information

Article

Deep Hybrid Network for Land Cover Semantic Segmentation in High-Spatial Resolution Satellite Images

Sultan Daud Khan [1,*], Louai Alarabi [2] and Saleh Basalamah [3]

[1] Department of Computer Science, National University of Technology, Islamabad 44000, Pakistan
[2] Department of Computer Science, Umm Al-Qura University, Makkah 24236, Saudi Arabia; lmarabi@uqu.edu.sa
[3] Department of Computer Engineering, Umm Al-Qura University, Makkah 24236, Saudi Arabia; smbasalamah@uqu.edu.sa
* Correspondence: sultandaud@nutech.edu.pk

Abstract: Land cover semantic segmentation in high-spatial resolution satellite images plays a vital role in efficient management of land resources, smart agriculture, yield estimation and urban planning. With the recent advancement in remote sensing technologies, such as satellites, drones, UAVs, and airborne vehicles, a large number of high-resolution satellite images are readily available. However, these high-resolution satellite images are complex due to increased spatial resolution and data disruption caused by different factors involved in the acquisition process. Due to these challenges, an efficient land-cover semantic segmentation model is difficult to design and develop. In this paper, we develop a hybrid deep learning model that combines the benefits of two deep models, i.e., DenseNet and U-Net. This is carried out to obtain a pixel-wise classification of land cover. The contraction path of U-Net is replaced with DenseNet to extract features of multiple scales, while long-range connections of U-Net concatenate encoder and decoder paths are used to preserve low-level features. We evaluate the proposed hybrid network on a challenging, publicly available benchmark dataset. From the experimental results, we demonstrate that the proposed hybrid network exhibits a state-of-the-art performance and beats other existing models by a considerable margin.

Keywords: land cover classification; remote sensing; semantic segmentation; deep learning

Citation: Khan, S.D.; Alarabi, L.; Basalamah, S. Deep Hybrid Network for Land Cover Semantic Segmentation in High-Spatial Resolution Satellite Images. *Information* **2021**, *12*, 230. https://doi.org/10.3390/info12060230

Academic Editor: Willy Susilo

Received: 9 April 2021
Accepted: 24 May 2021
Published: 28 May 2021

1. Introduction

With the recent advancement in remote sensing technologies, such as satellites, drones, and airborne vehicles, etc., high-resolution satellite images are easy to acquire [1]. This opens up new paradigms and research directions for the remote sensing community that offer different applications in diverse fields, for example, land cover segmentation [2–4], smart agriculture [5,6], traffic monitoring [7,8], disaster management [9], geo-localization [10], and urban planning [11,12]. Among these applications, land cover classification and segmentation is an important application that extracts useful information about the type of land covered by agriculture, water, forest, urban, etc., which is crucial for land resource managers.

Currently, large field surveys are conducted to obtain information about land cover [13,14]. A manual analysis of large fields is a laborious and time-consuming job and often misses valuable information about the land cover [15,16]. With recent advances in computer vision and the success of deep neural networks with regard to optical natural images, several automated models [3,17] have been proposed in the literature that automatically perform semantic labeling (assign class) of land cover in high-resolution remote sensing images.

The main goal of semantic segmentation is dense prediction, which involves classifying each pixel of an image into different categories. Generally, semantic segmentation is a two step process: (1) feature extraction and (2) classification. The first step extracts various

features, e.g., texture and appearance, from the image and transforms spatial or temporal information into a discriminative feature set. The second task involves training a classifier that classifies each feature set into a correct class.

Currently, most state-of-the-art deep learning models for semantic segmentation follow the basic architecture of Fully Convolutional Networks [18]. FCN consists of two convolutional neural networks, (1) encoder and (2) decoder networks. The encoder network takes an input image of arbitrary type and pass the image through series of convolutional and pooling layers that extract hierarchical features. In other words, the encoder down samples the feature map to capture more semantic and contextual information. Typical examples of encoder networks are VGG-16 [19], Alexnet [20], ResNet [21], Xception [22]. The decoder network consists of deconvolution layers that upsample the feature map to capture spatial information.

Although FCN exhibited a good performance in various segmentation tasks, it suffers from the following limitations that make it unsuitable for land cover segmentation tasks.

1. Single scale problem: Current state-of-the-art FCN networks [23–28] are single scale and cannot exploit multi-scale information that results in the loss of valuable information. Generally, high-resolution satellite images contain a wide variety of objects having aspect ratios and scales. Furthermore, satellite images of land covers often consist of irregular regions, such as agriculture areas, forests, water, etc. To acquire a precise and rich semantic map of land cover remote sensing images, multi-scale contextual information is required. This will discriminate the targets with similar appearances but distinct semantic classes.
2. Large number of parameters: FCN-based semantic segmentation models require a large number of parameters for training, which leads to computation and memory constraints.
3. Long training time: Large number of redundant convolutional layers cause gradient vanishing problems and take a long time to train [29].

To address the above mentioned limitations of existing deep learning networks, we proposed a hybrid network that consists of two deep neural network architectures, DenseNet [30] and U-Net [23]. DenseNet is widely adopted network and exhibits a good performance in different multi-class object detection and segmentation tasks. DenseNet consists of densely connected blocks with different output resolutions connected in a feed-forward fashion. The network exploits residual connections and extracts contextual features at multiple scales. The network reuses the features from downsampling layers and concatenates feature maps from different layers to provide a variety of inputs for subsequent layers. It should be noted that downsampling layers extract local features; however, the resolution of the feature map is reduced by half after passing through downsampling dense blocks, which results in loss of important information. To avoid the information loss, U-Net replaced the max-pooling layers by upsampling layers, which increase the spatial resolution of the feature maps. The upsampling layers also contain a large number of channels that allows the network to capture contextual information and pass it to the higher layers. One of the limitations of U-Net is that the network is of limited depth and therefore cannot extract multi-scale features. To utilize the advantages of both networks, we combined both the networks in an efficient manner.

Generally, the framework consists of two path, i.e., dense contraction path and dense expansion path. Both paths are symmetrical and skip connections are used to combine both paths. The contraction path captures the context while the expanding path helps in prediction of a precise segmentation mask. We first trained the DenseNet on ImageNet [31] and then used the pre-trained model as a convolutional encoder in the contracting branch of U-Net. This transfer learning allowed us to learn complex segmentation tasks, since the model was pre-trained on 14 million images and had already learnt the complex features of the images. To reduce the computation time during training and testing, we adopted a cascaded approach that has been widely used in other object recognition tasks [32–34].

We summarize the contribution of this work as follows:

- We design an efficient hybrid network for land cover classification in high-resolution satellite images by carefully integrating two networks.
- The proposed network learns low-level features and high-level contexts in an efficient manner for improved land cover segmentation in satellite images.
- The network is trained in an end-to-end manner and improves the flow of information and parameters and avoids the problem of a long training time.
- We evaluated the performance of the proposed framework on a publicly available benchmark dataset. From experiment results, we demonstrate that the proposed framework exhibits a superior performance compared to other state-of-the-art methods.

