applied sciences

MDPI

Article

Semantic Segmentation of Aerial Imagery Using U-Net with Self-Attention and Separable Convolutions

Bakht Alam Khan and Jin-Woo Jung *

Department of Computer Science and Engineering, Dongguk University, Seoul 04620, Republic of Korea; bakhtalam4687@gmail.com

* Correspondence: jwjung@dongguk.edu

Abstract: This research addresses the crucial task of improving accuracy in the semantic segmentation of aerial imagery, essential for applications such as urban planning and environmental monitoring. This study emphasizes the significance of maintaining the Intersection over Union (IOU) score as a metric and employs data augmentation with the Patchify library, using a patch size of 256, to effectively augment the dataset, which is subsequently split into training and testing sets. The core of this investigation lies in a novel architecture that combines a U-Net framework with self-attention mechanisms and separable convolutions. The introduction of self-attention mechanisms enhances the model's understanding of image context, while separable convolutions expedite the training process, contributing to overall efficiency. The proposed model demonstrates a substantial accuracy improvement, surpassing the previous state-of-the-art Dense Plus U-Net, achieving an accuracy of 91% compared to the former's 86%. Visual representations, including original patch images, original masked patches, and predicted patch masks, showcase the model's proficiency in semantic segmentation, marking a significant advancement in aerial image analysis and underscoring the importance of innovative architectural elements for enhanced accuracy and efficiency in such tasks.

Keywords: semantic segmentation; U-Net; self-attention; separable convolutions; aerial imagery; remote sensing

Citation: Khan, B.A.; Jung, J.-W. Semantic Segmentation of Aerial Imagery Using U-Net with Self-Attention and Separable Convolutions. *Appl. Sci.* **2024**, *14*, 3712. https://doi.org/10.3390/app14093712

Academic Editor: Thomas Lindner

Received: 5 March 2024
Revised: 9 April 2024
Accepted: 18 April 2024
Published: 26 April 2024

1. Introduction

The accurate semantic segmentation of aerial imagery holds immense significance in various practical applications, particularly in fields such as urban planning and environmental monitoring. As cities expand and landscapes evolve, the need for precise land cover classification becomes paramount for effective urban development and resource management. Furthermore, in environmental monitoring, the ability to differentiate between various land features aids in assessing ecological changes and ensuring sustainable practices. The integration of a U-Net framework with self-attention mechanisms and separable convolutions, as explored in this research, offers a promising avenue for advancing the accuracy of segmentation methods. The practical implications of this research extend to critical decision-making processes, where the improved model can provide detailed insights into urban infrastructure, land use patterns, and environmental changes. Such advancements are crucial for policymakers, urban designers, and environmental scientists who rely on accurate spatial information for informed decision making. By enhancing the precision of semantic segmentation, this study contributes to the broader goal of harnessing aerial imagery for practical applications that address contemporary challenges in urban development and environmental sustainability. The most difficult process that you can come across in the field of computer vision is the problem of semantic segmentation. The problem of semantic segmentation has been practiced and studied heavily in an image processing context. In the study of semantic segmentation, the goal is to connect or classify each pixel of an image into a certain class [1]. Land use and land cover maps are

frequently made from medium- and high-resolution satellite images like Sentinel [2] and Landsat [3]. Semantic segmentation holds crucial significance across a diverse range of applications within the field of geographic information. This technique plays a pivotal role in various geographic information applications, showcasing its multifaceted utility in extracting meaningful insights from spatial data [4]. The application of deep learning (DL) in remote sensing semantic segmentation is progressively gaining traction [5]. Earlier techniques relying on sliding windows and candidate regions are characterized by their time-consuming nature and a substantial number of redundant computations [6].

In recent times, an increasing number of methods based on deep learning (DL) have emerged in the realm of remote sensing, encompassing variations of the U-Net model and its derivatives [7]. By discerning the significance of specific elements within a given context, attention mechanisms contribute to the refinement of decision-making processes. Their ability to ignore extraneous or irrelevant data not only streamlines computational efficiency but also plays a crucial role in mitigating the impact of noise or distractors in complex datasets [8].

The deep learning architecture that is used herein is a self-attention U-Net with separable convolutions. The self-attention takes advantage of focusing on the key pixels in the images while ignoring the unnecessary pixel values. In order to better handle input images with high resolutions, we split input images into image patches and perform training and evaluation in a patch-by-patch manner [9]. The deep learning architecture that is used herein is a self-attention U-Net with separable convolutions. The self-attention takes advantage of focusing on the key pixels in the images while ignoring the unnecessary pixel values. The separable convolutions used in the proposed architecture speed up the training process. Several works in the literature have undertaken comprehensive reviews of research methodologies within the realm of semantic segmentation for remote sensing data. These works systematically classify and explore diverse perspectives, specifically delving into remote sensing image analysis within the broader domain of general deep learning algorithms. Through meticulous examination and categorization, these studies contribute to a nuanced understanding of the various approaches employed in the field, shedding light on the application of deep learning algorithms for the semantic segmentation of remote sensing data. The synthesis of these perspectives serves to consolidate knowledge and offer insights into the evolving landscape of research methods, providing a valuable resource for scholars, practitioners, and enthusiasts engaged in the dynamic field of remote sensing image analysis. At the heart of this investigation lies a groundbreaking architectural approach, seamlessly combining the U-Net framework with innovative self-attention mechanisms and separable convolutions.

The incorporation of self-attention mechanisms elevates the model's ability to grasp intricate image contexts, while the strategic use of separable convolutions accelerates the training process, culminating in heightened overall efficiency. Noteworthy is the model's unprecedented accuracy improvement, as evidenced by its outperformance of the prior state-of-the-art Dense Plus U-Net, achieving an impressive 91% accuracy compared to the former's 86%. These advancements, visually represented through original patch images, masked patches, and predicted patch masks, underscore the model's proficiency in semantic segmentation, marking a substantial stride forward in the realm of aerial image analysis. This work not only addresses the immediate need for enhanced accuracy but also emphasizes the pivotal role of novel architectural elements in shaping the future trajectory of this critical research domain. Due to the widespread adoption of deep learning, numerous traditional semantic segmentation models have been developed, among which is the fully convolutional network (FCN) [10]. Nevertheless, the FCN exhibits a fixed receptive field, posing a challenge in retaining intricate details as there is a susceptibility to information loss. The FCN had the problem of losing important pixel information; to solve this problem, the U-Net was used, in which the encoder–decoder architecture maintains the important pixels, and the results were more promising than the FCN [11]. Lately, advancements in remote sensing technologies have led to the development of highly sophisticated techniques capable of providing very-high-resolution (VHR) aerial images,

