remote sensing

MDPI

Article

Hierarchical Transfer Learning with Transformers to Improve Semantic Segmentation in Remote Sensing Land Use

Miaomiao Chen [1,2,†] and Lianfa Li [1,2,*,†]

[1] State Key Laboratory of Resources and Environmental Information Systems, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Datun Road, Beijing 100101, China; chenmiaomiao22@mails.ucas.ac.cn
[2] University of the Chinese Academy of Sciences, Beijing 100049, China
* Correspondence: lilf@lreis.ac.cn; Tel.: +86-10-64888362
† These authors contributed equally to this work.

Abstract: Land use classification remains a significant challenge in remote sensing semantic segmentation. While convolutional neural networks (CNNs) are widely used, their inherent limitations, such as restricted receptive fields, hinder their widespread application in remote sensing. Additionally, the scarcity of labeled remote sensing data and domain shift issues adversely impact deep learning model performance. This study proposes a hierarchical transfer learning framework for fine-category semantic segmentation tasks, leveraging the powerful global relationship modeling capabilities of Transformer models to classify land use in Dongpo District, Meishan City, in mainland China. Our framework represents multilevel transfer learning, progressing from non-remote sensing classification to coarse classification, then to the refined classification of remote sensing. We compared the performance of Transformer models with representative baseline CNNs like U-Net and DeepLab V3+. Results show that the Swin-Unet model outperforms the other models used in this study. It achieved the highest test mean intersection over union (MIoU) of 0.837 and 0.810 for residential and transportation in level 1 (coarse) classification, respectively, and 0.545 for irrigated land in level 2 (fine-grained) classification. Transfer learning from pre-trained models significantly enhanced semantic segmentation accuracy compared to random parameter initialization (ranging from 0.4% to 17.7%), with up to a 17.7% improvement in test MIoU for the public land category. The hierarchical transfer learning framework further improved segmentation accuracy for corresponding level 2 categories, leveraging pre-trained level 1 models. Our study shows the applicability of Transformer-based transfer learning in remote sensing land use classification.

Keywords: land use; semantic segmentation; transformer; hierarchical transfer learning

check for updates

Academic Editor: Salah Bourennane

Received: 2 December 2024
Revised: 9 January 2025
Accepted: 10 January 2025
Published: 15 January 2025

Citation: Chen, M.; Li, L. Hierarchical Transfer Learning with Transformers to Improve Semantic Segmentation in Remote Sensing Land Use. *Remote Sens.* **2025**, *17*, 290. https://doi.org/10.3390/rs17020290

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Land use patterns, reflecting both natural geography and human development, play a crucial role in urban planning [1], ecological evaluation [2], environmental monitoring [3,4] and agricultural production [5–7]. In recent decades, land use classification has emerged as a prominent research topic, with semantic segmentation technology becoming a pivotal approach for extracting information from remote sensing images. Semantic segmentation, or pixel-level classification, aims to recognize the semantic category of each pixel in an image. The remarkable achievements of deep learning in computer vision, particularly convolutional neural networks (CNNs) and encoder-decoder structures, have propelled

advancements in remote sensing image recognition [8]. Zhu et al. [9] introduced Generative Adversarial Networks (GANs) as a regularization technique to prevent overfitting in hyperspectral image classification. The U-Net model, a fully convolutional network, has gained significant attention due to its straightforward structure and robustness. Diakogiannis et al. [10] proposed an improved U-Net model integrating residual connections, dilated convolutions, pyramid pooling, and multi-task learning, while refining the Dice loss function to address sample imbalance. Furthermore, Ye et al. [11] combined spatial-spectral attention mechanisms with U-Net to improve feature expression consistency across convolutional layers, achieving high-precision building extraction. These advancements have significantly enhanced the accuracy and applicability of semantic segmentation in land use classification, addressing challenges such as overfitting, sample imbalance, and feature consistency. DeepLab V3+ adds a decoder module to DeepLabV3 to refine the boundary details, utilizing the deep separable convolution to reduce the computational complexity. This model has been extensively employed in the field of remote sensing semantic segmentation. Ren et al. [12] improved the DeepLab V3+ model to leverage remote sensing spectral information and spatial information, thereby enhancing the semantic segmentation accuracy of remote sensing datasets.

In summary, CNNs demonstrate certain advantages over classical machine learning in spatial location representation. However, modeling global semantic interactions and contextual information remains challenging due to the localized nature of convolutional operations [13]. The advent of Transformers has opened new avenues for research in global relationship modeling, applying a self-attention mechanism to assess the influence of disparate elements within the input data on the ultimate prediction outcome. This approach can capture intricate relationships and global dependencies within the data [14]. Chen et al. [15] designed an efficient dual-path Transformer structure capable of learning long-term dependencies of tokens in spatial and channel dimensions, achieving state-of-the-art performance on a benchmark building extraction dataset. Additionally, a coupled CNN-Transformer framework (CCTNet) was proposed, aiming to combine local details captured by CNN (e.g., edges and textures) with global contextual information obtained via Transformer for crop segmentation based on remotely sensed images [16]. Notably, Transformers have quadratic time complexity, necessitating longer training periods compared to CNNs, which have only linear time complexity [17]. Furthermore, while the incorporation of position embedding into Transformer models mitigates the inductive bias present in CNNs, it introduces an increased number of trainable parameters and a larger memory footprint, raising hardware requirements and demanding a greater volume of training data [14]. Currently, Transformers are prevalent in the field of computer vision, yet their applications in semantic segmentation for remote sensing remain relatively limited.

Deep learning has the capacity to adaptively extract the most relevant features for a given classification task, offering advantages such as robustness and ease of model transfer. However, these models typically require a substantial quantity of labeled data for training [18,19], the acquisition of which can be costly in remote sensing, especially for agricultural and forestry research involving labor-intensive field activities and destructive sampling. Furthermore, factors such as data collection, the geospatial environment, and the climate inevitably impact data quality [20]. In the absence of sufficient labeled data, models may exhibit a tendency to memorize the data instead of generalizing from it, leading to overfitting issues [21]. Moreover, models trained with labeled data from one domain often exhibit poor performance when directly applied to other domains [22], limiting the generalizability and transferability of deep learning models. Transfer learning techniques can assist in addressing these challenges by transferring knowledge acquired in the source domain to the target domain to accomplish relevant tasks [23]. However, the application

of transfer learning techniques in the field of land-use classification remains limited, with most studies focusing on convolutional networks rather than Transformer architectures. For example, Lin et al. proposed a method that can effectively transfer knowledge about the topology (i.e., relative position) of different crops in the spectral feature space to generate labels, thus supporting the classification of crops of different vintages [24]. Chew et al. collected 3 cm-resolution UAV images and fine-tuned the pre-trained VGG16 architecture to accomplish the classification task of field crops in Rwanda [25]. A literature review reveals that most existing studies on the semantic segmentation of remote sensing land use based on transfer learning are confined to a single classification level, with few conducting systematic investigations across diverse categories. Moreover, the majority of these studies often target specific land-use types or even individual crop types.

Traditional transfer learning methods often focus on knowledge transfer between tasks of similar granularity, overlooking the hierarchical relationships between the coarse-grained and fine-grained levels. However, many real-world applications involve multi-level hierarchical structures, where high-level categories share commonalities with their fine-grained subcategories. By leveraging these hierarchical relationships, hierarchical transfer learning enables the progressive refinement of knowledge across levels, enhancing model generalization and performance, particularly for fine-grained tasks. For instance, Min et al. [26] constructed encoders for different granularities and aligned them using subordinate entropy loss with a shared interpreter, effectively transferring knowledge learned from basic-level object recognition tasks to subordinate-level tasks, thereby improving recognition accuracy for fine-grained categories. Similarly, Luo et al. [27] mimicked the human process of learning new languages by training the model in the unrelated high-resource language pair, the similar intermediate language pair, and the low-resource language pair in turn. Through incremental transfer and the fine-tuning of model parameters, they improved the performance of neural machine translation for low-resource languages. However, most studies on hierarchical transfer learning have primarily concentrated on fields such as natural language processing, computer vision, and audio recognition [28], with limited attention to its application in remote sensing land-use classification. Leveraging transfer learning to apply the knowledge learned from coarse categories to fine-grained subcategories offers a promising approach to advancing fine-grained land-use classification.

This paper explores the potential of the Transformer models for semantic segmentation of remote sensing land use. Figure 1 shows a summary framework of our novel hierarchical Transformer-based transfer learning. We selected two representative Transformer architectures for transfer learning due to their robust performance in computer vision tasks: Swin-Unet and SegFormer. Swin-Unet, a pure Transformer model, features a symmetric encoder-decoder architecture akin to U-Net, achieving upsampling and an increase in feature dimensionality without convolution or interpolation operations, demonstrating excellent performance. SegFormer, in contrast, combines a hierarchical Transformer encoder with a lightweight all-MLP decoder, presenting a more complex architecture. To highlight the advantages of these Transformer models, we chose two widely-used and high-performing convolutional networks, U-Net and DeepLab V3+ (Figures A1 and A2), as baseline architectures for experimental comparison. We constructed a hierarchical transfer learning framework by enhancing and integrating full fine-tuning, partial fine-tuning, and low-rank adaptation (LoRA) strategies to examine the semantic segmentation effects of the selected models across two levels of land use categories. For level 1 coarse categories, we conducted adaptive training and prediction based on the original pre-trained models and various transfer learning strategies. For level 2 fine categories, we not only utilized the original pre-trained models but also employed the most effective models trained for the

level 1 categories as pre-trained models for the corresponding level 2 categories, enabling re-transfer application.

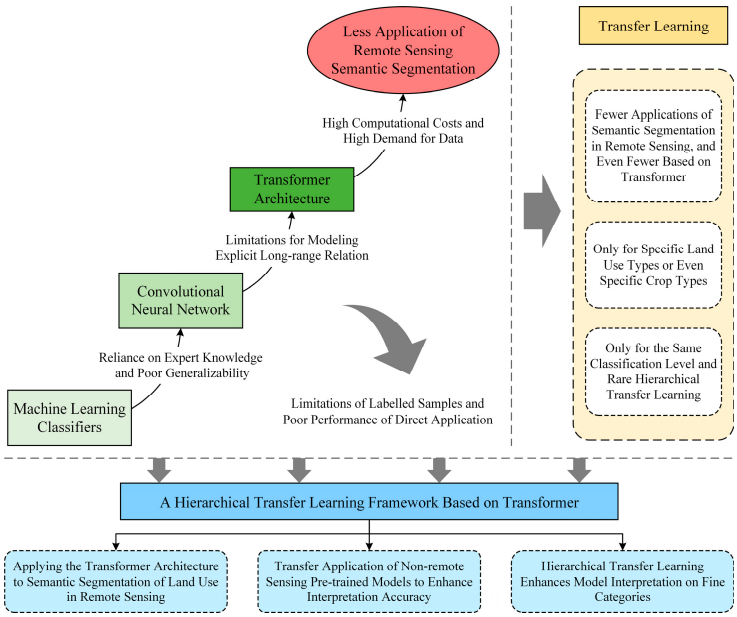


Figure 1. An overview of the research idea of this paper. Following the development of deep learning techniques, we analyzed the limitations of various semantic segmentation methods and identified the common shortcomings of deep learning approaches. Transfer learning has emerged as an effective solution to address these challenges. Through a comprehensive literature review, we summarized the current research status and gaps in the application of transfer learning in remote sensing semantic segmentation. Based on these findings, we proposed a hierarchical transfer learning framework based on Transformer and outlined the key innovations of this framework.

To comprehensively and intuitively evaluate and compare the semantic segmentation performance of these models, we employed four metrics: pixel accuracy (Acc), mean intersection over union (MIoU), Dice coefficient, and F1-score. Additionally, we visualized model prediction results on images of varying scales. We effectively mitigated the gradient explosion phenomenon during training through gradient clipping and adjusted the learning rate to further optimize the model training process and improve the model performance. In summary, this paper makes three contributions:

(1) We apply Transformer architectures, specifically Swin-Unet and SegFormer models, to remote sensing semantic segmentation, systematically verifying their applicability and superiority.
(2) We propose an effective hierarchical transfer learning framework based on Transformers, demonstrating the feasibility of transferring semantic segmentation pre-trained models from non-remote sensing domains to coarse-to-fine remote sensing image tasks, thereby enhancing the model generalization ability on fine categories.
(3) We conduct a comprehensive comparison between Transformer models and baseline architectures on semantic segmentation tasks across two levels of land use categories. Our experimental results reveal the markedly superior performance of the Transformer models, with transfer applications of pre-trained models further enhancing classification effectiveness.

2. Study Area and Datasets

Meishan City is a prefecture-level city under the jurisdiction of Sichuan Province, mainland China, located in the southwest region of the Chengdu Plain. Its topography

is characterized by a west-to-east elevation gradient, with higher elevations in the west and lower elevations in the east, as well as a south-to-north elevation gradient, with higher elevations in the south and lower elevations in the north. The city is characterized by long mountains, rolling hills and a dense river network. Dongpo District is subordinate to Meishan City (Figure 2) and serves as the political, economic and cultural center of Meishan. The terrain for the district is high in the northwest region and relatively low in the southeast corner, with plains, shallow hills and mountains each accounting for one-third. The rivers in the district are dendritically distributed, with forest area coverage exceeding 40%. This study conducts an experimental study on Dongpo District of Meishan City. The land use categories under investigation encompass nine primary categories: farmland, woodlands, residential, transportation, water, garden fields, industrial, commercial, and public land. The specific sub-category composition is shown in Table A1. The remote sensing image data used in this study are mainly Jilin-1 satellite images with a spatial resolution of 1 m, which is conducive to fine-scale land use classification tasks. The vector annotation dataset for ground feature type was obtained via land surveys. This dataset has undergone strict internal interpretation, quality verification, and field verification, boasting high accuracy, comprehensive ground coverage, and a complete classification system. We also checked these label data in practice and removed those with different label classes from the practical survey. The collection time of remote sensing images is from September to October 2022, and that of the annotation data is December 2022.

