remote sensing

MDPI

Article

Semantic Segmentation for Buildings of Large Intra-Class Variation in Remote Sensing Images with O-GAN

Shuting Sun [1], Lin Mu [2,3], Lizhe Wang [4], Peng Liu [5,*], Xiaolei Liu [6] and Yuwei Zhang [5]

1 Collage of Marine Science and Technology, China University of Geosciences (CUG), Wuhan 430074, China; 2201910378@cug.edu.cn
2 College of Life Sciences and Oceanography, Shenzhen University, Shenzhen 518060, China; mulin@szu.edu.cn
3 Southern Marine Science and Engineering Guangdong Laboratory (Guangzhou), Guangzhou 511458, China
4 School of Computer Science, China University of Geosciences (CUG), Wuhan 430074, China; lzwang@cug.edu.cn
5 Aerospace Information Research Institute, CAS, Beijing 100094, China; zhangyuwei20@mails.ucas.ac.cn
6 School of Information Science and Engineering, Yanshan University, Qinhuangdao 066012, China; liuxiao123456@hotmail.com
* Correspondence: liupeng@radi.ac.cn

Abstract: Remote sensing building extraction is of great importance to many applications, such as urban planning and economic status assessment. Deep learning with deep network structures and back-propagation optimization can automatically learn features of targets in high-resolution remote sensing images. However, it is also obvious that the generalizability of deep networks is almost entirely dependent on the quality and quantity of the labels. Therefore, building extraction performances will be greatly affected if there is a large intra-class variation among samples of one class target. To solve the problem, a subdivision method for reducing intra-class differences is proposed to enhance semantic segmentation. We proposed that backgrounds and targets be separately generated by two orthogonal generative adversarial networks (O-GAN). The two O-GANs are connected by adding the new loss function to their discriminators. To better extract building features, drawing on the idea of fine-grained image classification, feature vectors for a target are obtained through an intermediate convolution layer of O-GAN with selective convolutional descriptor aggregation (SCDA). Subsequently, feature vectors are clustered into new, different subdivisions to train semantic segmentation networks. In the prediction stages, the subdivisions will be merged into one class. Experiments were conducted with remote sensing images of the Tibet area, where there are both tall buildings and herdsmen's tents. The results indicate that, compared with direct semantic segmentation, the proposed subdivision method can make an improvement on accuracy of about 4%. Besides, statistics and visualizing building features validated the rationality of features and subdivisions.

Keywords: building extraction; GF-2; orthogonal generative adversarial networks; subdivision

check for updates

Citation: Sun, S.; Mu, L.; Wang, L.; Liu, P.; Liu, X.; Zhang, Y. Semantic Segmentation for Buildings of Large Intra-Class Variation in Remote Sensing Images with O-GAN. *Remote Sens.* **2021**, *13*, 475. https://doi.org/10.3390/rs13030475

Received: 6 December 2020
Accepted: 29 December 2020
Published: 29 January 2021

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

Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Building, a major artificial pastime closely related to humans, is an important symbol of human activities. Building information can further be used to research human activities, land use change, population estimation and prediction, disaster assessment, etc. Building extraction from remote sensing images has become an important way of acquiring information. Conventional building extraction methods mainly include traditional machine learning methods, such as the Normalized Difference Built-up Index (NDBI) [1], Morphological Shadow Index (MSI) [2], and Adaboost [3].

With the advent of high-resolution remote sensing images, a wealth of detailed information is provided, which makes it possible to acquire clear images of small buildings such as residences and temporary houses. Simultaneously, the complex texture and fragility of features is also revealed by very high resolutions, so it is often difficult to model building

features accurately in high-resolution remote sensing images through handcraft features or traditional shallow classifiers.

Today, deep learning has become a very successful method in many application fields. With a deep network structure and back-propagation optimization scheme, the classifier can automatically learn different levels of features and present powerful generalization abilities. Deep learning was hitherto widely used in the computer version of tasks such as image denoising [4], scene classification [5,6], target detection [7], and semantic segmentation [8]. With the development of remote sensing big data [9], deep learning is playing more and more important roles in the information extraction of remote sensing, such as hyperspectral image classification [6,10] and building extraction [11].

Because of the remarkable performance of feature learning, some architectures such as FCN [12] and a capsule network [13] were already widely used in the building extraction of remote sensing images. In current research, to extract the structural features of targets, at least two improvements were made on deep learning networks. One was to improve segmentation effects through enhancing edge information. For example, a signed distance function of building boundaries is introduced as output representation to enable the network to learn hierarchical features for segmenting individual objects [14]. Accordingly, we propose a class-boundary network to detect a set of scale-dependent class boundaries and fuse the result into the original image in the final semantic segmentation network [15]. Building outlines were added to the network as additional labels, and it proposed a new loss function considering building edges [16]. The masked-RCNN model, which adds post-processing after segmentation, implicitly regularizes noisy building boundaries in an iterative manner [17]. RiFCN, to achieve accurate boundary inference and semantic segmentation, a novel bidirectional network is proposed, by using a series of autoregressive recurrent connections so that boundary-aware feature maps and high-level features are orderly embedded [18]. The other is to enhance the versatility of the method through data enhancement or the introduction of external data. In [19], the author converts conventional binary labels to multi-value maps through an object mask network(OMN) to enhance robustness to inaccurate object candidates. A Siamese U-Net [20], which comprises two shared weights of a U-Net using original images, was also extended as a generalization of semantic segmentation. An original label was converted into a distance map, and an uncertainty weighted multi-task loss was introduced to improve semantic segmentation predictions of deep neural networks [21]. Some other studies introduced street view images [22], index information (e.g., IR, NDVI, and PCA) [11], or space-borne SAR tomography data as external information [23].

In most current research (e.g., [21,24]) and open data sets (e.g., UCMerced LandUse and WHU-RS19), buildings have small intra-class variations, and boundaries are accurately labeled so their research is to extract more accurate building boundaries. Conversely, in our study area, buildings have large intra-class variations and sometimes are coarsely labeled. Consequently, the purpose of this paper is to reduce the impact of intra-class variation. Intra-class difference is a common problem in the field of computer vision, and the usual solution is to subdivide with traditional computer vision techniques such as spectrum, texture, and shape [25–27].

In recent studies, different divide-and-conquer strategies are implemented on labels using the deep learning method, such as directly annotated targets or scenes in more diverse categories according to the backgrounds of land features [28]. OpenStreetMap data is used as a coarse approximation from the ground truth to produce a coarse to fine solution for the semantic labeling of satellite images [29] and correcting rural building annotations in OpenStreetMap through the alignment of the original annotations, removal of incorrect annotations, and addition of new annotations of buildings that appear for the first time in the updated imagery [30]. In this paper, we choose subdivision, which requires an effective image feature extraction method since a convolutional neural network can produce image representations that capture hierarchical patterns and attain global theoretical receptive fields. A deep learning network itself can provide plenty of incomprehensible

image features through the high-dimensional semantic representation of intermediate layer output. Previously, many studies have investigated intermediate layer interpretation or analysis [31–33].

In addition, in the field of deep learning, fine-grained image classification is a conventional way to resolve intra-class variance problems through features obtained from deep learning networks. Therefore, an important step of these methods is to integrate the features of the target from the deep network, which is similar to obtaining low-dimensional feature embedding with an anchor point using deep metric learning [34]. For example, a smooth network with a channel attention block and global average pooling is proposed to select more discriminative features [35] since a convolutional layer can express a neighborhood spatial connectivity pattern and it can be flexibly stacked into various structures. Owing to the concise and extensible structure, convolutional neural networks can easily implement feature extraction or analysis and outperform many traditional computer vision approaches marginally [36].

However, for deep architecture, the labeled samples may be often so insufficient that the advantages of deep learning [37] are decreased if we subdivide them further. A generative network meets such needs since it models real data distribution through training data. Currently, generative networks are mainly divided into two categories: The first is an auto-encoder (AE), which uses pointwise loss as an explicit measurement method. An AE can encode images well, but the generated image is usually fuzzy. The second is generative adversarial networks (GAN) [38], an implicit adaptive model that measures the distribution distance of different datasets through confrontation training. It generates high-quality data, and its intermediate results can be used for feature extraction. In addition, a large class of methods combined VAE with GAN, such as VAE-GAN [39], BIGAN [40], and O-GAN [41]. They can generate high-quality images in various scenarios.

For the building extraction problem, the complexity of backgrounds and large intra-class variance for targets coexist. Only one GAN cannot accurately model the relationship between backgrounds and targets. In this paper, inspired by the idea of O-GAN, we proposed a new architecture that consists of two connected GANs. The two GANs are separately responsible for generating backgrounds and buildings, and they exchange information by introducing a correlation loss function into their discriminator. Correspondingly, the building samples with large intra-class variance are augmented. Simultaneously, the differences between background and building can be well maintained, even enhanced, for the generated samples. For the features generated by O-GAN, we used the idea of fine-grained image classification, such as SCDA [42], to extract subtle differences in key parts of buildings through neural networks' hidden layers of discriminators. Finally, the proposed frame is applied to some conventional semantic segmentation architectures to validate its effectiveness. We will address the proposed method in detail in the following sections.

2. Method

In this paper, we mainly focus on the problem of too large intra-class variations in building semantic segmentation. As shown in Figure 1, the proposed method included four parts: There are two GANs, one for the background, and the other, for the building target. First, the two GANs are alternative trained, and their information is exchanged by adding the correlation terms into their loss function. Second, by using the SCDA method, we select an intermediate layer in the target GAN to form a convolutional activation tensor. Each channel of the intermediate layer is analyzed to screen out specific channels with large feedback values and large differences between the feedback value for the building and the background. Third, the feature vectors of building targets are clustered into different subdivisions, and then we give subdivisions new labels to form sub-classes. Finally, we train the semantic segmentation network using the samples with sub-class labels. In the testing stage, the prediction results are merged into building targets based on the subdivided labels. In the following subsections, each parts of the method will be addressed in detail.