characterized by a ground sampling distance ranging from 5 to 10 cm, both in spatial and spectral domains [12]. As a result, even minor objects become discernible, allowing for their segmentation with precision and accuracy [13]. While attention technology undeniably enhances segmentation accuracy, its widespread application is impeded by the substantial demand for computational resources. In recent times, a surge of refined methodologies has surfaced to address this challenge, notably the self-attention mechanism and fusion attention mechanism. This section provides a comprehensive synthesis and examination of linear attention and sub-attention mechanisms, delving into their nuanced contributions to optimizing computational efficiency in attention-based models. Additionally, it explores the intricacies of channel and spatial attention mechanisms, shedding light on their respective roles in refining segmentation accuracy without compromising computational scalability. As attention mechanisms continue to evolve, the nuanced discussion within this section aims to provide a comprehensive overview, laying the foundation for a deeper understanding of their applications and implications in the context of segmentation processes [4].

The major innovations of using the U-Net were its four layers of skip connections in between. Over the past few years, there has been a notable surge in the application of refined methodologies derived from classical deep learning approaches for semantic segmentation in the domain of remote sensing images. This influx of improved methods signifies a growing trend in leveraging advanced techniques to extract meaningful information from complex remote sensing datasets. The amalgamation of classical deep learning principles with innovative enhancements underscores a continuous effort to enhance the precision and efficiency of semantic segmentation in the context of remote sensing. These advancements play a pivotal role in addressing the evolving challenges of interpreting and understanding intricate spatial information embedded within remote sensing imagery. As researchers explore and implement these improved methods, the synergy between classical deep learning and remote sensing applications unfolds new possibilities for accurate and comprehensive image analysis [14].

In the dynamic interplay of urban expansion and environmental shifts, the need for accurate semantic segmentation of aerial imagery becomes even more pronounced, offering a crucial lens for understanding and navigating the complexities of contemporary landscapes. As urban centers burgeon into sprawling metropolises and natural landscapes undergo transformations, the demand for precise land cover classification intensifies, assuming a pivotal role in orchestrating effective urban development and judicious resource management strategies. This need resonates acutely in the realm of environmental monitoring, where discerning and categorizing various land features becomes an invaluable tool for assessing ecological changes and fostering sustainable practices. The fusion of a U-Net framework with self-attention mechanisms and separable convolutions, a core exploration in this research, signifies a transformative leap towards enhancing the accuracy of semantic segmentation methodologies, providing a progressive solution to the challenges posed by intricate and evolving landscapes. In practical terms, the ramifications of this research extend far beyond theoretical realms, permeating into the realms of critical decision-making processes. The refined model serves as a cognitive ally, unraveling detailed insights into the nuances of urban infrastructure, intricacies of land use patterns, and the dynamics of environmental changes. Stakeholders, including policymakers, urban designers, and environmental scientists, stand to benefit significantly from this innovative approach, relying on accurate spatial information to make informed decisions that reverberate through urban planning and environmental sustainability initiatives. In the face of contemporary challenges characterized by rapid urbanization, climate concerns, and the need for resilient urban ecosystems, the heightened accuracy achieved through improved semantic segmentation becomes a linchpin for navigating these complexities effectively. This study, by elevating the precision of aerial imagery analysis, not only contributes substantively to the overarching goal of harnessing advanced technologies but also empowers decision-makers to navigate the intricacies of modern urban development and environmental sustainability with foresight and precision. The fusion of a U-Net framework with self-attention and separable convolutions showcases a significant leap forward, addressing computational challenges while achieving unparalleled

precision. Through an exploration of linear and sub-attention, as well as channel and spatial attention mechanisms, this research contributes to the evolving landscape of image analysis, promising transformative advancements for practical applications in urban planning and environmental monitoring.

2. Literature Review

2.1. CNN Approach for Aerial Imagery Segmentation

Leveraging convolutional neural networks (CNNs), this approach involved precise pixel labeling for training and extracting intricate building areas. This method significantly contributed to the advancement of semantic segmentation in aerial images, utilizing CNNs' robust pattern recognition capabilities to enhance spatial understanding and provide valuable insights into the distribution of buildings. The refined methodology showcases ongoing progress in deploying deep learning techniques for extracting meaningful information from complex visual data [15].

2.2. Ensemble CNN-Based Semantic Segmentation in High-Resolution Aerial Imagery

In 2016, (Marmanis et al., 2016) elucidated a semantic segmentation model that leveraged high-resolution aerial imagery. The approach incorporated an ensemble of convolutional neural networks (CNNs), specifically the FCN, alongside adapted CNNs, highlighting their heightened efficiency when applied to standard datasets. This detailed exploration underscored the model's effectiveness in semantic segmentation and demonstrated the superior performance achieved through the integration of an ensemble of CNN architectures [16].

2.3. Improved Semantic Segmentation with Model Compression and ConvNet

In the realm of semantic segmentation modeling, (Holliday et al., 2017) directed their focus toward enhancing segmentation accuracy. Employing advanced model compression techniques, they strategically optimized the model for superior performance. The utilization of ConvNet played a pivotal role in evaluating the significance of segmentation within this specific context, contributing to a comprehensive understanding of the model's overall effectiveness [17].

2.4. CNN-Based Segmentation of High-Resolution Aerial Images with Pyramid Pooling

In the landscape of image segmentation, (Yu et al., 2018) innovatively devised a robust end-to-end methodology in 2018. Their comprehensive approach was aimed at semantically segmenting high-resolution aerial images and demonstrated a strategic fusion of a convolutional neural network (CNN) structure with a pyramid pooling phase. This thoughtful integration allowed for the extraction of feature maps across diverse scales, contributing to an enhanced and nuanced understanding of the semantic content embedded in the high-resolution aerial imagery [18].

2.5. Attention Dilation Link Neural (AD-Linknet)

In 2019, (Wu et al., 2019) delved into the examination of the attention dilation-linknet (AD-linknet) neural network. Their study embraced an encoder–decoder framework, featuring a pre-trained encoder, a channel-wise attention scheme, and a serial–parallel integrated dilated convolution. This research specifically focused on the application of these components for achieving semantic segmentation in the context of high-resolution satellite images [19].

