swin Transformer and combines them with the U-shaped architectural design. In Sub-Pixel Transformer, both the encoder and decoder components are constructed using Sub-Pixel Transformer blocks, resulting in a pure Transformer-based U-shaped architecture tailored for multispectral image analysis.

Figure 2 provides an overview of the architecture of the *Sub-Pixel Transformer*, highlighting the best performing version. The architecture shares a similar structure for stages 1 to 4 with the Swin Transformer, but it incorporates two key differences. First, spectral attention mechanism is introduced to enhance the model's ability to learn from multi-spectral images. Second, the Shifted window partitioning used in the original Swin Transformer is replaced with Sub-Pixel Partitioning. The patch partitioning module splits an input multi-spectral image into non-overlapping patches, like ViT and Swin. In the *Sub-Pixel Transformer*, individual patches are treated as ”tokens” with their features constructed by concatenating the raw values from five multi-spectral bands. Assuming a patch size of 4×4, each patch's feature dimension is calculated as $4 \times 4 \times 5 = 80$. To project these raw-valued features into an arbitrary dimension represented as C, a linear embedding layer is applied. The resulting patch tokens, along with their embedded features, are processed through multiple Swin Transformer blocks, which have modified self-attention computations.

The proposed *Sub-Pixel Transformer* architecture, applies several Transformer blocks with modified self-attention computation to patch tokens. To create a hierarchical representation, patch merging layers are used to reduce the number of tokens as the network goes deeper in a similar fashion as Swin Transformer. The first patch merging layer concatenates the features of neighboring 2×2 patches and applies a linear layer to the concatenated features. This reduces the number of tokens by a factor of 4 downsampling of resolution) and sets the output dimension to $2C$. More Transformer blocks are then applied for feature transformation, keeping the resolution at $\frac{H}{8} \times \frac{W}{8}$. This process is repeated twice for ”Stage 3” and ”Stage 4,” resulting in output resolutions of $\frac{H}{16} \times \frac{W}{16}$ and $\frac{H}{32} \times \frac{W}{32}$, respectively. By utilizing these stages, the Swin Transformer generates a hierarchical representation with the same feature map resolutions as traditional convolutional networks like VGG and ResNet. This allows the proposed architecture to easily replace the backbone networks in existing methods for various vision tasks.

4.2.1 Sub-Pixel Transformer block

Different from the conventional multi-head self attention (MSA) module, swin transformer block [7] is constructed based on shifted windows. In Figure 3, two consecutive Sub-Pixel transformer blocks are presented. Each Sub-Pixel transformer block is composed of LayerNorm (LN) layer, multi-head self attention module, residual connection and 2-layer MLP with GELU non-linearity. The based multi-head self attention (W-MSA) module and the *Sub-Pixel Multi-head Self Attention (SP-MSA)* module are applied in the two successive transformer blocks, respectively.

4.2.2 Sub-Pixel Window based Self-Attention

The use of sub-pixel shuffling window-based self-attention in the Swin Transformer instead of shifted window can be justified by several reasons:

- *Reduced Information Loss:* The sub-pixel shuffling window allows for more precise alignment of patches during the self-attention process. Unlike the shifted window approach, which shifts the patches by a fixed stride, the sub-pixel shuffling window rearranges the patches in a way that minimizes information loss, as illustrated in Figure 3. This ensures that the attention mechanism can capture fine-grained details and preserve spatial information more effectively.

- *Enhanced Local Context:* By shuffling the patches using sub-pixel shuffling, the self-attention mechanism can capture local context more accurately. This is because neighboring patches, which contain related information, are placed closer to each other, allowing the attention mechanism to better capture the dependencies and relationships within the local context. The shifted window approach may introduce misalignment and reduce the ability to capture precise local relationships.

- *Improved Global Context:* The sub-pixel shuffling window also facilitates the capture of global context in the self-attention mechanism. As the patches are rearranged, the attention mechanism can effectively attend to patches that are spatially distant but semantically related, as illustrated in Figure 3. This enables the model to capture long-range dependencies and incorporate global context information into the representation learning process.