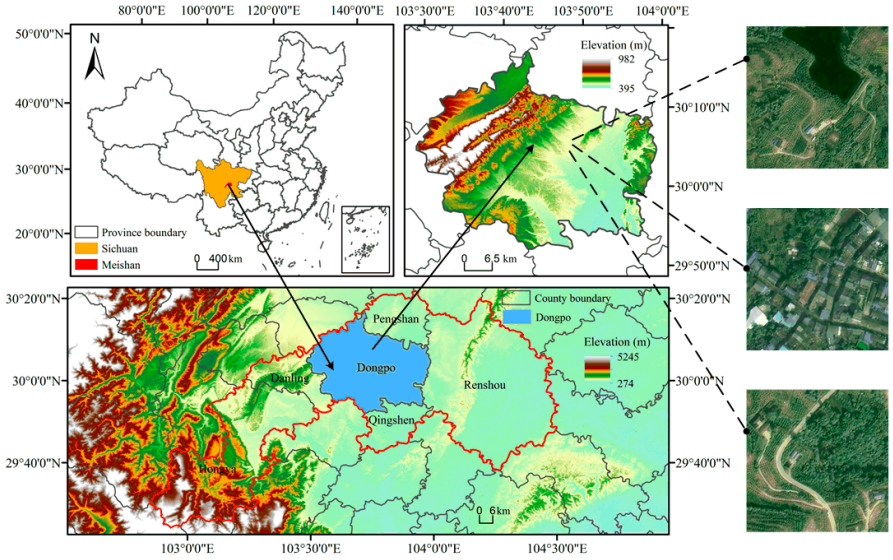


Figure 2. The location of the study area (Dongpo District, Meishan City, Sichuan Province, China).

3. Methodology

The hierarchical transfer learning framework is constructed as shown in the technical flow in Figure 3, which mainly consists of data processing and model training, evaluation, and prediction on level 1 and level 2 land use categories. In Section 3.1, we introduce the data processing process. In Section 3.2, we provide a detailed account of the Transformer models employed in this paper, namely the Swin-Unet and SegFormer models. In Section 3.3, we systematically describe our transfer learning methodology. In Section 3.4, we present the two approaches adopted in the study to optimize the training process, namely gradient clipping and adaptive learning rate. In Section 3.5, we list the evaluation metrics.

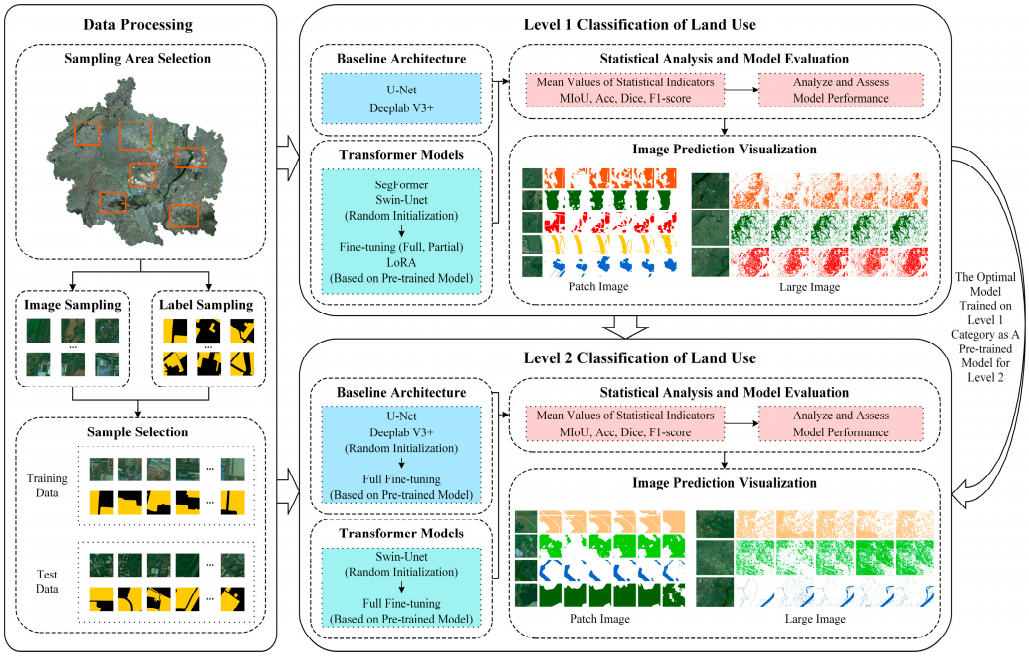


Figure 3. The technical flow of the hierarchical transfer learning framework proposed in this study.

3.1. Data Processing

Data processing mainly includes labelled data generation, sampling areas selection, patch sampling, and sample selection. Firstly, the labelled data sets of different land use categories in the study area were created based on the annotated data in vector format and matched with remote sensing image data. Notably, we manually checked the images to ensure that cloud contamination was not significant and had little effect on the main experimental results. Therefore, we did not analyze the remote sensing images for cloud contamination. Secondly, the sampling area was selected according to the land cover situation in Dongpo District to keep the sample categories as balanced as possible. Then, the remote sensing images and labelled data in the sampling areas were sliced into patches of 224 × 224 size. Finally, training and test samples were randomly selected, ensuring that the training and test data sets were not duplicated.

3.2. Transformer Models

In this study, the Swin-Unet framework and the SegFormer framework are employed as representative Transformer-based segment models.

The Swin-Unet model (Figure 4) is the first U-shaped network architecture based exclusively on Transformer, as proposed by Cao et al. in 2021 [29]. The architecture comprises an encoder, a bottleneck, a decoder, and skip connections, with the basic unit being the Swin Transformer block [30]. More specifically, the Swin-Unet model uses a hierarchical Swin Transformer with a shifted window and a patch merging layer as an encoder to extract contextual features, in which the Swin Transformer block is tasked with feature representation learning and the patch merging layer is responsible for downsampling and increasing the dimension. In addition, a Swin Transformer-based symmetric decoder with a patch expanding layer is designed to perform upsampling operations to restore the spatial resolution of the feature map. During the model prediction process, the image is segmented into image patches, each of which is regarded as a token and fed into the encoder to learn deep feature representation. Subsequently, the extracted contextual features are upsampled by the decoder with a patch expanding layer and fused with the multi-scale features from the encoder through skip connection to compensate for the loss of spatial information caused by downsampling, thereby restoring the resolution of the feature map to the input

image resolution. Finally, the pixel-level segmentation prediction results are output with a linear projection layer for these upsampled features.

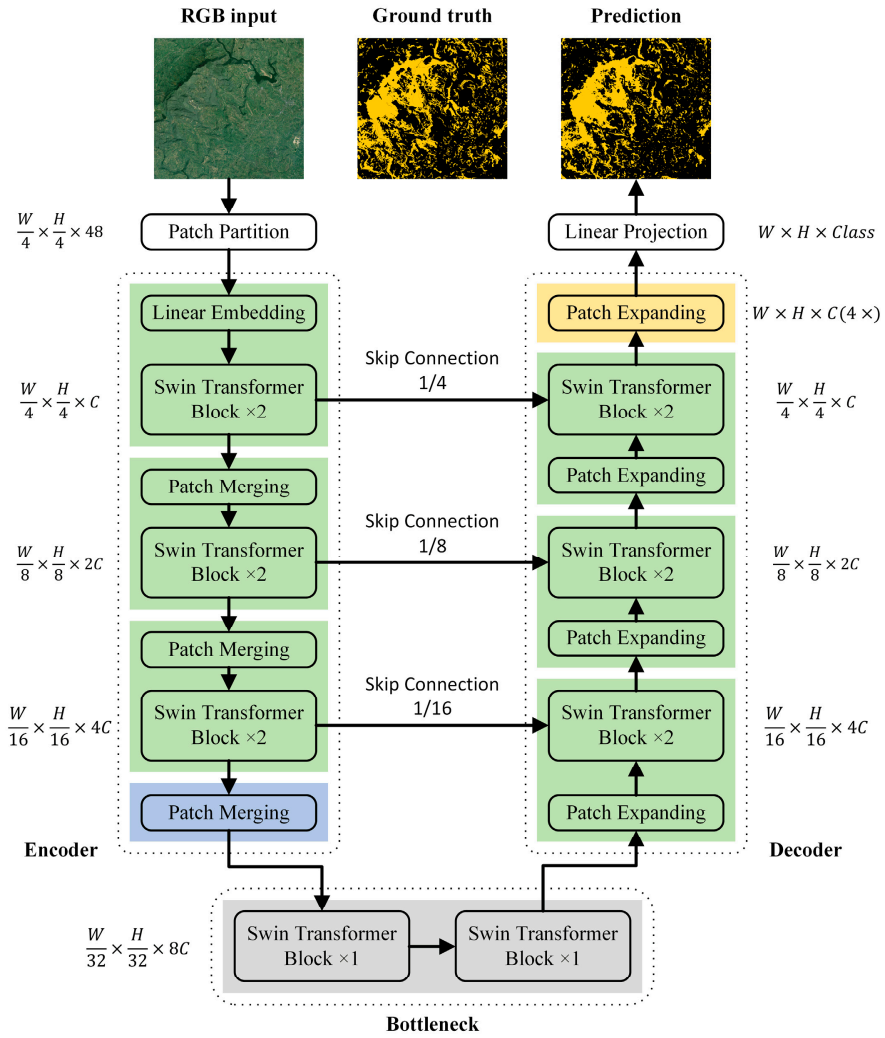


Figure 4. Swin-Unet model [29].

Unlike the purely Transformer-based Swin-Unet framework, the SegFormer framework (Figure 5), which is a combination of Transformer and a lightweight multi-layer perceptron (MLP) decoder, is a straightforward, efficient and robust semantic segmentation framework that has been applied to three publicly available semantic segmentation datasets and has reached the state-of-the-art level in terms of efficiency, accuracy, and robustness [31]. The SegFormer model builds a novel position-encoding-free and hierarchical Transformer encoder that can generate both high-resolution fine features and low-resolution coarse features, which is in sharp contrast to ViT [32], which can only generate low-resolution feature maps at a fixed resolution. Notably, the encoder of the SegFormer framework does not need to interpolate the positional encoding when reasoning about images with different resolutions from those utilized during training, so it can easily adapt to arbitrary test resolutions without affecting model performance. In addition, the lightweight MLP decoder proposed in the model is capable of aggregating information from different layers, thus combining local attention and global attention to generate robust representations.

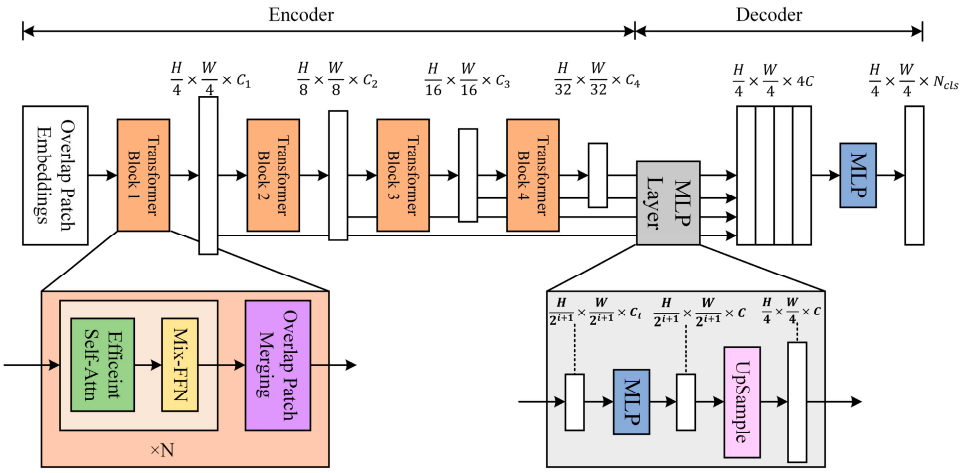


Figure 5. SegFormer model [31].

3.3. Transfer Learning

The transfer learning technique is a machine learning approach that enables the transfer of knowledge acquired in the source domain to the target domain, thereby facilitating the accomplishment of related tasks. It has emerged as a promising strategy that can address the domain transfer problem and reduce the reliance on labeled data. According to the different scenarios in which the transfer learning techniques are applied, they can be categorized into five groups: fine-tuning-based transfer learning, multi-task learning, few-shot learning, unsupervised domain adaptation, and self-supervised learning [22]. The optimal transfer learning method for a specific task depends on many factors, such as the availability of labeled samples, the relevance of the source and the target learning tasks, and the goal of the learning task. In this study, fine-tuning-based transfer learning (FTL) is employed for experimental investigation, which has the advantage of enabling pre-trained models to be transferred and applied to various domains, thereby solving all types of domain transfer problems. The specific strategies include full fine-tuning, partial fine-tuning, LoRA strategies, etc.

The basic strategy of transfer learning is full fine-tuning (Figure 6a), where the model is first pre-trained on the source dataset and then all model parameters are fine-tuned for the target task in the target domain dataset. Notably, it is crucial to pre-train the model with a large and diverse dataset in order for it to learn robust and generalizable representations. Therefore, the number of samples in the source domain is usually larger than that in the target domain. However, when there are insufficient labeled samples in the target domain, fine-tuning the entire neural network may lead to overfitting [33]. Furthermore, with the continuous development of pre-trained large models, the cost of fine-tuning and storing all parameters has become progressively more expensive. Fortunately, deep learning models have a layered architecture that can learn different features at varying layers. The initial layers tend to capture generic features, while the subsequent layers are more task-specific [34]. In addition, existing research results indicate that fine-tuning a small fraction of the parameters can effectively simulate the performance of the model while significantly reducing the computational and storage costs [35]. Therefore, specific layers of the model can be fine-tuned or re-trained to adapt to the target task [36]. This strategy is known as partial fine-tuning (Figure 6b). This study applies both full fine-tuning and partial fine-tuning of the decoder layers based on the pre-trained models.

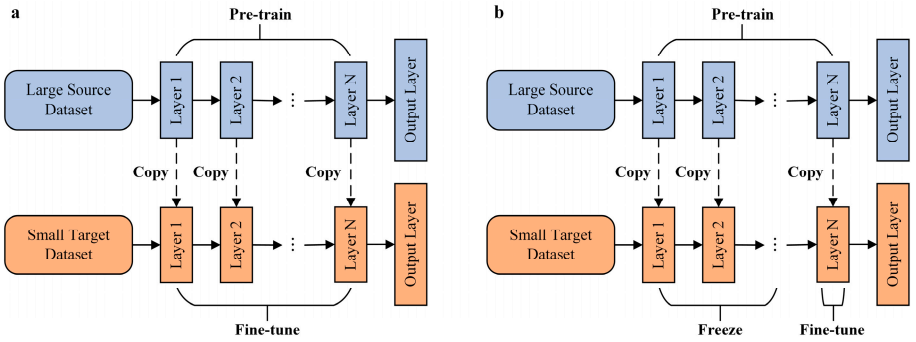


Figure 6. FTL technique flow: (**a**) full fine-tuning; (**b**) partial fine-tuning.