2.6. UAV Image Segmentation with Deep Learning Enhancements

In 2020, (Boonpook et al., 2020) introduced a novel approach for multifeature semantic segmentation derived from UAV photogrammetry images, leveraging deep learning techniques. The incorporation of the SegNet model notably enhanced the accuracy of building extraction in this context. Additionally, in the same year, Yang and colleagues conducted a study in a related field [20].

2.7. Land Cover Classification with Advanced Segmentation

In the year 2021, (Mehra et al., 2021) presented an innovative approach to semantic segmentation specifically tailored for the classification of land cover. Their method incorporated a selection of six deep learning architectures, namely pyramid scene parsing, U-Net, deeplabv3, a path aggregation network, an encoder–decoder network, and a feature pyramid network. Through the application of these advanced architectures, their methodology yielded exceptional results, showcasing a significant advancement in the accuracy of land cover classification in comparison to existing techniques [21].

2.8. Orchard Tree Identification: U-Net-Based Segmentation

In 2021, (Anagnostis et al., 2021) proposed a semantic segmentation method to identify orchard trees in aerial images. They employed U-Net to enhance the efficiency of the approach, particularly in terms of accuracy. The designed model prioritized the automatic localization and detection of orchard tree canopies under various constraints [22].

3. Proposed Method

3.1. Self-Attention Mechanism

In computer vision, self-attention acts as a guide for models to focus on relevant visual information. By allowing pixels or features to communicate and adapt their importance dynamically, self-attention enhances the model's ability to understand images, recognize objects, and extract meaningful patterns as can be seen in Figure 1. This adaptive and context-aware mechanism contributes to the overall effectiveness of computer vision models in various tasks. We can think of an image as a sequence of pixels. Self-attention helps each pixel consider the relevance of other pixels in the same image. This is akin to pixels "talking" to each other, focusing on what is important nearby. Self-attention allows a computer vision model to be flexible. It can pay attention to both the big picture (entire image) and small details (local features) based on what is important for the task at hand. In scenarios where images are part of a time series, self-attention enables the fusion of temporal and spatial information. This is beneficial for tasks like tracking changes in landscapes over time [23].

$$\text{Attention}\left(\mathbf{Q}, \mathbf{K}, \mathbf{V}\right) = \textbf{softmax}\left(\frac{\boldsymbol{Q} \cdot \boldsymbol{K}^{T}}{\sqrt{\boldsymbol{d}}}\right) \cdot \boldsymbol{V} \tag{1}$$

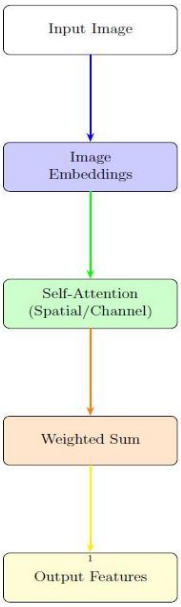


Figure 1. Self-attention mechanism representation in terms of computer vision.

3.2. Separable Convolutions

Separable convolutions are a type of smart convolutional operation designed to make convolutional processes simpler and faster as shown in Figure 2. They split the convolution into two steps: depth-wise convolution and point-wise convolution. Depth-wise convolution looks at how things are arranged in the data, but it only focuses on one channel at a time. It does not care about how different channels relate to each other. It is like zooming in on each channel's details separately. Point-wise convolution comes next. It looks at all the channels together at each point, combining the details from each channel. This step brings together the zoomed-in details from the depth-wise convolution to create a bigger picture of what is going on.

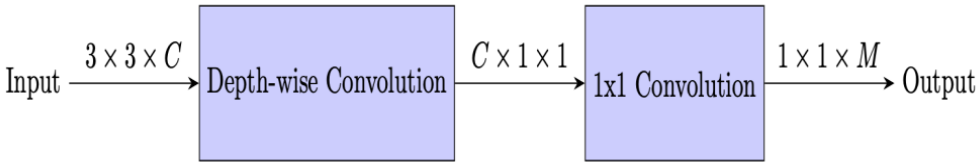


Figure 2. The process of depth-wise separable convolution. This process includes depth-wise convolution and 1×1 convolution.

The big advantage of separable convolutions is that they use fewer calculations and fewer numbers to remember. Depth-wise separable convolution can be decomposed into depth-wise convolution and 1×1 convolution (also known as point-by-point convolution) [24]. This makes them much faster and more efficient than standard convolutions. It is like using a simpler tool that does the same job but faster and with less effort. Because they are faster, separable convolutions are really useful when working with lots of data or if you have only little computing power. And even though they are simpler, they can still do a great job of picking out important details in the data. So, they are handy for tasks like figuring out what is in a picture or spotting objects in a video.

Another interesting thing about separable convolutions is that they help to prevent the model from becoming too specialized on the training data. With fewer things to remember, the model can be more flexible and better at dealing with new, unseen data. This makes it more reliable and able to handle different situations.

Standard convolution operations involve a large number of parameters and computations, making them resource-intensive. In contrast, separable convolutions break down the operation into depth-wise and point-wise convolutions, reducing the computational burden and enhancing efficiency. This decomposition is particularly advantageous in scenarios with limited resources, providing a more lightweight alternative to standard convolutions.

3.3. Workflow for Semantic Segmentation of Aerial Imagery

The proposed method as shown in Figure 3 leverages a comprehensive approach to the semantic segmentation of aerial imagery of Dubai, obtained through the use of satellites from the Mohammed Bin Rashid Space Centre (MBRSC). The dataset, consisting of 72 images grouped into 8 larger tiles, has been meticulously annotated for pixel-wise semantic segmentation across six classes: Building, Land, Road, Vegetation, Water, and Unlabeled. The task at hand involves effectively segmenting these diverse land cover features for applications such as urban planning, environmental monitoring, and infrastructure management.

To preprocess the dataset, a patch-based augmentation technique is employed. Large images are divided into smaller, manageable patches, each serving as a distinct section of the original image. These patches, with a specified size, are obtained through the Patchify library. The augmentation process includes loading the image, dividing it into patches, and determining whether these patches overlap or not. The resulting patches are instrumental in simplifying the handling of large datasets, enabling more efficient analysis and manipulation.