2. Related Work

In this section, we first review generic semantic segmentation models and then provide a concise review of different models for land cover segmentation.

With the success of deep learning models in multi-object detection and classification tasks, deep models are also considered as favorite and viable solutions in semantic segmentation tasks. For semantic segmentation tasks, the first deep learning network was proposed in [18], which consists of fully convolutional layers and is trained in an end-to-end manner. The network takes an input image of arbitrary size and classifies each pixel of the input image into corresponding class labels. This network lost its popularity due to the presence of pooling layers that reduce the resolution of the feature map and cause significant loss of spatial information. To address this problem, U-Net is proposed in [23], which consists of encoder and decoder paths. The decoder path recovers spatial information by combining skip connections with deconvolution layers that upsample the feature maps. Due to its unique architecture, U-Net exhibited a good performance and has drawn much attention from the research community of medical image analysis [25,35]. Most recently, Li et al. [36] proposed the hybrid densely connected U-Net (H-DenseUNet) to exploit spatial information along the third dimension to the maximum extent. H-DenseUNet consists of 2D-DenseUNet and 3D-DenseUNet, which work in a cooperative manner. Similarly, [37] proposed a stacked U-Nets to solve the image segmentation problem. A multi-path refinement network, namely, RefineNet, is proposed in [38] to further exploit spatial information along the contraction path and uses long-range residual connections for high-resolution feature map prediction. A ResNet-like network is proposed in [39], which extracts high-level semantic information without losing spatial details, further enhancing the performance. The network consists of two streams—(1) the pooling stream and (2) residual stream. The pooling streams result in a low-resolution feature map, but extract high-level semantic information. The residual stream outputs high-resolution feature maps that maintain spatial information by using the features learned from the pooling stream. A high-fused convolutional neural network is proposed in [40] for an image semantic segmentation task. The network generates feature maps by fusing and reusing the features from lower layers. Similarly, a Discriminative Feature Network (DFN) is proposed in [41] that consists of two small networks: (1) smooth network and (2) border network. Smooth networks learn discriminative features by using global average pooling and Channel Attention Block (CAB), while the border network distinguishes the boundaries of multiple semantic regions by using semantic boundary supervision. To handle the scale problem in semantic segmentation tasks, DeepLabv3 is proposed in [42], which uses atrous convolutions in a serial/parallel manner to capture multiple scales of objects. The performance of DeepLabv3 is further improved in DeepLabv3+ [28] by incorporating a decoder network that refines and smooths the segmentation results along multiple semantic boundaries. Densely connected Atrous Spatial Pyramid Pooling (DenseASPP) [43] generates multiple multi-scale feature maps by connecting different atrous convolutional layers [44] that significantly increase the semantic segmentation accuracy. SegNet [26] is another encoder–decoder network for semantic segmentation, where the decoder part of the network recovers the spatial information by using the indices of the pooling step and upsamples the feature map in a non-linear manner.

In addition to the above mentioned general methods for semantic segmentation, researchers have proposed specialized methods for land cover segmentation in high-resolution satellite images. Kuo et al. [3] proposed a deep aggregation network for land cover segmentation. The network extracts and fuses feature maps from multiple layers for semantic segmentation. A graph-based fine tuning method is introduced to further enhance segmentation accuracy. The Dense dilated convolution's merging network (DDCM-Net) is introduced in [45] for land cover segmentation in satellite images. The network uses dilated convolutions and combines feature maps with different dilation rates, which enhances the receptive field of the network, helping to extract local and global contextual information. A hallucination network is proposed in [46] for land cover classification. The network avoids all modalities required during the feature fusion. The Feature pyramid network (FPN) is adopted in [2] to address the land cover segmentation problem. A classical neural network is proposed in [17] that optimizes the Jaccard index for the land cover classification problem. An Uncertainty Gated Network [47] is proposed that models the multi-scale contexts by leveraging the heteroscedastic measure of uncertainty for the classification of all pixels of a satellite image. The Dense Fusion Classmate Network (DFCNet) [48] incorporates mid-level information by using an auxiliary road dataset in addition to the deepglobe dataset [49] for land cover classification. An approach based on U-Net is used in [4] that uses Lovasz-Softmax loss to compensate for incomplete and incorrect labeling of data and data imbalance, problems that are commonly observed in land cover classification problems.

3. Methodology

The architecture of the proposed network for land segmentation is shown in Table 1. We adopted DenseNet-201 [30] as a feature extractor in our framework, which consists of 201 layers that are densely connected in the form of dense blocks. The output feature maps of each dense block has a different resolution and captures different contextual features of multiple scales. The network consists of four dense blocks and, within each dense block, convolutional layers are directly connected to other subsequent layers. The network takes an input of arbitrary size and applies a convolutional layer of kernel size 7×7, stride 2, followed by a max-pooling layer of kernel size 3×3 and stride of 2. The resultant feature maps are then passed through four densely connected convolutional blocks, namely, $denseblock_1$, $denseblock_2$, $denseblock_3$ and $denseblock_4$. Each dense block consists of set of two convolutional layers, where the kernel size of the first convolutional layer is 1×1 and size of the second convolutional layer is 3×3. Each $denseblock_i$ is repeated d times. In our architecture, the first dense block, $denseblock_1$ is repeated 6 times,and thus consists of $6 \times 2 = 12$ convolutional layers. The second dense block $denseblock_2$ is repeated 12 times and contains 24 convolutional layers. In the same way, $denseblock_3$ and $denseblock_4$ contain 96 layers and 64 convolutional layers, respectively. Each dense block is followed by a transition layer that consists of a set of one convolutional layer of size 1×1 followed by a pooling layer with kernel size of 2×2 and stride of 2. Such a dense connection among the layers within block improves the flow of information among the layers and avoids gradient vanishing, which is a common problem in shallow network architectures.

DenseNet-201 achieved significant performance gain in a multi-class classification task. The network takes an input image of fixed size and incorporates fully connected layers to the output classification score, while our problem is similar to a segmentation problem, where pixel-wise classification is required. To employ DenseNet-201 for the segmentation task, we replace fully connected layers with convolutional layers. Such a configuration allows the network to accept an input of arbitrary size and outputs a feature map instead of a classification score. However, the size of the output feature map is small and loses a significant amount of information of low-level features after passing through a series of max-pooling layers. To address this problem, in the decoder part, several upsampling operations are applied to produce dense feature map equal to size of input image. For example, one of the popular segmentation networks is U-Net, which was

originally proposed for segmentation of cell images. U-Net adopts a U-shape structure that consists of two paths, a contraction path and expansion path. The contraction path of U-Net consists of four layers to extract features. However, we observed that such a limited depth of U-Net can not extract feature of multiple scales.