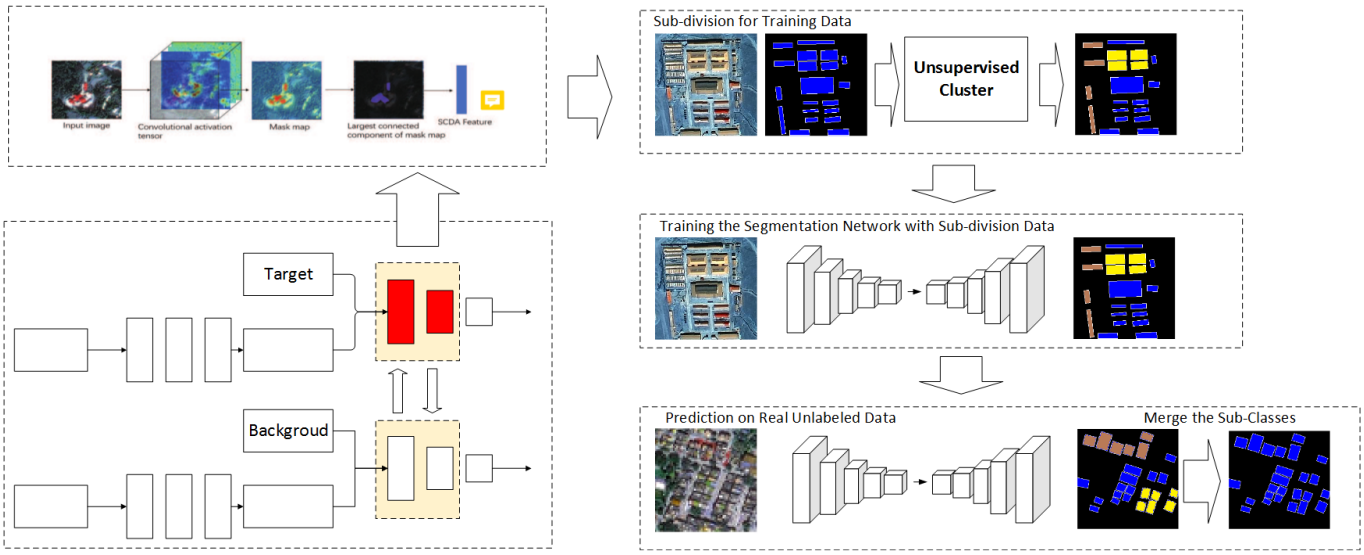


Figure 1. Flow chart of the proposed method.

2.1. O-GAN Encoding for Target and Background Data

In building extraction for remote sensing images, there are two obvious difficulties: the complexity of the background and the diversity of the building target. Furthermore, in most study areas, image labels are insufficient, especially if we categorized them into subdivisions. It is necessary to generate more samples identically distributed as the statistical characteristics of both the background and building target dataset. With the generated datasets, a fine-grained classification method can be applied to analyzing intermediate layers of a deep learning network, so we can select the appropriate features to model the complexity of the background and the diversity of the building target.

In this paper, we promote GAN to generate the samples for semantic segmentation. We did not adopt VAE [41], which is also an effective method for generating samples, because the image quality generated by VAE often is not very high, while GAN can generate high-quality simulated images. We proposed that the features in the discriminator of GAN can be used as the input of a sub-division before the training of segmentation network. However, conventional GAN is not good at separately representing the relationship between complex backgrounds and diverse building targets. Furthermore, the discriminator of conventional GAN is at risk of degradation, so it is inappropriate to directly analyze its intermediate layer features for subdivision.

Therefore, we proposed a new architecture with two GANs, which are connected and can exchange information with their introduction into an orthogonality loss function (O-GAN). By making modifications to the loss function, the two connected GANs can well represent the complex background and diverse building target. Meanwhile, the two discriminators as coders will supervise each other so there is no degradation during training.

Before addressing O-GAN, we first review the object function of conventional GAN. Without a loss of generality, the loss function of conventional GAN is denoted as

$$\begin{aligned} D &= \arg\min_{D} E_{x\sim p(x), z\sim q(z)}[f(D(x) + g(D(G(z)))] \\ G &= \arg\min_{G} E_{z\sim q(z)}[h(D(G(z)))] \end{aligned} \tag{1}$$

where $x \in \mathbb{R}^{n_x}$ is the real data and $z \in \mathbb{R}^{n_z}$ is the noise data. Function $p(x)$ is the distribution of real data, and $q(z)$ is the distribution of noise data. (In this paper, we set $q(z)$ as Gaussian distribution with a zero mean and variance 1). $G(z) : \mathbb{R}^{n_z} \to \mathbb{R}^{n_x}$, which

is the generator, and $D(x) : \mathbb{R}^{n_x} \rightarrow \mathbb{R}$, which is the discriminator. If we let $f = log(x)$ and $g = h = log(1 - x)$, there are

$$\begin{aligned} D &= \arg\max_{D} E_{x\sim p(x)}[log(D(x))] + E_{z\sim q(z)}[log(1 - D(G(z)))] \\ G &= \arg\min_{G} E_{z\sim q(z)}[log(1 - D(G(z)))] \end{aligned} \tag{2}$$

It is also denoted as

$$\arg\min_{G}\arg\max_{D} E_{x\sim p(x)}[log(D(x))] + E_{z\sim q(z)}[log(1 - D(G(z)))] \tag{3}$$

From the perspective of network structure functioning, a discriminator is very similar to an encoder, and a discriminator can be regarded as the decomposition of two parts, as in Equation (4).

$$D(x) \triangleq T(E(x)) \tag{4}$$

where $T(\cdot)$ function is a relatively simple function, and $E(\cdot)$ function is an encoder. A good decoder $E(\cdot)$ means that the output of $E(x)$ will be as similar as possible to the corresponding latent variable z. In [41], the author proved that, when measuring the correlation of decoding in GAN, we can discard $T(\cdot)$ and use $D(x)$ directly as $E(x)$. It was guaranteed that the GAN model has the function of an encoder and decoder at the same time. For the problem of building sample generation, we hope that the target and background would be as discriminative as possible. However, it is not easy to generate build images and background images at the same time with only one GAN. In this paper, we proposed to use two GANs: one for builds and the other for backgrounds. Inspired by the idea of O-GAN [41], the encoder part for targets should be unrelated to the encoder part of backgrounds. Therefore, we need to add a new term to the loss function of the two GANs, which can represent the relationship of encoded features between buildings and backgrounds. We also called the proposed two GANs O-GAN, although they are different from the structures as in reference [41]. The object functions of the proposed O-GAN are defined as

$$\arg\min_{G_1}\arg\max_{D_1} E_{x_1\sim p(x_1)}[log(D(x_1))] + E_{z_1\sim q(z_1)}[log(1 - D_1(G_1(z_1))) - \lambda\rho(z_2, D_1(G_1(z_1)))] \tag{5}$$

$$\arg\min_{G_2}\arg\max_{D_2} E_{x_2\sim p(x_2)}[log(D_2(x_2))] + E_{z_2\sim q(z_2)}[log(1 - D_2(G_2(z_2))) - \lambda\rho(z_1, D_2(G_2(z_2)))] \tag{6}$$

where $D(x_1)$ and $G_1(z_1)$ are the discriminator and generator for building images, $D_2(x_2)$ and $G_2(z_2)$ are the discriminator and generator for background images, and ρ is the correlation coefficient as

$$\rho(z, \hat{z}) = \frac{\sum_{i=1}^{n_z}(z_i - avg(z))(\hat{z}_i - avg(\hat{z}))/n_z}{std(z) \times std(\hat{z})} \tag{7}$$

where $avg(z)$ is the mean value and $std(z)$ is the standard deviation.

To verify the effectiveness of the proposed O-GAN in generating building and background samples, 1000 samples with more than 70% of building area are selected to train building target GANs, and 1000 background samples without building are selected to train the background GANs. The structure of the discriminator (encoder) of the target-GAN is shown in Table 1. The architecture of O-GAN is shown in Figure 2.

Table 1. Layers of discriminator (encoder) of building target generative adversarial network (GAN) in orthogonal GAN (O-GAN).

Layers of O-GANs	Output Shape
input_1	(None,256,256,3)
conv2d_1	**(None,128,128,64)**
leaky_re_lu_1	(None,128,128,64)
conv2d_2	**(None,64,64,128)**
batch_normalization_1	(None,64,64,128)
leaky_re_lu_2	(None,64,64,128)
conv2d_3	**(None,32,32,256)**
batch_normalization_2	(None,32,32,256)
leaky_re_lu_3	(None,32,32,256)
conv2d_4	(None,16,16,512)
batch_normalization_3	(None,16,16,512)
leaky_re_lu_4	(None,16,16,512)
conv2d_5	(None,8,8,1024)
batch_normalization_4	(None,8,8,1024)
leaky_re_lu_5	(None,8,8,1024)
conv2d_6	(None,4,4,2048)
batch_normalization_5	(None,4,4,2048)
leaky_re_lu_6	(None,4,4,2048)
flatten_1	(None,32768)
dense_1	(None,256)

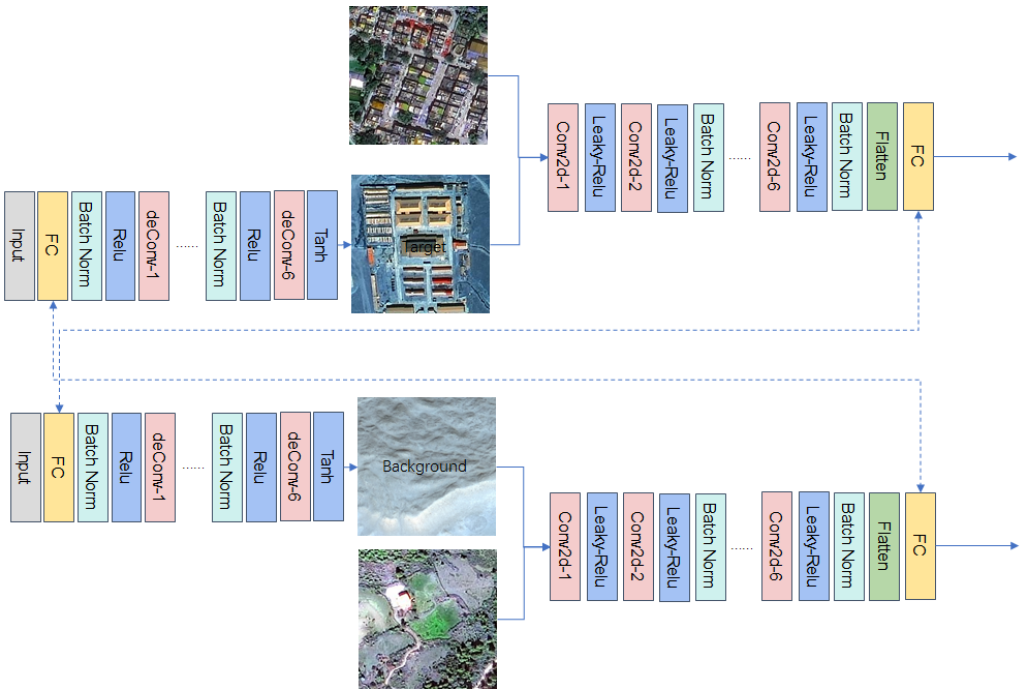


Figure 2. The Architecture of O-GAN.