Full fine-tuning and partial fine-tuning both re-train the original parameters of the model, while the LoRA (Figure 7) proposed by Hu et al. [37] freezes the pre-trained weights and injects the re-trainable rank decomposition matrices into the Transformer layers. During fine-tuning, only the reparametrized portion—namely, the LoRA parameters—undergoes training. The central idea of this approach is to indirectly train specific dense layers in the neural network by optimizing the rank decomposition matrices of the dense layers' change during the process of model adaptation while simultaneously allowing the pre-trained weights to remain unchanged. The LoRA strategy can significantly reduce the number of training parameters for downstream tasks, storage requirements, and task costs. Furthermore, it can improve the feasibility of pre-trained large models for transfer applications, which has already achieved good results with RoBERTa, DeBERTa, GPT-2, and GPT-3 [37]. In this study, the LoRA strategy is employed to fine-tune the decoder layers and the entire Transformer model (the encoder and decoder layers), respectively. However, unlike Hu et al., we also train the model's original parameters while fine-tuning the LoRA parameters, as well as compare the results with those obtained from training only the LoRA parameters.

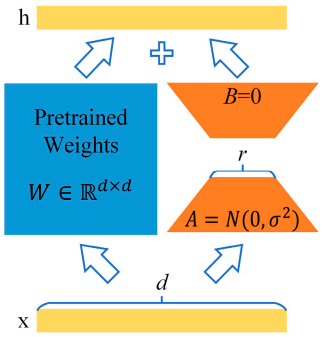


Figure 7. LoRA parameterization [37]. x denotes the input feature vector with a dimensionality of d. A represents a randomly Gaussian-initialized matrix with rank r, containing trainable parameters designed to adapt or augment the pretrained weights during fine-tuning. B is a matrix initialized to 0, also containing trainable parameters similar to A. h represents the output feature vector computed after the parameterized operations.

3.4. Gradient Clipping and Adaptive Learning Rate

Transformer-based pretrained models, compared to CNNs, involve more complex matrix operations, such as computing query, key, and value items in the multi-head attention mechanism. These operations include computationally intensive functions like softmax, which can potentially lead to gradient explosion and subsequently compromise model performance. To mitigate these challenges, we employed two key techniques: gradient clipping and adaptive learning rate.

Gradient clipping is a technique that sets a predefined threshold to prevent the excessive accumulation of gradients, which can cause instability during model training. This method has been widely adopted in natural language processing (NLP) to stabilize learning [38]. By constraining gradient values within a specified range, we can prevent the model from experiencing extreme updates that might derail the training process.

Simultaneously, we implemented an adaptive learning rate strategy. This approach dynamically adjusts the learning rate throughout the training process, typically reducing it from larger to smaller values. The learning rate is configured based on the total number of epochs and the current epoch index. Through multiple sensitivity analyses, we systematically selected the optimal gradient clip value and initial learning rate to enhance model performance and training stability.

Ultimately, we identified the gradient clipping value and initial learning rate to be both 0.01 for the Swin-Unet model and 0.75 and 0.004 for the SegFormer model, respectively. By carefully implementing gradient clipping and adaptive learning rate, we can effectively manage the computational complexities inherent in Transformer-based models and optimize their training dynamics in transfer learning for remote sensing.

3.5. Evaluation Metrics

For the purpose of evaluating the semantic segmentation accuracy of the models, we employed the Acc and MIoU indicators to evaluate the target classes, respectively.

(1) Acc is defined as the ratio of the number of correctly classified pixels to the total number of pixels:

$$Acc = \frac{\sum_{k=1}^{C} n_{kk}}{\sum_{k=1}^{C} t_k} \tag{1}$$

where C denotes the number of classes, n_{kk} is the total number of pixels correctly classified as class k, and t_k denotes the total number of pixels labeled as class k.

(2) The intersection over union (IoU)—also known as the Jaccard index (JI)—which is defined as the size of the intersection of sample sets divided by the size of the union, is a metric used to measure the degree of overlap between two sample sets. It is one of the most frequently utilized metrics for semantic segmentation evaluation. The MIoU is the average of the IoU values of all categories. This study uses MIoU as one of the evaluation indicators:

$$\mathrm{IoU}(c) = \frac{|\hat{\gamma} \cap \gamma|}{|\hat{\gamma} \cup \gamma|} = \frac{|\hat{\gamma} \cap \gamma|}{|\hat{\gamma}| + |\gamma| - |\hat{\gamma} \cap \gamma|} = \frac{n_{cc}}{\sum_{i=1}^{C} n_{ic} + \sum_{j=1}^{C} n_{cj} - n_{cc}}, \ \mathrm{MIoU} = \frac{1}{C}\sum_{i=1}^{C} IoU(i) \tag{2}$$

where γ is the set of ground truth masks, $\hat{\gamma}$ is the set of predicted results, n_{cc} is the total number of pixels correctly classified as target class c, n_{ic} is the total number of pixels whose ground truth is class i and is classified into target class c, n_{cj} is the total number of pixels whose ground truth is class c and is classified into target class i, and C denotes the number of classes.

(3) The Dice coefficient, also known as the Sørensen–Dice coefficient, is defined as twice the size of the intersection of the sample sets divided by the sum of the sample sizes of the sets. It is an important indicator for measuring the similarity of two sample sets and is also used to evaluate the semantic segmentation effect of the model:

$$Dice = \frac{2|\hat{\gamma} \cap \gamma|}{|\hat{\gamma}| + |\gamma|} = \frac{2n_{cc}}{\sum_{i=1}^{C} n_{ic} + \sum_{j=1}^{C} n_{cj} - 2n_{cc}} \tag{3}$$

where γ is the set of ground truth masks, $\hat{\gamma}$ is the set of predicted results, n_{cc} is the total number of pixels correctly classified as target class c, n_{ic} is the total number of

pixels whose ground truth is class i and is classified into target class c, n_{cj} is the total number of pixels whose ground truth is class c and is classified into target class i, and C denotes the number of classes.

(4) F1-score is the harmonic mean of precision and recall, providing a balance between the two:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (4)$$

where *Precision* is the proportion of correctly predicted positive samples among all predicted positive samples and *Recall* is the proportion of correctly predicted positive samples among all actual positive samples.

4. Results

4.1. Semantic Segmentation Results for Level 1 Classification

4.1.1. Evaluation Metrics of Model Testing for Level 1 Classification

The interpretation accuracies of the baseline models are presented in Tables 1 and A2. The DeepLab V3+ model achieves a higher Acc, F1-score, and Dice coefficient compared to the U-Net model, but with relatively lower MIoUs. This indicates that while the DeepLab V3+ model demonstrates greater overall prediction reliability, it may exhibit reduced boundary alignment in the semantic segmentation results. Conversely, the U-Net model achieves test MIoU values consistently above 0.660, with its best performance observed for the residential category, where the test MIoU reaches 0.798. Furthermore, the DeepLab V3+ model achieves test F1-scores exceeding 0.682 across all categories, with an outstanding F1-score of 0.909 for the industrial category.

Table 1. Test metrics of the DeepLab V3+ model and the U-Net model for level 1 classification.

Land Cover	DeepLab V3+ MIoU	U-Net MIoU	DeepLab V3+ Acc	U-Net Acc	DeepLab V3+ F1-Score	U-Net F1-Score	DeepLab V3+ Dice	U-Net Dice
Farmland	0.621	0.660	0.877	0.833	0.770	0.695	0.770	0.694
Woodlands	0.645	0.722	0.929	0.904	0.786	0.718	0.786	0.717
Residential	0.693	0.798	0.958	0.947	0.823	0.795	0.823	0.796
Transportation	0.589	0.776	0.970	0.968	0.748	0.743	0.748	0.743
Water	0.653	0.795	0.958	0.953	0.804	0.786	0.804	0.785
Garden Fields	0.746	0.687	0.874	0.816	0.856	0.797	0.856	0.794
Industrial	0.834	0.697	0.933	0.832	0.909	0.786	0.909	0.787
Commercial	0.501	0.663	0.916	0.887	0.682	0.635	0.682	0.633
Public Land	0.669	0.695	0.925	0.876	0.808	0.704	0.808	0.703

Tables 2 and A3 show the semantic segmentation accuracy of the Swin-Unet model for nine level 1 land use categories under different model fine-tuning strategies. The experimental results demonstrate that the semantic segmentation accuracies are consistently higher when applying the pre-trained model than when it is not applied. Moreover, the balance between precision and recall is significantly improved with the use of the pre-trained model. As an illustration, the use of the pre-trained model enhances the test MIoU of the Swin-Unet model from 0.742 to 0.782 for the farmland land use type, and from 0.756 to 0.810 for the transportation type. Notably, the utilization of the pre-trained model results in a remarkable 9.6% increase in the test MIoU, along with an impressive 13.7% improvement in the test F1-score. In comparison to full fine-tuning based on the pre-trained model, fine-tuning only the decoder layers of the model shows a diminished impact on enhancing the model interpretation accuracy. However, it still yields good interpretation results for some categories. For example, the test MIoU for the industrial category increases by 7.5%, the test Acc improves by 3.4%, and the test F1-score achieves a remarkable improvement of 12.2%. Surprisingly, the test MIoU for the water category under this strategy is consistent with the results when the parameters are randomly initialized. When compared to the results with the baseline architectures, it is evident that the Swin-Unet model outperforms both the DeepLab V3+ and U-Net models of all land use categories.

The overall performance of the Swin-Unet model under the LoRA strategy is comparatively inferior to that of the aforementioned transfer learning strategies. As demonstrated in Tables 3 and A4, the semantic segmentation performance of the Swin-Unet model is relatively high when the LoRA parameters are incorporated into the decoder layers and all model parameters, including the LoRA parameters, are updated for training. Under this configuration, the test MIoU for the residential category reaches 0.821, while the test F1-score reaches 0.834. However, the highest test MIoU for commercial land reaches only 0.578, which is 15.5% lower than the optimal result of 0.733 shown in Table 2. Furthermore, the model demonstrates superior performance when applying and fine-tuning the LoRA parameters only at the decoder layers during training, as compared to applying and fine-tuning the LoRA parameters at both the encoder and decoder layers. Although the interpretation accuracy of the Swin-Unet model under the LoRA strategy is reduced, it still achieves favorable interpretation results in comparison to random initialization. For instance, the test MIoU for the industrial category is enhanced from 0.756 to 0.795 and the test F1-score is improved from 0.833 to 0.863. Notably, the test F1-score for certain categories under the LoRA strategy falls below 0.5, indicating significant challenges in balancing precision and recall. This issue is likely a result of a class imbalance within the dataset.

Table 2. Test metrics of the Swin-Unet model under different fine-tuning strategies for level 1 classification.

Land Cover	Swin-Unet MIoU	Swin-Unet [1] MIoU	Swin-Unet [1,2] MIoU	Swin-Unet Acc	Swin-Unet [1] Acc	Swin-Unet [1,2] Acc	Swin-Unet F1-Score	Swin-Unet [1] F1-Score	Swin-Unet [1,2] F1-Score	Swin-Unet Dice	Swin-Unet [1] Dice	Swin-Unet [1,2] Dice
Farmland	0.742	0.782	0.714	0.878	0.899	0.862	0.793	0.837	0.776	0.790	0.835	0.775
Woodlands	0.761	0.780	0.714	0.920	0.928	0.898	0.775	0.803	0.727	0.773	0.801	0.726
Residential	0.792	0.837	0.774	0.948	0.960	0.939	0.794	0.852	0.781	0.792	0.851	0.782
Transportation	0.756	0.810	0.734	0.967	0.974	0.958	0.724	0.803	0.696	0.723	0.802	0.700
Water	0.785	0.808	0.785	0.950	0.957	0.950	0.776	0.813	0.787	0.775	0.812	0.787
Garden Fields	0.686	0.769	0.702	0.816	0.872	0.826	0.792	0.864	0.822	0.786	0.861	0.818
Industrial	0.756	0.816	0.812	0.869	0.904	0.903	0.833	0.880	0.887	0.833	0.879	0.888
Commercial	0.700	0.733	0.653	0.911	0.924	0.892	0.679	0.735	0.639	0.676	0.732	0.638
Public Land	0.663	0.759	0.738	0.865	0.910	0.899	0.656	0.793	0.778	0.652	0.789	0.774

[1] The pre-trained model is used for training. [2] Fine-tune only the decoder layer.

Table 3. Test metrics of Swin-Unet model under the LoRA strategy for level 1 classification.

Land Cover	Swin-Unet [1,3] MIoU	Swin-Unet [1] MIoU	Swin-Unet [2,3] MIoU	Swin-Unet [2] MIoU	Swin-Unet [1,3] Acc	Swin-Unet [1] Acc	Swin-Unet [2,3] Acc	Swin-Unet [2] Acc	Swin-Unet [1,3] F1-Score	Swin-Unet [1] F1-Score	Swin-Unet [2,3] F1-Score	Swin-Unet [2] F1-Score	Swin-Unet [1,3] Dice	Swin-Unet [1] Dice	Swin-Unet [2,3] Dice	Swin-Unet [2] Dice
Farmland	0.651	0.732	0.593	0.697	0.826	0.872	0.791	0.853	0.683	0.748	0.609	0.746	0.673	0.743	0.591	0.742
Woodlands	0.667	0.771	0.590	0.709	0.876	0.923	0.843	0.899	0.654	0.791	0.528	0.713	0.648	0.790	0.508	0.709
Residential	0.697	0.821	0.716	0.723	0.910	0.955	0.916	0.923	0.673	0.834	0.699	0.708	0.678	0.833	0.703	0.706
Transportation	0.667	0.795	0.468	0.739	0.945	0.972	0.936	0.962	0.583	0.784	0.377	0.700	0.591	0.784	0.376	0.699
Water	0.733	0.781	0.640	0.721	0.936	0.950	0.910	0.935	0.710	0.775	0.551	0.679	0.710	0.774	0.549	0.674
Garden Fields	0.640	0.717	0.612	0.716	0.782	0.838	0.762	0.836	0.753	0.814	0.735	0.824	0.736	0.808	0.724	0.819
Industrial	0.726	0.795	0.630	0.686	0.851	0.893	0.783	0.825	0.804	0.863	0.727	0.785	0.801	0.860	0.719	0.782
Commercial	0.539	0.578	0.492	0.526	0.837	0.857	0.827	0.828	0.415	0.506	0.289	0.388	0.406	0.496	0.233	0.373
Public Land	0.548	0.728	0.399	0.619	0.807	0.895	0.764	0.842	0.532	0.759	0.352	0.621	0.509	0.756	0.316	0.609

[1] Applying the LoRA parameters only at the decoder layer. [2] Applying the LoRA parameters at the encoder and decoder layers. [3] Fine-tuning only the LoRA parameters.