Table 1. Architecture of the proposed network. The network consists of two parts, i.e., encoder and decoder. The encoder part consists of denseblocks and transition layers. The decoder part consists of upsampling layers. $Denseblock_i \times d$ represents denseblock i, where d represents the repetition of denseblock.

Layer	Operation	Kernel Size	# of Channels	Stride	Feature Size
Input	-	-	-	-	256 × 256
Encoder Part					
Convolution	Conv	7 × 7	96	2	128 × 128
Pooling	Max pooling	3 × 3	-	2	64 × 64
Denseblock1 × 6	Conv	1 × 1	192	1	64 × 64
	Conv	3 × 3	48	1	64 × 64
Transition Layer1	Conv	1 × 1	48	1	64 × 64
	Avg Pooling	2 × 2	-	2	32 × 32
Denseblock2 × 12	Conv	1 × 1	192	1	32 × 32
	Conv	3 × 3	48	1	32 × 32
Transition Layer2	Conv	1 × 1	48	1	32 × 32
	Avg Pooling	2 × 2	-	2	16 × 16
Denseblock3 × 48	Conv	1 × 1	192	1	16 × 16
	Conv	3 × 3	48	1	16 × 16
Decoder Part					
Transition layer3	Conv	1 × 1	48	1	16 × 16
	Avg Pooling	2 × 2	-	2	8 × 8
Denseblock4 × 32	Conv	1 × 1	192	1	8 × 8
	Conv	3 × 3	48	1	8 × 8
Up sampling layer 1	D-conv	2 × 2	-	-	16 × 16
	Conv	3 × 3	768	1	16 × 16
Up sampling layer 2	D-conv	2 × 2	-	-	32 × 32
	Conv	3 × 3	384	1	32 × 32
Up sampling layer 3	D-conv	2 × 2	-	-	64 × 64
	Conv	3 × 3	384	1	64 × 64
Up sampling layer 3	D-conv	2 × 2	-	-	128 × 128
	Conv	3 × 3	96	1	128 × 128
Up sampling layer 3	D-conv	2 × 2	-	-	256 × 256
	Conv	3 × 3	96	1	256 × 256
Convolution	Conv	1 × 1	17	1	256 × 256

Considering the above mentioned limitations, we developed a hybrid network that combines the advantages of both DenseNet and U-Net. We replaced the contraction path of U-Net with DenseNet-201 to extract features of multiple scales, while using the long-range connection of the U-Net concatenate encoding and decoding path to conserve

low-level features. The contraction path (encoder part) consists of four dense blocks, where each dense block is followed by a transition layer. A contraction mechanism was also implemented inside the transition layer to control the expansion of the feature maps. The expansion pa0 .3th involves upsampling and merging operati.0 o0ns followed by a convolutional operation that expands the resolution of the current feature map. To predict the segmentation mask, the expansion path utilizes skip connections [23] to merge the upsampled feature map with its corresponding feature map from the contraction path. The Softmax layer then assigns class probability to each pixel and outputs a 2-dimensional segmentation mask.

3.1. Loss Function, Training and Testing Strategies

In this section, we now discuss details of loss function, training and testing strategies.

3.1.1. Loss Function

Generally, deep learning networks use cross-entropy loss to optimize the cost function. Cross-entropy loss performs well in different object classification, detection and segmentation tasks. However, cross-entropy loss can not handle class imbalance problems, which are commonly observed in multi-class semantic segmentation problems. In these problems, the training data are always limited and expensive to acquire. Each sample of the training dataset affects the loss function regardless of the training scheme; therefore, the number of samples per class can change the shape of the loss function. For example, a dominant class that contains more samples compared to other classes will affect the loss function more and bias the overall training process.

To address the above problem, we used multi-class hybrid loss (L_{mchl}), which is the linear combination of cross-entropy loss (also termed as local loss) defined in Equation (1) and dice loss (also termed as global loss) defined in Equation (2).

$$L_c(y, \hat{y}) = -\frac{1}{N_c}\sum_{n}^{N_c} y \ln(\hat{y}) + (1-y)\ln(1-\hat{y}) \tag{1}$$

where L_c is the local loss and it measures the sum of cross-entropy loss for each pixel, N_c is the number of classes, y is the ground truth label of pixel and $\hat{y}$ is the predicted label.

Dice loss L_d is the global loss and measures the segmentation score by comparing the similarity between two images and defined as follows:

$$L_d = \frac{2|G \cap P|}{|G| + |P|} \tag{2}$$

where G is the ground truth image and P is the predicted image. We then computed the multi-class hybrid loss (L_{mchl}) as defined in Equation (3).

$$L_{mchl} = L_c + L_d \tag{3}$$

3.1.2. Training Scheme

For training the network, we used images and corresponding ground truth segmentation masks and optimized the objective function by stochastic gradient descent implemented in Pytorch. We used the images and their corresponding ground truth segmentation masks for training the network. Instead of using batch size 1, as adopted in [23], we kept the batch size as 8 to converge the network to global optima. We started with a learning rate of 0.001 and used a cyclical schedule learning rate strategy as in [50], where learning rate linearly decreased in each cycle. We trained the network for 100 epochs.

3.1.3. Testing Scheme

During the testing phase, we provided an image of size 256 $\times$ 256 as an input to the network. However, before the input step, we performed normalization in order to boost the testing speed. We normalized the input image by subtracting the mean (calculated from the

training set) from each pixel of the input image and then divided the result by the standard deviation. Furthermore, we converted the RGB image into a single channel gray scale using the weighted sum of the R, G, and B channels. The network then predicted the segmentation mask, f_m, where each pixel is assigned class probability. We then performed morphological operations on the mask f_m. In the following are the reasons for performing morphological operations: (1) to remove imperfections caused by camera motion and other random noises; (2) to group the adjacent objects belonging to the same class as a single one; (3) to separate the foreground and background pixels. For applying morphological operations, we first converted the mask f_m into a binary mask f_b by labeling all the foreground pixels as 1 and background pixels as 0. We first applied morphological closing by using the structuring element of size 3×3 to a binary image to fill the small holes. Then, we performed erosion with structuring element of size 3×3, followed by dilation with structuring element of the size 5×5. We then obtained a refined segmented mask f_r by employing a fusion operation, as denoted by $f_m \odot f_b$, where $\odot$ denotes the element wise product. We then computed the area ratio α, which is the ratio of the area of the blob i to the maximum area of for all the blobs formulated as $\frac{A_i}{\arg\max(\sum_{i=1}^{K} A_i)}$. We then defined a threshold ω and removed blobs with areas less than the threshold. In all our experiments, we fixed the value of ω to 0.05.

4. Experiment Results

In this section, we evaluate and compare the the performance of proposed framework in qualitative and quantitative ways. We first introduce the publicly available dataset used to evaluate the proposed framework and then compare performance with other state-of-the-art methods.