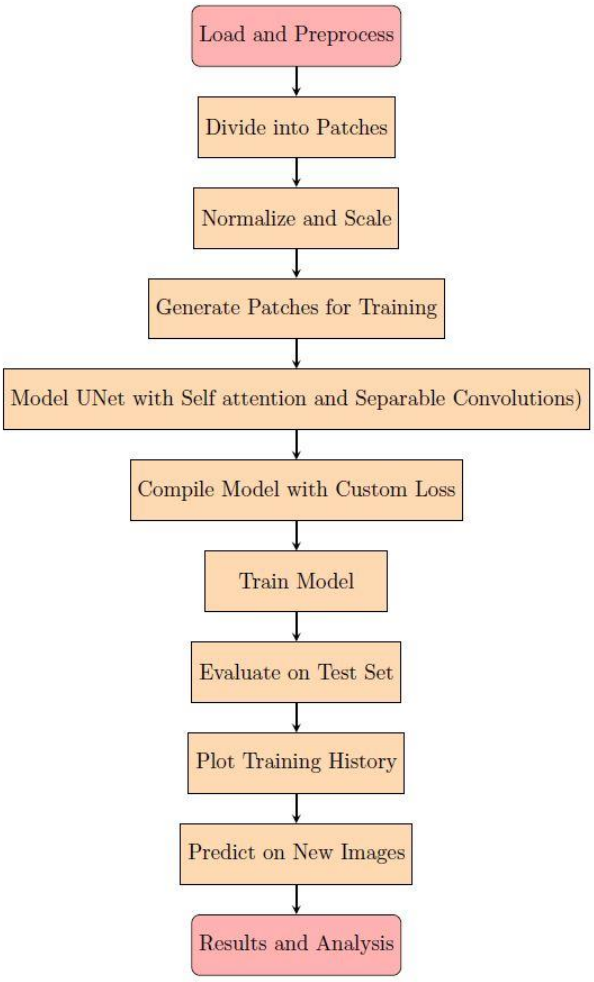


Figure 3. Workflow for semantic segmentation of aerial imagery.

Following dataset preprocessing, the method proceeds to model development. The architecture chosen for this task is the U-Net with self-attention and separable convolutions, a deep learning model known for its effectiveness in semantic segmentation tasks. The model is compiled with a custom loss function, combining Dice Loss and Focal Loss, aiming to enhance the training process and improve the model's ability to handle imbalanced classes. The model is trained on the prepared dataset, and its performance is evaluated on a separate test set. Training history is visualized through plots to assess the model's learning progress.

Ultimately, the trained model is applied to predict land cover classes in new, unseen images, facilitating the generation of valuable insights for further analysis. The proposed method concludes with a comprehensive analysis of results and their implications, providing a robust framework for semantic segmentation in aerial imagery and contributing to advancements in remote sensing applications for urban landscapes.

3.4. Proposed Model

The innovative model in Figure 4 proposed herein represents a nuanced evolution of the widely acclaimed U-Net architecture, specifically honed to address the intricate challenges inherent in satellite imagery segmentation. In this tailored adaptation, a meticulous amalgamation of state-of-the-art techniques is employed, synergizing the computational efficiency offered by separable convolutions with the profound contextual awareness instilled by the self-attention mechanism.

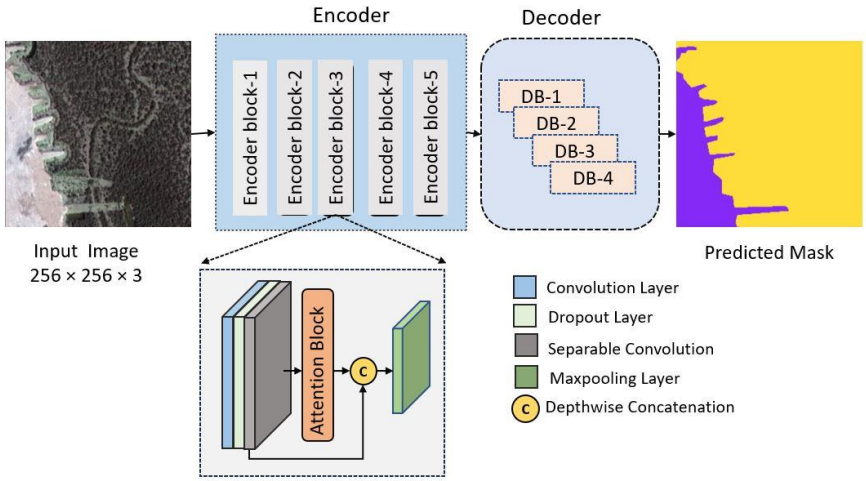


Figure 4. Proposed SA-SC-U-Net: U-Net with self-attention and separable convolutions.

3.4.1. Contextualizing U-Net

The U-Net architecture, renowned for its prowess in image segmentation, serves as the foundational framework for this advanced model. Its inherent ability to capture hierarchical features through encoder–decoder structures forms the backbone of the proposed design. Recognizing the unique demands posed by satellite imagery, this model takes a deliberate step further to enhance its segmentation capabilities. Satellite imagery, characterized by vast spatial intricacies, necessitates a model that goes beyond conventional segmentation approaches. The proposed variant is meticulously tailored to accommodate and capitalize on the distinctive features of satellite images. The integration of separable convolutions and self-attention mechanisms represents a strategic choice aimed at addressing the challenges specific to this domain. The efficiency of separable convolutions becomes particularly pertinent in the context of processing expansive satellite image datasets. This technique, characterized by its parameter-efficient design, empowers the model to navigate the complexities of high-resolution imagery while maintaining computational tractability. Each layer strategically utilizes separable convolutions, optimizing feature extraction without compromising on model efficiency.

3.4.2. Infusing Contextual Awareness with Self-Attention

Recognizing the need for nuanced contextual awareness in satellite image segmentation, self-attention mechanisms are seamlessly woven into the fabric of the model. The introduction of attention layers, specifically tailored to focus on salient features, enables the model to discern intricate patterns and long-range dependencies. This strategic fusion of contextual understanding and computational efficiency sets the proposed model apart in the realm of satellite imagery analysis.

Explaining the Components of the Model Architecture:

3.4.3. Encoder Refinement

The model's encoder is meticulously refined to capture intricate spatial hierarchies. Each encoder block is crafted with precision, incorporating convolutional layers, dropout regularization, and separable convolutions. The sequential application of attention mechanisms further refines the representation, allowing the model to learn and emphasize crucial features.

3.4.4. Decoding Spatial Details

The decoder intricately reconstructs spatial details, facilitating the faithful reconstruction of the input image. Transposed convolutions, coupled with concatenation techniques, enable the model to upsample efficiently, capturing details at multiple scales. This process is replicated across multiple decoder blocks, ensuring a comprehensive understanding of spatial relationships within the satellite imagery.

