remote sensing

MDPI

Article

U-Net Ensemble for Enhanced Semantic Segmentation in Remote Sensing Imagery

Ivica Dimitrovski *, Vlatko Spasev, Suzana Loshkovska and Ivan Kitanovski

Faculty of Computer Science and Engineering, University Ss Cyril and Methodius, 1000 Skopje, North Macedonia; vlatko.spasev@finki.ukim.mk (V.S.); suzana.loshkovska@finki.ukim.mk (S.L.); ivan.kitanovski@finki.ukim.mk (I.K.)

* Correspondence: ivica.dimitrovski@finki.ukim.mk

Abstract: Semantic segmentation of remote sensing imagery stands as a fundamental task within the domains of both remote sensing and computer vision. Its objective is to generate a comprehensive pixel-wise segmentation map of an image, assigning a specific label to each pixel. This facilitates in-depth analysis and comprehension of the Earth's surface. In this paper, we propose an approach for enhancing semantic segmentation performance by employing an ensemble of U-Net models with three different backbone networks: Multi-Axis Vision Transformer, ConvFormer, and EfficientNet. The final segmentation maps are generated through a geometric mean ensemble method, leveraging the diverse representations learned by each backbone network. The effectiveness of the base U-Net models and the proposed ensemble is evaluated on multiple datasets commonly used for semantic segmentation tasks in remote sensing imagery, including LandCover.ai, LoveDA, INRIA, UAVid, and ISPRS Potsdam datasets. Our experimental results demonstrate that the proposed approach achieves state-of-the-art performance, showcasing its effectiveness and robustness in accurately capturing the semantic information embedded within remote sensing images.

Keywords: remote sensing imagery; U-Net; ensemble learning; semantic segmentation; land cover

check for updates

Citation: Dimitrovski, I.; Spasev, V.; Loshkovska, S.; Kitanovski, I. U-Net Ensemble for Enhanced Semantic Segmentation in Remote Sensing Imagery. *Remote Sens.* **2024**, *16*, 2077. https://doi.org/10.3390/rs16122077

Academic Editors: Benoit Vozel, Orhan Altan, Min Huang, Changjiang Xiao, Nengcheng Chen and Runze Li

Received: 5 April 2024
Revised: 31 May 2024
Accepted: 4 June 2024
Published: 8 June 2024

1. Introduction

Remote sensing images encompass imagery captured from a distance by sensors or instruments mounted on various platforms such as satellites, aircraft, drones, and other vehicles. These images serve to gather information about the Earth's surface, atmosphere, and other objects or phenomena without requiring direct physical contact [1]. There are two main types of remote sensing images: aerial and satellite. Both aerial and satellite images are valuable sources of data, but they differ in how they collect data and their characteristics [2]. Satellites travel around Earth, collecting data from large areas at regular intervals. This provides a broad view of the entire planet. Aerial images are captured from airplanes or drones flying closer to the ground. They cover smaller areas but with much finer detail, making them ideal for studying specific locations. The choice between aerial and satellite imagery depends on the specific needs of the application at hand, such as the level of detail required and the availability of data [3].

The recent surge in sophisticated machine learning techniques, coupled with the ever-growing availability of remote sensing data, has significantly empowered image analysis and interpretation [4–7]. Semantic segmentation, a technique that assigns a specific class/annotation/label to each pixel in an image, has become a major research focus for remote sensing imagery [3]. This approach allows for a highly detailed analysis of ground objects and their spatial relationships. Unlike object detection, which focuses on identifying and roughly locating objects, semantic segmentation provides fine-grained pixel-level identification, enabling a deeper understanding of the image's content [8–10]. Semantic segmentation in remote sensing has far-reaching implications across various domains, such as urban planning [11–13], disaster management [14–16], environmental monitoring, and precision agriculture [17–20].

The emergence of deep learning, particularly convolutional neural networks (CNNs) and fully convolutional networks (FCNs), revolutionized the field of semantic segmentation by enabling the automatic learning of hierarchical representations from data [21]. Methods based on FCNs in combination with encoder–decoder architectures have become the dominant approach for semantic segmentation. Early methods utilized a series of successive convolutions followed by spatial pooling to achieve dense predictions [22]. Subsequent approaches, for example, U-Net [23] and SegNet [24], employed upsampling of high-level feature maps combined with low-level ones during decoding, aiming to capture global context and restore precise object boundaries. To enhance the receptive field of convolutions in initial layers, various techniques such as DeepLab [25] introduced dilated or atrous convolutions. Subsequent advancements integrated spatial pyramid pooling to capture multiscale contextual details in higher layers, as seen in models like PSPNet [26] and UperNet [27]. Incorporating these advancements, including atrous spatial pyramid pooling, DeepLabV3+ introduced a straightforward yet efficient encoder–decoder FCN architecture [28]. Subsequent developments, such as those seen in PSANet [29] and DRANet [30], replaced traditional pooling with attention mechanisms atop encoder feature maps to better grasp long-range dependencies.

Most recently, the adoption of transformer architectures, which utilize self-attention mechanisms and capture long-range dependencies, has marked additional advancement in semantic segmentation [31]. Transformer encoder–decoder architectures like Segmenter [32], SegFormer [33], and MaskFormer [34] harness transformers to enhance performance. Segmenter, a transformer encoder–decoder architecture designed for semantic image segmentation, utilizes a vision transformer (ViT) backbone and incorporates a mask decoder. SegFormer is a semantic segmentation framework that combines mix transformer encoders with lightweight multilayer perceptron (MLP) decoders, offering a simple, efficient, yet powerful solution. MaskFormer is a versatile semantic segmentation architecture inspired by DEtection TRansformer (DETR) [35]. It employs a transformer decoder to generate masks for individual objects in an image, utilizing a single decoder for various segmentation tasks. In response to the constraints of MaskFormer, Mask2Former was developed, featuring a multiscale decoder and a masked attention mechanism [36].

Despite the advancements in deep learning and semantic segmentation techniques, accurately segmenting remote sensing images remains a challenging task due to factors like complex spatial structures, diverse object sizes, data imbalance, and high background complexity [3]. Remote sensing images can have a wide range of resolutions and object orientations, posing challenges for consistent segmentation. Objects within a single image can vary dramatically in size. Buildings might stand tall next to tiny patches of vegetation. Effective models need to handle this wide spectrum of scales and accurately segment objects regardless of their relative size. Densely packed objects (like urban buildings) or very small objects (like individual trees) can be difficult for models to identify and segment accurately. Certain classes within a remote sensing image might be under-represented compared to others. For example, rare land cover types might have far fewer pixels compared to common vegetation classes. This data imbalance can make it difficult for models to learn accurate representations of all classes and lead to biased segmentation results. The foreground (objects of interest) might occupy a much smaller area compared to the background, making segmentation lopsided. Moreover, backgrounds in aerial images can be intricate and cluttered (e.g., urban environments), further complicating object segmentation [37].

In the context of semantic segmentation of remote sensing images, several approaches were proposed to overcome these challenges [38–40]. AerialFormer utilizes a hierarchical structure, in which a transformer encoder generates multiscale features, while a multidilated convolutional neural network (MD-CNN) decoder aggregates information from these diverse inputs [38]. The UNetFormer model incorporates a global–local transformer block

(GLTB) within its decoder to construct the output [39]. Additionally, it utilizes a feature refinement head (FRH) for optimizing the extracted global–local context. For efficiency, the transformer decoder is combined with a lightweight CNN-based encoder. Uncertainty-Aware Network (UANet) introduces the concept of uncertainty and proposes a novel network that leverages this concept [40]. UANet improves the accuracy of distinguishing building footprints from the complex distribution of surrounding ground objects. These advancements demonstrate the ongoing efforts to improve the semantic segmentation of remote sensing images through innovative deep learning architectures.

This paper examines the impact of incorporating powerful backbones into the U-Net architecture to potentially improve the semantic segmentation of remote sensing imagery. We leverage three such backbone networks: Multi-Axis Vision Transformer (MaxViT), ConvFormer, and EfficientNet. MaxViT is a hybrid vision transformer architecture, ConvFormer utilizes convolutional layers within a transformer framework, and EfficientNet is a CNN architecture known for its balance between accuracy and computational cost. We further enhance the performance by employing an ensemble learning approach. The diverse backbones within the ensemble extract complementary features from the images, leading to a richer and more comprehensive understanding of the scene. The ensemble benefits from the unique representations learned by each backbone within the U-Net architecture, leading to more robust and accurate segmentation results. To combine the base predictions, we employ a geometric mean ensemble strategy. To this end, we investigate the potential of combining backbones with different strengths to improve U-Net's performance for remote sensing semantic segmentation tasks. By exploiting the diversity and utilizing a geometric mean ensemble strategy, we managed to achieve state-of-the-art performance over several semantic segmentation datasets for remote sensing imagery. Our research can be summarized by the following primary contributions:

- Integration of three strong backbone networks within U-Net: the Multi-Axis Vision Transformer (MaxViT), ConvFormer, and EfficientNet. These models exhibit exceptionally high performance, surpassing previous models with similar model sizes.
- Introduction of an ensemble learning approach that leverages the complementary strengths of each backbone, tailored to enhance semantic segmentation of remote sensing imagery. Limiting the ensemble to three base models provides a good balance between performance gains and computational cost.
- Conducting of a comprehensive comparison of our approach with existing methods on various remote sensing image datasets. The experimental results show the superior performance of our models. Visual results further validate the effectiveness of our approach by showcasing accurate segmentation maps in remote sensing images.

The subsequent sections of this paper are organized as follows: In Section 2, we introduce the evaluation datasets, outlining their main characteristics and preprocessing steps. This section also provides insights into the selected backbones and how they are integrated into the ensemble. Additionally, Section 2 elaborates on the experimental setup employed for conducting the experiments. Moving on to Section 3, we present and summarize the obtained results, with detailed analyses and discussions provided. Finally, Section 4 offers concluding remarks.

2. Materials and Methods

2.1. Data Description

Remote sensing has become a powerful tool for analyzing and understanding the Earth's surface. Semantic segmentation datasets play a crucial role in this field, providing high-resolution satellite and aerial imagery alongside detailed pixel-level annotations [41]. By leveraging these datasets, researchers can train machine learning models to automatically extract meaningful information from vast amounts of remote sensing imagery. In our experimental study, we consider five datasets for semantic segmentation of remote sensing images: LoveDA [42], LandCover.ai [12], UAVid [43], INRIA [11], and ISPRS Potsdam, https://www.isprs.org/education/benchmarks/UrbanSemLab/semantic-labeling.aspx (accessed on 5 June 2024). These datasets present a diverse landscape, with variations in the amount of imagery, types of scenes captured, image resolution, and file formats. This diversity can be both an advantage and a challenge for researchers.

The LoveDA (Land-cOVEr Domain Adaptive) semantic segmentation dataset, https://github.com/Junjue-Wang/LoveDA (accessed on 5 June 2024), consists of 5987 high-resolution images sourced from Google Earth at 0.3 m spatial resolution and a pixel resolution of 1024 × 1024 divided into training, validation, and test sets by the creators [42]. Each pixel is annotated with one of the following labels: background, building, road, water, barren, forest, and agriculture. Training and validation images and ground-truth masks are downloadable, while for the test set, masks are withheld, prompting participants to submit predictions to the LoveDA Semantic Segmentation Challenge on CodaLab, https://codalab.lisn.upsaclay.fr/competitions/421 (accessed on 5 June 2024).

The UAVid dataset, https://uavid.nl (accessed on 5 June 2024), contains 420 high-resolution images captured by unmanned aerial vehicles (UAVs), each measuring 4096 × 2160 or 3840 × 2160 pixels [43], with 200 images for training, 70 for validation, and the rest for testing. UAVid features eight classes, including buildings, roads, static and moving cars, trees, low vegetation, humans, and background clutter. The large original images in the UAVid dataset were preprocessed using 512-pixel clips with a 256-pixel stride, ensuring full coverage. This resulted in 8000 training images and 2800 validation images, all at 512 × 512 pixels resolution. Test images remained unchanged.

