Article

Semantic Segmentation of High-Resolution Airborne Images with Dual-Stream DeepLabV3+

Ozgun Akcay ^{1,*} , Ahmet Cumhur Kinaci ² , Emin Ozgur Avsar ¹ and Umut Aydar ¹

¹ Department of Geomatics Engineering, Faculty of Engineering, Çanakkale Onsekiz Mart University, Çanakkale 17100, Turkey; ozguravsar@comu.edu.tr (E.O.A.); umutaydar@comu.edu.tr (U.A.)

² Department of Computer Engineering, Faculty of Engineering, Çanakkale Onsekiz Mart University, Çanakkale 17100, Turkey; cumhur.kinaci@comu.edu.tr

* Correspondence: akcay@comu.edu.tr

Abstract: In geospatial applications such as urban planning and land use management, automatic detection and classification of earth objects are essential and primary subjects. When the significant semantic segmentation algorithms are considered, DeepLabV3+ stands out as a state-of-the-art CNN. Although the DeepLabV3+ model is capable of extracting multi-scale contextual information, there is still a need for multi-stream architectural approaches and different training approaches of the model that can leverage multi-modal geographic datasets. In this study, a new end-to-end dual-stream architecture that considers geospatial imagery was developed based on the DeepLabV3+ architecture. As a result, the spectral datasets other than RGB provided increments in semantic segmentation accuracies when they were used as additional channels to height information. Furthermore, both the given data augmentation and Tversky loss function which is sensitive to imbalanced data accomplished better overall accuracies. Also, it has been shown that the new dual-stream architecture using Potsdam and Vaihingen datasets produced 88.87% and 87.39% overall semantic segmentation accuracies, respectively. Eventually, it was seen that enhancement of the traditional significant semantic segmentation networks has a great potential to provide higher model performances, whereas the contribution of geospatial data as the second stream to RGB to segmentation was explicitly shown.

Citation: Akcay, O.; Kinaci, A.C.; Avsar, E.O.; Aydar, U. Semantic Segmentation of High-Resolution Airborne Images with Dual-Stream DeepLabV3+. *ISPRS Int. J. Geo-Inf.* **2022**, *11*, 23. <https://doi.org/10.3390/ijgi11010023>

Academic Editor: Wolfgang Kainz

Received: 5 October 2021

Accepted: 26 December 2021

Published: 30 December 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Keywords: deep learning; semantic segmentation; photogrammetry; multi-spectral aerial imagery; digital surface model; vegetation index; land cover classification

1. Introduction

Airborne imagery is an important data type in Earth Observation (EO) studies in terms of their capability of containing detailed geomorphological information with high resolution and fast acquisition features. In order to extract meaningful information from images, different land surface properties must be classified and labeled in accordance with the ground truth. This process, called semantic segmentation, is defined as classifying pixels having similar properties and assigning appropriate class labels. Accurate semantic segmentation of airborne images has the utmost importance in many EO applications like environmental monitoring and change detection, land use in urban areas, automatization in map update for Geographical Information Systems, fast monitoring of the effects, and quick response in natural hazards, etc. [1] as an environmental monitoring study developed an image segmentation architecture for Unmanned Aerial Vehicle (UAV) infrared thermal images with a focus on the segmentation of ground vehicles. On the other hand, semantic segmentation in Land Use and Land Cover (LULC) change detection takes a considerable part in the RS literature. Venugopal [2] explained an automatic semantic segmentation-based change detection that produces a final change between the given two input images, while [3] described a semantic segmentation method with category boundary for LULC mapping. Touzani and Granderson [4], which can be considered in the scope of GIS applications, made improvements to existing accuracy of automatic building footprint

extractions from RS images using a deep learning model. Bragagnolo et al. [5], which is one of the natural hazard studies regarding deep learning, formed and trained an image database with U-Net to identify landslide scars in a reliable manner.

For decades, traditional semantic segmentation methods have been employed in the Remote Sensing (RS) community for semantic data extraction. Widely used techniques are mostly machine learning approaches based on hand-crafted features [6], such as Maximum Likelihood Estimation (MLE), Support Vector Machines (SVMs), Random Forests (RFs), and Artificial Neural Networks (ANNs). Cortes and Vapnik [7] described SVMs data classification method which is able to map input vectors into high-dimensional feature spaces non-linearly. The two land cover classification experiments for the test areas showed that the SVMs achieve a higher level of accuracy than either the MLE or the ANNs classifier for high-dimensional data [8]. RFs as the other most used classification method in machine learning apply tree predictors and produce a class prediction based on the votes [9]. Furthermore, Mas and Flores [10] discussed the classification method ANNs which can be expressed as the basis of deep learning, and then compared the implementations of ANNs in some of the image processing software packages.

Semantic segmentation as a wide-ranging research field in computer vision is utilized not only in remote sensing but also in many other applications such as medical and biological imaging, retail services, autonomous driving, and facial recognition. Moen et al. [11] mentioned available annotated fluorescent and pathology images for single-cell segmentation techniques and indicated considerable progress in adapting semantic segmentation to biological and medical imagery. Moreover, the retail industry, which is a sector of the economy, needs operating semantic segmentation for selling their products. Hameed et al. [12] carried out a score-based mask edge improvement of Mask-RCNN to segment fruit and vegetable images in a supermarket environment. Wei et al. [13] proposed a dataset reflecting the large scale and the fine-grained nature of the product categories for the realistic checkout scenario. In autonomous driving, Hamian et al. [14] combined classical segmentation methods with deep learning to improve segmentation accuracy when compared to the model without post-processing. Müller et al. [15] created a training database and multiclass segmentation mask annotations for different participants and varying face and head poses with infrared images. Then, the authors developed a deep neural network architecture with the training database in order to detect and analyze face temperature in the context of the COVID-19 pandemic.

Semantic segmentation of images is slightly different from image classification in that prior information about the visual concepts or objects is not essential. Image classification techniques essentially require object definitions to be classified and create classes of pre-defined labels, while semantic segmentation classifies all properties within the image. An ideal image segmentation algorithm will also segment unknown objects, that is, objects which are new or unknown [16]. Airborne images contain generally homogeneous objects and this enables successful delineation by segmentation. As a result of successful image segmentation, the number of elements as a basis for the following image classification is enormously reduced. The quality of classification is directly affected by segmentation quality [17,18]. In the last decade, deep learning methods have proven great success especially in the computer vision area, and became a standard tool for many applications like object detection, artificial intelligence, scene recognition for automation, etc. One of the key innovations in these methods is that they operate directly from raw images instead of requiring the extraction of features from the images [19].

