remote sensing

MDPI

Article

Hierarchical Disentangling Network for Building Extraction from Very High Resolution Optical Remote Sensing Imagery

Jianhao Li [1], Yin Zhuang [1], Shan Dong [1,2,*], Peng Gao [3], Hao Dong [4], He Chen [1], Liang Chen [1] and Lianlin Li [4]

[1] Beijing Key Laboratory of Embedded Real-Time Information Processing Technology, Beijing Institute of Technology, Beijing 100081, China; 3220200570@bit.edu.cn (J.L.); yzhuang@bit.edu.cn (Y.Z.); chenhe@bit.edu.cn (H.C.); chenl@bit.edu.cn (L.C.)
[2] State Key Laboratory of Media Convergence and Communication, Communication University of China, Beijing 100024, China
[3] Shanghai AI Laboratory, Shanghai 200232, China; gaopeng@pjlab.org.cn
[4] Center on Frontiers of Computing Studies and School of Electronic Engineering and Computer Science, Peking University, Beijing 100087, China; hao.dong@pku.edu.cn (H.D.); lianlin.li@pku.edu.cn (L.L.)
* Correspondence: dongshan@cuc.edu.cn

Abstract: Building extraction using very high resolution (VHR) optical remote sensing imagery is an essential interpretation task that impacts human life. However, buildings in different environments exhibit various scales, complicated spatial distributions, and different imaging conditions. Additionally, with the spatial resolution of images increasing, there are diverse interior details and redundant context information present in building and background areas. Thus, the above-mentioned situations would create large intra-class variances and poor inter-class discrimination, leading to uncertain feature descriptions for building extraction, which would result in over- or under-extraction phenomena. In this article, a novel hierarchical disentangling network with an encoder–decoder architecture called HDNet is proposed to consider both the stable and uncertain feature description in a convolution neural network (CNN). Next, a hierarchical disentangling strategy is set up to individually generate strong and weak semantic zones using a newly designed feature disentangling module (FDM). Here, the strong and weak semantic zones set up the stable and uncertain description individually to determine a more stable semantic main body and uncertain semantic boundary of buildings. Next, a dual-stream semantic feature description is built to gradually integrate strong and weak semantic zones by the designed component feature fusion module (CFFM), which is able to generate a powerful semantic description for more complete and refined building extraction. Finally, extensive experiments are carried out on three published datasets (i.e., WHU satellite, WHU aerial, and INRIA), and the comparison results show that the proposed HDNet outperforms other state-of-the-art (SOTA) methods.

Keywords: building extraction; convolution neural networks; encoding–decoding method; hierarchical disentangling; optical remote sensing imagery; very high resolution

Citation: Li, J.; Zhuang, Y.; Dong, S.; Gao, P.; Dong, H.; Chen, H.; Chen, L.; Li, L. Hierarchical Disentangling Network for Building Extraction from Very High Resolution Optical Remote Sensing Imagery. *Remote Sens.* **2022**, *14*, 1767. https://doi.org/10.3390/rs14071767

Academic Editors: Tais Grippa, Lei Ma, Claudio Persello and Arnaud Le Bris

Received: 25 February 2022
Accepted: 4 April 2022
Published: 6 April 2022

1. Introduction

Building extraction plays an important role in several applications, such as in urban planning, illegal building supervision, and geographical information surveying, mapping, and updating [1–4], and these applications are significant to societal development and daily human life. With the rapid development of VHR optical remote sensing imaging, a lot of optical remote sensing data are available and can be used for building investigations. Yet, when facing amounts of optical remote sensing data, the manual investigation of buildings is arduous. It is necessary to develop an automatic and powerful building extraction algorithm to eliminate this time-consuming labor. However, referring to several published building extraction datasets [5–7], several statistical and visual results are shown in Figure 1, and there are three main aspects (e.g., buildings in various environments

