applied sciences

MDPI

Article

Semantic Segmentation of Cabbage in the South Korea Highlands with Images by Unmanned Aerial Vehicles

Yongwon Jo [1], Soobin Lee [1], Youngjae Lee [1], Hyungu Kahng [1], Seonghun Park [1], Seounghun Bae [2], Minkwan Kim [2], Sungwon Han [1] and Seoungbum Kim [1,*]

[1] School of Industrial and Management Engineering, Korea University, Seoul 02841, Korea; gyj4318@korea.ac.kr (Y.J.); log0629@korea.ac.kr (S.L.); jae601@korea.ac.kr (Y.L.); hgkahng@korea.ac.kr (H.K.); seonghun@vt.edu (S.P.); swhan@korea.ac.kr (S.H.)

[2] Korea Land and Geospatial Informatix Corporation Spatial Information Research Institute, Wanju-Gun 55365, Korea; shbae29@gmail.com (S.B.); toughmk82@gmail.com (M.K.)

* Correspondence: sbkim1@korea.ac.kr

Abstract: Identifying agricultural fields that grow cabbage in the highlands of South Korea is critical for accurate crop yield estimation. Only grown for a limited time during the summer, highland cabbage accounts for a significant proportion of South Korea's annual cabbage production. Thus, it has a profound effect on the formation of cabbage prices. Traditionally, labor-extensive and time-consuming field surveys are manually carried out to derive agricultural field maps of the highlands. Recently, high-resolution overhead images of the highlands have become readily available with the rapid development of unmanned aerial vehicles (UAV) and remote sensing technology. In addition, deep learning-based semantic segmentation models have quickly advanced by recent improvements in algorithms and computational resources. In this study, we propose a semantic segmentation framework based on state-of-the-art deep learning techniques to automate the process of identifying cabbage cultivation fields. We operated UAVs and collected 2010 multispectral images under different spatiotemporal conditions to measure how well semantic segmentation models generalize. Next, we manually labeled these images at a pixel-level to obtain ground truth labels for training. Our results demonstrate that our framework performs well in detecting cabbage fields not only in areas included in the training data but also in unseen areas not included in the training data. Moreover, we analyzed the effects of infrared wavelengths on the performance of identifying cabbage fields. Based on the results of our framework, we expect agricultural officials to reduce time and manpower when identifying information about highlands cabbage fields by replacing field surveys.

Keywords: land-cover classification; semantic segmentation; unmanned aerial vehicles

Citation: Jo, Y.; Lee, S.; Lee, Y.; Kahng, H.; Park, S.; Bae, S.; Kim, M.; Han, S.; Kim, S. Semantic Segmentation of Cabbage in the South Korea Highlands with Images by Unmanned Aerial Vehicles. *Appl. Sci.* **2021**, *11*, 4493. https://doi.org/10.3390/app11104493

Academic Editor: Tobias Meisen

Received: 19 April 2021
Accepted: 11 May 2021
Published: 14 May 2021

1. Introduction

Monitoring distribution and changes in a region of interest (RoI) is a fundamental task in land-cover classification and has been part of many applications such as urban management [1], land-used management [2], and crop classification [3]. Land-cover classification generates information on the status of land use. It has been used especially in the South Korea highlands for a study of classifying cabbage as well as potatoes [4]. The reason for identifying cabbage in the highlands is that these regions, despite their short growing seasons, are the major cultivation areas for these cabbages. Therefore, it is important to develop land-cover classification methods to investigate the agricultural highlands of South Korea.

Traditionally, manual field surveys have been used to identify agricultural lands and derive land maps [4]. However, this method requires a significant amount of time and manpower. To replace manual field surveys, many studies have been conducted with remote sensing (RS) imagery using satellites and aircrafts [5]. However, this approach is disadvantageous in that they are mostly low-resolution images and can be degraded by

weather conditions or shadows that complicate the collection of accurate information [6,7]. Especially, these issues have been a major problem for the highlands of South Korea [4]. Recently, the development of RS technology has made capable high-resolution photography with unmanned aerial vehicles (UAV) [8] has created a trend away from satellite and aircraft imagery. UAV imagery is less likely to be affected by weather because it is taken at lower altitudes with higher spatial and spectral resolution capabilities that improve accuracy in rendering ROI [8]. In this study, we operated UAVs in the highlands and collected high-resolution photographs taken under different spatiotemporal conditions to generate information on cabbage cultivation fields.

In general, a multispectral sensor has been equipped with satellites, aircrafts, and UAVs to generate information on RoI [9]. Multispectral sensors capture not only visible wavelength reflectances but also infrared wavelengths, which enable the monitoring of vegetation growth [10,11]. Based on theoretical backgrounds for agriculture, we equipped UAVs with a multispectral sensor and collected various information on cabbage cultivation fields in South Korea highlands. In addition, we analyze the effect on infrared wavelengths in terms of performance on detecting cabbage fields at a pixel-level.

Recently, convolutional neural networks (CNN) have led to advances for computer vision tasks such as image classification [12], object detection [13], and semantic segmentation [14]. In particular, the semantic segmentation task aims to assign a class label to each pixel in an image [14]. Accordingly, CNN-based semantic segmentation algorithms have also been successfully applied to analyzing UAV imagery. For instance, semantic segmentation algorithms were applied to UAV imageries to ascertain the ratio of roads in cities and to assess pavement crack segmentation on an airport runway [15,16].

In this study, we propose a semantic segmentation framework based on UAV imagery to automate the process of identifying the cabbage fields in the South Korea highlands. First of all, UAV multispectral images of cabbage fields collected under various conditions were annotated with the help of agricultural experts to generate information on the important cabbage. Second, we trained several different semantic segmentation models and compared their performances. Moreover, we made extensive studies to analyze the generalization performance of our models. Finally, we analyzed the effects of infrared wavelengths on identifying cabbage fields at a pixel-level. The main contributions of this study can be summarized as follows:

- To measure how well our framework generalizes well despite differences for highlands and shooting dates, we operated UAVs equipped with a multispectral sensor and collected multispectral images under different spatiotemporal conditions.
- Our proposed framework shows exceptional detection performance for test images collected from the areas included in the training data but on different dates. Moreover, our method generalizes well to unseen areas not used during training.
- To analyze which wavelength in multispectral images has a positive effect on detection performance, we experimented with four different combinations of input wavelengths and compared their detection performances. Based on the results, we demonstrate that the semantic segmentation model trained with blue, green, red, and red edge wavelengths is the most suitable for automating the process of identifying cabbage cultivation fields.

The remainder of this paper is organized as follows. In Section 2, we review the papers associated with semantic segmentation models based on CNN, land-cover classification, and applications of CNN to UAV imagery. In Section 3, we describe the proposed framework used to detect cabbage cultivation. In Section 4, we give a thorough analysis of the experimental results of different semantic segmentation models. In Section 5, we discuss our results in terms of semantic segmentation models and input wavelengths. In Section 6, we summarize our study with conclusions and future research directions.

2. Related Work

2.1. Semantic Segmentation Models

Semantic segmentation is the task of assigning each pixel in an image to a class label [14]. Various studies based on CNNs have been proposed to tackle the semantic segmentation task. First, Fully convolutional networks (FCNs) [17] have demonstrated performance improvement on PASCAL visual object classes dataset [18], which is the popular image segmentation benchmark.The models after FCNs have been based on encoder-decoder architectures. More recently, CNN encoder-decoder architectures [19,20] have been studied in depth. The encoder module maps raw image pixels to a rich representation which is a collection of feature maps with smaller spatial dimensions. The decoder module makes pixel-wise predictions by taking the representation emitted by the encoder and mapping it back to the original input image size. Meanwhile, spatial pyramid pooling and *atrous* convolution have been developed to extract multi-scale information from input images [21,22]. In PSPNet [21], original feature maps are calculated by max-pooling operations with different scales and outputs, formed by different regions of original maps. Based on spatial pyramid pooling (SPP), each output is upsampled to match the size of original feature maps. *Atrous* convolution, or dilated convolution [22], is another convolution operation with different factors to expand receptive fields of CNN representation vectors without losing resolution. Extracting multi-scale information, including SPP and *atrous* convolution, is helpful in that the resulting representations comprise of feature maps of different scales which prevents contextual information from vanishing. In this study, we experiment with models based on an encoder-decoder architecture which is capable of extracting multi-scale information.

