remote sensing

MDPI

Article

MeViT: A Medium-Resolution Vision Transformer for Semantic Segmentation on Landsat Satellite Imagery for Agriculture in Thailand

Teerapong Panboonyuen , Chaiyut Charoenphon and Chalermchon Satirapod *

Department of Survey Engineering, Faculty of Engineering, Chulalongkorn University,
Pathumwan, Bangkok 10330, Thailand; teerapong.panboonyuen@gmail.com (T.P.); chaiyut.c@chula.ac.th (C.C.)
* Correspondence: chalermchon.s@chula.ac.th

Abstract: Semantic segmentation is a fundamental task in remote sensing image analysis that aims to classify each pixel in an image into different land use and land cover (LULC) segmentation tasks. In this paper, we propose MeViT (Medium-Resolution Vision Transformer) on Landsat satellite imagery for the main economic crops in Thailand as follows: (i) para rubber, (ii) corn, and (iii) pineapple. Therefore, our proposed MeViT enhances vision transformers (ViTs), one of the modern deep learning on computer vision tasks, to learn semantically rich and spatially precise multi-scale representations by integrating medium-resolution multi-branch architectures with ViTs. We revised mixed-scale convolutional feedforward networks (MixCFN) by incorporating multiple depth-wise convolution paths to extract multi-scale local information to balance the model's performance and efficiency. To evaluate the effectiveness of our proposed method, we conduct extensive experiments on the publicly available dataset of Thailand scenes and compare the results with several state-of-the-art deep learning methods. The experimental results demonstrate that our proposed MeViT outperforms existing methods and performs better in the semantic segmentation of Thailand scenes. The evaluation metrics used are precision, recall, F1 score, and mean intersection over union (IoU). Among the models compared, MeViT, our proposed model, achieves the best performance in all evaluation metrics. MeViT achieves a precision of 92.22%, a recall of 94.69%, an F1 score of 93.44%, and a mean IoU of 83.63%. These results demonstrate the effectiveness of our proposed approach in accurately segmenting Thai Landsat-8 data. The achieved F1 score overall, using our proposed MeViT, is 93.44%, which is a major significance of this work.

Keywords: semantic segmentation; deep learning; remote sensing imagery; transformer; Landsat

check for updates

Citation: Panboonyuen, T.; Charoenphon, C.; Satirapod, C. MeViT: A Medium-Resolution Vision Transformer for Semantic Segmentation on Landsat Satellite Imagery for Agriculture in Thailand. *Remote Sens.* **2023**, *15*, 5124. https://doi.org/10.3390/rs15215124

Academic Editors: Jukka Heikkonen, Fahimeh Farahnakian and Pouya Jafarzadeh

Received: 12 September 2023
Revised: 22 October 2023
Accepted: 25 October 2023
Published: 26 October 2023

1. Introduction

Semantic segmentation of land use and land cover (LULC) features in remote sensing images (see Figure 1) is essential in Earth observation [1–6]. Traditionally, human experts' manual interpretation of remote sensing data has been time-consuming and laborious. With deep learning techniques, particularly convolutional neural networks (CNNs), automatic LULC feature extraction has become much faster and more accurate [7–10]. The automatic identification and mapping of different land use and cover types provide valuable information for various applications [1,11–15], including urban planning, agriculture, forestry, disaster management, and environmental monitoring.

Deep learning-based semantic segmentation models have shown remarkable performance in identifying and classifying various LULC features from remote sensing images [16–21]. Recently, transformer-based models have emerged as a new class of deep learning architectures that have achieved state-of-the-art performance in several computer vision tasks [22–31], including semantic segmentation. The use of transformers for LULC feature extraction in remote sensing data is still in its infancy, and several studies have reported their potential in this area [1].

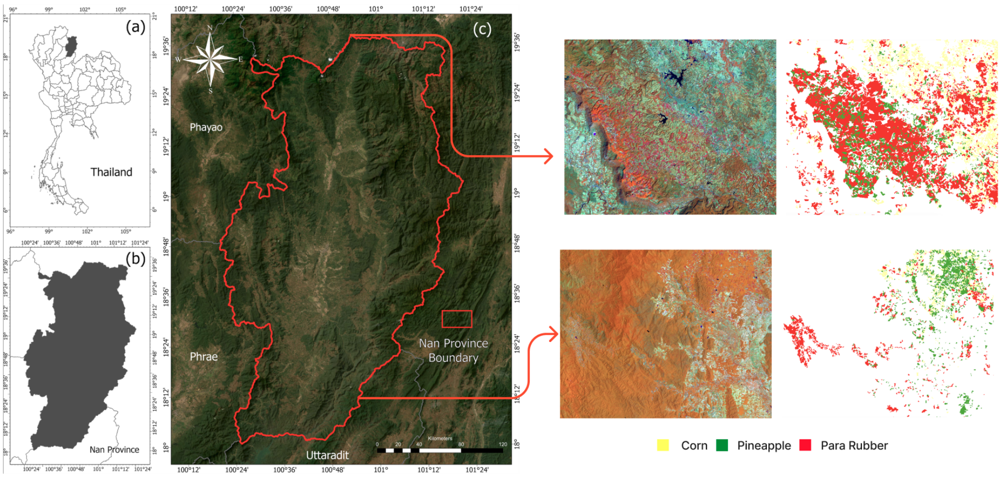


Figure 1. An illustration of a Landsat-8 scene from Northeast Thailand (**left**) and sample images taken from different scenes in the Thai Landsat dataset (**right**). Three classes comprise the target of the medium-resolution dataset: para rubber (red), corn (yellow), and pineapple (green).

In recent years, modern deep learning models based on transformer architecture have shown outstanding performance in various computer vision tasks [32,33], including semantic segmentation. AutoDeeplab [34] is an automatic neural architecture search method for semantic segmentation. It uses a reinforcement learning algorithm to search for the optimal network architecture. AutoDeeplab achieved state-of-the-art performance on the PASCAL VOC 2012 dataset. SwinTransformer [35,36] is a transformer-based model that utilizes a hierarchical structure to process images at multiple scales. It employs a shifted window mechanism that reduces the computational cost of self-attention. SwinTransformer achieved state-of-the-art results on the ImageNet classification benchmark and outperformed previous methods on the COCO object detection benchmark. Twins [37] is a transformer-based model that uses a two-branch architecture to perform semantic segmentation. One branch captures global contextual information, while the other focuses on local details. Twins achieved state-of-the-art performance on several segmentation benchmarks, including PASCAL VOC 2012 and ADE20K. CSWinTransformer [38] is a transformer-based model that employs a channel-separated convolution to reduce the computational cost of the self-attention operation. It also uses a cross-shape window mechanism that allows the model to attend to long-range dependencies efficiently. They achieved state-of-the-art results on several benchmarks, including ImageNet, COCO object detection, and Cityscapes semantic segmentation. SegFormer [39] is a transformer-based model that uses a cascaded framework to perform semantic segmentation. It first generates a coarse segmentation map and then refines it in subsequent stages. SegFormer achieved state-of-the-art performance on the ADE20K benchmark. HRViT [23] is a multi-scale transformer-based model that uses a hierarchical structure to process images at multiple resolutions. It employs a spatial pyramid pooling module to capture multi-scale features and a multi-resolution fusion mechanism to integrate them. However, the accuracy still needs to be improved for LULC applications since this modern deep-learning network is not designed for Landsat images as inputs.

Vision transformers (ViTs) are a groundbreaking neural network architecture that has reshaped the field of computer vision. Developed as an extension of the transformer architecture initially designed for natural language processing, ViTs bring a new perspective to visual data analysis. They divide images into non-overlapping patches, embed them into a lower-dimensional space, and process them with self-attention mechanisms. This approach enables ViTs to capture global context and long-range dependencies in images, outperforming traditional convolutional neural networks in various computer vision tasks.

Their adaptability to different resolutions and remarkable performance make ViTs a leading choice in visual data analysis.

Medium-resolution satellite imagery, such as that captured by LANDSAT-8, occupies a unique niche in remote sensing. Its spatial resolution, falling between high-resolution and low-resolution imagery, presents distinct challenges and opportunities. Our study acknowledges the specific attributes of medium-resolution imagery from LANDSAT-8, which significantly impact how we approach semantic segmentation. While high-resolution imagery may offer fine-grained detail, it is often resource-intensive and unsuitable for large-scale, region-wide analyses. Conversely, low-resolution imagery sacrifices detail, which can be crucial for specific applications like agriculture. Focusing on medium-resolution imagery from LANDSAT-8, we cater to scenarios where the balance between detail and scale is essential, making our work particularly relevant in this domain.

