energies

MDPI

Review

Deep Learning-Based Building Extraction from Remote Sensing Images: A Comprehensive Review

Lin Luo [1], Pengpeng Li [1,2] and Xuesong Yan [1,*]

[1] School of Computer Science, China University of Geosciences, Wuhan 430074, China; luolin_computerscience@cug.edu.cn (L.L.); lipp@cug.edu.cn (P.L.)
[2] Wuhan Geomatics Institute, Wuhan 430021, China
* Correspondence: yanxs@cug.edu.cn

Abstract: Building extraction from remote sensing (RS) images is a fundamental task for geospatial applications, aiming to obtain morphology, location, and other information about buildings from RS images, which is significant for geographic monitoring and construction of human activity areas. In recent years, deep learning (DL) technology has made remarkable progress and breakthroughs in the field of RS and also become a central and state-of-the-art method for building extraction. This paper provides an overview over the developed DL-based building extraction methods from RS images. Firstly, we describe the DL technologies of this field as well as the loss function over semantic segmentation. Next, a description of important publicly available datasets and evaluation metrics directly related to the problem follows. Then, the main DL methods are reviewed, highlighting contributions and significance in the field. After that, comparative results on several publicly available datasets are given for the described methods, following up with a discussion. Finally, we point out a set of promising future works and draw our conclusions about building extraction based on DL techniques.

Keywords: deep learning; convolutional neural network; building extraction; high resolution; remote sensing

Citation: Luo, L.; Li, P.; Yan, X. Deep Learning-Based Building Extraction from Remote Sensing Images: A Comprehensive Review. *Energies* **2021**, *14*, 7982. https://doi.org/10.3390/en14237982

Academic Editor: Francisco Manzano Agugliaro

Received: 22 October 2021
Accepted: 24 November 2021
Published: 29 November 2021

1. Introduction

With the rapid development of imaging technology, high-resolution remote sensing (RS) imagery is becoming more and more readily available. Therefore, research within the field of RS has flourished, and automatic building segmentation from high-resolution images has received widespread attention [1–15]. The process of extracting buildings from RS images is shown in Figure 1, which is essentially a pixel-level classification of RS images to obtain binary images with contents of building or non-building, and this process can be modeled as a semantic segmentation problem [16–29].

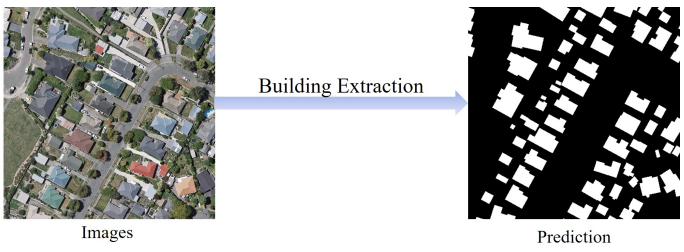


Figure 1. Illustration of extracting buildings from remote sensing images. The white and black pixels in prediction denote buildings and background respectively.

Deep learning (DL), with convolutional neural networks (CNN) [30–34] as its representative, is an automated artificial intelligence technique that has emerged in recent years,

specializing in learning general patterns from large amounts of data as well as exploiting the knowledge learned to solve unknown problems. It has been successfully applied and rapidly developed in areas such as image classification [35], target detection [36], boundary detection [37], semantic segmentation [16], and instance segmentation [38] in the field of computer vision. Proving to be a powerful tool for breakthroughs in many fields, DL techniques applied to building extraction in RS have emerged and become the mainstream technical tools. Although there are some reviews on RS image building extraction [39–42] or DL-based RS image processing [43–45], there is still a lack of a research that summarizes the latest results of RS image building extraction based on DL techniques. In this paper, we extensively review the DL-based building extraction from RS images, excluding the extraction of roads and other man-made features, in which the processing inputs include aerial images, satellite images, and other multi-source data such as light detection and ranging (LiDAR) point cloud data and elevation data.

As a fundamental task in the field of RS, automatic building extraction is of great significance in a wide range of application areas such as urban planning, change detection, map services and disaster management [46–56]. It is the basis for accomplishing these applications to have efficient and accurate building information. Building extraction has some unique features and challenges, which mainly include the following:

- Building types are in general highly changeable. They differ in interior tones and textures and have a variety of spatial scales. In addition, their shapes and colors may vary from building to building.
- Buildings generally stand in close proximity to features of similar materials such as roads, and can easily be confused with other elements. The segmentation quality of boundary contours is particularly important.
- The long-distance association relationship between buildings and surrounding objects is an important concern due to a variety of complex factors that may cause foreground occlusions, such as shadows, artificial non-architectural features, and heterogeneity of building surfaces.
- RS images have more complex and diverse backgrounds and scenes, and the shapes of buildings are more regular and well-defined than those of natural objects, rendering boundary issues particularly critical.

DL techniques have breathed new momentum into meeting these challenges and have sparked a wave of new promising research. In this paper, we review these research advances, with the following core contributions:

1. A detailed description of existing high-quality public datasets applied to building segmentation problems and commonly used evaluation metrics, which are key elements in judging the effectiveness of building segmentation methods.
2. Structured review of existing DL-based building segmentation methods, including their characteristic structures or contributions.
3. The quantitative experimental results provided in the literature are discussed, including their problems as well as some properties.
4. An outlook based on extensive literature research and summary is given for the possible directions of future work.

As shown in Figure 2, the remaining parts of the article are organized as follows. The introduction of DL techniques related to the building segmentation problem is presented in Section 2. In Section 3, descriptions of the most widely used public building datasets and evaluation metrics for segmentation models are presented. Then, an overview of existing methods is presented in Section 4. In Section 5, the proposed methods are briefly discussed based on quantitative results on the datasets, and future research directions are provided. Finally, Section 6 concludes the paper.

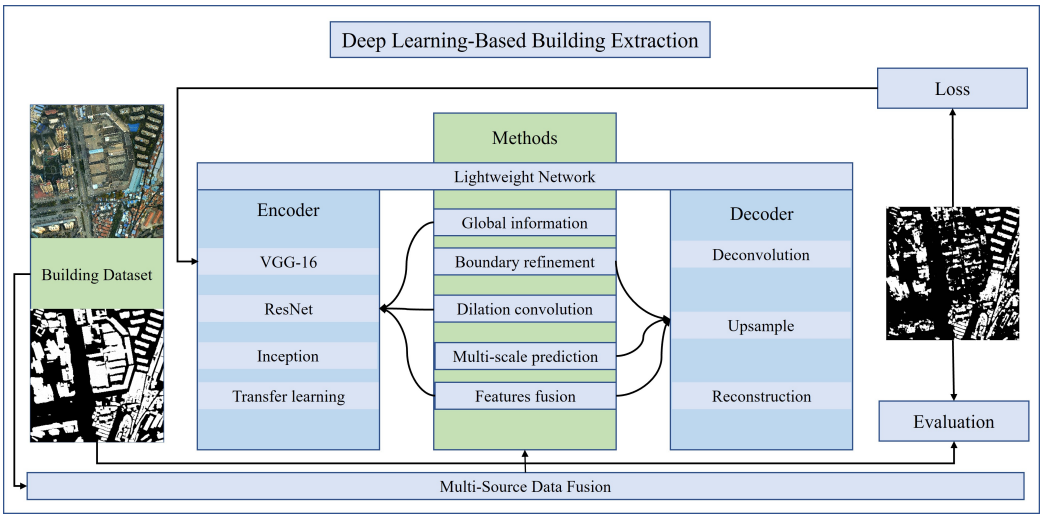


Figure 2. Visualization of the reviewed contents.

2. DL Techniques

Semantic segmentation, one of the research directions of DL most closely related to building extraction, is not an isolated research area, but a natural step in the process from coarse to refined inference. It is a downstream task of image classification, a fundamental computer vision task, for which image classification models provide feature extractors that extract rich semantic features from different layers. In this section, firstly, we recall the classic deep CNNs and design inspirations used as deep semantic segmentation systems, and point out its enlightening role for subsequent segmentation networks. Then, transfer learning, an important means of training DL models, is introduced. Finally, we introduce the loss functions used to train segmentation networks.

2.1. Deep CNNs

As one of the most fundamental tasks in computer vision, the image classification task assigns labels based on the input image and predefined categories. CNN-based image classification methods have matured in recent years and have become an important part of the downstream task of semantic segmentation. Here, we briefly review some classical CNN architectures for image processing, which include VGG, GooLeNet, and ResNet.

2.1.1. VGG Networks

In 2014, Visual Geometry Group (VGG) [57] at the University of Oxford proposed a network with more than 10 layers with concise design principles to build deeper neural network models. The structure of the VGG network is shown in Figure 3, with the main components being a 3×3 convolution operation and a 2×2 max-pooling operation. The small-sized convolutional layer has a smaller number of parameters and computations than the convolutional layer with large-sized convolutional kernels (e.g., 5×5 or 7×7 convolution operations in AlexNet [35]) to obtain a similar perceptual field. In addition, a remarkable feature is to increase the number of feature maps after using the pooling layer, reducing the phenomenon of useful information loss in feature maps after downsampling.

VGG network is one of the most influential CNN models because of its reinforcing the important idea in DL that CNNs with deeper architectures can facilitate hierarchical feature representation of visual data. It could be a guide to the structural design of subsequent deep CNN models. Meanwhile, VGG with 16 layers (VGG-16) has become one of the common feature extractors for downstream tasks.

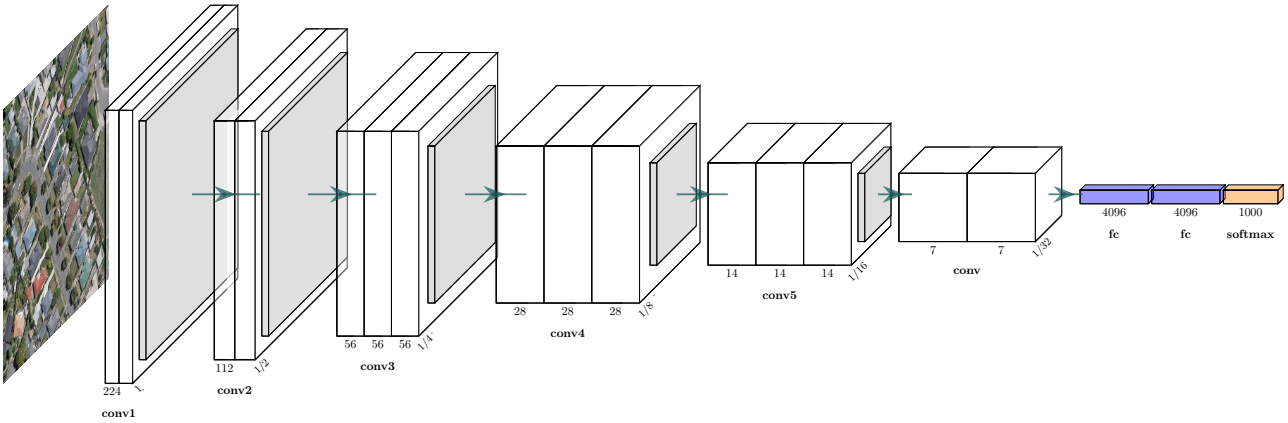


Figure 3. VGG network architecture.