2.2. Land Cover Classification

The task of land cover classification is an important task for the RS community, which aims to observe the Earth and monitoring its changes [5]. Early works mostly relied on low-resolution RS images to generate information about the Earth [23]. To extract information, each pixel in RS images was classified by a patch around each pixel [24]. However, this method is computationally inefficient. Another disadvantage was that boundary information in RS images was fuzzy, leading to incorrect information [25]. To monitor a target study area, it is difficult to extract accurate information from low-resolution RS images. Recently, the development of RS technology has made capable high-resolution images for land cover classification, such as disease detection [26], yield estimation [27], and weed detection [28]. With high-resolution RS images, semantic segmentation methods were applied to classify whether each pixel in vineyards is healthy or diseased [25]. Likewise, various information can be extracted by the semantic segmentation methods at a pixel-level depending on what researchers aim to classify. In this study, we take a first approach to classifying cabbage fields in the highlands.

2.3. Applications of Semantic Segmentation for Agriculture

Overall, several studies have successfully applied semantic segmentation models to the agricultural data domain. One study applied a feature pyramid network to classify seven land types such as urban land, agricultural land, and water [29]. Other works include but are not limited to applying semantic segmentation models to detect weed plants or identifying cranberry fields [30,31]. In our study, we propose a framework based on state-of-the-art semantic segmentation models including U-Net [19], SegNet [20], and DeepLab V3+ [32], to detect the cabbage fields at a pixel-level in the South Korea highlands. We conducted various experiments to compare model performances for identifying cabbage pixels under different spatiotemporal conditions unseen from the training dataset.

3. Materials and Methods

3.1. Target Crop and Regions

The purpose of this study is to automate the process of identifying cabbage fields at a pixel level in the South Korea highlands. As shown in Figure 1, vegetables on the left and right are different. The left vegetable is called napa cabbage or kimchi cabbage, which is mainly used to make a traditional Korean food called kimchi. The right vegetable is a common cabbage in the West. Our work focuses on detecting cabbages like those in the left figure. We propose a semantic segmentation framework to classify cabbage-growing regions that are invariant to spatiotemporal changes.

Figure 1. (**Left**) Napa cabbage, which we classify in fields; (**Right**) common cabbage in the West.

We selected three major highland cabbage cultivation areas in *Gangwon Province*, South Korea (Figure 2); Maebongsan (37°13′05″ N, 128°57′58″ E), Gwinemi (37°20′21″ N, 129°00′20″ E), and Anbanduck (37°37′31″ N, 128°44′21″ E). In South Korea, high summer temperature, humidity, and the proliferation of insects from June to August restrict the successful cultivation of cabbages except for these three highlands [4]. UAV images taken in Maebongsan on 24 July 2019, and 5 August 2019; Gwinemi on 5 August 2019; and in Anbanduck on 6 August 2019, were used. We attached a multispectral sensor, RedEdge-MX (MicaSense, Inc, Seattle, WA, USA), to a rotary-wing UAV (DJI M2010) to collect images from the study areas. The sensor collects image data within specific wavelength ranges consisting of blue (B, 475 ± 20 nm), green (G, 560 ± 20 nm), red (R, 668 ± 10 nm), red edge (RE, 717 ± 10 nm), and near-infrared (NIR, 840 ± 40 nm) wavelengths. Each image size with ground sampling distance (GSD) of approximately 12 cm is $1280(width) \times 960(height)$.

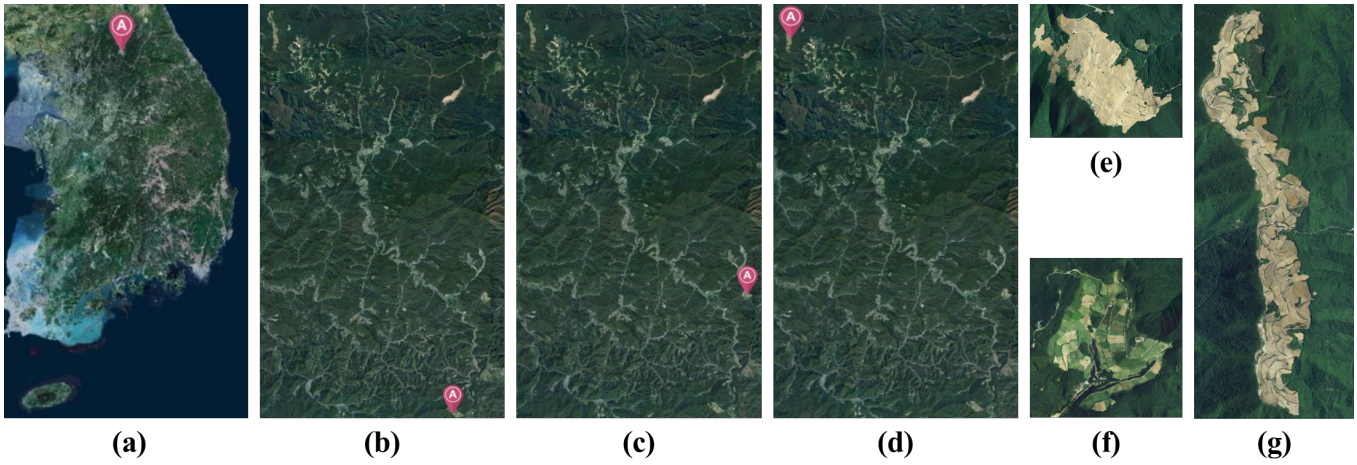

Figure 2. Location of target regions [33]. (**a**) Gangwon Province in South Korea (highlands); (**b**) Maebongsan (bottom red point); (**c**) Gwinemi (middle red point); (**d**) Anbanduck (top red point); (**e**) Maebongsan; (**f**) Gwinemi; and (**g**) Anbanduck.

3.2. Image Preprocessing

Misalignment between the five lenses of RedEdge-MX requires an additional data processing step prior to model training. A visual example of bandwidth misalignment is shown in Figure 3. Simply stacking the five bandwidths along the channel dimension results in counterfactual images such as Figure 3f. To cope with the misalignment and create RGB images visible to the human eye, we applied enhanced correlation coefficient (ECC) transformation [34,35]. Next, with the help of agricultural experts, we visually examined the RGB images and generated pixel-wise ground truth labels of the cabbage fields, as shown in Figure 3h. We used four different combinations of input wavelengths to train the semantic segmentation models: RGB, RGB with RE, RGB with NIR, and RGB with both RE and NIR. We give comparisons with respect to the input types in Sections 4.3 and 5.2.

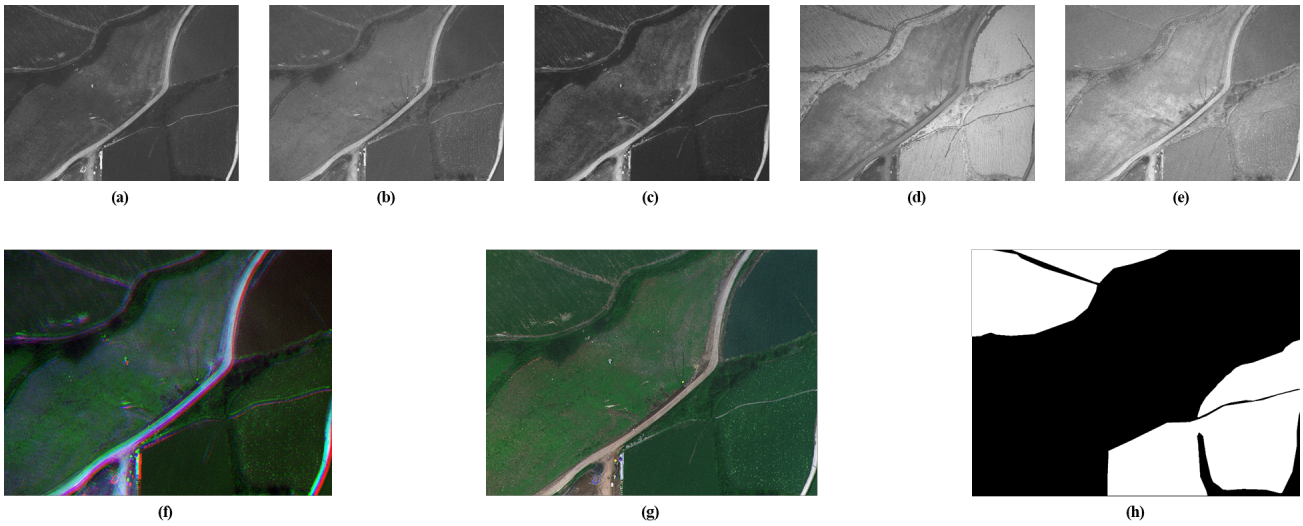

Figure 3. Registration from multi spectral images to an RGB image. Using ECC transformation, (**a**) blue, (**b**) green, (**c**) red, (**d**) red edge, and (**e**) near-infrared wavelength values in gray scale are aligned to (**g**) RGB image. (**f**,**h**) is an example of incorrect registration and pixel-wise ground truth label.