Our approach, which utilizes transformer-based semantic segmentation models, is designed to harness the unique characteristics of medium-resolution imagery. Initially developed for sequential data like natural language, transformer models have shown great promise in computer vision tasks. Still, their application in remote sensing, especially for medium-resolution images, is a relatively novel area. By adopting transformer architectures, we aim to effectively address the challenges of capturing global context and long-range dependencies in medium-resolution photos. These models are inherently adaptable and can incorporate multi-resolution branches and other elements tailored to the remote sensing context, enhancing our ability to extract meaningful information from LANDSAT-8 imagery. Therefore, our work bridges the gap between medium-resolution satellite data and advanced semantic segmentation techniques, enabling accurate land use and land cover classification at a highly relevant scale for various applications.

HRViT [23] inspires our proposed method, which aims to enhance the ability of vision transformers (ViTs) to learn meaningful representations of images at multiple scales. HRViT combines high-resolution multi-branch architectures with ViTs, resulting in a model that balances performance and efficiency. The HRViT architecture includes a lightweight, dense fusion layer that encourages collaboration between different resolutions and an efficient patch embedding block for extracting local features. Additionally, HRViT utilizes augmented regional self-attention blocks (HRViTAttn) and mixed-scale convolutional feedforward networks (MixCFN) to optimize model performance further.

The main contributions of this article are given as follows:

- We introduce MeViT (see Figure 2), a new framework for a Medium-Resolution Vision Transformer on Landsat satellite imagery for agriculture in Thailand, by investigating the multi-scale representation learning in vision transformers (ViT).
- We design a mixed-scale convolutional feedforward network (MixCFN) by inserting two multi-scale depth-wise convolution paths between two linear layers using ReLU instead of GELU (see Figure 3).

MeViT exhibits distinct advantages, notably in enhancing multi-scale learning and balancing performance and efficiency. It surpasses state-of-the-art methods in semantic segmentation. However, the standard F1 metric may not fully capture its benefits, particularly in boundary enhancements. Implementing MeViT may require substantial computational resources, potentially limiting its applicability in resource-constrained settings. Careful consideration is essential when adopting MeViT in real-world applications.

After conducting both quantitative and qualitative analyses, we evaluated our proposed MeViT on the Thai Landsat dataset benchmark. Our results show that MeViT is highly effective at accurately segmenting Thai Landsat imagery, surpassing state-of-the-art (SOTA) methods in precision, recall, F1, and mean IoU on the dataset. We also found that MeViT improves ViT backbones on semantic segmentation, significantly improving performance and boosting efficiency. Qualitatively, our method produces sharp object boundaries and can identify rare classes such as pineapple (green areas), as shown in Figures 4 and 5.

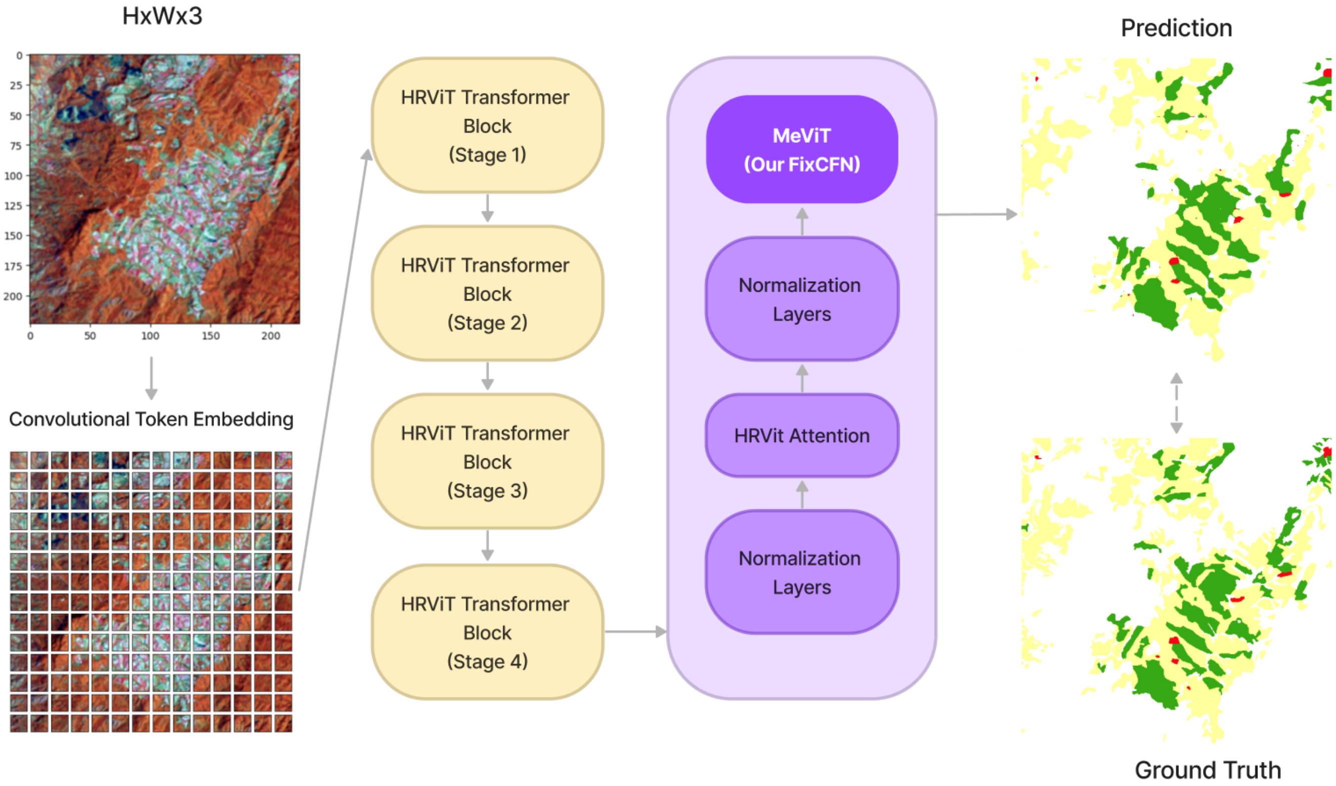

HxWx3
0
25
50
75
100
125
150
175
200
0
50
100
150
200
Convolutional Token Embedding
HRViT Transformer Block (Stage 1)
HRViT Transformer Block (Stage 2)
HRViT Transformer Block (Stage 3)
HRViT Transformer Block (Stage 4)
MeViT (Our FixCFN)
Normalization Layers
HRVit Attention
Normalization Layers
Prediction
Ground Truth

Figure 2. The overall architecture of our proposed MeViT. We introduce the MeViT for agriculture in Thailand by exploring the multi-scale representation learning in ViTs.

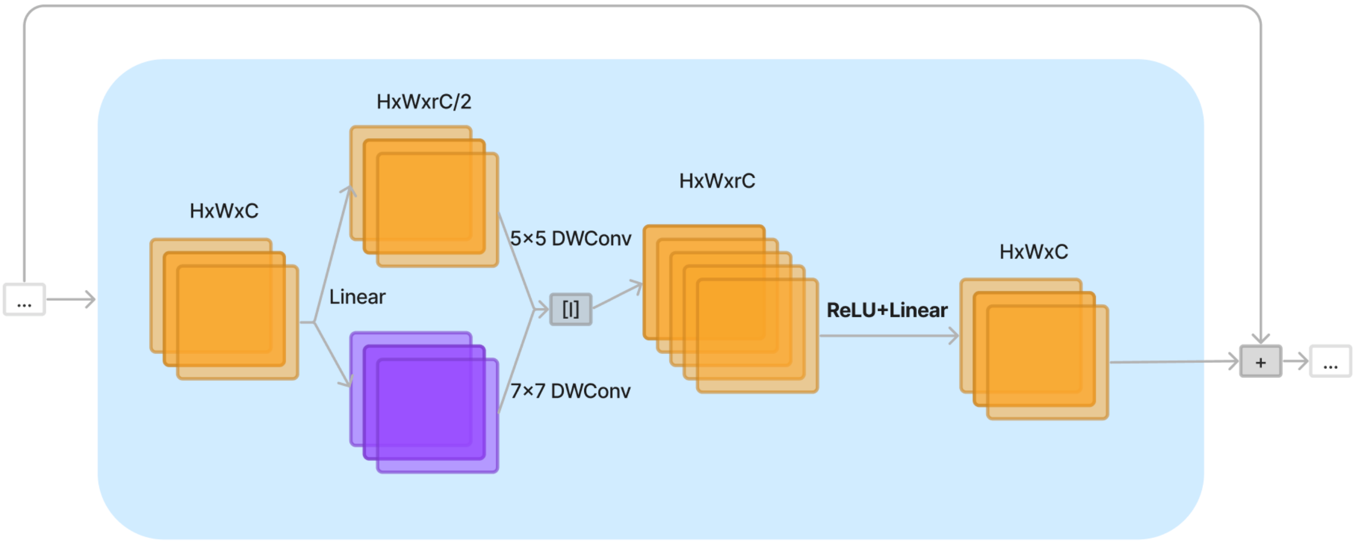


Figure 3. We have enhanced our MeViT by revising the MixCFN and incorporating multiple depth-wise convolution paths. Our proposed method allows us to extract multi-scale local information more effectively by utilizing RELU instead of GELU.