The LandCover.ai (Land Cover from Aerial Imagery) dataset, https://landcover.ai.linuxpolska.com/download/landcover.ai.v1.zip (accessed on 5 June 2024), supports automated mapping of land features like buildings, woodlands, water, and roads using aerial imagery from Poland [12]. It includes 41 orthophoto tiles covering 216.27 km^2 with 25–50 cm per-pixel resolution. The dataset is divided into 512 × 512 pixel tiles and split into 70% for training, 15% for validation, and 15% for testing. The label distribution is skewed, with background and woodlands dominating.

The Potsdam dataset features high-resolution aerial images spanning Potsdam City, Germany. With 38 images, each 6000 × 6000 pixels at 5 cm per pixel, it classifies land cover into six categories: impervious surface, building, low vegetation, tree, car, and clutter/background. Annotations are either eroded (without boundaries) or non-eroded (with boundaries), with eroded used for evaluation. The dataset is split into 24 training and 14 testing images. The original image tiles were cropped into smaller images (512 × 512 pixels) with an overlap of 256 pixels, resulting in 3456 training images (2790 training, 666 validation) and 2016 test images. Performance metrics are provided for two scenarios: with and without clutter, aligning with previous studies [38,39].

The INRIA Aerial Image Labeling (INRIA) dataset is crafted for semantic segmentation tasks with aerial imagery, providing annotations for buildings and non-buildings [11]. The dataset comprises 180 training images with corresponding masks. About 16% of the pixels are labeled as buildings, while the rest are labeled as background. Each image in the dataset measures 5000 × 5000 pixels at a resolution of 30 cm. Test set images match the training set in size but lack publicly available ground-truth labels. However, predictions for the test set can be submitted for evaluation via the dataset creators' contest platform, https://project.inria.fr/aerialimagelabeling/contest/ (accessed on 5 June 2024), with results displayed on a public leaderboard. To manage the large images in the INRIA dataset, we preprocessed them into 512-pixel clips with a 256-pixel stride, ensuring full coverage. This process generated 18,000 images from the training set, with 15,500 for training and 2500 for validation, all at 512 × 512 resolution. Test images remained unchanged. The statistics for each dataset, including the number of images per training, validation, and test sets, image size, and number of labels present in each dataset, are given in Table 1.

Table 1. Statistics for each dataset including the number of images per set, image size, and number of labels present in each dataset.

Dataset		**LoveDA**	**LandCover.ai**	**UAVid**	**Inria**	**Potsdam**
No. of images	train	2522	7470	8000	15,500	2790
	val	1669	1602	2800	2500	666
	test	1796	1602	150	180	2016
image size	train	1024 × 1024	512 × 512	512 × 512	512 × 512	512 × 512
	val	1024 × 1024	512 × 512	512 × 512	512 × 512	512 × 512
	test	1024 × 1024	512 × 512	4096 (3840) × 2160	5000 × 5000	512 × 512
No. of labels		7	5	8	2	6

2.2. Model Architecture

U-Net is a convolutional neural network architecture designed for semantic segmentation tasks, particularly in medical image analysis [23]. U-Net is characterized by a U-shaped architecture that consists of a contracting path for feature extraction (encoder) followed by an expansive path for precise localization (decoder) as depicted in Figure 1.

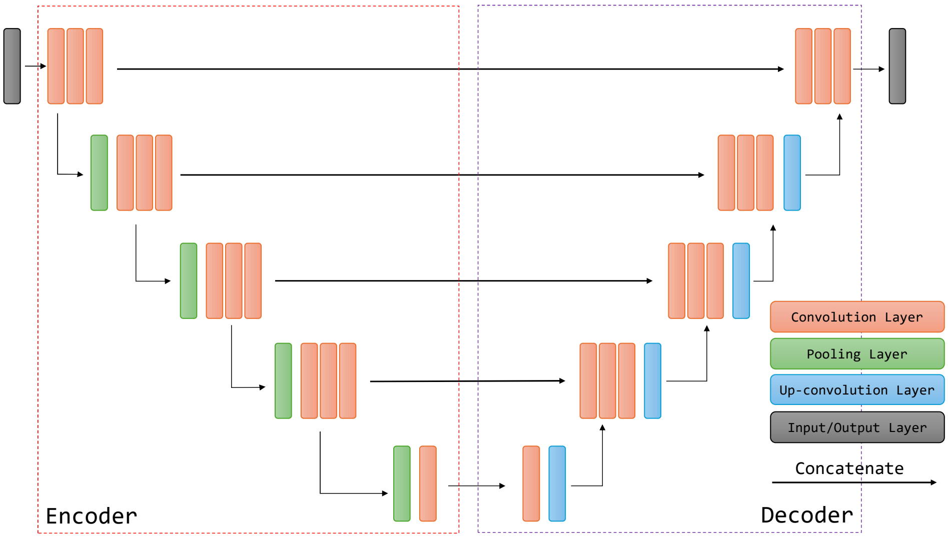

Figure 1. Illustration of the U-Net architecture. The displayed U-Net is an encoder–decoder network with a contracting path (encoding part, left side) that reduces the height and width of the input images and an expansive path (decoding part, right side) that recovers the original dimensions of the input images.

The encoder processes the input image and extracts features by using repeated stacks of convolutional layers with pooling operations (like max pooling). Pooling operations reduce the spatial resolution of the image while capturing higher-level features. Each stack typically increases the number of filters, allowing for learning more complex features. The decoder aims to recover the spatial resolution while preserving the extracted features by using upsampling operations (like transposed convolution) to increase the resolution. Each upsampling step combines the upsampled feature map with a corresponding feature map from the contracting path (via skip connections). Skip connections directly provide detailed information from the earlier stages, helping the decoder accurately localize features. The decoder culminates in a final convolutional layer with several filters equivalent to the predefined number of segmentation classes. This layer effectively acts as a classifier, generating a probability map as its output. Each pixel value within this map represents the probability of that specific pixel belonging to a particular class.

The combination of contracting and expanding paths with skip connections allows U-Net to learn both high-level semantic features and low-level spatial details crucial for accurate segmentation. U-Net architecture is known to perform well even with limited training data compared to other segmentation models [23]. This is partly due to the effective use of skip connections. U-Net can be adapted to various segmentation tasks by changing the number of classes and the final layer configuration. Overall, U-Net's U-shaped structure with its contracting and expanding paths makes it a powerful and widely used architecture for semantic segmentation tasks.

The U-Net architecture is flexible and allows for the incorporation of different encoders as feature extractors/backbones. One of the strengths of U-Net is its adaptability to various encoder structures. The original U-Net architecture is often modified by replacing the default encoder with pretrained models like VGG and ResNet, depending on the specific requirements of the tasks or datasets at hand. In this study, three distinct backbone networks are utilized: Multi-Axis Vision Transformer (MaxViT) [44], ConvFormer [45], and EfficientNet [46].

2.2.1. Multi-Axis Vision Transformer

MaxViT is a recent advancement in vision transformer architectures that addresses the challenge of capturing both local and global information within an image [44]. MaxViT utilizes a hierarchical architecture where each stage in the hierarchy consists of a MaxViT block, which combines a multi-axis self-attention (Max-SA) block with a convolutional layer. This combination leverages the strengths of both approaches: Max-SA for global context and convolutions for efficient local feature extraction. The use of Max-SA makes MaxViT computationally efficient compared to full self-attention in standard ViTs [47].

The architecture of the MaxViT network is illustrated in Figure 2. The network begins by downsampling the input through Conv3x3 layers in the stem stage (S0). The body of the network contains four stages (S1–S4), with each stage having half the resolution of the previous one with a doubled number of channels (hidden dimensions). The MaxViT model can be scaled up by increasing the number of blocks per stage and the channel dimension. There are several MaxViT variants including MaxViT-T, MaxViT-S, MaxViT-B, MaxViT-L, and MaxViT-XL. These variants progressively increase in complexity (number of blocks and channels) and likely performance, potentially reaching a trade-off between accuracy and efficiency [44]. In this study, we use MaxViT-S as an encoder in the U-Net architecture. MaxViT's architecture allows for easy scaling to handle large datasets and complex tasks. By combining efficient global attention with the ability to capture local details, MaxViT offers a promising approach for various computer vision applications, including the task of semantic segmentation explored in this work.

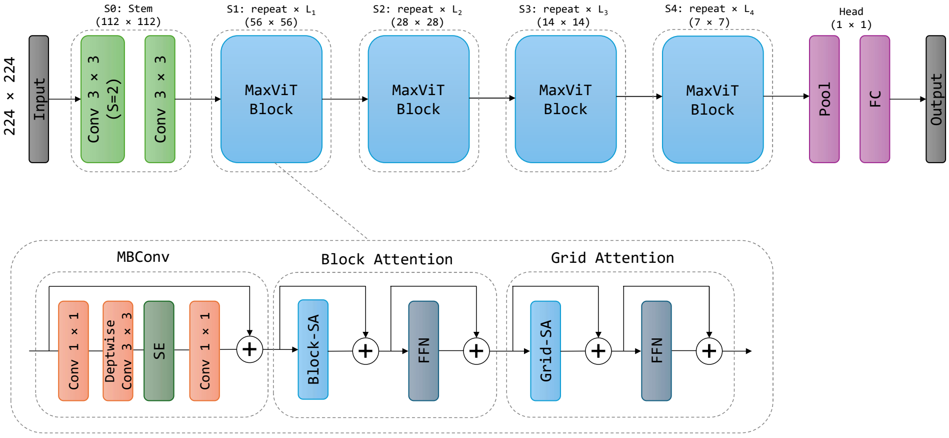

Figure 2. MaxViT architecture with hierarchical design and basic building block that unifies MBConv, block, and grid attention layers.

2.2.2. ConvFormer

The MetaFormer concept is not tied to a particular model; rather, it is a generalized architecture abstracted from the transformer, omitting the specification of token mixers [48].

The instantiation of MetaFormer into specific models occurs by specifying concrete token mixers as modules. One way to instantiate the token mixer within the MetaFormer is by using separable depthwise convolutions. This specific implementation results in a model called ConvFormer [45]. ConvFormer leverages the MetaFormer structure for efficient feature processing but relies solely on convolutional operations, making it functionally a type of CNN. ConvFormer adopts a hierarchical architecture of 4 stages [49], as illustrated in Figure 3. Based on the number of channels and the number of blocks, several model configurations of different sizes are defined: ConvFormer-S12, ConvFormer-S24, ConvFormer-S36, ConvFormer-M36, and ConvFormer-M48. In this study, we adopt the ConvFormer-M36 as an encoder in the U-Net architecture. The notation "M36" means the model is of medium size of embedding dimensions (the total number of channels in all four stages) with 36 ConvFormer blocks in total. ConvFormer demonstrates superior performance compared to the robust CNN model ConvNeXt [50] and achieves comparable accuracy to another powerful CNN model, EfficientNetV2-L [51]. Additionally, ConvFormer surpasses several strong attention-based or hybrid models. Notably, ConvFormer-M36 outperforms Swin-B [52] and CoAtNet-2 [53] while utilizing fewer parameters.

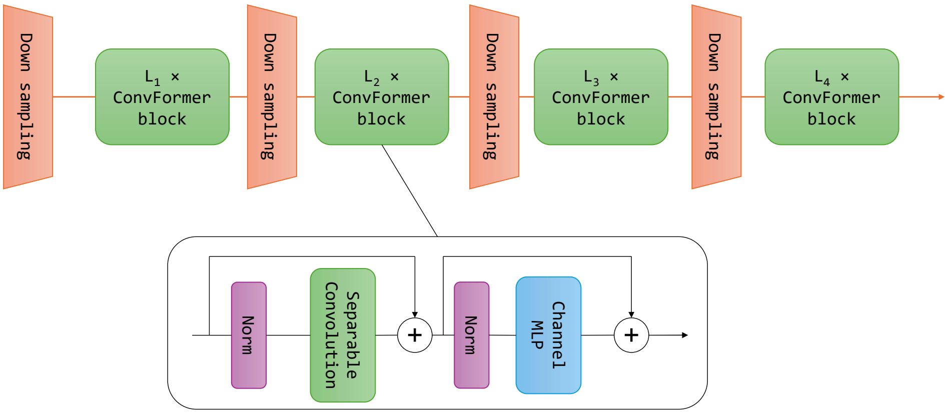

Figure 3. Overall framework of ConvFormer and architecture of the ConvFormer block, which has a token mixer of separable depthwise convolutions.