The interpretation effects of the SegFormer model under different transfer learning strategies are shown in Tables 4, 5, A5 and A6. Similar to the Swin-Unet model, the semantic segmentation accuracy of the SegFormer model for the majority of the land use categories is higher when utilizing the pre-trained model than when not utilizing the pre-trained model. For example, the pre-trained model results in an increase in the test MIoU for the transportation category from 0.677 to 0.723, and the test MIoU for the industrial category from 0.615 to 0.731. This finding provides further evidence of the rationality and effectiveness of pre-trained weights in improving the model's adaptability to diverse downstream tasks. When only the decoder layers are fine-tuned based on the pre-trained model, the overall performance of the SegFormer model is optimal, exhibiting superior test MIoUs and F1-scores, such as for the farmland category. For instance, the test MIoU for the water, garden fields, and commercial categories are 0.754, 0.639, and 0.590, respectively, while the corresponding test F1-scores are 0.742, 0.768, and 0.517. Notably, unlike the Swin-Unet model, the interpretation accuracy of the SegFormer model when fully fine-tuned based on the pre-trained model in some categories is lower than when only the decoder layers are fine-tuned, and even lower than when the model's parameters are randomly initialized. As an illustration, the test MIoU for the water land use type is enhanced by 2.7% when only the decoder layers are fine-tuned, and by 1.4% when the parameters are randomly initialized, in comparison to 0.727 when full fine-tuning is conducted based on the pre-trained model. Under the LoRA strategy, the SegFormer model exhibits suboptimal performance in semantic segmentation. However, it achieves a test MIoU of 0.753 and an F1-score of 0.736 for the residential category, representing the model's best performance for this specific class.

4.1.2. Visualization of Predicted Results for Level 1 Classification

Based on the evaluation results of the nine level 1 land use classes, it is evident that the overall performance of the Swin-Unet model in this semantic segmentation task sur-passes that of the SegFormer model. However, its interpretation accuracy under the LoRA strategy is relatively lower compared to other transfer learning strategies. Figure 8 presents the visualized prediction results of the baseline architectures and the Swin-Unet model under the non-LoRA strategies trained and saved by this study for various land use categories on large-sized images. It also includes the model's predicted MIoU metrics for this image. As illustrated, the interpretation integrity of the DeepLab V3+ is comparatively poor. For relatively fragmented patches of land objects or small details, the model struggles to accurately predict them. In contrast, the U-Net mode demonstrates superior semantic segmentation performance compared to the DeepLab V3+ model, effectively capturing the overall outline and some small details of the target category. However, the U-Net model is prone to misclassifications, often incorporating features that do not belong to the target category into its prediction results. For instance, regarding the farmland cate-gory, the DeepLabV3+ model captures only a small portion of the target features and fails to delineate the overall outline of the target category, which also exhibits poor details in the prediction results. The visual effect of semantic segmentation of the U-Net model is better than that of the DeepLabV3+ model, but there is an obvious misclassification. In contrast, the Swin-Unet model achieves the highest level of detail and contour integrity in semantic segmentation. For the industrial category, its predicted MIoU reaches 0.896, representing improvements of 11.9% and 9.0% over the DeepLab V3+ model and the U-Net model, respectively. In terms of visual effects, the Swin-Unet model, which is fully fine-tuned based on the pre-trained model, achieves significant outcomes. On the one hand, it predicts the overall contour of the industrial category with greater completeness. On the other hand, through its inner attention mechanism, it effectively preserves the detailed

information of the target features and accurately interprets the fragmented patches and their edge information. Notably, the visual semantic segmentation effect of the Swin-Unet model with randomly initialized parameters is poor, even lower than that of the DeepLab V3+ model. The application of pre-trained models and transfer strategies not only leads to an enhancement in the metric accuracy of the model prediction, but also significantly improves the completeness, level of detail, and accuracy of the visualizations generated by the models.

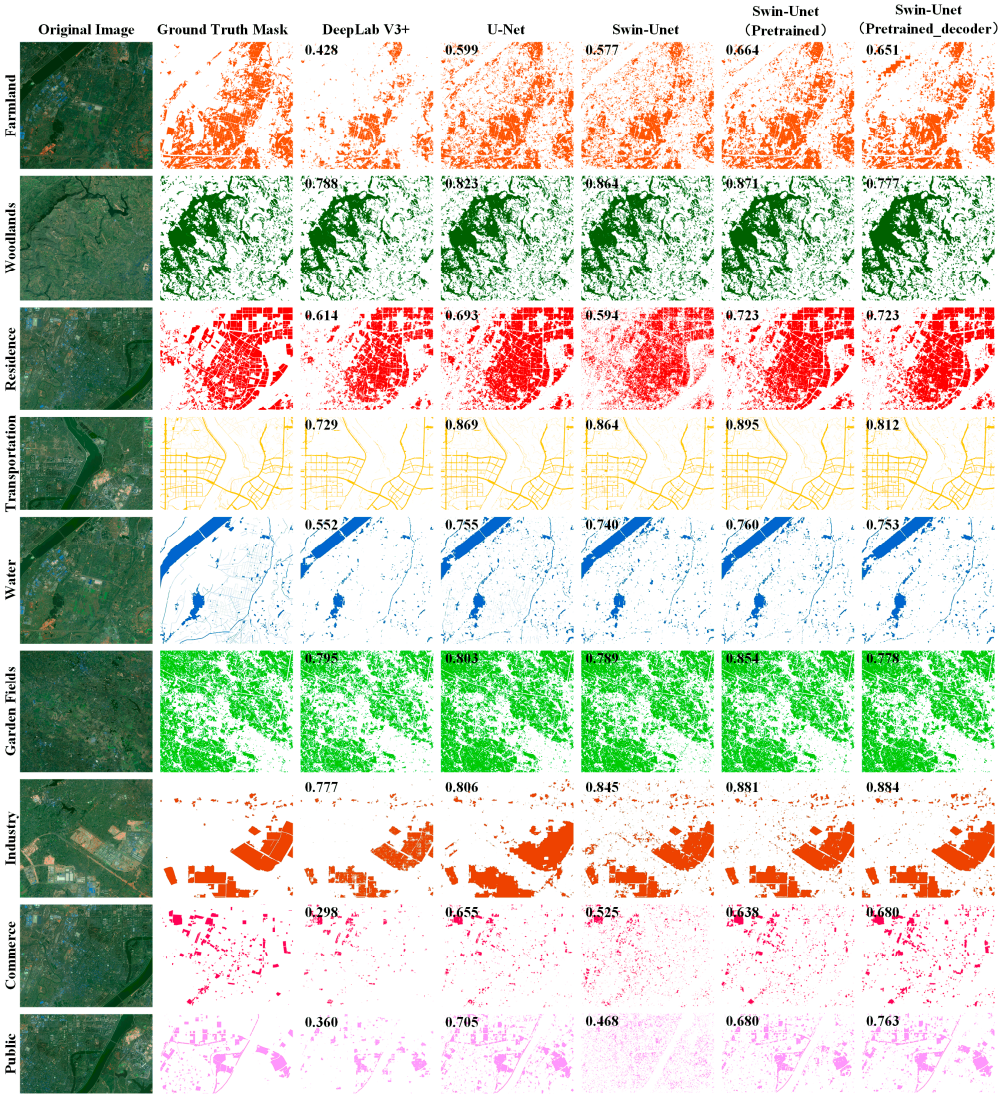


Figure 8. Visualization of model prediction results for nine level 1 land use classifications and their predicted MIoU metrics on large-scale images. "Pretrained" stands for fully fine-tuned based on the pre-trained model. "Pretrained_decoder" stands for fine-tuning the decoder layers based on the pre-trained model.

The baseline architectures and Transformer models with higher accuracies were selected for visual prediction analysis to compare the semantic segmentation effects of different models. Figure 9 illustrates that the Swin-Unet model exhibits the highest MIoU, while the accuracy of the SegFormer model is lower than that of the U-Net model, with a predicted MIoU of only 0.679. Furthermore, the SegFormer model delineates the boundaries of ground objects poorly and performs a low completeness in predicting the target content. In contrast, the Swin-Unet model can semantically segment the target objects more completely and accurately, which may be caused by the model's architectures.

Table 4. Test metrics of the SegFormer model under different fine-tuning strategies for level 1 classification.

Land Cover	SegFormer MIoU	SegFormer [1] MIoU	SegFormer [1,2] MIoU	SegFormer Acc	SegFormer [1] Acc	SegFormer [1,2] Acc	SegFormer F1-Score	SegFormer [1] F1-Score	SegFormer [1,2] F1-Score	SegFormer Dice	SegFormer [1] Dice	SegFormer [1,2] Dice
Farmland	0.633	0.631	0.653	0.811	0.815	0.829	0.667	0.661	0.699	0.661	0.651	0.690
Woodlands	0.672	0.679	0.667	0.877	0.881	0.877	0.655	0.665	0.654	0.654	0.661	0.645
Residential	0.745	0.749	0.726	0.933	0.932	0.925	0.723	0.734	0.712	0.715	0.731	0.705
Transportation	0.677	0.723	0.720	0.948	0.959	0.959	0.584	0.667	0.668	0.581	0.662	0.664
Water	0.741	0.727	0.754	0.938	0.937	0.944	0.713	0.687	0.742	0.712	0.686	0.736
Garden Fields	0.625	0.634	0.639	0.772	0.779	0.782	0.741	0.751	0.768	0.725	0.739	0.755
Industrial	0.615	0.646	0.731	0.777	0.796	0.856	0.707	0.748	0.814	0.695	0.746	0.807
Commercial	0.520	0.494	0.590	0.823	0.803	0.865	0.361	0.333	0.517	0.341	0.322	0.506
Public Land	0.459	0.444	0.636	0.750	0.716	0.849	0.317	0.323	0.632	0.285	0.302	0.622

[1] The pre-trained model is used for training. [2] Fine-tuning only the decoder layer.

Table 5. Test metrics of the SegFormer model under the LoRA strategy for level 1 classification.

Land Cover	SegFormer [1,3] MIoU	SegFormer [1] MIoU	SegFormer [2,3] MIoU	SegFormer [2] MIoU	SegFormer [1,3] Acc	SegFormer [1] Acc	SegFormer [2,3] Acc	SegFormer [2] Acc	SegFormer [1,3] F1-Score	SegFormer [1] F1-Score	SegFormer [2,3] F1-Score	SegFormer [2] F1-Score	SegFormer [1,3] Dice	SegFormer [1] Dice	SegFormer [2,3] Dice	SegFormer [2] Dice
Farmland	0.634	0.630	0.642	0.625	0.811	0.817	0.824	0.814	0.667	0.652	0.668	0.647	0.650	0.635	0.648	0.631
Woodlands	0.631	0.682	0.649	0.668	0.849	0.887	0.872	0.879	0.614	0.665	0.615	0.646	0.591	0.658	0.598	0.641
Residential	0.683	0.753	0.712	0.738	0.901	0.935	0.921	0.927	0.652	0.736	0.680	0.720	0.648	0.731	0.668	0.718
Transportation	0.648	0.721	0.651	0.700	0.933	0.960	0.942	0.956	0.547	0.657	0.549	0.624	0.549	0.648	0.539	0.618
Water	0.693	0.739	0.737	0.734	0.915	0.938	0.938	0.936	0.664	0.710	0.707	0.706	0.683	0.706	0.706	0.705
Garden Fields	0.631	0.630	0.626	0.605	0.776	0.776	0.772	0.755	0.755	0.748	0.748	0.736	0.735	0.732	0.732	0.721
Industrial	0.694	0.665	0.664	0.654	0.831	0.809	0.812	0.804	0.775	0.741	0.746	0.738	0.768	0.736	0.732	0.724
Commercial	0.530	0.501	0.509	0.486	0.803	0.825	0.819	0.814	0.435	0.309	0.326	0.278	0.419	0.282	0.299	0.257
Public Land	0.583	0.459	0.541	0.460	0.798	0.735	0.782	0.735	0.560	0.298	0.502	0.322	0.567	0.285	0.492	0.297

[1] Applying the LoRA parameters only at the decoder layer. [2] Applying the LoRA parameters at the encoder and decoder layers. [3] Fine-tuning only the LoRA parameters.

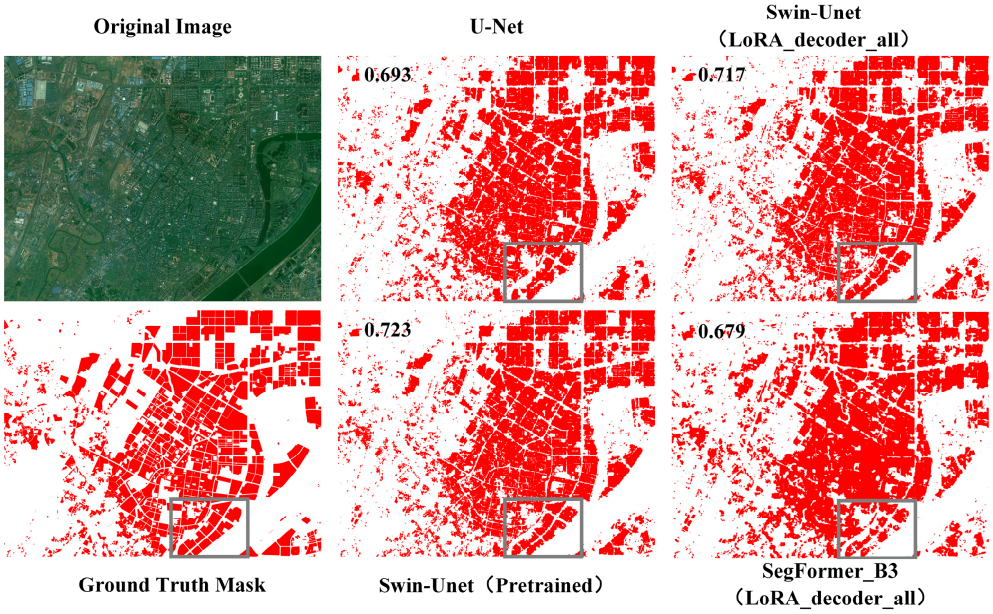


Figure 9. Visualization of model prediction results for the residential category and their predicted MIoU metrics (one area of contrast is circled by the grey square). "Pretrained" stands for fully fine-tuned based on the pre-trained model. "LoRA_decoder_all" stands for applying the LoRA parameters only at the decoder layers and then fine-tuning all parameters.