- *Better Integration with Hierarchical Structure:* The sub-pixel shuffling window aligns well with the hierarchical structure of the Swin Transformer. As the network progresses through different stages, the patches are merged and the resolution decreases. The use of sub-pixel shuffling allows for a consistent alignment of patches across different stages, maintaining the coherence of attention patterns and facilitating information flow between different levels of the hierarchy.

Overall, the sub-pixel shuffling window-based self-attention offers improved information preservation, enhanced local and global context modeling, and better integration with the hierarchical structure of the network. These advantages justify its use over the shifted window approach.

The Sub-Pixel Transformer blocks, depicted in Figure 3 are computed as:

$$\begin{aligned}
\hat{Z}^{l} &= \text{W-MSA}\ (\text{LN}(Z^{l-1})) + Z^{l-1}, \\
Z^{l} &= \text{MLP}\ (\text{LN}(\hat{Z}^{l})) + \hat{Z}^{l}, \\
\hat{Z}^{l+1} &= \text{SP-MSA}\ (\text{LN}(Z^{l})) + Z^{l}, \\
Z^{l+1} &= \text{MLP}\ (\text{LN}(\hat{Z}^{l+1})) + \hat{Z}^{l+1}
\end{aligned} \tag{1}$$

where $\hat{z}^{l}$ and z^{l} denote the output features of the *W-MSA* and *SP-MSA* modules and the *MLP* module for block

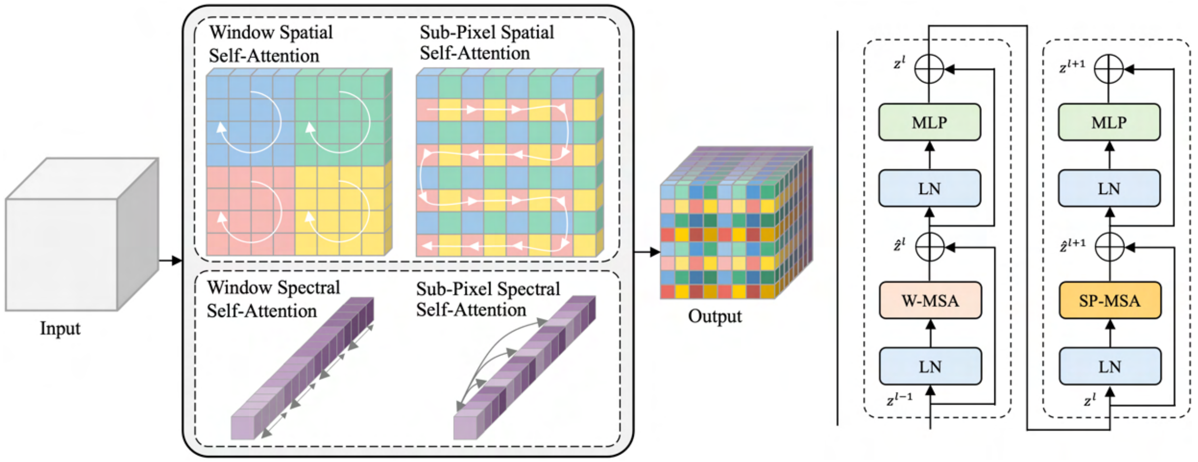
Window Spatial Self-Attention
Sub-Pixel Spatial Self-Attention
Input
Window Spectral Self-Attention
Sub-Pixel Spectral Self-Attention
Output
z^l
MLP
LN
$\hat{z}^l$
W-MSA
LN
z^{l-1}
z^{l+1}
MLP
LN
$\hat{z}^{l+1}$
SP-MSA
LN
z^l

Figure 3. Left: Regular window-based and Sub-Pixel partitioning (spectral and spatial), and Right: A pair of Sub-Pixel Transformer Blocks.

l, respectively. *W-MSA* and *SP-MSA* denote window based multi-head self-attention (spatial and spectral) using regular and sub-pixel partitioning configurations, respectively. The sub-pixel partitioning approach introduces connections between neighboring non-overlapping windows in the previous layer and is found to be effective in image classification, object detection, and semantic segmentation, as discussed in Sec. 5.