The great achievement in the computer vision domain as well as flexibility for adaptation to any image-related operation, deep learning, attracted the RS community for image segmentation studies in recent years. Many different Deep Convolutional Neural Network (DCNN) approaches have been employed in the studies for the semantic segmentation of RS images. Sherrah [20] is the first study that proposes Fully Convolutional Networks (FCN) for the semantic segmentation of airborne images. In the study, a no-downsampling strategy is proposed in order to maintain the original resolution of the input image by replacing the downsampling layers of FCN with dilated convolutions. Also, in the study,

a dual-stream hybrid model is used in order to be able to combine RGB data and DSM, but the authors state that they did not observe any contribution of DSM. Wang et al. [21] is the paper that presents the best results in the ISPRS semantic segmentation contest in terms of overall accuracy and F1 scores. In the study, a multi-connection strategy has been applied in order to fuse multiple features from different layers of FCN. Also, the model uses a so-called class-specific attention model which combines multi-scale features in the scale space. Varying scaled images are processed in the model simultaneously in order to generate a class-specific weight map and the weighted sum of score maps is then utilized for further classification. The study also includes a comparison of the proposed architecture with the other state-of-the-art networks. Sun et al. [22] address two problems and solutions of application of encoder-decoder type DCNN for the remote sensing images and reaches a good performance in the ISPRS online contest in overall accuracy. The first problem defined in the study is the structural stereotype which causes edge deterioration at image patches due to padding and pooling operations. The study proposes a random sampling training strategy and ensemble inference strategy as a solution. The difficulty of training deep networks which is the result of vanishing gradients is stated as the second problem and a novel residual topology is presented for the solution. In [10] a sequentially aggregated self-cascaded architecture that is based on the network by Visual Geometry Group (VGG) has been applied for semantic segmentation of airborne images. The proposed DCNN is a global-to-local context approach and it proves the capability of handling confusing man-made objects. Marcu and Leordeanu [23] has a dual-stream architecture that aims to obtain multi-context information by using different patch sizes which take the global context and local context into account in each stream and fuse the obtained information in later stages. The mentioned study employs adjusted VGG-Net and AlexNet for local and global contexts respectively. Piramanayagam et al. [24] is a dual-stream DCNN which is a composition of FCN and SegNet using RGB and nDSM + IR + NDVI data sets for each stream respectively. The main focus of the given study is to propose an optimal fusion stage of the features. Marmanis et al. [25] also utilize a dual-stream architecture instead of classical architectures by including height data as the second input data in education. The proposed study is aimed to minimize the edge defects occurring due to semantic segmentation by employing the Holistically-nested Edge Detection (HED) [26] and SegNet architecture for the remote sensing images in an integrated network for spectral, DSM, and nDSM datasets in a multi-scale approach. Du et al. [27] uses an incorporating DeepLabV3+ [28] deep learning model and Object Based Image Analysis (OBIA) image classification for semantic segmentation purposes of aerial images. In the combined model, at the first step, they train an aerial image and obtain predictions by using the DeepLabV3+ model. The second step utilizes a random field classification with hand-crafted features to the concatenated image that is composed of aerial image and nDSM. Afterward, the results are fused by Dempster-Shafer theory and finally, an object-constrained higher-order conditional random field optimization is applied in order to refine the classifications. Audebert et al. [29] describes a multi-stream and multi-scale network based on SegNet and ResNet architectures that extracts semantic maps at multiple resolutions. The authors conclude that late fusion is more suitable when combining several strong classifiers, while the early fusion approach is more appropriate for the FuseNet. Also, they evaluate nDSM contribution for the object detectability such as cars and buildings, while testing the early and late fusion strategies. Song and Kim [30] is a combined deep learning network that aims to leverage large-scale natural-image datasets with different characteristics in order to increase the segmentation accuracy of RS images. The proposed study utilizes multi-modal datasets in parallel by sharing the three encoder blocks. Although the model performs well and obtains promising results, it suffers from processing time. The authors also state that the model performance was dependent on the number of auxiliary data and weights. Another shortage of the model is the lack of usage of multi-sensor data (NIR, DSM) together which may effectively increase the accuracy. In [31,32], a more detailed review of deep learning methods in semantic segmentation of geospatial data is addressed.

Many traditional CNN architectures such as FastFCN, Gated-SCNN, DeepLab, Mask R-CNN [33–36] focus on semantic segmentation of ordinary RGB images which have comprehensive daily life contents. Also, the traditional CNN architectures are not initially considered for multi-modal data such as multi-spectral orthorectified images, spectral indices, and three-dimensional digital terrain models in EO applications. Therefore, semantic segmentation of remote sensing data requires optimized architectures so as to extract geospatial objects precisely. When the noteworthy semantic segmentation algorithms [37] having the potential to specialize in geospatial data are considered, DeepLabV3+ [28] stands out as a state-of-the-art CNN which is able to obtain sharper object boundaries by adding a decoder module to DeepLabV3 [38]. Moreover, DeepLabV3+ is capable of maintaining the rich contextual information from the encoder part. Although the DeepLabV3+ model has achieved great success in semantic segmentation in recent years, there is still a need for multi-stream architectural approaches of the model that can leverage multi-modal geographic datasets. Besides customizing the architecture, the effect of different training approaches on semantic segmentation performance should also be scrutinized. One of the main problems considering geospatial data classification can be expressed as the lack of high-resolution remote sensing benchmark and uncertainty in the quality of the existing datasets such as shadow effects, orthophoto distortions, and ground truth errors. Another important point is that there is no standard for patch size and it has a significant effect on the results as mentioned in [22]. Furthermore, two issues stand out, especially in the handling of geospatial data. The first is the determination of the optimal fusion phase and the other one is the usage of scale-variant data. The approaches mentioned here should be considered to provide a balance between accuracy and computational load.

In this study, the contribution of multi-modal geospatial data such as RGB, IR, and DSM to the classification accuracy was discussed by comparing them within single-stream architectures. In addition, the improvement of different data augmentation and loss functions in land cover classification accuracy were examined. Moreover, a new end-to-end dual-stream architecture that considers geospatial imagery was developed based on the DeepLabV3+ architecture. The proposed architecture is able to utilize multi-scale contextual information with atrous spatial pyramid pooling, exploit two discrete multi-channel data simultaneously in different streams and present a late fusion layer combining different high-level features. As a result, it has been shown that the new dual-stream architecture that can process rich geospatial datasets improves semantic segmentation performance with 88.87% overall accuracy. This obtained performance demonstrates that the proposed architecture provides competitive results with respect to the other state of the art models. Eventually, it was seen that enhancement of the multi-stream of the significant semantic segmentation networks has a great potential to provide higher model performances for the multi-modal geospatial datasets. Whereas the contribution of geospatial data as the second stream to RGB to segmentation was explicitly shown.

The paper is organized as follows. Section 2 explains the applied data augmentation techniques, the structure of the proposed dual-stream DeepLabV3+ network, and the employed training approaches, respectively. The results section presents the segmentation accuracies of the experiments using the Potsdam dataset and also submits the validation with the Vaihingen dataset. Section 4 as the discussion compares the findings and their implications with the previous studies. The paper then ends with the conclusions section.

2. Methods

The open-access ISPRS Potsdam dataset [39], which has been used to measure the performance of many developed semantic segmentation algorithms, was initially preferred as the dataset due to its various geospatial data supply. During the training data preparation process, two data augmentation techniques based on basic image manipulations [40] were applied to enlarge the size and improve the variety of the datasets. For the algorithm of the geospatial data-assisted proposed model, DeepLabV3+ is considered since it is a new generation network that presents high semantic segmentation performance. The existing single-stream architecture of DeepLabV3+ has been modified to dual-stream for

geospatial data as well as RGB datasets. As a result of the proposed architecture, models were produced with different training approaches and their performances were measured. Lastly, the method was also trained with the Vaihingen dataset in order to validate the performances [41].