2.1.2. GoogLeNet

GoogLeNet based on inception modules was proposed by Google in 2014 [58], which won the ImageNet competition that year. It has been improved several times in the following years, leading to InceptionV2 [59], InceptionV3 [60] and InceptionV4 [61]. The structure of the inception module is shown in Figure 4, presenting a net-in-network (NIN) architecture. The same network layer including large-size convolution, small-size convolution and pooling operations can capture feature information separately in a parallel manner. In addition, inception modules control the number of channels with 1×1 convolution and enhance the network representation by fusing information from different sensory domains or scales. Due to these modules, the number of parameters and operations is greatly reduced, while the network advances in terms of storage footprint and time consumption. The idea of inception provides a new way of stacking networks for CNN architecture design, rather than just sequential stacking, as well as it can be designed to be much wider. For the same number of parameters, inception-based networks are wider and more expressive, providing a fundamental direction for lightweight design of deep neural networks.

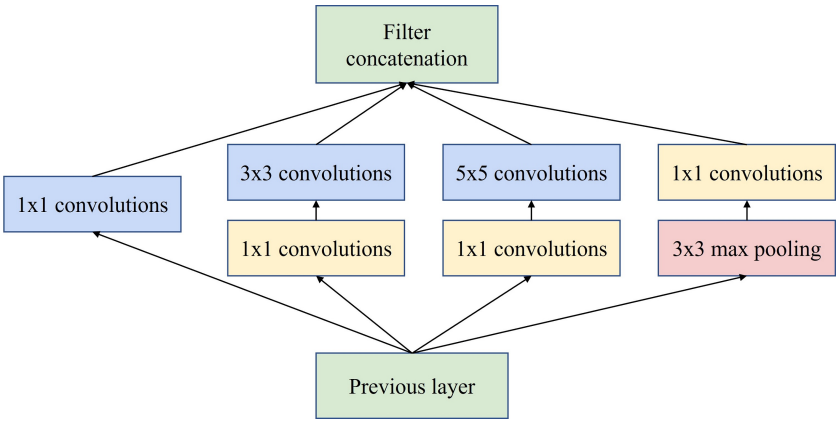


Figure 4. Inception module with dimensionality reduction from the GoogLeNet architecture.

2.1.3. ResNet

Presented in 2015, ResNet [62] is a landmark research result that pushed neural networks to deeper layers. 152-layer ResNet ranked in the top five at ILSVRC 2015 with an error rate of 3.6% and achieved a new record with respect to classification, detection, and localization in a single network architecture. Through experiments and analysis of several deep CNN models, it was found that deep networks experience network degradation during layer deepening and could not necessarily perform better than shallow networks. In response, a deep residual structure was proposed as shown in Figure 5, allowing the

network to shift to learning of residuals. The residual network learns new information different from what was previously available, relieving the pressure on the deep network to learn feature representations and update parameters. It allows the DL model to move once again in a deeper and better direction.

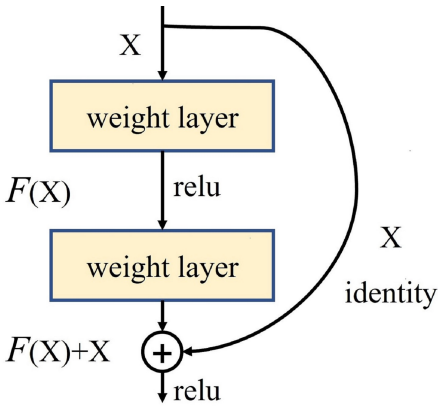


Figure 5. Residual block from the ResNet architecture.

2.2. Transfer Learning

Training a deep neural network from scratch is often difficult for two reasons. On the one hand, training a deep network from scratch requires enough dataset, and the dataset of the target task is not large enough. On the other hand, it takes a long time for the network to reach convergence. Even if a sufficiently large dataset is obtained and the network can reach convergence in a short time, it is much better to start the training process with weights from previous training results than with randomly initialized weights [63,64]. Yosinski et al. [65] demonstrated that even features learned by migration from less relevant tasks are better than those learned directly from random initialization. It also takes into account that the transferability will decrease as the difference between the source task and the target task increases.

However, the application of transfer learning techniques is not so straightforward. The use of pre-trained networks must satisfy the network architecture constraint of using existing network architectures or network components for transfer learning. Then, the training process itself in transfer learning is very small compared to the training process from scratch, so it can pave the way for fast convergence of downstream tasks. An important practice in transfer learning is to continue the training process from a previously trained network to fine-tune the weight values of the model. It is important to choose the layers for fine-tuning wisely, generally choosing the higher layers in the network, as the underlying layers generally tend to retain more general features.

ImageNet [66,67] is a large image classification dataset in the field of computer vision and is often used to train the feature extraction network part of segmentation networks. VGG-16 and ResNet pre-trained by ImageNet are easily available to be used as the encoder part of the segmentation network as well. In addition, a large collection of RS image segmentation data has also been collected and merged into a large dataset and used to pre-train the segmentation network [68].

2.3. Loss Function

Deep neural network models are trained with the loss-gradient back-propagation algorithm, so that the design of the loss function is also directly related to the efficiency of the network training and the performance of the model on the target task. Then the rest of this section describes several commonly used loss functions in building segmentation networks. To facilitate the expression of the computational process, y and p denote the ground truth label and prediction result, respectively.

- Cross entropy loss: Cross entropy loss (CE) is the most commonly used loss function in dense semantic annotation tasks. It can be described as:

$$CE(p, y) = -\sum y \log(p) \tag{1}$$

- Weighted Cross Entropy loss: Weighted cross entropy loss (WCE) is obtained by summing over all pixel losses and can not actively cope with application scenarios such as building extraction where the categories are unbalanced. Therefore, WCEs that consider category imbalance, such as median frequency balancing (MFB) [69,70] CEs, have emerged.

$$MFBCE(p, y) = -\sum y \log(p) \times w \tag{2}$$

  where w is the category balance weight in median form, expressed by the ratio of the median of the pixel frequencies of all the categories to the pixel frequencies of that category.
- Dice loss: Dice loss is designed for the intersection over union (IoU), an important evaluation metric in semantic segmentation, and is designed to improve the performance of the model by increasing the value of this evaluation metric.

$$DiceLoss(p, y) = 1 - \frac{2\sum(y \times p)}{\sum(y) + \sum(p)} \tag{3}$$

- Focal loss: Focal loss (FL) is improved from CE loss. To address class imbalance, an intuitive idea is to use weighting coefficients to further reduce the loss of the easy classification category. FL can be expressed as:

$$FL(p) = -\alpha(1-p)^{\gamma} \log(p) \tag{4}$$

  where α is the weighting factor for the classes and $\gamma \geq 0$ is a tunable parameter.

3. Datasets and Evaluation Metrics

In the case of DL, data is an extremely important component, specially with the deepening of the network and the increasing number of parameters. The establishment of each new building extraction method for basic DL requires the validation of a dataset. In this section, several publicly available datasets and important evaluation metrics for evaluating deep segmentation networks are presented.

3.1. Open Datasets

The data sources used to validate the building extraction methods are numerous including datasets compiled by several research institutions and data obtained by literature authors from publicly available websites (e.g., Google Earth, OpenStreetMap, and United States Geological Survey [71–77]). The former is of higher quality, while the latter is relatively more confusing and less generalized. Therefore, only three high quality datasets with considerable applications are described next in this section.

- Massachusetts Buildings Dataset [78]: The datasets, available on the website of Toronto University (https://www.cs.toronto.edu/~vmnih/data/, 15 August 2021), consists of 151 high-resolution aerial images of Boston's urban and suburban areas. The image size in Massachusetts Buildings Dataset is 1500 × 1500 pixels, and each image covers a widespread area of 2250 × 2250 m^2. The dataset was randomly divided into three subsets: a training set of 137 images, a validation set of 4 images and a test set of 10 images. It is worth mentioning that these data are restricted to regions where the average missed noise level is about 5% or lower. An example is shown in Figure 6.
- Inria Aerial Dataset [79]: This dataset, available on https://project.inria.fr/aerialimagelabeling/ (15 August 2021), consists of 360 high-resolution RGB aerial images covering different cities, including Austin, Chicago, Gitza, West/East Tyrol, Vienna,

Bellingham, Bloomington, and San Francisco. The areas cover urban buildings with different characteristics. For example, most of the buildings in Chicago and San Francisco are densely distributed and usually smaller in shape, while the buildings in Kitsap are scattered. The images have a spatial resolution of 0.3 m and an image size of 5000 × 5000 pixels, each covering a widespread surface of 1500 × 1500 m^2. Only 180 images were provided with public pixel annotation (ground truth), and the remaining 180 images were reserved for testing, where users could submit predicted images and obtain scores on the official website. To test the performance of the segmentation method more easily and quickly, by convention, the first five images of each region from the training set can be selected for validation. It is worth mentioning that all image data are of high quality as they are derived from different aerially captured orthorectified images of the landscape that are officially available locally, ignoring data such as Open Street Maps (OSM). An example is shown in Figure 7.

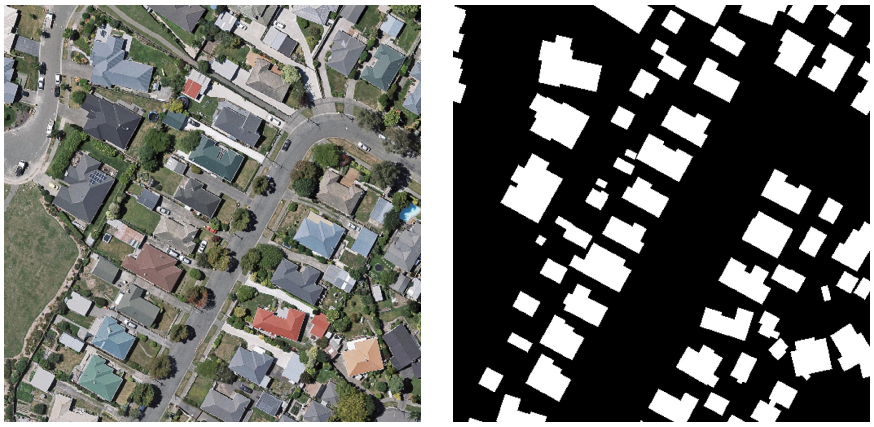
- WHU Building Dataset [80]: The whole dataset, available on the website of Photogrammetry and Computer Vision (GPCV) at Wuhan University (http://gpcv.whu.edu.cn/data/, 15 August 2021), contains both aerial image dataset and satellite image dataset. The WHU aerial dataset covers 18,700 buildings of diverse shapes and colors. The entire image and the corresponding vector shapefiles were seamlessly cropped into 8189 patches of 512 × 512 pixels with a ground resolution of 0.3 m. The WHU satellite dataset consists of six adjacent satellite images covering 550 km^2 in East Asia with a ground resolution of 2.7 m. Images of different colors from different sensors and seasons constitute a challenging case for automated building extraction. The vector building map contains 29,085 buildings. The entire image is also seamlessly cropped into 17,388 slices for training and testing, processed in the same way as the aerial dataset. Of these, 21,556 buildings (13,662 tiles) were used for isolated training and the remaining 7,529 buildings (3,726 tiles) were used for testing. An example is shown in Figure 8.