5. Experimental Results

5.1. Experimental Setup

5.1.1 Dataset

We created a dataset for landcover classification from Google Earth Engine (GEE). We extracted our input bands from several satellites as described in 3, and obtained the landcover labels from the Copernicus CORINE which consists of 44 landcover classes. We used a separate set of countries for training and test sets with around 1m patches for training and 100k for test purposes. These bands were selected to capture different aspects of the target environment and enable effective training of our model.

5.1.2 Implementation Details

The proposed *SubPixel Trans.* model is implemented based on Python 3.8 and $Pytorch 1.8.0$. In order to enhance the diversity of the training data, a comprehensive range of data augmentation techniques, including flips, scaling, translation, brightness, Gaussian noise, synthetic clouds, and rotations, are applied to all training patches. The input image size and patch size are set as 224×224 and 128, respectively. We train our model on four NVIDIA Tesla T4 with a total of 64GB memory on Google cloud in northamerica-northeast2-Toronto region. During the training period, the SGD optimizer with momentum 0.9 for 100 epochs using a cosine decay learning rate scheduler and 20 epochs of linear warm-up.

5.2. Quantitative Results

5.2.1 Landcover Classification Performance

To highlight the effectiveness of the proposed model, several widely adopted architectures such as Baseline *UNet* [12] and TernausNet [6] (winner of the Carvana challenge), PSPNet [16], DeepLab V3 [3], Blast-Net [10], Trans UNet [2], and Swin UNet [1] are considered. Table 1 compares the performance of the proposed *Sub-Pixel Transformer* model to that of other models considered in this study. Both the model performance and the ecological aspects of the experimentation are presented. The *Dice Score or F*1 values for landcover classification indicates the accuracy of the models in delineating the regions. Both the mean and standard deviation of Dice Score or F1 are reported, providing insights into the consistency and reliability of the segmentation results. Additionally, the table includes information on the Energy Consumption (EC) and carbon footprint (CO2e) of the models. This transparency is important to promote sustainable machine learning applications by considering the environmental impact of training and running the models.

5.2.2 Ablution Study

Furthermore, in our study of the proposed *Sub-Pixel Transformer* model, we investigated the individual contributions of its two key components: Sub-Pixel Window based partitioning and the spectral self-attention. To highlight the significance of these components, we conducted two separate

Table 1. landcover classification performance comparison (*Dice Score or F*1 %). Best results are in bold **black** and second-best ones are in **teal**.

Models	Size	Dice ± stdev	EC (kWh)	CO$_2$e (lbs)
UNet [12]	**6m**	79.83 ±2.8	**42.0**	**40.1**
TernausNet [6]	**10m**	81.98 ±4.9	58.2	55.7
PSPNet [16]	35m	84.56 ±5.0	51.5	49.1
DeepLab V3 [3]	40m	85.25 ±4.7	48.1	45.99
Blast-Net [10]	25m	85.85 ±4.8	43.3	41.4
Trans UNet [2]	42m	86.37 ±4.4	57.8	55.2
Swin UNet [1]	29m	87.75 ±4.3	54.8	52.3
Ours w/o Sub-Pixel	17m	88.19 ±3.9	40.7	38.9
Ours w/o Spectral	16m	**89.01 ±3.7**	43.7	41.8
Ours	18m	**89.97** ±2.9	**40.5**	**38.6**

Table 2. Pairwise comparisons between models using the *Tukey's HSD* test.

Model 1	Model 2	p-value (adj)	Reject
Ours	UNet [12]	0.0000	True
Ours	TernausNet [6]	0.0000	True
Ours	PSPNet [16]	0.0010	True
Ours	DeepLab V3 [3]	0.0019	True
Ours	Blast-Net [10]	0.0024	True
Ours	ViT [5]	0.0041	True
Ours	Trans UNet [2]	0.0052	True
Ours	Swin UNet [1]	0.2765	False

experiments, eliminating one component at a time.