2.2.3. EfficientNet

EfficientNet is a family of convolutional neural network (CNN) architectures designed to achieve state-of-the-art accuracy while maintaining computational efficiency [46]. It addresses the challenge of finding an optimal balance between three key factors influencing model capacity and performance: depth, width, and resolution. EfficientNet proposes a novel scaling method that systematically balances these three factors. It utilizes a compound coefficient that uniformly scales depth, width, and resolution in a principled way. This approach ensures efficient resource allocation and avoids overemphasizing any single factor. EfficientNet models have achieved remarkable accuracy on various image classification tasks, often surpassing previous CNN architectures [5]. Several EfficientNet models can be scaled from the baseline EfficientNet-B0 model using different compound coefficients [46], offering a good trade-off between accuracy and resource consumption. Variants like B1 to B7 progressively increase in complexity. The EfficientNet architecture relies on the MBConv layer, which is an inverted residual block. Figure 4 illustrates the basic building blocks of EfficientNet-B0, highlighting the MBConv layers.

The specific choice of the EfficientNet variant depends on the task at hand and the available computational resources. By offering a spectrum of complexity levels, EfficientNet provides a versatile set of CNN architectures for various image recognition applications [5,46]. In this study, we adopt EfficientNet-B7 as an encoder in the U-Net architecture.

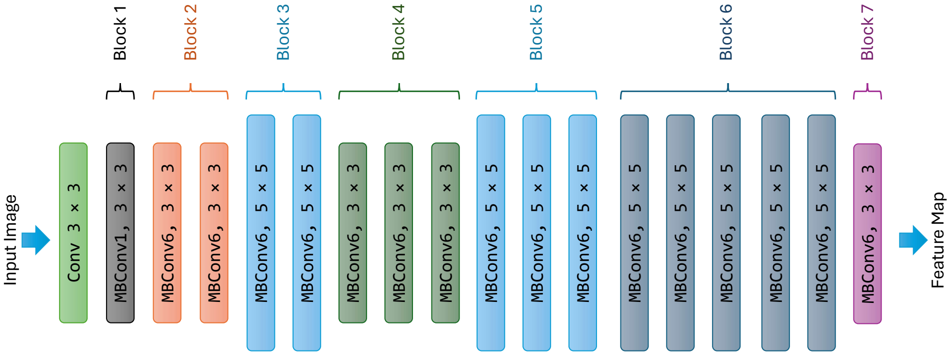

Figure 4. Architecture of EfficientNet-B0 with MBConv as basic building blocks.

2.2.4. Ensemble Method

Ensemble learning is a powerful technique in machine learning that involves combining the predictions of multiple models to achieve better performance than any single model could on its own [54]. Ensemble methods can help reduce the variance of the predictions, making the model less prone to overfitting or poor performance on unseen data. Ensembles are also less susceptible to errors or biases present in any single base model. Ensemble learning offers a valuable approach to enhancing the performance and robustness of deep learning models. Strategically combining multiple models has the potential to yield superior results across a range of deep learning tasks, such as image classification [55], semantic segmentation [56,57], and object detection [58,59]. The ensemble approach in the context of deep learning starts by training multiple deep learning models, often referred to as base models. These base models can have identical architectures trained with different initial weights or hyperparameters to promote diversity in the learning process, or they may have different architectures altogether to leverage the strengths of diverse approaches to feature extraction and representation. Once trained, the predictions from each base model are combined in some way to generate the final ensemble predictions. Common combination strategies are averaging, weighted averaging, and voting [60].

This paper investigates the potential of ensemble learning for semantic segmentation tasks in remote sensing imagery. We built the ensemble by combining base models as diversely as possible. The key strength lies in the diversity of the employed backbones in the base models. The ensemble utilizes U-Net architecture with three distinct backbone networks: Multi-Axis Vision Transformer (MaxViT), ConvFormer, and EfficientNet. The full setup is illustrated in Figure 5. Each backbone network possesses unique advantages. For example, MaxViT is a hybrid vision transformer architecture, ConvFormer utilizes convolutional layers within a transformer framework, and EfficientNet is a CNN architecture known for its balance between accuracy and efficiency. This diversity in the feature extraction process allows the ensemble to capture a richer and more comprehensive understanding of the image, potentially leading to improved segmentation performance and more robust and accurate segmentation results.

To generate the final segmentation maps, we employ a geometric mean ensemble strategy, assuming the ensemble has k (in our experiments, $k = 3$) base models, denoted by $[M_1(x), ..., M_k(x)]$, where x is the input image. Each model outputs a raw score vector, denoted as $[z_1, ..., z_k]$. We apply the $softmax$ function to each raw score vector z_i to convert them into probability distributions (typically between 0 and 1) over the class labels as expressed in Formula (1).

$$p_i(x) = softmax(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{k} e^{z_j}} \tag{1}$$

where $p_i(x)$ represents the probability vector for the i-th base model given the input image x. The final probability scores, $E(x)$, are obtained by combining the probability scores for each base model using the geometric mean function as in Formula (2). The segmentation map is calculated by applying the *argmax* function on the ensemble probability scores $E(x)$.

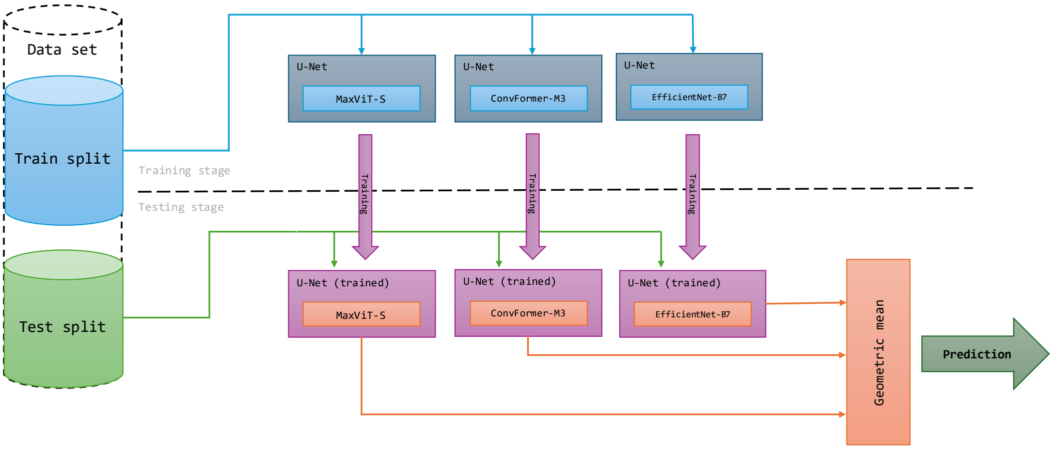

Figure 5. Geometric mean ensemble training and testing strategy of three base models: MaxViT-S, ConvFormer-M36, and EfficientNet-B7.

$$E(x) = \left(\prod_{i=1}^{k} p_i(x) \right)^{\frac{1}{k}} \tag{2}$$

In essence, this work explores the potential of ensemble learning to enhance the capabilities of U-Net architecture for semantic segmentation tasks in remote sensing imagery. By utilizing a diverse set of backbones and a geometric mean ensemble strategy, we aim to achieve superior segmentation performance compared to existing/published methods.

2.3. Experimental Setup

This study investigates the performance of three U-Net variations for semantic segmentation of remote sensing images. Each variant utilizes different pretrained encoders: MaxViT-S, ConvFormer-M36, and EfficientNet-B7. The encoders are pretrained on the ImageNet-1K dataset, which helps to learn general visual features that can be beneficial for semantic segmentation tasks. Additionally, we assess the effectiveness of an ensemble approach, where the predictions from base models are combined for improved performance. In this ensemble, the final prediction for each pixel is derived by calculating the geometric mean of the probabilities predicted by each U-Net variant.

The models were trained using the training set, with parameter selection/search performed using the validation set. To overcome overfitting, we performed early stopping on the validation set if no improvements in the validation loss were observed over 20 epochs. The best checkpoint/model found (with the highest evaluation metric) was saved and then applied to the original test set to obtain the final assessment of the predictive performance. The maximum number of epochs was set to 100.

In our experiments, we used a fixed value for the batch size, set to 12. For optimization, we employed the AdamW optimizer and a learning rate set to 0.0001 [61]. Next, we used a learning rate scheduler that used a polynomial decay schedule. It is commonly observed that a monotonically decreasing learning rate, whose degree of change is carefully chosen, results in a better performing model. This schedule applied a polynomial decay function to an optimizer step, given a provided initial_learning_rate (1×10^{-4}), to reach an end_learning_rate (1×10^{-7}) in the given decay_steps. We used a combination of cross-entropy loss and dice loss to achieve a more balanced approach to semantic segmentation. Cross-entropy loss measures the pixel-wise similarity between the predicted and ground-truth masks, while the dice loss ensures accurate boundary localization. The combination of these loss functions is effective in achieving both accurate object localization and overall segmentation accuracy.

To further enhance the model's robustness, we incorporated data augmentation techniques during training. The process involved random horizontal and vertical flips, followed by random rotation and random changes in the brightness and contrast. Additionally, we randomly selected one of the following transformations to apply: contrast limited adaptive histogram equalization to the input image, grid distortion, or optical distortion [62]. Finally, normalization was applied using the mean and standard deviation derived from the ImageNet-1k dataset. The augmentation strategy for the UAVid dataset differed, as it excluded vertical flipping and random rotation techniques. For the LoveDA dataset, we first utilized the RandomCrop transformation to randomly extract smaller image patches from the original images [62]. Specifically, we cropped 512 × 512 pixel regions from the original LoveDA images, which have a larger dimension of 1024 × 1024 pixels. We ensured that all datasets had a consistent input resolution of 512 × 512 pixels for training the models. However, during the evaluation of validation or test data, the input images were only normalized.

The inference approach used depends on the image size of the test set within each dataset. The inference for the LoveDA, LandCover.ai, and ISPRS Potsdam datasets was directly applied to the entire image due to their relatively small resolution. For the INRIA and UAVid datasets, because of the larger test images, we employed a sliding-window inference technique [63]. This method partitions the image into smaller overlapping patches, processes each patch with the model, and then assembles the predictions to create the final segmentation map. The sliding window was configured with dimensions of 1024 pixels, and the overlap was set to 128 pixels. For areas covered by multiple sliding windows, the final prediction was determined by averaging the predictions from each window. This distinction ensured efficient processing for smaller images while effectively handling larger ones with overlapping patches to capture the entire scene. Moreover, we implemented test-time augmentation (TTA) through horizontal flipping and rotation. The final prediction was computed as the mean of predictions obtained from both the original and augmented images.

In terms of evaluation measures for predictive performance, we report the label-wise intersection over union (IoU) as an evaluation measure, which denotes the area of the overlap between the ground truth and predicted label divided by the total area. We also report the mean intersection over union ($mIoU$) averaged across the different labels. All models were trained on NVIDIA A100-PCIe GPUs with 40 GB of memory running CUDA version 11.5. We configured and ran the experiments using the deep learning framework PyTorch Lightning [64].

3. Results and Discussion

The summarized results of the base U-Net models and the ensemble models across the different evaluation datasets are shown in Table 2. We report the mean intersection over union (%) and also report the rank of the models, averaged over the respective datasets. For the INRIA dataset, the provided metric is the intersection over union (%) specifically for the building label, which is a common practice in building extraction [40]. Notably, the U-Net ensemble models ranked the best overall and achieved the best performance on six (out of the six) tasks/datasets. Among the base models, those utilizing MaxViT-S as a backbone ranked second, closely followed by the U-Net models with EfficientNet-B7 as a backbone. While the U-Net models featuring the ConvFormer-M36 backbone ranked last, they demonstrated comparable performance across all datasets.