(**a**) (**b**)

Figure 6. An example of the Massachusetts Building Dataset. (**a**) Original image; (**b**) Ground truth label.

(**a**)

Figure 7. An example of the Inria dataset. (**a**) Original image; (**b**) Ground truth label.



(**a**) (**b**)

Figure 8. An example of the WHU dataset. (**a**) Original image; (**b**) Ground truth label.

3.2. Evaluation Metrics

In order to evaluate the performance of segmentation methods, it is usually necessary to select some quantitative evaluation metrics to evaluate the accuracy of different methods. In this section, we introduce the commonly used evaluation metrics, including pixel accuracy (PA), precision (Pre), recall (Rec), F1 score (F1), and IoU. In the building extraction task, the building is the positive case and the background category is the negative case. These four metrics are defined as:

$$PA = \frac{tp + tn}{tp + tn + fp + fn} \tag{5}$$

$$Pre = \frac{tp}{tp + fp} \tag{6}$$

$$Rec = \frac{tp}{tp + fn} \tag{7}$$

$$F1 = 2 \cdot \frac{Pre \cdot Rec}{Pre + Rec} \tag{8}$$

$$IoU = \frac{tp}{tp + fp + fn} \tag{9}$$

where tp, tn, fp and fn are the number of true positives, true negatives, false positives and false negatives pixels, respectively.

4. Building Extraction Methods Based on DL

DL techniques represented by CNNs have been developed for a long time in the direction of building extraction under the field of RS, whose processing of the input and output can be shown in Figure 9. Various deep neural network architectures for solving building extraction problems have emerged one after another.

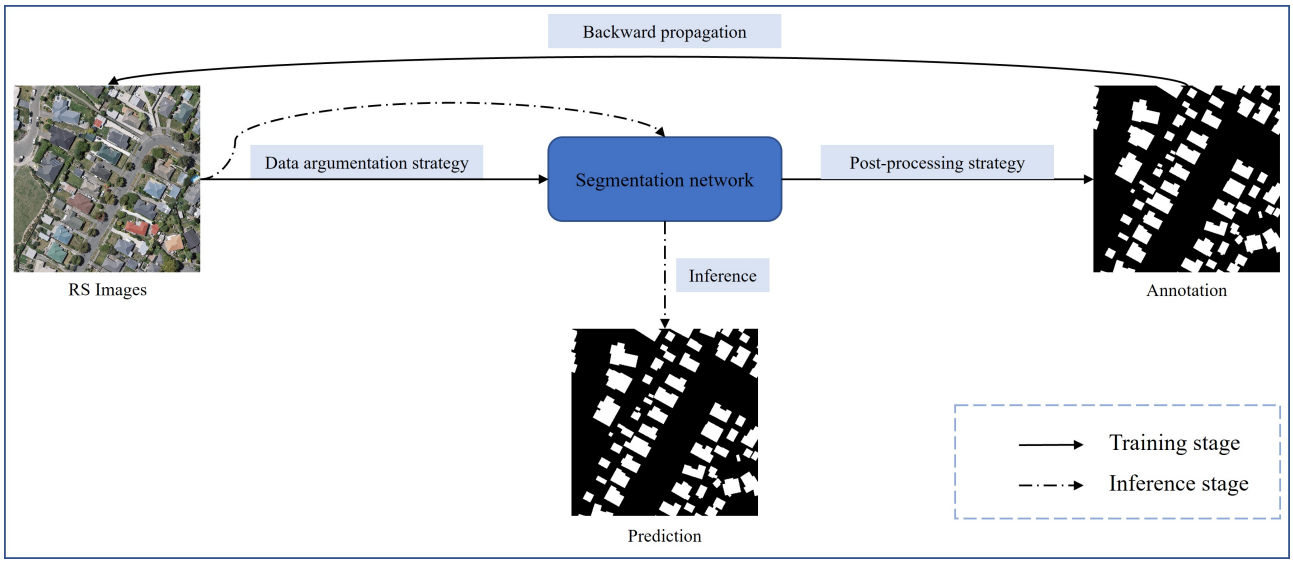


Figure 9. Processing of input and output on DL-based building extraction study. The dataset is partially enhanced to supply the model for back-propagation training until a certain termination condition is reached, such as iteration time and number of iterations. After that, the model can enter the application phase, where inference on unseen data produces predictions that match the requirements.

Patch-based annotation networks [81,82] are the key process of the adoption of DL into the building segmentation problem, with the main advantage of helping researchers to free themselves from complex manual feature design and perform automated building extraction for high- and even ultra-high-resolution RS images. The patch-based approach is essentially an image classification network that assigns a specified label to each patch, where the last layer of the network is usually a fully connected layer. The method cuts the image into a number of sub-images much smaller than the original size, i.e. patches, after which a CNN is applied to process the individual patches and give a single classification for each one, and finally stitch them together to form a complete image. The patch-level annotation method does not require high capacity of the network, and the network is usually uncomplicated in structure and easy to design. Saito et al. [81] designed a simple neural network containing three convolutional layers and two fully connected layers to accomplish automatic extraction of buildings, in which the feasibility and effectiveness of the method was confirmed by experiments. However, the patch-based classification method has two inherent defects that can not be avoided. On the one hand, the features of neighboring patches are similar and the proportion of overlapping regions is extremely large. Thus, there is a large amount of redundant computation, resulting in wasted resources and low efficiency. On the other hand, there is a lack of long-distance information exchange. As a result, the method can not fully exploit the contextual information in high-resolution RS images, with difficulty in completely and accurately extracting buildings from complex backgrounds.

Fully convolutional network (FCN) [16] is a landmark pixel-based segmentation method proposed to provide new inspiration for applying CNNs to advance building extraction research. The core idea is to use existing CNNs as encoders to generate hierarchical features and use upsampling means such as deconvolution as decoders to reconstruct

images and generate the semantic labels, eliminating the fully connected layer exclusively. A classical encoder-decoder structure is formed, which can theoretically accept images of different sizes as network inputs and output semantically labeled images at pixel level with the same resolution. There exists a common point in current segmentation networks, i.e., feature extraction is performed by the encoder in the process of performing multi-stage downsampling, and the decoder gradually recovers the size and structure of the image in the process of upsampling and generates semantic annotations. Based on this starting point consideration, a popular approach is to use the image classification network with the fully connected layer removed directly as a feature extraction network, i.e., encoder, such as VGG-16, GoogLeNet and ResNet; the decoder part is composed of upsampling modules such as deconvolution, which eventually generates dense pixel-level labels.

However, in spite of a robust approach, there are limitations in the classical FCN model for building extraction from RS images:

1. RS images are usually high-resolution with rich contextual semantic information, while the classical classification network is not sufficient for mining global contextual information.
2. CNNs do well in mining local features, but not in modeling long-distance association information. It is difficult for the plain decoder structure to reconstruct the structured hierarchical detail information, such as building boundaries and contours, which is lost due to the decrease of feature map resolution caused by the encoder downsampling.
3. The RS images are informative, so the processing of building extraction problem should focus on the model operation efficiency while ensuring the segmentation accuracy.

In this section, we first give an overview of the baseline methods such as FCN, SegNet [17], and U-Net [18] applied to building extraction. These problems mentioned above are then reviewed along with the current feasible solutions to these problems. Table 1 shows the main methods involved in the review (in the order in which they appear in this section), containing the architectures involved in the methods, their main contributions, and a hierarchy based on their task objectives: accuracy (ACC) and reusability (Reu) of the model structure. Specifically, Reu indicates whether the advanced network modules proposed in the literature can be reused relatively easily by other partitioned networks or studies. Each objective is divided into three grades, depending on the degree of focus of the corresponding work on that objective. From the perspective of accuracy, aggregating multi-scale contextual information, considering boundary information, iterative refinement and adopting appropriate post-processing strategies are aspects to be considered. Network components with good reusability are usually robust while not producing large changes in the size of the input and output feature maps, such as attention modules. Moreover, they are usually designed with the objective of aggregating certain elements that are useful for accomplishing the target task.

Table 1. Deep learning-based methods on building extraction.

Methods	Acc.	Reu.	Contributions
DeconvNet-Fusion [2]	**	**	Multi-source data post-fusion
FCN [83]	*	*	Early CNNs
ConvNet [84]	*	*	Signed distance
Fused-FCN4s [85]	**	**	Multi-source data post-fusion
SegNet-Dist [86]	*	*	Signed distance
MC-FCN [87]	*	**	Multi-scale architecture
MFRN [88]	*	**	Multi-scale architecture
BR-Net [89]	**	**	Boundary extraction, multiple tasks
GMEDN [90]	***	**	NB, multi-scale architecture
ENRU-Net [91]	**	**	APNB
PISANet [92]	**	**	Pyramid self-attention module

Table 1. *Cont.*

Methods	Acc.	Reu.	Contributions
ELU-FCN-CRFs [93]	*	*	ELU, CRFs
FC-DenseNet-FPCRF [71]	**	**	FPCRFs, GCNs
CNN-RNN [94]	***	***	Iterative refinement of RNN architecture
EANet [95]	***	***	Boundary-aware networks
Networks with BP loss [96]	**	**	BP loss
BRRNet [97]	***	***	Residual refinement module
DSFE-GGCN [98]	**	***	Gated GCN, deep feature embedding
FCN with LFE [99]	*	*	Local feature extraction module
EU-Net [100]	*	**	DSPP, category balanced loss
ScasNet [101]	***	***	Multi-scale aggregation
SR-FCN [102]	*	**	Multiscale prediction, ASPP
Building-A-Nets [103]	**	***	GAN
P-LinkNet [104]	*	**	Multi-scale structure LinkNet
MA-FCN [105]	*	**	Boundary constraints, multiscale prediction
GAN-SCA [106]	**	***	SCA, GAN
HFSA-Unet [107]	***	***	Two-stage channel attention
ESFNet [108]	*	***	Separable factorized residual block
ACR-Net [109]	**	***	RBAC
SegNet-Dist-Fused [110]	*	*	Signed distance, multi-source data fusion
CFCN [111]	**	**	Boundary constraint networks

4.1. Baseline Methods

FCN, SegNet, and U-Net all employ an encoder-decoder architecture, but offer different aspects of design mindsets that are reflected in the encoder, upsampling, and skip connection, respectively.

- The encoders for FCN and SegNet are usually obtained by removing fully connected layers using classification networks such as VGG-16 and ResNet, and the encoder for U-Net is designed to be symmetric with the decoder, allowing the depth of the network to be increased or decreased depending on the complexity of the task.
- The decoder structure of FCN is the simplest and contains only one deconvolution operation, while U-Net and SegNet adopt multiple upsampling to organize the decoder structure.
- There is a feature fusion by FCN with feature maps organized by pixel-by-pixel summing, U-Net with feature map stitching, and SegNet with pooling indices generated by pooling operation embedded in the decoder feature map to solve the problem of insufficient recovery information in the upsampling process.

These three types of basic methods have been applied to the building extraction problem since a few years ago [3,83,84,86–89], which have recently been used mainly as a baseline to motivate new methods and to compare their effectiveness.