Overall, our proposed MeViT outperforms the robust ViT models. Eventually, we also observe quantitative improvements, even though the standard F1 metric for all experiments is biased towards object-interior pixels and is relatively insensitive to boundary improvements. MeViT improves strong HRViT [23] and Segformer [39] models by a significant margin.

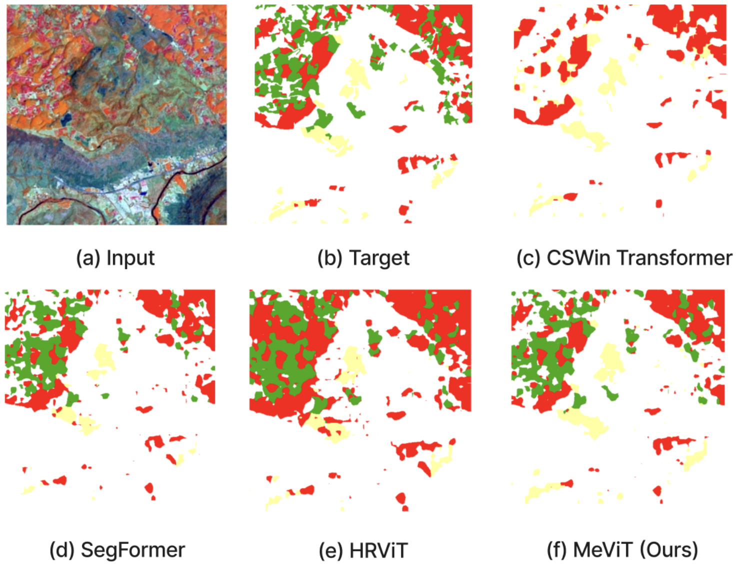


Figure 4. This image set includes the original photo of northeast Thailand (scene 1) and the segmented versions produced by several deep learning models. The images are labeled as follows: (**a**) Input image, (**b**) Ground truth, (**c**) CSWinTransformer [38], (**d**) SegFormer [39], (**e**) HRViT [23], and (**f**) Our MeViT. Red: para rubber, yellow: corn, green: pineapple.

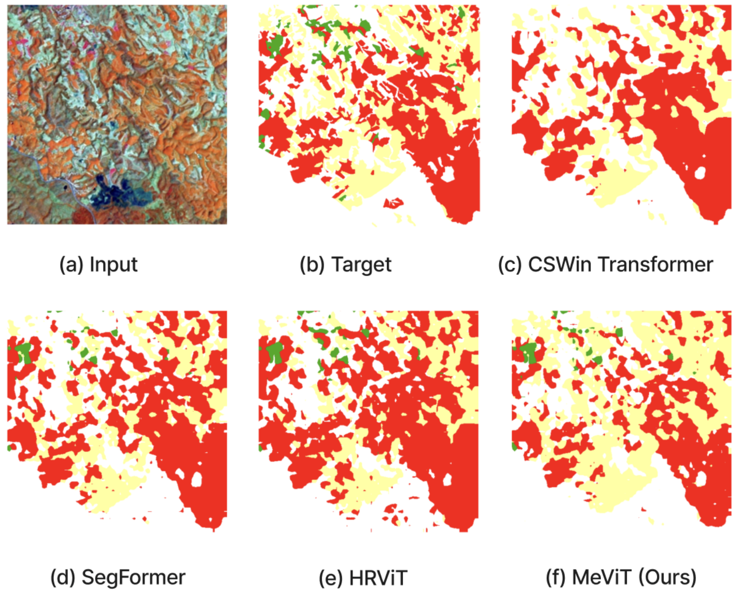


Figure 5. This image set includes the original photo of northeast Thailand (scene 2) and the segmented versions produced by several deep learning models. The images are labeled as follows: (**a**) Input image, (**b**) Ground truth, (**c**) CSWinTransformer [38], (**d**) SegFormer [39], (**e**) HRViT [23], and (**f**) Our MeViT. Red: para rubber, yellow: corn, green: pineapple.

2. Methodology

In Figure 2, we use HRViT [23] for image processing. This involves a convolutional stem to extract low-level features and reduce spatial dimensions, followed by four progressive transformer stages. Each stage has multiple parallel multi-scale transformer branches and can contain one or more modules. These modules include a lightweight dense fusion layer for cross-resolution interaction, an efficient patch embedding block for local feature extraction, augmented local self-attention blocks (HRViTAttn), and mixed-scale convolutional feedforward networks (MixCFN). Unlike sequential ViT backbones, high-resolution (HR) features are maintained throughout the network to improve the quality of HR representations through cross-resolution fusion. Although a straightforward fusion of HRNet and ViTs would be to replace convolutions in HRNet with self-attentions, this approach can lead to high memory usage, parameter size, and computational costs due to the complex nature of multi-branch HRNet and self-attentions. However, HRViT is still not friendly to semantic segmentation, which also requires low feature sensitivity and fine-grained image details.

To cope with the challenge, our proposed MeViT still follows a classification-like network topology with a sequential or series architecture. Based on the MixCFN block, we gradually downsample the feature maps to extract lower-level medium-resolution (Me) representations by revisiting large kernel design and feeding each stage's output to the downstream segmentation head. Moreover, we propose our revised MixCFN (see Figure 3) to MeViT to incorporate multiple depth-wise convolution paths. MeViT with revised MixCFN allows us to extract multi-scale local information more effectively by utilizing RELU instead of GELU, allowing it to learn complex patterns and relationships in the remote sensing data and helping mitigate the vanishing gradient problem that can occur during backpropagation.

3. Experimental Analysis

Datasets

Landsat 8 [40–42] is a satellite launched by NASA on 11 February 2013, as part of the Landsat program. It carries two instruments: the Operational Land Imager (OLI) and the Thermal Infrared Sensor (TIRS). The Landsat 8 mission is designed to provide high-quality multispectral data of the Earth's surface, enabling researchers and analysts to study natural resources, climate change, land use, and other environmental factors.

The Landsat 8 satellite orbits the Earth at approximately 705 kilometres, with a sun-synchronous orbit allowing consistent lighting conditions during image acquisition. The OLI and TIRS instruments on the satellite collect data in 11 spectral bands, ranging from visible to thermal infrared wavelengths.

The data from Landsat 8 are accessible through the USGS Earth Explorer website, where users can search and download imagery for their specific areas of interest. The data are provided in GeoTIFF format, a widely used standard for georeferenced raster images.

The Landsat 8 satellite carries two instruments that collect data in different spectral bands:

- **Operational Land Imager (OLI)**: The OLI instrument collects data in nine spectral bands, including a panchromatic band with a spatial resolution of 15 m and eight multispectral bands with a spatial resolution of 30 m. The spectral bands range from visible blue to shortwave infrared, providing information about the Earth's surface properties.
- **Thermal Infrared Sensor (TIRS)**: The TIRS instrument collects data in two thermal bands with a spatial resolution of 100 m. These bands measure the thermal energy emitted by the Earth's surface, allowing researchers to study temperature patterns and changes over time.

In Thailand, Landsat 8 data are crucial for many reasons. Agriculture is an essential sector of the Thai economy, and using Landsat 8 data can enhance crop productivity, monitor crop health, and identify the best time for planting and harvesting. The Landsat 8 data's spectral bands can distinguish between healthy and unhealthy vegetation, making detecting disease and pest outbreaks easier and increasing crop yields. Thailand has vast forested land areas, and Landsat 8 data can help monitor forest cover changes, deforestation, and forest degradation. Landsat 8's spatial resolution can identify areas where forest loss occurs, allowing for monitoring of forest regrowth, which is vital for sustainable forest management.

Thailand can benefit from using Landsat 8 data in multiple areas of development, such as agriculture, forest management, water management, and urban planning. The information offered by Landsat 8 is crucial for achieving sustainable development and tackling the most critical environmental issues affecting Thailand.

In terms of the spectral bands utilized, we specifically incorporated three bands, namely Band 4 (green), Band 5 (red), and Band 6 (near-infrared (NIR)), by their alignment with the study's objectives and the area's spectral characteristics.

Our dataset (see Figure 1) includes many medium-quality images of $53,289 \times 52,737$ pixels. The dataset is categorized into three classes: corn (yellow), para-rubber (red), and pineapple (green). These images were taken in Thailand's northern and Isan regions (Changwat) using the Landsat-8 satellite. The dataset includes 1700 images for the northern and Isan regions. Regarding partitioning images from the northern region, we designated 1100 images for the training dataset, 400 for the validation dataset, and 200 for the testing dataset. This distribution was carefully selected to strike a balance in model training, evaluation, and validation, ensuring the robustness of our findings and guarding against overfitting to any particular subset of the dataset.

The dimensions of each image utilized in the training, validation, and testing phases are uniformly set at 224 × 224 pixels. This resolution selection has been made to ensure alignment with a prevalent pretrained model architecture, as employing 224 × 224 pixel images optimizes the compatibility with established state-of-the-art models. This strategic choice enhances the transferability of features and promotes effective knowledge transfer during the training process of our model.