Table 2. Mean intersection over union (mIoU %) of the U-Net models across the different semantic segmentation datasets. For the INRIA dataset, the provided metric is the intersection over union (IoU) specifically for the building label. Bold indicates the best-performing model and underline second-best model for a given dataset. We report the average rank of a model (lower is better), ranked based on the performance and averaged across the datasets.

Dataset\Model	U-Net (MaxViT-S)	U-Net (ConvFormer-M36)	U-Net (EfficientNet-B7)	U-Net Ensemble
LoveDA	56.16	54.80	55.07	**57.36**
UAVid	71.88	70.79	71.42	**73.34**
LandCover.ai	87.41	87.64	87.06	**88.02**
Potsdam (without clutter)	88.17	89.45	88.59	**89.9**
Potsdam (with clutter)	79.7	79.77	79.84	**80.82**
INRIA	80.84	79.89	80.6	**81.43**
Avg. Rank	2.83	3.17	3.00	1.00

The ensemble models consistently outperform across all datasets. For the LoveDA dataset, the achieved mIoU stands at 57.36%, marking the best result and securing the top rank on the publicly available leaderboard of the LoveDA semantic segmentation challenge. Similarly, for the UAVid dataset, the mIoU reaches 73.34%, which also represents the best reported result to date for this dataset. For the LandCover.ai datasets, the ensemble surpasses all previously reported results, achieving mIoU values of 88.02%. We analyzed the segmentation performance on the Potsdam dataset in two cases, with and without clutter (background). The label clutter is the most challenging as it can contain anything except for the five main labels defined for this dataset. Our U-Net ensemble stands out as a robust segmentation model, achieving top performance in both scenarios, with and without clutter, boasting mIoU values of 80.82% and 89.9%, respectively. Moreover, the obtained IoU value of 81.43% for the INRIA dataset aligns with the best-performing methods and reported results observed on the contest platform hosted by the dataset creators. These results strongly validate the effectiveness of our proposed ensemble approach in enhancing semantic segmentation performance on remote sensing imagery. Table 3 presents the performance of various backbone combinations on the selected semantic segmentation datasets. While all combinations outperform the base models, the ensemble utilizing all three backbones (MaxViT, ConvFormer, EfficientNet) consistently achieves the best results.

Table 3. Mean intersection over union (mIoU %) for various combinations of backbones across different semantic segmentation datasets. For the INRIA dataset, the provided metric is the intersection over union (IoU) specifically for the building label.

Dataset\Model	MaxViT-S+ConvFormer-M36	MaxViT-S+EfficientNet-B7	ConvFormer-M36+EfficientNet-B7
LoveDA	56.98	57.19	56.11
UAVid	72.91	72.0	72.61
LandCover.ai	88.01	87.76	87.83
Potsdam (without clutter)	89.72	89.44	89.72
Potsdam (with clutter)	80.49	80.57	80.55
INRIA	81.15	81.4	80.96

Having established the overall effectiveness of U-Net models, both base models and ensemble models, we now examine how well they perform on each dataset. This analysis aims to identify the potential strengths and weaknesses of the models across different types of remote sensing imagery and segmentation tasks. Subsequently, we explore any notable observations or patterns based on the performance of each specific dataset. Our analysis includes per-label IoU values, confusion matrices, and sample inference masks for each dataset. These insights will aid in understanding performance fluctuations across different labels or regions within the images (e.g., urban versus rural areas) and pinpoint potential challenges such as data imbalances or complex object shapes. We also present a detailed comparison of our models performance against established methods/models for all considered remote sensing image datasets.

In Table 4, we report performance comparisons with existing methods on the test set of the LoveDA dataset. This comparison is based on the IoU metric calculated for each semantic class/label. Our models (base and ensemble) consistently outperform existing state-of-the-art methods. Notably, this includes competitive approaches like AerialFormer-B [38], DC-Swin [65], multitask pretraining using the InternImage-XL model [66], UperNet (Swin small) [27], and the foundation model trained with UperNet and a vision transformer as a backbone [67]. To explore the impact of encoder selection on performance, we compared our trained U-Net models to a variant equipped with a ResNet50 encoder. The U-Net model with a ResNet50 encoder achieves an mIoU value of 47.84%. Our U-Net models demonstrate significant improvements: the MaxViT backbone outperforms this result by 8.32% in mIoU, the ConvFormer by 6.96%, and EfficientNet by 7.23%. This demonstrates the effectiveness of our backbone selection.

The top-performing model is the U-Net ensemble, achieving an mIoU value of 57.36%, marking a substantial improvement of 2.81% over existing state-of-the-art methods. This result also secures the highest rank on the publicly available leaderboard of the LoveDA semantic segmentation challenge. Notably, the U-Net ensemble outperforms the existing methods by 2.33% IoU for the road label, 2.26% IoU for the barren label, 1.39% IoU for the forest label, 1.82% IoU for the agriculture label, 0.11% IoU for the water label, and 1.29% IoU for the background label. The U-Net ensemble model exhibits a lower IoU value only for the building label, with a decrease of 1.48% compared to existing methods.

Table 4. Performance comparison on LoveDA dataset between our U-Net models and other existing semantic segmentation approaches. We report the mean intersection over union (mIoU %) and intersection over union (IoU %) for each label. Bold indicates the best-performing model and underline the second-best model for a specific label.

Model	IoU per Label							mIoU
	Background	Building	Road	Water	Barren	Forest	Agriculture	
U-Net (ResNet50) [42]	43.06	52.74	52.78	73.08	10.33	43.05	59.87	47.84
HRNet (W32) [42]	44.61	55.34	57.42	73.96	11.07	45.25	60.88	49.79
DeepLabV3+ (ResNet50) [42]	42.97	50.88	52.02	74.36	10.4	44.21	58.53	47.62
AerialFormer-B [38]	47.8	60.7	59.3	81.5	17.9	47.9	64.0	54.1
SegFormer-B5 [33]	46.54	57.46	58.91	80.09	27.89	46.14	61.0	54.01
DC-Swin small [65]	45.9	57.97	61.38	80.64	24.15	46.59	60.33	53.85
UperNet (Swin small) [27]	46.18	59.97	57.4	81.2	26.71	47.21	63.18	54.55
UperNet (ViT-G12×4) [67]	47.57	61.6	59.91	81.79	18.6	47.3	64.0	54.4
UNetFormer (ResNet18) [39]	44.7	58.8	54.9	79.6	20.1	46.0	62.5	52.4
MTP (InternImage-XL) [66]	46.8	**62.6**	58.96	82.25	17.49	47.63	63.44	54.17
U-Net (MaxViT-S)	48.59	60.47	63.4	81.17	27.02	48.1	64.4	56.16
U-Net (ConvFormer-M36)	45.17	60.81	58.0	81.48	28.27	46.9	62.96	54.8
U-Net (EfficientNet-B7)	45.88	57.57	58.92	80.69	28.24	47.88	**66.31**	55.07
U-Net Ensemble	**49.09**	61.12	**63.71**	**82.36**	**30.15**	**49.29**	65.82	**57.36**

A confusion matrix was not calculated for this dataset as the ground-truth masks for the test images are not publicly accessible. Figure 6 displays sample images alongside the inference masks generated by the U-Net ensemble, the best-performing model, and other models for comparison. The mIoU value of 57.36% for this dataset indicates its high level of difficulty. The regions labeled as background exhibit significant intraclass variance due to the complexity of the scenes, leading to substantial false alarms. Identifying small-scale objects like buildings and scattered trees poses challenges. Additionally, distinguishing between forest and agricultural labels is difficult due to their similar spectra. However, the water label achieves the highest IoU value, suggesting good recognition performance for this label.

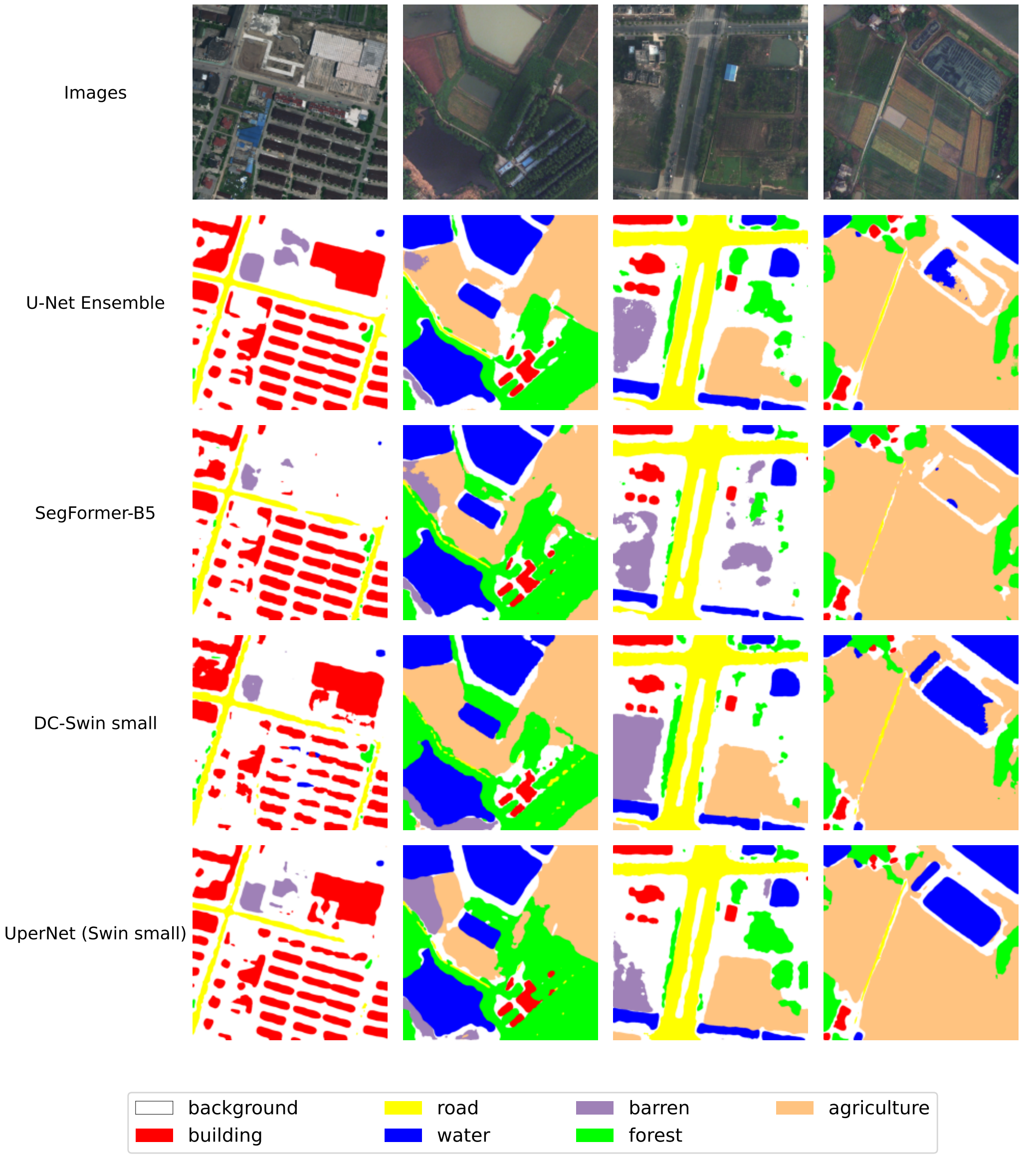

Figure 6. Example images and inference masks from LoveDA dataset. The first row displays example images. Each subsequent row shows the segmentation masks generated by a different model for the corresponding image in the first row.