5.2.3 Statistical Significance

To validate the significance of these differences, we employed statistical tests, specifically an analysis of variance (ANOVA) followed by Tukey's Honestly Significant Difference (HSD) test. The post hoc test results provided in Figure 2 offer insights into which pairs of methods have statistically significant performance differences and which differences might be considered marginal. While certain performance differences may not exhibit statistical significance, it's important to note that even modest improvements hold the potential to provide insights in the critical context of climate change.

5.2.4 Performance and Training Size

Figure 4 illustrates the performance of different models concerning the training ratio, which represents the proportion of available data used for training. This visualization showcases how well the models can learn from limited data. Remarkably, some models demonstrate exceptional learning capabilities even with minimal training data, suggesting their potential for efficient knowledge extraction and utilization. On the other hand, the performance of DCNNs tends to saturate earlier in comparison. The vision transformer-based models [1, 2], despite benefiting from 2-3 times more data availability, still face certain limitations. However, the proposed *SubPixel Trans.* model outperforms all the other considered models across all training ratios, signifying its superiority in handling limited data scenarios.

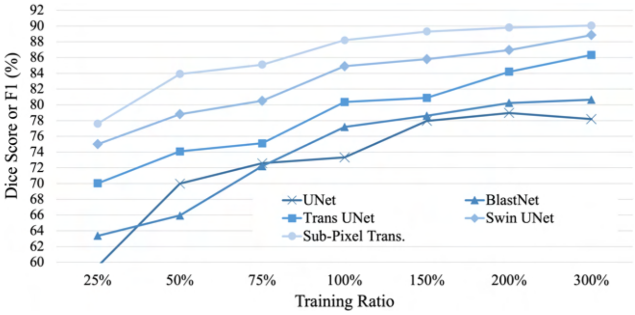

Figure 4. Performance comparison of models with different size training set.

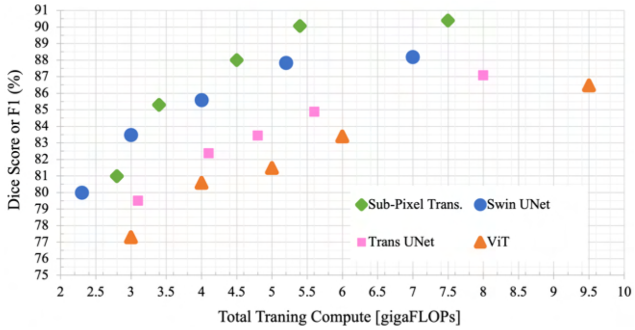

Figure 5. Performance versus training compute for different models: Vision Transformers [5], Swin UNet [1], Trans UNet [2], and our Sub-Pixel Trans.

5.2.5 Performance Gains and Computational Demands

Our experimental results demonstrate the effectiveness of the *Sub-Pixel Transformer* vision transformer in improving wildfire detection accuracy, particularly in scenarios where long-range dependencies are essential. However, it is important to consider the potential trade-offs introduced by the increased computational complexity compared to other models considered. Figure 5 contains the performance versus total training compute. First, it can be observed that *Sub-Pixel Transformer* outperforms *ViT* and *Trans UNet* on the performance/compute trade-off. Second, *Sub-Pixel Transformer* uses approximately $10\% - 15\%$ more compute to attain the same performance as *Swin UNet*. Third, at larger computational budgets *Sub-Pixel Transformer* outperforms *Swin UNet* by 2% in Dice Score, but that improvement comes with 8% more computational demand.

5.2.6 Latency and Throughput

Furthermore, the throughput in terms of images processed per second (image/s) is measured. This metric provides insights into the model's efficiency during inference which is particularly relevant to real-time applications such as wildfire detection. It turns out that our *Sub-Pixel Transformer* vision transformer achieves an inference throughput of 681.8 images/sec, surpassing the performance of the *ViT* model, which achieves 632.1 images/sec by a margin of 50. Furthermore, the *Sub-Pixel Transformer* falls short of *Swin UNet* (with 715.3 images/sec.) by a mere 35 images/sec.