3.3. Semantic Segmentation Models

We use three different semantic segmentation models that have demonstrated acceptable performance across various image domains: U-Net [19], SegNet [20], and DeepLab V3+ [32]. U-Net was originally developed to distinguish neuronal structures in electron microscopic stacks. The model architecture consists of two modules; one is a contracting path (encoder), and the other an expanding path (decoder). In the contracting path, context information is extracted by a sequence of neural network blocks. Each block in the path consists of two convolution layers, followed by a batch normalization layer, a rectified linear unit (ReLU), and a 2×2 max-pooling layer. Five blocks constitute a contracting path. In the expanding path, rich context information from a contracting path is restored to the original input size. The expanding path consists of four blocks in which each block consists of a transposed convolution layer, followed by a convolution layer, a batch normalization layer, and a ReLU. In every transposed convolution operation, context information from each block in the contracting path is cropped and copied to alleviate the loss of border information of an input image when calculating convolution and max-pooling layers. Combining cropped feature maps in the contracting path with feature maps in the expanding path enables aggregation of multi-scale feature maps. Another method to prevent the loss of border information is a mirroring strategy, which extrapolates the border regions of an input image as shown in Figure 4. This technique allows the border regions of the input to be accurately predicted.

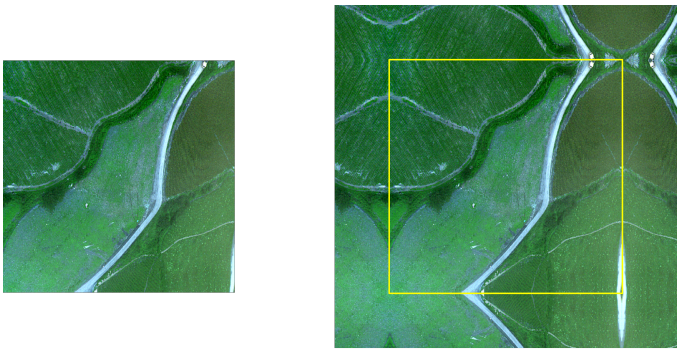

Figure 4. Mirroring extrapolation strategy to improve prediction result about border regions of an input image.

SegNet has an encoder-decoder architecture and was developed to use pixels to recognize objects on roads for use in an autonomous driving system [20]. The encoder of SegNet is identical to the feature extractor of VGG-16 [36]. Moreover, the model records the max-pooling indices of the encoder and uses them during upsampling layers in the decoder. Using max-pooling indices shortens the inference time of SegNet. We expected the use of max-pooling indices in SegNet to make the prediction boundaries of cabbage fields sharper.

Finally, DeepLab V3+ combines advantages from encoder-decoder structures and a multi-scale feature extraction module [32]. The encoder-decoder structure generates an output image with pixel-wise prediction results equal to the original input size. This structure enables accurate detection of object boundary surfaces by combining information from the encoder during decoder upsampling [32]. The multi-scale feature extraction module, called an *atrous* spatial pyramid pooling (ASPP), consists of several *atrous* convolutions with different dilation rates. The ASPP module enlarges receptive fields of representation vectors and enables DeepLab V3+ to recognize objects of different sizes [32,37]. We assume that DeepLab V3+ is different from U-Net and SegNet in terms of multi-scale feature extraction method with *atrous* convolution and has an invariant performance under different conditions from the training dataset. Figure 5 depicts the DeepLab V3+ structure used in our study.

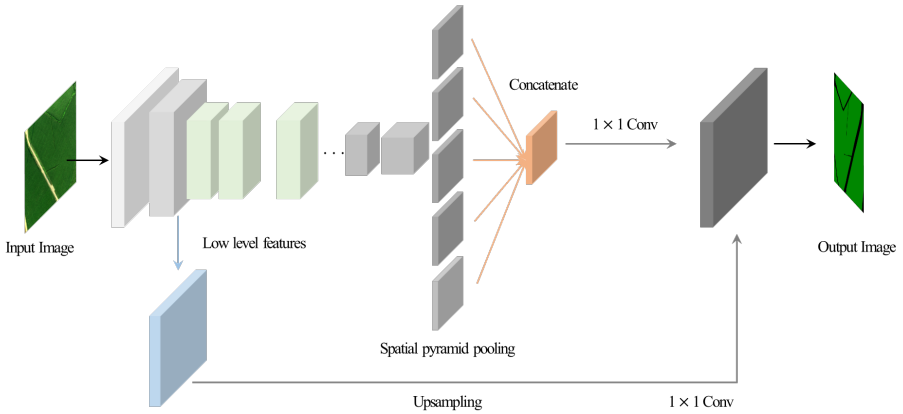

Figure 5. The DeepLab V3+ structure for detecting cabbage fields. Although there are differences in the sizes of cabbage field, an atrous convolution in the encoder can accommodate the differences and extracts multi-scale context information.

DeepLab V3+ comprises three modules: an encoder, an ASPP module, and a decoder. The encoder is an Xception model [38], which extracts representation vectors from RGB and multispectral images. Next, the ASPP module performs *atrous* depth-wise separable

convolution operated for each channel, followed by a 1×1 convolution [39]. The advantage of the ASPP module is that it reduces computational complexity while maintaining predictive performance. Finally, the decoder stacks the feature maps obtained from the second block of the encoder and also the feature maps obtained from the ASPP module. The model performs 1×1 convolution on the stacked feature maps before upsampling to the input resolution space for the final pixel-wise prediction.

Our loss function for learning models, $\mathcal{L}_{cls}$, is a categorical cross entropy function and formulated as follows:

$$\mathcal{L}_{cls} = -\sum_{c=0}^{1} \alpha_c * \sum_{i=0}^{n} y_{i,c} \log \hat{y_{i,c}} \tag{1}$$

where c is an index of ground truth class, α is weighting factor calculated for ground truths in our training dataset, n is the number of pixel in images of an mini batch, $y_{i,c}$ is a target class value, and $\hat{y_{i,c}}$ is a predicted score of ith pixel which belongs to class c. We give comparisons with respect to the semantic segmentation models in Sections 4.3 and 5.1.

4. Experiments

4.1. Dataset

We took 1240 UAV images of Maebongsan recorded on 24 July 2019, and used 1032 for training and 208 for model validation. In order to improve the diversity of the training and validation dataset, five data augmentation methods, such as horizontal flips, vertical flips, and rotation in 90, 180, and 270 degrees, were applied to both the training and validation sets to obtain a final training set of 6192 images and a validation set of 1248. We aim to train robust semantic segmentation models for cabbage fields of different scales through data augmentations. Next, we composed three different testing sets to measure the generalization performance of the proposed method. The first testing set, denoted as the MBS dataset, consisted of 471 images taken from the Maebongsan area on 5 August 2019. Evaluation on the MBS dataset is necessary to measure how well our models generalize to data collected under different temporal conditions. The second testing set, denoted as the GNM dataset, consisted of 156 images taken from the Gwinemi area on 5 August 2019. Lastly, 143 images of the Anbanduck area on 6 August 2019, comprised the third testing set, denoted as the ABD dataset. Evaluations on the GNM and ABD datasets are necessary to measure how well our models perform on data collected under different spatiotemporal conditions.

4.2. Hyperparameters

We used three different semantic segmentation models in our experiments: U-Net, SegNet, and DeepLab V3+. These models were initialized by Xavier initialization [40]. We trained U-Net with the AdamW optimizer [41], using a learning rate of 0.01 for 100 epochs with a batch size of 12. The input heights and widths were fixed to 572. SegNet was trained with stochastic gradient descent for 100 epochs with a batch size of 16, using a learning rate of 0.01, and a momentum factor of 0.9. The input heights and widths were set to 224. Finally, we trained DeepLab V3+ with the AdamW optimizer , using a learning rate of 0.001 and a weight decay factor of 0.01 for 70 epochs with a batch size of 8. The input heights and widths were fixed to 513. All experiments were implemented with PyTorch 1.4.0 and conducted on a single NVIDIA TITAN RTX GPU.