2.1. Data Augmentation

ISPRS Potsdam dataset offers IR band and DSM/nDSM images as well as RGB true orthophoto with a ground sampling distance of 5 cm. The data set contains 38 patches with 6000×6000 pixel size and corresponding ground truth data consisting of impervious surface (white), buildings (blue), low vegetation (cyan), trees (green), cars (yellow), and clutter (red) classes. Two different data augmentation (DA) techniques have been achieved in order to generate input dataset. The First DA method, which is called DA-I in the paper, has been implemented in the GNU Octave environment before the model training. Each 6000×6000 image is divided sequentially to form 800×800 sub-images with overlapping of 150 pixels for avoiding data missing at the edges of cropped images. For each patch; flipping (non and flipped), rotation (non, 90° , 180° , and 270°) were applied in all combinations. Furthermore, contrast (non, -50% and $+50\%$) operations were applied to RGB and IR images considering the effect of reflection differences according to different image acquisition angles and shadows. Consequently, 24 augmented images for each cropped RGB and IR patches, and 8 augmented images for each cropped nDSM, NDVI, and label patches were generated. In the previous study, [42] which elaborated an example of the augmentation operations for an 800×800 RGB image depicted the results in graphic detail. For DA-II, smaller patch size is implemented as well as coincidental image operations were applied for higher batch size and faster training. In order to generate an input image, a 600×600 sub-image is randomly cropped, and flipping (non or flipped) and rotation (non, 90° , 180° , 270°) were applied. Additionally, brightness, contrast, hue, and saturation values were randomly changed between -0.05 and 0.05 , between 0.7 and 1.3 , between -0.08 and 0.08 , and between 0.6 and 1.6 , respectively (Figure 1). The main purpose of DA-II is to use more randomized inputs for the learning process of the network. The augmentation approaches and total numbers of sub-images are summarized in Table 1. The two patches 2_10 and 5_12 of the Potsdam dataset reflecting the class distribution of the full dataset were chosen as sub-dataset to evaluate the contribution of the combined use of nDSM, IR, and NDVI data in a reasonable time (Figure 2). Class distributions of the full and sub-dataset are given in Table 2. While patch 5_12 provides more building content, patch 2_10 involves intensive vegetation composition. The patches 2_10 and 5_12 were used together, rather than a single patch, to yield more balance among the class distributions in the sub-dataset. Therefore, the sub-dataset was essentially constructed to make a quick pre-evaluation of different experiments. For dual-stream models, RGB data is the input for the first stream, and the rest for the second with different options to analyze their contributions to model performance.

Figure 1. DA-II Sample images.

Figure 2. ISPRS Potsdam full-datasets and sub-datasets in UTM WGS84 datum.

Table 1. The applied augmentation operations for the Potsdam dataset.

Augmen- Tation	Dataset	Input Data	# of Sub-Images for Each Patch	Operations						# of Total Sub-Images	
				Rotation	Flipping	Contrast	Brightness	Hue	Saturation	Training	Test
DA-I	sub	RGB IR	81	non, 90°, 180°, 270°	yes/no	non, −50%, +50%	-	-	-	3499	389
		nDSM NDVI Label	81	non, 90°, 180°, 270°	yes/no	-	-	-	-	1166	130
	full	RGB IR	81	non, 90°, 180°, 270°	yes/no	non, −50%, +50%	-	-	-	46,656	1134
		nDSM NDVI Label	81	non, 90°, 180°, 270°	yes/no	-	-	-	-	15,552	1134
DA-II	full	RGB IR	100	non, 90°, 180°, 270° randomly	yes/no randomly	(0.7–1.3)	(−0.05, 0.05)	(−0.08, 0.08)	(0.6, 1.6)	96,000	1134
		nDSM NDVI Label	100	non, 90°, 180°, 270° randomly	yes/no randomly	-	-	-	-	~19,200	1334

Table 2. Class distribution of Potsdam training dataset.

Usage Status	Patch Coverage	Imp. Surf. (%)	Building (%)	Low Veg. (%)	Tree (%)	Car (%)	Clutter (%)
-	2_10	20.65	15.13	42.17	7.44	0.87	13.73
-	5_12	31.23	48.62	8.92	5.92	3.06	2.26
sub-dataset for Training	2_10 and 5_12	25.94	31.87	25.55	6.68	1.96	8.00
Full dataset for Training	24 patches including 2_10 and 5_12	28.46	26.72	23.54	14.62	1.69	4.96
Test dataset	14 patches	31.47	23.92	21.01	17.04	1.96	4.59

The ISPRS Vaihingen dataset was used as a second validation dataset since it supplies the necessary data for the proposed algorithm and has similar classes. Although there are some feature differences, it is another open-access dataset that has been used in many studies, like the ISPRS Potsdam dataset. The main differences are; the true orthophotos of the dataset have a combination of IRRG instead of RGB, the ground sampling distance of both true orthophoto and DSM is 9 cm, and the dataset has 33 patches with different dimensions [41]. Moreover, the dataset only provides DSM, therefore nDSM patches generated by [43] were used in this study.

2.2. Dual-Stream DeepLabV3+

DeepLabV3+ which has encoding and decoding stages is a state-of-the-art deep learning model for semantic image segmentation [28,35]. The encoding stage extracts basic information from the image using a CNN whereas the decoding stage reconstructs outputs based on information obtained from the encoder stage. The decoder resamples low resolution encoded images to the original image size in order to produce better segmentation results across object boundaries. DeepLabV3+ supports the following network backbones: MobileNetv2, Xception, ResNet, PNASNet, Auto-DeepLab. McNeely-White et al. [44] indicated that minor accuracy effects are expected in the case of the usage of some different backbones. In this respect, a particular research study should be envisaged for the assessment of backbones. In this study, the Xception network backbone was chosen to train the DeepLabV3+ model because authors of DeepLabV3+ noted that a modified version of Xception as the backbone promised better ImageNet accuracy.

Most of the common machine learning models for semantic segmentation are trained with images with RGB channels. [45] listed some popular semantic segmentation datasets developed for different tasks in computer vision such as remote sensing, autonomous driving, and indoor applications. However semantic segmentation studies in LULC specifically need to utilize benchmarks including main land cover classes, high-resolution multi-spectral images, and the elevation of the terrain. SkyScapes provided annotations for 31 semantic categories such as buildings, roads, and vegetation, to perform semantic segmentation tasks [46]. However aerial images of the dataset do not include any band out of visible range although the labels present rich semantic land-use class categories. Also, Boguszewski et al. [47] introduced annotated land cover dataset with four class objects as buildings, woodlands, water, and roads using RGB aerial images. Therefore, except for medium spatial resolution satellite image datasets, the Potsdam dataset stands out due to its geospatial content. In our case, the Potsdam dataset provides IR band and elevation data additionally when compared to ordinary RGB image benchmarks such as SkyScapes and ImageNet. A training process with the mentioned additional data has considerable potential to improve semantic segmentation performance. Also generated data like NDVI can be informative to help improve model accuracy. However, handling different data types besides ordinary three-band RGB images is crucial as it needs some special approaches in deep learning. The main problem with the additional data which constructs an image stack

having more than three channels cannot be fed directly to the state-of-the-art semantic segmentation architectures. The first solution which addresses the obstacle is to use the additional data with RGB image as one combined dataset and to train defined models with this data. Nevertheless, the higher batch sizes due to an increase in input dimensions will inevitably suffer the computation process. Also, transfer learning cannot be used because of incompatible input dimensions. For these reasons, in this study, RGB images and nDSM + IR + NDVI were considered separate datasets during the training process in two independent models. Then, outputs of these two models are combined in order to obtain a final semantic map. Figure 3 depicts the proposed dual-stream DeepLabV3+ which is designed to deal with all input geospatial datasets.