After the alternative training of the O-GAN, the two types of generated samples will be encoded by $D_1(G_1(z))$ and $D_2(G_2(z))$. As in Table 1, the feedback values of three intermediate layers (conv2d-1, conv2d-2, and conv2d-3) are selected to validate the relationship between targets and buildings. The Structural Similarity Index(SSIM), a commonly used image similarity calculation method, is used to measure the similarity of the outputs of the two GANs. The structural features of the image signal can present a similarity in more specific situations [43]. The SSIM values between different buildings and SSIM values between buildings and backgrounds are calculated and shown in Figure 3.

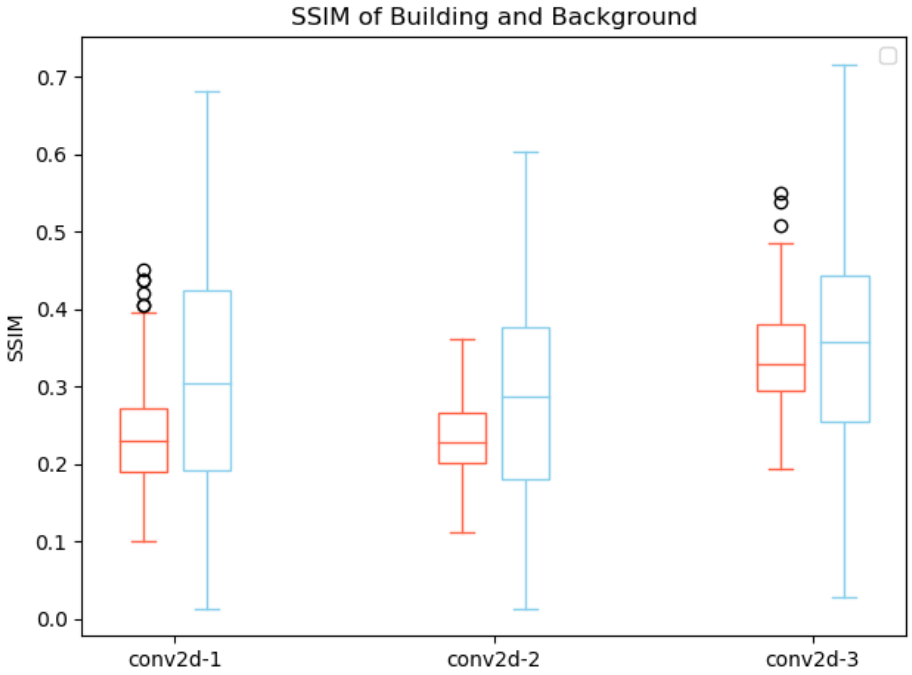


Figure 3. Structural Similarity Index (SSIM) result of interior buildings (red) and building backgrounds (blue).

As in Figure 3, the feedback value SSIMs of three intermediate layers of O-GAN indicate that the similarity between building and building is obviously larger than between buildings and the background. This phenomenon is more obvious on layer conv2d-1 and conv2d-2. Except for quantity comparison, we also validate the effectiveness of O-GAN by visualizing feature layers. Figure 4 visualizes the representative channels in intermediate layers on the building GAN. For each layer, there is a representative channel with higher feedback in the building area and lower feedback in the background area. We can observe that layer conv2d-1 and conv2d-2 show larger feedback values. Therefore, in the next section, layer conv2d-2 is selected for feature analysis and subdivision.

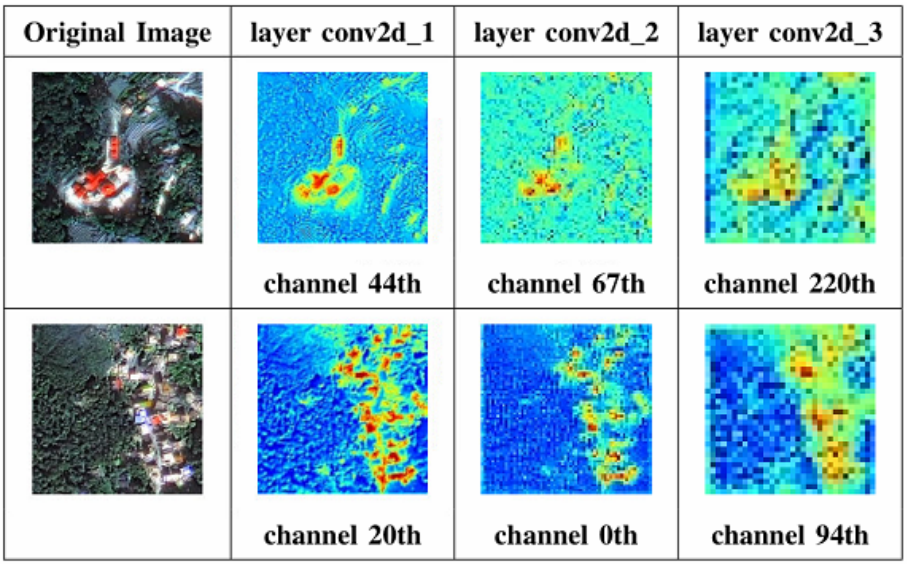


Figure 4. Response of convolutional layers of different shapes on two training samples.

2.2. Feature Extraction and Sub-Division

Not all features of building in the discriminator are suitable for subdivision, although they are generated by building GAN. The information exchange between building GAN and backgound GAN can only tell the two types of features apart but not produce subdivision for buildings. In this paper, for subdivision, building features in the decoder are extracted through SCDA [42]. It provides an efficient feature extraction method through analyzing the intermediate outputs of a deep neural network.

The original SCDA directly locates targets through extracting the relatively high feedback value of certain channels on one or several intermediate layers of a trained deep network. There are usually three steps: first, acquire the convolutional activation tensor through the feedback value of convolutional layers; second, take the largest connected component of each channel as a descriptor; finally, obtain the SCDA feature through aggregating these descriptors.

In our research, the n-dimensional building features reflecting different locations and scales are obtained for the result. With the help of labels in the training dataset, the building target area of the intermediate layer output can be accurately determined. For this scenario, on one hand, appropriate channels can be selected according to building feedback values. On the other hand, labels enable building features to be compressed into a one-dimensional feature in the spatial domain. In addition, the original SCDA did not specify which architecture of neural network they used. In this paper, the features of a discriminator in O-GAN are selected as the input of the SCDA scheme. The feature integration process is shown in Figure 5.

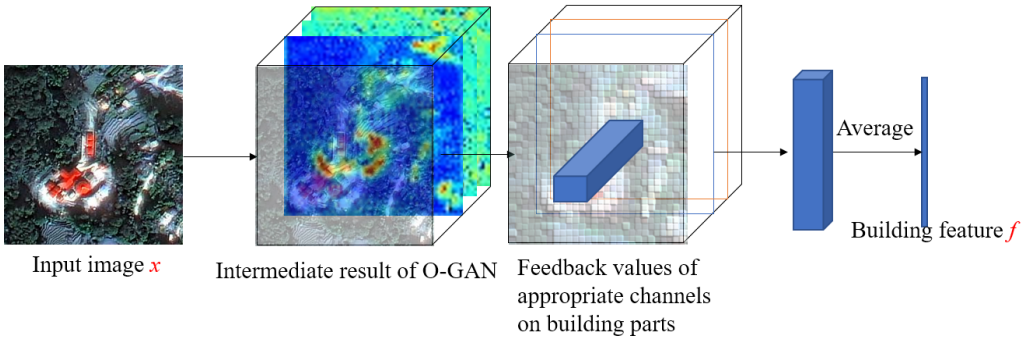


Figure 5. Building feature extraction process.

Obviously, not all features need to join SCDA feature integration. We need to select a number of significant channels to perform SCDA. As mentioned in the latest section, layer conv2d-2 usually activates more building features. A visualized feedback map of typical feature channels of layer conv2d-2 is shown in Figure 6. For the two sample images, the 0th channel is suitable for the feature layer of the upper image, the 67th layer is suitable for the lower one, and the 112th layer does not perform well on both of the two images. The feature maps indicate that, for each type of building, there is a corresponding channel with a high value on building areas. Meanwhile, some channels are not suitable for extracting building features.

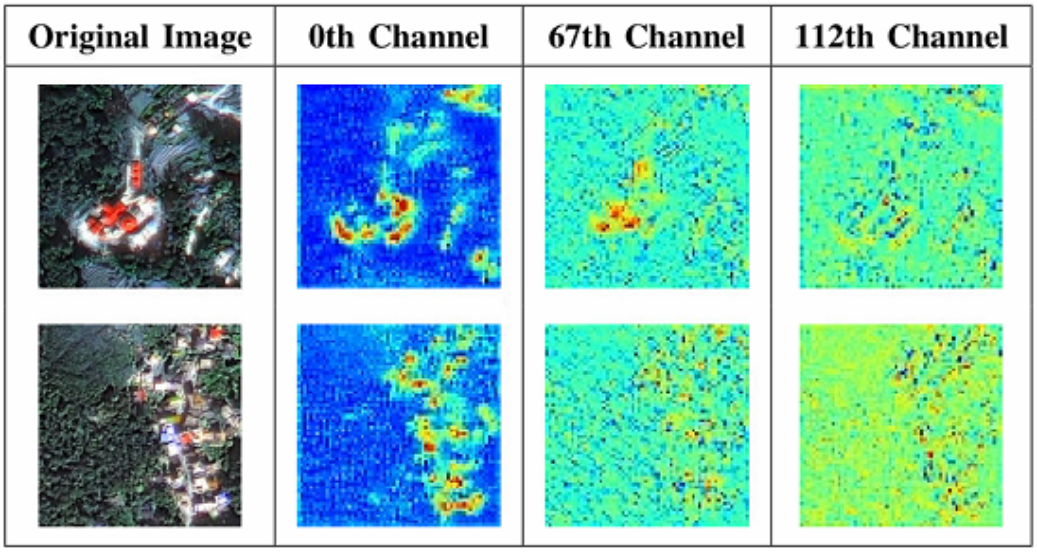


Figure 6. Values in different convolutional channels of building images.