For the purpose of more meticulously reflecting the visual effect of the model's semantic segmentation, the prediction results of the models on patch images are visualized in Figure 10. Taking the traffic type as an illustration, the predicted MIoU of the U-Net model for the patch is 0.822, which is higher than the 0.747 of the DeepLab V3+ model. The results also demonstrate that the DeepLab V3+ model exhibits deficiencies in interpreting the details of the ground objects, which may be due to the fact that the model finally outputs the result through bilinear upsampling. In comparison with the baseline architectures, the Swin-Unet model shows superior performance. The model with random initialization of parameters exhibits the highest semantic segmentation accuracy with an MIoU of 0.880, which is 13.3% and 5.8% higher than the DeepLab V3+ model and the U-Net model, respectively. In terms of visual quality, the Swin-Unet model demonstrates higher integrity and better detail in the interpretation of patch images. It not only interprets the main roads more accurately, but also depicts the general outline and orientation of the surrounding traffic branches.

4.2. Semantic Segmentation Results for Level 2 Classification

4.2.1. Evaluation Metrics of Model Testing for Level 2 Classification

Based on the experimental results of the level 1 classifications, the hierarchical transfer learning framework is employed to explore the performance of the models on the semantic segmentation task of the subclasses contained in the major categories. Tables 6, 7, A7 and A8 present the results of the evaluation metrics of the baseline architectures for seven level 2 land use categories. The semantic segmentation accuracy of the baseline architectures with randomly initialized parameters shows a decline for level 2 classifications. However, the use of pre-trained models from level 1 categories significantly enhances the model's semantic segmentation performance for subcategories, particularly in terms of MIoU and F1-score. For the paddy category, the utilization of the pre-trained model trained from the primary category improves the test MIoU of the DeepLab V3+ model from 0.557 to 0.646 and the test F1-score from 0.717 to 0.787. A similar trend is observed in the U-Net model, where the test MIoU increases from 0.557 to 0.617, and the F1-score rises from 0.618

to 0.708. These results indicate that the transfer of knowledge from primary categories aids in enhancing the accuracy of boundary segmentation while helping to balance precision and recall in the segmentation process. However, the overall interpretation performance of the models for some categories such as bamboo and shrub remains poor, which may be attributed to issues such as the imbalance of sample categories, overfitting, the fragmented nature of the features, etc.

Table 6. Test metrics of the DeepLab V3+ model under hierarchical transfer learning strategy for level 2 classification.

Land Cover	DeepLab V3+ MIoU	DeepLab V3+ [1] MIoU	DeepLab V3+ Acc	DeepLab V3+ [1] Acc	DeepLab V3+ F1-Score	DeepLab V3+ [1] F1-Score	DeepLab V3+ Dice	DeepLab V3+ [1] Dice
Paddy	0.557	0.646	0.735	0.804	0.717	0.787	0.717	0.787
Irrigated	0.231	0.241	0.736	0.767	0.386	0.411	0.386	0.411
Orchard	0.358	0.481	0.758	0.838	0.525	0.641	0.525	0.641
Bamboo	0.143	0.139	0.891	0.909	0.257	0.248	0.257	0.248
River	0.708	0.783	0.917	0.940	0.810	0.866	0.810	0.866
Arboreal	0.220	0.341	0.887	0.899	0.370	0.522	0.370	0.522
Shrub	0.026	0.002	0.748	0.946	0.049	0.003	0.049	0.003

[1] The pre-trained model of level 1 classification is used for training.

Table 7. Test metrics of the U-Net model under the hierarchical transfer learning strategy for level 2 classification.

Land Cover	U-Net MIoU	U-Net [1] MIoU	U-Net Acc	U-Net [1] Acc	U-Net F1-Score	U-Net [1] F1-Score	U-Net Dice	U-Net [1] Dice
Paddy	0.557	0.617	0.736	0.776	0.618	0.708	0.570	0.686
Irrigated	0.476	0.477	0.819	0.791	0.274	0.277	0.197	0.273
Orchard	0.525	0.580	0.732	0.781	0.539	0.601	0.531	0.601
Bamboo	0.557	0.577	0.892	0.889	0.364	0.432	0.350	0.411
River	0.777	0.814	0.901	0.922	0.833	0.869	0.837	0.867
Arboreal	0.566	0.590	0.874	0.878	0.429	0.482	0.423	0.482
Shrub	0.520	0.489	0.914	0.901	0.242	0.097	0.202	0.063

[1] The pre-trained model of level 1 classification is used for training.

As shown in Tables 8 and A9, the Swin-Unet model shows superior overall performance compared to the baseline architectures. For the paddy, irrigated, and orchard categories, the Swin-Unet model achieves the highest test MIoU values of 0.708, 0.545, and 0.664, which are 6.2%, 6.8%, and 8.4% higher than those of the baseline architectures, respectively. The corresponding test F1-scores for these categories are 0.813, 0.413, and 0.697. In comparison to the randomly initialized model, the utilization of the original pre-trained model has been observed to enhance the interpretation accuracy of the model. Furthermore, the application of the pre-trained model saved from the level 1 category can further improve the interpretation effect of the model. For example, the test MIoU for the bamboo category of the randomly initialized Swin-Unet model is 0.504, and the test F1-score is 0.251. The use of the original pre-trained model improves the test MIoU to 0.563 (an improvement of 5.9%) and the test F1-score to 0.408 (an improvement of 15.7%). Furthermore, the pre-trained model from level 1 classification improves the test MIoU by an additional 3.0% and the test F1-score by an additional 7.7%. Notably, for the paddy category, the use of the pre-trained model of level 1 classification improves the test MIoU by 9.8% and the test F1-score by 9.0% compared to the original pre-trained model. Although the Swin-Unet model achieves the best semantic segmentation performance for level 2 land use categories, its interpretation accuracy remains inferior to that of its corresponding level 1 classification. Moreover, the overlap between the predicted segmentation regions and the ground truth regions is found to be minimal for some categories.

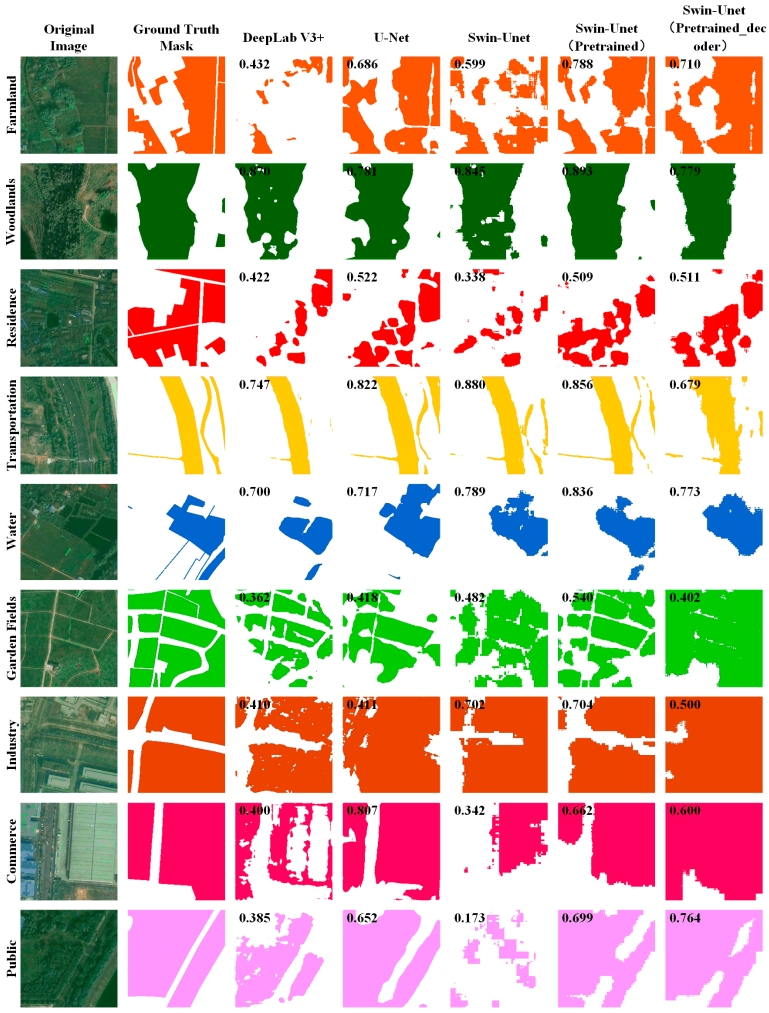


Figure 10. Visualization of model prediction results for nine level 1 land use classifications and their predicted MIoU metrics on patch images. "Pretrained" stands for fully fine-tuned based on the pre-trained model. "Pretrained_decoder" stands for fine-tuning the decoder layers based on the pre-trained model.

Table 8. Test metrics of the Swin-Unet model under the hierarchical transfer learning strategy for level 2 classification.

Land Cover	Swin-Unet MIoU	Swin-Unet [1] MIoU	Swin-Unet [2] MIoU	Swin-Unet Acc	Swin-Unet [1] Acc	Swin-Unet [2] Acc	Swin-Unet F1-Score	Swin-Unet [1] F1-Score	Swin-Unet [2] F1-Score	Swin-Unet Dice	Swin-Unet [1] Dice	Swin-Unet [2] Dice
Paddy	0.605	0.610	0.708	0.757	0.765	0.832	0.723	0.723	0.813	0.698	0.669	0.812
Irrigated	0.421	0.545	0.526	0.843	0.841	0.851	0.253	0.413	0.334	0.244	0.399	0.328
Orchard	0.521	0.587	0.664	0.774	0.815	0.844	0.461	0.558	0.697	0.373	0.525	0.689
Bamboo	0.504	0.563	0.596	0.868	0.881	0.892	0.251	0.408	0.485	0.224	0.404	0.477
River	0.775	0.800	0.830	0.905	0.917	0.929	0.801	0.824	0.862	0.791	0.821	0.854
Arboreal	0.521	0.592	0.614	0.849	0.892	0.889	0.346	0.459	0.536	0.339	0.454	0.532
Shrub	0.473	0.501	0.502	0.947	0.930	0.929	0.097	0.148	0.134	0.083	0.141	0.119

[1] The initial pre-trained model is used for training. [2] The pre-trained model of level 1 classification is used for training.

4.2.2. Visualization of Predicted Results for Level 2 Classification

Due to the sparse sample distribution of the three categories of irrigated, bamboo, and arboreal, the visualization effect and analyzability of their results is poor. Therefore, this study only conducts visual prediction displays for the remaining four categories. Figure 11 shows the visualization results predicted by the model for different categories on large-sized images. It can be seen that the semantic segmentation accuracy of the model is relatively low when the parameters are randomly initialized. Furthermore, the overall performance of the Swin-Unet model is superior to that of the DeepLabV3+ and the U-Net models.

Taking paddy type as an example, when the model parameters are randomly initialized, the predicted MIoU of the DeepLabV3+ model, U-Net model, and Swin-Unet model are 0.575, 0.546, and 0.626, respectively. A comparison reveals obvious the misclassification and omission of clustering phenomena between the visualization results and the real label data. In addition, the model fails to accurately present the edge information of the features. In contrast, the utilization of the pre-trained models obtained from the level 1 categories can significantly improve the prediction accuracy of the model by up to 33.9% (DeepLab V3+ model for orchard). Moreover, the application of the pre-trained model can enhance the model's ability to predict the fragmented parcels and feature boundaries, thereby improving the accuracy, completeness, and level of detail of the model's semantic segmentation results. Zooming in to show the prediction results for the river category (Figure 12), it can be observed that the use of the hierarchical pre-trained model improves the interpretation of the fine features, enabling the model to predict the minute branches of the river more completely (gray box). Notably, there are no river branches in the yellow box in the real label data. However, the prediction results all show the river branches. This is due to the fact that the characteristics of each subcategory of the water type are extremely similar in the remote sensing images, making it challenging to perform precise semantic segmentation. This is a major limitation of the application of remote sensing semantic segmentation in the field of land use.

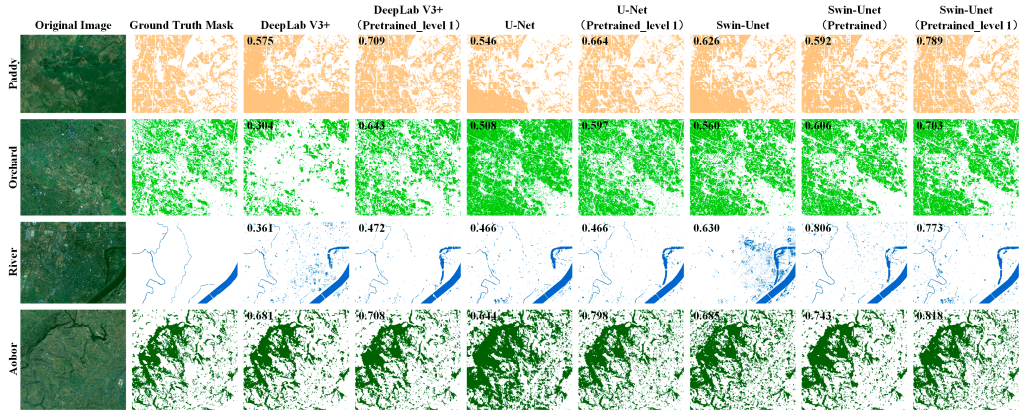


Figure 11. Visualization of model prediction results for four level 2 land use classifications and their predicted MIoU metrics on large-scale images. "Pretrained" stands for fully fine-tuned based on the initial pre-trained model. "Pretrained_level 1" stands for fully fine-tuned based on the pre-trained model of level 1 classification.