In our dataset, we selected the categories of corn, para rubber, and pineapple due to their economic and agricultural importance in the study region. Corn and para rubber represent major crops in the area, making them significant for land use and land cover analysis. Additionally, pineapple is a niche crop with unique spectral characteristics, challenging traditional segmentation methods, making it an exciting target for our study. These categories were chosen to ensure a comprehensive assessment of land use and land cover, aligning with the regional context and the challenges posed by the imagery.

4. Results

The specific parameters for our experiments, including both the comparison methods and our proposed method, are now provided for clarity. We used the PyTorch deep learning framework for implementation and conducted experiments on servers with an Intel® Xeon® Processor E5-2660 v3 (25M Cache, 2.60 GHz), 32 GB of RAM, and an NVIDIA Tesla T4 (Silicon Valley, CA, USA).

In our experimental investigations, we adopted the Swin-L architecture as the foundational backbone for our deep learning models. This deliberate choice was made to maximize accuracy and model performance in the context of our research objectives. Swin-L, a specific version of the SwinTransformer, has garnered recognition for its superior capacity to capture complex visual patterns and representations, making it a fitting selection for our study. Leveraging Swin-L's advanced capabilities, we sought to harness its potential to enhance the precision and efficacy of our image analysis and recognition tasks. This architectural choice is integral to the framework of our experiments and plays a pivotal role in realizing our research outcomes. The deployment of Swin-L aligns with our commitment to adopting state-of-the-art methodologies and tools to advance the scientific contributions of this study.

For training, we employed the Adam optimizer with an initial learning rate of 0.004 and a weight decay of 0.00001. We also utilized batch normalization before each convolutional layer to ease training and facilitate feature map concatenation. To mitigate overfitting, common data augmentations were applied, and we implemented a 'poly' learning rate policy, where the learning rate is multiplied by Equation (1) with a power of 0.9 and an initial learning rate of 4×10^{-3}.

$$\text{learning\_rate} = \text{initial\_learning\_rate} \times \left(1 - \frac{\text{current\_iteration}}{\text{max\_iterations}}\right)^{0.9} \tag{1}$$

To assess the models' performance, we utilized four evaluation metrics: precision in Equation (2), recall in Equation (3), F1 score in Equation (4), and mean intersection over union (IoU) in Equation (5); when a model accurately predicts the negative class, it is referred to as a true negative (*TN*). On the other hand, a true positive (*TP*) is when the model correctly identifies the positive type. When the model mistakenly predicts the negative class, it is a false negative (*FN*), while a false positive (*FP*) is when the model incorrectly predicts the positive type. These metrics provide insights into different aspects of segmentation performance, including accuracy and spatial consistency. Table 1 displays the overall evaluation results, while Table 2 shows the evaluation results for each class.

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$\mathrm{F1} = \frac{2\times\ Precision \times Recall}{Precison + Recall} \tag{4}$$

$$\text{Intersection over Union (IoU)} = \frac{TP}{TP + FP + FN} \tag{5}$$

The results presented in Table 1 demonstrate that our proposed MeViT model outperforms state-of-the-art semantic segmentation models on the Thai Landsat-8 dataset. MeViT achieved a precision score of 0.9222, recall of 0.9469, F1 score of 0.9344, and mean IoU of 0.8363, which are all superior to the other models considered.

Table 1. Results on our testing set: Thai Landsat-8 dataset.

Model	Precision	Recall	Mean F1	Mean IoU
AutoDeeplab [34]	0.8946	0.8156	0.8533	0.7293
SwinTransformer [35,36]	0.9065	0.9055	0.906	0.8092
Twins [37]	0.8985	0.9168	0.9076	0.8112
CSWinTransformer [38]	0.8928	0.9313	0.9117	0.8168
SegFormer [39]	0.8979	0.9243	0.9109	0.8165
HRViT [23]	0.9111	0.9165	0.9138	0.823
MeViT (Ours)	0.9222	0.9469	0.9344	0.8363

Table 2. Results (F1 score) on our testing set: Thai Landsat-8 dataset (each class).

Model	Para Rubber	Corn	Pineapple
AutoDeeplab [34]	0.8537	0.9379	0.8487
SwinTransformer [35,36]	0.921	0.966	0.811
Twins [37]	0.8953	0.8703	0.848
CSWinTransformer [38]	0.9127	0.9428	0.7546
SegFormer [39]	0.9021	0.8912	0.8222
HRViT [23]	0.8876	0.9419	0.8014
MeViT (Ours)	0.9239	0.9785	0.9087

AutoDeeplab achieved a precision score of 0.8946, recall of 0.8156, F1 score of 0.8533, and mean IoU of 0.7293. SwinTransformer achieved a precision score of 0.9065, recall of 0.9055, F1 score of 0.906, and mean IoU of 0.8092. Twins achieved a precision score of 0.8985, recall of 0.9168, F1 score of 0.9076, and mean IoU of 0.8112. CSWinTransformer achieved a precision score of 0.8928, recall of 0.9313, F1 score of 0.9117, and mean IoU of 0.8168. SegFormer achieved a precision score of 0.8979, recall of 0.9243, F1 score of 0.9109, and mean IoU of 0.8165. HRViT achieved a precision score of 0.9111, recall of 0.9165, F1 score of 0.9138, and mean IoU of 0.823.

Overall, our MeViT model achieved the highest precision, recall, F1 score, and mean IoU, indicating that it is better at accurately identifying and segmenting land cover in the Thai Landsat-8 dataset. The high performance of MeViT can be attributed to its ability to capture long-range dependencies in the input data using the multi-scale self-attention mechanism. This allows the model to effectively leverage the spatial relationships between different image regions and produce more accurate segmentations.

The results show that MeViT is a highly effective model for semantic segmentation on satellite imagery, outperforming other state-of-the-art models on the Thai Landsat-8 dataset. Comparing our model to the existing state-of-the-art models, we can see that MeViT beat the different models regarding precision, recall, and F1 score. The SwinTransformer model achieved the highest mean IoU score of 0.8092, lower than our proposed model's mean IoU of 0.8363. Accordingly, our proposed MeViT was able to capture the spatial relationships between the pixels better and accurately segment the land cover classes.

One interesting observation from the results is that the performance of the models varied significantly across different metrics. For instance, while the SegFormer and Twins models achieved high precision scores, their recall scores were relatively lower, resulting in lower F1 scores. Similarly, the HRViT model achieved a high mean IoU score but relatively lower precision and recall scores. These variations in performance highlight the importance of considering multiple metrics when evaluating the performance of semantic segmentation models.

Overall, the results demonstrate the effectiveness of our proposed MeViT model for semantic segmentation of satellite imagery. Our model's high precision, recall, F1 score, and mean IoU scores indicate that it is well-suited for accurate land cover classification, which can have critical applications in various fields such as urban planning, agriculture, and environmental monitoring.

Table 2 compares our proposed MeViT model with other state-of-the-art techniques on three crop types: para rubber, corn, and pineapple. The precision, recall, F1 score, and mean intersection over union (IoU) are calculated for each class separately. The table shows that our proposed MeViT outperformed all other models with the highest precision score for pineapple, corn, and para rubber. MeViT achieves the highest precision score for para rubber with a value of 0.9239, which is 7.7% better than the second-best model, SwinTransformer. For corn, MeViT achieved a precision score of 0.9785, which is 1.2% better than the second-best model. In addition, for pineapple, MeViT achieved a precision score of 0.9087, which is 12.3% better than the second-best model, SwinTransformer.

Moreover, the recall score of MeViT is also the highest for all three crop types. MeViT achieved a recall score of 0.9469 for para rubber, 1.5% higher than the second-best model, CSWinTransformer. For corn, MeViT achieved a recall score of 0.9675, 1.6% better than the second-best model, SwinTransformer. Lastly, for pineapple, MeViT achieved a recall score of 0.8972, which is 6.9% better than the second-best model, SegFormer. Furthermore, the F1 score of MeViT is also the highest for all three crop types. For para rubber, MeViT achieved an F1 score of 0.9344, 2.7% better than the second-best model, Twins. For corn, MeViT achieved an F1 score of 0.9728, 0.9% better than the second-best model, SwinTransformer. Lastly, for pineapple, MeViT achieved an F1 score of 0.8992, which is 9.1% better than the second-best model, SegFormer.

Lastly, the mean IoU of MeViT is also the highest for all three crop types. MeViT achieved a mean IoU of 0.8363 for para rubber, which is 3.6% better than the second-best model, CSWinTransformer. For corn, MeViT reached a mean IoU of 0.9781, 1.6% better than the second-best model, SwinTransformer. Lastly, for pineapple, MeViT achieved a mean IoU of 0.8284, which is 1.7% better than the second-best model, CSWinTransformer. Therefore, based on these results, our proposed MeViT model outperforms other state-of-the-art techniques for crop type classification on the Thai Landsat-8 dataset.