For many building samples, there are still small areas of background. We also need to avoid their influences in feature selection. To select appropriate layers for SCDA integration, feedback maps of each channel are divided into building area $M \in Building$ and non-building $M \notin Building$ areas with the help of label information. With average feedback values within the building area $avg(M \in Building) > T_{avg}$ and the difference of feedback values between building and non-building areas $WD(hist(M \in Building), hist(M \notin Building)) > T_{WD}$, the layers will be selected as candidates for SCDA, where WD is the Wasserstein Distance [44], $hist$ is the value distribution, and T_{avg} and T_{WD} are thresholds. In this paper, take 0.6 and 0.006, respectively.

Figure 7 is the frequency statistic of 5000 training samples, where the x axis is the channel's serial number and the y axis counts the number of samples that meet the criteria. The results indicate that about 20 channels can fit more than 10% of the samples. Moreover, due to the complexity of buildings in the study area, there is no single channel suitable for all samples. To reduce parameters and avoid a dimension disaster, 20 high-frequency channels are selected as feature channels.

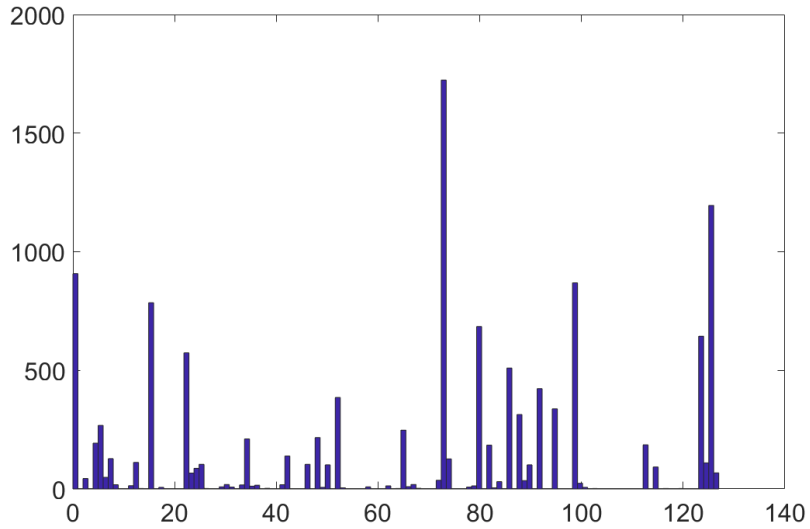


Figure 7. Statistics of samples that meet the criteria in each channel.

Figure 8 exhibits the average feedback value of the building mask ($avg(M \in Building)$) of each sample in channels 70(x axis) and 0(y axis), which are the two most representative

channels of building features. To exhibit the effects of subdivision, points are marked in four colors corresponding to four categories. We can observe that intra-class variations are obvious; however, features become separable after feature integration. After feature selection and SCDA for all samples (both original and generated), they are projected onto a new feature space with relatively low dimensionality. In this situation, the SCDA building feature can be clustered into several new subdivisions. In this paper, we use k-means to produce a new sub-class and give each sub-class a new label. The trainings of building extraction are based on the new sub-classes with new labels, but not old labels. In the next section, we will introduce the subdivided result into conventional segmentation networks.

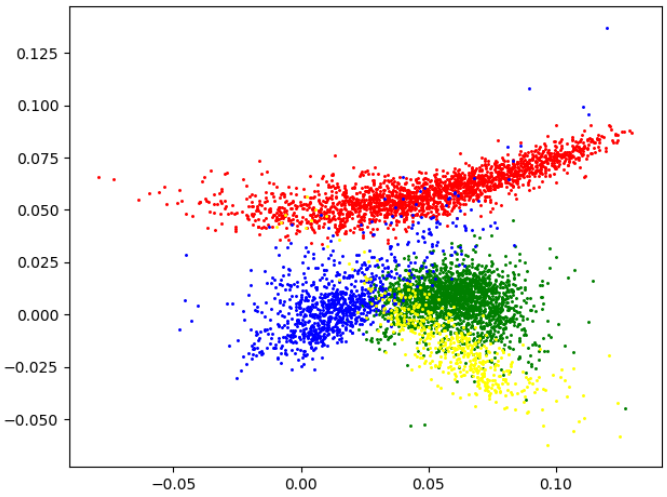


Figure 8. Average feedback value on building mask of each sample on channel 70th (x-axis) and 0th (y-axis).

2.3. Introducing the Subdivided Schemes into Semantic Segmentation

Conventional semantic segmentation has been successfully applied to plenty remote sensing applications, including building extraction. However, in our study area, buildings have large intra-class variations and are often imperfectly labeled without fine-grained class labels. In this situation, the accuracy of the conventional segmentation model would decrease or even cause non-convergence. Therefore, in the previous two sections, we generated more samples with two orthogonal GANs to solve the problem of intra-class variations and an insufficient sample. At the same time, the features of the building and the background are well separated by O-GANs, and we integrate features with SCDA for building and cluster them into different subdivisions.

In this section, we train the semantic segmentation network with new subdivision labels for building extraction. We selected three types of semantic segmentation models, specifically as FCN, Unet, and Deeplab-V3. Among them, the structures of FCN and Unet are similar; both consist of a network with convolution, pooling, and deconvolution layers. The core feature of FCN is a skip architecture procedure and upsampling pooling layers to restore them to the size of input data.

In Unet, the semantic segmentation result is extracted directly through two convolutional processes after the last copy-and-crop operation. When facing large intra-class variation, the result of Unet may be unstable. Some specifics like small buildings will not be extracted. For Deeplab-V3, the most distinctive feature is the Atrous convolution, which allows us to effectively enlarge the field of view to incorporate multi-scale information. However, the edge of the segmentation result of Deeplab-V3 may be vague due to the enlarged field of view. In this paper, to accurately extract building areas, these semantic segmentation networks are trained on subdivision labels generated by the proposed method. After performing an initial building prediction, the subcategories are merged to obtain final results.

In summary, we proposed a building feature extraction and subdivision method using a discriminator of O-GANs. With the value feedback of the O-GAN discriminator, the feature extraction process was analyzed, and the subdivision schemes were developed statistically. Finally, we introduced the subdivision scheme into different semantic segmentation algorithms. The algorithmic steps of the proposed building extraction method are summarized as Algorithm 1.

Algorithm 1 Sematic Segmentation with O-GAN

Input: Original Image:x

1: Training the two orthogonal GANs with building samples and background samples.

2: Generating a feedback value map of a building with the discriminator of building GANs.

3: Extracting the feedback values of 20 channels according to the threshold T_{avg} and T_{WD}.

4: Averaging the spatial dimension of SCDA and obtaining the final building feature.

5: Unsupervised clustering of the final building feature and obtaining subdivided labels.

6: Training the semantic segmentation network on a subdivided dataset from step 5.

7: Processing semantic segmentation on a predication dataset.

8: Merging subcategories to obtain final results.

3. Study Area and Experimental Data

The study area is on the southern Qinghai-Tibet plateau, including China, India, Bhutan, and Nepal, as shown in Figure 9. Influenced by southwest monsoons in the Indian Ocean, the southern Qinghai-Tibet plateau belongs to a subtropical and tropical monsoon climate that has high rainfall and is warm and humid all year round, with excellent hydrothermal conditions. This area is the best natural area on the Qinghai-Tibet plateau and contains one of the most densely populated and economically developed areas, virgin forests, and border guard posts.



Figure 9. Research area of this paper.

Due to the influence of climate, topography, economy, and culture, the style, morphology, area, distribution, and surrounding environments of buildings are quite different. As the data obtained by satellite sensors, these differences are also reflected in remote sensing images. Buildings in the study area are as follows: From the view of the surrounding environment, the backgrounds include wasteland, woodland, near farms, and towns. From the view of morphology and distribution, some are in a wasteland, separately or sporadically distributed, some of them are in a village with sparse distribution, some of them are in new urban area where the buildings are dense and neat, and some of them are in old urban areas where the buildings are denser and messy. In addition, buildings have various styles and areas such as large, small, and regular with clear outlines or a fuzzy boundary, as is shown in Figure 10. In this paper, we separately carried out experiments on different categories with different styles.