Dual-stream Xception based DeepLabV3+ considered two parallel streams as encoders so as to exploit both RGB and nDSM + IR + NDVI data sets together. Atrous Spatial Pyramid Pooling (ASPP) which is utilized in the architecture is capable of encoding multi-scale contextual information. However, the main differences of dual-stream structure from the original DeepLabV3+ occur at the decoder stage. First, the decoder of the proposed model employs 1×1 convolutions for the extracted features from both streams and then applies bilinear upsampling by a factor of 4. In two separate streams, 3×3 convolutions were operated after upsampling and 1×1 convolutions were harnessed. Afterward, the decoder concatenated features from two streams, then applied a 1×1 convolution and a bilinear upsampling by a factor of 4 to extract final semantic segmentation images.

Figure 3. Dual-stream DeepLabV3+.

2.3. Training Approaches

In deep learning, training strategies substantially influence the segmentation and classification performance of the networks. In particular, four factors of the training phase should be considered delicately in order to produce more efficient results. In this study, the effect of various training implementations was also examined to determine more accurate models. The first training feature is to resolve an appropriate loss function in order to measure cost between the estimated value and true value. Categorical cross-entropy loss which requires softmax activation is a widely used function for multi-class classification problems as implemented in semantic segmentation. However, unequal distribution of classes in the training dataset as seen in Table 2 suffers performances of

categorical cross-entropy function. The other drawback of the function is ineffectiveness in the F1 score. For these reasons, the Focal Tversky Loss (*FTL*) function is tested for the final model training phases. *FTL* is a generalized focal loss function based on the Tversky Index (*TI*) [48]. *TI*, which is seen as Equation (1), is an asymmetric similarity measure calculated with True Positive (*TP*), False Negative (*FN*), and False Positive (*FP*) metrics. α and β are the coefficients which control *FN*, *FP* respectively, where $\alpha, \beta > 0$. While $\alpha \neq \beta$, the sum of α and β is expected to be 1 in *TI* as a generalization of the Sorensen–Dice coefficient and the Jaccard index [49]. Provided that any value of α is greater than β , the imbalanced image segmentation datasets can be handled more effectively. On the other hand, the γ parameter that determines the behavior of the *FTL* function in Equation (2) varies in the range [1,3]. Although the effect of γ relatively changes depending on *TI* value, in their experiments, Abraham and Khan [48] stated that the best performance was observed with $\gamma = 4/3$. Subsequently, parameter values were set as $\alpha = 0.7$, $\beta = 0.3$ and $\gamma = 4/3$.

$$TI = \frac{TP}{TP + \alpha FN + \beta FP} \quad (1)$$

$$FTL = (1 - TI)^{\frac{1}{\gamma}}, \quad (2)$$

Transfer learning, which is a machine learning methodology to reuse a model developed for one task to another task, is considered a second training factor. Transfer learning significantly improves network performance while reducing learning time. However, transfer learning may not be applicable because the structure of the source models is not suitable. Therefore, the status of transfer learning was also regarded as an indicative method in the model training. In the study, transfer learning was accomplished using original DeepLabV3+ model weights trained with Pascal VOC dataset and Xception backbone as shown in Figure 4. Also, the proposed models which utilize transfer learning were trained using the Adam optimization algorithm with a learning rate of 0.01 on the last layers particular to classification.

Figure 4. Training steps.

After transfer learning, fine-tuning was implemented as the third training factor. All layers of the models were trained using stochastic gradient descent (SGD) with a low learning rate of 0.0001 and momentum of 0.9. Lastly, patch dimensions used as network input were revised in order to reduce training durations and optimize memory usage.

3. Results

In order to monitor the improvement in semantic segmentation, networks were trained with different georeferenced data sets such as multi-spectral orthophotos, nDSM, and NDVI. Network analysis was performed, with one stream up to three channels and two streams for six channels data input requirements. In this way, it is aimed to clearly reveal the contribution of data sets to semantic segmentation for land cover mapping. In order to reduce training durations, some models were initially produced using sub-dataset due to the huge amount of data processing needed. Then, some of these models were created using the full dataset according to the preliminary evaluation results obtained with sub-dataset. All produced models are numbered with the initial letter M for reference and comparison in the text, and as a result, 16 models from M1 to M16 are presented. Sections 3.1 and 3.2 presents results obtained using the Potsdam dataset, whereas Section 3.3 indicates the evaluations with the Vaihingen dataset.

3.1. Evaluation of Single-Stream Models

Single-stream networks produce models that are fed by data consisting of one, two, and three channels as input. All models from M1 to M14 were designed as a single-stream network. Models up to M8 of these were only trained with the sub-dataset as seen in Table 3, while the models from M9 to M14 were implemented with the full dataset.

Table 3. Features of different training approaches and inputs using sub-datasets for single-stream segmentation models.

Model	Loss Function	Transfer Learning	Fine Tuning	Patch Size	Data Augmentation (DA)	Inputs Based on Stream Structure		Dataset
						First Stream	Second Stream	
M1	cross entropy	-	-	800 × 800	DA-I	nDSM	-	sub
M2	cross entropy	-	-	800 × 800	DA-I	nDSM + IR	-	sub
M3	cross entropy	-	-	800 × 800	DA-I	nDSM + NDVI	-	sub
M4	cross entropy	-	-	800 × 800	DA-I	nDSM + IR + NDVI	-	sub
M5	cross entropy	+	-	800 × 800	DA-I	nDSM + IR + NDVI	-	sub
M6	cross entropy	+	+	800 × 800	DA-I	nDSM + IR + NDVI	-	sub
M7	cross entropy	+	-	800 × 800	DA-I	R + G + B	-	sub
M8	cross entropy	+	+	800 × 800	DA-I	R + G + B	-	sub

Since there are no specific pre-trained models with nDSM, IR, and NDVI datasets, the first four models M1, M2, M3, and M4 were trained with randomly initialized weights without transfer learning. This situation, in which transfer learning was not implemented, has affected the model successes negatively, as can be seen in Table 4. On the other hand, the second half of the first eight models from M5 to M8 were examined with transfer learning.