4.2. Contextual Information Mining

The key points of building extraction are mining local information (short-distance contextual information around pixels such as building outline and boundary) and global information (long-distance contextual information between buildings and background and overall association relationship between buildings and buildings with other pixels in the image). Rich local information helps to improve the accuracy of pixel-level annotation, while complete global information is also essential to resolve local blur. It is the concern of all DL-based building extraction methods to balance and fuse these two aspects.

4.2.1. Global Information Mining

An encoder that contains only backbone network such as VGG-16 and ResNet for feature extraction is imperfect. It is specialized in extracting local information at short

distances, but not able to adequately extract contextual information at long distances, which is mainly due to the inherent properties of CNNs.

A popular approach to get out of this dilemma is to develop network modules based on self-attention mechanisms. Non-local block (NB) [112] is a self-attention network module developed to extract global information and efficiently capture global contextual information by computing the similarity relationship between each pair of pixels. At the same time, NB ensures that the input and output shapes are the identical and can be migrated to different networks in a "plug-and-play" manner. The global information is properly fused with the local information extracted by CNN to enhance the building extraction capability of the model. However, the computational effort of NB is closely related to the resolution of the input image with the drawback of high time complexity and high spatial complexity of computation, which reaches $O(H \times W \times H \times W)$. Asymmetric pyramid non-local block (APNB) [113] was introduced by Wang et al. [91] for contextual global information extraction, and it employs four adaptive pooling layers of different scales to reduce the number of pixels involved in the relational computation. Due to these pooling layers, the time and space complexity of APNB is reduced to $O(H \times W \times N)$, where N is much smaller than $H \times W$. To better balance the conflict between extracting local and global contextual information and coordinating the relationship between different scales and levels, Zhou et al. [92] combined atrous spatial pyramid pooling (ASPP) [114] and NB both structures to propose a pyramidal self-attentive module for convenient embedding in the network, which further enhances the information processing capability of FCNs. Overall, the self-attention mechanism can be cleverly attached to other convolutional networks, but it suffers from an inherent drawback that it can destroy the short-range association and detailed information between buildings to some extent, and should be used in conjunction with other means.

4.2.2. Boundary Contour Refinement

CNNs have spatial transformation invariance, which makes it impossible to accurately locate spatial locations when modeling long distances, so FCNs fail to accurately label the boundary contours of buildings. However, properties such as contours and boundaries are of great significance for buildings.

A feasible way to optimize the segmentation results is to introduce a second-trained conditional random field (CRF) as a post-processing module in the last layer of the segmentation network. The common implementation of CRF is a fully connected network that mines the interrelationships between long-range pixels without distance constraints, a characteristic that is difficult to be taken into account by CNNs. Shrestha et al. [93] argues that FCN involves upsampling operations that produce rough borders for building segmentation and extraction results. In contrast, among RS image segmentation, two pixels with similar location and color features have a high probability of being assigned the same label, and vice versa is less likely to be segmented. Hence, the fully connected CRF can be used to improve the FCN-8s with VGG-16 as the encoder to exploit the association characteristics between pixels and improve the blurring at the building boundaries. CRFs in the form of fully connected networks are separated from the previous CNN module when used as a post-processing module for segmentation results, and the previous CNN module is usually fixed during training of CRFs without information interaction with the features extracted by the CNN. Li et al. [71] proposed the feature pairwise conditional random field (FPCRF) based on the graph convolutional network (GCN) [115–119], which is a CRF for pairs of potential pixels with local constraints, incorporating the feature maps extracted by the CNN. FPCRF module can be added to the building segmentation network as a plug-and-play component to improve the segmentation performance of the model without significantly increasing the training and inference time. In addition, the training efficiency of FPCRF is significantly higher than that of the fully connected form of CRF. Sharing a similar design idea with CRF, Maggiori et al. [94] designed a recurrent neural network-based post-processing module for segmentation results by analyzing the

mathematical process of partial differential equations for the refinement process when considering the boundary refinement of building segmentation, which enhances the the confidence level of pixel classification and thus optimize the segmentation details.

It is time consuming to find a better CNN network architecture with no guarantee of robustness of the new CNN model. Another way to solve the boundary problem is to design boundary-aware networks in a targeted manner, and boundary labeling can be simply obtained by applying morphological erosion operations through building labeling. The branch network for boundary extraction is used to guide the semantic segmentation network to learn more information about the building boundaries that contribute to image segmentation. Bringing in a branch network focusing on boundary information processing in the segmentation network [95,96], the model achieved fine segmentation results by training the network with a joint loss of multiple branches, such as boundary-aware perceptual loss (BP loss).

Shao et al. [97] provides an idea of segmentation refinement by drawing on the design of residual networks, and the proposed network consists of two parts: a prediction module and a residual refinement module. The prediction part is a conventional segmentation network with encoder-decoder structure, and the residual refinement module is a residual learning module that accepts the coarse segmentation results generated by the prediction module to learn the different information between the coarse segmentation results and the true annotations. The whole network is trained jointly by a multi-branch loss function. The residual refinement module has a deeper hierarchy and large perceptual field with easy portability to other fully convolutional neural networks, thus improving the refinement capability of the basic segmentation network for boundary contour. Shi et al. [98] analyzed the conflict between the downsampling operation of deep CNN and the accurate segmentation of boundaries, and introduced a gated GCN based on GCN design into the CNN structure, which is able to refine the coarse semantic prediction results to generate clear boundaries and high fine-grained pixel-level classification results.

It is good to take into account the boundary information and some specific structures and train the network with joint loss, but the richness of the boundary samples can also be an important factor that should not be ignored to limit the performance exploration of such methods.

4.2.3. Dilated Convolution

Dilated convolution [120], also known as atrous convolution, has a convolutional kernel that encompasses a larger region in a mesh without significantly increasing the computational effort. It is able to rapidly increase the perceptual field while maintaining image resolution, alleviating the conflict that deep network features have a theoretical perceptual field much smaller than the actual perceptual field [121], thus making it easier to model multi-scale contextual information. Hamaguchi et al. [99] enhanced the extraction of local features with a reasonable stacking of small dilation rate dilation convolutions, effectively reducing the cases of ambiguous results for small-sized building segmentation.This property of that is used to extract multi-scale contextual information, and the most popular design structure is ASPP, as shown in Figure 10. ASPP mines feature information at different scales in parallel with multiple sets of dilation convolution kernels of different sizes. The proper dilation rate design ensures that each pixel location is involved in the computational side. Feature maps generated by multiple parallel branches are stitched together.

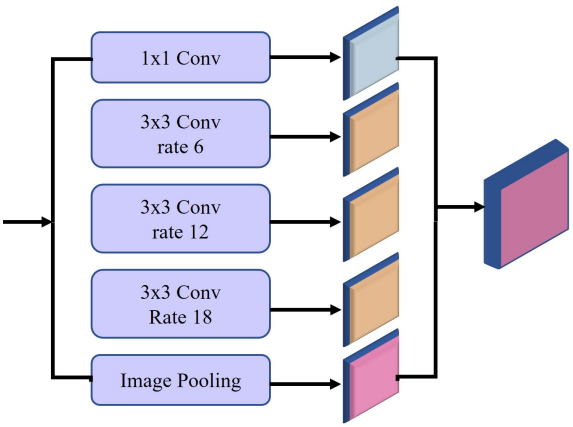


Figure 10. Atrous spatial pyramid pooling block.

The success of the DeepLab models [21–24] can be attributed to the innovation of the ASPP module, which has made a splash in the segmentation of natural images and has also been widely introduced in the study of building extraction [102,104]. Kang et al. [100] proposed a dense ASPP for quantitative analysis of image input size and perceptual field reality, named dense spatial pyramid pooling (DSPP) module, which consists of five parallel convolution or pooling layers. It consists of two dilated convolutions with dilation rates of 3 and 6, two standard convolutional layers with sizes of 1×1 and 3×3, and one pooling layer, obtaining dense multi-scale contextual information better than ASPP.

Different from the dilation convolution module with parallel structure like ASPP, Liu et al. [101] designed a self-cascaded multi-scale context aggregation module as shown in Figure 11. The semantic extraction module at each level is implemented by dilation convolution with different dilation rates, fusing multi-scale semantic information at different resources level by level. The self-cascaded module aims to aggregate global to local contextual information while well preserving hierarchical dependencies, i.e., the underlying containment and relative location relationships between objects and scenes at different scales.

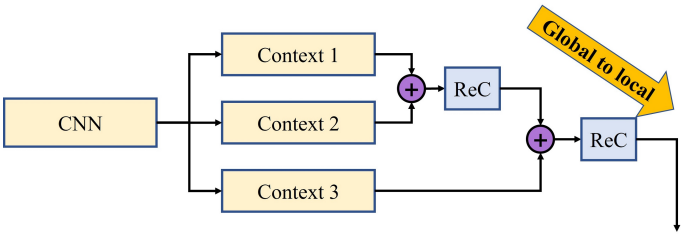


Figure 11. Multi-scale contextual information aggregation module.

4.2.4. Multi-Scale Prediction

An alternative approach to aggregate contextual information is to use a multi-scale prediction strategy, i.e., to guide the feature information towards the general direction of accurately labeled buildings as it propagates through the decoder step by step. In the process of decoder recovering feature map size to produce labeled images, the parameters of each convolutional block influence the change of feature mapping, so that the process of decoder expanding feature map size is full of many uncertainties. For this reason, a multi-scale prediction can be adopted to guide this process, which can be shown in Figure 12. Feature maps of different sizes are gradually generated during the upsampling process, where each prediction map aggregates the scale information of the previous level and requires the network to gradually produce better predictions.

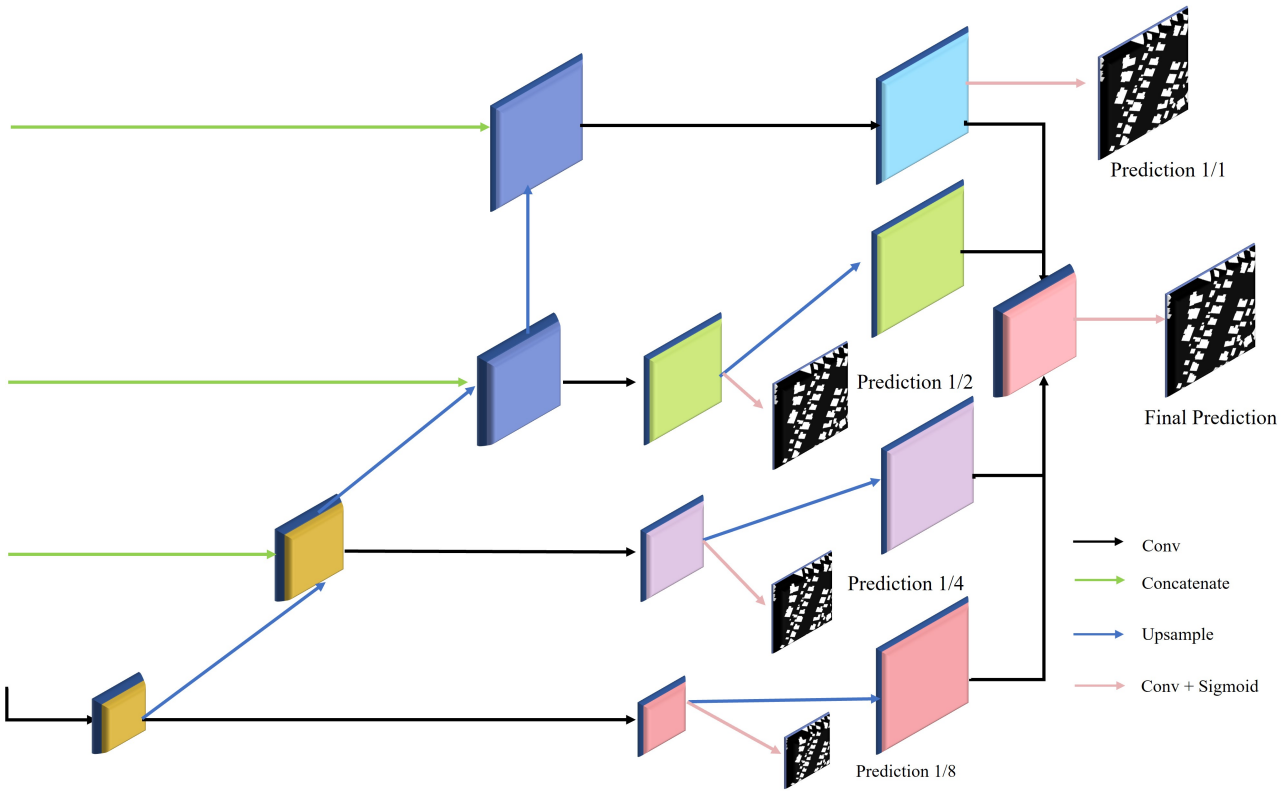


Figure 12. Multi-scale Prediction.