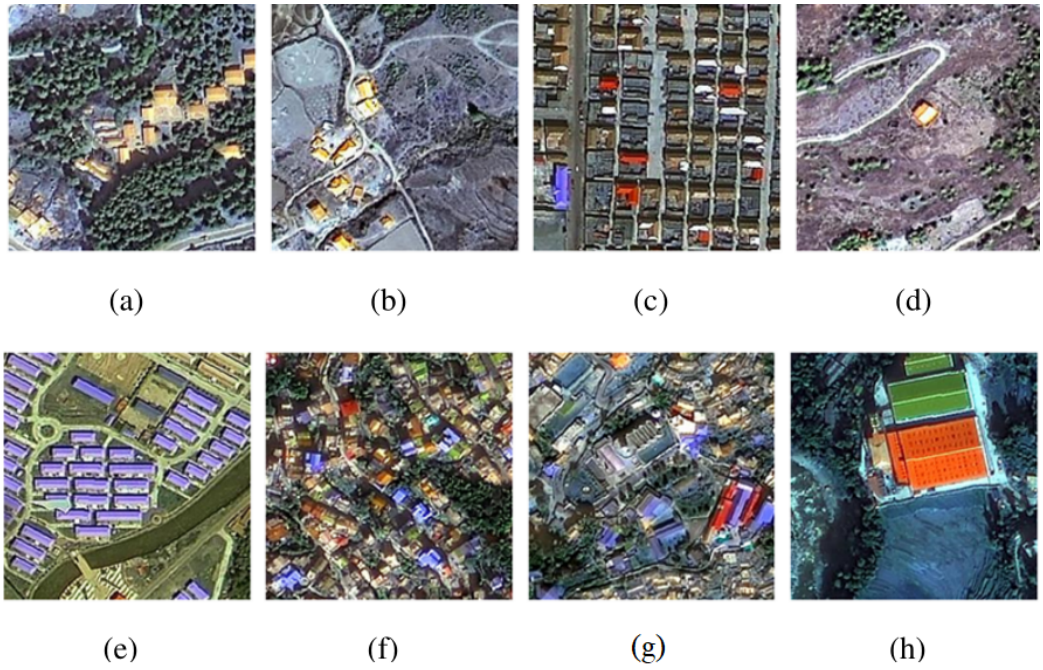


Figure 10. Variety of building types in the study area: (**a**) buildings in a forest, (**b**) buildings in wasteland, (**c**) buildings in an urban area, (**d**) buildings separately distributed, (**e**) buildings dense and neatly organized, (**f**) buildings dense and messy, (**g**) small buildings with a fuzzy boundary, and (**h**) large buildings with a clear boundary. We suggest that they should be further subdivided because of their complex composition.

Experimental data in this paper is GF-2 imagery [45], with a resolution of 0.8 m in panchromatic and 3.2 m in multi-spectrum. These remote sensing images were all obtained from the China Centre for Resources Satellite Data and Application (CRESDA) [46]. In the experiments, a visible light data product with a resolution of 0.8 m is obtained after orthorectification [47] and image fusion. We labeled approximately 8000 samples and generated 8000 samples using O-GANs as the dataset for training and validation.

4. Experimental Results

In this paper, the O-GAN model is implemented on Keras (Version 2.2.4) and an NVIDIA RTX 2080Ti GPU platform. During training, the Root Mean Square Prop (RMSprop) optimization method was used. The initial learning rate was set to 0.0001, and then in the fine tuning, the learning rate was set as 0.00001. While the parameter ϵ was set to 0.000001, and the batch size was set to eight for images with a size of 256 $\times$ 256 and a total of 100 epochs were trained.

All semantic segmentation implementations are based on Tensorflow (Version 1.14) and an NVIDIA RTX 2080Ti GPU. In the procedure, the Adaptive Moment Estimation (Adam) optimization method was used. The original learning rate was set to 0.00001, and the decay was set to 0.000001, while β_1 and β_2 were correspondingly 0.9 and 0.999. The batch size was set to eight for images with a size of 256 × 256, and a total of 50 epochs were trained. All models used pretrained weights. Specifically, we used Deeplab with the weight of ResNet50, Unet with the weight of the VGG16, both of them are from the model library of Tensorflow and FCN with weights of the ADE20K from MIT.

4.1. Training Result of O-GANs

As can be seen from Figure 11, the image generated by O-GAN is very similar to the real image. Therefore, in addition to extracting building features, complementary building images are also generated by O-GAN to alleviate the problem of the reduced sample size of each sub-category after subdivision.

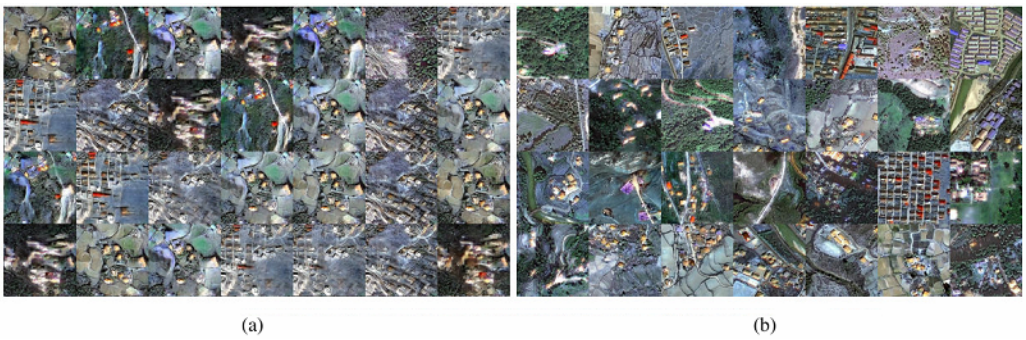


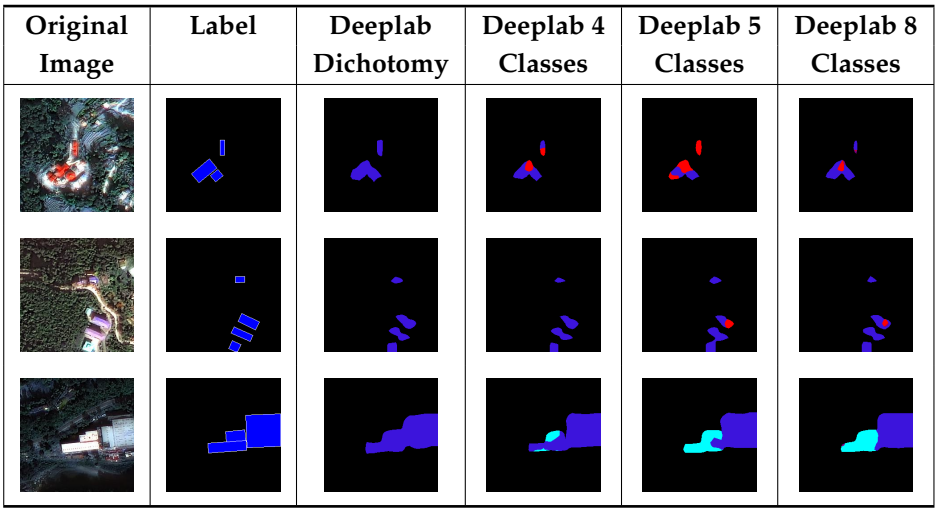
Figure 11. Fake images generated by O-GAN (**a**) and real images in the study area (**b**). O-GAN can generate simulated images similar to real images.

4.2. Building Extraction Results and Accuracy

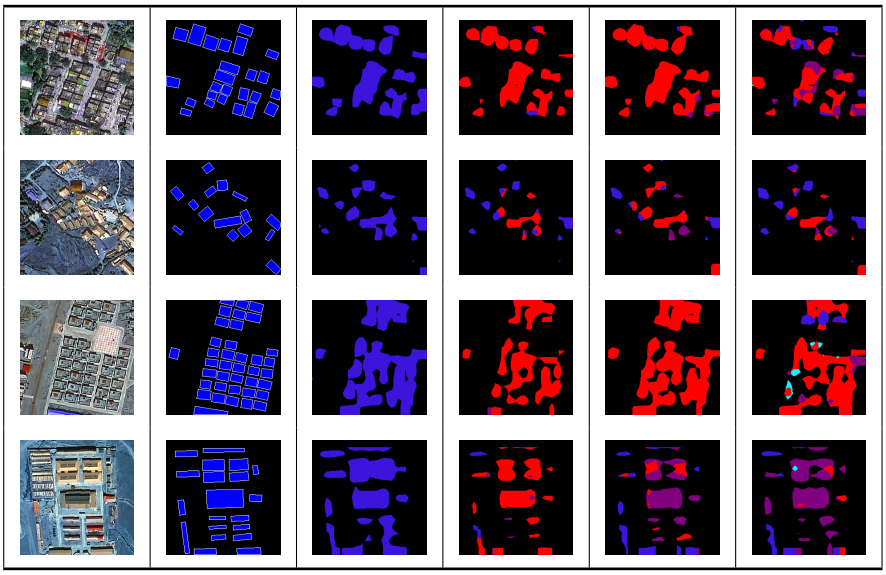
In this section, building samples are subdivided into four, five, and eight categories though the proposed method for subsequent training and prediction. The mean intersection over union (MIOU), a widely used accuracy metric in the field of semantic segmentation, is adopted to measure results. Subcategories are merged when calculating the MIOU of the final results. Due to the complexity of experimental data, other than evaluating all types of scenes, three types of samples (forest land, dense buildings, farm or waste land) were selected to test all methods and show the subdivision results. In addition, the complex environment and imperfect marking would inevitably reduce the overall MIOU to make the accuracy of the evaluation more comprehensive. The results of some typical local regions and corresponding accuracy indexes are also exhibited.

In terms of the control group, the results of the dichotomy ofeach semantic segmentation method are retained. Deeplab-V3, FCN, and Unet are used to validate the effectiveness of the proposed subdivision method by O-GAN. Besides, BRNet [16], a cutting-edge, open-source method, is also compared with the proposed method. It uses building outlines as an auxiliary label and performs well on conventional building dataset such as Aerial Imagery for Roof Segmentation(AIRS).

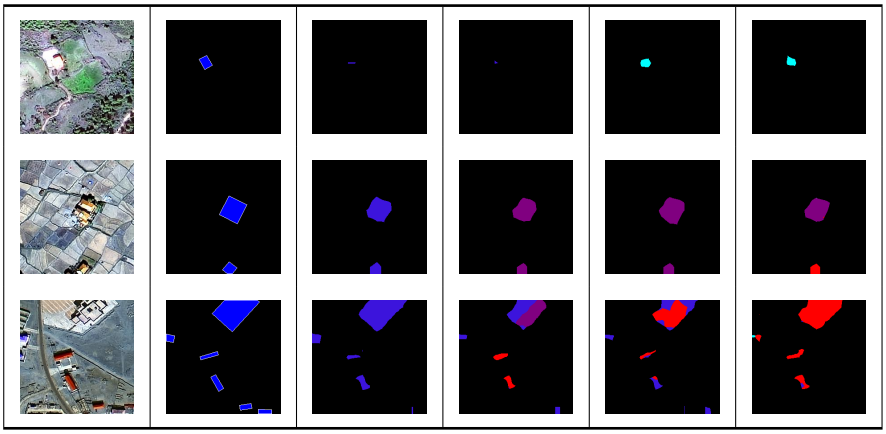
Figure 12a–c are the results of Deeplab-V3 for three type of buildings. Since Deeplab-V3 uses the Atrous convolution to obtain large-scale information, it is easily affected by local conditions. Consequently, the accuracy of the dichotomy method is lower than that of FCN and Unet, as in Table 2. We can observe that, when the number of subdivision categories is four or five, it can identify more building targets, and the overall accuracy is also improved. When the number of categories reaches eight, the subdivisions would mislead the classifier so the accuracy is often higher than the dichotomy but lower than that of four or five categories.