Table 5 presents a performance comparison of the UAVid dataset test set, evaluated using the intersection over union (IoU) metric for each semantic class/label. Our ensemble model surpasses existing state-of-the-art methods such as EMNet [68], DCDNet [69], and UperNet (Swin small) [27], while our base U-Net models achieve competitive results. We investigated the influence of encoder selection on performance by comparing our U-Net models to a variant equipped with a ResNet50 encoder. This baseline U-Net achieved an mIoU value of 67.22%. However, even stronger results were obtained by employing more powerful backbones within the U-Net architecture. The MaxViT backbone significantly outperformed the ResNet50 by 4.66% in mIoU, while ConvFormer and EfficientNet also achieved improvements of 3.57% and 4.2%, respectively. These findings demonstrate the effectiveness of our backbone selection strategy in boosting U-Net's performance for remote sensing semantic segmentation tasks. Our top-performing U-Net ensemble achieves a state-of-the-art mIoU of 73.34%, exceeding the methods presented in Table 5 and the best results of 73.21% mIoU from the public UAVid competition, https://competitions.codalab.org/competitions/25224 (accessed on 5 June 2024).

Table 5. Performance comparison on UAVId dataset between our U-Net models and other existing semantic segmentation approaches. We report the mean intersection over union (mIoU %) and intersection over union (IoU %) for each label. Bold indicates the best-performing model and underline the second-best model for a specific label.

Model	**IoU per Label**								**mIoU**
	Clutter	**Building**	**Road**	**Tree**	**Vegetation**	**Mov. Car**	**Sta. Car**	**Human**	
U-Net (ResNet50) [23]	67.69	87.28	80.2	79.69	63.46	70.72	58.11	30.62	67.22
DeepLabV3+ (ResNet50) [25]	67.86	87.87	80.23	79.74	62.03	71.51	62.99	29.5	67.72
SegFormer-B5 [33]	70.21	88.41	82.54	80.81	64.54	76.38	66.31	32.61	70.23
EMNet [68]	73.27	**92.03**	**85.13**	81.05	64.73	73.32	67.01	**35.12**	71.46
CAGNet [70]	69.8	88.4	82.7	80.6	64.6	76.0	57.8	32.1	69.0
UNetFormer (ResNet18) [39]	68.4	87.4	81.5	80.2	63.5	73.6	56.4	31.0	67.8
DC-Swin small [65]	70.72	89.66	83.42	80.75	65.23	74.97	59.77	32.02	69.57
UperNet (Swin small) [27]	71.23	89.54	83.44	81.27	65.59	77.08	69.86	30.95	71.12
DCDNet [69]	72.11	90.56	83.63	82.15	66.52	77.68	74.7	31.69	72.38
U-Net (MaxViT-S)	71.45	89.44	81.93	81.52	66.39	77.6	73.33	33.4	71.88
U-Net (ConvFormer-M36)	71.34	89.65	82.33	81.1	64.71	75.16	69.88	32.1	70.79
U-Net (EfficientNet-B7)	71.15	90.19	81.80	81.75	66.17	77.25	71.22	31.84	71.42
U-Net Ensemble	**73.8**	91.05	84.05	**82.58**	**67.74**	**79.02**	**74.98**	33.49	**73.34**

Examining the IoU values, it becomes evident that the most accurate predictions are achieved for the building label. Moreover, satisfactory results are obtained for the road, tree, and moving car labels. The label human exhibits the lowest performance. The confusion matrix calculated for the U-Net ensemble as the best-performing model is visualized in Figure 7. The confusion matrix reveals the frequent misclassifications by the model. Notably, it tends to confuse the labels moving car and static car, as well as tree and low vegetation. This is rather expected given the semantic similarity between these labels. Additionally, there is a tendency to misclassify the human label as the low vegetation label, likely a result of pixel overlap between these two labels. There are misclassifications with the background clutter label for nearly all labels, as expected. This is because diverse objects and regions not belonging to any specific label are designated with the background clutter label.

Figure 8 shows example images and ground-truth masks from the UAVid dataset. Additionally, it presents the inference masks generated by the U-Net ensemble model. Despite the high scene complexity, characterized by the number of objects and varied object configurations in the UAVid dataset, the U-Net ensemble model demonstrates excellent performance. Figure 9 showcases a cropped region of an example image along with its corresponding ground-truth mask, featuring the labels human and moving cars. The predicted mask, shown as the last image in Figure 9, is generated by the U-Net ensemble model. A comparison between the ground-truth mask and the predicted mask reveals segments where the model misclassifies moving cars as static cars. In this particular scenario, distinguishing between these two labels is challenging because, for instance, the cars are stationary, waiting at the pedestrian crossing. Additionally, there is a lack of fine-grained segmentation for the human label due to overlapping and dense objects/segments marked with this label in the ground-truth mask.

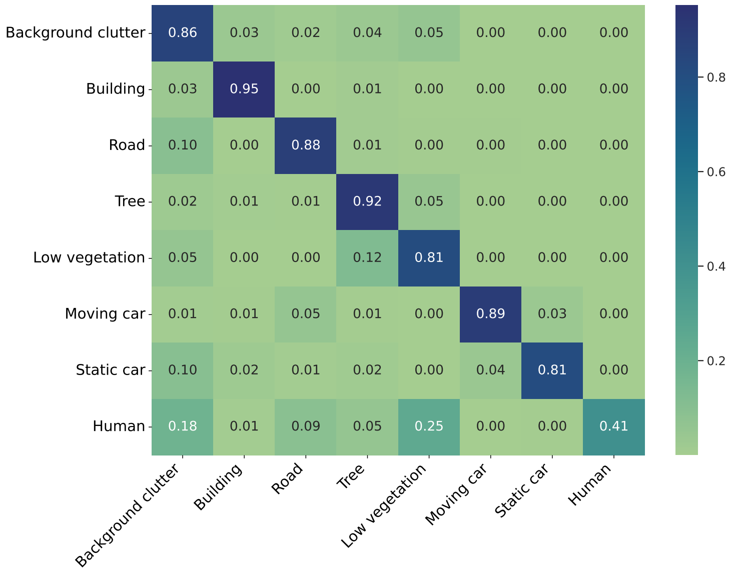

Figure 7. Confusion matrix for the U-Net ensemble model on the UAVid dataset.

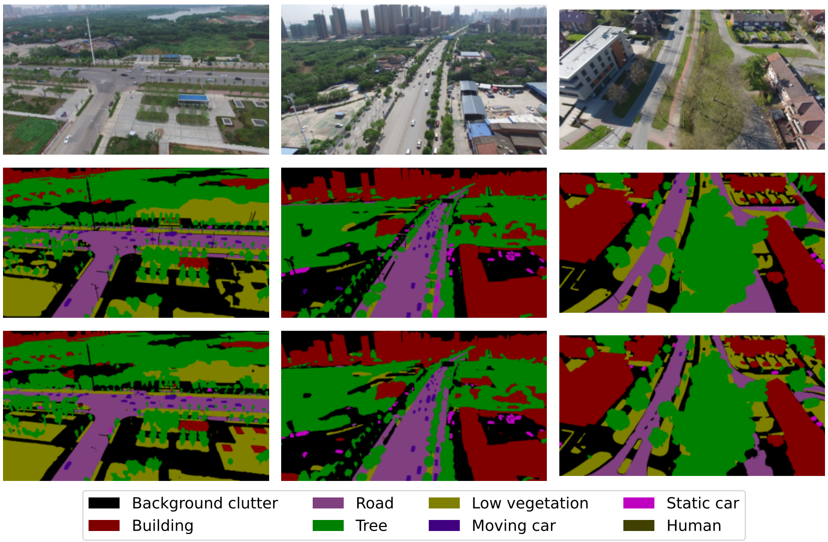

Figure 8. Example images, ground-truth masks, and inference masks from UAVid dataset. First row shows example images. Second row shows the corresponding ground-truth masks. Third row shows the prediction results of U-Net ensemble model as in Table 2.

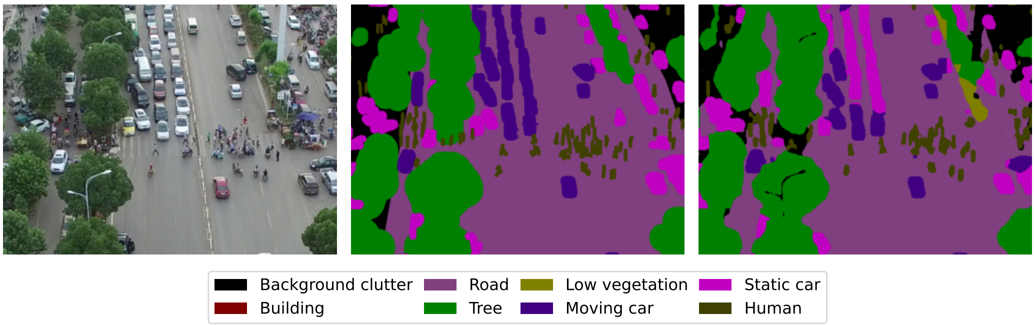

Figure 9. Cropped image, ground-truth mask, and predicted mask from the U-Net ensemble model, as outlined in Table 2. The images highlight a predominant region containing the labels human and moving cars from the UAVid dataset.

In Table 6, we report performance comparisons with existing methods on the test set of the LandCover.ai dataset. This comparison is based on the IoU metric calculated for each semantic class/label. Our models (base and ensemble) consistently outperform existing state-of-the-art methods. Notably, this includes competitive approaches like DC-Swin small [65], UperNet (Swin small), and SegFormer-B5 [33]. In the comparison for this dataset, we also included a U-Net model with ResNet50 as the backbone, achieving an mIoU value of 85.66%. Compared to this result, our U-Net models show notable improvements with different backbones: a 1.75% mIoU gain for MaxViT, a 1.98% gain for ConvFormer, and a 1.4% gain for EfficientNet. These results underscore the advantages of our chosen backbones. Our top-performing model is the U-Net ensemble, achieving an mIoU value of 88.02%, marking an improvement of 0.64% over existing state-of-the-art methods.

Table 6. Performance comparison on LandCover.ai dataset between our U-Net models and other existing semantic segmentation approaches. We report the mean intersection over union (mIoU %) and intersection over union (IoU %) for each label. Bold indicates the best-performing model and underline the second-best model for a specific label.

Model	**IoU per Label**					**mIoU**
	Background	**Buildings**	**Woodlands**	**Water**	**Road**	
U-Net (ResNet50) [23]	93.35	81.71	91.23	94.54	67.45	85.66
DeepLabV3+ (ResNet50) [25]	93.81	81.84	91.79	95.24	69.41	86.42
HRNet (W32) [71]	93.9	84.28	91.93	94.95	70.68	87.15
SegFormer-B5 [33]	94.08	83.58	92.1	**95.56**	70.9	87.24
UperNet (Swin small) [27]	94.12	83.68	92.13	95.51	71.44	87.38
UNetFormer (ResNet18) [39]	93.24	80.67	91.12	94.64	67.51	85.43
DC-Swin small [65]	94.19	82.47	92.31	95.52	70.91	87.08
U-Net (MaxViT-S)	93.93	84.51	91.89	95.13	71.6	87.41
U-Net (ConvFormer-M36)	94.18	84.51	92.22	95.33	71.93	87.64
U-Net (EfficientNet-B7)	93.7	84.11	91.66	94.94	70.9	87.06
U-Net Ensemble	**94.3**	**85.33**	**92.43**	95.42	**72.62**	**88.02**

The IoU values for the labels road and building are lower compared to the other labels. These labels are more challenging in the context of semantic segmentation because they are usually narrow, in the case of roads, or often small, in the case of buildings. Additionally, they are sometimes obscured by other objects, such as trees. Figure 10 illustrates the confusion matrix for the LandCover.ai dataset calculated for the U-Net ensemble model. The confusion matrix further supports this observation, indicating that the road label is frequently misclassified as woodlands, and in general, all labels exhibit confusion with the background label, as expected due to its diverse composition, which may include fields, grass, or pavement.

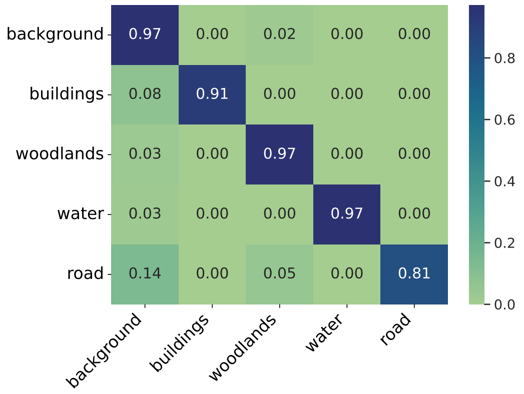

Figure 10. Confusion matrix for the U-Net ensemble model applied on the LandCover.ai dataset.