Figure 13 illustrates the model's more detailed visual prediction effect on four level 2 land use types, which proves that the utilization of a pre-trained model can enhance the adaptability of the model to different downstream tasks. It also demonstrates the effectiveness of the hierarchical transfer learning strategy. In the baseline architectures, the use of the pre-trained model from a level 1 category results in an improvement in the predicted MIoU of the DeepLab V3+ model from 0.838 to 0.888 for the paddy category, and the predicted MIoU of the U-Net model from 0.545 to 0.701 for the orchard category. For the Swin-Unet model, the predicted MIoU for the arboreal category when fully fine-tuned based on the pre-trained model from the level 1 category is 29.9% higher than when the parameters were randomly initialized, and 19.4% higher than when fully fine-tuned based on the original pre-trained model. Furthermore, the use of the pre-trained model saved from level 1 types can improve the completeness of the model's prediction results for its subclasses and enhance the model's representation of detailed information, both for the baseline architectures and the Swin-Unet model.

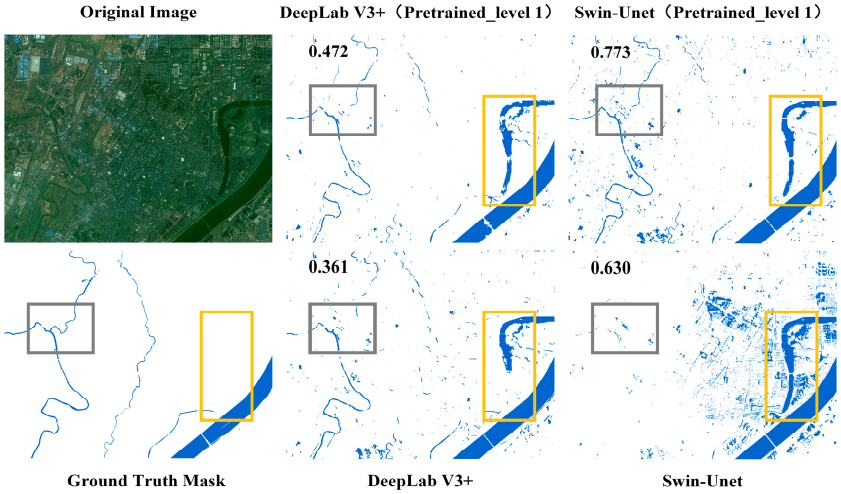


Figure 12. Visualization of model prediction results for the river category and their predicted MIoU metrics (two different areas of contrast are circled by the grey and yellow squares). "Pretrained" stands for fully fine-tuned based on the pre-trained model. "Pretrained_level 1" stands for fully fine-tuned based on the pre-trained model of level 1 classification.

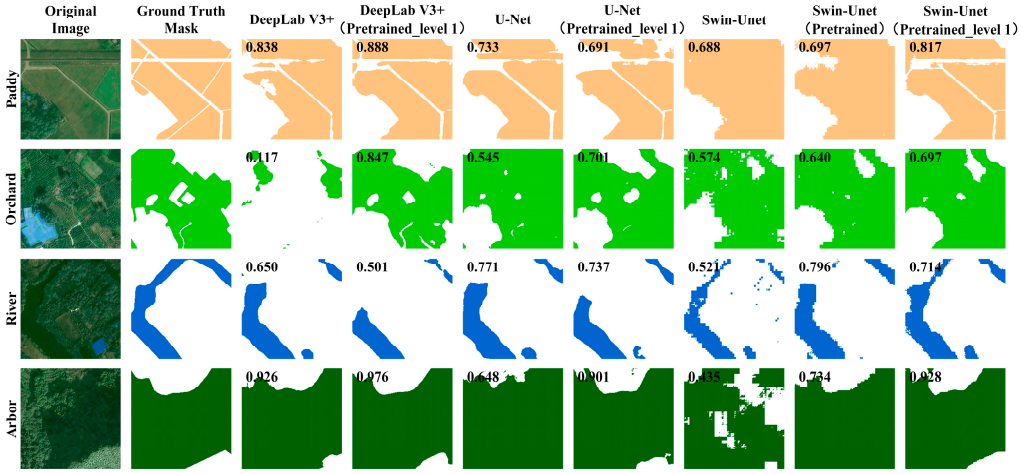


Figure 13. Visualization of model prediction results for four level 2 land use classifications and their predicted MIoU metrics on patch images. "Pretrained" stands for fully fine-tuned based on the initial pre-trained model. "Pretrained_level 1" stands for fully fine-tuned based on the pre-trained model of level 1 classification.

5. Discussion

CNN is one of the most successful and widely used architectures for deep learning algorithms in remote sensing. However, the local receptive fields inherent to CNN constrain their capacity to model long-range dependencies in images [13,39]. Furthermore, convolution is content-independent, with the same weights be applied to all inputs given that the weights of convolution filters are fixed [13,39]. In contrast, Transformer has the ability to model long-distance interactions in a content-dependent manner, with the flexibility to adjust the sensory field in order to counteract interference in the data and learn effective feature representations [40,41]. This study demonstrates the successful application of the Transformer architectures to the field of remote sensing semantic segmentation of land use, achieving excellent results. Notably, the SegFormer model exhibits suboptimal interpretation results in general, which may be attributable to the model architecture or the pre-trained dataset [42]. Although Transformer models significantly outperform CNNs in this study, their computational efficiency remains a critical concern. Experimental results reveal that, under the same number of iterations and training dataset size, the runtime of

the Swin-Unet model is approximately 2.5 times that of the CNNs, while the SegFormer model exceeds 5 times that of the CNNs. This increased computational cost not only prolongs training time but also imposes greater demands on hardware capabilities. Regarding memory usage, the SegFormer model requires the largest amount of memory, followed by the Swin-Unet model, with the CNNs requiring the least. These findings suggest that while Transformer models achieve improved performance, they do so at the expense of significantly higher computational resource consumption. Notably, we found that Swin-Unet offers a well-balanced choice, delivering superior performance while demanding less hardware resources compared to SegFormer. Furthermore, the Transformer models are more susceptible to gradient explosion during the training phase [43–45], which can result in issues such as a reduction in the training speed and poor classification accuracy. Through our experimental analysis, it has been demonstrated that the application of gradient clipping and adaptive learning rate can effectively mitigate this phenomenon.

The pre-trained datasets used for pre-training the models comprise Image-Net 1K and medical image datasets, the image features of which diverge considerably from those of the dataset utilized in this study [29,31]. However, the experimental results indicate that the application of the pre-trained weights enhances the model's ability to adapt to diverse downstream tasks and improves the semantic segmentation accuracy. This phenomenon indicates that despite significant differences in the source dataset features, the feature representations learned through the model pre-training still facilitate performance on an alternative dataset. In particular, the utilization of pre-trained models obtained from the level 1 categories can further enhance the model's interpretation accuracy for corresponding level 2 categories when compared to the non-remote sensing pre-trained model. This may be attributed to the existence of an intersection between the level 1 and level 2 category features, which enables the model to learn more similar visual representations shared in semantic segmentation. Furthermore, significant discrepancies in the model's interpretation accuracy are observed when different transfer learning strategies are employed. The accuracy of the Swin-Unet model when partially fine-tuned on the decoder layers is lower than that achieved when fully fine-tuned. Conversely, the classification accuracy of the SegFormer model for some land use categories when partially fine-tuned on the decoder layers is enhanced in comparison to when fully fine-tuned. This may be attributed to the pre-training datasets of the models [46]. These findings also confirm that fine-tuning all parameters of the model may not necessarily result in optimal outcomes. Alternatively, partial fine-tuning is likely to effectively enhance the model's performance while reducing costs [35]. Notably, the model under LoRA exhibits the poorest performance among all transfer learning strategies. It is hypothesized that LoRA is an adaptive strategy proposed for fine-tuning larger natural language models, which may not be applicable to smaller visual models or to the Transformer models selected in this study. Further investigation may be warranted in the future.

6. Conclusions

In this paper, we constructed a hierarchical transfer learning framework for the task of semantic segmentation of fine categories. Based on this transfer learning framework, we transferred the Transformer models to the task of land use classification. Two typical Transformer models in the field of computer vision were selected for experimental investigation: the Swin-Unet model and the SegFormer model. The structural features and advantages of the two models are distinct. In addition, the U-Net and DeepLab V3+ convolutional networks, which are widely used and representative of typical architectures, were employed as baselines for comparison. In the hierarchical migration learning framework, this study improved and applied full fine-tuning, partial fine-tuning, and

LoRA strategies to the application of the pre-trained Transformer models. Notably, the adaptation training and prediction for the level 1 categories were based solely on the original pre-trained models. In contrast, for the level 2 category, in addition to the original pre-trained model, the optimal model trained for the corresponding level 1 category was also employed as the pre-trained model for secondary transfer. The results of the systematic experiments demonstrate that the Transformer models perform better than the baseline architectures on the land use dataset of Dongpo District, especially the Swin-Unet model. The proposed hierarchical transfer learning framework is both feasible and effective. On the one hand, applying non-remote sensing pre-trained models with different strategies can significantly enhance semantic segmentation accuracy. Notably, we found that gradient clipping and adaptive learning rates played a crucial role in improving the performance of transfer learning-based approaches for land-use classification in remote sensing. On the other hand, the re-transfer utilization of the trained model for the level 1 categories further improves the model's classification accuracy for the corresponding level 2 categories. This study provides comprehensive evidence of the applicability of the Transformer models and their non-remote sensing pre-trained models in the field of remote sensing land use classification. It also validates the efficacy of hierarchical transfer in enhancing the classification accuracy of fine categories. These contributions build upon existing research and represent a novel advancement in the field of remote sensing semantic segmentation. Nevertheless, it must be acknowledged that our approach does have certain limitations. Firstly, Transformer models incur higher computational costs when calculating global self-attention, necessitating a trade-off between performance and computational efficiency in practical applications. To address this challenge, future research could explore optimization techniques, such as model compression and parameter pruning, to reduce computational complexity and memory usage, thereby promoting the practical adoption of Transformer models in resource-constrained environments. Secondly, while pre-training has demonstrated significant advantages in improving model performance, several limitations remain. Pre-trained weight matrices are often optimized for specific datasets and regions, which may restrict the model's ability to generalize effectively to areas with significant geographical differences or dynamic changes. Moreover, pre-trained models show limited adaptability to future landscape alterations. When faced with repeat imagery that includes significant landscape changes, pre-trained models may struggle to capture new environmental features adequately. To address these limitations, future work could incorporate techniques such as incremental learning or periodic model updates, enabling the model to progressively adapt to evolving landscapes, thereby improving its long-term effectiveness and adaptability. Finally, while the transfer learning approach used in this study has shortened model training times and reduced the number of parameters, there is still room for improvement. Future research could focus on exploring more efficient and streamlined methods for applying pre-trained models through transfer learning to further enhance model efficiency.

Author Contributions: Funding acquisition, L.L.; data curation, M.C.; methodology, M.C. and L.L.; software, M.C. and L.L.; investigation, M.C. and L.L.; validation, M.C. and L.L.; formal analysis, M.C. and L.L.; writing—original draft preparation, M.C.; writing—review and editing, M.C. and L.L.; visualization, M.C.; project administration, L.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Key Research and Development Program of China (grant number 2021YFB3900501), National Natural Science Foundation of China (grant number 42071369), LREIS Independent Innovation Project (grant number 05Z5006JYA), and the Strategic Priority Research Program of Chinese Academy of Sciences (grant number XDB0740100).

Data Availability Statement: The datasets presented in this article are available on request from the corresponding author. The data are not publicly available due to privacy considerations.

Conflicts of Interest: The authors declare that they have no known competing financial interest or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A

Table A1. Land use classification table.

Level 1 Classification	Level 2 Classification
Farmland	Paddy Irrigated Dry Land
Woodlands	Arboreal Bamboo Shrub Other Woodlands
Residential	Urban Residential Land Rural Homesteads
Transportation	Highways Roads In Towns And Villages Transportation Service Stations Rural Roads
Water	Breeding Pits River Pond Water Inland Tidal Flats Ditch Hydraulic Construction
Garden Fields	Orchard Tea Garden Other Gardens
Industrial	Industrial Land Mining Land
Commercial	Logistics And Warehousing Commercial Service Facilities
Public Land	Utility Land Parks And Green Spaces Press And Publication By Organs And Groups UNESCO

Appendix B

The U-Net network (Figure A1) was first proposed by Olaf Ronneberger, Philipp Fischer, and Thomas Brux, and has achieved very good performance in the application of biomedical segmentation [47]. The network mainly consists of two parts: the first part is the contraction path, which employs a downsampling module composed of multiple convolutional blocks to extract semantic and contextual features; the second part is the symmetric expansion path, which applies a set of convolutional blocks equipped with upsampling operations to progressively increase the spatial resolution of the feature map while reducing the dimensions to produce pixel classification scores. In addition, the network utilizes the decoder to learn the spatial correlation of the corresponding encoding stage through jump connections to mitigate the loss of spatial information caused by downsampling. Currently, the U-Net network plays an important role in the fields of

image segmentation [48], image synthesis [49,50], image denoising [51,52]. and image reconstruction [53].

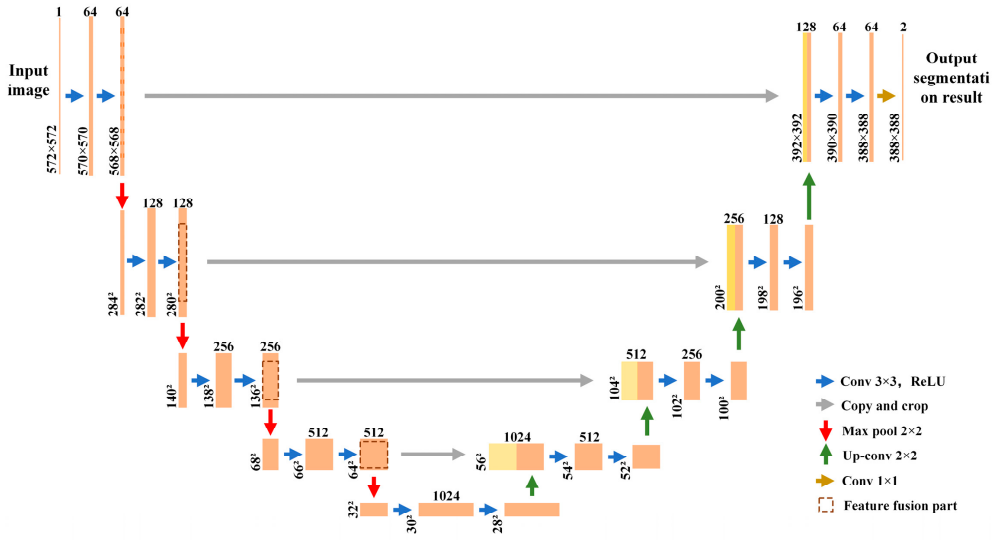


Figure A1. U-Net architecture.