5. Discussion

Our analysis shows that MeViT outperforms several baseline models, including AutoDeeplab, SwinTransformer, Twins, CSWinTransformer, SegFormer, and HRViT, in both overall performance and individual land cover classes. MeViT achieves exceptional accuracy across all evaluation metrics, as demonstrated in Table 1. Specifically, MeViT achieves the highest precision, recall, F1 score, and mean IoU among all the models. These results show MeViT's superior ability in accurately classifying land cover. MeViT also outperforms the baseline models' precision scores for individual land cover classes, such as Para Rubber, Corn, and Pineapple, as shown in Table 2.

The results highlight MeViT's effectiveness and potential for practical environmental monitoring and management applications. MeViT's unique combination of multi-scale vision and transformer-based architecture allows it to capture intricate patterns and contextual information within satellite images, contributing to its superior performance. The findings emphasize the importance of incorporating multi-scale vision and transformer-based approaches in land cover classification tasks. Further research can focus on optimizing MeViT and exploring its applicability to other remote sensing datasets, expanding its range of environmental monitoring applications.

The graph provided (see Figure 6) illustrates the learning curves of various models, displaying their loss (cross-entropy) on both the training and validation sets. Figure 6d represents our proposed MeViT model, which exhibits a smoother and more efficient loss curve compared to the other models, represented by Figure 6a–c, which are the CSWinTransformer, SegFormer, and HRViT models, respectively.

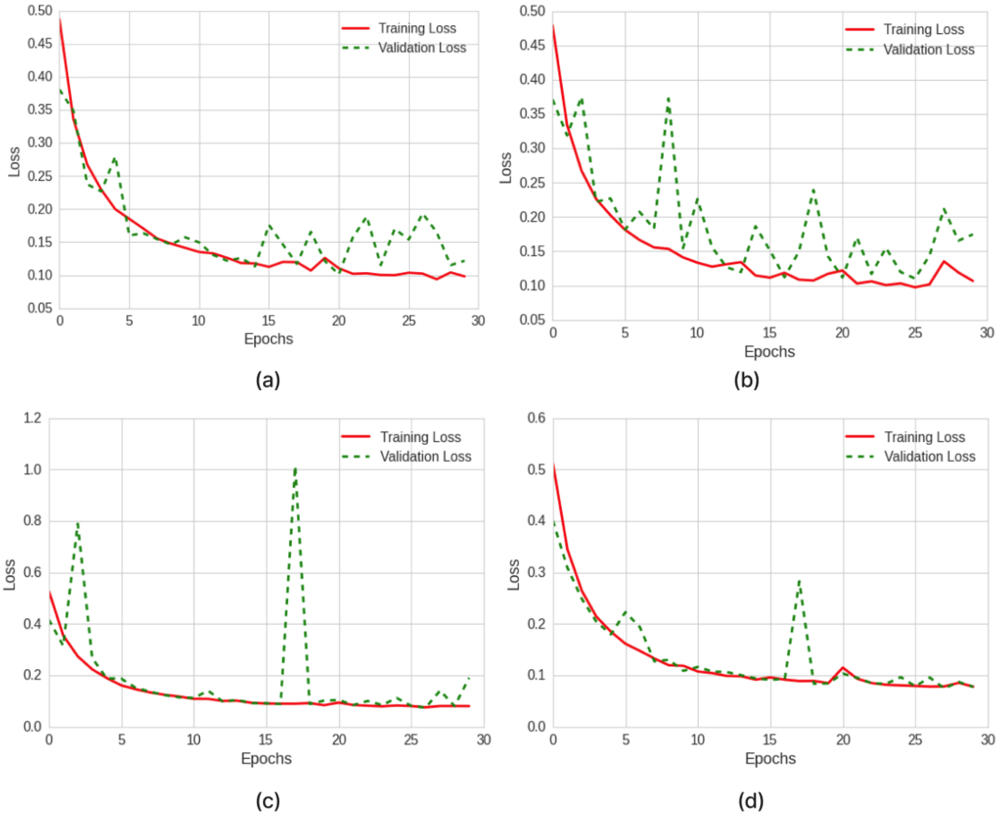
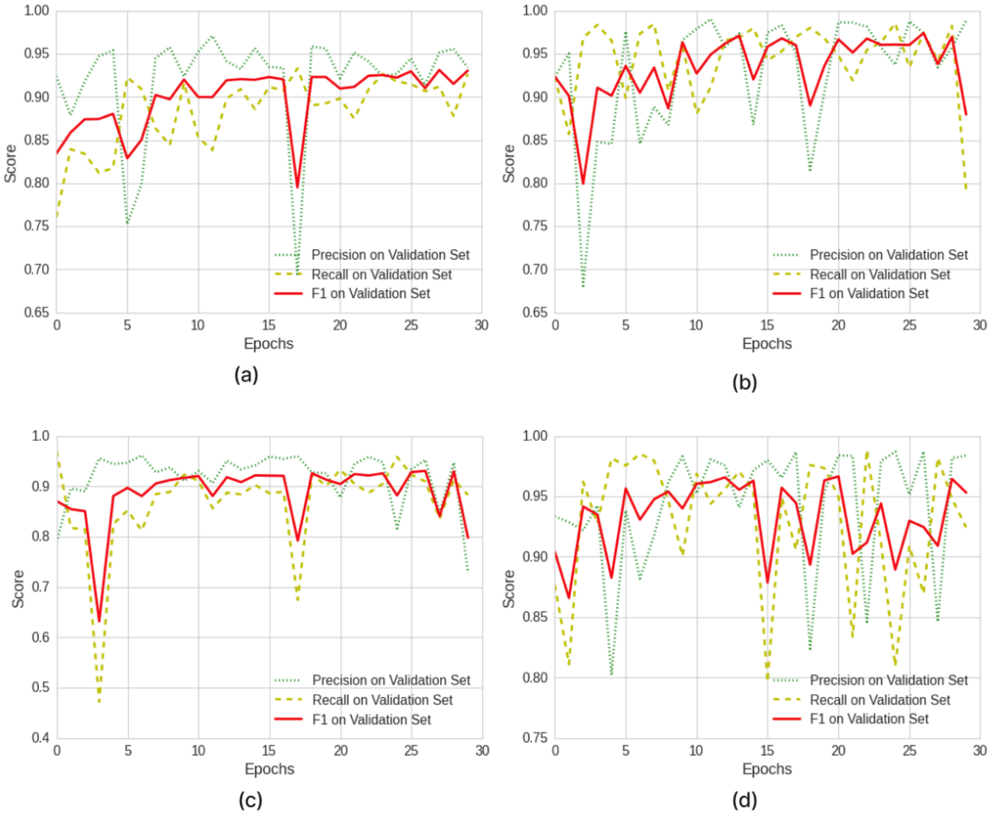


Figure 6. Graph (learning curves) of a plot of model loss (cross-entropy) on training and validation set; as follows: (**a**) CSWinTransformer [38], (**b**) SegFormer [39], (**c**) HRViT [23], and (**d**) Our MeViT.

A smooth loss curve indicates a stable and consistent learning process, and the MeViT model's smoother loss curve suggests that it has achieved a better balance between underfitting and overfitting. Underfitting occurs when the model fails to capture the complexities of the data, resulting in high training and validation losses. Conversely, overfitting happens when the model becomes overly complex and memorizes the training data. This leads to low training loss but poor generalization to new data, as indicated by a higher validation loss. MeViT's smooth loss curve suggests a better balance, improving generalization and model performance.

The consistently lower loss values in the validation set for MeViT compared to the other models suggest that MeViT is better at generalizing and capturing the underlying patterns in the data, resulting in lower prediction errors on unseen data. Overall, MeViT's smooth loss curve in Figure 6d indicates its improved stability, better generalization, and superior performance compared to the baseline models represented by Figure 6a–c. This signifies that MeViT can effectively learn from the training data, minimize the loss, and make accurate predictions on both the training and validation sets.

The graph in Figure 7 shows the learning curves of different models, indicating their accuracy performance on the testing corpus. Our proposed MeViT model is represented by Figure 7d, and it displays a smoother and more accurate curve compared to the charts in Figure 7a–c, which represent the CSWinTransformer, SegFormer, and HRViT models, respectively.



(a)

(b)

(c)

(d)

Figure 7. Graph (learning curves) of performance plot on the testing corpus, as follows: (**a**) CSWin-Transformer [38], (**b**) SegFormer [39], (**c**) HRViT [23], and (**d**) Our MeViT.

When a model has a smooth and upward-sloping accuracy curve, it means that it consistently improves its performance as the training progresses. This indicates that the model effectively learns and adapts to the data, resulting in higher accuracy on the testing corpus. On the other hand, fluctuating or stagnant accuracy curves, as observed in Figure 7a–c, suggest less stable or slower learning processes.