Figure 11 shows example images and ground-truth masks from the LandCover.ai dataset. Additionally, it displays the inference masks generated by the U-Net ensemble, the best-performing model, and other models for comparison. The identified regions within the inference masks typically exhibit smoother boundaries, aligning well with reality, particularly in the context of woodlands and water. However, this smoothness can lead to inaccuracies for buildings, potentially omitting smaller structures and producing slightly less defined/precise building edges/boundaries. Conversely, the model excels at identifying individual trees that might have been missed by human annotators.

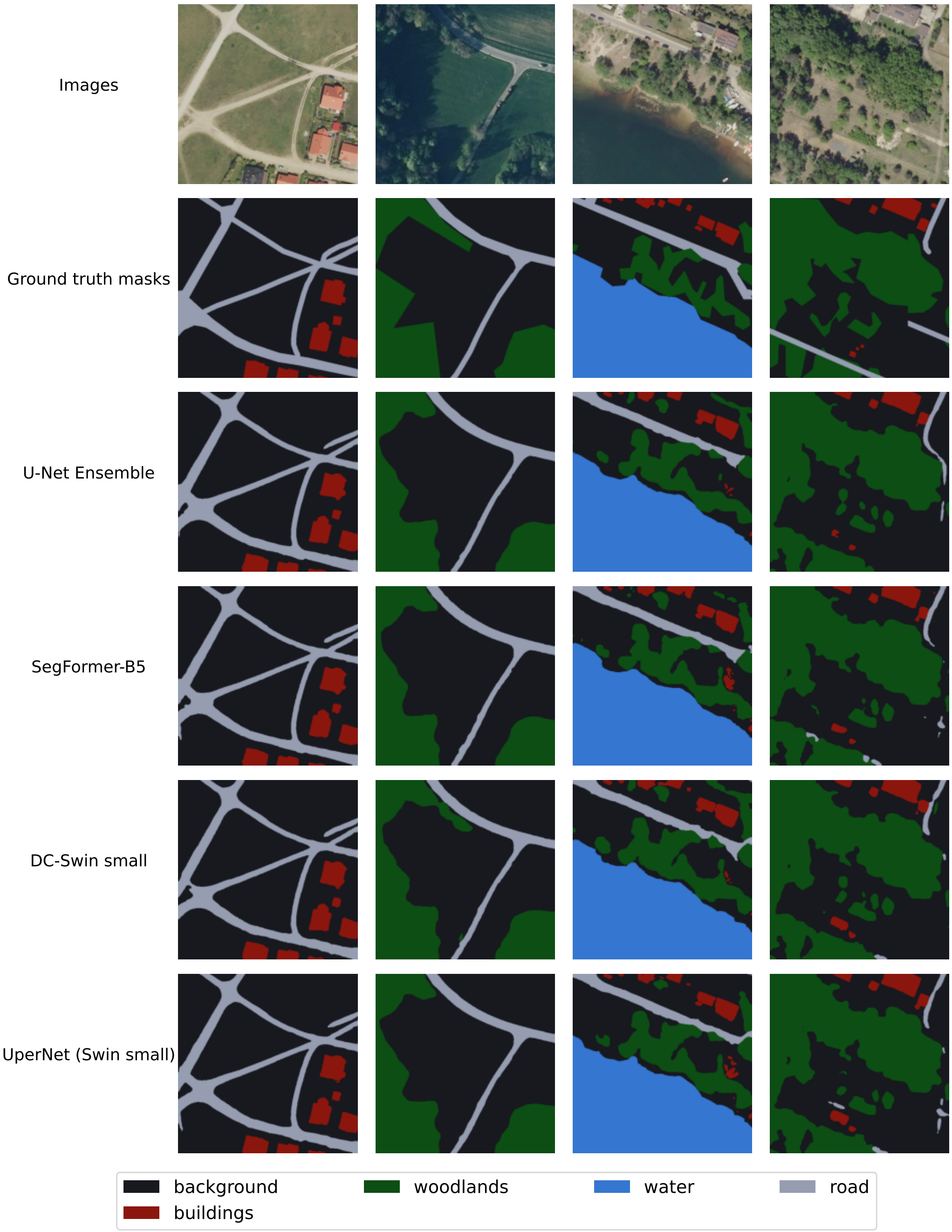

Figure 11. Example images, ground-truth masks, and inference masks from LandCover.ai dataset. The first row shows example images. The second row shows the corresponding ground-truth masks. Each subsequent row shows the segmentation masks generated by a different model for the corresponding image in the first row.

We evaluate the segmentation performance on the Potsdam dataset under two conditions: with and without clutter. For this dataset, we report per-label F1 scores, mean F1 scores, and mean intersection over union (IoU). We include F1 scores in this comparison because previous studies on this dataset utilized this metric to assess predictive performance. The results for the two evaluation scenarios are presented in Tables 7 and 8. In this experimental setup, the analysis indicates that U-Net (ConvFormer-M36) and the U-Net ensemble achieved the highest scores in mIoU and mF1 metrics when the clutter label was excluded from consideration. However, when the clutter label was included, U-Net (EfficientNet-B7) and the U-Net ensemble emerge as the top performers. Interestingly, the exclusion of clutter seems to reduce ambiguity among the remaining labels.

Table 7. Performance comparison on Potsdam dataset (without clutter) between our U-Net models and other existing semantic segmentation approaches. We report the mean intersection over union (mIoU %), mean F1 score (mF1 %), and F1 score (%) for each label. Bold indicates the best-performing model and underline the second-best model for a specific label.

Model	F1 per Label					mF1	mIoU
	Imp. Surf.	Car	Tree	Low Veg.	Building		
U-Net (ResNet50) [23]	95.13	97.2	88.97	89.2	97.92	93.69	88.37
DeepLabV3+ (ResNet50) [25]	95.09	97.41	88.75	88.64	97.57	93.49	88.04
SegFormer-B5 [33]	95.26	97.4	89.21	89.39	98.01	93.85	88.66
UperNet (Swin small) [27]	95.35	97.65	89.51	89.54	98.01	94.01	88.94
AerialFormer-B [38]	95.4	97.4	89.7	89.6	98.0	94.0	89.0
UNetFormer (ResNet18) [39]	95.04	97.39	89.09	88.99	97.68	93.64	88.28
DC-Swin small [65]	95.28	97.57	89.67	89.61	98.1	94.05	88.99
U-Net (MaxViT-S)	94.55	97.44	89.51	89.18	97.26	93.59	88.17
U-Net (ConvFormer-M36)	95.46	97.85	90.14	90.0	98.11	94.31	89.45
U-Net (EfficientNet-B7)	95.16	97.38	89.29	89.21	98.02	93.81	88.59
U-Net Ensemble	**95.78**	**97.96**	**90.27**	**90.46**	**98.34**	**94.56**	**89.9**

Table 8. Performance comparison on Potsdam dataset (with clutter) between our U-Net models and other existing semantic segmentation approaches. We report the mean intersection over union (mIoU %), mean F1 score (mF1 %), and F1 score (%) for each label. Bold indicates the best-performing model and underline the second-best model for a specific label.

Model	F1 per Label						mF1	mIoU
	Imp. Surf.	Clutter	Car	Tree	Low Veg.	Building		
U-Net (ResNet50) [23]	93.31	57.7	96.52	88.94	87.47	96.78	86.79	78.81
DeepLabV3+ (ResNet50) [25]	93.3	59.91	96.45	88.86	87.37	96.83	87.12	79.12
SegFormer-B5 [33]	93.52	58.66	96.5	89.12	87.63	96.84	87.04	79.13
UperNet (Swin small) [27]	93.54	56.87	96.08	88.59	87.35	97.02	86.57	78.55
AerialFormer-B [38]	93.5	61.9	95.7	89.3	88.1	97.2	87.6	79.7
UNetFormer (ResNet18) [39]	93.29	57.8	96.32	88.64	87.18	96.7	86.65	78.58
DC-Swin small [65]	93.36	53.84	96.09	88.86	87.5	97.08	86.12	78.15
U-Net (MaxViT-S)	93.56	60.74	96.66	89.27	87.8	97.06	87.51	79.7
U-Net (ConvFormer-M36)	93.8	59.41	96.71	89.57	88.01	97.27	87.46	79.77
U-Net (EfficientNet-B7)	93.53	61.97	96.35	89.42	87.78	97.04	87.68	79.84
U-Net Ensemble	**94.1**	**62.08**	**96.95**	**90.12**	**88.81**	**97.48**	**88.26**	**80.82**

Figure 12 shows example images and ground-truth masks from the Potsdam dataset. Additionally, it displays the inference masks generated by the U-Net ensemble, the best-performing model, and other models for comparison. The U-Net ensemble model exhibits remarkable resilience, achieving excellent performance despite the high scene complexity of the Potsdam dataset. The confusion matrices for the Potsdam dataset (with and without clutter) are visualized in Figure 13. The clutter label in the Potsdam dataset presents a unique challenge as it encompasses anything outside the five main categories. This ambiguity can lead to confusion with labels like impervious_surface, low_vegetation, and building. Excluding the clutter label significantly reduces ambiguity among the remaining labels. This is evident in the F1 values exceeding 90% for all the labels when clutter is ignored. As expected, some confusion persists between tree and low_vegetation labels, likely due to their inherent similarities.

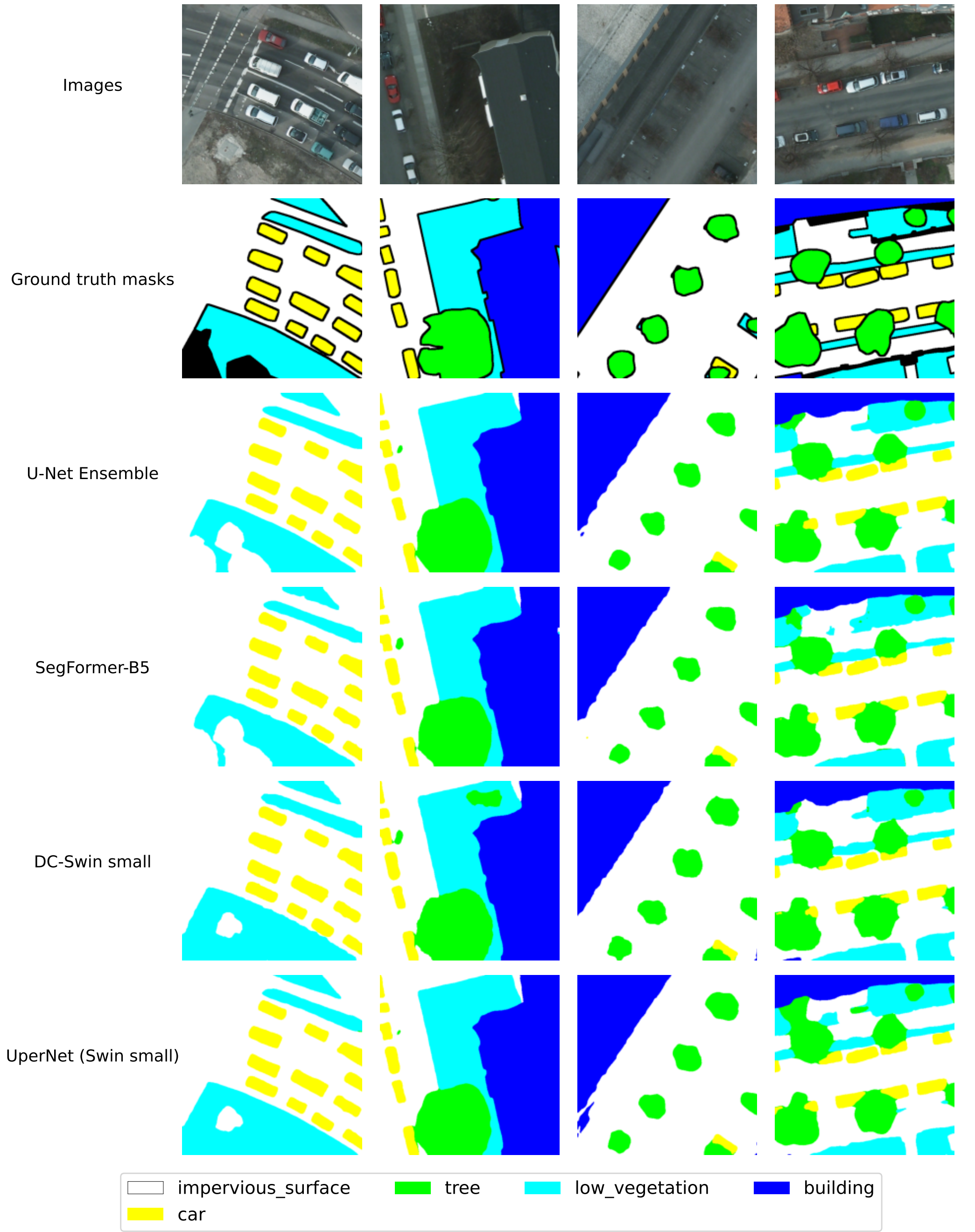

Figure 12. Example images, ground-truth masks, and inference masks from the Potsdam dataset. The first row shows example images. The second row shows the corresponding ground-truth masks. Each subsequent row shows the segmentation masks generated by a different model for the corresponding image in the first row.