For performance evaluation, we used the Dense labeling remote sensing dataset (DLRSD). The DLRSD dataset is densely labeled by Shao et al. [51] using the images from the UC Merced archive [52], where the spatial resolution of the image is 30 cm (or 1 foot). Specifically, each image in UC Merced archive [52] was manually labeled into 17 distinct classes by using eCognition 9.0 software (http://www.ecognition.com). These classes were assigned labels in the following order: 1: airplane, 2: bare soil, 3: building, 4: car, 5: chaparral, 6: court, 7: dock, 8: field, 9:grass, 10: mobile home, 11: pavement, 12: sand, 13: sea, 14: ship, 15: tank, 16: tree, 17: water. Since each image is labeled pixel-wise, this dataset can be used for evaluating semantic segmentation tasks in remote sensing images. The dataset consists of 2100 images with each image having the resolution of 256×256 pixels. Figure 1 shows the distribution of number of images per class. From the figure, it is obvious that the dataset suffers from a class imbalance problem that may result in poor generalization of the network. In order to address this problem, we evenly selected 100 samples from each class. Figure 2 shows some sample frames with their corresponding ground truth segmentation masks.

To evaluate the performance of different models, we use following performance metrics: Hamming loss, precision, recall, accuracy, and F-score. For comprehensive comparison, we divide the state-of-the-art models into two groups: (1) hand-crafted feature models and (2) deep learning models.

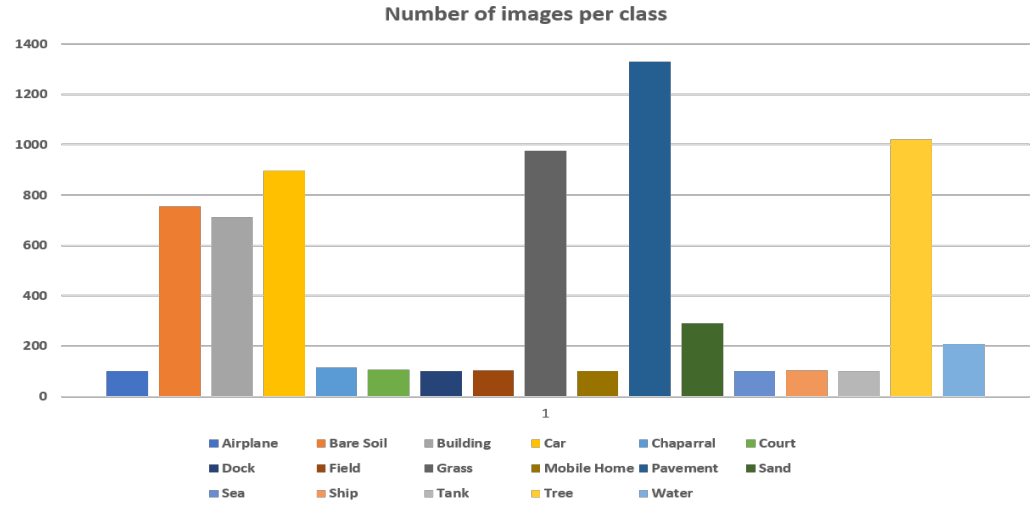

Figure 1. Distribution of number of images per class.

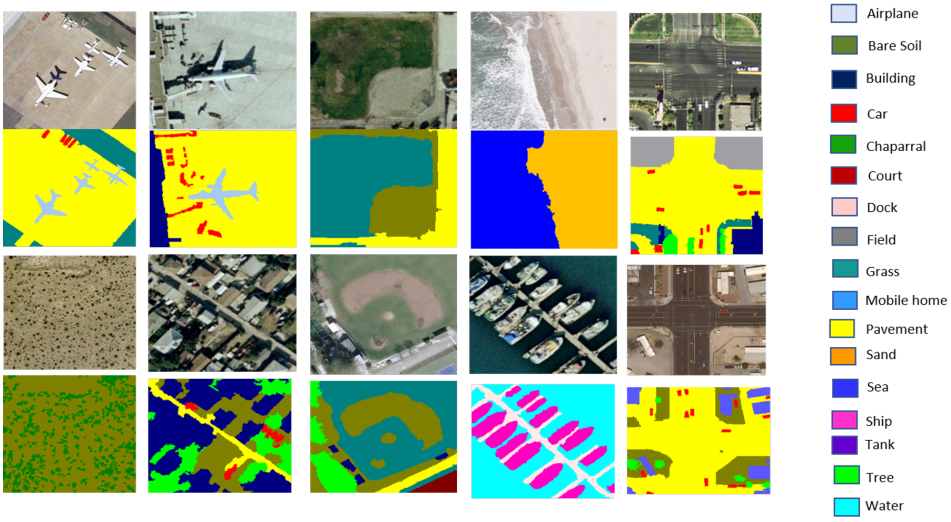

Figure 2. Sample images from the dataset and their corresponding ground truth segmented masks encoded in different colors. The first and third rows show the samples frames. Second and fourth rows show the corresponding segmentation masks.

4.1. Hand-Crafted Feature Models

Hand-crafted feature models include, (1) local binary pattern (LBP) [53], (2) Gabor filter [54], (3) GIST features [55], (4) Bag-of-Visual-Words (BoVW) [56], (5) color histogram.

In the first method, we first divided the image into small non-overlapping regions. We then computed the LBP feature of each region and concatenated them into a single feature vector and trained a multi-class classifier using a Support vector machine (SVM). In the second method, a spectral approach, as adopted in [57], was used to texture satellite images. The texture images were then converted into feature images by employing Gabor filters. A unique set of feature vectors were generated from feature images, where each feature vector points to one dimension of feature space. An unsupervised fuzzy c-means clustering method was then adopted to classify each pixel of an image into a specific category based on the associated feature vector. In the third method, we extracted global feature GIST [55] features from satellite images. These features were obtained by convoluting a kernel (filter) with an image at different orientations and scales to obtain high- and low-repetitive structures of an image. The feature space was then reduced by ranking through principal component analysis (PCA) to select discriminating features. In the fourth method, we uniformly sampled points from an input image. Then, with each point as the center, we

extracted a patch of size 28 × 28. We then extracted SIFT features from each patch and generated a visual dictionary. Next, K-means clustering was employed to generate visual dictionary vocabulary. Similarly, in the last method, we divided an input image into non-overlap patches and then extracted a color histogram after quantizing the RGB channel of each patch into 32 bins. Three histograms for each channel were then concatenated and trained via the SVM classifier.

4.2. Deep Learning Models

To compare the performance of proposed framework with other state-of-the-art deep learning models, we selected popular segmentation models that include U-Net [23], U-Net++ [25], SegNet [26], Multi-scale fully convolutional network (MSFCN) [58], Tiramisu [59], FGC [60], CE-Net [61], DenseNet [30], and U-NetPPL [62].