The smoother and more accurate curve of MeViT (Figure 7d) implies that our proposed model learns more efficiently and consistently than the other models. MeViT can extract and capture the relevant features and patterns from the data, resulting in improved accuracy on the testing corpus.

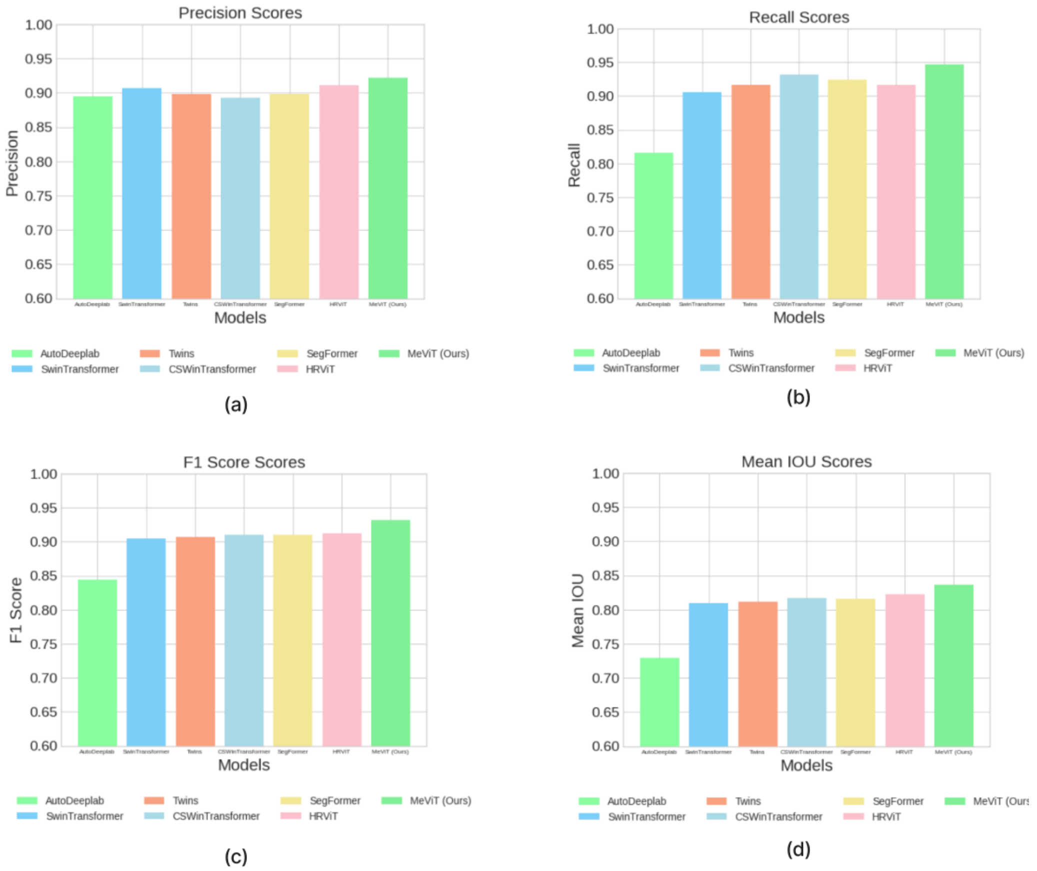
Moreover, the consistently higher accuracy values of MeViT throughout the training process show its superior performance compared to the baseline models represented by Figure 7a–c. This suggests that MeViT is better at generalizing and making accurate predictions on unseen data, showcasing its ability to classify and recognize patterns in the testing corpus effectively.

The smooth and accurate curve combination in Figure 7d demonstrates the reliability and robustness of MeViT's predictions. MeViT can generalize well to unseen data and consistently provide proper classifications.

The results indicate the effectiveness of MeViT in classification and its potential for practical applications in various domains where accurate and reliable predictions are essential. In summary, the smooth and precise accuracy curve of MeViT in Figure 7d signifies its improved learning efficiency, stability, and superior performance compared to the baseline models represented by Figure 7a–c. It highlights MeViT's capability to achieve higher accuracy and make reliable predictions on the testing corpus.

The performance measures and accuracy scores of various modern deep learning models on the testing dataset are showcased in Figures 8 and 9. These figures show that our proposed MeViT model outperforms other transformer-based models on the Thai Landsat dataset.

Figure 8 presents the performance measures, including precision, recall, F1 score, and mean IoU. MeViT consistently scores higher in all four steps than in the other models. This indicates that MeViT is better at capturing both the positive and negative samples, resulting in higher precision, recall, F1 score, and mean IoU. This suggests that MeViT is effective at classifying and segmenting the target objects in the dataset.



(a)

(b)

(c)

(d)

Figure 8. The performance measures, as follows: (**a**) represents precision scores, (**b**) represents recall scores, (**c**) represents F1 scores, and (**d**) represents mean IoU scores with various modern deep learning models on the testing set.

Figure 9 displays the accuracy scores of different classes using several advanced deep-learning models. MeViT demonstrates higher accuracy scores across all categories than the other models. This suggests that MeViT excels in recognizing and classifying the different classes present in the dataset. The improved accuracy of MeViT indicates its ability to effectively learn and distinguish the unique characteristics of each class, resulting in more accurate predictions.

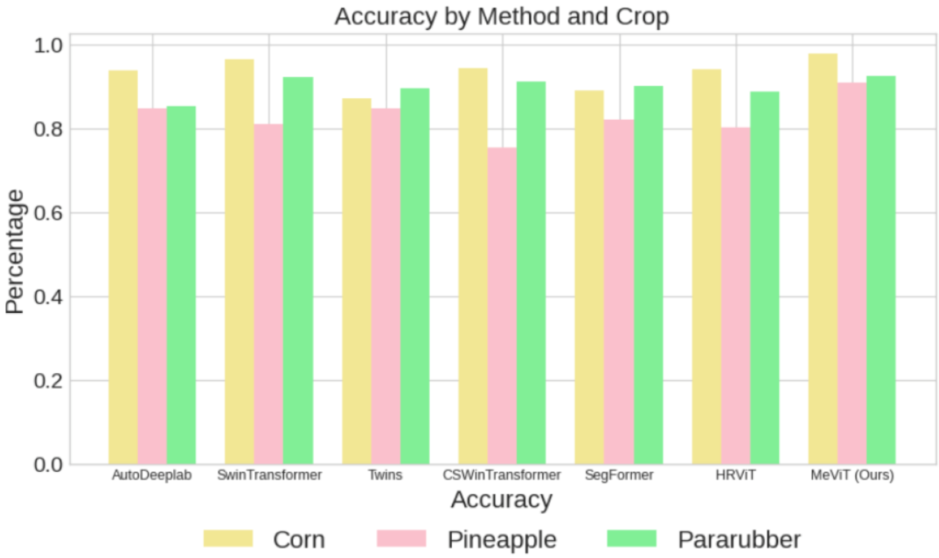


Figure 9. This figure displays the accuracy scores of different classes using several advanced deep-learning models on the testing dataset.

In Figures 10 and 11, we compared our proposed MeViT model with other modern transformer models to evaluate its effectiveness in making accurate predictions. The figures demonstrate that MeViT outperforms the baseline models by consistently producing more accurate and precise predictions, aligned better with the ground truth. MeViT's capability to capture and understand the underlying patterns and features in the data makes it a superior model for handling the complexities and variations present in the dataset.

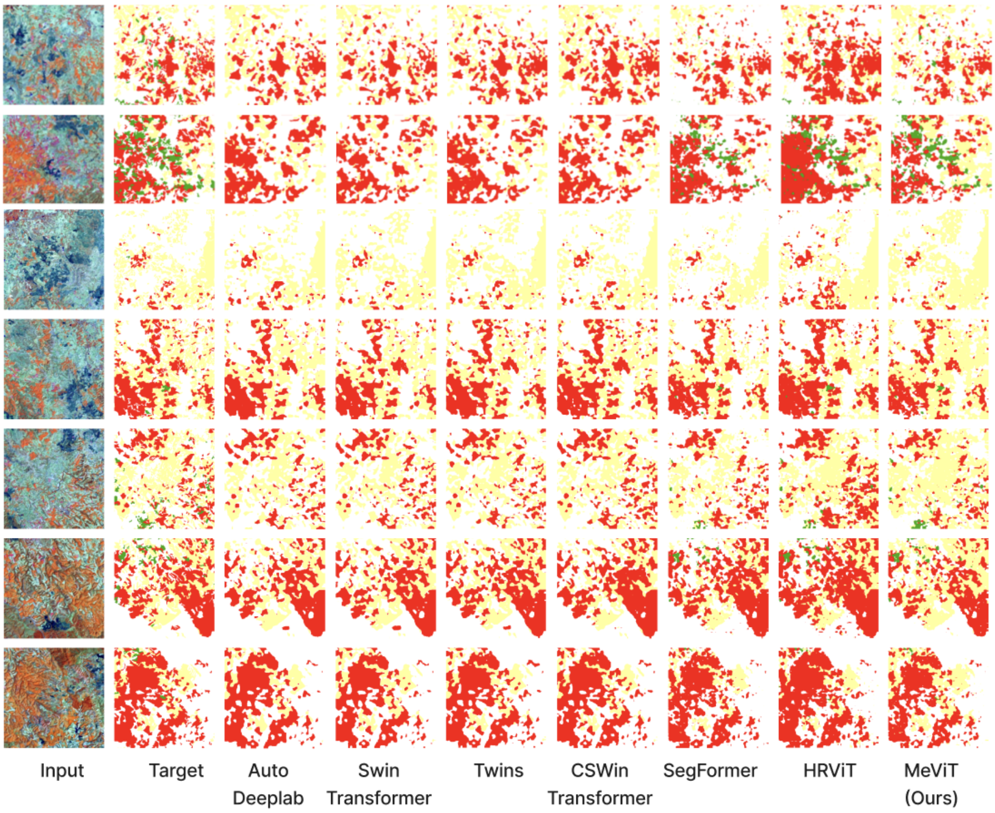


Figure 10. We compare the effectiveness of our proposed MeViT with modern transformer models, emphasizing the accurate prediction of the rubber and maize classes (red and yellow area, respectively), where our model outperforms the baselines. Red: para rubber, yellow: corn, green: pineapple.