showing great differences in scale, spatial distribution, and imaging conditions) affecting the refined building extraction results. First, in Figure 1a, the horizontal axis represents the proportion of different-scale buildings from datasets, and the vertical axis indicates the scale distribution of each counted building. From the statistical and visual results in Figure 1a, we can see that a large-scale spanning of buildings would create challenges when setting up a robustness feature description to carefully extract the full scale of the buildings. Therefore, these circumstances often result in the incomplete extraction of large-scale buildings and missing extraction for small-scale buildings. Next, in addition to large-scale spanning, the arbitrary spatial distribution of buildings is also a challenging situation that impacts building extraction performance, which is shown in Figure 1b. Here, the horizontal axis represents the different spatial distributions of buildings according to the initial datasets, and the vertical axis represents the number of buildings per image. Subsequently, for densely distributed buildings, despite each building being able to provide strong spatial relevance to support each other during building extraction, this makes it hard to provide satisfactory results with precise boundary information due to the smaller intervals of the densely distributed buildings. Then, when buildings have a sparse and isolated spatial distribution, the redundant background context information emerges in large quantities, and there is a big risk of missing extraction or false alarms due to the uncertain semantic feature description for building extraction. Moreover, as presented in Figure 1c–e, there are three typical cases of shadows, occlusions and low contrast ratio to make a challenge for effective semantic feature description and refined building extraction. Moreover, in VHR optical remote sensing imagery, the interior details of buildings and the redundant context of the surrounding background become clearer, which can further enlarge the intra-class variance and worsen inter-class discrimination for building extraction from complex scenes. Thus, when jointly considering these mentioned challenges, traditional hand-crafted feature-based methods [8–16] (e.g., spectrum [8], shape [9], color [10], texture [11], polygon [12,13], and shadow [14,15]) demonstrate poor building description and extraction results. In recent years, because of the booming development of deep learning techniques, a lot of convolution neural networks (CNNs) based encoder–decoder architectures have been designed in the computer vision field, such as FCN [17], SegNet [18], GCN [19], PSPnet [20], Unet series [21–23], and Deeplab series [24–27]. Since these networks can provide more competitive results than traditional methods [8–16], they are widely used and studied for automatic building extraction tasks in optical remote sensing scenes.

For example, for environmental riverbank protection in Chongqing, China, W. Boonpook et al. [28] utilized SegNet to achieve automatic building extraction from unmanned aerial vehicle (UAV)-based images. J. Liu et al. [29] referred to Unet and combined the residual block into a simplified U-shape encoder–decoder to achieve building extraction for urban planning. G. Wu et al. [30] utilized a designed multi-scale supervision strategy for FCN training and set up a multi-constraint fully convolutional network (MC-FCN) to generate good results for varying scales of building extraction. Tremendous efforts have demonstrated that CNN-based encoder–decoder methods can generate a stable pixel-level semantic feature description for building extraction. However, when encountering the challenge cases in Figure 1, there are no good enough results that can be generated from VHR optical remote sensing imagery. Consequently, a lot of studies have been dedicated to improving the semantic feature description ability of these methods to enhance building extraction performance. Y. Liu et al. [31] and W. Kang et al. [32] introduced the spatial pyramid pooling (SPP) module into the encoder–decoder of Unet that was able to further enlarge the receptive fields (RFs) and aggregate multi-scale contextual information for building description. Then, L. Hao et al. [33] employed inception downsampling modules in the encoding procedure, and the dense upsampling convolution (DUC) was adopted instead of bilinear interpolation or de-convolution to strengthen both information preservation and discrimination for refined building extraction. Next, Y. Liu et al. [34] took the irregular shapes of buildings into account and adopted the U-shape encoder–decoder method combined with asymmetric convolutions, providing more appropriate feature extraction.