U-Net [23] was initially proposed for biomedical image segmentation tasks and won ISBI cell tracking challenge in 2015. The network consists of contracting path that extracts contextual feature and expanding path allows precise localization of objects. U-Net++ [25] is the extension of U-Net that addresses the problems of U-Net and was originally proposed for medical image segmentation problems. U-NetPPL [62] extended and improved U-Net by incorporating pyramid pooling layers (PPL) for multi-object segmentation task. U-Net++ is a deep encoder–decoder network, where the skip pathways of the original U-Net are re-designed to minimize the semantic gap between the feature maps of two paths. Similarly, SegNet [26] is also an encoder–decoder network, where the decoder part is followed by a classification layer that classifies pixels into specific categories. The architecture of the encoder part of the network is similar to VGG16 [19], while the decoder part is modified in a manner that upsamples the feature maps by using the pooling indices of the max-pooling layer of the encoder. Dense feature maps are then obtained by convolving the upsampled maps with pre-trained filters. Multi-scale fully convolutional network (MSFCN) [58] uses 3D-CNN to incorporate both spatial and temporal features. Similarly, FGC [60] consists of a 3D fully convolutional network to extract both spatial and temporal features for crop classification from temporal remote sensing images. The network uses a 3D channel attention module to enable channel consistency between the feature maps of the encoder and decoder parts and the 3D global pooling method is used for selecting discriminating features. Tiramisu [59] extended DenseNet [30] for semantic segmentation problem and achieved promising results. CE-Net [61] is proposed for 2D medical image segmentation tasks. The network consists of three major modules, encoder, context extraction module and decoder module. The network covers the limitations of conventional U-Net by capturing more high-level information and retains spatial information that enables precise localization of a target.

We randomly selected and utilized 80% for training and 30% for testing. We utilized the pre-trained models of the above mentioned deep learning methods and used transfer learning to train the models on the DLRSD dataset. During training, we used a learning rate of 0.0001 with a batch size of 10. We used NVIDIA TITAN Xp GPU with RAM of 12 GB for experiments.

The comparison results of different hand-crafted feature methods and deep learning models are reported in Tables 2 and 3, respectively. From both tables, it is obvious that deep learning models outperform hand-crafted feature methods. This is due to the fact that hand-crafted feature models rely on the computation of complex features that are affected by natural factors, such as changes in illumination, scale, and object size. Local binary patterns compute the local structures that lead to performance gain in texture analysis. However, we observed that this method suffers from the following limitations.

LBP generates a long histogram by computing texture values from a small neighbourhood (3 × 3 pixels). This small neighbourhood extracts a limited amount of texture information from the local patch of an image and loses significant amount of information, which decreases the segmentation accuracy. Furthermore, the method directly computes the difference of neighbouring pixels and, therefore, is highly sensitive to illumination and

noise. A small change in illumination may bring a significant change in the texture of the image, which may be a challenge for LBP to discriminate actual texture from that of noise. The Gabor filter computes a high-dimension matrix that contains many redundant features. This redundancy of features decreases the performance of semantic segmentation process. Bag-of-Visual-Words could also not produce comparable results. This is due to fact that BoVW cannot capture rich contextual information, which is required for segmentation of high-resolution satellite images. Moreover, BoVW avoids the spatial relationship among different patches of an image and generates a high-dimension feature vector that avoids utilization of co-occurrence statistics among the words of the visual vocabulary. GIST uses low-level features to capture high-level semantic information of the scene; however, the method avoids local objects and their relationships with the scene. Similarly, the color histogram method is significantly affected by illumination and its appearance changes and does not exploit spatial information, which results in a reduced performance.

Table 2. Performance comparison of hand-crafted feature methods.

	Accuracy	Precision	Recall	F1-Score
Local Binary Pattern	49.04	44.71	42.83	43.75
Gabor Filter	51.29	49.75	43.29	46.30
GIST Features	39.26	37.19	39.62	38.37
Bag-of-Visual-Words	54.54	45.23	51.34	48.09
Color Histogram	48.95	40.33	42.39	41.33
Proposed	77.67	75.20	70.54	72.80

Table 3. Performance comparison of different deep learning models.

	Accuracy	Precision	Recall	F1-Score
U-Net	65.73	64.27	57.24	60.55
U-Net++	70.29	61.75	70.25	65.73
SegNet	63.24	65.46	57.27	61.09
MS-FCN	71.52	68.95	65.29	67.07
CE-Net	69.79	59.29	64.95	61.99
U-NetPPL	68.55	55.67	66.38	60.56
FGC	63.29	52.37	65.43	58.18
Tiramisu	69.42	60.89	62.28	61.58
DenseNet	57.12	49.65	55.24	52.30
Proposed	77.67	75.20	70.54	72.80

In contrast to all hand-crafted feature methods, deep learning models extract hierarchical features directly from the raw images and capture context-rich and semantic information by directly learning the discriminating features automatically during the training process. To summarize the discussion, we report the following findings from the experiment results reported in Tables 2 and 3. Deep hierarchical features achieve superior and relatively consistent results compared to hand-craft features for semantic segmentation tasks. Deep features, in contrast to hand-crafted features, are more robust and are not affected by appearance and illumination changes.

The performances of different deep learning models are reported in Table 3. The performance of DenseNet is lower compare to other state-of-the-art methods in the segmentation problems. This is due to fact that DenseNet suffers from downsampling problems, where the resolution of feature maps is reduced after passing through each dense block, which

results in loss of important information on the small objects. However, DenseNet extracts rich contextual information, which is why we used it as an encoder in our framework. To avoid the information loss, we replaced max-pooling layers by upsampling layers in the decoder part of the proposed framework, which not only increased the spatial resolution of the feature maps but also retained spatial information of all objects. This is due to reason that proposed method outperforms other state-of-the-art deep learning models by a considerable margin. The proposed method achieves near 8% and 9.5% gains in accuracy compared to MS-FCN and U-Net++ models, respectively. These state-of-the-art models achieve comparable performances in terms of accuracy, precision, recall and F-score on DLRSD dataset. These state-of-the-art models are the variants of U-Net and Fully convolutional network (FCN) that suffer from the following limitations: (1) The depth of these networks is limited due to which they capture contextual features in limited scales. For example, for the input image of size 256×256, only four convolutional layers are applied to capture semantic information. (2) The skip connection of U-Net enables a fusion scheme that allows the aggregation of feature maps at the same scale, which adversely affects the segmentation process. The proposed framework overcomes these limitations and exhibited a state-of-the-art performance by replacing the stack of convolutional and pooling layers with deep stack of dense blocks and transition layers. The skip connection of the proposed framework enables a flexible and effective fusion scheme that aggregates a feature map of a multiple semantic scale. The framework effectively re-uses the feature map by linking each dense block to the previous blocks in feed-forward mode.