(**a**)



(**b**)



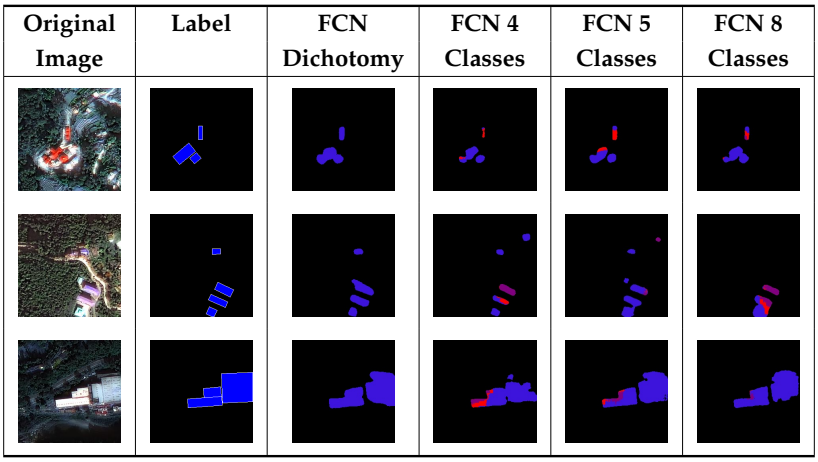
(**c**)

Figure 12. (**a**–**c**) The segmentation results of Deeplab-V3.

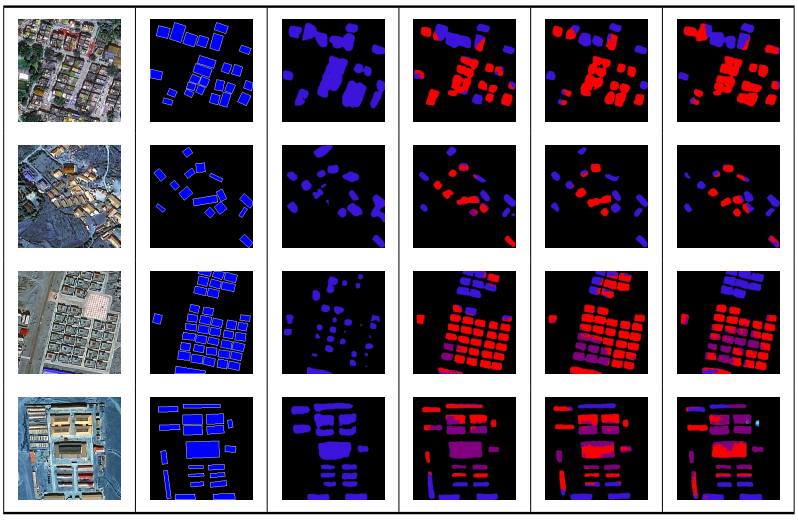
Table 2. Detailed mean intersection over union (MIOU) Result of Each Method.

MIOU	Overall	Forest	Dense Buildings	Farm and Wasteland
Deeplab Dichotomy	67.8%	68.9%	66.7%	79.1%
Deeplab 4 categories	69.5%	69.9%	65.1%	81.3%
Deeplab 5 categories	71.3%	72.4%	63.6%	82.3%
Deeplab 8 categories	68.6%	69.0%	57.0%	78.4%
FCN Dichotomy	71.3%	66.2%	75.6%	82.3%
FCN 4 categories	74.3%	69.5%	79.1%	85.3%
FCN 5 categories	76.0%	72.6%	78.6%	86.3%
FCN 8 categories	73.5%	68.8%	78.0%	85.8%
Unet Dichotomy	68.2%	65.4%	70.3%	77.7%
Unet 4 categories	68.4%	64.1%	70.0%	78.3%
Unet 5 categories	72.0%	66.6%	73.6%	80.3%
Unet 8 categories	68.6%	65.8%	71.0%	79.9%
BRNet	66.3%	69.1%	60.2%	68.9%

Figure 13a–c are the results of FCN for three type of buildings. In this scenario, for almost all types of scenes, FCN is shown higher than Deeplab-V3 of Figure 12. We can observe that subdivision methods can help FCN to identify more building targets. This advantage is more obvious when buildings are grayish white, a farm or wasteland, or densely placed. When the number of categories reaches eight, the accuracy is decreased, but it is still higher than the dichotomy method.

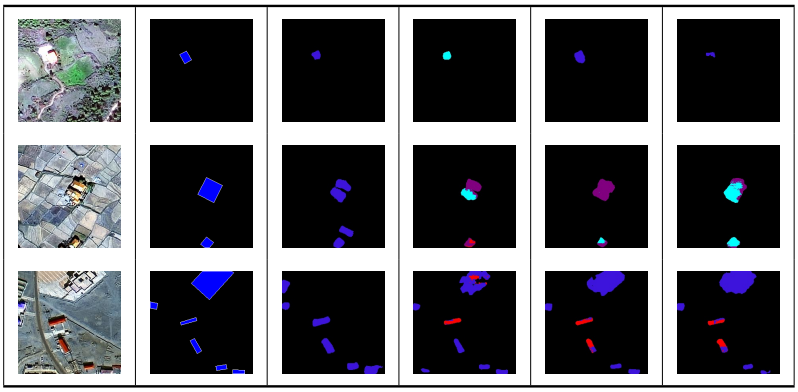


(**a**)



(**b**)

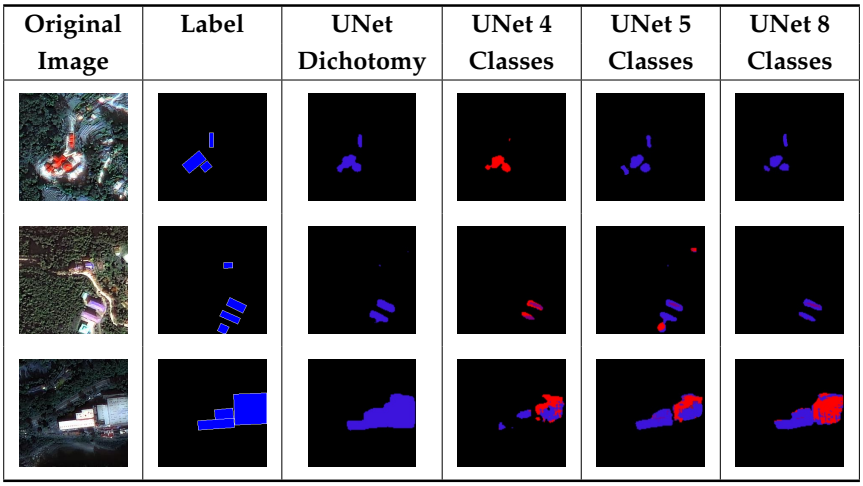
Figure 13. *Cont.*



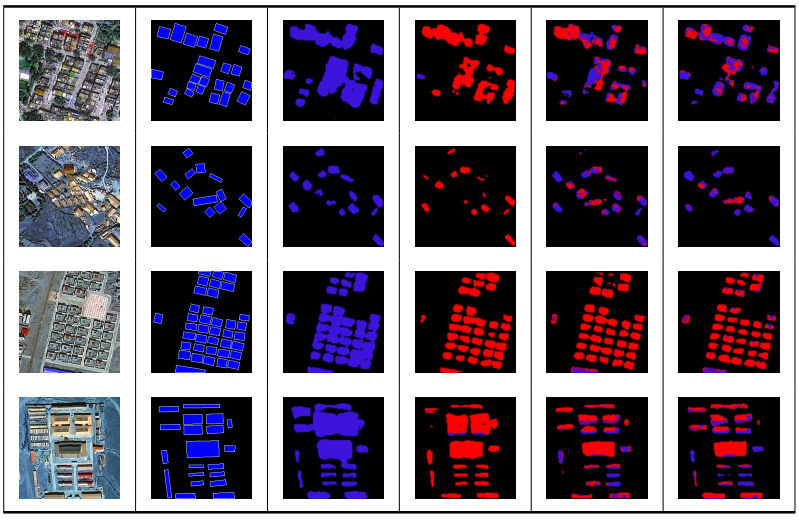
(**c**)

Figure 13. (**a**–**c**) The segmentation results of FCN.

Figure 14a–c are the results of Unet for three type of buildings. Since the structure of the Unet is similar to that of FCN, the effectiveness of subdivision for Unet is also similar to that of FCN. Relatively speaking, the impact of the number of sub-categories on Unet is smaller than that of FCN. We can observe that the proposed method benefits FCN, Unet, and Deeplab-V3.

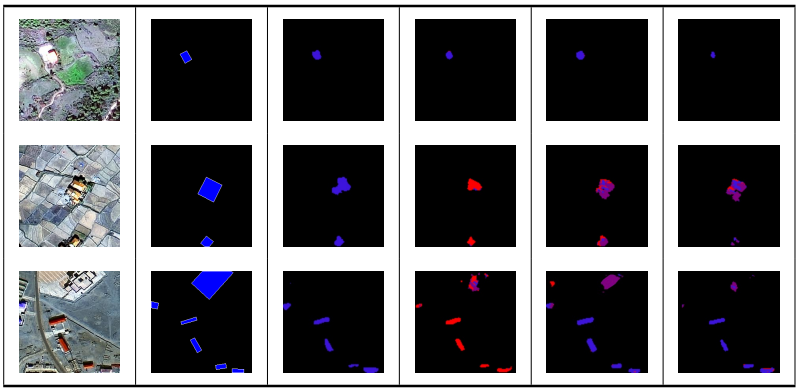


(**a**)



(**b**)

Figure 14. *Cont.*



(**c**)

Figure 14. (**a**–**c**) The segmentation results of Unet.

Figure 15 is the segmentation result of BRNet, since it used a building outline as auxiliary labels. In our imperfectly labeled dataset, it cannot accurately extract the building outline as in previous high-quality datasets. The MIOU of this BRNet method is slightly lower than other methods.