Moreover, atrous convolutions with different dilation rates were also employed during deep layer feature extraction, enhancing the generalization ability of semantic description. Furthermore, referring to the Deeplab series [24–27], J. Cai et al. [35] adopted DUC for a decoder and designed multipath hybrid-dilated convolution to aggregate multiscale context features in the encoder, obtaining acceptable building extraction results. Subsequently, in view of a powerful encoder–decoder architecture-based semantic description method and to alleviate losses in the shape or boundary information caused by stacked convolutions and to achieve refined building delineation, Y. Yu et al. [36] designed the capsule feature pyramid network (CapFPN), in which capsule layers are used as the encoder to preserve the information that may be lost by downsampling consecutive convolutions. Then, multi-scale capsule feature fusing is used as the decoder to generate more refined building delineation results. Q. Hu et al. [37] followed the Unet architecture and utilized an attention mechanism to set up an attention gate to facilitate feature fusion for the skip connection in the encoder–decoder, and then the squeeze-and-extraction operation was adopted to leverage the semantic feature generalization procedure for refined building delineation using a residual recurrent convolutional neural network. A. Abdollahi et al. [38] proposed a Multi-Res-Unet to optimize the skip connections in the Unet architecture and to assimilate the features learned from the same scale level of the encoder and decoder. Moreover, they also integrated semantic edge information with semantic polygons to retain the irregular shape and refined boundary information. H. Ye et al. [39] also focused on the feature fusion of the skip connections in U-shaped encoder–decoders, and they utilized the feature reuse mechanism to refine the fused features by introducing more reasonable low-level spatial information to alleviate fuzzy building boundary prediction to some extent. Furthermore, in addition to studies [36–39] achieving building extraction from data with a more refined shape or boundary, some research has considered building extraction tasks in which there is a great amount of variation in the building size. R. Hamaguchi et al. [40] designed size-specific prediction branches for small-, medium-, and large-scale building extraction tasks followed by an Unet structure, and then they proposed a post-processing technique for integrating the output probability maps from each scale of branches. Similar to [40], H. Guo et al. [41] proposed a scale-robust deep-supervision network for building extraction by designing a scale attention module combined with multi-scale deep supervision to facilitate multi-scale decision fusion at the decoder. Then, S. Ji et al. [5] and Y. Liao et al. [42] both utilized the dual-stream of a Siamese U-net architecture to individually model the global and local information and finally fused them for building extraction tasks at different scales. Next, E. Maggiori et al. [6], P. Liu et al. [43], S. Wei et al. [44], Q. Zhu et al. [45], and F. Chen et al. [46] all designed specific multi-scale feature extraction and fusion networks for building extraction, and [44,45] also incorporated polygon regularization post-processing or spatial and channel attention mechanisms to refine the building extraction results at a wide range of scales. In general, although substantial efforts [5,6,28–46] have been made to refine building extraction, almost all of these methods focus on trading off the classification and localization performance by individually optimizing and fusing deep- and shallow-layer features. Unfortunately, there is antinomy in the feature fusion of the deep and shallow layers. As the receptive field continues to increase due to stacked convolutions, it inevitably introduces optimized deep layer features that can improve classification ability while worsening the localization ability. On the other hand, it also inevitably introduces optimized shallow layer features to improve the localization ability while degenerating classification ability. According to this antinomy in the deep- and shallow-layer feature fusion, the above-mentioned methods often result in limitations and uncertain predictions during building extraction. Therefore, there is still much room for improvement in refined building extraction using VHR optical remote sensing imagery.