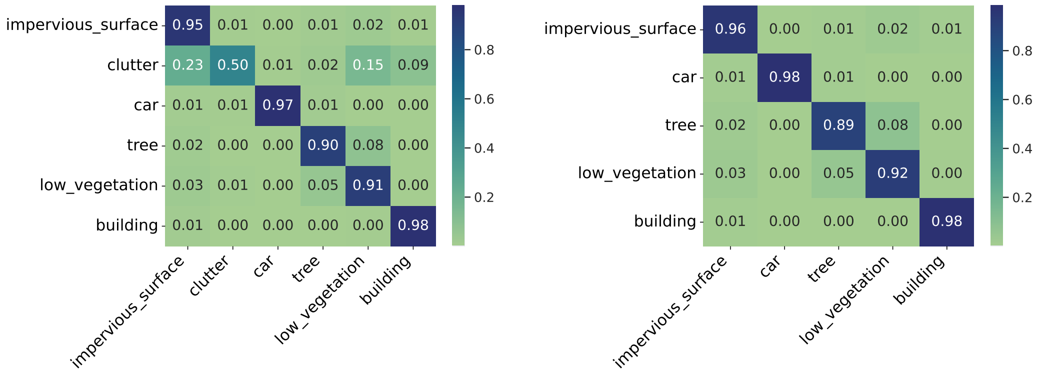

Figure 13. Confusion matrix for the U-Net ensemble model on the Potsdam dataset, with clutter (**left**) and without clutter (**right**).

For the INRIA dataset, the provided metric is the intersection over union (%) specifically for the building label, which is a common practice in building extraction [40]. Our model achieves a competitive IoU value of 81.43% on the building extraction task, even without incorporating postprocessing techniques for boundary refinement. While this aligns with current state-of-the-art results (the highest reported on the contest platform being 81.91%), there is room for further improvement. For instance, the UANet method [40] achieves an IoU of 83.08% by introducing uncertainty awareness into the network. This allows UANet to maintain high confidence in predictions across diverse scales, complex backgrounds, and various building appearances. As future work, we plan to explore similar uncertainty-handling techniques to potentially boost the performance of our models. Figure 14 showcases close-ups of the segmentation outcomes on the test set. The U-Net ensemble model adeptly identifies buildings across diverse images, showcasing its capability to detect structures of varying sizes and shapes.

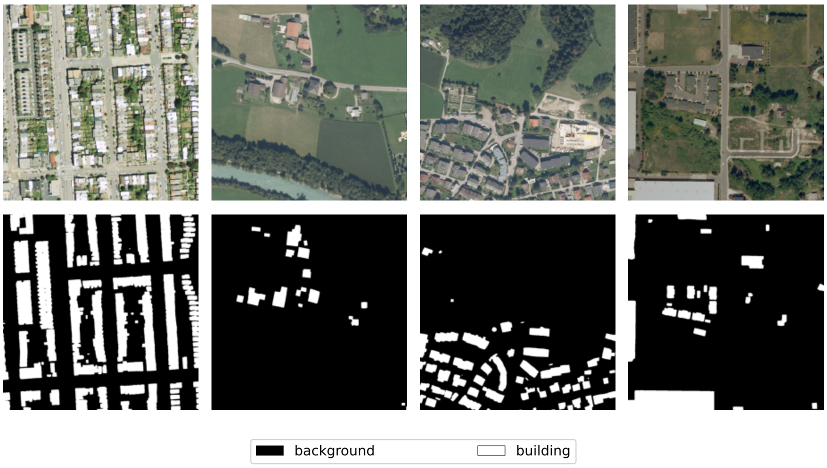

Figure 14. Example images and inference masks for INRIA dataset. First row shows example images and second row shows the prediction results of U-Net ensemble model as in Table 2.

Balancing Performance and Efficiency

Our approach leverages medium-sized backbones, MaxViT-S (∼69M parameters), ConvFormer-M36 (∼57M parameters), and EfficientNet-B7 (∼66M parameters), within the U-Net architecture. This choice strikes a balance between performance and computational efficiency. Notably, the individual base models achieve competitive results on most datasets, showcasing the effectiveness of these backbones. For instance, compared to state-of-the-art methods like AerialFormer-B (∼114M parameters), SegFormer-B5 (∼84M parameters), DC-Swin small (∼67M parameters), multitask pretraining using the InternImage-XL model (∼335M), and UperNet with Swin (small variant) (∼81M parameters), our models achieve similar performance with a comparable number of parameters. This demonstrates that our chosen medium-sized backbones offer a strong trade-off between accuracy and resource consumption.

To push the performance of the base U-Net models even further, we introduce a three-model ensemble strategy. This approach capitalizes on the complementary strengths of each backbone by combining their diverse feature representations, aiming to surpass the capabilities of base models. Ensemble methods present a trade-off: they can boost performance but come with a higher computational burden. Limiting the ensemble to three models ensures a good balance between performance gains and computational cost. This selection provides a diverse set of feature extraction capabilities while maintaining computational tractability. To further improve efficiency, we can explore techniques for optimizing the inference time of the base models. These techniques include pruning the model architecture [72] and quantizing the network parameters [73].

4. Conclusions

In this paper, we explored the effectiveness of incorporating strong backbones within the U-Net architecture to enhance the semantic segmentation of remote sensing imagery. We investigated three distinct backbone networks—Multi-Axis Vision Transformer, ConvFormer, and EfficientNet. These backbones extract complementary features, leading to a richer understanding of the scene. Further, we introduced an ensemble learning approach that leverages the unique representations learned by each backbone, resulting in more robust and accurate segmentation. We utilized a geometric mean ensemble strategy to integrate the base predictions effectively. Through a comprehensive comparison with existing methods on various remote sensing image datasets commonly used in semantic segmentation tasks for remote sensing imagery, including LandCover.ai, LoveDA, INRIA, UAVid, and ISPRS Potsdam datasets, our approach consistently achieved state-of-the-art performance. Notably, the ensemble secured the top rank on the LoveDA leaderboard and established new benchmarks for UAVid, LandCover.ai, and ISPRS Potsdam datasets.

To gain a more comprehensive understanding of the ensemble's performance across the various datasets, we analyzed per-label IoU values, confusion matrices, and sample inference masks. This detailed analysis provided insights into the model's performance across different types of remote sensing imagery and segmentation tasks, highlighting its strengths and areas for improvement. Overall, our study demonstrates the effectiveness and robustness of the proposed U-Net ensemble approach in enhancing semantic segmentation performance in remote sensing imagery.

Author Contributions: Conceptualization, I.D., I.K., V.S. and S.L.; methodology, I.D., I.K., V.S. and S.L.; software, I.D.; validation, I.D., I.K., V.S. and S.L.; formal analysis, I.D., I.K., V.S. and S.L.; investigation, I.D., I.K., V.S. and S.L.; resources, I.D.; data curation, I.D.; writing—original draft preparation, I.D.; writing—review and editing, I.D., I.K., V.S. and S.L.; visualization, I.D.; supervision, S.L.; project administration, I.K.; funding acquisition, I.D., I.K., V.S. and S.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The LoveDA dataset in this study is openly and freely available at https://github.com/Junjue-Wang/LoveDA (accessed on 29 May 2024). The UAVid dataset in this study is openly and freely available at https://uavid.nl (accessed on 29 May 2024). The LandCover.ai dataset in this study is openly and freely available at https://landcover.ai.linuxpolska.com/download/landcover.ai.v1.zip (accessed on 29 May 2024). The Potsdam dataset in this study is openly and freely available at https://www.isprs.org/education/benchmarks/UrbanSemLab/semantic-labeling.aspx (accessed on 29 May 2024). The INRIA dataset in this study is openly and freely available at https://project.inria.fr/aerialimagelabeling/ (accessed on 29 May 2024). The code to reproduce the results from the paper will be available at https://github.com/ivicadimitrovski (accessed on 5 June 2024) upon submission decision.

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

References

1. Toth, C.; Jóźków, G. Remote sensing platforms and sensors: A survey. *ISPRS J. Photogramm. Remote Sens.* **2016**, *115*, 22–36. [CrossRef]
2. Tupin, F.; Inglada, J.; Nicolas, J.M. *Remote Sensing Imagery*; John Wiley & Sons: Hoboken, NJ, USA, 2014.
3. Spasev, V.; Dimitrovski, I.; Kitanovski, I.; Chorbev, I. Semantic Segmentation of Remote Sensing Images: Definition, Methods, Datasets and Applications. In Proceedings of the ICT Innovations 2023. Learning: Humans, Theory, Machines, and Data, Ohrid, North Macedonia, 24–26 September 2024; pp. 127–140.
4. Kemker, R.; Salvaggio, C.; Kanan, C. Algorithms for semantic segmentation of multispectral remote sensing imagery using deep learning. *ISPRS J. Photogramm. Remote Sens.* **2018**, *145*, 60–77. [CrossRef]
5. Dimitrovski, I.; Kitanovski, I.; Kocev, D.; Simidjievski, N. Current trends in deep learning for Earth Observation: An open-source benchmark arena for image classification. *ISPRS J. Photogramm. Remote Sens.* **2023**, *197*, 18–35. [CrossRef]
6. Dimitrovski, I.; Kitanovski, I.; Simidjievski, N.; Kocev, D. In-Domain Self-Supervised Learning Improves Remote Sensing Image Scene Classification. *IEEE Geosci. Remote Sens. Lett.* **2024**, *21*, 1–5. [CrossRef]
7. Dimitrovski, I.; Kitanovski, I.; Panov, P.; Kostovska, A.; Simidjievski, N.; Kocev, D. AiTLAS: Artificial Intelligence Toolbox for Earth Observation. *Remote Sens.* **2023**, *15*, 2343. [CrossRef]
8. Yuan, X.; Shi, J.; Gu, L. A review of deep learning methods for semantic segmentation of remote sensing imagery. *Expert Syst. Appl.* **2021**, *169*, 114417. [CrossRef]
9. Kotaridis, I.; Lazaridou, M. Remote sensing image segmentation advances: A meta-analysis. *ISPRS J. Photogramm. Remote Sens.* **2021**, *173*, 309–322. [CrossRef]
10. Neupane, B.; Horanont, T.; Aryal, J. Deep Learning-Based Semantic Segmentation of Urban Features in Satellite Images: A Review and Meta-Analysis. *Remote Sens.* **2021**, *13*, 808. [CrossRef]
11. Maggiori, E.; Tarabalka, Y.; Charpiat, G.; Alliez, P. Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark. In Proceedings of the 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Fort Worth, TX, USA, 23–28 July 2017; pp. 3226–3229.
12. Boguszewski, A.; Batorski, D.; Ziemba-Jankowska, N.; Dziedzic, T.; Zambrzycka, A. LandCover.ai: Dataset for Automatic Mapping of Buildings, Woodlands, Water and Roads from Aerial Imagery. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, Nashville, TN, USA, 20–25 June 2021; pp. 1102–1110.
13. Toker, A.; Kondmann, L.; Weber, M.; Eisenberger, M.; Camero, A.; Hu, J.; Hoderlein, A.P.; Şenaras, C.; Davis, T.; Cremers, D.; et al. DynamicEarthNet: Daily Multi-Spectral Satellite Dataset for Semantic Change Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 21158–21167.
14. Hernández, D.; Cecilia, J.M.; Cano, J.C.; Calafate, C.T. Flood Detection Using Real-Time Image Segmentation from Unmanned Aerial Vehicles on Edge-Computing Platform. *Remote Sens.* **2022**, *14*, 223. [CrossRef]
15. Cui, L.; Jing, X.; Wang, Y.; Huan, Y.; Xu, Y.; Zhang, Q. Improved Swin Transformer-Based Semantic Segmentation of Postearthquake Dense Buildings in Urban Areas Using Remote Sensing Images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *16*, 369–385. [CrossRef]
16. Rashkovetsky, D.; Mauracher, F.; Langer, M.; Schmitt, M. Wildfire Detection From Multisensor Satellite Imagery Using Deep Semantic Segmentation. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 7001–7016. [CrossRef]
17. Merdjanovska, E.; Kitanovski, I.; Kokalj, Ž.; Dimitrovski, I.; Kocev, D. Crop Type Prediction Across Countries and Years: Slovenia, Denmark and the Netherlands. In Proceedings of the IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium, Kuala Lumpur, Malaysia, 17–22 July 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 5945–5948.
18. Dadsetan, S.; Rose, G.L.; Hovakimyan, N.; Hobbs, J. Detection and Prediction of Nutrient Deficiency Stress using Longitudinal Aerial Imagery. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020.
19. 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]
20. Moazzam, S.I.; Khan, U.S.; Qureshi, W.S.; Nawaz, T.; Kunwar, F. Towards automated weed detection through two-stage semantic segmentation of tobacco and weed pixels in aerial Imagery. *Smart Agric. Technol.* **2023**, *4*, 100142. [CrossRef]
21. Guo, Y.; Liu, Y.; Georgiou, T.; Lew, M.S. A review of semantic segmentation using deep neural networks. *Int. J. Multimed. Inf. Retr.* **2018**, *7*, 87–93. [CrossRef]
22. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
23. 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, Munich, Germany, 5–9 October 2015; pp. 234–241.
24. Badrinarayanan, V.; Kendall, A.; Cipolla, R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [CrossRef]
25. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 834–848. [CrossRef]
26. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2881–2890.