4.3. Model Performance

We used the mean intersection over union (MIoU) metric [17] to quantify the performance of the semantic segmentation models used in this study. The formula of MIoU for a dataset of N images is defined as follows:

$$MIoU = \frac{1}{N} \sum_{i=1}^{N} \frac{n_{11}^{(i)}}{n_{1.}^{(i)} + n_{.1}^{(i)} - n_{11}^{(i)}} \tag{2}$$

where $n_{11}^{(i)}$ is the number of correctly classified pixels in ith image, $n_{1.}^{(i)}$ is the number of ground truth pixels labeled as cabbages, and $n_{.1}^{(i)}$ is the number of pixels predicted as cabbage fields. Best hyperparamters were chosen based on the validation MIoU metric. We checked the MIoU score after every training epoch, and saved the best model checkpoint for further inference on the MBS, GNM, and ABD datasets. We compared the train and validation learning curves of our experiments in Figure 6. It could be observed that the learning curves of DeepLab V3+ shown in Figure 6a,d,g,j exhibited the fastest rate of convergence compared to other models. Also, the fluctuation ranges of the DeepLab V3+ validation MIoU were smaller than other models. In addition, as could be seen in Figure 6f,i,l, SegNet had wide fluctuation ranges and were unstable after epoch 80. As shown from the first row to the fourth row in Figure 6, the more the number of input wavelengths were, the wider fluctuation ranges occurred. Reasons are to be given in Section 5.2. As shown in Figure 6j–l, it can be evaluated that models with NIR wavelength suffer from overfitting. To demonstrate the applicability of our semantic segmentation framework, we should measure how the framework generalizes well for images in three test datasets under different spatiotemporal conditions.

Each combination of input data and models was repeated ten times with different random seeds, and we report the average and standard deviations of the MIoU metric. As shown in Table 1, DeepLab V3+ with RGB outperformed both U-Net and SegNet on the MBS dataset. Moreover, we observed better performances in all three models when using RGB images rather than the other input wavelengths. Although the best validation MIoU was the performance of U-Net with RGB and RE, we demonstrated that DeepLab V3+ has more generalizability than U-Net and SegNet based on the MBS dataset MIoU.

Table 1. Comparisons of model's performance on the *validation* and the *MBS* dataset. The best performance of MIoU is **boldfaced**. Standard deviations are given in parentheses.

Model	Input Wavelengths	The Number of Parameters	Validation MIoU	MBS Dataset MIoU
DeepLab V3+	**RGB**	**54,700,434**	**0.9021 (0.0028)**	**0.8997 (0.0183)**
U-Net	RGB	40,446,786	0.8979 (0.0023)	0.8455 (0.0187)
SegNet	RGB	29,444,166	0.8506 (0.0198)	0.4851 (0.1298)
DeepLab V3+	RGB + RE	54,700,722	0.8974 (0.0025)	0.8483 (0.0466)
U-Net	RGB + RE	40,447,362	0.9108 (0.0028)	0.8042 (0.0215)
SegNet	RGB + RE	29,444,742	0.8676 (0.0051)	0.6604 (0.0482)
DeepLab V3+	RGB + NIR	54,700,722	0.8768 (0.0097)	0.8741 (0.0396)
U-Net	RGB + NIR	40,447,362	0.9085 (0.0029)	0.7613 (0.0742)
SegNet	RGB + NIR	29,444,742	0.7833 (0.0155)	0.1856 (0.0537)
DeepLab V3+	RGB+RE+NIR	54,701,010	0.8923 (0.0045)	0.5700 (0.1291)
U-Net	RGB+RE+NIR	40,447,938	0.8878 (0.0120)	0.5312 (0.0864)
SegNet	RGB+RE+NIR	29,445,318	0.8136 (0.0301)	0.4829 (0.1887)

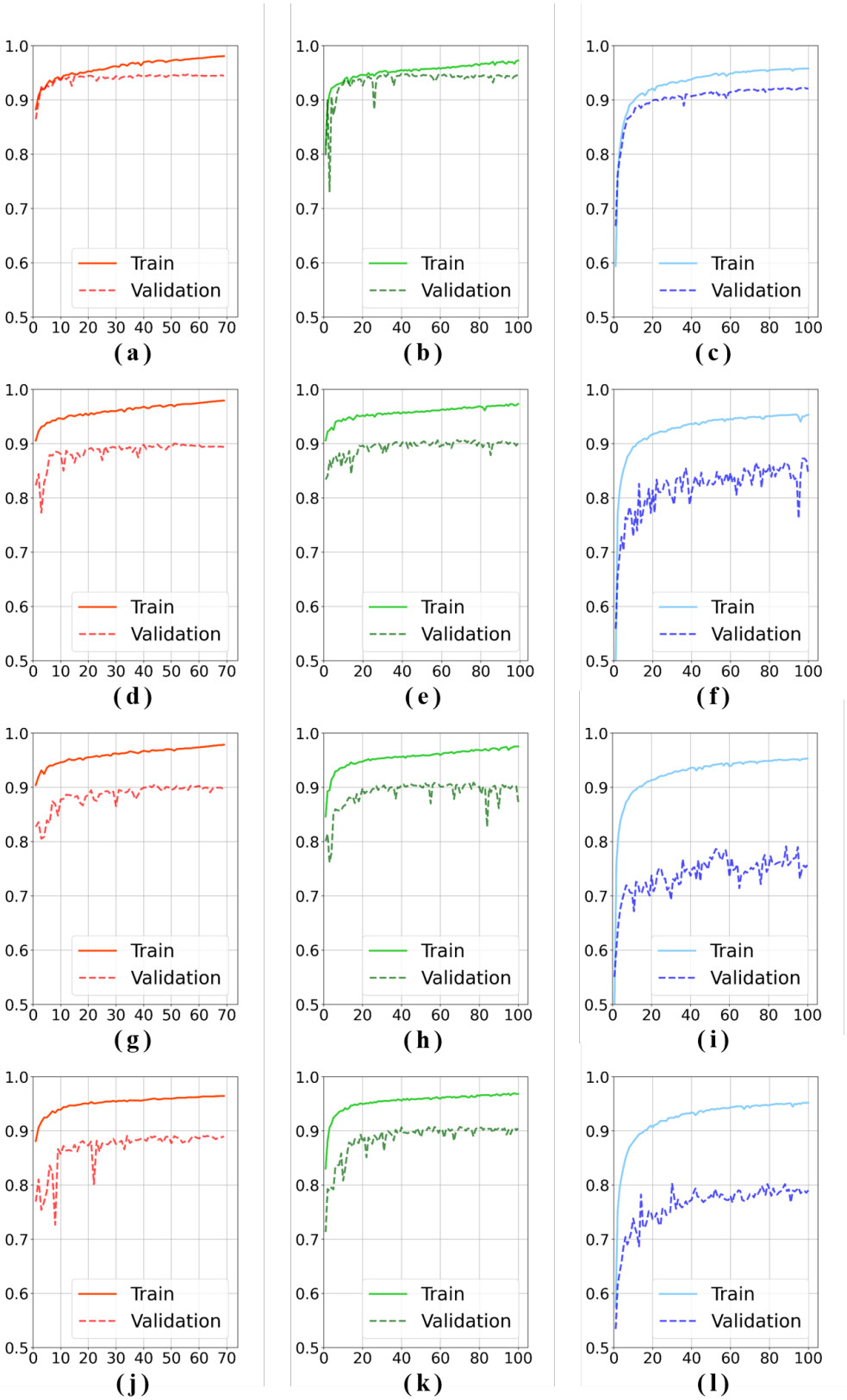

Figure 6. Train (solid) and validation (dashed) learning curves of our experiments. The *x*-axis and *y*-axis mean epoch and validation MIoU, respectively. Each column corresponds to DeepLab V3+, U-Net, and SegNet. Each row corresponds to RGB images, RGB with RE, RGB with NIR, and RGB with RE and NIR, respectively. Note that the *x*-axis limits of the first and other columns are set differently.

Figure 7 provides visualizations of model predictions, along with their ground truth labels for three representative examples from the MBS dataset. It is observable that the DeepLab V3+ predictions using RGB inputs most closely match the ground truth.