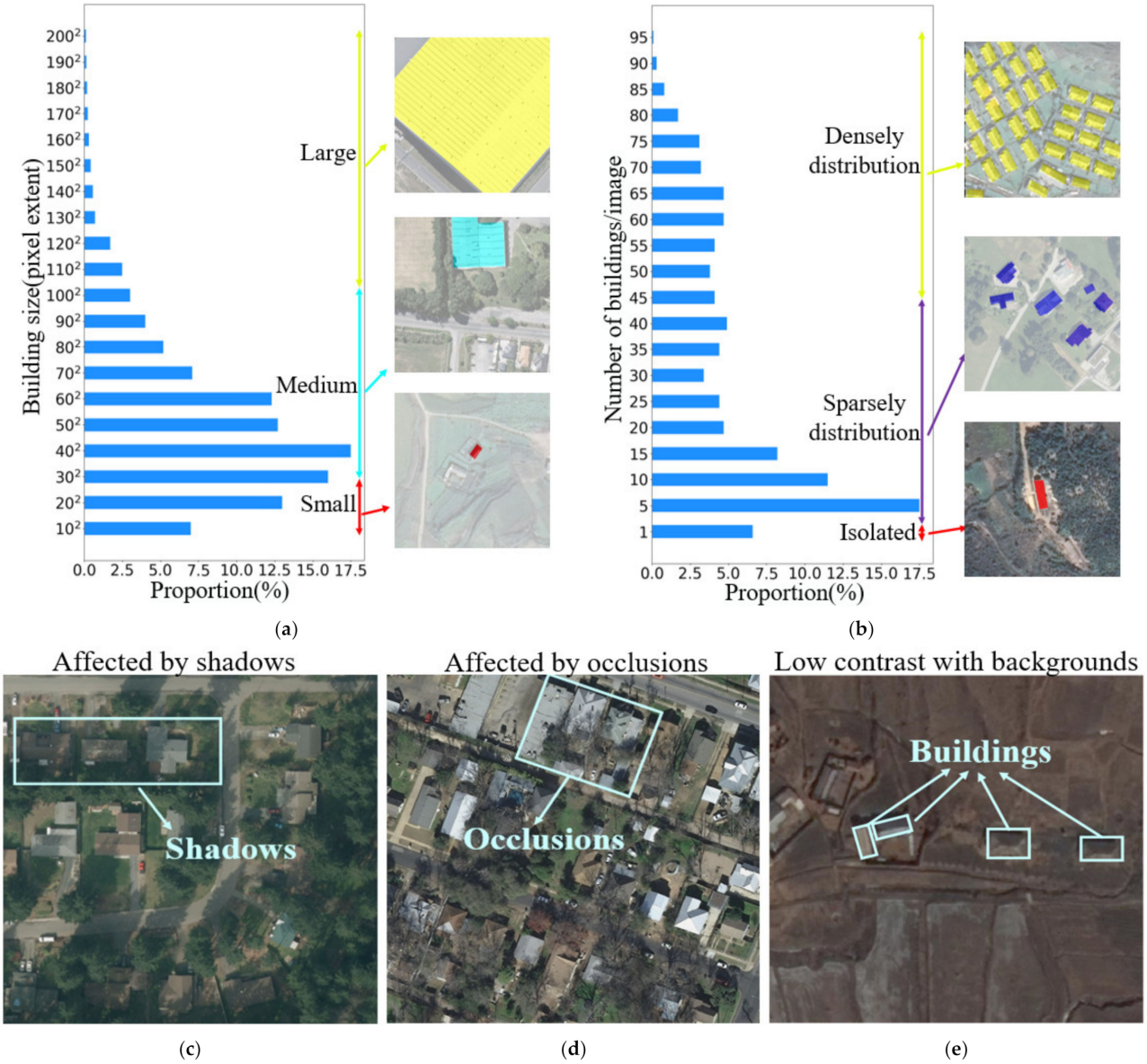


Figure 1. Statistical charts of the size and number buildings and examples of buildings with different imaging conditions. (**a**) is the distribution of building size; (**b**) is the distribution of the number of buildings in each image; (**c**) is building extraction affected by shadows; (**d**) is building extraction affected by occlusions; (**e**) is a building extraction scenario with low contrast between the buildings and background. All statistics and examples are taken from the three datasets (i.e., WHU satellite, WHU aerial, and INRIA) that we used in this work.