Ma et al. [90] designed a distillation decoder containing an upsampling branch and a multi-scale prediction branch. The upsampling branch contains five deconvolution layers, and the multi-scale branch aggregates the multi-scale feature maps generated by the last four upsamplings without considering the information generated by the first upsampling since it contains only little high-level semantic information. Model ablation experiments show that the aggregation of multi-scale information for prediction is effective and significantly outperforms baseline methods such as FCN, SegNet, and U-Net for building extraction on RS images. Ji et al. [102] introduced a multi-scale prediction module based on FCN, which generates multiple independent prediction maps simultaneously in the process of aggregating multi-scale information, imposing restrictions on the generation and transmission of each scale information level by level, guiding them in the direction that is conducive to producing better prediction maps. In addition to the restrictions imposed step-by-step inside the network, it is possible to bootstrap externally during the training phase without affecting the inference speed. Li et al. [103] and Pan et al. [106] proposed the models by applies the design idea of generative adversarial network (GAN), where a discriminative network is attached to the segmentation network to guide the segmentation network to generate annotated images that are very similar to the real annotations, and make the discriminative network unable to distinguish whether the input annotated images are real annotations or not. This adversarial training approach is shown to be effective and can be transferred to other models to improve model performance.

Multi-branch networks prove to be an important and effective means of extracting multi-scale information, but can limit the speed when inferring. A suitable network post-processing technique is the way to break this limitation.

4.2.5. Feature Fusion

Feature fusion aims to integrate global feature maps from different layers and relatively local feature maps by introducing contextual information in the network architecture in the form of skip connections. A commonly used approach is primarily the pixel-by-pixel position summation of feature maps from FCN and the stitching of feature maps along

the channel dimension from U-Net, which are widely used in RS image segmentation networks [104,105]. However, it is clear that the capability of the feature maps from different layers to provide information is varied and even in conflict, which may accumulate noisy information irrelevant to the segmentation task.

Bringing in attention mechanism is an important means to improve the way of feature fusion. Pan et al. [106] chained spatial attention module (SAM) and channel attention module (CAM) in fusing low-level features, enabling segmentation networks to selectively enhance more useful features in specific locations and channels. He et al. [107] designed a two-stage attention structure for exploring the correlation between different feature channels of intermediate features. The network module contains two parallel channel attention modules of different forms for extracting channel correlation features, which can be easily embedded into segmentation networks such as U-Net to enhance the segmentation capability.

4.3. Lightweight Network Design

Building extraction is usually performed in high-resolution images so that the design of segmentation networks has to take into account the consumption of computational resources such as GPU memory and the inference speed of the prediction phase. However, most existing methods usually require a large number of parameters and floating-point operations to obtain high accuracy, which leads to high computational resource consumption and low inference speed.

In order to achieve a better balance between accuracy and efficiency, a common approach is to apply an existing lightweight network or adopt a more efficient convolutional module to develop a lightweight network as a feature extraction network [122–124]. Lin et al. [108] and Liu et al. [109] developed new feature extraction backbone networks with deep separable convolutional asymmetric convolution respectively, incorporating decoder networks to achieve segmentation results with accuracy no less than mainstream networks such as U-Net, SegNet, and earlier lightweight networks such as ENet [125], with significantly lower number of parameters and computational effort.

4.4. Multi-Source Data

Extracting buildings from RGB images is currently the most widely used method. However, DSM elevation data and LiDAR data are widely used as auxiliary data to correct the angle of the building and pinpoint the location of the building to improve the accuracy of building segmentation. In other words, RGB data provides extensive background color information and building shape information, while DSM elevation data and LIDAR data provide accurate relative position information and three-dimensional spatial information. The fusion of RGB images with data from other sources exists in two main stages, the preprocessing stage before input to the network and the post-processing stage of the network.

Data fusion in the previous stage typically involves attaching multi-source data such as DSM as an additional channel to RGB images to form multi-channel data [110,111]. However, the approach of direct fusion of data ignores the variability of different data sources. Huang et al. [2] utilizes independent FCNs to provide segmentation results based on data from different data sources and performs feature fusion at the final layer with confidence votes to obtain the results. Bittner et al. [85] employs multiple mutually independent encoder networks for feature extraction from multiple sources of data separately, and the segmentation results are derived with a decoder after fusing the features.

5. Discussion and Outlook

In the previous sections we have reviewed the existing methods qualitatively, that is, we have not considered any quantitative results. In this section, we gather the results of the runs of these methods on the three most representative datasets (Massachusetts Building Dataset, Inria building dataset and WHU aerial imagery dataset) in terms of the

metrics described in Section 3 and perform analysis. In addition, we provide an outlook on possible directions for future research.

5.1. Analysis of Quantitative Experimental Results

After an extensive survey of the literature in recent years, it can be found that the experimental results of many methods are conducted on some non-standard datasets, which undoubtedly makes it difficult to engage in comparison of methods from different literature. Therefore, in this section, three widely used publicly available datasets are selected for the summary of quantitative experimental results. The collected experimental results are shown in Tables 2–4.

Table 2. Results on Massachusetts Building Dataset (%).

Methods	PA	Pre	Rec	F1	IoU
Patch-based CNN [82]	-	94.6	95.5	-	-
MFRN [88]	94.51	-	-	85.01	-
GMEDN [90]	93.78	-	-	-	70.39
ENRU-Net [91]	94.18	-	-	84.41	73.02
ELU-FCN-CRFs [93]	-	95.07	93.40	93.93	89.08
ResNet ScasNet [101]	-	-	-	85.58	74.34
EU-Net [100]	-	86.70	83.40	85.01	73.93
HFSA-Unet [107]	-	84.75	79.08	81.75	69.23

Table 3. Results on Inria building dataset (%).

Methods	PA	Pre	Rec	F1	IoU
Ensemble FCNs [15]	96.46	-	-	-	76.27
GMEDN [90]	96.43	-	-	-	76.69
PISANet [92]	94.50	85.92	88.68	87.27	77.45
UNet+BP Loss [96]	96.52	-	-	-	76.62
EU-Net [100]	-	90.28	88.14	89.20	80.50
Building-A-Nets [103]	96.71	-	-	-	78.73
P-LinkNet [104]	-	91.50	-	-	84.48
GAN-SCA [106]	96.60	-	-	-	77.52
HFSA-Unet [107]	-	92.30	89.89	91.07	83.63
ARC-Net [109]	92.5	89.6	86.8	87.5	77.9

Table 4. Results on WHU aerial building dataset (%).

Methods	PA	Pre	Rec	F1	IoU
ENRU-Net [91]	98.92	-	-	95.16	90.77
PISANet [92]	96.15	94.20	92.94	93.55	87.97
UNet+BP Loss [96]	98.84	95.06	94.89	94.97	90.78
EU-Net [100]	-	94.98	95.10	95.04	90.56
SR-FCN [102]	-	94.4	93.9	-	88.9
MA-FCN [105]	-	95.2	95.1	-	90.7
HFSA-Unet [107]	-	95.09	95.18	95.13	90.72
ESFNet [108]	-	-	-	-	85.34
ARC-Net [109]	97.5	96.4	95.1	95.7	91.8
MAP-Net [76]	-	95.62	94.81	95.21	90.86

However, the different settings of the experimental hyperparameters (e.g., training time and number of iteration rounds) still make it difficult to compare the newly developed and diverse approaches fairly even if these methods use the same dataset for experiments. Most of the approaches in the literature are compared with well-known baseline methods such as FCN, U-Net, and SegNet to demonstrate the effectiveness of the new methods.

It is worth mentioning that there are some approaches (e.g., ESFNet) whose research goal is to limit the complexity of the model or to pursue the speed of network inference, i.e.,

to make the model lightweight without significantly reducing its segmentation capability and to make the study of building extraction more relevant to practical application scenarios. Table 5 shows the amount of FLOPS and the number of parameters for some of the networks involved in this paper. Even though we do not compare the computational complexity measures of all the networks for completeness because the complexity-related measures are not clearly given in the original paper, we can still find the following patterns. Real-time semantic segmentation networks such as ENet, ERFNet and EDANet, as well as the efficient building segmentation network ESFNet, are dedicated to developing lightweight models with much lower number of parameters and FLOPS than other segmentation networks. The remaining networks, except for FCN-8s and Deeplabv3, are almost comparable in terms of computation and number of parameters, and are all controlled within an acceptable range. In addition, some works have reported their training time and inference time, such as GMEDN and ARC-Net. based on their structure and parameter size one can make an estimate of the computing time on GPU, the general model takes about 6 hours to train on the Massachusetts building dataset and Inria dataset, and about 0.26 seconds to infer a patch of 256$\times$ 256 size. But the training time consumption of lightweight model will be greatly reduced. For a larger dataset like WHU, the network may take ten hours or more to train, but random cropping of the input image to a smaller size would shorten that time.

Table 5. Model Complexity.

Methods	FLOPS (G)	Parameters (M)
FCN-8s	73.49	134.27
U-Net	53.51	29.55
SegNet	79.89	39.87
PSPNet	93.48	46.73
HRNetv2	59.20	29.54
Deeplabv3	121.06	60.99
Deeplabv3+	19.12	40.47
SRFCN	96.66	35.00
GMEDN	29.85	127.43
ENRU-Net [91]	51.87	73.71
MAP-Net [76]	48.09	24.00
ENet [125]	2.22	0.36
ERFNet [126]	14.67	2.06
EDANet [127]	4.410	0.68
ESFNet [108]	2.514	0.18

Considering the experimental comparison of methods and their reproducibility, each method should evaluate its results on a standard dataset, including important evaluation metrics such as IoU and F1. An excellent route is to describe their training process comprehensively and disclose their models with corresponding training weights.