5.3. Qualitative Results

Figure 6 presents the qualitative results comparing the landcover classification maps generated by the proposed *Sub-Pixel Transformer* model and those of *Trans UNet* [2] and *Swin UNet* [1]. The figure consists of four rows, each representing an entire test country. The first column displays false-color Landsat-8 patches and the second column shows the landcover ground truth data obtained from the Copernicus CORINE. Columns c and d showcase the classification error of *Swin UNet* [1] and our proposed vision transformer, respectively. Each black pixel represents a wrong prediction (false positive and false negative). The qualitative results provide a visual assessment of the model's performance in accurately delineating landcover errors and highlight the effectiveness of the proposed approach in learning from multispectral images for landcover classification task.

6. Conclusion

Our proposed Sub-Pixel Transformer architecture demonstrated improved accuracy and efficiency in modeling complex spatial patterns and spectral relationships present in multispectral imagery. The model's ability to capture global dependencies and long-range interactions contributed to its success in handling multispectral data with limited training samples. While we focused on landcover classification to gain better insights on and monitor the impacts of climate change, we recognize the importance of adapting and assessing our approach for other multi-spectral image processing tasks with varying characteristics. Future research directions may include exploring ways to further enhance the model's performance in handling limited data scenarios and extending its application to other remote sensing tasks. Additionally, investigating techniques for unsupervised or weakly supervised learning with vision transformers could unlock new possibilities for landcover classification and environmental analysis.

References

[1] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. In Leonid Karlinsky, Tomer Michaeli, and Ko Nishino, editors, *Computer Vision – ECCV 2022 Workshops*, pages 205–218, Cham, 2022. Springer Nature Switzerland. 2, 5, 6, 7

[2] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L. Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *CoRR*, abs/2102.04306, 2021. 2, 5, 6, 7

[3] L. C. Chen, G. Papandreou, F. Schroff, and H. Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 2, 5, 6

[4] Gabriel Henrique de Almeida Pereira, Andre Minoro Fusioka, Bogdan Tomoyuki Nassu, and Rodrigo Minetto. Active fire detection in landsat-8 imagery: A large-scale dataset and a deep-learning study. *ISPRS Journal of Photogrammetry and Remote Sensing*, 178:171–186, 2021. 3

[5] 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. *CoRR*, abs/2010.11929, 2020. 2, 6

[6] V. Iglovikov and A. Shvets. Ternausnet: U-net with vgg11 encoder pre-trained on imagenet for image segmentation. *arXiv preprint arXiv:1801.05746*, 2018. 2, 5, 6

[7] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, and Baining Guo. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12009–12019, June 2022. 2, 3, 4

[8] W. Luo, Y. Li, R. Urtasun, and R. Zemel. Understanding the effective receptive field in deep convolutional neural networks. In *Advances in Neural Info. Process. Syst.*, pages 4898–4906, 2016. 2

[9] R. M. Rad, P. Saeedi, J. Au, and J. Havelock. Multiresolutional ensemble of stacked dilated u-net for inner cell mass segmentation in human embryonic images. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 3518–3522, 2018. 2

[10] R. M. Rad, P. Saeedi, J. Au, and J. Havelock. Blast-net: Semantic segmentation of human blastocyst components via cascaded atrous pyramid and dense progressive upsampling. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 1865–1869, 2019. 2, 5, 6

[11] R. M. Rad, P. Saeedi, J. Au, and J. Havelock. Cell-net: Embryonic cell counting and centroid localization via residual incremental atrous pyramid and progressive upsampling convolution. *IEEE Access*, 7:81945–81955, 2019. 2