In Table 4, we report the performance of the proposed framework in terms of class wise accuracy, precision, recall and F1-score. From the table, it is obvious that the proposed framework exhibits a higher performance when segmenting small, medium and large object remote sensing images. From the experiments, we observed that the proposed method achieved good results for all classes. It is important to note that the proposed method obtained the best results for classes of large and medium objects, such as, "Airplane (89.56%)", "Building (83.43%)", "Dock (82.67%)", "Grass (88.47%)", and "Ship (87.54%)". We noticed that the method produced a considerable number of true positives compared to false positive and negatives; therefore, the framework obtained a reasonable balance between precision an recall for most of the classes as well as achieved a good F1-score (greater than 70%) for more than half of the classes. Furthermore, the performance of the proposed method is also illustrated in Figure 3, where we compared the ground truth and predicted masks. From the figure, it is obvious that the proposed method predicts high-quality segmentation masks by precisely classifying different categories and accurately differentiating complicated class boundaries in various satellite images.

Another advantage of the proposed network is that it effectively accelerates the learning by effectively handling the gradient vanishing problem. Stacking multiple convolutional layers causes the gradient vanishing problem, making it difficult for a model to converge. The proposed network handles the gradient vanishing problem by integrating U-Net, which uses long-range skip connections between the encoding and decoding parts. Although there is no theoretical justification, long-range skip connections perform well in dense prediction tasks [63]. This may be attributed to the fact that long-range skip connections simplify the network, improve the flow of information through a few layers and accelerate the learning process. The models that use long-range skip connections can converge faster and achieve better performances. To demonstrate the effectiveness of the proposed network, we compared the training losses of the proposed network with/without long-range skip connections, as reported in Figure 4. We used two variants of the proposed network—one without using skip-connections and the other with a long-range skip connection. From the figure, it is obvious that the proposed network converges faster and achieves lower loss values than its counterpart without long-range skip connections.

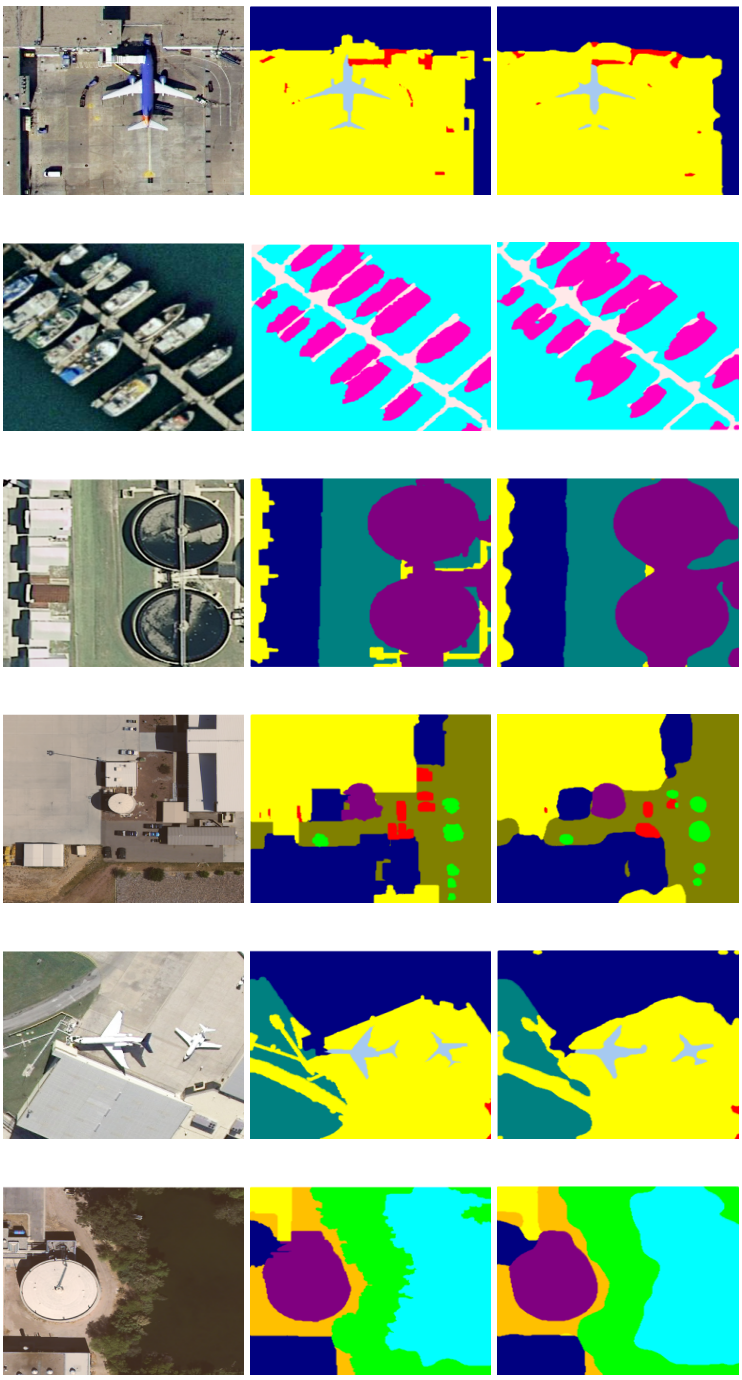

Figure 3. Visualization of the land cover semantic segmentation by the proposed method. The first column shows samples' frames from the dataset. Second column is the ground truth masks and third column is the predicted segmentation masks.

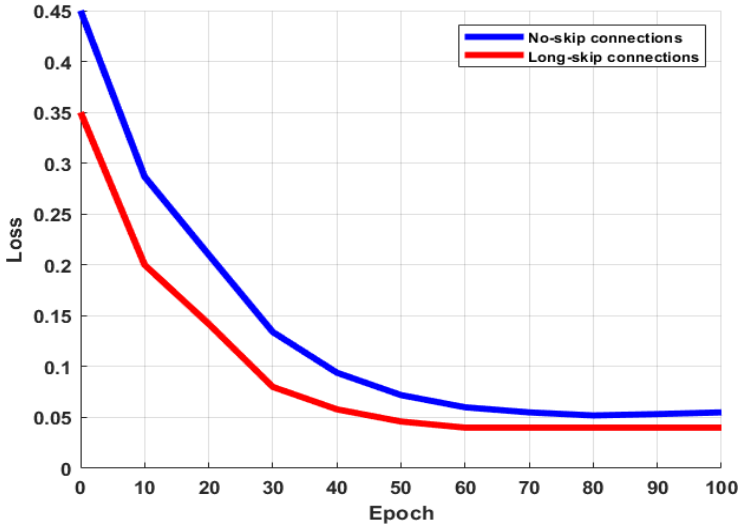

Figure 4. Long-range skip connections vs. no-skip connections. The figure demonstrates the effectiveness of using long-range skip connections to avoid gradient vanishing and accelerate the convergence.

Table 4. Class-wise performance of the proposed method.