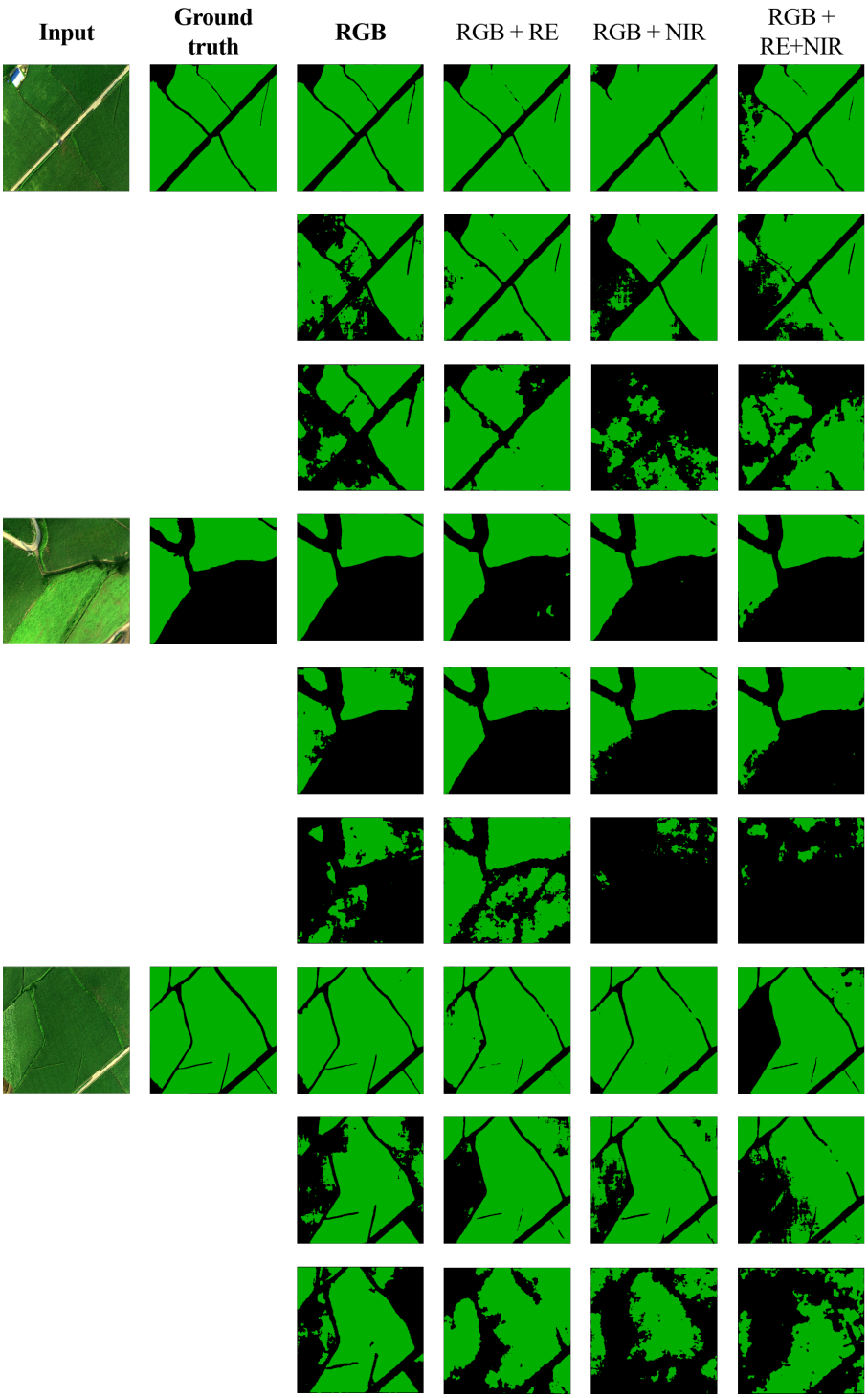

Figure 7. Visualization of model predictions for three representative examples from the MBS dataset. There were input images and ground truth in the first and second columns. Prediction results are represented from the third column to the sixth column. Results of DeepLab V3+, U-Net, and SegNet are repeatably visualized from the first row to the final row. The area occupied by each input in Maebongsan is about 1.77 ha. Each input image width and height of DeepLab V3+, U-Net, and SegNet is 513, 572, and 224, respectively.

To verify whether DeepLab V3+ generalizes across different times and regions, we evaluated its predictive performance on the GNM dataset collected from Gwinemi. As shown in Table 2, DeepLab V3+ with RGB and RE performed the best. We found that the RE wavelength has a positive effect on the detection performance of images collected from different spatiotemporal conditions. In addition, the standard deviation of DeepLab V3+ with RGB and RE was the smallest. Further, in Figure 8, we provided visualizations of predictions on the GNM dataset. The example in the first row included greenhouses. We found that DeepLab V3+ with RGB and RE wavelengths succeeds in distinguishing the greenhouses from cabbage fields. In addition, the second example of Figure 8 demonstrates that the DeepLab V3+ is capable of distinguishing cabbage fields from weeds (dark green). Last but not least, we could see from the third row's example that the land left fallow in the lower left part of the image is correctly classified as not growing cabbage.

Table 2. Comparisons of model's performance on the *GNM* (third column) and *ABD* (fourth column) dataset. The best performance of MIoU is **boldfaced**. Standard deviations are given in parentheses.

Input Wavelengths	Model	GNM Dataset MIoU	ABD Dataset MIoU
DeepLab V3+	RGB	0.9072 (0.0045)	0.7294 (0.0764)
U-Net	RGB	0.8999 (0.0076)	0.4734 (0.0689)
SegNet	RGB	0.8191 (0.0210)	0.4873 (0.2066)
textbfDeepLab V3+	**RGB + RE**	**0.9097 (0.0030)**	**0.8223 (0.0483)**
U-Net	RGB +RE	0.8983 (0.0030)	0.7459 (0.0605)
SegNet	RGB + RE	0.7921 (0.0288)	0.6435 (0.0611)
DeepLab V3+	RGB + NIR	0.8605 (0.0221)	0.7812 (0.0340)
U-Net	RGB +NIR	0.8912 (0.0174)	0.5933 (0.0976)
SegNet	RGB + NIR	0.7054 (0.0214)	0.1440 (0.0873)
DeepLab V3+	RGB + RE + NIR	0.8525 (0.0246)	0.3084 (0.1622)
U-Net	RGB + RE + NIR	0.8665 (0.0326)	0.3121 (0.1901)
SegNet	RGB + RE + NIR	0.7048 (0.0443)	0.5444 (0.2246)

The ABD dataset was composed of images photographed in Anbanduck. As can be seen from the data in Table 2, DeepLab V3+ with RGB and RE performed the best like the GNM dataset. Also, we observed that adding RE to the input improves the detection performance. Figure 9 shows that our model successfully distinguishes fields of weeds, forest regions, and wind turbines from cabbage fields.