The spatial pyramid pooling module can encode multi-scale contextual information by using filters or pooling operations as detection incoming features at multiple rates and multiple effective fields of view, while the encoder-decoder structure can capture clearer object boundaries by gradually restoring spatial information. In 2018, Chen et al. combined the strengths of both approaches—the spatial pyramid pooling module and the encoder-decoder structure—proposed the DeepLab V3+ model (Figure A2), and demonstrated the superior performance of the model on the PASCAL VOC 2012 and Cityscapes datasets [54]. Notably, the network not only detects convolutional features at multiple scales by applying different ratios of dilated convolution [55], but is also capable of arbitrarily controlling the resolution of the extracted encoder features through null convolution to trade-off accuracy and runtime.

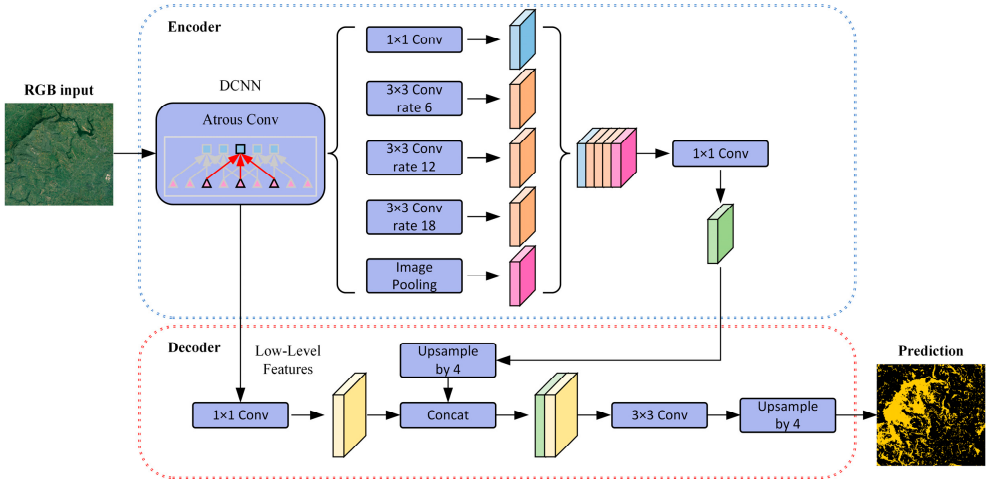


Figure A2. DeepLab V3+ architecture. The DeepLab V3+ model processes contextual information using multi-scale convolutions. The purple squares indicate different types of convolutions or processing operations, and the cubes in different colors denote feature representations at different hierarchical levels.

U-Net and DeepLab V3+ network architectures are classic semantic segmentation frameworks with wide applications and excellent performance. Therefore, this study analyzes them as baseline architectures for comparative analysis with the Transformer networks.

Appendix C

Table A2. Training metrics of the DeepLab V3+ model and the U-Net model for level 1 classification.

Land Cover	DeepLab V3+ MIoU	U-Net MIoU	DeepLab V3+ Acc	U-Net Acc	DeepLab V3+ F1-Score	U-Net F1-Score	DeepLab V3+ Dice	U-Net Dice
Farmland	0.889	0.913	0.967	0.962	0.944	0.937	0.944	0.939
Woodlands	0.903	0.914	0.983	0.973	0.950	0.947	0.950	0.948
Residential	0.906	0.937	0.988	0.985	0.956	0.942	0.956	0.945
Transportation	0.806	0.920	0.988	0.990	0.900	0.922	0.900	0.923
Water	0.844	0.940	0.982	0.987	0.924	0.947	0.924	0.947
Garden Fields	0.932	0.929	0.970	0.964	0.964	0.958	0.964	0.959
Industrial	0.980	0.742	0.993	0.862	0.990	0.812	0.990	0.810
Commercial	0.918	0.964	0.987	0.990	0.962	0.969	0.962	0.970
Public Land	0.943	0.960	0.988	0.985	0.973	0.972	0.973	0.972

Table A3. Training metrics of Swin-Unet model under different fine-tuning strategies for level 1 classification.

Land Cover	Swin-Unet MIoU	Swin-Unet [1] MIoU	Swin-Unet [1,2] MIoU	Swin-Unet Acc	Swin-Unet [1] Acc	Swin-Unet [1,2] Acc	Swin-Unet F1-Score	Swin-Unet [1] F1-Score	Swin-Unet [1,2] F1-Score	Swin-Unet Dice	Swin-Unet [1] Dice	Swin-Unet [1,2] Dice
Farmland	0.908	0.934	0.810	0.959	0.971	0.912	0.941	0.964	0.875	0.940	0.964	0.876
Woodlands	0.940	0.934	0.802	0.982	0.980	0.932	0.961	0.961	0.846	0.961	0.961	0.848
Residential	0.950	0.944	0.832	0.989	0.987	0.957	0.965	0.963	0.859	0.965	0.964	0.862
Transportation	0.910	0.921	0.767	0.989	0.991	0.966	0.924	0.944	0.746	0.925	0.944	0.755
Water	0.883	0.894	0.838	0.974	0.977	0.962	0.893	0.910	0.856	0.893	0.910	0.858
Garden Fields	0.881	0.920	0.804	0.938	0.959	0.893	0.934	0.965	0.903	0.932	0.964	0.903
Industrial	0.955	0.959	0.902	0.979	0.981	0.953	0.975	0.980	0.951	0.975	0.980	0.953
Commercial	0.949	0.954	0.824	0.985	0.987	0.944	0.963	0.971	0.871	0.963	0.972	0.874
Public Land	0.946	0.948	0.863	0.980	0.981	0.947	0.961	0.971	0.915	0.961	0.971	0.917

[1] The pre-trained model is used for training. [2] Fine-tuning only the decoder layer.

Table A4. Training metrics of the Swin-Unet model under the LoRA strategy for level 1 classification.

Land Cover	Swin-Unet [1,3] MIoU	Swin-Unet [1] MIoU	Swin-Unet [2,3] MIoU	Swin-Unet [2] MIoU	Swin-Unet [1,3] Acc	Swin-Unet [1] Acc	Swin-Unet [2,3] Acc	Swin-Unet [2] Acc	Swin-Unet [1,3] F1-Score	Swin-Unet [1] F1-Score	Swin-Unet [2,3] F1-Score	Swin-Unet [2] F1-Score	Swin-Unet [1,3] Dice	Swin-Unet [1] Dice	Swin-Unet [2,3] Dice	Swin-Unet [2] Dice
Farmland	0.908	0.934	0.810	0.864	0.959	0.971	0.912	0.873	0.689	0.860	0.607	0.879	0.678	0.856	0.587	0.877
Woodlands	0.940	0.934	0.802	0.859	0.982	0.980	0.932	0.907	0.675	0.947	0.534	0.865	0.669	0.947	0.515	0.864
Residential	0.950	0.944	0.832	0.882	0.989	0.987	0.957	0.947	0.685	0.945	0.720	0.735	0.690	0.945	0.724	0.733
Transportation	0.910	0.921	0.767	0.843	0.989	0.991	0.966	0.967	0.595	0.913	0413	0.777	0.605	0.914	0.421	0.778
Water	0.883	0.894	0.838	0.888	0.974	0.977	0.962	0.956	0.728	0.835	0.576	0.691	0.727	0.835	0.573	0.688
Garden Fields	0.881	0.920	0.804	0.841	0.938	0.959	0.893	0.832	0.757	0.895	0.773	0.922	0.741	0.892	0.763	0.920
Industrial	0.955	0.959	0.902	0.940	0.979	0.981	0.953	0.907	0.816	0.947	0.709	0.834	0.812	0.945	0.700	0.831
Commercial	0.949	0.954	0.824	0.904	0.985	0.987	0.944	0.896	0.640	0.700	0.339	0.772	0.642	0.699	0.293	0.772
Public Land	0.946	0.948	0.863	0.912	0.980	0.981	0.947	0.899	0.597	0.931	0.499	0.873	0.579	0.931	0.496	0.872

[1] Applying the LoRA parameters only at the decoder layer. [2] Applying the LoRA parameters at the encoder and decoder layers. [3] Fine-tuning only the LoRA parameters.

Table A5. Training metrics of the SegFormer model under different fine-tuning strategies for level 1 classification.

Land Cover	SegFormer MIoU	SegFormer [1] MIoU	SegFormer [1,2] MIoU	SegFormer Acc	SegFormer [1] Acc	SegFormer [1,2] Acc	SegFormer F1-Score	SegFormer [1] F1-Score	SegFormer [1,2] F1-Score	SegFormer Dice	SegFormer [1] Dice	SegFormer [1,2] Dice
Farmland	0.712	0.670	0.802	0.859	0.838	0.909	0.756	0.712	0.880	0.751	0.701	0.873
Woodlands	0.771	0.728	0.745	0.919	0.902	0.910	0.782	0.732	0.778	0.784	0.730	0.773
Residential	0.794	0.797	0.816	0.948	0.947	0.953	0.791	0.797	0.843	0.784	0.796	0.841
Transportation	0.696	0.764	0.797	0.954	0.968	0.973	0.618	0.730	0.793	0.618	0.727	0.793
Water	0.782	0.746	0.859	0.949	0.941	0.969	0.770	0.718	0.887	0.770	0.717	0.883
Garden Fields	0.689	0.685	0.693	0.819	0.815	0.821	0.796	0.794	0.818	0.785	0.785	0.811
Industrial	0.627	0.766	0.911	0.788	0.876	0.957	0.701	0.833	0.960	0.679	0.829	0.956
Commercial	0.657	0.729	0.887	0.882	0.909	0.966	0.633	0.731	0.939	0.625	0.730	0.937
Public Land	0.615	0.513	0.894	0.839	0.755	0.960	0.617	0.471	0.952	0.583	0.434	0.951

[1] The pre-trained model is used for training. [2] Fine-tuning only the decoder layer.

Table A6. Training metrics of the SegFormer model under the LoRA strategy for level 1 classification.

Land Cover	SegFormer [1,3] MIoU	SegFormer [1] MIoU	SegFormer [2,3] MIoU	SegFormer [2] MIoU	SegFormer [1,3] Acc	SegFormer [1] Acc	SegFormer [2,3] Acc	SegFormer [2] Acc	SegFormer [1,3] F1-Score	SegFormer [1] F1-Score	SegFormer [2,3] F1-Score	SegFormer [2] F1-Score	SegFormer [1,3] Dice	SegFormer [1] Dice	SegFormer [2,3] Dice	SegFormer [2] Dice
Farmland	0.641	0.652	0.644	0.642	0.815	0.830	0.824	0.823	0.672	0.686	0.678	0.675	0.653	0.666	0.659	0.658
Woodlands	0.641	0.745	0.658	0.691	0.852	0.912	0.875	0.888	0.639	0.751	0.634	0.684	0.618	0.745	0.619	0.679
Residential	0.685	0.795	0.722	0.787	0.902	0.947	0.924	0.943	0.657	0.793	0.694	0.786	0.652	0.790	0.681	0.786
Transportation	0.647	0.759	0.657	0.723	0.935	0.968	0.946	0.962	0.546	0.718	0.558	0.662	0.550	0.711	0.550	0.657
Water	0.704	0.787	0.754	0.763	0.918	0.951	0.942	0.943	0.682	0.779	0.730	0.747	0.699	0.775	0.729	0.746
Garden Fields	0.639	0.701	0.634	0.620	0.782	0.827	0.778	0.767	0.762	0.807	0.756	0.748	0.743	0.795	0.741	0.734
Industrial	0.776	0.847	0.777	0.737	0.884	0.924	0.886	0.861	0.842	0.865	0.829	0.809	0.835	0.868	0.811	0.793
Commercial	0.637	0.623	0.640	0.588	0.859	0.875	0.876	0.860	0.628	0.579	0.598	0.524	0.621	0.554	0.576	0.497
Public Land	0.646	0.651	0.684	0.571	0.834	0.851	0.864	0.802	0.646	0.617	0.717	0.548	0.651	0.590	0.710	0.531

[1] Applying the LoRA parameters only at the decoder layer. [2] Applying the LoRA parameters at the encoder and decoder layers. [3] Fine-tuning only the LoRA parameters.

Table A7. Training metrics of the DeepLab V3+ model under the hierarchical transfer learning strategy for level 2 classification.

Land Cover	DeepLab V3+ MIoU	DeepLab V3+ [1] MIoU	DeepLab V3+ Acc	DeepLab V3+ [1] Acc	DeepLab V3+ F1-Score	DeepLab V3+ [1] F1-Score	DeepLab V3+ Dice	DeepLab V3+ [1] Dice
Paddy	0.635	0.785	0.909	0.949	0.770	0.872	0.770	0.872
Irrigated	0.333	0.504	0.921	0.937	0.339	0.635	0.339	0.635
Orchard	0.479	0.807	0.758	0.923	0.530	0.895	0.530	0.895
Bamboo	0.656	0.772	0.975	0.984	0.755	0.874	0.755	0.874
River	0.813	0.954	0.951	0.988	0.906	0.977	0.906	0.977
Arboreal	0.647	0.684	0.934	0.937	0.794	0.819	0.794	0.819
Shrub	0.006	0.538	0.930	0.971	0.076	0.720	0.076	0.720

[1] The pre-trained model of level 1 classification is used for training.

Table A8. Training metrics of the U-Net model under the hierarchical transfer learning strategy for level 2 classification.

Land Cover	U-Net MIoU	U-Net [1] MIoU	U-Net Acc	U-Net [1] Acc	U-Net F1-Score	U-Net [1] F1-Score	U-Net Dice	U-Net [1] Dice
Paddy	0.617	0.786	0.826	0.915	0.387	0.653	0.333	0.606
Irrigated	0.500	0.976	0.856	0.995	0.290	0.973	0.247	0.974
Orchard	0.574	0.831	0.746	0.914	0.666	0.832	0.666	0.836
Bamboo	0.606	0.698	0.927	0.943	0.407	0.537	0.394	0.550
River	0.876	0.934	0.948	0.973	0.903	0.948	0.904	0.948
Arboreal	0.679	0.912	0.895	0.975	0.641	0.885	0.646	0.887
Shrub	0.517	0.636	0.916	0.945	0.213	0.496	0.148	0.491

[1] The pre-trained model of level 1 classification is used for training.