In this article, we consider the existing challenges in refined building extraction and rethink the current semantic description paradigm. A novel hierarchical disentangling network called HDNet is proposed for refined building extraction. Different from previous encoder–decoder architectures that trade off the classification and localization performance by fusing specific deep and shallow layer features, the proposed HDNet sets up semantic description using a hierarchical disentangling strategy and dual-stream semantic feature description that focuses on individually describing the strong and weak semantic zones and achieving effective feature fusion. Then, the newly designed feature disentangling module (FDM) is used for the hierarchical disentangling strategy to decompose the hierarchical features into strong semantic zones with a more stable semantic main body and weak semantic zone with uncertain semantic boundaries among the stacked convolution

layers. Here, the more stable semantic main body and uncertain semantic boundary can be considered to be more effective feature descriptions for the main body and boundary information of buildings. Next, after the hierarchical feature disentangling strategy, the newly designed component feature fusion module (CFFM) is used to gradually integrate a more stable semantic main body and uncertain semantic boundary followed by a dual-stream semantic feature description. After dual-stream semantic feature description, the strong and weak semantic zones can be elegantly fused to generate powerful semantic descriptions for refined building extraction. Finally, extensive experiments are carried out on three published datasets (i.e., WHU satellite [5], WHU aerial [5], and INRIA [6]) and the comparison results show that when using an effective and new semantic description paradigm for building extraction, the proposed HDNet not only outperforms the baseline of Deeplabv3 [26], but also has more superior performance than Deeplabv3+ [27] regardless of whether satellite or aerial data are used. Meanwhile, the proposed HDNet can provide more competitive building extraction results than the other state-of-the-art (SOTA) building extraction methods. In general, the main contributions of our study can be summarized as follows:

- A novel hierarchical disentangling strategy and effective dual-stream semantic feature description method are proposed for refined building extraction using VHR optical remote sensing imagery that elaborates on the defects of previous semantic feature description methods and provides a novel semantic feature description paradigm.
- A novel FDM is designed for the hierarchical disentangling strategy in the proposed HDNet that can effectively decompose each feature into an uncertain semantic boundary for the weak semantic zone and a more stable semantic main body for the strong semantic zone.
- An effective CFFM is designed for dual-stream semantic feature description in the proposed HDNet that can achieve better feature fusion for both strong and weak semantic zones and can further facilitate final fine-scale semantic feature generation.
- A powerful convolution network called HDNet is set up for refined building extraction, and extensive experiments demonstrate that the proposed HDNet method can provide competitive building extraction results for both satellite and aerial datasets. Thus, it can also be widely used for automatic building extraction applications, such as building change detection, urban area investigations, illegal building supervision, and so on.

The rest of this article is organized as follows: Section 2 introduces works related to this research; Section 3 elaborates on the HDNet in detail; Section 4 reports the extensive experiments. Finally, the discussion and conclusions are presented in Sections 5 and 6, respectively.

2. Related Works

2.1. Building Extraction from VHR Optical Remote Sensing Imagery

In recent years, thanks to the development of CNN-based semantic segmentation, a lot of CNN building extraction methods have emerged that can be roughly divided into two categories: object-based [47–49] and pixel-based [5,6,28–46,50–61] methods. First, object-based methods mainly rely on CNN object detectors. Y. Xiong et al. [47] employed ResNet-50 as the backbone of the one-stage object detector Yolov-2 to achieve building detection using several specific bounding boxes. Y. Liu et al. [48] constructed a multi-scale U-shape structure combined with a region proposal network (RPN) to achieve building instance extraction by means of multi-task supervision. H. Guo et al. [49] designed an object-based scene classification branch using prior information to leverage various types of building extraction. Second, in terms of pixel-based methods, apart from the methods mentioned in Section 1, there are other aspects of studies focusing on refined building extraction. X. Pan et al. [50] introduced spatial and channel attention mechanisms into the Unet architecture, which then had a generative adversarial network (GAN) incorporated into it to generate refined building extraction results. A. Abdollahi et al. [51] also utilized the GAN to facilitate SegNet with a bidirectional convolutional LSTM to generate more