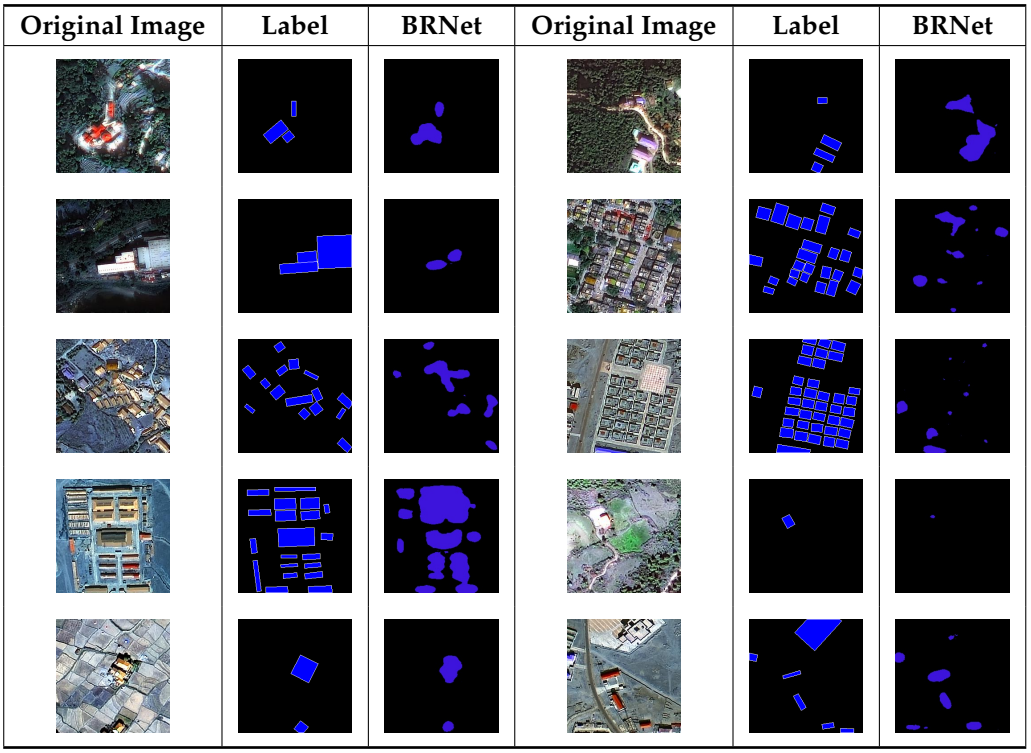


Figure 15. The segmentation results of BRNet.

Figure 16 is the quantitative results of overall MIOU for three type of scenes. For subdivision, it obviously improves the extraction accuracy of Deeplab and FCN, whereas the improvement of Unet is relatively small. Among them, buildings in the farm or wasteland display low intra-class variation so they are relatively easy to extract and obtain higher accuracy. The number of subdivision categories has a certain impact on accuracy. For our study, for most scenes, five sub-categories are often the best number for subdivision, based on the results in Figures 11–14 and the statistics in Figure 15. Moreover, over-classification occurs when the number of categories reaches eight, but it still improved the overall accuracy compared with no subdivisions.

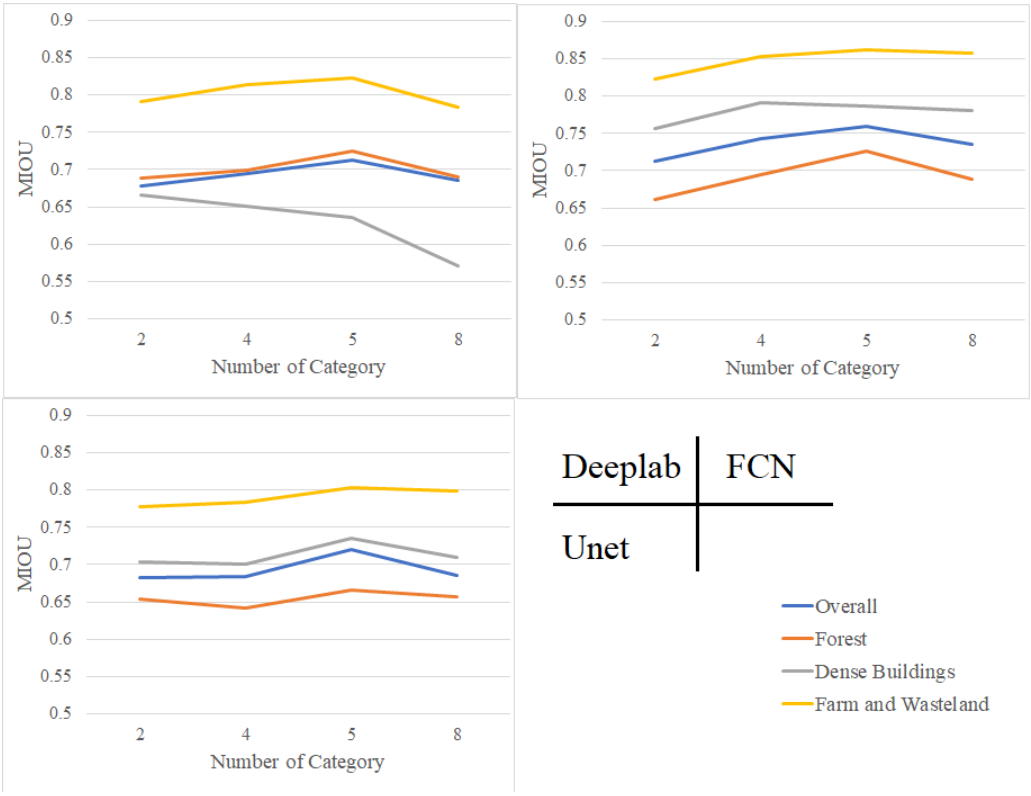


Figure 16. Mean Intersection over Union (MIOU) Result of Each Method. A proper segmentation improves MIOU results in many scenarios.

The results in Figures 12–15 indicated all methods can identify most buildings, but due to imperfect labels and the complexity of experimental data, they cannot accurately extract the building's outline. Furthermore, due to the complexity of backgrounds and targets and the interference of coarse labels, the overall accuracy of total areas could not be very high. Nevertheless, in some local areas, the method proposed in this paper can achieve high extraction accuracy. As in Figure 17, the result of FCN with five categories' subdivision is shown. In this area, the footprints of most buildings are accurately extracted. MIOU and the pixel accuracy of this area are 88.9% and 92.2%, respectively. Specific statistical results are shown in Table 2.

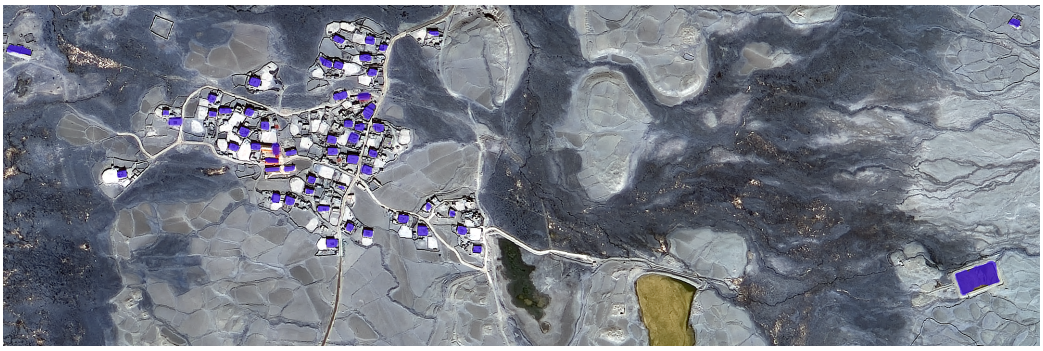


Figure 17. Result of FCN with five sub-categories.

5. Discussion

Since a convolutional neural network can construct informative features by aggregating spatial and channel-wise information within local receptive fields at each layer, the research on fine-grained image classification and explainable neural networks indicated that the information can be extracted. To improve the building extraction performance of conventional semantic segmentation, a subdivision method using intermediate layer

output as a feature is proposed to reduce the intra-class variation of a building. The core of the method is to analyze the intermediate output of O-GAN based on fine-grained image classification. Currently, the method shows good performance, whereas such a method still requires several annotated labels. There are also human factors in the process of data analysis. Therefore, in subsequent research, the attention mechanism would be involved, such as SeNet [48], a new architectural unit composed of a pooling and full connection layer, which aggregates feature maps across spatial dimensions. Such a simple structure can be easily embedded into the continuous convolution structure of O-GAN. In this way, since interdependencies between channels can be modeled explicitly through the direct output of SeNet, an automatic and objective layer analysis method can be realized. Moreover, SeNet can also be used as the part input of the next convolution layer, so that when going further than preprocessing, a feature learning network that can adaptively recalibrate can be proposed.

In the future, the proposed method can be applied to the building extraction with complex backgrounds and different forms. Large numbers of building images are hard to extract because their forms vary due to cultures, climates, terrains, and even local politics or economies. Usually, we only label them as buildings of one class; however, their intra-class variations seriously decrease the generalization of the training. The proposed method considered intra-class variations so we can partly avoid labelling too many fine-grained subclasses in a training dataset. In the future, the proposed method can also be extended to other datasets with large intra-class variations, such as car classification, boat classification, clothes classification, etc. In the case of complex background and intra-class variation, the advances of the proposed method will be manifested.

6. Conclusions

In this paper, to improve the performance of building extraction in buildings with high intra-class differences, subdivision preprocessing is proposed. Specifically, two GANs are connected by an orthogonal decomposition loss function and used to extract a convolutional activation tensor. By compressing the space domain and screening layers and channels, an N-dimensional building feature vector is obtained and clustered into subdivisions to enhance segmentation networks.

To better evaluate the performance of building extraction and the effectiveness of subdivision preprocessing, we implement four typical segmentation methods on three types of scenes. In addition, the process and results of building feature extraction are analyzed through statistics and visualization results. Experiments indicated that a proposed subdivision method with O-GAN leads to a visible improvement over such dichotomous work, which is especially obvious for extracting white buildings on wasteland. The statistical results also indicate that when the number of subdivisions is four or five the building extraction method will obtain higher accuracy in most cases in this study area.