Table A9. Training metrics of Swin-Unet model under the hierarchical transfer learning strategy for level 2 classification.

Land Cover	Swin-Unet MIoU	Swin-Unet [1] MIoU	Swin-Unet [2] MIoU	Swin-Unet Acc	Swin-Unet [1] Acc	Swin-Unet [2] Acc	Swin-Unet F1-Score	Swin-Unet [1] F1-Score	Swin-Unet [2] F1-Score	Swin-Unet Dice	Swin-Unet [1] Dice	Swin-Unet [2] Dice
Paddy	0.632	0.624	0.856	0.834	0.841	0.946	0.635	0.613	0.900	0.625	0.594	0.899
Irrigated	0.448	0.631	0.969	0.896	0.903	0.994	0.343	0.587	0.978	0.321	0.574	0.978
Orchard	0.544	0.610	0.718	0.725	0.768	0.841	0.613	0.708	0.820	0.546	0.693	0.819
Bamboo	0.547	0.618	0.629	0.889	0.917	0.907	0.351	0.499	0.532	0.354	0.507	0.549
River	0.767	0.916	0.883	0.900	0.966	0.952	0.794	0.943	0.914	0.789	0.942	0.912
Arboreal	0.686	0.728	0.796	0.878	0.908	0.931	0.686	0.763	0.821	0.702	0.763	0.823
Shrub	0.471	0.652	0.647	0.942	0.947	0.938	0.435	0.558	0.565	0.462	0.560	0.574

[1] The initial pre-trained model is used for training. [2] The pre-trained model of level 1 classification is used for training.

References

1. Luo, H.; Chen, C.; Fang, L.; Khoshelham, K.; Shen, G. MS-RRFSegNet: Multiscale regional relation feature segmentation network for semantic segmentation of urban scene point clouds. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 8301–8315. [CrossRef]
2. Mohamed, A.; Worku, H. Quantification of the land use/land cover dynamics and the degree of urban growth goodness for sustainable urban land use planning in Addis Ababa and the surrounding Oromia special zone. *J. Urban Manag.* **2019**, *8*, 145–158. [CrossRef]
3. Rottensteiner, F.; Sohn, G.; Jung, J.; Gerke, M.; Baillard, C.; Benitez, S.; Breitkopf, U. The ISPRS benchmark on urban object classification and 3D building reconstruction. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* **2012**, *1*, 293–298. [CrossRef]
4. Wen, D.; Ma, S.; Zhang, A.; Ke, X. Spatial pattern analysis of the ecosystem services in the Guangdong-Hong Kong-Macao Greater Bay Area using Sentinel-1 and Sentinel-2 imagery based on deep learning method. *Sustainability* **2021**, *13*, 7044. [CrossRef]
5. Blaschke, T.; Lang, S.; Lorup, E.; Strobl, J.; Zeil, P. Object-oriented image processing in an integrated GIS/remote sensing environment and perspectives for environmental applications. *Environ. Inf. Plan. Politics Public* **2000**, *2*, 555–570.
6. Sheikh, R.; Milioto, A.; Lottes, P.; Stachniss, C.; Bennewitz, M.; Schultz, T. Gradient and log-based active learning for semantic segmentation of crop and weed for agricultural robots. In Proceedings of the 2020 IEEE International Conference on Robotics and Automation (ICRA), Paris, France, 31 May–31 August 2020; pp. 1350–1356.
7. Muhadi, N.A.; Abdullah, A.F.; Bejo, S.K.; Mahadi, M.R.; Mijic, A. Deep learning semantic segmentation for water level estimation using surveillance camera. *Appl. Sci.* **2021**, *11*, 9691. [CrossRef]
8. Zhu, X.X.; Tuia, D.; Mou, L.; Xia, G.-S.; Zhang, L.; Xu, F.; Fraundorfer, F. Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE Geosci. Remote Sens. Mag.* **2017**, *5*, 8–36. [CrossRef]
9. Zhu, L.; Chen, Y.; Ghamisi, P.; Benediktsson, J.A. Generative adversarial networks for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 5046–5063. [CrossRef]
10. 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]
11. Ye, Z.; Fu, Y.; Gan, M.; Deng, J.; Comber, A.; Wang, K. Building extraction from very high resolution aerial imagery using joint attention deep neural network. *Remote Sens.* **2019**, *11*, 2970. [CrossRef]
12. Ren, Y.; Zhang, X.; Ma, Y.; Yang, Q.; Wang, C.; Liu, H.; Qi, Q. Full Convolutional Neural Network Based on Multi-Scale Feature Fusion for the Class Imbalance Remote Sensing Image Classification. *Remote Sens.* **2020**, *12*, 3547. [CrossRef]
13. Chen, J.; Lu, Y.; Yu, Q.; Luo, X.; Adeli, E.; Wang, Y.; Lu, L.; Yuille, A.L.; Zhou, Y. TransUNet: Transformers make strong encoders for medical image segmentation. *arXiv* **2021**, arXiv:2102.04306. Available online: https://arxiv.org/abs/2102.04306 (accessed on 1 January 2025).
14. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is ill you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5998–6008.
15. Chen, K.; Zou, Z.; Shi, Z. Building extraction from remote sensing images with sparse token transformers. *Remote Sens.* **2021**, *13*, 4441. [CrossRef]
16. Wang, H.; Chen, X.; Zhang, T.; Xu, Z.; Li, J. CCTNet: Coupled CNN and transformer network for crop segmentation of remote sensing images. *Remote Sens.* **2022**, *14*, 1956. [CrossRef]
17. Keles, F.D.; Wijewardena, P.M.; Hegde, C. On the computational complexity of self-attention. In Proceedings of the International Conference on Algorithmic Learning Theory, Singapore, 20–23 February 2023; pp. 597–619.
18. Gadiraju, K.K.; Vatsavai, R.R. Comparative analysis of deep transfer learning performance on crop classification. In Proceedings of the 9th ACM SIGSPATIAL International Workshop on Analytics for Big Geospatial Data, Seattle, WA, USA, 3 November 2020; pp. 1–8.
19. Lu, G.; Li, S.; Mai, G.; Sun, J.; Zhu, D.; Chai, L.; Sun, H.; Wang, X.; Dai, H.; Liu, N. Agi for agriculture. *arXiv* **2023**, arXiv:2304.06136. Available online: https://arxiv.org/abs/2304.06136 (accessed on 1 January 2025).
20. Michau, G.; Fink, O. Unsupervised transfer learning for anomaly detection: Application to complementary operating condition transfer. *Knowl.-Based Syst.* **2021**, *216*, 106816. [CrossRef]
21. Oikonomidis, A.; Catal, C.; Kassahun, A. Deep learning for crop yield prediction: A systematic literature review. *New Zealand J. Crop Hortic. Sci.* **2023**, *51*, 1–26. [CrossRef]
22. Ma, Y.; Chen, S.; Ermon, S.; Lobell, D.B. Transfer learning in environmental remote sensing. *Remote Sens. Environ.* **2024**, *301*, 113924. [CrossRef]
23. Pan, S.J.; Yang, Q. A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* **2009**, *22*, 1345–1359. [CrossRef]
24. Lin, C.; Zhong, L.; Song, X.-P.; Dong, J.; Lobell, D.B.; Jin, Z. Early-and in-season crop type mapping without current-year ground truth: Generating labels from historical information via a topology-based approach. *Remote Sens. Environ.* **2022**, *274*, 112994. [CrossRef]
25. Chew, R.; Rineer, J.; Beach, R.; O'Neil, M.; Ujeneza, N.; Lapidus, D.; Miano, T.; Hegarty-Craver, M.; Polly, J.; Temple, D.S. Deep neural networks and transfer learning for food crop identification in UAV images. *Drones* **2020**, *4*, 7. [CrossRef]

26. Min, S.; Xie, H.; Yao, H.; Deng, X.; Zha, Z.-J.; Zhang, Y. Hierarchical granularity transfer learning. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 11650–11661.
27. Luo, G.; Yang, Y.; Yuan, Y.; Chen, Z.; Ainiwaer, A. Hierarchical transfer learning architecture for low-resource neural machine translation. *IEEE Access* **2019**, *7*, 154157–154166. [CrossRef]
28. Chang, J.-W.; Ma, H.-S.; Hu, Z.-Y. Multi-Level Transfer Learning using Incremental Granularities for environmental sound classification and detection. *Appl. Soft Comput.* **2024**, *112*, 112619. [CrossRef]
29. Cao, H.; Wang, Y.; Chen, J.; Jiang, D.; Zhang, X.; Tian, Q.; Wang, M. Swin-unet: Unet-like pure transformer for medical image segmentation. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; pp. 205–218.
30. 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, Virtual, 11–17 October 2021; pp. 10012–10022.
31. 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.
32. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929. Available online: https://arxiv.org/abs/2010.11929 (accessed on 1 January 2025).
33. Ghazi, M.M.; Yanikoglu, B.; Aptoula, E. Plant identification using deep neural networks via optimization of transfer learning parameters. *Neurocomputing* **2017**, *235*, 228–235. [CrossRef]
34. Yosinski, J.; Clune, J.; Bengio, Y.; Lipson, H. How transferable are features in deep neural networks? In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014; Volume 27.
35. Ding, N.; Qin, Y.; Yang, G.; Wei, F.; Yang, Z.; Su, Y.; Hu, S.; Chen, Y.; Chan, C.-M.; Chen, W. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nat. Mach. Intell.* **2023**, *5*, 220–235. [CrossRef]
36. Abdalla, A.; Cen, H.; Wan, L.; Rashid, R.; Weng, H.; Zhou, W.; He, Y. Fine-tuning convolutional neural network with transfer learning for semantic segmentation of ground-level oilseed rape images in a field with high weed pressure. *Comput. Electron. Agric.* **2019**, *167*, 105091. [CrossRef]
37. Hu, E.J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W. LoRA: Low-Rank Adaptation of Large Language Models. *arXiv* **2021**, arXiv:2106.09685. Available online: https://arxiv.org/abs/2106.09685 (accessed on 1 January 2025).
38. Shen, S.; Li, L.H.; Tan, H.; Bansal, M.; Rohrbach, A.; Chang, K.-W.; Yao, Z.; Keutzer, K. How Much Can CLIP Benefit Vision-and-Language Tasks? *arXiv* **2021**, arXiv:2107.06383. Available online: https://arxiv.org/abs/2107.06383 (accessed on 1 January 2025).
39. Aleissaee, A.A.; Kumar, A.; Anwer, R.M.; Khan, S.; Cholakkal, H.; Xia, G.-S.; Khan, F.S. Transformers in remote sensing: A survey. *Remote Sens.* **2023**, *15*, 1860. [CrossRef]
40. Naseer, M.M.; Ranasinghe, K.; Khan, S.H.; Hayat, M.; Shahbaz Khan, F.; Yang, M.-H. Intriguing properties of vision transformers. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 23296–23308.
41. Park, N.; Kim, S. How Do Vision Transformers Work? *arXiv* **2022**, arXiv:2202.06709. Available online: https://arxiv.org/abs/2202.06709 (accessed on 1 January 2025).
42. Wang, D.; Zhang, J.; Du, B.; Xia, G.-S.; Tao, D. An empirical study of remote sensing pretraining. *IEEE Trans. Geosci. Remote Sens.* **2022**, *61*, 1–20. [CrossRef]
43. Saxe, A.M.; McClelland, J.L.; Ganguli, S. Exact Solutions to the Nonlinear Dynamics of Learning in Deep Linear Neural Networks. *arXiv* **2013**, arXiv:1312.6120. Available online: https://arxiv.org/abs/1312.6120 (accessed on 1 January 2025).
44. Dong, Y.; Cordonnier, J.-B.; Loukas, A. Attention is not all you need: Pure attention loses rank doubly exponentially with depth. In Proceedings of the International Conference on Machine Learning, Virtual, 18–24 July 2021; pp. 2793–2803.
45. Noci, L.; Anagnostidis, S.; Biggio, L.; Orvieto, A.; Singh, S.P.; Lucchi, A. Signal propagation in transformers: Theoretical perspectives and the role of rank collapse. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 27198–27211.
46. Clark, C.; Yatskar, M.; Zettlemoyer, L. Don't Take the Easy Way Out: Ensemble-Based Methods for Avoiding Known Dataset Biases. *arXiv* **2019**, arXiv:1909.03683. Available online: https://arxiv.org/abs/1909.03683 (accessed on 1 January 2025).
47. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, 5–9 October 2015; Proceedings, Part III 18. pp. 234–241.
48. Minaee, S.; Boykov, Y.; Porikli, F.; Plaza, A.; Kehtarnavaz, N.; Terzopoulos, D. Image segmentation using deep learning: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 3523–3542. [CrossRef]
49. Sun, B.; Jia, S.; Jiang, X.; Jia, F. Double U-Net CycleGAN for 3D MR to CT image synthesis. *Int. J. Comput. Assist. Radiol. Surg.* **2023**, *18*, 149–156. [CrossRef] [PubMed]
50. Costa, P.; Galdran, A.; Meyer, M.I.; Abramoff, M.D.; Niemeijer, M.; Mendonça, A.M.; Campilho, A. Towards Adversarial Retinal Image Synthesis. *arXiv* **2017**, arXiv:1701.08974. Available online: https://arxiv.org/abs/1701.08974 (accessed on 1 January 2025).

51. Reymann, M.P.; Würfl, T.; Ritt, P.; Stimpel, B.; Cachovan, M.; Vija, A.H.; Maier, A. U-Net for SPECT image denoising. In Proceedings of the 2019 IEEE Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC), Manchester, UK, 26 October–2 November 2019; pp. 1–2.
52. Lee, S.; Negishi, M.; Urakubo, H.; Kasai, H.; Ishii, S. Mu-net: Multi-scale U-net for two-photon microscopy image denoising and restoration. *Neural Netw.* **2020**, *125*, 92–103. [CrossRef]
53. Guan, S.; Hsu, K.-T.; Eyassu, M.; Chitnis, P.V. Dense Dilated U-Net: Deep Learning for 3D Photoacoustic Tomography Image Reconstruction. *arXiv* **2021**, arXiv:2104.03130. Available online: https://arxiv.org/abs/2104.03130 (accessed on 1 January 2025).
54. Chen, L.-C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European conference on computer vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 801–818.
55. Papandreou, G.; Kokkinos, I.; Savalle, P.-A. Modeling local and global deformations in deep learning: Epitomic convolution, multiple instance learning, and sliding window detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 390–399.

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.