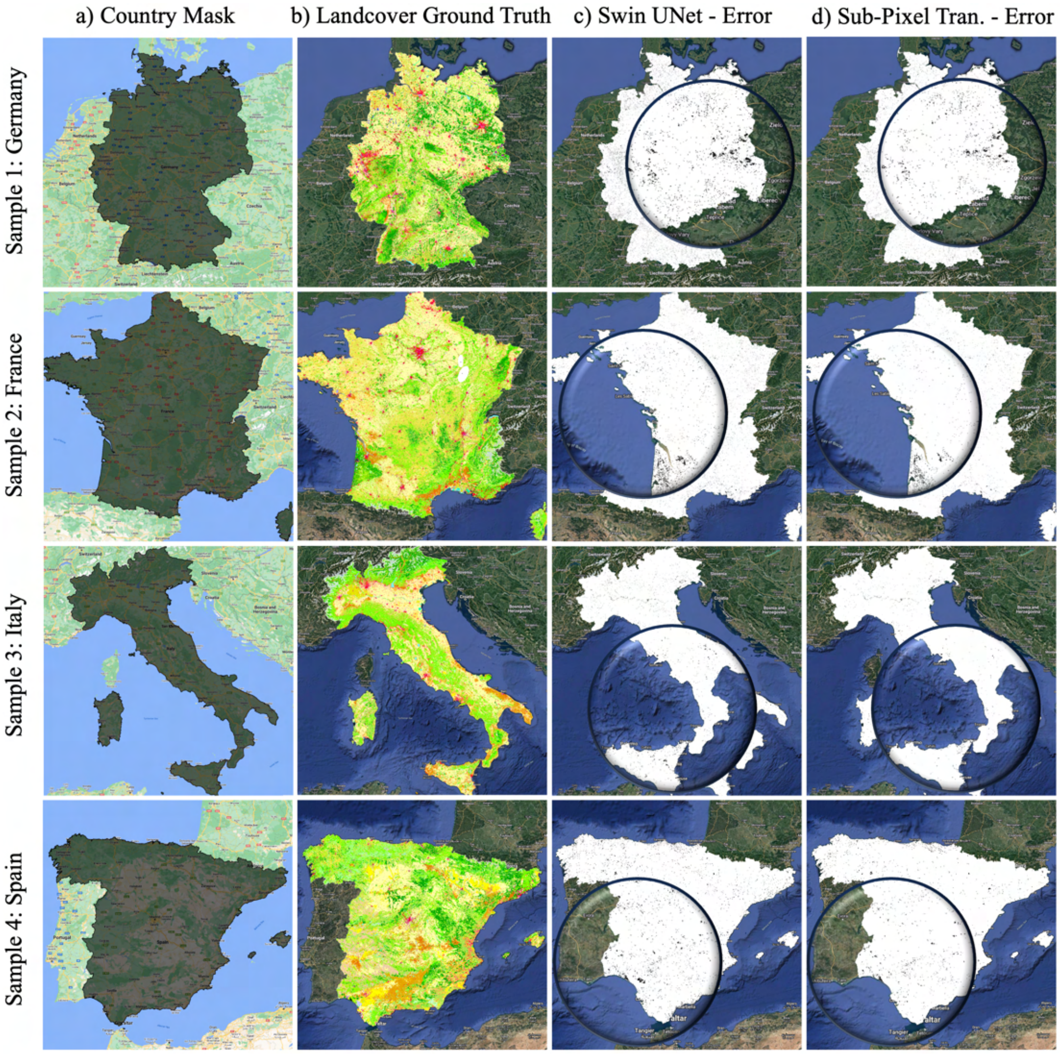

Figure 6. Qualitative comparison of the our produced landcover maps (by *Sub-Pixel Transformer*) and the ground truth in some European countries.

[12] O. Ronneberger, P. Fischer, and T. Brox. *U-Net: Convolutional Networks for Biomedical Image Segmentation*, pages 234–241. Springer International Publishing, Cham, 2015. 2, 5, 6

[13] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. *CoRR*, abs/2012.12877, 2020. 2

[14] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017. 2

[15] Yundong Zhang, Huiye Liu, and Qiang Hu. Transfuse: Fusing transformers and cnns for medical image segmentation. *CoRR*, abs/2102.08005, 2021. 2

[16] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid scene parsing network. In *Proc. IEEE Conf. on Comput. Vision and Pattern Recognition*, pages 2881–2890, 2017. 2, 5, 6

[17] Decheng Zhou, Jingfeng Xiao, Stefania Bonafoni, Christian Berger, Kaveh Deilami, Yuyu Zhou, Steve Frolking, Rui Yao, Zhi Qiao, and José A. Sobrino. Satellite remote sensing of surface urban heat islands: Progress, challenges, and perspectives. *Remote Sensing*, 11(1), 2019. 1