Given the performance with respect to the methods, we draw the following points:

1. ResNet and VGG-16 are the most widely used backbone networks for local feature extraction.
2. Dilation convolution is the main underlying module for rapidly increasing the perceptual field and extracting multi-scale information.
3. Skip connection is a necessary way for the intersection of different levels of semantic information.
4. Multi-scale prediction and boundary constraints are powerful tools for improving the information processing capability of the decoder.

When considering the choice of model structure, the following factors can be considered and selected. U-Net can be the first model that can be tried to solve the building extraction problem for a specific region, and usually acceptable results can be achieved. DeeplabV3+ can be an advanced building segmentation network, and some feasible improvements can be developed on it, such as GMEDN, because they are highly scalable.

Lightweight real-time segmentation networks based on structures such as channel grouping strategy, deep separable convolution and dilated convolution are an optimal choice when hardware resources are limited within the scene, such as ESFNet.

5.2. Future Research Directions

Based on extensive literature research and summary, it is believed that DL techniques remain the mainstream approach to investigate building extraction and will continue to evolve in the future. However, technical challenges still exist, and the remainder of the section gives some research outlooks.

- Multi-source data fusion: RS image segmentation using the combination of RGB images and multi-source data such as LiDAR is an important research direction, but there still exist numerous challenges. There are many multi-source data noises, such as mis-matching of images with LiDAR data, so the robustness of the method is important. Exploring an effective and robust way of multi-source data fusion remains an important research point for the future [4,13].
- Feature maps fusion: Almost all current FCN segmentation methods use CNN for feature map upsampling and downsampling, which can directly lead to changes in feature map resolution information and thus cause information loss. A feasible approach is to take multiple feature maps from different locations of the network for fusion and complementary information loss [90,97,104,107], while the specific way of fusing multiple features is an open problem to be studied systematically.
- Multi-scale contextual information: Mining of multi-scale contextual information [104, 105,107] is a key component of building extraction networks. Although some modules based on dilated convolution attached to feature extraction backbone networks have been developed, further research on more efficient approaches remains necessary.
- Boundary optimization: Buildings are usually with a certain regular geometry or a combination of multiple geometries. Therefore, the annotation of boundary contours is extremely important, and improving the quality of boundary segmentation is an obvious research direction to enhance the model segmentation capability. It is a common practice to apply CRF to post-process segmentation results to further improve the accuracy of boundary contour annotation, which is simple in design but has limited room for improvement. A new class of approaches is to exploit a branch network specifically for boundary refinement [97,98] as well as a boundary loss function [95,96], for which appropriate training strategies are required.
- Lightweight network structure: The building extraction study faces a high-resolution RS image application scenario, where the inference speed of the application equipment is limited. Hence, it is of great importance to consider lightweight factors when designing segmentation networks, so that the networks can consume less computational and storage resources while ensuring little performance loss of the segmentation models [108,109]. One possible research direction is to separate the training and inference phases of the network. Multi-branch networks are easier to train and get good segmentation capability, while single-branch networks are faster in inference [128,129]. Therefore, the network structure can be merged or pruned during inference to improve the application capability in real scenarios.

6. Conclusions

In this paper, we focus on RS-based building extraction using DL semantic segmentation methods. Compared to other surveys on traditional building extraction methods or RS, this review paper is more devoted to the popular topic of DL semantic segmentation, covering the state-of-the-art and latest work. We analyse the building extraction problem and the basics of DL, providing the reader with the basic features of RS related to building distribution and the basic deep learning knowledge to conduct this research. We cover multiple types of popular DL semantic segmentation methods for solving building extraction and their associated elements such as datasets and evaluation methods, illustrating

their important organization and characteristics so that researchers can grasp the flow of research and current research progress. We investigate the methods involved from two perspectives: qualitative contributions and quantitative performance comparisons. In addition, we provide some helpful insights into the problems and methodological developments in the field, including the limitations and advantages of existing methods as well as the design and selection of networks. However, the work in this paper also suffers from the following two main limitations: no quantitative comparison of those non-open-source studies because of the difficulty of a fair comparison and less attention to downstream tasks relevant to building extraction, such as building instance segmentation and damage change detection. In conclusion, DL methods to solve the automatic building extraction problem from RS images are very powerful and mainstream means to promote practical application scenarios. Therefore, we strongly encourage researchers to make the implementation code of their proposed models and then their application datasets as open as possible, and look forward to a large number of innovations and research lines in the upcoming years.

Author Contributions: Conceptualization, L.L., X.Y. and P.L.; methodology, L.L.; formal analysis, L.L. and X.Y.; writing—original draft preparation, L.L.; writing—review and editing, L.L., X.Y. and P.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by the National Natural Science Foundation of China (U1911205).

Acknowledgments: We thank the anonymous reviewers for their insights and constructive comments, which helped to improve the paper.

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

Abbreviations

The following abbreviations are used in this manuscript:

RS	remote sensing
DL	deep learning
CNN	convolutional neural network
LiDAR	light detection and ranging
VGG	visual geometry group
NIN	net-in-network
CE	cross entropy
WCE	weighted cross entropy
MFB	median frequency balancing
IoU	intersection over union
FL	focal loss
PA	pixel accuracy
Pre	precision
Rec	recall
FCN	fully convolutional network
ACC	accuracy
NB	non-local block
APNB	asymmetric pyramid non-local block
ASPP	atrous spatial pyramid pooling
CRF	conditional random field
FPCRF	feature pairwise conditional random field
GCN	graph convolutional network
BP	boundary-aware perceptual
DSPP	dense spatial pyramid pooling
GAN	generative adversarial network
SAM	spatial attention module
CAM	channel attention module

References

1. Li, E.; Femiani, J.; Xu, S.; Zhang, X.; Wonka, P. Robust Rooftop Extraction From Visible Band Images Using Higher Order CRF. *IEEE Trans. Geosci. Remote Sens.* **2015**, *53*, 4483–4495. [CrossRef]
2. Huang, Z.; Cheng, G.; Wang, H.; Li, H.; Shi, L.; Pan, C. Building extraction from multi-source remote sensing images via deep deconvolution neural networks. In Proceedings of the 2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Beijing, China, 10–16 June 2016; pp. 1835–1838. [CrossRef]
3. Bittner, K.; Cui, S.; Reinartz, P. Building Extraction from Remote Sensing Data Using Fully Convolutional Networks. In Proceedings of the International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Hannover, Germany, 6–9 June 2017; pp. 481–486.
4. 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. [CrossRef]
5. Maltezos, E.; Doulamis, A.; Doulamis, N.; Ioannidis, C. Building Extraction From LiDAR Data Applying Deep Convolutional Neural Networks. *IEEE Geosci. Remote Sens. Lett.* **2019**, *16*, 155–159. [CrossRef]
6. Sun, G.; Huang, H.; Zhang, A.; Li, F.; Zhao, H.; Fu, H. Fusion of Multiscale Convolutional Neural Networks for Building Extraction in Very High-Resolution Images. *Remote Sens.* **2019**, *11*, 227. [CrossRef]
7. Wu, G.; Guo, Z.; Shao, X.; Shibasaki, R. GEOSEG: A Computer Vision Package for Automatic Building Segmentation and Outline Extraction. In Proceedings of the 2019 IEEE International Geoscience and Remote Sensing Symposium, Yokohama, Japan, 28 July–2 August 2019; pp. 158–161. [CrossRef]
8. Huang, J.; Zhang, X.; Xin, Q.; Sun, Y.; Zhang, P. Automatic building extraction from high-resolution aerial images and LiDAR data using gated residual refinement network. *ISPRS J. Photogramm. Remote Sens.* **2019**, *151*, 91–105. doi: 10.1016/j.isprsjprs.2019.02.019. [CrossRef]
9. Li, W.; He, C.; Fang, J.; Zheng, J.; Fu, H.; Yu, L. Semantic Segmentation-Based Building Footprint Extraction Using Very High-Resolution Satellite Images and Multi-Source GIS Data. *Remote Sens.* **2019**, *11*, 403. [CrossRef]
10. Zhang, Y.; Gong, W.; Sun, J.; Li, W. Web-Net: A Novel Nest Networks with Ultra-Hierarchical Sampling for Building Extraction from Aerial Imageries. *Remote Sens.* **2019**, *11*, 1897. [CrossRef]
11. Davari Majd, R.; Momeni, M.; Moallem, P. Transferable Object-Based Framework Based on Deep Convolutional Neural Networks for Building Extraction. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2019**, *12*, 2627–2635. [CrossRef]
12. Zhang, Z.; Wang, Y. JointNet: A Common Neural Network for Road and Building Extraction. *Remote Sens.* **2019**, *11*, 696. [CrossRef]
13. Chen, S.; Shi, W.; Zhou, M.; Zhang, M.; Chen, P. Automatic Building Extraction via Adaptive Iterative Segmentation With LiDAR Data and High Spatial Resolution Imagery Fusion. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 2081–2095. [CrossRef]
14. Erdem, F.; Avdan, U. Comparison of different U-net models for building extraction from high-resolution aerial imagery. *Int. J. Environ. GeoInform.* **2020**, *7*, 221–227. [CrossRef]
15. Milosavljevi, A. Automated Processing of Remote Sensing Imagery Using Deep Semantic Segmentation: A Building Footprint Extraction Case. *ISPRS Int. J. Geo-Inf.* **2020**, *9*, 486. [CrossRef]
16. Long, J.; Shelhamer, E.; Darrell, T. Fully Convolutional Networks for Semantic Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015.
17. 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]
18. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015, Munich, Germany, 5–9 October 2015; Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F., Eds.; Springer International Publishing: Cham, Switzerland, 2015; pp. 234–241.
19. Zheng, S.; Jayasumana, S.; Romera-Paredes, B.; Vineet, V.; Su, Z.; Du, D.; Huang, C.; Torr, P.H.S. Conditional Random Fields as Recurrent Neural Networks. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 10–13 December 2015.
20. Peng, C.; Zhang, X.; Yu, G.; Luo, G.; Sun, J. Large Kernel Matters—Improve Semantic Segmentation by Global Convolutional Network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
21. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv* **2014**, arXiv:1412.7062.
22. 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.* **2018**, *40*, 834–848. [CrossRef]
23. Chen, L.C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking atrous convolution for semantic image segmentation. *arXiv* **2017**, arXiv:1706.05587.
24. 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.