	Accuracy	Precision	Recall	F1-Score
Airplane	89.56	86.24	82.76	84.46
Bare soil	78.94	79.14	69.45	73.98
Building	83.43	78.62	81.02	79.80
Car	79.38	70.92	79.79	75.09
Chaparral	65.95	71.69	52.76	60.79
Court	78.16	77.76	69.19	73.23
Dock	82.67	83.48	72.79	77.77
Field	73.25	78.94	55.72	65.33
Grass	88.47	84.64	82.79	83.70
Mobile home	67.73	67.65	58.76	62.89
Pavement	82.19	79.23	75.64	77.39
Sand	76.92	68.47	75.08	71.62
Sea	74.02	73.97	65.48	69.47
Ship	87.54	92.17	75.46	82.98
Tank	73.32	64.39	67.42	65.87
Tree	64.74	56.75	60.37	58.50
Water	74.25	64.38	74.76	69.18

5. Conclusions

In this work, we proposed a hybrid network for land cover semantic segmentation from high-spatial resolution satellite images. The proposed hybrid network combines the benefits of two deep learning models and has the following key features: (1) The network learns low-level features and high-level contexts by replacing the contraction path of U-Net with a stack of dense blocks. (2) The skip connections of the network enable effective fusion, which aggregates the features map on multiple scales. This enables the network to create pixel-wise segments of small, medium and large objects. We evaluated and compared the proposed network using a publicly available challenging dataset. From the experiment

results, we demonstrated that the proposed network produces an accurate segmentation map and beats other state-of-the-art methods by a considerable margin.

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

Funding: This research received no external funding.

Data Availability Statement: The dataset used in this research are publicly available and can be downloaded from http://weegee.vision.ucmerced.edu/datasets/landuse.html.

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

References

1. Mboga, N.; Georganos, S.; Grippa, T.; Lennert, M.; Vanhuysse, S.; Wolff, E. Fully convolutional networks and geographic object-based image analysis for the classification of VHR imagery. *Remote Sens.* **2019**, *11*, 597. [CrossRef]
2. 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.
3. Kuo, T.S.; Tseng, K.S.; Yan, J.W.; Liu, Y.C.; Frank Wang, Y.C. Deep aggregation net for land cover classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 18–22 June 2018; pp. 252–256.
4. Rakhlin, A.; Davydow, A.; Nikolenko, S. Land cover classification from satellite imagery with u-net and lovász-softmax loss. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 18–22 June 2018; pp. 262–266.
5. Chiu, M.T.; Xu, X.; Wei, Y.; Huang, Z.; Schwing, A.G.; Brunner, R.; Khachatrian, H.; Karapetyan, H.; Dozier, I.; Rose, G.; et al. Agriculture-vision: A large aerial image database for agricultural pattern analysis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 2828–2838.
6. Maddikunta, P.K.R.; Hakak, S.; Alazab, M.; Bhattacharya, S.; Gadekallu, T.R.; Khan, W.Z.; Pham, Q.V. Unmanned aerial vehicles in smart agriculture: Applications, requirements, and challenges. *IEEE Sens. J.* **2021**. [CrossRef]
7. Larsen, S.Ø.; Salberg, A.B.; Eikvil, L. Automatic system for operational traffic monitoring using very-high-resolution satellite imagery. *Int. J. Remote Sens.* **2013**, *34*, 4850–4870. [CrossRef]
8. Drouyer, S.; de Franchis, C. Highway traffic monitoring on medium resolution satellite images. In Proceedings of the IGARSS 2019—2019 IEEE International Geoscience and Remote Sensing Symposium, Yokohama, Japan, 28 July–2 August 2019; pp. 1228–1231.
9. Wheeler, B.J.; Karimi, H.A. Deep learning-enabled semantic inference of individual building damage magnitude from satellite images. *Algorithms* **2020**, *13*, 195. [CrossRef]
10. Hu, S.; Lee, G.H. Image-based geo-localization using satellite imagery. *Int. J. Comput. Vis.* **2020**, *128*, 1205–1219. [CrossRef]
11. Sirmacek, B.; Unsalan, C. A probabilistic framework to detect buildings in aerial and satellite images. *IEEE Trans. Geosci. Remote Sens.* **2010**, *49*, 211–221. [CrossRef]
12. Kazemzadeh-Zow, A.; Darvishi Boloorani, A.; Samany, N.N.; Toomanian, A.; Pourahmad, A. Spatiotemporal modelling of urban quality of life (UQoL) using satellite images and GIS. *Int. J. Remote Sens.* **2018**, *39*, 6095–6116. [CrossRef]
13. Su, M.; Guo, R.; Chen, B.; Hong, W.; Wang, J.; Feng, Y.; Xu, B. Sampling Strategy for Detailed Urban Land Use Classification: A Systematic Analysis in Shenzhen. *Remote Sens.* **2020**, *12*, 1497. [CrossRef]
14. MohanRajan, S.N.; Loganathan, A.; Manoharan, P. Survey on Land Use/Land Cover (LU/LC) change analysis in remote sensing and GIS environment: Techniques and Challenges. *Environ. Sci. Pollut. Res.* **2020**, *27*, 29900–29926. [CrossRef]
15. Zhang, C.; Han, Y.; Li, F.; Gao, S.; Song, D.; Zhao, H.; Fan, K.; Zhang, Y. A new CNN-Bayesian model for extracting improved winter wheat spatial distribution from GF-2 imagery. *Remote Sens.* **2019**, *11*, 619. [CrossRef]
16. Basso, B.; Liu, L. Seasonal crop yield forecast: Methods, applications, and accuracies. *Adv. Agron.* **2019**, *154*, 201–255.
17. Davydow, A.; Nikolenko, S. Land cover classification with superpixels and jaccard index post-optimization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 18–22 June 2018; pp. 280–284.
18. 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.
19. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
20. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **2012**, *25*, 1097–1105. [CrossRef]

21. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
22. 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.
23. 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.
24. Paszke, A.; Chaurasia, A.; Kim, S.; Culurciello, E. Enet: A deep neural network architecture for real-time semantic segmentation. *arXiv* **2016**, arXiv:1606.02147.
25. Zhou, Z.; Siddiquee, M.M.R.; Tajbakhsh, N.; Liang, J. Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 3–11.
26. 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]
27. Wang, Y.; Zhou, Q.; Liu, J.; Xiong, J.; Gao, G.; Wu, X.; Latecki, L.J. Lednet: A lightweight encoder-decoder network for real-time semantic segmentation. In Proceedings of the 2019 IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, 22–25 September 2019; pp. 1860–1864.
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. Huang, G.; Sun, Y.; Liu, Z.; Sedra, D.; Weinberger, K.Q. Deep networks with stochastic depth. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 646–661.
30. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
31. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.* **2015**, *115*, 211–252. [CrossRef]
32. Farag, A.; Lu, L.; Roth, H.R.; Liu, J.; Turkbey, E.; Summers, R.M. A bottom-up approach for pancreas segmentation using cascaded superpixels and (deep) image patch labeling. *IEEE Trans. Image Process.* **2016**, *26*, 386–399. [CrossRef] [PubMed]
33. Zhou, Y.; Xie, L.; Fishman, E.K.; Yuille, A.L. Deep supervision for pancreatic cyst segmentation in abdominal CT scans. In Proceedings of the InternationaL Conference on Medical Image Computing and Computer-Assisted Intervention, Quebec City, QC, Canada, 11–13 September 2017; pp. 222–230.
34. Roth, H.R.; Lu, L.; Lay, N.; Harrison, A.P.; Farag, A.; Sohn, A.; Summers, R.M. Spatial aggregation of holistically-nested convolutional neural networks for automated pancreas localization and segmentation. *Med. Image Anal.* **2018**, *45*, 94–107. [CrossRef]
35. Çiçek, Ö.; Abdulkadir, A.; Lienkamp, S.S.; Brox, T.; Ronneberger, O. 3D U-Net: Learning dense volumetric segmentation from sparse annotation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Athens, Greece, 17–21 October 2016; pp. 424–432.
36. Li, X.; Chen, H.; Qi, X.; Dou, Q.; Fu, C.W.; Heng, P.A. H-DenseUNet: Hybrid densely connected UNet for liver and tumor segmentation from CT volumes. *IEEE Trans. Med. Imaging* **2018**, *37*, 2663–2674. [CrossRef]
37. Shah, S.; Ghosh, P.; Davis, L.S.; Goldstein, T. Stacked U-Nets: A no-frills approach to natural image segmentation. *arXiv* **2018**, arXiv:1804.10343.
38. Lin, G.; Milan, A.; Shen, C.; Reid, I. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1925–1934.
39. Pohlen, T.; Hermans, A.; Mathias, M.; Leibe, B. Full-resolution residual networks for semantic segmentation in street scenes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4151–4160.
40. Yang, T.; Wu, Y.; Zhao, J.; Guan, L. Semantic segmentation via highly fused convolutional network with multiple soft cost functions. *Cogn. Syst. Res.* **2019**, *53*, 20–30. [CrossRef]
41. Yu, C.; Wang, J.; Peng, C.; Gao, C.; Yu, G.; Sang, N. Learning a discriminative feature network for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 1857–1866.
42. Chen, L.C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking atrous convolution for semantic image segmentation. *arXiv* **2017**, arXiv:1706.05587.
43. Yang, M.; Yu, K.; Zhang, C.; Li, Z.; Yang, K. Denseaspp for semantic segmentation in street scenes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 3684–3692.
44. 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]

45. Liu, Q.; Kampffmeyer, M.; Jenssen, R.; Salberg, A.B. Dense dilated convolutions' merging network for land cover classification. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 6309–6320. [CrossRef]
46. Kampffmeyer, M.; Salberg, A.B.; Jenssen, R. Urban land cover classification with missing data modalities using deep convolutional neural networks. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2018**, *11*, 1758–1768. [CrossRef]
47. Pascual, G.; Seguí, S.; Vitria, J. Uncertainty gated network for land cover segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 18–22 June 2018; pp. 276–279.
48. Tian, C.; Li, C.; Shi, J. Dense fusion classmate network for land cover classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 18–22 June 2018; pp. 192–196.
49. Demir, I.; Koperski, K.; Lindenbaum, D.; Pang, G.; Huang, J.; Basu, S.; Hughes, F.; Tuia, D.; Raskar, R. Deepglobe 2018: A challenge to parse the earth through satellite images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 18–22 June 2018; pp. 172–181.
50. Garipov, T.; Izmailov, P.; Podoprikhin, D.; Vetrov, D.; Wilson, A.G. Loss surfaces, mode connectivity, and fast ensembling of dnns. *arXiv* **2018**, arXiv:1802.10026.
51. Shao, Z.; Yang, K.; Zhou, W. Performance evaluation of single-label and multi-label remote sensing image retrieval using a dense labeling dataset. *Remote Sens.* **2018**, *10*, 964. [CrossRef]
52. Yang, Y.; Newsam, S. Geographic image retrieval using local invariant features. *IEEE Trans. Geosci. Remote Sens.* **2012**, *51*, 818–832. [CrossRef]
53. Ahonen, T.; Hadid, A.; Pietikäinen, M. Face recognition with local binary patterns. In Proceedings of the European Conference on Computer Vision, Prague, Czech Republic, 11–14 May 2004; pp. 469–481.
54. Mehrotra, R.; Namuduri, K.R.; Ranganathan, N. Gabor filter-based edge detection. *Pattern Recognit.* **1992**, *25*, 1479–1494. [CrossRef]
55. Oliva, A.; Torralba, A. Modeling the shape of the scene: A holistic representation of the spatial envelope. *Int. J. Comput. Vis.* **2001**, *42*, 145–175. [CrossRef]
56. Sivic, J.; Zisserman, A. Video Google: A text retrieval approach to object matching in videos. In Proceedings of the IEEE International Conference on Computer Vision, Nice, France, 13–16 October 2003; Volume 3, p. 1470.
57. Idrissa, M.; Acheroy, M. Texture classification using Gabor filters. *Pattern Recognit. Lett.* **2002**, *23*, 1095–1102. [CrossRef]
58. Li, R.; Zheng, S.; Duan, C. Land cover classification from remote sensing images based on multi-scale fully convolutional network. *arXiv* **2020**, arXiv:2008.00168.
59. Jégou, S.; Drozdzal, M.; Vazquez, D.; Romero, A.; Bengio, Y. The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, HI, USA, 21–26 July 2017; pp. 11–19.
60. Ji, S.; Zhang, Z.; Zhang, C.; Wei, S.; Lu, M.; Duan, Y. Learning discriminative spatiotemporal features for precise crop classification from multi-temporal satellite images. *Int. J. Remote Sens.* **2020**, *41*, 3162–3174. [CrossRef]
61. Gu, Z.; Cheng, J.; Fu, H.; Zhou, K.; Hao, H.; Zhao, Y.; Zhang, T.; Gao, S.; Liu, J. Ce-net: Context encoder network for 2d medical image segmentation. *IEEE Trans. Med. Imaging* **2019**, *38*, 2281–2292. [CrossRef] [PubMed]
62. Kim, J.H.; Lee, H.; Hong, S.J.; Kim, S.; Park, J.; Hwang, J.Y.; Choi, J.P. Objects segmentation from high-resolution aerial images using U-Net with pyramid pooling layers. *IEEE Geosci. Remote Sens. Lett.* **2018**, *16*, 115–119. [CrossRef]
63. Hoang, H.H.; Trinh, H.H. Improvement for Convolutional Neural Networks in Image Classification Using Long Skip Connection. *Appl. Sci.* **2021**, *11*, 2092. [CrossRef]