3.4.5. Final Predictive Layer

Culminating in the final layer, a 1×1 convolutional operation with softmax activation consolidates the learned features into a pixel-wise segmentation map. This top-level synthesis of information encapsulates the model's predictive capabilities, offering a nuanced delineation of diverse classes within the satellite imagery.

3.4.6. Training Strategies

The training strategy for this model is as sophisticated as its architecture. The categorical cross-entropy loss is chosen, aligning with the model's objective of multi-class segmentation in satellite imagery. Leveraging the Adam optimization algorithm, the model efficiently adapts its parameters during training, ensuring convergence towards an optimal solution. The proposed model stands as a testament to the intricate interplay between architectural refinement and strategic integration of advanced techniques. By tailoring the U-Net architecture for satellite imagery segmentation and infusing it with the efficiency of separable convolutions and the contextual awareness of self-attention mechanisms, this model emerges as a potent tool for unraveling the complexities inherent in remote sensing applications. As the dimensions of satellite imagery continue to evolve, so does the adaptability and sophistication of this proposed model, making it a formidable asset in the domain of satellite image analysis.

4. Results and Discussions

4.1. Dataset (MBRSC)

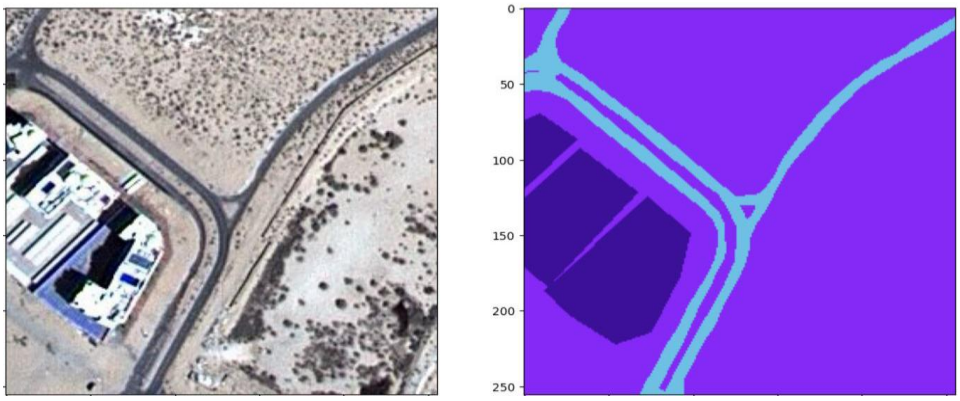
The dataset utilized in this study comprises 72 aerial images of Dubai obtained by MBRSC satellites, meticulously annotated through pixel-wise semantic segmentation across six distinct classes. These images are organized into eight larger tiles, enabling efficient handling of the dataset. The semantic segmentation classes include Building, Land, Road, Vegetation, Water, and Unlabeled, with each pixel assigned a specific label. The Unlabeled class caters to ambiguous or unannotated regions. The dataset is instrumental for training and evaluating models tailored for semantic segmentation tasks in urban environments. The MBRSC satellite-derived imagery holds significance for applications in urban planning, environmental monitoring, and infrastructure development. Additionally, the dataset allows for insights into the composition of Dubai's landscape, and the annotation of building structures, roads, and vegetation facilitates detailed analysis for various research purposes. This study underscores the dataset's importance in remote sensing and urban landscape analysis; considering both its spatial and semantic richness, the dataset is openly available on Kaggle. The original image and the mask generated can be seen in Figure 5.

(**a**) Image before preprocessing (**b**) Ground truth before preprocessing

Figure 5. Dataset semantic segmentation of aerial imagery.

4.2. Dataset Preprocessing

In the preprocessing phase of this dataset, we implemented a technique known as patching to streamline the handling of large images. Initially, we loaded substantial images, such as high-resolution photographs, and subsequently broke them down into more manageable components called patches or chunks. The size of these patches was determined by the parameter named image_patch_size. An interesting aspect of the method is the ability to choose whether these patches overlap or not. In the code, we opted for non-overlapping patches to maintain distinct sections. The resulting patches, akin to puzzle pieces, amounted to a total of two, forming a structured 1×1 grid. Each patch possessed dimensions of 256×256 pixels as can be seen in Figure 6, and comprised three color channels (R, G, B). This patching technique proves invaluable for image-related tasks, facilitating the independent analysis or modification of smaller portions of an image. Ultimately, it serves to simplify the management and processing of extensive datasets.



(**a**) Input preprocessed image (**b**) Input preprocessed mask

Figure 6. Preprocessed images after being patched with patch size of 256 with no overlapping.

The images from the original published dataset in Kaggle are preprocessed in a way so that the deep learning model can learn the features in an easy manner and the images are not heavy.

4.3. Custom Loss Combination

The custom combines the Dice Loss and Focal Loss using a weighted sum. The weights (0.5 for Dice Loss and 0.5 for Focal Loss) determine the contribution of each loss term. The combined loss is $0.5 \times$ Dice Loss + $0.5 \times$ Focal Loss. The custom loss is a hybrid of Dice Loss and Focal Loss, aiming to benefit from both methods. It addresses class imbalance through Focal Loss and encourages accurate localization through Dice Loss. The specific choice of weights (0.5 for each) reflects an equal emphasis on both components. Adjusting these weights could be explored to fine-tune the model based on the characteristics of your dataset and task.

4.4. Model Evaluation Performance Metrics

The model was trained on a labeled dataset of aerial imagery from Dubai, grouped into eight larger tiles with a total of 72 images. Each image was annotated with pixel-wise semantic segmentation in six classes: Building, Land, Road, Vegetation, Water, and Unlabeled. The dataset underwent preprocessing, including patch-based data augmentation and normalization. The number of images after augmentation was increased to 945. There were 803 training images, which makes 85 percent, and 142 were testing images. The predictions were performed on original training images and masked images. The trained model exhibited promising performance on both the training and validation sets, achieving competitive IoU scores and demonstrating its ability to segment different land features accurately. A visual inspection of the predicted segmentation masks on new images further

confirmed the model's effectiveness in capturing complex spatial patterns in the aerial imagery of Dubai. Furthermore, the utilization of pixel-wise semantic segmentation enabled the model to precisely delineate boundaries between different land classes, contributing to its overall accuracy and reliability. This capability is crucial for tasks such as urban planning, environmental monitoring, and infrastructure development in Dubai and similar urban environments.