25. Wang, P.; Chen, P.; Yuan, Y.; Liu, D.; Huang, Z.; Hou, X.; Cottrell, G. Understanding Convolution for Semantic Segmentation. In Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Tahoe, NV, USA, 12–15 March 2018; pp. 1451–1460. [CrossRef]
26. 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 (CVPR), Honolulu, HI, USA, 21–26 July 2017.
27. Chen, J.; Lu, Y.; Yu, Q.; Luo, X.; Adeli, E.; Wang, Y.; Lu, L.; Yuille, A.L.; Zhou, Y. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv* **2021**, arXiv:2102.04306.
28. Zheng, S.; Lu, J.; Zhao, H.; Zhu, X.; Luo, Z.; Wang, Y.; Fu, Y.; Feng, J.; Xiang, T.; Torr, P.H.; Zhang, L. Rethinking Semantic Segmentation From a Sequence-to-Sequence Perspective With Transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 19–25 June 2021; pp. 6881–6890.
29. Strudel, R.; Garcia, R.; Laptev, I.; Schmid, C. Segmenter: Transformer for Semantic Segmentation. *arXiv* **2021**, arXiv:2105.05633.
30. Ning, F.; Delhomme, D.; LeCun, Y.; Piano, F.; Bottou, L.; Barbano, P. Toward automatic phenotyping of developing embryos from videos. *IEEE Trans. Image Process.* **2005**, *14*, 1360–1371. [CrossRef]
31. Ciresan, D.; Giusti, A.; Gambardella, L.; Schmidhuber, J. Deep neural networks segment neuronal membranes in electron microscopy images. *Adv. Neural Inf. Process. Syst.* **2012**, *25*, 2843–2851.
32. Farabet, C.; Couprie, C.; Najman, L.; LeCun, Y. Learning Hierarchical Features for Scene Labeling. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 1915–1929. [CrossRef] [PubMed]
33. Hariharan, B.; Arbeláez, P.; Girshick, R.; Malik, J. Simultaneous Detection and Segmentation. In Proceedings of the Computer Vision—ECCV 2014, Zurich, Switzerland, 6–12 September 2014; Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T., Eds.; Springer International Publishing: Cham, Switzerland, 2014; pp. 297–312.
34. Gupta, S.; Girshick, R.; Arbeláez, P.; Malik, J. Learning Rich Features from RGB-D Images for Object Detection and Segmentation. In Proceedings of the Computer Vision—ECCV 2014, Zurich, Switzerland, 6–12 September 2014; Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T., Eds.; Springer International Publishing: Cham, Switzerland, 2014; pp. 345–360.
35. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **2012**, *25*, 1097–1105. [CrossRef]
36. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 91–99. [CrossRef] [PubMed]
37. Xie, S.; Tu, Z. Holistically-nested edge detection. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 10–13 December 2015; pp. 1395–1403.
38. Li, Y.; Qi, H.; Dai, J.; Ji, X.; Wei, Y. Fully convolutional instance-aware semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2359–2367.
39. Feng, T.; Zhao, J. Review and Comparison: Building Extraction Methods Using High-Resolution Images. In Proceedings of the 2009 Second International Symposium on Information Science and Engineering, Shanghai, China, 26–28 December 2009; pp. 419–422. [CrossRef]
40. Jozdani, S.; Chen, D. On the versatility of popular and recently proposed supervised evaluation metrics for segmentation quality of remotely sensed images: An experimental case study of building extraction. *ISPRS J. Photogramm. Remote Sens.* **2020**, *160*, 275–290. doi: 10.1016/j.isprsjprs.2020.01.002. [CrossRef]
41. Bo, Z.; Chao, W.; Hong, Z.; Fan, W. A review on building extraction and Reconstruction from SAR image. *Remote Sens. Technol. Appl.* **2012**, *4*.
42. Mishra, A.; Pandey, A.; Baghel, A.S. Building detection and extraction techniques: A review. In Proceedings of the 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, India, 16–18 May 2016; pp. 3816–3821.
43. Zhu, X.X.; Tuia, D.; Mou, L.; Xia, G.S.; Zhang, L.; Xu, F.; Fraundorfer, F. Deep Learning in Remote Sensing: A Comprehensive Review and List of Resources. *IEEE Geosci. Remote Sens. Mag.* **2017**, *5*, 8–36. [CrossRef]
44. Zhang, L.; Zhang, L.; Du, B. Deep Learning for Remote Sensing Data: A Technical Tutorial on the State of the Art. *IEEE Geosci. Remote Sens. Mag.* **2016**, *4*, 22–40. [CrossRef]
45. Huang, B.; Reichman, D.; Collins, L.M.; Bradbury, K.; Malof, J.M. Dense labeling of large remote sensing imagery with convolutional neural networks: a simple and faster alternative to stitching output label maps. *arXiv* **2018**, arXiv:1805.12219.
46. Clark, R.N.; Roush, T.L. Reflectance spectroscopy: Quantitative analysis techniques for remote sensing applications. *J. Geophys. Res. Solid Earth* **1984**, *89*, 6329–6340. [CrossRef]
47. Zhou, G.; Song, C.; Simmers, J.; Cheng, P. Urban 3D GIS From LiDAR and digital aerial images. *Comput. Geosci.* **2004**, *30*, 345–353. doi: 10.1016/j.cageo.2003.08.012. [CrossRef]
48. Tang, J.; Wang, L.; Yao, Z. Analyzing urban sprawl spatial fragmentation using multi-temporal satellite images. *Giscience Remote Sens.* **2006**, *43*, 218–232. [CrossRef]
49. Wu, S.s.; Qiu, X.; Wang, L. Population estimation methods in GIS and remote sensing: A review. *Giscience Remote Sens.* **2005**, *42*, 80–96. [CrossRef]
50. Tian, J.; Cui, S.; Reinartz, P. Building Change Detection Based on Satellite Stereo Imagery and Digital Surface Models. *IEEE Trans. Geosci. Remote Sens.* **2014**, *52*, 406–417. [CrossRef]

51. Montoya-Zegarra, J.A.; Wegner, J.D.; Schindler, K. Semantic segmentation of aerial images in urban areas with class-specific higher-order cliques. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* **2015**, *2*, 127–133. [CrossRef]
52. Grinias, I.; Panagiotakis, C.; Tziritas, G. MRF-based segmentation and unsupervised classification for building and road detection in peri-urban areas of high-resolution satellite images. *ISPRS J. Photogramm. Remote Sens.* **2016**, *122*, 145–166. doi: 10.1016/j.isprsjprs.2016.10.010. [CrossRef]
53. Liu, Y.; Li, Z.; Wei, B.; Li, X.; Fu, B. Seismic vulnerability assessment at urban scale using data mining and GIScience technology: application to Urumqi (China). *Geomat. Nat. Hazards Risk* **2019**, *10*, 958–985. [CrossRef]
54. Li, X.; Li, Z.; Yang, J.; Liu, Y.; Fu, B.; Qi, W.; Fan, X. Spatiotemporal characteristics of earthquake disaster losses in China from 1993 to 2016. *Nat. Hazards* **2018**, *94*, 843–865. [CrossRef]
55. Zhang, B.; Chen, Z.; Peng, D.; Benediktsson, J.A.; Liu, B.; Zou, L.; Li, J.; Plaza, A. Remotely sensed big data: Evolution in model development for information extraction [point of view]. *Proc. IEEE* **2019**, *107*, 2294–2301. [CrossRef]
56. Liu, Y.; So, E.; Li, Z.; Su, G.; Gross, L.; Li, X.; Qi, W.; Yang, F.; Fu, B.; Yalikun, A.; et al. Scenario-based seismic vulnerability and hazard analyses to help direct disaster risk reduction in rural Weinan, China. *Int. J. Disaster Risk Reduct.* **2020**, *48*, 101577. [CrossRef]
57. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
58. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.E.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going Deeper with Convolutions. *arXiv* **2014**, arXiv:1409.4842.
59. Ioffe, S.; Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In Proceedings of the 32nd International Conference on Machine Learning, Lille, France, 7–9 July 2015; Bach, F., Blei, D., Eds.; PMLR: Lille, France, 2015; Volume 37, pp. 448–456.
60. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the Inception Architecture for Computer Vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26–30 June 2016.
61. Szegedy, C.; Ioffe, S.; Vanhoucke, V.; Alemi, A.A. Inception-v4, inception-resnet and the impact of residual connections on learning. In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017.
62. 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 (CVPR), Las Vegas, NV, USA, 26–30 June 2016.
63. Ahmed, A.; Yu, K.; Xu, W.; Gong, Y.; Xing, E. Training Hierarchical Feed-Forward Visual Recognition Models Using Transfer Learning from Pseudo-Tasks. In Proceedings of the Computer Vision—ECCV 2008, Marseille, France, 12–18 October 2008; Forsyth, D.; Torr, P.; Zisserman, A., Eds.; Springer: Berlin/Heidelberg, Germany, 2008; pp. 69–82.
64. Oquab, M.; Bottou, L.; Laptev, I.; Sivic, J. Learning and Transferring Mid-Level Image Representations using Convolutional Neural Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 23–28 June 2014.
65. Yosinski, J.; Clune, J.; Bengio, Y.; Lipson, H. How transferable are features in deep neural networks? *arXiv* **2014**, arXiv:1411.1792.
66. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. ImageNet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255. [CrossRef]
67. 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]
68. Kaiser, P.; Wegner, J.D.; Lucchi, A.; Jaggi, M.; Hofmann, T.; Schindler, K. Learning Aerial Image Segmentation From Online Maps. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 6054–6068. [CrossRef]
69. Kampffmeyer, M.; Salberg, A.B.; Jenssen, R. Semantic Segmentation of Small Objects and Modeling of Uncertainty in Urban Remote Sensing Images Using Deep Convolutional Neural Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, Las Vegas, NV, USA, 26 June–1 July 2016.
70. Eigen, D.; Fergus, R. Predicting Depth, Surface Normals and Semantic Labels With a Common Multi-Scale Convolutional Architecture. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Sydney, Australia, 1–8 December 2016.
71. Li, Q.; Shi, Y.; Huang, X.; Zhu, X.X. Building Footprint Generation by Integrating Convolution Neural Network With Feature Pairwise Conditional Random Field (FPCRF). *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 7502–7519. [CrossRef]
72. Hosseinpoor, H.; Samadzadegan, F. Convolutional Neural Network for Building Extraction from High-Resolution Remote Sensing Images. In Proceedings of the 2020 International Conference on Machine Vision and Image Processing (MVIP), Qom, Iran, 18–20 February 2020; pp. 1–5. [CrossRef]
73. Yu, Y.; Ren, Y.; Guan, H.; Li, D.; Yu, C.; Jin, S.; Wang, L. Capsule feature pyramid network for building footprint extraction from high-resolution aerial imagery. *IEEE Geosci. Remote Sens. Lett.* **2020**, *18*, 895–899. [CrossRef]
74. Deng, W.; Shi, Q.; Li, J. Attention-Gate-Based Encoder–Decoder Network for Automatical Building Extraction. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 2611–2620. [CrossRef]
75. Guo, H.; Shi, Q.; Du, B.; Zhang, L.; Wang, D.; Ding, H. Scene-driven multitask parallel attention network for building extraction in high-resolution remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 4287–4306. [CrossRef]
76. Zhu, Q.; Liao, C.; Hu, H.; Mei, X.; Li, H. MAP-Net: Multiple Attending Path Neural Network for Building Footprint Extraction From Remote Sensed Imagery. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 6169–6181. [CrossRef]