Especially in the first four models, the importance of nDSM, IR, and NDVI in semantic segmentation was discussed by sequentially increasing the number of channels. When Tables 3 and 4 are evaluated together, it is understood that the contribution of IR and nDSM datasets to segmentation performance is higher than the NDVI dataset. In other

words, the IR band and nDSM can be seen as more essential geospatial datasets in semantic segmentation. Since NDVI data includes the IR channel, it can be considered that the simultaneous presence of NDVI and IR as the input channels of architecture will not contribute to the overall accuracy. However, a 3.87% increase between M2 and M4 in Table 4 indicates extra improvement when NDVI and IR are combined. On the other hand, it may be possible for the M2 to reach the accuracy of the M4, but this may require a longer training time.

Table 4. Overall test accuracies obtained using sub-dataset for single-stream segmentation models.

Model	M1	M2	M3	M4	M5	M6	M7	M8
Overall (%)	30.15	48.52	35.30	52.39	68.62	86.67	69.85	87.74

Since input data of M5 and M6 has three channels, these models were considered to use transfer learning with pre-trained RGB models. Although data types of M5 and M6 are different from conventional RGB data, transfer learning contributed to their model accuracies. In addition, model M6, using fine-tuning which outperformed the models M4 and M5, explicitly outlined the improvement with 86.67% accuracy among models with three geospatial data channels. Therefore, the best overall accuracy is obtained from M6 which is trained with data containing nDSM, IR, and NDVI bands (Table 4). One stream models utilizing RGB data were trained with transfer learning and fine-tuning as M7 and M8, respectively. When Table 4 is examined, the fact that M5 and M7 offer accuracies that are close to each other shows that RGB-based pre-trained models can be used in transfer learning for geospatial data. Moreover, it was also found that M6 almost achieved as much accuracy as M8. This can be seen as another indicator of the contribution of RGB-based transfer learning and fine-tuning in geospatial data.

As the full dataset provides much richer content for the networks, it is expected that the performance of test data will be increased on the trained models. Models from M9 to M14 are single-stream networks generated with the full dataset as explained in Table 5. After eight experiments from M1 to M8 with the sub-dataset were implemented and assessed, the models from M5 to M8 were reconsidered for further training in order to evaluate them with full dataset experiments. The models M5, M6, M7, and M8 were renamed for the full dataset assessment as M9, M10, M12, and M13, respectively. Besides the renamed models, new models M11 and M14 which managed focal Tversky loss function were created with DA-II augmentation type.

Table 5. Features of different training approaches and inputs using full dataset for single-stream segmentation models.

Model	Loss Function	Transfer Learning	Fine Tuning	Patch Size	Data Augmentation (DA)	Inputs Based on Stream Structure		Dataset
						First Stream	Second Stream	
M9	cross entropy	+	-	800 × 800	DA-I	nDSM + IR + NDVI	-	full
M10	cross entropy	+	+	800 × 800	DA-I	nDSM + IR + NDVI	-	full
M11	focal Tversky	+	+	600 × 600	DA-II	nDSM + IR + NDVI	-	full
M12	cross entropy	+	-	800 × 800	DA-I	R + G + B	-	full
M13	cross entropy	+	+	800 × 800	DA-I	R + G + B	-	full
M14	focal Tversky	+	+	600 × 600	DA-II	R + G + B	-	full

Table 6 indicates class-based F1 scores and overall accuracies which are obtained after training of full dataset with single-stream models. When examining the contribution of fine-tuning in Table 6, the fine-tuning using the full dataset also yielded better semantic

segmentation in the models M10 and M13, as similarly discussed before in sub-dataset experiments. Moreover, it is clearly seen that overall accuracies of M11 and M14 surpassed accuracies of M10 and M13, respectively. Three main factors which are loss function, patch size, and data augmentation method lead to boost the mentioned accuracy values. To constitute this optimization, three changes made can be elaborated as follows: (a) a different augmentation scheme (DA-II) was tested; (b) focal Tversky loss function was used which balances false positives and false negatives; (c) batch size was increased by reducing input size from 800×800 to 600×600 for a more stable training.

Table 6. Test scores obtained using full-dataset for single-stream segmentation models. All quality measures except for overall accuracy are F1 scores in [%].

Model	Imp. Surf. (%)	Building (%)	Low Veg. (%)	Tree (%)	Car (%)	Overall (%)
M9	59.38	54.34	49.15	42.53	8.75	51.74
M10	82.97	93.61	75.91	77.14	63.82	81.08
M11	89.95	95.83	83.34	84.07	88.71	87.46
M12	65.25	69.94	57.42	65.40	53.13	63.19
M13	88.52	92.69	82.44	83.59	84.97	85.80
M14	90.95	95.12	83.86	85.12	88.81	87.98

In addition, quite close overall accuracy values which are seen between M11 and M14 directly comparing nDSM + IR + NDVI datasets with R + G + B datasets can be stated as remarkable. The closeness in segmentation performance indicates that geospatial datasets supported with IR channel information might be considered to substitute conventional RGB imagery datasets. On the other hand, when the class-based inspection is carried out, F1 scores of M11 and M14 largely demonstrate similar segmentation performances. Particularly the score of the lower vegetation class highlights the need for improved segmentation while the score of building class outweighs any class-based achievement.

3.2. Evaluation of Dual-Stream Models

In order to construct a dual-stream model capable of processing six-channel inputs totally, a single-stream network that utilizes R + G + B datasets was associated with a nDSM + IR + NDVI based network at once as explained in Figure 3. In this study, two dual-stream models M15 and M16 were implemented to examine the contribution of geospatial and imagery datasets at the same training session. Model M15 employed the network structure of M10 and M13 together, while the networks of models M11 and M14 were handled in model M16.

Two dual-stream models M15 and M16 are trained with properties given in Table 7. The results show that the dual-stream model M16 has a better overall accuracy value than all single-stream models (Table 8). M16 which is different from M15 in terms of input size, loss function, and data augmentation performed better on test data. In order to refine segmentation results, a center-cropped selection method while selecting an input was also applied. The performances obtained from M16 with the center-cropped selection method were indicated as M16* in Table 8. For a given image, multiple input images were created and only the central region of output was selected and then patched to create the final segmentation map. This method is employed only after model training and has no effect on the training process. The main reason for using center cropped results is that model outputs may have unstable results on the borders of input images and on the contrary the center of outputs is more stable. Eventually, the center-cropped prediction which was given as M16* helped to get slightly better overall accuracy from model M16. On the other hand, improvements in the class-based F1 scores are also observed in Table 8. Besides F1 scores, class accuracies that are obtained from the model M16* are illustrated with the confusion matrix in Table 9. The confusion matrix explicitly presents that the clutter class revealed the lowest accuracy value due to incorrect assignment. Clutter class

substantially appeared as an impervious surface or lower vegetation as a result of the commission error. The main reason for some class misalignments is regarded as difficulty during the preparation of ground truth data instead of failure in the model. Finally, Figure 5 which enables visual comparison between ground truth and prediction depicts output results of semantic segmentation using model M-16* for the given Potsdam test dataset. The accuracy and loss curves of the M11, M14, and M16 models which enable monitoring changes in learning performance during the training process were demonstrated in Figure 6. To illustrate in the figure, M16 was chosen as the best-trained model while M11 and M14 were chosen as employed by M16.