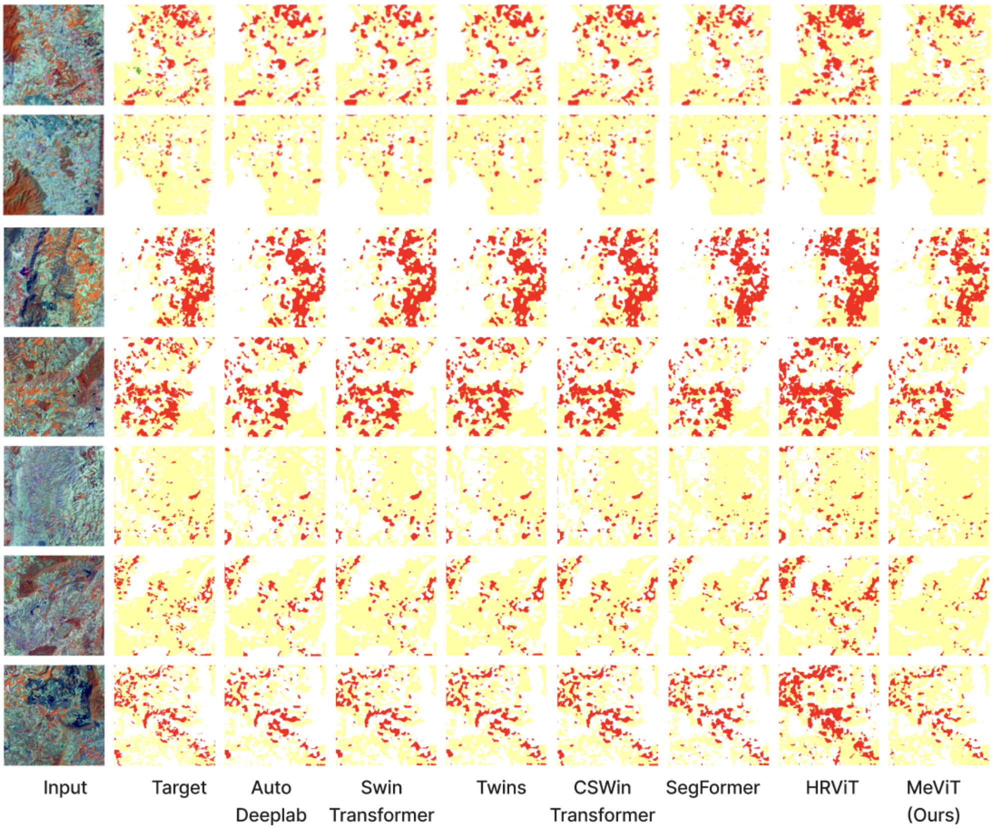


Figure 11. Our proposed MeViT model (rightmost column) is compared to modern transformer models to determine its effectiveness. We aim to showcase instances where our model successfully identifies rare categories, such as pineapples (green area). Red: para rubber, yellow: corn, green: pineapple.

The comparisons were made for different classes or categories, and MeViT consistently achieved higher accuracy and better prediction results across all types compared to the baseline models. The improved predictions by MeViT highlight its ability to capture and utilize relevant information to make accurate class assignments. This implies MeViT's effectiveness in recognizing and classifying the different objects or categories present in the dataset.

Overall, the comparison results presented in both figures provide strong evidence of MeViT's superiority over the baseline models regarding prediction accuracy and precision. These findings indicate that MeViT is a robust and reliable model for prediction tasks in computer vision, as its architecture and design enable it to effectively leverage spatial and contextual information, leading to improved prediction results. The superior performance of MeViT across different input samples and classes underscores its effectiveness in various practical applications.

6. Conclusions

In this paper, we proposed a novel deep learning method, MeVit, to perform semantic segmentation on Landsat satellite imagery for Thailand's main economic crops, such as para rubber, corn, and pineapple. Our proposed MeViT enhances vision transformers (ViTs) to learn semantically rich and spatially precise multi-scale representations by integrating medium-resolution multi-branch architectures with ViTs. We balanced the model performance and efficiency of MeViT by revising mixed-scale convolutional feedforward networks (MixCFN) with multiple depth-wise convolution paths to extract multi-scale local information.

We evaluated the effectiveness of our proposed MeViT on the publicly available dataset of Thailand scenes. We compared the results with several state-of-the-art deep learning methods such as AutoDeeplab, SwinTransformer, Twins, CSWinTransformer, SegFormer, and HRViT. Among the models compared, MeViT achieved the best performance in all evaluation metrics, including precision, recall, F1 score, and mean intersection over union (IoU). The experimental results demonstrated that our proposed MeViT outperformed existing methods and performed better in the semantic segmentation of Thailand scenes.

Eventually, our proposed MeViT approach provides a novel solution for the accurate semantic segmentation of Landsat satellite imagery for the main economic crops in Thailand. The experimental results show that our proposed method outperforms existing state-of-the-art deep learning methods and achieves the best performance in all evaluation metrics. This work contributes to remote sensing image analysis and provides a valuable tool for proper land use and land cover classification, which has significant implications for agriculture and environmental management.

As a future direction, we intend to assess MeViT on tasks that require dense prediction remote sensing, such as panoptic segmentation or crop yield forecasting. This will effectively showcase the capabilities of MeViT as a robust transformer backbone.

Author Contributions: T.P.; formal analysis, T.P.; investigation, T.P.; methodology, T.P.; project administration, T.P.; resources, T.P.; software, T.P.; supervision, T.P.; validation, T.P., C.C. and C.S.; visualization, T.P.; writing—original draft, T.P.; writing—review and editing, T.P.; funding acquisition, C.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Ratchadapisek Somphot Fund for Postdoctoral Fellowship, Chulalongkorn University.

Data Availability Statement: If you would like to obtain the data that were used in this study, please contact the corresponding author.

Acknowledgments: This research was supported by the Ratchadapisek Somphot Fund for Postdoctoral Fellowship, Chulalongkorn University. Teerapong Panboonyuen, also known as Kao Panboonyuen, appreciates (thanks) and acknowledges the scholarship from Ratchadapisek Somphot Fund for Postdoctoral Fellowship, Chulalongkorn University, Thailand. This research was supported by the NASA Land Cover Land Use Change Grant. It was also funded by the CGIAR Global Rice Science Partnership (GRiSP) program. We greatly acknowledge the Geo-informatics and Space Technology Development Agency (GISTDA), Thailand, for providing satellite imagery used in this study.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

CNNs	Convolutional Neural Networks
HeViT	High-Resolution Vision Transformer
LULC	Land Use and Land Cover
MeViT	Medium-Resolution Vision Transformer
ViT	Vision Transformer

References

1. Scheibenreif, L.; Hanna, J.; Mommert, M.; Borth, D. Self-supervised vision transformers for land-cover segmentation and classification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–20 June 2022; pp. 1422–1431.
2. Pang, K.; Weng, L.; Zhang, Y.; Liu, J.; Lin, H.; Xia, M. SGBNet: An ultra light-weight network for real-time semantic segmentation of land cover. *Int. J. Remote Sens.* **2022**, *43*, 5917–5939. [CrossRef]
3. Chen, J.; Sun, B.; Wang, L.; Fang, B.; Chang, Y.; Li, Y.; Zhang, J.; Lyu, X.; Chen, G. Semi-supervised semantic segmentation framework with pseudo supervisions for land-use/land-cover mapping in coastal areas. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *112*, 102881. [CrossRef]
4. Pasquarella, V.J.; Arévalo, P.; Bratley, K.H.; Bullock, E.L.; Gorelick, N.; Yang, Z.; Kennedy, R.E. Demystifying LandTrendr and CCDC temporal segmentation. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *110*, 102806. [CrossRef]
5. Toker, A.; Kondmann, L.; Weber, M.; Eisenberger, M.; Camero, A.; Hu, J.; Hoderlein, A.P.; Şenaras, Ç.; Davis, T.; Cremers, D.; et al. Dynamicearthnet: Daily multi-spectral satellite dataset for semantic change segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 21158–21167.