refined building extraction results. Q. Li et al. [52] integrated a CNN and graph model together to implement the pairwise conditional random field for the features that were able to preserve sharp boundaries and fine-grained building segmentation results. N. Girard et al. [53] added frame field output to a CNN-based segmentation model to leverage precise polygon generation for building extraction. W. Li et al. [54] proposed a multi-task segmentation network for pixel-wise building extraction via joint semantic and geometric learning (e.g., multi-class corner and edge orientation predictions). In addition, [54] also proposed a polygon refinement network to generate refined building extraction results. As mentioned above, object- and pixel- based methods both are widely used in automatic building extraction tasks. However, since object-based methods have a large risk of not detecting buildings at the RPN stage, this results in a poor recall metric during building extraction. Thus, most of the current research mainly focuses on pixel-based methods, but in the face of various challenges, the semantic feature description of these methods is not robust enough to solve uncertain building extraction problems in the main body and boundaries of buildings in complex scenes. Thus, in this article, we also focus on exploring a more effective and novel pixel-level semantic description paradigm to improve building extraction performance without employing post-processing.

2.2. Boundary-Aware Semantic Segmentation

Recently, the boundary-aware semantic segmentation paradigm has started to attract more and more attention in the semantic segmentation field. T. Takikawa et al. [62] introduced dense image representation, where color, shape, and texture information are all processed together in a deep CNN. However, they contain a very different type of information that is relevant for pixel-wise classification. Thus, [62] proposed a two-stream CNN architecture for semantic segmentation that deals with the boundary shape and classification streams separately for semantic segmentation tasks in natural environments. H. Ma et al. [63] also indicated that boundary information is very important for semantic segmentation and can supplement semantic details. Then, to better leverage boundary information and fuse it with mainstream features, [63] proposed a two-stream boundary-guided context aggregation network for semantic segmentation tasks in natural environments. Moreover, H. He et al. [64] proposed a new bi-directional flow-based warping process to squeeze the object boundary from the stable main body and then designed a specific boundary loss label for supervision that is able to obtain accurate mask representation for both instance and semantic segmentation tasks. Next, for optical remote sensing semantic segmentation tasks, Y. Wang et al. [65] designed a boundary-aware multitask learning framework to perform semantic segmentation, height estimation, and boundary detection within a unified model, and a boundary attentive module was proposed to build cross-task interaction for master tasks, enabling their network to generate more refined segmentation results. A. Li et al. [66] also proposed a semantic boundary awareness network for land cover classification and a boundary attention module to capture refined sharp information from hierarchical feature aggregation. Furthermore, in the refined building extraction works from VHR optical remote sensing, H. Yin et al. [55] and C. Liao et al. [56] both introduced boundary information with joint learning during building ground truth (GT) to improve precise boundary generation during building extraction. Y. Zhu et al. [57] proposed two cascaded networks (e.g., Edge-Net and Detail-Net) for automatic building extraction from VHR aerial images. Here, in [57], the Edge-Net network that was designed captured and preserved boundary information for building detection, and then the Detail-Net network was designed to refine the results obtained by Edge-Net. X. Jiang et al. [58] also proposed two stage-refined building extraction frameworks based on encoder–decoder prediction and a residual boundary refinement module. K. Lee et al. [59] designed a new metric called the boundary-oriented intersection over union, which was used for boundary-oriented losses to generate precise building extraction results followed by a two-stage extractor. Following the Deeplabv3+ framework, A. Jiwani et al. [60] designed an F-Beta measure to balance the relationship between the boundaries and GT to generate refined building

extraction results. In addition, facing the lack of consideration of semantic boundaries, Y. Jin et al. [67] proposed a novel network embedded with a special boundary-aware loss to alleviate the huge uncertainty predictions at the building boundaries. In summary, in recent years, researchers have gradually realized that boundary-aware information is very important for the semantic segmentation field. Specifically, in automatic building extraction applications using VHR optical remote sensing imagery, accurate boundary and consistent building area predictions can help to generate the expected results. Although existing methods mainly focus on learning boundary-aware information using a specific loss function, they lack the specific consideration of the semantic features to solve the challenge of inaccurate boundary information description in complex scenarios. In this case, it is often difficult to obtain satisfactory boundaries, and this should continue to be explored.