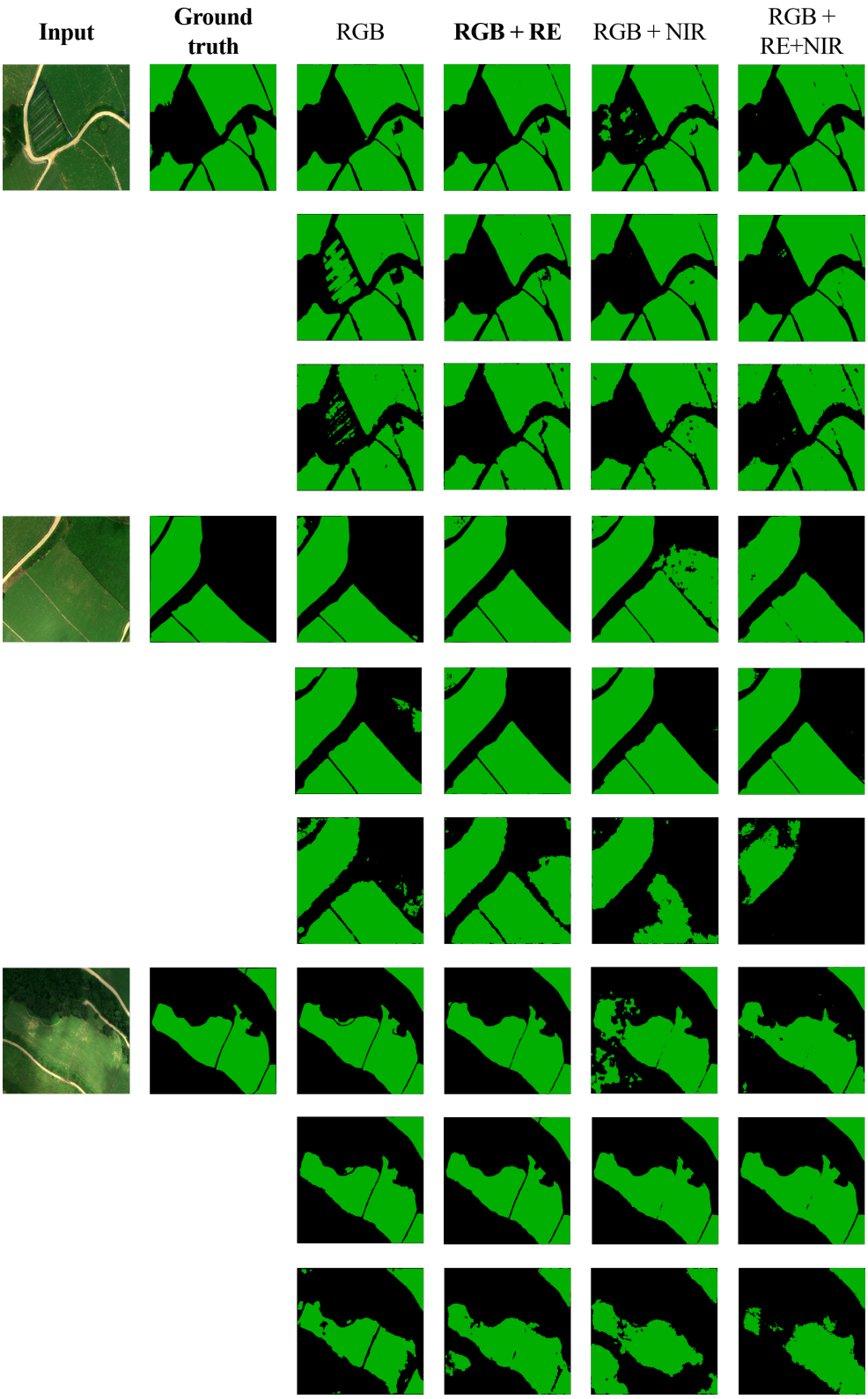

Figure 8. Visualization of model predictions for three representative examples from the GNM dataset. There were input images and ground truth in the first and second columns. Prediction results are represented from the third column to the sixth column. Results of DeepLab V3+, U-Net, and SegNet are repeatably visualized from the first row to the final row. The area occupied by each input in Gwinemi is about 1.77 ha. Each input image width and height of DeepLab V3+, U-Net, and SegNet is 513, 572, and 224, respectively.

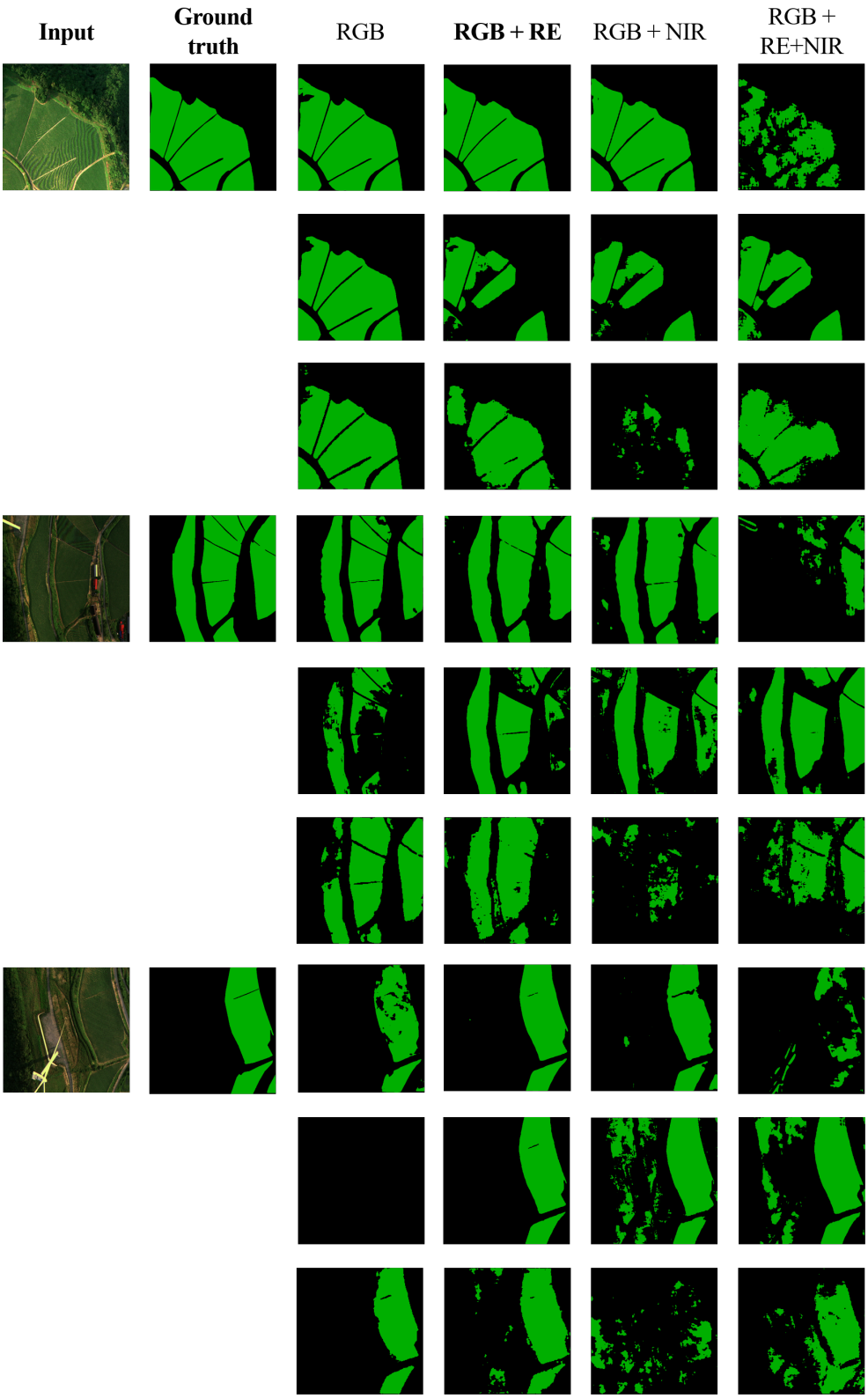

Figure 9. Visualization of model predictions for three representative examples from the ABD dataset. There were input images and ground truth in the first and second columns. Prediction results are represented from the third column to the sixth column. Results of DeepLab V3+, U-Net, and SegNet are repeatably visualized from the first row to the final row. The area occupied by each input in Anbanduck is about 1.77 ha. Each input image width and height of DeepLab V3+, U-Net, and SegNet is 513, 572, and 224, respectively.

5. Discussion

5.1. Comparisons of Models

We made several assumptions on semantic segmentation models as mentioned in Section 3.3. First, we assumed that DeepLab V3+ has the best invariant performance under different spatiotemporal conditions. This was demonstrated in Tables 1 and 2. The ASPP module in DeepLab V3+ enabled the extraction of multi-scale information of cabbage fields by applying *atrous* convolutions with different dilation rates. Receptive fields of representation vectors had a positive influence on the detection performance [42]. As can be seen from Figures 7–9, DeepLab V3+ accurately predicted cabbage fields of different sizes. Next, the assumption on U-Net's mirroring strategy did not hold. As shown in the U-Net visualizations of Figure 7, pixels in the border regions were still misclassified. Finally, SegNet's max-pooling indices received from each encoder did not sharpen prediction boundaries from Figures 7–9. In summary, DeepLab V3+ demonstrated superior performance compared to U-Net and SegNet. Especially, we claim that multi-scale feature map aggregation induced by the ASPP module was critical when identifying the complex environment of the South Korea highlands under different spatiotemporal conditions. This has also been emphasized in several previous studies on extracting multi-scale information in UAV images for agriculture [42,43].