The proposed SA-SC-U-Net outperforms the benchmark models, including FCN, U-Net, and Dense + U-Net, across various performance metrics. Notably, SA-SC-U-Net achieves the highest scores in accuracy (91.78%) and IoU (81.82%) can be seen in Table 1. The superior performance of the proposed model can be attributed to its unique combination of self-attention mechanisms and separable convolutions. The self-attention mechanism allows the model to focus on more relevant image regions, capturing long-range dependencies and enhancing feature extraction. Additionally, the use of separable convolutions reduces computational complexity while preserving model efficiency, enabling more effective learning of spatial hierarchies in the data. This combination results in improved segmentation accuracy and robustness, making SA-SC-U-Net a compelling choice for semantic segmentation tasks in aerial imagery compared to the traditional architectures like FCN and U-Net. Post completion of model training, predictions were made on new images using the SA-SC U-Net architecture. The model demonstrated its ability to segment objects effectively, providing pixel-wise predictions for each class—Building, Land, Road, Vegetation, Water, and Unlabeled. These predictions showcased the model's capacity to comprehend complex visual information and accurately assign semantic labels to various regions within the images. By visualizing the predicted masks alongside the original images, it becomes evident how well the SA-SC U-Net captures intricate patterns and nuances, making it a reliable tool for image segmentation tasks. These predictions serve as a testament to the model's practical utility and its potential application in diverse domains, from urban planning to environmental monitoring. The successful execution of predictions validates the model's learning capabilities and signifies its readiness for real-world deployment. The SA-SC-U-Net model stands out for its exceptional performance, achieving the highest accuracy (91.78%) and IoU (81.82%) scores compared to other benchmark models. This success is due to its unique combination of self-attention mechanisms and separable convolutions. The self-attention mechanism helps the model focus on important parts of the image, capturing distant relationships and improving feature extraction. At the same time, separable convolutions reduce the complexity of calculations while keeping the model efficient, allowing it to better understand the structure of the data.

Table 1. Performance metrics SA-SC U-Net (self-attention with separable convolutions U-Net).

Model	Accuracy	Validation Accuracy	Loss	Validation Loss	IoU	Validation IoU
FCN	78.2%	71.4%	86.5%	90.3%	65.3%	61.8%
U-Net	82.3%	78.9%	89.6%	93.2%	71.03%	68.36%
CNN	83.3%	79.4%	88.5%	92.2%	72.9%	70.2%
DeepLabV3	85.3%	81.5%	87.2%	91.5%	74.3%	71.9%
Dense + U-Net	86.45%	81.76%	81.78	87.56%	76.90%	72.43%
SA-SC U-Net (Ours)	**91.78%**	**87.34%**	**80.28%**	**85.56%**	**81.82%**	**77.45%**

The scale employed for image processing involves measuring the spatial resolutions for the red, green, and blue color channels. In Figure 7a–c the spatial resolutions are approximately 11, 10, and 11 LP/mm for the red, green, and blue channels, respectively. For Figure 7d–f, these resolutions are approximately 10, 9, and 11 LP/mm for the respective channels, while Figure 7g–i shows approximately 9 LP/mm for all three channels. Across all

images, the radiometric resolution remains fixed at eight bits per pixel, ensuring consistent color depth. Additionally, each image maintains a constant number of bands at three, representing the presence of red, green, and blue channels.

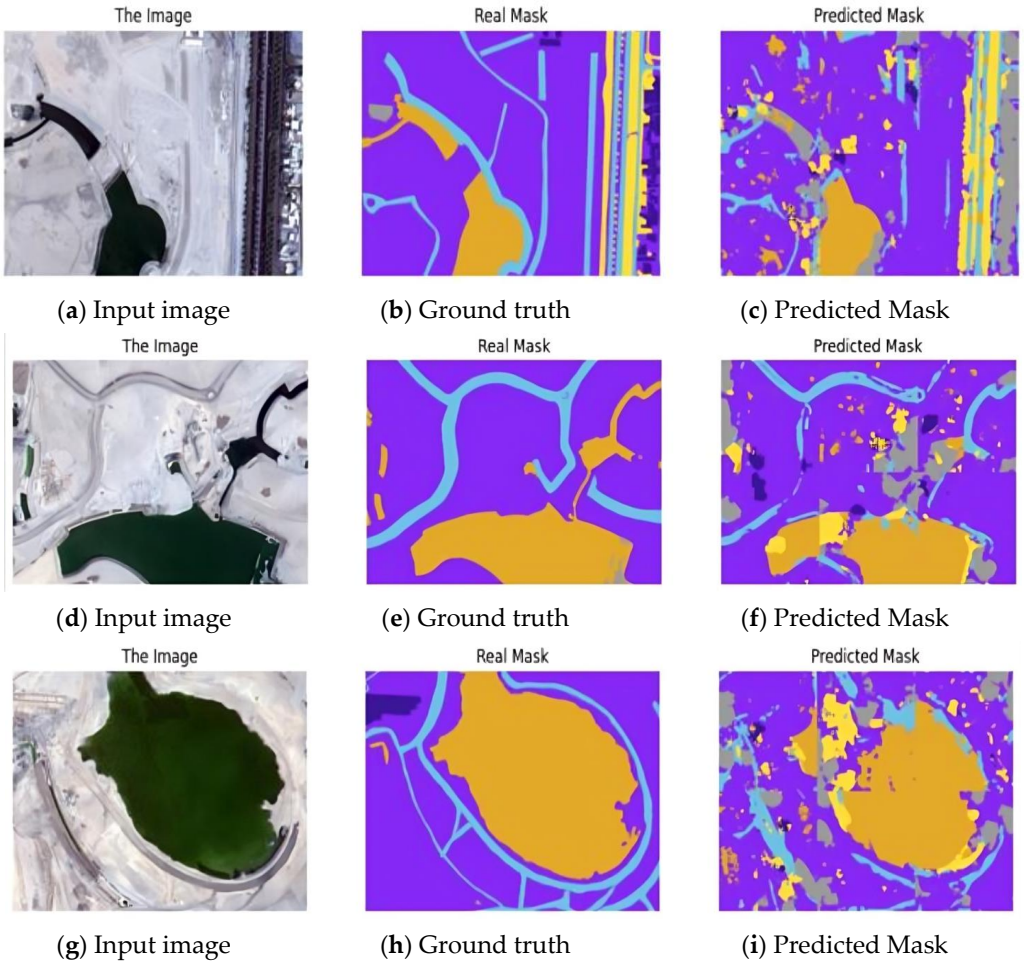


Figure 7. Predicted Images of SA-SC U-Net.