Figure 5. Potsdam test dataset semantic segmentation output of model M-16* with center cropped predictions. **White:** Impervious surfaces, **Blue:** Building, **Cyan:** Low vegetation, **Green:** Tree, **Yellow:** Car, **Red:** Clutter/background.

Figure 6. Potsdam models M11, M14, and M16 training accuracy and loss.

Table 7. Features of different training approaches and inputs using the full datasets for dual-stream segmentation models.

Model	Transfer Learning	Fine Tuning	Patch Size	Loss Function	Data Augmentation (DA)	Inputs Based on Stream Structure		Dataset
						First Stream	Second Stream	
M15	+	+	800 × 800	cross entropy	DA-I	R + G + B	nDSM + IR + NDVI	full
M16	+	+	600 × 600	focal Tversky	DA-II	R + G + B	nDSM + IR + NDVI	full

Table 8. Test scores obtained using full-dataset for dual-stream segmentation models. All quality measures except for overall accuracy are F1 scores in [%].

Model	Imp. Surf. (%)	Building (%)	Low Veg. (%)	Tree (%)	Car (%)	Overall (%)
M15	89.71	89.71	83.49	84.97	86.39	87.20
M16	91.18	96.14	84.85	85.56	89.07	88.71
M16*	91.34	96.07	85.17	85.80	89.58	88.87

Table 9. Confusion matrix produced from test data predictions using the model M16*.

	Imp. Surf. (%)	Building (%)	Low Veg. (%)	Tree (%)	Car (%)	Clutter (%)
Imp. surf.	91.92	1.39	3.47	1.46	0.50	1.26
Building	1.33	96.92	0.69	0.25	0.01	0.80
Low veg.	4.03	0.80	88.69	5.20	0.03	1.24
Tree	3.39	0.60	12.57	82.66	0.39	0.39
Car	5.33	0.24	0.23	0.56	92.73	0.92
Clutter	23.74	9.71	15.48	1.85	1.02	48.19

3.3. Evaluation of the Vaihingen Dataset

The Vaihingen dataset having different spatial resolutions was used in order to validate the proposed end-to-end dual-stream architecture. Firstly M11 and M14 were essentially implemented to fulfill the training strategies explained in Figure 4. The results in Table 10 demonstrated that the single-stream M11 utilizing nDSM, IR, and NDVI achieved the accuracy with 84.86% and then, M14 exploiting visible bands RGB produced a considerable accuracy with 87.18%. On the other hand, M16 and M16* which carried out dual-stream architecture accomplished higher accuracy rates with 87.33% and 87.39% than each single-stream.

Table 10. Test scores obtained using Vaihingen dataset for dual-stream segmentation models. All quality measures except for overall accuracy are F1 scores in [%].

Model	Imp. Surf. (%)	Building (%)	Low Veg. (%)	Tree (%)	Car (%)	Overall (%)
M11	86.51	90.84	77.83	85.59	20.78	84.86
M14	89.06	93.40	79.32	86.61	78.75	87.18
M16	89.21	93.53	79.26	86.87	78.64	87.33
M16*	89.58	93.72	78.88	86.74	78.85	87.39

4. Discussion

In this study, a new end-to-end dual-stream architecture that considers multi-band geospatial imagery based on the DeepLabV3+ was proposed. The ISPRS 2-D Semantic Labelling datasets were used to evaluate our model. The model operates as two parallel streams and fuses the final features from each stream at the end. With the help of this multi-modal dual structure, the architecture gains the capability of handling IR band, DSM, and NDVI datasets beside RGB images. On the other hand, another motivation of the study is to investigate the effects of the varied loss functions, data augmentation, and input

selection strategies on classification accuracy. At the evaluation phase, we conducted tests with different combinations which are given through Tables 3–8. As a result of the tests, 88.9% overall accuracy was obtained with the model named M16*. In order to compare our results, we choose the studies given in Table 11. All these studies use the same dataset and have similarities with our proposal in terms of architecture and data usage. Wang et al. [21] achieves the best overall accuracy result in the ISPRS 2-D Semantic Labelling contest. The study proposes a multi-connection ResNet architecture to fuse multi-level deep features corresponding to different layers and also a fusion strategy for multi-scale features. Audebert et al. [29] presents a multi-stream, multi-scale deep fully convolutional neural network based on SegNet and ResNet architectures that extracts semantic maps at multiple resolutions. The study also investigates the effects of early and late fusion of DSM and multi-spectral data on overall accuracy. Piramanayagam et al. [24] proposes a FCN or SegNet based fusion network consisting of two or more streams to be able to use multi-channel data (IR, R, G for one stream and IR, NDVI, nDSM for the second). The study investigates the optimal fusion time for the features. For this reason, authors conduct varied configurations which they call late fusion (LaFSN), proposed composite fusion (CoFSN), and fusion after n^{th} layer (LnFsn). It should be noted that the given results in Table 11 belong to late fusion configuration and do not reflect their best results which are 90.3% in overall accuracy. The reason that we choose this configuration is that it is identically similar to our proposal except for the used DCNN architecture. In [30], a combined deep learning network is proposed in order to improve the segmentation accuracy by using oblique urban images [50] and RS images together. The proposed architecture handles two inputs in parallel by sharing the three encoder blocks. Then the network is trained with the weighted loss which is defined as the weighted sum of losses of the two paths. The study also includes the segmentation results of other networks such as SegNet, DeepLab-V3+ with 4 band data (RGB + NIR). Apart from the used network architectures themselves, there are some strategies such as multi-scale data usage, multi-connection, and optimal fusion time, which seem the most prominent ones which we also consider to adapt for our dual-stream model as future work. These obtained results show that the produced model based on our architecture provides competitive results with respect to the other state-of-the-art models. As it is seen that the four-band DeepLabV3+ [30] had obtained a lower overall accuracy when compared with our single-stream models M11 and M14 presented in the result section, although they utilize the Potsdam datasets on DeepLabV3+ architectures as well. Therefore, some deviation should be considered during the direct comparison of the accuracies produced from such different architectures and models due to data augmentation or loss function. We believe that the segmentation accuracies of our proposed architecture can be increased with some modifications especially by focusing on the fusion time as it is proposed in [21,24,29].

As seen in Figure 7, the test results were evaluated from the models obtained by training Potsdam and Vaihingen datasets on the related architectures. Accordingly, it is understood that the model results show a compatible performance in different datasets. On the other hand, it is observed that the Vaihingen data set produces relatively less accuracy. In addition, in other related studies, it is also seen that the Potsdam dataset provides better accuracy than the Vaihingen dataset. This is considered to occur due to the lower spatial resolution of the Vaihingen dataset.

Table 11. Performance metrics of various methods for ISPRS Potsdam dataset.