27. Xiao, T.; Liu, Y.; Zhou, B.; Jiang, Y.; Sun, J. Unified perceptual parsing for scene understanding. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 418–434.
28. 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.
29. Zhao, H.; Zhang, Y.; Liu, S.; Shi, J.; Loy, C.C.; Lin, D.; Jia, J. Psanet: Point-wise spatial attention network for scene parsing. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 267–283.
30. Fu, J.; Liu, J.; Jiang, J.; Li, Y.; Bao, Y.; Lu, H. Scene segmentation with dual relation-aware attention network. *IEEE Trans. Neural Netw. Learn. Syst.* **2020**, *32*, 2547–2560. [CrossRef] [PubMed]
31. Zheng, S.; Lu, J.; Zhao, H.; Zhu, X.; Luo, Z.; Wang, Y.; Fu, Y.; Feng, J.; Xiang, T.; Torr, P.H.; et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 6881–6890.
32. Strudel, R.; Garcia, R.; Laptev, I.; Schmid, C. Segmenter: Transformer for semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 7262–7272.
33. 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.
34. Cheng, B.; Schwing, A.; Kirillov, A. Per-pixel classification is not all you need for semantic segmentation. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 17864–17875.
35. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 213–229.
36. Cheng, B.; Choudhuri, A.; Misra, I.; Kirillov, A.; Girdhar, R.; Schwing, A.G. Mask2former for video instance segmentation. *arXiv* **2021**, arXiv:2112.10764.
37. Zheng, Z.; Zhong, Y.; Wang, J.; Ma, A. Foreground-aware relation network for geospatial object segmentation in high spatial resolution remote sensing imagery. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 4096–4105.
38. Yamazaki, K.; Hanyu, T.; Tran, M.; Garcia, A.; Tran, A.; McCann, R.; Liao, H.; Rainwater, C.; Adkins, M.; Molthan, A.; et al. AerialFormer: Multi-resolution Transformer for Aerial Image Segmentation. *arXiv* **2023**, arXiv:2306.06842.
39. Wang, L.; Li, R.; Zhang, C.; Fang, S.; Duan, C.; Meng, X.; Atkinson, P.M. UNetFormer: A UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery. *ISPRS J. Photogramm. Remote Sens.* **2022**, *190*, 196–214. [CrossRef]
40. He, W.; Li, J.; Cao, W.; Zhang, L.; Zhang, H. Building extraction from remote sensing images via an uncertainty-aware network. *arXiv* **2023**, arXiv:2307.12309.
41. Xiong, Z.; Zhang, F.; Wang, Y.; Shi, Y.; Zhu, X.X. EarthNets: Empowering AI in Earth Observation. *arXiv* **2022**, arXiv:2210.04936.
42. Wang, J.; Zheng, Z.; Ma, A.; Lu, X.; Zhong, Y. LoveDA: A Remote Sensing Land-Cover Dataset for Domain Adaptive Semantic Segmentation. In Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks, Virtual, 6–14 December 2021; Volume 1.
43. Lyu, Y.; Vosselman, G.; Xia, G.S.; Yilmaz, A.; Yang, M.Y. UAVid: A semantic segmentation dataset for UAV imagery. *ISPRS J. Photogramm. Remote Sens.* **2020**, *165*, 108–119. [CrossRef]
44. Tu, Z.; Talebi, H.; Zhang, H.; Yang, F.; Milanfar, P.; Bovik, A.; Li, Y. MaxViT: Multi-axis Vision Transformer. In Proceedings of the Computer Vision—ECCV 2022, Tel Aviv, Israel, 23–27 October 2022; Springer Nature: Cham, Switzerland, 2022; pp. 459–479.
45. Yu, W.; Si, C.; Zhou, P.; Luo, M.; Zhou, Y.; Feng, J.; Yan, S.; Wang, X. MetaFormer Baselines for Vision. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *46*, 896–912. [CrossRef] [PubMed]
46. Tan, M.; Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In Proceedings of the International Conference on Machine Learning, Beach, CA, USA, 9–15 June 2019; pp. 6105–6114.
47. 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 16 × 16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
48. Yu, W.; Luo, M.; Zhou, P.; Si, C.; Zhou, Y.; Wang, X.; Feng, J.; Yan, S. Metaformer is actually what you need for vision. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 10819–10829.
49. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778.
50. Liu, Z.; Mao, H.; Wu, C.; Feichtenhofer, C.; Darrell, T.; Xie, S. A ConvNet for the 2020s. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 11966–11976.
51. Tan, M.; Le, Q. Efficientnetv2: Smaller models and faster training. In Proceedings of the International Conference on Machine Learning, Virtual, 18–24 July 2021; pp. 10096–10106.
52. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 10012–10022.
53. Dai, Z.; Liu, H.; Le, Q.V.; Tan, M. Coatnet: Marrying convolution and attention for all data sizes. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 3965–3977.

54. Dong, X.; Yu, Z.; Cao, W.; Shi, Y.; Ma, Q. A survey on ensemble learning. *Front. Comput. Sci.* **2020**, *14*, 241–258. [CrossRef]
55. Jozdani, S.E.; Johnson, B.A.; Chen, D. Comparing deep neural networks, ensemble classifiers, and support vector machine algorithms for object-based urban land use/land cover classification. *Remote Sens.* **2019**, *11*, 1713. [CrossRef]
56. Marmanis, D.; Wegner, J.D.; Galliani, S.; Schindler, K.; Datcu, M.; Stilla, U. Semantic segmentation of aerial images with an ensemble of CNSS. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* **2016**, *3*, 473–480. [CrossRef]
57. Abdollahi, A.; Pradhan, B.; Alamri, A.M. An ensemble architecture of deep convolutional Segnet and Unet networks for building semantic segmentation from high-resolution aerial images. *Geocarto Int.* **2022**, *37*, 3355–3370. [CrossRef]
58. Ding, J.; Xue, N.; Xia, G.S.; Bai, X.; Yang, W.; Yang, M.Y.; Belongie, S.; Luo, J.; Datcu, M.; Pelillo, M.; et al. Object detection in aerial images: A large-scale benchmark and challenges. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 7778–7796. [CrossRef] [PubMed]
59. Albaba, B.M.; Ozer, S. Synet: An ensemble network for object detection in uav images. In Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 10–15 January 2021; pp. 10227–10234.
60. Ganaie, M.; Hu, M.; Malik, A.; Tanveer, M.; Suganthan, P. Ensemble deep learning: A review. *Eng. Appl. Artif. Intell.* **2022**, *115*, 105151. [CrossRef]
61. Loshchilov, I.; Hutter, F. Decoupled weight decay regularization. *arXiv* **2017**, arXiv:1711.05101.
62. Buslaev, A.; Iglovikov, V.I.; Khvedchenya, E.; Parinov, A.; Druzhinin, M.; Kalinin, A.A. Albumentations: Fast and Flexible Image Augmentations. *Information* **2020**, *11*, 125. [CrossRef]
63. Sermanet, P.; Eigen, D.; Zhang, X.; Mathieu, M.; Fergus, R.; LeCun, Y. OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks. *arXiv* **2014**, arXiv:1312.6229.
64. Falcon, W.; The PyTorch Lightning team. *PyTorch Lightning/Pytorch-Lightning: 2.1.2 Release*; Zenodo: Geneva, Switzerland, 2019. [CrossRef]
65. Wang, L.; Li, R.; Duan, C.; Zhang, C.; Meng, X.; Fang, S. A Novel Transformer Based Semantic Segmentation Scheme for Fine-Resolution Remote Sensing Images. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 1–5. [CrossRef]
66. Wang, D.; Zhang, J.; Xu, M.; Liu, L.; Wang, D.; Gao, E.; Han, C.; Guo, H.; Du, B.; Tao, D.; et al. MTP: Advancing Remote Sensing Foundation Model via Multi-Task Pretraining. *arXiv* **2024**, arXiv:2403.13430.
67. Cha, K.; Seo, J.; Lee, T. A billion-scale foundation model for remote sensing images. *arXiv* **2023**, arXiv:2304.05215.
68. Li, X.; Li, Y.; Ai, J.; Shu, Z.; Xia, J.; Xia, Y. Semantic segmentation of UAV remote sensing images based on edge feature fusing and multi-level upsampling integrated with Deeplabv3+. *PLoS ONE* **2023**, *18*, e0279097. [CrossRef] [PubMed]
69. Ding, Y.; Zheng, X.; Chen, Y.; Shen, S.; Xiong, H. Dense context distillation network for semantic parsing of oblique UAV images. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *114*, 103062. [CrossRef]
70. Wang, S.; Hu, Q.; Wang, S.; Zhao, P.; Li, J.; Ai, M. Category attention guided network for semantic segmentation of Fine-Resolution remote sensing images. *Int. J. Appl. Earth Obs. Geoinf.* **2024**, *127*, 103661. [CrossRef]
71. Yuan, Y.; Chen, X.; Wang, J. Object-Contextual Representations for Semantic Segmentation. In Proceedings of the Computer Vision—ECCV 2020, Glasgow, UK, 23–28 August 2020; pp. 173–190.
72. Frankle, J.; Carbin, M. The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks. *arXiv* **2018**, arXiv:1803.03635.
73. Liang, T.; Glossner, J.; Wang, L.; Shi, S.; Zhang, X. Pruning and quantization for deep neural network acceleration: A survey. *Neurocomputing* **2021**, *461*, 370–403. [CrossRef]

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