4.5. Plotting IOU vs. Validation IOU

Analyzing the Intersection over Union (IoU) and validation IoU of the SA-SC U-Net model provides crucial insights into its segmentation accuracy. IoU measures the overlap between predicted and true masks, while validation IoU extends this assessment to unseen data. Elevated IoU values, especially in the validation set, signify the model's proficiency in accurately delineating object boundaries. Consistent improvement in both IoU and validation IoU during training indicates effective learning and robust generalization. Monitoring these metrics for the SA-SC U-Net ensures a comprehensive understanding of its segmentation performance, guiding adjustments to enhance accuracy and reliability in real-world applications. A higher IoU indicates better alignment between predicted and actual segmentation, signifying the model's ability to precisely identify and delineate objects in unseen data. Monitoring the IoU and IoUv trends throughout training offers insights into the model's generalization capability, ensuring its effectiveness in real-world scenarios. These metrics serve as valuable benchmarks for evaluating and fine-tuning the model's segmentation performance during the development and validation phases.

4.6. Plotting Loss vs. Validation Loss

Examining the loss and validation loss of the proposed SA-SC U-Net model is essential for gaging its learning dynamics and performance on unseen data. The loss metric as-

sesses the dissimilarity between predicted and actual segmentation masks during training, with a decreasing trend indicating effective learning. Concurrently, the validation loss offers insights into the model's generalization capabilities as can be seen in Figure 8. A diminishing validation loss, coupled with a decreasing training loss, highlights the model's ability to learn and generalize well. Monitoring these metrics is critical for refining training parameters and addressing challenges like overfitting or underfitting, ensuring optimal performance of the SA-SC U-Net in semantic segmentation tasks. Examining the loss and validation loss of the proposed SA-SC U-Net model is essential for gaging its learning dynamics and performance on unseen data. The loss metric assesses the dissimilarity between predicted and actual segmentation masks during training, with a decreasing trend indicating effective learning. Concurrently, the validation loss offers insights into the model's generalization capabilities. A diminishing validation loss, coupled with a decreasing training loss, highlights the model's ability to learn and generalize well as demonstrated and shown in Figure 9. Monitoring these metrics is critical for refining training parameters and addressing challenges like overfitting or underfitting, ensuring optimal performance of the SA-SC U-Net in semantic segmentation tasks.

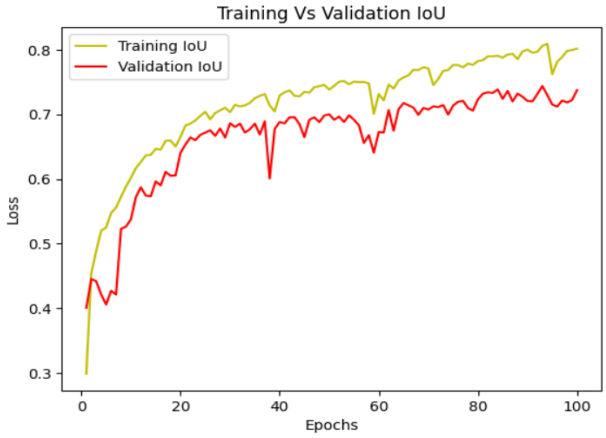


Figure 8. IoU vs. validation IoU.

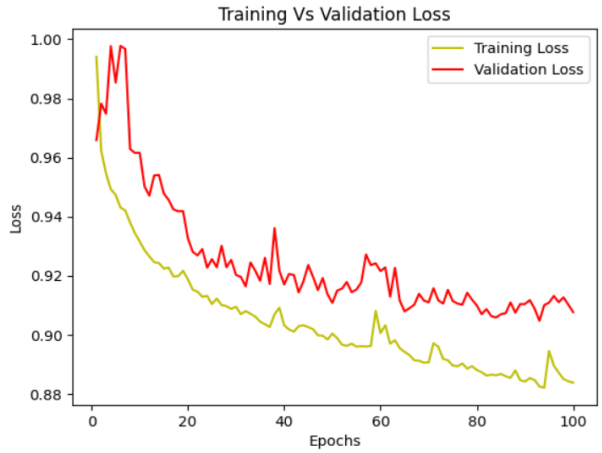


Figure 9. Loss vs. validation loss.

5. Conclusions

In conclusion, the proposed methodology, employing an SA-SC U-Net architecture for semantic segmentation, has proven to be highly effective in handling large-scale image datasets. The workflow, from dataset preprocessing to model training and prediction, showcased a systematic approach to extracting meaningful information from complex visual data. By dividing large images into manageable patches and incorporating self-attention mechanisms along with separable convolutions, the SA-SC U-Net demonstrated

superior performance compared to other established architectures like FCN, U-Net, and Dense + U-Net. The custom loss function, a combination of Dice Loss and Focal Loss, contributed to the model's robustness and accuracy, ensuring precise segmentation across multiple classes. The model's parameters, including image dimensions, class weights, and training settings, were carefully tuned to achieve optimal results.

The comparative evaluation against existing models revealed the SA-SC U-Net's superiority in terms of Intersection over Union (IoU) across various classes. Its consistently high IoU scores, especially for critical classes like Building and Water, underscored its capability to discern fine details and complex spatial relationships. The training history, illustrated by the plots of model loss and IoU over epochs, provided insights into the learning process. The model exhibited steady convergence and superior generalization on the validation set, demonstrating its capacity to learn intricate patterns and generalize well to unseen data. In summary, the proposed SA-SC U-Net stands as a powerful tool for semantic segmentation tasks, offering state-of-the-art performance and showcasing its potential for real-world applications in fields such as urban planning, environmental monitoring, and beyond. The success of this project highlights the significance of tailored architectures and thorough parameter tuning in achieving superior results in computer vision tasks.

6. Future Work

Certainly, in future work, exploring real-time segmentation and prediction would be an exciting and valuable avenue to pursue. Real-time applications often present unique challenges, such as the need for low-latency processing and efficient model deployment. Implementing the SA-SC U-Net architecture in real-time scenarios could involve optimizing the model for inference speed, possibly leveraging hardware accelerators like GPUs or TPUs. Additionally, integrating the model into a live-streaming or video-processing pipeline could open up possibilities for applications in autonomous systems, surveillance, or augmented reality. Fine-tuning the model parameters, considering the dynamic nature of real-time data, and addressing any computational constraints would be crucial aspects of this future exploration.