5.2. Difference in Input Wavelengths

Based on detection performance comparisons in Section 4.3, a histogram-based analysis was included to identify why models trained with RGB and RE inputs showed better performance than other wavelength combinations under spatiotemporal conditions. To visualize distributions of each wavelength reflection in Figure 10, we randomly selected 100 multispectral images in the train, MBS, GNM, and ABD datasets, respectively. Two-sample t-tests were then performed between each wavelength distribution in the training dataset and three testing datasets. The null hypothesis states that means of each wavelength reflection for the training and testing dataset are equal. As shown in Table 3, p-values were zero, indicating that we could reject the null hypothesis. It demonstrated that a discrepancy exists between the wavelength distributions of the training and test sets. From these results, we claim that such differences caused the performance to degrade in models trained with RGB, RE, and NIR naively stacked along the channel dimension. [44]. Furthermore, we believe that these differences caused multispectral models to overfit to the training data as well. A recent study pointed out that naively stacking RGB with NIR/RE bandwidths can result in redundant information and mutual interference when training CNNs [45]. As shown in Table 2, we also observed a similar phenomenon. Because models with RGB and RE outperformed models with the other combinations of wavelengths, while models with RGB, RE, and NIR showed the worst detection performance. In addition, improvement for the performance of models with RGB and RE was larger than the improvement with RGB and NIR. Based on the results, we claim that RE wavelength could have a more positive effect on identifying cabbage fields under spatial and temporal differences than NIR wavelength. We believe that our proposed method can be improved by incorporating several encoders to learn separate low-level representations for each wavelength including RE and NIR.

Table 3. Two-sample t-test results between each wavelength reflection distribution in the training and testing sets, including MBS, GNM, and ABD.

Dataset	MBS					GNM					ABD				
	B	**G**	**R**	**RE**	**NIR**	**B**	**G**	**R**	**RE**	**NIR**	**B**	**G**	**R**	**RE**	**NIR**
T-statistic	989.70	602.73	858.70	3540.27	6767.05	−280.60	−298.04	419.88	155.37	2394.65	1309.56	2190.63	1251.48	4305.60	7144.72
p-value	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

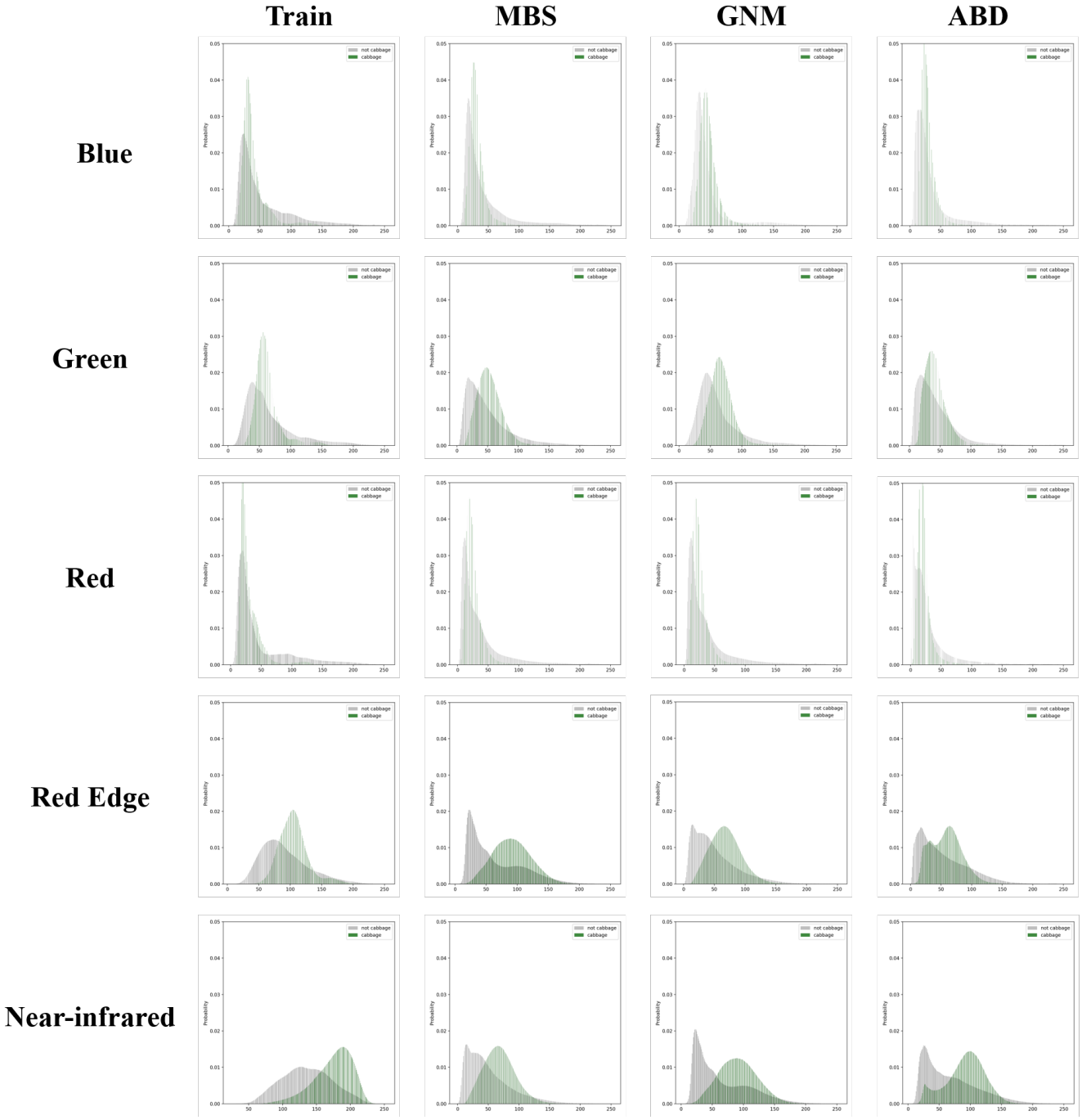

Figure 10. Reflection histograms of the each channel. Each row contains histograms for a single wavelength. Each column corresponds to a dataset. Green and gray distributions correspond to the distribution of cabbage-pixel reflection and other-pixel reflection, respectively.

6. Conclusions

In this study, we have proposed a semantic segmentation framework based on UAV images to automate the process of identifying cabbage fields in the highlands of South Korea. First, we collected high-resolution multispectral images by operating UAVs over different highlands in South Korea to generate information on the important cabbage. Second, we applied ECC transformation to handle misalignment between channels and generated pixel-wise ground truth labels. We compared the performances of detecting cabbage cultivation fields by three semantic segmentation models and four combinations of input wavelengths and concluded that DeepLab V3+, which was trained on RGB and

RE wavelengths, performed the best. We demonstrated that the model was effective in distinguishing between the cabbage fields, fields of weeds, and buildings despite changes in operational dates and regions. Based on the results of our proposed framework, we expect agricultural officials to save time and manpower when collecting information about cabbage cultivating fields in South Korea highlands by replacing manual field surveys. In future studies, we plan to apply semantic segmentation models to detect multiple crops such as cabbages, peppers, and beans. We also expect to make use of additional infrared wavelengths, such as RE and NIR, by incorporating several CNN-based encoders to learn low-level representations from each wavelength and enhance model performance.

Author Contributions: Conceptualization, S.B., S.H. and S.K.; methodology, Y.J., S.L., Y.L., H.K. and S.P.; investigation, Y.J., S.L., Y.L., H.K., S.P., M.K., S.H. and S.K.; data curation, S.B. and M.K.; writing—original draft preparation, Y.J., Y.L. and H.K.; writing—review and editing, S.H. and S.K.; supervision, S.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by a grant from the Korea Land and Geospatial Informatix Corporation Spatial Information Research Institute (LX SIRI), the Brain Korea 21 FOUR, Ministry of Science and ICT (MSIT) in Korea under the ITRC support program (IITP-2020-0-01749) supervised by the IITP, the National Research Foundation of Korea grant funded by the MSIT (NRF-2019R1A4A1024732), and the Ministry of Culture, Sports and Tourism and Korea Creative Content Agency (R2019020067).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are not publicly available due to privacy and legal restrictions.

Acknowledgments: The authors would like to thank the Editor and reviewers for their useful comments and suggestions, which were greatly helpful in improving the quality of the paper.