Model	Motivation	Imp. Surf. (%)	Building (%)	Low Veg. (%)	Tree (%)	Car (%)	Clutter (%)	Based Architecture
SWJ_2 [21]	(1) Multi-connection (2) Fusion of Multi-scale features	94.4	97.4	87.8	87.6	94.7	91.7	Resnet
V-FuseNet [29]	(1) Multi-modal architecture (2) Optimal fusion time	92.7	96.3	87.3	88.5	95.4	90.6	Segnet Resnet
Our Proposal	(1) Dual-stream architecture (2) Different training, loss function and training strategies	91.3	96.1	85.2	85.8	89.6	88.9	DeepLabV3+
RIT (LaFSN) [24]	(1) Dual-stream architecture (2) Investigation of optimal fusion time	90.6	95.9	83.5	83.3	93.1	87.9	FCN-8 Segnet
Combined U-net Case2 [30]	From U-net study	89.2	92.9	79.7	79.8	88.5	87.2	U-net
DeepLabV3+ (4 Band) [30]	From U-net study 4Band	88.8	91.9	80.1	82.4	87.7	86	DeepLabV3+

Figure 7. Comparison of overall accuracies of the trained architectures.

5. Conclusions

The contribution of IR and NDVI data to segmentation was explicitly observed in the single-stream models. It has been realized that using two data sets together boosts the performance when compared to the separate uses of IR and NDVI. When we examine the triple usage of nDSM, IR, and NDVI data sets, models achieved as high accuracy as single-stream RGB models, and this situation should be considered in more detail in future works. Another point determined as a result of the study shows the importance of training processes. Transfer learning, fine-tuning, and the selection of loss function, data augmentation method contributed to model performance. While the Travesky method stands out as the loss function, it has been observed that the randomization of data feeding is a considerable technique for achieving higher accuracies.

Since the single-stream semantic segmentation models yield successful results with both RGB inputs and nDSM + IR + NDVI inputs, a dual-stream architectural design has also been implemented. One of the most important achievements of the study is that dual-stream deep learning models produce higher accuracy than single-stream models. As a result, it was revealed that the photogrammetric products such as nDSM and NDVI obtained using four-band multi-spectral aerial photographs should be considered in semantic segmentation beside RGB datasets. In order to determine the contribution level of NDVI, the performance

of NDVI + GB and IR + RG can be compared to single-stream architectures in the future. Also, in dual-stream, while RGB is given as input of the first stream; NDVI and IR can be considered separately as inputs of the second stream.

This study is based on DeepLabV3+ as a pioneering semantic segmentation architecture, however, other segmentation architectures can also be considered to assess multi-spectral image datasets in the land cover classification. Besides, different backbones which enable transfer learning might be adapted to DeepLabV3+ architecture. The analyses considering five land use classes were evaluated in this study, however, contributions of multi-spectral bands and different indices to the classification can be dealt with based on only one land class such as buildings. On the other hand, the architecture was designed only as the late fusion between the streams. However higher accuracy is observed in the studies implementing early fusion. Therefore late fusion is considered as the limitation of our proposed models, although it needs less computation during training.

Author Contributions: Conceptualization, Ozgun Akcay, Ahmet Cumhur Kinaci, Emin Ozgur Avsar and Umut Aydar; methodology, Ozgun Akcay and Ahmet Cumhur Kinaci; software, Ahmet Cumhur Kinaci; validation, Ahmet Cumhur Kinaci; data curation, Emin Ozgur Avsar; writing—original draft preparation, Ozgun Akcay, Ahmet Cumhur Kinaci, Emin Ozgur Avsar and Umut Aydar; writing—review and editing, Ozgun Akcay, Ahmet Cumhur Kinaci, Emin Ozgur Avsar and Umut Aydar; visualization, Ozgun Akcay and Emin Ozgur Avsar; supervision, Ozgun Akcay; project administration, Ozgun Akcay; All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by The Scientific and Technological Research Council of Turkey (TÜBİTAK), Project No: 119Y363.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data used in this study is available in [39] upon request.

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

References

1. Masouleh, M.K.; Shah-Hosseini, R. Development and evaluation of a deep learning model for real-time ground vehicle semantic segmentation from UAV-based thermal infrared imagery. *ISPRS J. Photogramm. Remote Sens.* **2019**, *155*, 172–186. [\[CrossRef\]](#)
2. Venugopal, N. Automatic semantic segmentation with DeepLab dilated learning network for change detection in remote sensing images. *Neural Processing Lett.* **2020**, *51*, 2355–2377. [\[CrossRef\]](#)
3. Xu, Z.; Su, C.; Zhang, X. A semantic segmentation method with category boundary for Land Use and Land Cover (LULC) mapping of Very-High Resolution (VHR) remote sensing image. *Int. J. Remote Sens.* **2021**, *42*, 3146–3165. [\[CrossRef\]](#)
4. Touzani, S.; Granderson, J. Open Data and Deep Semantic Segmentation for Automated Extraction of Building Footprints. *Remote Sens.* **2021**, *13*, 2578. [\[CrossRef\]](#)
5. Bragagnolo, L.; Rezende, L.; da Silva, R.; Grzybowski, J. Convolutional neural networks applied to semantic segmentation of landslide scars. *CATENA* **2021**, *201*, 105189. [\[CrossRef\]](#)
6. Kanwal, S.; Uzair, M.; Ullah, H. A Survey of Hand Crafted and Deep Learning Methods for Image Aesthetic Assessment. *arXiv* **2021**, arXiv:2103.11616.
7. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297. [\[CrossRef\]](#)
8. Pal, M.; Mather, P. Support vector machines for classification in remote sensing. *Int. J. Remote Sens.* **2005**, *26*, 1007–1011. [\[CrossRef\]](#)
9. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [\[CrossRef\]](#)
10. Mas, J.F.; Flores, J.J. The application of artificial neural networks to the analysis of remotely sensed data. *Int. J. Remote Sens.* **2008**, *29*, 617–663. [\[CrossRef\]](#)
11. Moen, E.; Bannan, D.; Kudo, T.; Graf, W.; Covert, M.; Van Valen, D. Deep learning for cellular image analysis. *Nat. Methods* **2019**, *16*, 1233–1246. [\[CrossRef\]](#) [\[PubMed\]](#)
12. Hameed, K.; Chai, D.; Rassau, A. Score-based mask edge improvement of Mask-RCNN for segmentation of fruit and vegetables. *Expert Syst. Appl.* **2021**, *190*, 116205. [\[CrossRef\]](#)
13. Wei, X.S.; Cui, Q.; Yang, L.; Wang, P.; Liu, L. RPC: A large-scale retail product checkout dataset. *arXiv* **2019**, arXiv:1901.07249.
14. Hamian, M.H.; Beikmohammadi, A.; Ahmadi, A.; Nasersharif, B. Semantic Segmentation of Autonomous Driving Images by the combination of Deep Learning and Classical Segmentation. In Proceedings of the 2021 26th International Computer Conference, Computer Society of Iran (CSICC), Tehran, Iran, 3–4 March 2021; pp. 1–6.