77. Hu, Q.; Zhen, L.; Mao, Y.; Zhou, X.; Zhou, G. Automated building extraction using satellite remote sensing imagery. *Autom. Constr.* **2021**, *123*, 103509. [CrossRef]
78. Mnih, V. Machine Learning for Aerial Image Labeling. Ph.D. Thesis, University of Toronto, Toronto, ON, Canada, 2013.
79. Maggiori, E.; Tarabalka, Y.; Charpiat, G.; Alliez, P. Can semantic labeling methods generalize to any city? The inria aerial image labeling benchmark. In Proceedings of the 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Worth, TX, USA, 23–28 June 2017; pp. 3226–3229.
80. 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.* **2019**, *57*, 574–586. [CrossRef]
81. Saito, S.; Yamashita, T.; Aoki, Y. Multiple object extraction from aerial imagery with convolutional neural networks. *Electron. Imaging* **2016**, *2016*, 1–9. [CrossRef]
82. Alshehhi, R.; Marpu, P.R.; Woon, W.L.; Mura, M.D. Simultaneous extraction of roads and buildings in remote sensing imagery with convolutional neural networks. *ISPRS J. Photogramm. Remote Sens.* **2017**, *130*, 139–149. [CrossRef]
83. Maggiori, E.; Tarabalka, Y.; Charpiat, G.; Alliez, P. Fully convolutional neural networks for remote sensing image classification. In Proceedings of the 2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Beijing, China, 10–16 June 2016; pp. 5071–5074. [CrossRef]
84. Yuan, J. Automatic Building Extraction in Aerial Scenes Using Convolutional Networks. *arXiv* **2016**, arXiv:1602.06564.
85. Bittner, K.; Adam, F.; Cui, S.; K?rner, M.; Reinartz, P. Building Footprint Extraction From VHR Remote Sensing Images Combined With Normalized DSMs Using Fused Fully Convolutional Networks. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2018**, *11*, 2615–2629. [CrossRef]
86. Yang, H.L.; Lunga, D.; Yuan, J. Toward country scale building detection with convolutional neural network using aerial images. In Proceedings of the 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Worth, TX, USA, 23–28 June 2017; pp. 870–873. [CrossRef]
87. Wu, G.; Shao, X.; Guo, Z.; Chen, Q.; Yuan, W.; Shi, X.; Xu, Y.; Shibasaki, R. Automatic building segmentation of aerial imagery using multi-constraint fully convolutional networks. *Remote Sens.* **2018**, *10*, 407. [CrossRef]
88. Li, L.; Liang, J.; Weng, M.; Zhu, H. A multiple-feature reuse network to extract buildings from remote sensing imagery. *Remote Sens.* **2018**, *10*, 1350. [CrossRef]
89. Wu, G.; Guo, Z.; Shi, X.; Chen, Q.; Xu, Y.; Shibasaki, R.; Shao, X. A boundary regulated network for accurate roof segmentation and outline extraction. *Remote Sens.* **2018**, *10*, 1195. [CrossRef]
90. Ma, J.; Wu, L.; Tang, X.; Liu, F.; Zhang, X.; Jiao, L. Building Extraction of Aerial Images by a Global and Multi-Scale Encoder-Decoder Network. *Remote Sens.* **2020**, *12*, 2350. [CrossRef]
91. Wang, S.; Hou, X.; Zhao, X. Automatic Building Extraction From High-Resolution Aerial Imagery via Fully Convolutional Encoder-Decoder Network With Non-Local Block. *IEEE Access* **2020**, *8*, 7313–7322. [CrossRef]
92. Zhou, D.; Wang, G.; He, G.; Long, T.; Yin, R.; Zhang, Z.; Chen, S.; Luo, B. Robust Building Extraction for High Spatial Resolution Remote Sensing Images with Self-Attention Network. *Sensors* **2020**, *20*, 7241. [CrossRef]
93. Shrestha, S.; Vanneschi, L. Improved Fully Convolutional Network with Conditional Random Fields for Building Extraction. *Remote Sens.* **2018**, *10*, 1135. [CrossRef]
94. Maggiori, E.; Charpiat, G.; Tarabalka, Y.; Alliez, P. Recurrent Neural Networks to Correct Satellite Image Classification Maps. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 4962–4971. [CrossRef]
95. Yang, G.; Zhang, Q.; Zhang, G. EANet: Edge-Aware Network for the Extraction of Buildings from Aerial Images. *Remote Sens.* **2020**, *12*, 2161. [CrossRef]
96. Zhang, Y.; Li, W.; Gong, W.; Wang, Z.; Sun, J. An Improved Boundary-Aware Perceptual Loss for Building Extraction from VHR Images. *Remote Sens.* **2020**, *12*, 1195. [CrossRef]
97. Shao, Z.; Tang, P.; Wang, Z.; Saleem, N.; Yam, S.; Sommai, C. BRRNet: A Fully Convolutional Neural Network for Automatic Building Extraction From High-Resolution Remote Sensing Images. *Remote Sens.* **2020**, *12*, 1050. [CrossRef]
98. Shi, Y.; Li, Q.; Zhu, X.X. Building segmentation through a gated graph convolutional neural network with deep structured feature embedding. *ISPRS J. Photogramm. Remote Sens.* **2020**, *159*, 184–197. doi: 10.1016/j.isprsjprs.2019.11.004. [CrossRef]
99. Hamaguchi, R.; Fujita, A.; Nemoto, K.; Imaizumi, T.; Hikosaka, S. Effective Use of Dilated Convolutions for Segmenting Small Object Instances in Remote Sensing Imagery. In Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Tahoe, NV, USA, 12–15 March 2018; pp. 1442–1450. [CrossRef]
100. Kang, W.; Xiang, Y.; Wang, F.; You, H. EU-Net: An Efficient Fully Convolutional Network for Building Extraction from Optical Remote Sensing Images. *Remote Sens.* **2019**, *11*, 2813. [CrossRef]
101. Liu, Y.; Fan, B.; Wang, L.; Bai, J.; Xiang, S.; Pan, C. Semantic labeling in very high resolution images via a self-cascaded convolutional neural network. *ISPRS J. Photogramm. Remote Sens.* **2018**, *145*, 78–95.10.1016/j.isprsjprs.2017.12.007. [CrossRef]
102. Ji, S.; Wei, S.; Lu, M. A scale robust convolutional neural network for automatic building extraction from aerial and satellite imagery. *Int. J. Remote Sens.* **2019**, *40*, 3308–3322. [CrossRef]
103. Li, X.; Yao, X.; Fang, Y. Building-A-Nets: Robust Building Extraction From High-Resolution Remote Sensing Images With Adversarial Networks. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2018**, *11*, 3680–3687. [CrossRef]
104. Ding, Y.; Wu, M.; Xu, Y.; Duan, S. P-linknet: Linknet with spatial pyramid pooling for high-resolution satellite imagery. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2020**, *43*, 35–40. [CrossRef]

105. Wei, S.; Ji, S.; Lu, M. Toward Automatic Building Footprint Delineation From Aerial Images Using CNN and Regularization. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 2178–2189. [CrossRef]
106. Pan, X.; Yang, F.; Gao, L.; Chen, Z.; Zhang, B.; Fan, H.; Ren, J. Building Extraction from High-Resolution Aerial Imagery Using a Generative Adversarial Network with Spatial and Channel Attention Mechanisms. *Remote Sens.* **2019**, *11*, 917. [CrossRef]
107. He, N.; Fang, L.; Plaza, A. Hybrid first and second order attention Unet for building segmentation in remote sensing images. *Sci. China Inf. Sci.* **2020**, *63*, 1–12. [CrossRef]
108. Lin, J.; Jing, W.; Song, H.; Chen, G. ESFNet: Efficient Network for Building Extraction From High-Resolution Aerial Images. *IEEE Access* **2019**, *7*, 54285–54294. [CrossRef]
109. Liu, Y.; Zhou, J.; Qi, W.; Li, X.; Gross, L.; Shao, Q.; Zhao, Z.; Ni, L.; Fan, X.; Li, Z. ARC-Net: An Efficient Network for Building Extraction From High-Resolution Aerial Images. *IEEE Access* **2020**, *8*, 154997–155010. [CrossRef]
110. Yang, H.L.; Yuan, J.; Lunga, D.; Laverdiere, M.; Rose, A.; Bhaduri, B. Building Extraction at Scale Using Convolutional Neural Network: Mapping of the United States. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2018**, *11*, 2600–2614. [CrossRef]
111. Liu, W.; Yang, M.; Xie, M.; Guo, Z.; Li, E.; Zhang, L.; Pei, T.; Wang, D. Accurate Building Extraction from Fused DSM and UAV Images Using a Chain Fully Convolutional Neural Network. *Remote Sens.* **2019**, *11*, 2912. [CrossRef]
112. Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-Local Neural Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018.
113. Zhu, Z.; Xu, M.; Bai, S.; Huang, T.; Bai, X. Asymmetric Non-Local Neural Networks for Semantic Segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019.
114. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid Scene Parsing Network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
115. You, J.; Leskovec, J.; He, K.; Xie, S. Graph Structure of Neural Networks. In Proceedings of the 37th International Conference on Machine Learning, Virtual, 13–18 July 2020; Daumé, H., III, Singh, A., Eds.; PMLR: Lille, France, 2020; Volume 119, pp. 10881–10891.
116. Zhang, L.; Li, X.; Arnab, A.; Yang, K.; Tong, Y.; Torr, P.H. Dual graph convolutional network for semantic segmentation. *arXiv* **2019**, arXiv:1909.06121.
117. Li, X.; Yang, Y.; Zhao, Q.; Shen, T.; Lin, Z.; Liu, H. Spatial Pyramid Based Graph Reasoning for Semantic Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020.
118. Zhang, L.; Xu, D.; Arnab, A.; Torr, P.H. Dynamic Graph Message Passing Networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020.
119. Hu, H.; Cui, J.; Zha, H. Boundary-aware Graph Convolution for Semantic Segmentation. In Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 10–15 January 2021; pp. 1828–1835. [CrossRef]
120. Yu, F.; Koltun, V. Multi-scale context aggregation by dilated convolutions. *arXiv* **2015**, arXiv:1511.07122.
121. Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; Torralba, A. Object detectors emerge in deep scene cnns. *arXiv* **2014**, arXiv:1412.6856.
122. Anwar, S.; Hwang, K.; Sung, W. Structured Pruning of Deep Convolutional Neural Networks. *ACM J. Emerg. Technol. Comput. Syst.* **2015**, *13*. [CrossRef]
123. Han, S.; Mao, H.; Dally, W.J. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *arXiv* **2015**, arXiv:1510.00149.
124. Molchanov, P.; Tyree, S.; Karras, T.; Aila, T.; Kautz, J. Pruning convolutional neural networks for resource efficient transfer learning. *arXiv* **2016**, arXiv:1611.06440.
125. Paszke, A.; Chaurasia, A.; Kim, S.; Culurciello, E. Enet: A deep neural network architecture for real-time semantic segmentation. *arXiv* **2016**, arXiv:1606.02147.
126. Romera, E.; Alvarez, J.M.; Bergasa, L.M.; Arroyo, R. Erfnet: Efficient residual factorized convnet for real-time semantic segmentation. *IEEE Trans. Intell. Transp. Syst.* **2017**, *19*, 263–272. [CrossRef]
127. Lo, S.Y.; Hang, H.M.; Chan, S.W.; Lin, J.J. Efficient dense modules of asymmetric convolution for real-time semantic segmentation. In Proceedings of the ACM Multimedia Asia, Beijing, China, 15–18 December 2019; pp. 1–6.
128. Ding, X.; Xia, C.; Zhang, X.; Chu, X.; Han, J.; Ding, G. Repmlp: Re-parameterizing convolutions into fully-connected layers for image recognition. *arXiv* **2021**, arXiv:2105.01883.
129. Ding, X.; Zhang, X.; Ma, N.; Han, J.; Ding, G.; Sun, J. Repvgg: Making vgg-style convnets great again. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 13733–13742.