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

Funding: This research was partially supported by the Ministry of Trade, Industry and Energy (MOTIE) and the Korea Institute for Advancement of Technology (KIAT) through the International Cooperative R&D program (Project No. P0026318). This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2018R1A5A7023490); by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2024-2020-0-01789) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation); by the Artificial Intelligence Convergence Innovation Human Resources Development (IITP-2024-RS-2023-00254592) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation); and by the Police-Lab 2.0 Program (www.kipot.or.kr) funded by the Ministry of Science and ICT (MSIT, Korea) and the Korean National Police Agency (KNPA, Korea) [Project Name: Development of Intelligent CCTV System using Multi-Sensor Fusion for Police Station Jail Environment/Project Number: 220122M02].

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in the study are included in the article.

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

References

1. Long, J.; Shelhamer, E.; Darrell, T. Fully Convolutional Networks for Semantic Segmentation. *arXiv* **2014**, arXiv:1411.4038.
2. Drusch, M.; Del Bello, U.; Carlier, S.; Colin, O.; Fernandez, V.; Gascon, F.; Hoersch, B.; Isola, C.; Laberinti, P.; Martimort, P.; et al. Sentinel-2: ESA's optical high-resolution mission for GMES operational services. *Remote Sens. Environ.* **2012**, *120*, 25–36. [CrossRef]
3. Irons, J.R.; Dwyer, J.L.; Barsi, J.A. The next Landsat satellite: The Landsat data continuity mission. *Remote Sens. Environ.* **2012**, *122*, 11–21. [CrossRef]
4. Lv, J.; Shen, Q.; Lv, M.; Li, Y.; Shi, L.; Zhang, P. Deep learning-based semantic segmentation of remote sensing images: A review. *Front. Ecol. Evol.* **2023**, *11*, 1201125. [CrossRef]
5. Piramanayagam, S.; Saber, E.; Schwartzkopf, W.; Koehler, F. Supervised classification of multisensor remotely sensed images using a deep learning framework. *Remote Sens.* **2018**, *10*, 1429. [CrossRef]
6. Davis, L.S.; Rosenfeld, A.; Weszka, J.S. Region extraction by averaging and thresholding. *IEEE Trans. Syst. Man Cybern.* **1975**, *3*, 383–388. [CrossRef]
7. Yue, K.; Yang, L.; Li, R.; Hu, W.; Zhang, F.; Li, W. TreeUNet: Adaptive tree convolutional neural networks for subdecimeter aerial image segmentation. *ISPRS J. Photogramm.* **2019**, *156*, 1–13. [CrossRef]
8. Li, R.; Zheng, S.Y.; Duan, C.X.; Su, J.L.; Zhang, C. Multistage attention ResU-net for semantic segmentation of fine-resolution remote sensing images. *IEEE Geosci. Remote Sens. Lett.* **2021**, *19*, 8009205. [CrossRef]
9. Wang, Y.; Wang, L.; Lu, H.; He, Y. Segmentation based rotated bounding boxes prediction and image synthesizing for object detection of high resolution aerial images. *Neurocomputing* **2020**, *388*, 202–211. [CrossRef]
10. 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]
11. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional networks for biomedical image segmentation. In Proceedings of the MICCAI 2015, Munich, Germany, 5–9 October 2015; Springer: Berlin/Heidelberg, Germany, 2015; Volume 2015, pp. 234–241.
12. Guo, D.; Weeks, A.; Klee, H. Robust approach for suburban road segmentation in high-resolution aerial images. *Int. J. Remote Sens.* **2007**, *28*, 307–318. [CrossRef]
13. Wei, W.; Xin, Y. Feature extraction for manmade objects segmentation in aerial images. *Mach. Vis. Appl.* **2008**, *19*, 57–64. [CrossRef]
14. Zhou, Z.W.; Siddiquee, M.M.; Tajbakhsh, N.; Liang, J.M. Unet++: A nested u-net architecture for medical image segmentation. In Proceedings of the DLMIA 2018, Granada, Spain, 20 September 2018. [CrossRef]
15. Saito, S.; Arai, R.; Aoki, Y. Seamline determination based on semantic segmentation for aerial image mosaicking. *IEEE Access* **2015**, *3*, 2847–2856. [CrossRef]
16. Marmanis, D.; Schindler, K.; Wegner, J.D.; Galliani, S.; Datcu, M.; Stilla, U. Classification with an Edge: Improving Semantic Image Segmentation with Boundary Detection. *ISPRS J. Photogramm. Remote Sens.* **2016**, *135*, 158–172. [CrossRef]
17. Holliday, A.; Barekatain, M.; Laurmaa, J.; Kandaswamy, C.; Prendinger, H. Speedup of Deep Learning Ensembles for Semantic Segmentation Using a Model Compression Technique. *Comput. Vis. Image Underst.* **2017**, *164*, 16–26. [CrossRef]
18. Yu, Y.; Wang, C.; Fu, Q.; Kou, R.; Huang, F.; Yang, B.; Yang, T.; Gao, M. Techniques and Challenges of Image Segmentation: A Review. *Electronics* **2023**, *12*, 1199. [CrossRef]
19. Wu, M.; Zhang, C.; Liu, J.; Zhou, L.; Li, X. Towards accurate high resolution satellite image semantic segmentation. *IEEE Access* **2019**, *7*, 55609–55619. [CrossRef]
20. Boonpook, W.; Tan, Y.; Xu, B. Deep learning-based multi-feature semantic segmentation in building extraction from images of UAV photogrammetry. *Int. J. Remote Sens.* **2021**, *42*, 1–19. [CrossRef]
21. Mehra, A.; Mandal, M.; Narang, P.; Chamola, V. ReViewNet: A Fast and Resource Optimized Network for Enabling Safe Autonomous Driving in Hazy Weather Conditions. *IEEE Trans. Intell. Transp. Syst.* **2021**, *22*, 4256–4266. [CrossRef]
22. Anagnostis, A.; Tagarakis, A.C.; Kateris, D.; Moysiadis, V.; Sørensen, C.G.; Pearson, S.; Bochtis, D. Orchard mapping with deep learning semantic segmentation. *Sensors* **2021**, *21*, 3813. [CrossRef] [PubMed]
23. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. *arXiv* **2017**, arXiv:1706.03762.
24. Dang, L.; Pang, P.; Lee, J. Depth-Wise Separable Convolution Neural Network with Residual Connection for Hyperspectral Image Classification. *Remote Sens.* **2020**, *12*, 3408. [CrossRef]