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

References

1. Gebrehiwot, A.; Hashemi-Beni, L.; Thompson, G.; Kordjamshidi, P.; Langan, T.E. Deep convolutional neural network for flood extent mapping using unmanned aerial vehicles data. *Sensors* **2019**, *19*, 1486. [CrossRef] [PubMed]
2. Scott, G.J.; England, M.R.; Starms, W.A.; Marcum, R.A.; Davis, C.H. Training deep convolutional neural networks for land—Cover classification of high-resolution imagery. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 549–553. [CrossRef]
3. Rustowicz, R.M.; Cheong, R.; Wang, L.; Ermon, S.; Burke, M.; Lobell, D. Semantic segmentation of crop type in africa: A novel dataset and analysis of deep learning methods. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Long Beach, CA, USA, 16–20 June 2019; pp. 75–82.
4. Kwak, G.H.; Park, N.W. Impact of texture information on crop classification with machine learning and UAV images. *Appl. Sci.* **2019**, *9*, 643. [CrossRef]
5. Vali, A.; Comai, S.; Matteucci, M. Deep learning for land use and land cover classification based on hyperspectral and multispectral earth observation data: A review. *Remote Sens.* **2020**, *12*, 2495. [CrossRef]
6. Kussul, N.; Lavreniuk, M.; Skakun, S.; Shelestov, A. Deep learning classification of land cover and crop types using remote sensing data. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 778–782. [CrossRef]
7. Maes, W.H.; Steppe, K. Perspectives for remote sensing with unmanned aerial vehicles in precision agriculture. *Trends Plant Sci.* **2019**, *24*, 152–164. [CrossRef]
8. Sarkar, T.K.; Ryu, C.S.; Kang, Y.S.; Kim, S.H.; Jeon, S.R.; Jang, S.H.; Park, J.W.; Kim, S.G.; Kim, H.J. Integrating UAV remote sensing with GIS for predicting rice grain protein. *J. Biosyst. Eng.* **2018**, *43*, 148–159.
9. Zhou, X.; Zheng, H.; Xu, X.; He, J.; Ge, X.; Yao, X.; Cheng, T.; Zhu, Y.; Cao, W.; Tian, Y. Predicting grain yield in rice using multi-temporal vegetation indices from UAV-based multispectral and digital imagery. *ISPRS J. Photogramm. Remote Sens.* **2017**, *130*, 246–255. [CrossRef]
10. Li, Z.; Hu, H.M.; Zhang, W.; Pu, S.; Li, B. Spectrum characteristics preserved visible and near-infrared image fusion algorithm. *IEEE Trans. Multimed.* **2020**, *23*, 306–319. [CrossRef]
11. Pôças, I.; Calera, A.; Campos, I.; Cunha, M. Remote sensing for estimating and mapping single and basal crop coefficientes: A review on spectral vegetation indices approaches. *Agric. Water Manag.* **2020**, *233*, 106081. [CrossRef]
12. Rawat, W.; Wang, Z. Deep convolutional neural networks for image classification: A comprehensive review. *Neural Comput.* **2017**, *29*, 2352–2449. [CrossRef]

13. Zhao, Z.Q.; Zheng, P.; Xu, S.T.; Wu, X. Object detection with deep learning: A review. *IEEE Trans. Neural Netw. Learn. Syst.* **2019**, *30*, 3212–3232. [CrossRef] [PubMed]
14. 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]
15. Li, Y.; Peng, B.; He, L.; Fan, K.; Tong, L. Road segmentation of unmanned aerial vehicle remote sensing images using adversarial network with multiscale context aggregation. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2019**, *12*, 2279–2287. [CrossRef]
16. Jiang, L.; Xie, Y.; Ren, T. A deep neural networks approach for pixel-level runway pavement crack segmentation using drone-captured images. *arXiv* **2020**, arXiv:2001.03257.
17. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
18. Everingham, M.; Eslami, S.A.; Van Gool, L.; Williams, C.K.; Winn, J.; Zisserman, A. The pascal visual object classes challenge: A retrospective. *Int. J. Comput. Vis.* **2015**, *111*, 98–136. [CrossRef]
19. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; pp. 234–241.
20. Badrinarayanan, V.; Kendall, A.; Cipolla, R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [CrossRef] [PubMed]
21. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2881–2890.
22. Yu, F.; Koltun, V. Multi-scale context aggregation by dilated convolutions. *arXiv* **2015**, arXiv:1511.07122.
23. Rembold, F.; Atzberger, C.; Savin, I.; Rojas, O. Using low resolution satellite imagery for yield prediction and yield anomaly detection. *Remote Sens.* **2013**, *5*, 1704–1733. [CrossRef]
24. Alam, M.; Wang, J.F.; Guangpei, C.; Yunrong, L.; Chen, Y. Convolutional Neural Network for the Semantic Segmentation of Remote Sensing Images. *Mob. Netw. Appl.* **2021**, *26*, 200–215. [CrossRef]
25. Kerkech, M.; Hafiane, A.; Canals, R. Vine disease detection in UAV multispectral images using optimized image registration and deep learning segmentation approach. *Comput. Electron. Agric.* **2020**, *174*, 105446. [CrossRef]
26. Wang, T.; Thomasson, J.A.; Yang, C.; Isakeit, T.; Nichols, R.L. Automatic classification of cotton root rot disease based on UAV remote sensing. *Remote Sens.* **2020**, *12*, 1310. [CrossRef]
27. Yang, M.D.; Tseng, H.H.; Hsu, Y.C.; Tsai, H.P. Semantic segmentation using deep learning with vegetation indices for rice lodging identification in multi-date UAV visible images. *Remote Sens.* **2020**, *12*, 633. [CrossRef]
28. Sa, I.; Popović, M.; Khanna, R.; Chen, Z.; Lottes, P.; Liebisch, F.; Nieto, J.; Stachniss, C.; Walter, A.; Siegwart, R. WeedMap: A large-scale semantic weed mapping framework using aerial multispectral imaging and deep neural network for precision farming. *Remote Sens.* **2018**, *10*, 1423. [CrossRef]
29. Seferbekov, S.; Iglovikov, V.; Buslaev, A.; Shvets, A. Feature pyramid network for multi-class land segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 18–22 June 2018; pp. 272–275.
30. Revanasiddappa, B.; Arvind, C.; Swamy, S. Real-time early detection of weed plants in pulse crop field using drone with IoT. *Technology* **2020**, *16*, 1227–1242.
31. Akiva, P.; Dana, K.; Oudemans, P.; Mars, M. Finding berries: Segmentation and counting of cranberries using point supervision and shape priors. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 14–19 June 2020; pp. 50–51.
32. 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), Seattle, WA, USA, 14–19 June 2018; pp. 801–818.
33. National Geographic Information Institute. Available online: http://map.ngii.go.kr/ (accessed on 15 March 2020).
34. Evangelidis, G.D.; Psarakis, E.Z. Parametric image alignment using enhanced correlation coefficient maximization. *IEEE Trans. Pattern Anal. Mach. Intell.* **2008**, *30*, 1858–1865. [CrossRef]
35. MicaSense Imageprocessing. Available online: https://github.com/micasense/imageprocessing (accessed on 10 August, 2019).
36. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
37. Chen, L.C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking atrous convolution for semantic image segmentation. *arXiv* **2017**, arXiv:1706.05587.
38. Chollet, F. Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1251–1258.
39. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv* **2017**, arXiv:1704.04861.
40. Glorot, X.; Bengio, Y. Understanding the difficulty of training deep feedforward neural networks. In Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, JMLR Workshop and Conference Proceedings, Sardinia, Italy, 13–15 May 2010; pp. 249–256.
41. Loshchilov, I.; Hutter, F. Decoupled weight decay regularization. *arXiv* **2017**, arXiv:1711.05101.

42. Qin, J.; Wang, B.; Wu, Y.; Lu, Q.; Zhu, H. Identifying Pine Wood Nematode Disease Using UAV Images and Deep Learning Algorithms. *Remote Sens.* **2021**, *13*, 162. [CrossRef]
43. You, J.; Liu, W.; Lee, J. A DNN-based semantic segmentation for detecting weed and crop. *Comput. Electron. Agric.* **2020**, *178*, 105750. [CrossRef]
44. Kemker, R.; Salvaggio, C.; Kanan, C. Algorithms for semantic segmentation of multispectral remote sensing imagery using deep learning. *ISPRS J. Photogramm. Remote Sens.* **2018**, *145*, 60–77. [CrossRef]
45. Jiang, J.; Liu, F.; Xu, Y.; Huang, H. Multi-spectral RGB-NIR image classification using double-channel CNN. *IEEE Access* **2019**, *7*, 20607–20613. [CrossRef]