6. Zhang, M.; Singh, H.; Chok, L.; Chunara, R. Segmenting across places: The need for fair transfer learning with satellite imagery. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 2916–2925.
7. Ettehadi Osgouei, P.; Sertel, E.; Kabadayı, M.E. Integrated usage of historical geospatial data and modern satellite images reveal long-term land use/cover changes in Bursa/Turkey, 1858–2020. *Sci. Rep.* **2022**, *12*, 9077. [CrossRef] [PubMed]
8. Chaves, M.E.; Soares, A.R.; Mataveli, G.A.; Sánchez, A.H.; Sanches, I.D. A Semi-Automated Workflow for LULC Mapping via Sentinel-2 Data Cubes and Spectral Indices. *Automation* **2023**, *4*, 94–109. [CrossRef]
9. Duarte, D.; Fonte, C.; Patriarca, J.; Jesus, I. Geographical Transferability of Lulc Image-Based Segmentation Models Using Training Data Automatically Generated from Openstreetmap–Case Study in Portugal. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* **2022**, *3*, 25–31. [CrossRef]
10. Li, Z.; Li, E.; Samat, A.; Xu, T.; Liu, W.; Zhu, Y. An Object-Oriented CNN Model Based on Improved Superpixel Segmentation for High-Resolution Remote Sensing Image Classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 4782–4796. [CrossRef]
11. Desai, S.; Ghose, D. Active Learning for Improved Semi-Supervised Semantic Segmentation in Satellite Images. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–8 January 2022; pp. 553–563.
12. Wang, Z.; Yang, P.; Liang, H.; Zheng, C.; Yin, J.; Tian, Y.; Cui, W. Semantic segmentation and analysis on sensitive parameters of forest fire smoke using smoke-unet and landsat-8 imagery. *Remote Sens.* **2022**, *14*, 45. [CrossRef]
13. Chen, X.; Pan, S.; Chong, Y. Unsupervised domain adaptation for remote sensing image semantic segmentation using region and category adaptive domain discriminator. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 4412913. [CrossRef]
14. Ma, X.; Zhang, X.; Wang, Z.; Pun, M.O. Unsupervised domain adaptation augmented by mutually boosted attention for semantic segmentation of VHR remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5400515. [CrossRef]
15. Wu, L.; Lu, M.; Fang, L. Deep covariance alignment for domain adaptive remote sensing image segmentation. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5620811. [CrossRef]
16. Diakogiannis, F.I.; Waldner, F.; Caccetta, P.; Wu, C. ResUNet-a: A deep learning framework for semantic segmentation of remotely sensed data. *ISPRS J. Photogramm. Remote Sens.* **2020**, *162*, 94–114. [CrossRef]
17. Yuan, K.; Zhuang, X.; Schaefer, G.; Feng, J.; Guan, L.; Fang, H. Deep-learning-based multispectral satellite image segmentation for water body detection. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 7422–7434. [CrossRef]
18. Chen, Z.; Li, D.; Fan, W.; Guan, H.; Wang, C.; Li, J. Self-attention in reconstruction bias U-Net for semantic segmentation of building rooftops in optical remote sensing images. *Remote Sens.* **2021**, *13*, 2524. [CrossRef]
19. Li, P.; Zhang, H.; Guo, Z.; Lyu, S.; Chen, J.; Li, W.; Song, X.; Shibasaki, R.; Yan, J. Understanding rooftop PV panel semantic segmentation of satellite and aerial images for better using machine learning. *Adv. Appl. Energy* **2021**, *4*, 100057. [CrossRef]
20. Yeung, H.W.F.; Zhou, M.; Chung, Y.Y.; Moule, G.; Thompson, W.; Ouyang, W.; Cai, W.; Bennamoun, M. Deep-learning-based solution for data deficient satellite image segmentation. *Expert Syst. Appl.* **2022**, *191*, 116210. [CrossRef]
21. Wurm, M.; Stark, T.; Zhu, X.X.; Weigand, M.; Taubenböck, H. Semantic segmentation of slums in satellite images using transfer learning on fully convolutional neural networks. *ISPRS J. Photogramm. Remote Sens.* **2019**, *150*, 59–69. [CrossRef]
22. Zhang, W.; Huang, Z.; Luo, G.; Chen, T.; Wang, X.; Liu, W.; Yu, G.; Shen, C. TopFormer: Token pyramid transformer for mobile semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 12083–12093.
23. Gu, J.; Kwon, H.; Wang, D.; Ye, W.; Li, M.; Chen, Y.H.; Lai, L.; Chandra, V.; Pan, D.Z. Multi-scale high-resolution vision transformer for semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 12094–12103.
24. Xu, L.; Ouyang, W.; Bennamoun, M.; Boussaid, F.; Xu, D. Multi-class token transformer for weakly supervised semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 4310–4319.
25. He, X.; Zhou, Y.; Zhao, J.; Zhang, D.; Yao, R.; Xue, Y. Swin transformer embedding UNet for remote sensing image semantic segmentation. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 4408715. [CrossRef]
26. Wang, L.; Li, R.; Zhang, C.; Fang, S.; Duan, C.; Meng, X.; Atkinson, P.M. UNetFormer: A UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery. *ISPRS J. Photogramm. Remote Sens.* **2022**, *190*, 196–214. [CrossRef]
27. Zhang, B.; Tian, Z.; Tang, Q.; Chu, X.; Wei, X.; Shen, C. Segvit: Semantic segmentation with plain vision transformers. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 4971–4982.
28. Zhou, B.; Krähenbühl, P. Cross-view transformers for real-time map-view semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 13760–13769.
29. Ru, L.; Zhan, Y.; Yu, B.; Du, B. Learning affinity from attention: End-to-end weakly-supervised semantic segmentation with transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 16846–16855.
30. Zhang, J.; Yang, K.; Ma, C.; Reiß, S.; Peng, K.; Stiefelhagen, R. Bending reality: Distortion-aware transformers for adapting to panoramic semantic segmentation. In Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 16917–16927.

31. Lazarow, J.; Xu, W.; Tu, Z. Instance segmentation with mask-supervised polygonal boundary transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 4382–4391.
32. Zhang, C.; Jiang, W.; Zhang, Y.; Wang, W.; Zhao, Q.; Wang, C. Transformer and CNN hybrid deep neural network for semantic segmentation of very-high-resolution remote sensing imagery. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 4408820. [CrossRef]
33. Qiu, C.; Li, H.; Guo, W.; Chen, X.; Yu, A.; Tong, X.; Schmitt, M. Transferring transformer-based models for cross-area building extraction from remote sensing images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 4104–4116. [CrossRef]
34. Liu, C.; Chen, L.C.; Schroff, F.; Adam, H.; Hua, W.; Yuille, A.L.; Fei-Fei, L. Auto-deeplab: Hierarchical neural architecture search for semantic image segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 82–92.
35. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 10012–10022.
36. Panboonyuen, T.; Jitkajornwanich, K.; Lawawirojwong, S.; Srestasathiern, P.; Vateekul, P. Transformer-based decoder designs for semantic segmentation on remotely sensed images. *Remote Sens.* **2021**, *13*, 5100. [CrossRef]
37. Chu, X.; Tian, Z.; Wang, Y.; Zhang, B.; Ren, H.; Wei, X.; Xia, H.; Shen, C. Twins: Revisiting the design of spatial attention in vision transformers. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 9355–9366.
38. Dong, X.; Bao, J.; Chen, D.; Zhang, W.; Yu, N.; Yuan, L.; Chen, D.; Guo, B. Cswin transformer: A general vision transformer backbone with cross-shaped windows. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022, pp. 12124–12134.
39. Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J.M.; Luo, P. SegFormer: Simple and efficient design for semantic segmentation with transformers. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 12077–12090.
40. Roy, D.P.; Wulder, M.A.; Loveland, T.R.; Woodcock, C.E.; Allen, R.G.; Anderson, M.C.; Helder, D.; Irons, J.R.; Johnson, D.M.; Kennedy, R.; et al. Landsat-8: Science and product vision for terrestrial global change research. *Remote Sens. Environ.* **2014**, *145*, 154–172. [CrossRef]
41. Knight, E.J.; Kvaran, G. Landsat-8 operational land imager design, characterization and performance. *Remote Sens.* **2014**, *6*, 10286–10305. [CrossRef]
42. Loveland, T.R.; Irons, J.R. Landsat 8: The plans, the reality, and the legacy. *Remote Sens. Environ.* **2016**, *185*, 1–6. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.