15. Müller, D.; Ehlen, A.; Valeske, B. Convolutional neural networks for semantic segmentation as a tool for multiclass face analysis in thermal infrared. *J. Nondestruct. Eval.* **2021**, *40*, 1–10. [\[CrossRef\]](#)
16. 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. [\[CrossRef\]](#)
17. Neubert, M.; Herold, H.; Meinel, G. Evaluation of remote sensing image segmentation quality—further results and concepts. In Proceedings of the International Conference on Object-Based Image Analysis (ICOIA), Salzburg, Austria, 4–5 July 2006; pp. 1–6.
18. Akcay, O.; Avsar, E.; Inalpulat, M.; Genc, L.; Cam, A. Assessment of Segmentation Parameters for Object-Based Land Cover Classification Using Color-Infrared Imagery. *ISPRS Int. J. Geo-Inf.* **2018**, *7*, 424. [\[CrossRef\]](#)
19. Schwartzman, A.; Kagan, M.; Mackey, L.; Nachman, B.; De Oliveira, L. *Image Processing, Computer Vision, and Deep Learning: New Approaches to the Analysis and Physics Interpretation of LHC Events*; IOP Publishing: Bristol, UK, 2016; Volume 762, p. 012035.
20. Sherrah, J. Fully Convolutional Networks for Dense Semantic Labelling of High-Resolution Aerial Imagery. *arXiv* **2016**, arXiv:1606.02585.
21. Wang, J.; Shen, L.; Qiao, W.; Dai, Y.; Li, Z. Deep feature fusion with integration of residual connection and attention model for classification of VHR remote sensing images. *Remote Sens.* **2019**, *11*, 1617. [\[CrossRef\]](#)
22. Sun, Y.; Tian, Y.; Xu, Y. Problems of encoder-decoder frameworks for high-resolution remote sensing image segmentation: Structural stereotype and insufficient learning. *Neurocomputing* **2019**, *330*, 297–304. [\[CrossRef\]](#)
23. Marcu, A.; Leordeanu, M. Dual Local-Global Contextual Pathways for Recognition in Aerial Imagery. *arXiv* **2016**, arXiv:1605.05462.
24. Piramanayagam, S.; Saber, E.; Schwartzkopf, W.; Koehler, F.W. Supervised classification of multisensor remotely sensed images using a deep learning framework. *Remote Sens.* **2018**, *10*, 1429. [\[CrossRef\]](#)
25. 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.* **2018**, *135*, 158–172. [\[CrossRef\]](#)
26. Xie, S.; Tu, Z. Holistically-nested edge detection. In Proceedings of the IEEE International Conference on Computer Vision, Washington, DC, USA, 7–13 December 2015; pp. 1395–1403.
27. Du, S.; Du, S.; Liu, B.; Zhang, X. Incorporating DeepLabv3+ and object-based image analysis for semantic segmentation of very high resolution remote sensing images. *Int. J. Digit. Earth* **2021**, *14*, 357–378. [\[CrossRef\]](#)
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. Audebert, N.; Le Saux, B.; Lefèvre, S. Beyond RGB: Very high resolution urban remote sensing with multimodal deep networks. *ISPRS J. Photogramm. Remote Sens.* **2018**, *140*, 20–32. [\[CrossRef\]](#)
30. Song, A.; Kim, Y. Semantic Segmentation of Remote-Sensing Imagery Using Heterogeneous Big Data: International Society for Photogrammetry and Remote Sensing Potsdam and Cityscape Datasets. *ISPRS Int. J. Geo-Inf.* **2020**, *9*, 601. [\[CrossRef\]](#)
31. Yuan, X.; Shi, J.; Gu, L. A review of deep learning methods for semantic segmentation of remote sensing imagery. *Expert Syst. Appl.* **2020**, *169*, 114417. [\[CrossRef\]](#)
32. Nikparvar, B.; Thill, J.C. Machine Learning of Spatial Data. *ISPRS Int. J. Geo-Inf.* **2021**, *10*, 600. [\[CrossRef\]](#)
33. Wu, H.; Zhang, J.; Huang, K.; Liang, K.; Yu, Y. Fastfcn: Rethinking dilated convolution in the backbone for semantic segmentation. *arXiv* **2019**, arXiv:1903.11816.
34. Takikawa, T.; Acuna, D.; Jampani, V.; Fidler, S. Gated-scnn: Gated shape cnns for semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27–28 October 2019; pp. 5229–5238.
35. 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. [\[CrossRef\]](#) [\[PubMed\]](#)
36. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2961–2969.
37. Minaee, S.; Boykov, Y.Y.; Porikli, F.; Plaza, A.J.; Kehtarnavaz, N.; Terzopoulos, D. Image segmentation using deep learning: A survey. *arXiv* **2020**, arXiv:2001.05566.
38. Chen, L.C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking atrous convolution for semantic image segmentation. *arXiv* **2017**, arXiv:1706.05587.
39. ISPRS. International Society for Photogrammetry and Remote Sensing. 2D Semantic Labeling Challenge. 2016. Available online: <http://www2.isprs.org/commissions/comm3/wg4/semantic-labeling.html> (accessed on 5 October 2021).
40. Shorten, C.; Khoshgoftaar, T.M. A survey on image data augmentation for deep learning. *J. Big Data* **2019**, *6*, 1–48. [\[CrossRef\]](#)
41. Cramer, M. The DGPF-test on digital airborne camera evaluation overview and test design. *PFG Photogramm. Fernerkund. Geoinf.* **2010**, *2010*, 73–82. [\[CrossRef\]](#) [\[PubMed\]](#)
42. Akcay, O.; Kinaci, A.C.; Avsar, E.O.; Aydar, U. Boundary Extraction Based on Dual Stream Deep Learning Model in High Resolution Remote Sensing Images. *J. Adv. Res. Nat. Appl. Sci.* **2021**, *7*, 358–368. [\[CrossRef\]](#)
43. Gerke, M. *Use of the Stair Vision Library within the ISPRS 2D Semantic Labeling Benchmark (Vaihingen)*; Technical Report; University of Twente: Enschede, The Netherlands, 2015; doi:10.13140/2.1.5015.9683. [\[CrossRef\]](#)
44. McNeely-White, D.; Beveridge, J.R.; Draper, B.A. Inception and ResNet features are (almost) equivalent. *Cogn. Syst. Res.* **2020**, *59*, 312–318. [\[CrossRef\]](#)
45. Hao, S.; Zhou, Y.; Guo, Y. A brief survey on semantic segmentation with deep learning. *Neurocomputing* **2020**, *406*, 302–321. [\[CrossRef\]](#)

46. Azimi, S.M.; Henry, C.; Sommer, L.; Schumann, A.; Vig, E. Skyscapes fine-grained semantic understanding of aerial scenes. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27–28 October 2019; pp. 7393–7403.
47. Boguszewski, A.; Batorski, D.; Ziemba-Jankowska, N.; Zambrzycka, A.; Dziedzic, T. Landcover. ai: Dataset for automatic mapping of buildings, woodlands and water from aerial imagery. *arXiv* **2020**, arXiv:2005.02264
48. Abraham, N.; Khan, N.M. A novel focal tversky loss function with improved attention u-net for lesion segmentation. In Proceedings of the 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), Venice, Italy, 8–11 April 2019; pp. 683–687.
49. Gragera, A.; Suppakitpaisarn, V. Semimetric properties of sørensen-dice and tversky indexes. In *International Workshop on Algorithms and Computation*; Springer: Cham, Switzerland, 2016; pp. 339–350.
50. Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; Schiele, B. The cityscapes dataset for semantic urban scene understanding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 3213–3223.