Author Contributions: Conceptualization, L.W. and P.L.; Funding acquisition, L.W., L.M., P.L.; Methodology, S.S.; Resources, L.M.; Writing—review & editing, P.L., Y.Z. and X.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the Key-Area Research and Development Program of Guangdong Province under Grant 2020B1111020005 and in part by NSFC under Grant U1711266, Grant 61731022, Grant U2006210, Grant 41925007, and Grant 41971397.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from China resources satellite application center.

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

References

1. Zha, Y.; Gao, J.; Ni, S. Use of normalized difference built-up index in automatically mapping urban areas from TM imagery. *Int. J. Remote Sens.* **2003**, *24*, 583–594. [CrossRef]
2. Huang, X.; Zhang, L. Morphological building/shadow index for building extraction from high-resolution imagery over urban areas. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2011**, *5*, 161–172. [CrossRef]
3. Wei, Y.; Yao, W.; Wu, J.; Schmitt, M.; Stilla, U. Adaboost-based feature relevance assessment in fusing lidar and image data for classification of trees and vehicles in urban scenes. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* **2012**, *1*, 323–328. [CrossRef]
4. Liu, P.; Wang, M.; Wang, L.; Han, W. Remote-sensing image denoising with multi-sourced information. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2019**, *12*, 660–674. [CrossRef]
5. 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.
6. Liu, P.; Zhang, H.; Eom, K.B. Active deep learning for classification of hyperspectral images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2017**, *10*, 712–724. [CrossRef]
7. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. Ssd: Single shot multibox detector. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2016; pp. 21–37.
8. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*; Springer: Cham, Switzerland, 2015; pp. 234–241.
9. Liu, P.; Di, L.; Du, Q.; Wang, L. Remote sensing big data: Theory, methods and applications. *Remote Sens.* **2018**, *10*, 711. [CrossRef]
10. Huang, F.; Lu, J.; Tao, J.; Li, L.; Tan, X.; Liu, P. Research on Optimization Methods of ELM Classification Algorithm for Hyperspectral Remote Sensing Images. *IEEE Access* **2019**, *7*, 108070–108089. [CrossRef]
11. Xu, Y.; Wu, L.; Xie, Z.; Chen, Z. Building extraction in very high resolution remote sensing imagery using deep learning and guided filters. *Remote Sens.* **2018**, *10*, 144.
12. Maggiori, E.; Tarabalka, Y.; Charpiat, G.; Alliez, P. Convolutional neural networks for large-scale remote-sensing image classification. *IEEE Trans. Geosci. Remote Sens.* **2016**, *55*, 645–657. [CrossRef]
13. Paoletti, M.E.; Haut, J.M.; Fernandez-Beltran, R.; Plaza, J.; Plaza, A.; Li, J.; Pla, F. Capsule networks for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2018**, *57*, 2145–2160. [CrossRef]
14. Yuan, J. Learning building extraction in aerial scenes with convolutional networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 2793–2798. [CrossRef] [PubMed]
15. Marmanis, D.; Schindler, K.; Wegner, J.D.; Galliani, S.; Datcu, M.; Stilla, U. Classification with an edge: Improving semantic image segmentation with boundary detection. *ISPRS J. Photogramm. Remote Sens.* **2018**, *135*, 158–172. [CrossRef]
16. Wu, G.; Guo, Z.; Shao, X.; Shibasaki, R. Geoseg: A Computer Vision Package for Automatic Building Segmentation and Outline Extraction. In Proceedings of the IGARSS 2019—2019 IEEE International Geoscience and Remote Sensing Symposium, Yokohama, Japan, 28 July–2 August 2019; pp. 158–161.
17. Zhao, K.; Kang, J.; Jung, J.; Sohn, G. Building Extraction From Satellite Images Using Mask R-CNN With Building Boundary Regularization. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Salt Lake City, UT, USA, 18–22 June 2018; pp. 247–251.
18. Mou, L.; Zhu, X.X. RiFCN: Recurrent network in fully convolutional network for semantic segmentation of high resolution remote sensing images. *arXiv* **2018**, arXiv:1805.02091.
19. Hayder, Z.; He, X.; Salzmann, M. Boundary-aware instance segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 5696–5704.
20. Ji, S.; Wei, S.; Lu, M. Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set. *IEEE Trans. Geosci. Remote Sens.* **2018**, *57*, 574–586. [CrossRef]
21. Bischke, B.; Helber, P.; Folz, J.; Borth, D.; Dengel, A. Multi-task learning for segmentation of building footprints with deep neural networks. In Proceedings of the 2019 IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, 22–25 September 2019; pp. 1480–1484.
22. Kang, J.; Körner, M.; Wang, Y.; Taubenböck, H.; Zhu, X.X. Building instance classification using street view images. *ISPRS J. Photogramm. Remote Sens.* **2018**, *145*, 44–59. [CrossRef]
23. Shahzad, M.; Maurer, M.; Fraundorfer, F.; Wang, Y.; Zhu, X.X. Buildings detection in VHR SAR images using fully convolution neural networks. *IEEE Trans. Geosci. Remote Sens.* **2018**, *57*, 1100–1116. [CrossRef]
24. Xu, S.; Pan, X.; Li, E.; Wu, B.; Bu, S.; Dong, W.; Xiang, S.; Zhang, X. Automatic building rooftop extraction from aerial images via hierarchical rgb-d priors. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 7369–7387. [CrossRef]
25. Zhang, Y. Optimisation of building detection in satellite images by combining multispectral classification and texture filtering. *ISPRS J. Photogramm. Remote Sens.* **1999**, *54*, 50–60.
26. Cheng, L.; Gong, J.; Chen, X.; Han, P. Building boundary extraction from high resolution imagery and lidar data. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2008**, *37*, 693–698.
27. Jabari, S.; Zhang, Y.; Suliman, A. Stereo-based building detection in very high resolution satellite imagery using IHS color system. In Proceedings of the 2014 IEEE Geoscience and Remote Sensing Symposium, Quebec City, QC, Canada, 13–18 July 2014; pp. 2301–2304.

28. Pashaei, M.; Starek, M.J. Fully Convolutional Neural Network for Land Cover Mapping in A Coastal Wetland with Hyperspatial UAS Imagery. In Proceedings of the IGARSS 2019—2019 IEEE International Geoscience and Remote Sensing Symposium, Yokohama, Japan, 28 July–2 August 2019; pp. 6106–6109.
29. Audebert, N.; Le Saux, B.; Lefèvre, S. Joint learning from earth observation and openstreetmap data to get faster better semantic maps. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, HI, USA, 21–26 July 2017; pp. 67–75.
30. Vargas-Muñoz, J.E.; Lobry, S.; Falcão, A.X.; Tuia, D. Correcting rural building annotations in OpenStreetMap using convolutional neural networks. *ISPRS J. Photogramm. Remote Sens.* **2019**, *147*, 283–293. [CrossRef]
31. Zeiler, M.D.; Fergus, R. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2014; pp. 818–833.
32. Samek, W.; Wiegand, T.; Müller, K.R. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv* **2017**, arXiv:1708.08296.
33. Kingma, D.P.; Welling, M. Auto-encoding variational bayes. *arXiv* **2013**, arXiv:1312.6114.
34. Cui, Y.; Zhou, F.; Lin, Y.; Belongie, S. Fine-grained categorization and dataset bootstrapping using deep metric learning with humans in the loop. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1153–1162.
35. 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, 18–23 June 2018; pp. 1857–1866.
36. Kindermans, P.J.; Schütt, K.T.; Alber, M.; Müller, K.R.; Erhan, D.; Kim, B.; Dähne, S. Learning how to explain neural networks: Patternnet and patternattribution. *arXiv* **2017**, arXiv:1705.05598.
37. Liu, P.; Choo, K.K.R.; Wang, L.; Huang, F. SVM or deep learning? A comparative study on remote sensing image classification. *Soft Comput.* **2017**, *21*, 7053–7065. [CrossRef]
38. Goodfellow, I.J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.C.; Bengio, Y. Generative Adversarial Networks. *arXiv* **2014**, arXiv:abs/1406.2661.
39. Larsen, A.B.L.; Sønderby, S.K.; Larochelle, H.; Winther, O. Autoencoding beyond pixels using a learned similarity metric. *Int. Conf. Mach. Learn. PMLR* **2016**, *48*, 1558–1566.
40. Donahue, J.; Krähenbühl, P.; Darrell, T. Adversarial feature learning. *arXiv* **2016**, arXiv:1605.09782.
41. Su, J. O-GAN: Extremely concise approach for auto-encoding generative adversarial networks. *arXiv* **2019**, arXiv:1903.01931.
42. Wei, X.S.; Luo, J.H.; Wu, J.; Zhou, Z.H. Selective convolutional descriptor aggregation for fine-grained image retrieval. *IEEE Trans. Image Process.* **2017**, *26*, 2868–2881. [CrossRef]
43. Wang, Z.; Simoncelli, E.P.; Bovik, A.C. Multiscale structural similarity for image quality assessment. In Proceedings of the Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, Pacific Grove, CA, USA, 9–12 November 2003; Volume 2, pp. 1398–1402.
44. Rüschendorf, L. The Wasserstein distance and approximation theorems. *Probab. Theory Relat. Fields* **1985**, *70*, 117–129. [CrossRef]
45. Gaofen-2—Satellite Missions—eoPortal Directory. Available online: https://directory.eoportal.org/web/eoportal/satellite-missions/g/gaofen-2 (accessed on 27 December 2020).
46. China Centre for Resources Satellite Data and Application. Available online: http://www.cresda.com/EN/ (accessed on 27 December 2020).
47. Triggs, B.; McLauchlan, P.F.; Hartley, R.I.; Fitzgibbon, A.W. Bundle adjustment—A modern synthesis. In *International Workshop on Vision Algorithms*; Springer: Berlin/Heidelberg, Germany, 1999, pp. 298–372.
48. Hu, J.; Shen, L.; Albanie, S.; Sun, G.; Wu, E. Squeeze-and-Excitation Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 2011–2023. [CrossRef] [PubMed]