2.3. Effective Multi-Scale Feature Fusion

Effective feature fusion is a core topic in computer vision tasks and has been widely studied in a lot of fields [68–70]. Initially, feature map concatenation is a simple feature fusion method in CNNs that is often seen in various multi-scale feature fusion modules. For example, P. Liu et al. [43] designed a spatial residual inception module to aggregate multi-scale contexts for building extraction by successively fusing multi-level features. S. Wei et al. [44] concatenated multi-scale decoding features to achieve multi-scale descriptions for various scale buildings. H. Yin et al. [55] utilized concatenation operations to fuse boundary and semantic branch features for refined building extraction. Then, due to the wide and successful application of an attention mechanism in CNN networks [70,71], more effective feature fusion methods have emerged. Q. Zhu et al. [45] and C. Liao et al. [56] both adopted the squeeze mechanism during channel attention to optimize and concatenate multi-scale features for building extraction. W. Deng et al. [61] designed an attention gate-based encoder–decoder architecture for optimizing the skip connection in multi-scale encoder and decoder feature fusion as well as in the balancing classification and localization performance for final building extraction. A similar ideal was also utilized in J. Huang et al. [70], where an attention-guided feature fusion module was designed for multi-scale feature fusion in an encoder–decoder architecture to achieve effective multi-scale semantic feature representation. In conclusion, an effective multi-scale feature fusion way is a very important element for building extraction at varying scales. It can effectively utilize the hierarchical expression characteristics of a CNN to combine the advantages of multi-scale features. Moreover, it can also avoid the interference of inaccurate information and improve the expressiveness of semantic features to promote refined building extraction performance. Consequently, effective multi-scale feature fusion has always been a research hotspot.

3. Method

In this section, an overview of the proposed HDNet is provided in Section 3.1. Then, in Section 3.1, the hierarchical disentangling strategy, dual-stream semantic feature description, and multitask supervision are individually elaborated upon in Sections 3.2–3.4.

3.1. Overview

Figure 2 shows the training framework of the proposed HDNet, which includes three parts: (a) a hierarchical disentangling strategy, (b) dual-stream semantic feature description, and (c) multitask supervision. First, different from previous studies [31,32,34,61] in which classification and localization performance are traded off by fusing the specific deep and shallow layer features, our research focuses on stable and uncertain semantic feature descriptions for buildings. Thus, we decompose each residual layer of the feature map during hierarchical feature representation, which can be defined as two multi-scale feature sets with a strong semantic zone (i.e., a more stable semantic main body) and a weak semantic zone (i.e., uncertain semantic boundary). Then, an FDM is designed for feature disentangling and is coupled with hierarchical feature representation in the encoder, as shown in Figure 2a. Here, ResNet-50 is employed as the encoder, but it can be replaced

by any other powerful backbone. Next, after the hierarchical disentangling, the two disentangled multi-scale feature sets, the more stable semantic main body and the uncertain semantic boundary, are fused and generalized by the designed CFFM and parallel multi-rate atrous convolutions during dual-stream semantic feature description, which can generate two strong and weak semantic zones that can then be fused by a feature fusion (FF) module to obtain the final semantic description, as shown in Figure 2b. Finally, corresponding to a dual-stream sematic feature description, the multitask supervision module shown in Figure 2c is designed to jointly supervise the more stable semantic main body, the uncertain semantic boundary, and their fused features based on three specific supervising signals. Then, the proposed HDNet shown in Figure 2 can obtain refined building extraction results for both consistent building region predictions and precise boundary generations. Moreover, based on the proposed method, a novel semantic feature generalization method called HDNet, there is no need to introduce additional features or operations into the final fused semantic feature to balance the classification and localization capabilities, and the proposed method can also prevent the semantic gap from being influenced and is followed by the proposed hierarchical disentangling and dual-stream semantic feature description methods. Then, the hierarchical detangling strategy, dual-stream sematic feature description, and multitask supervision can be elaborated upon as follows:

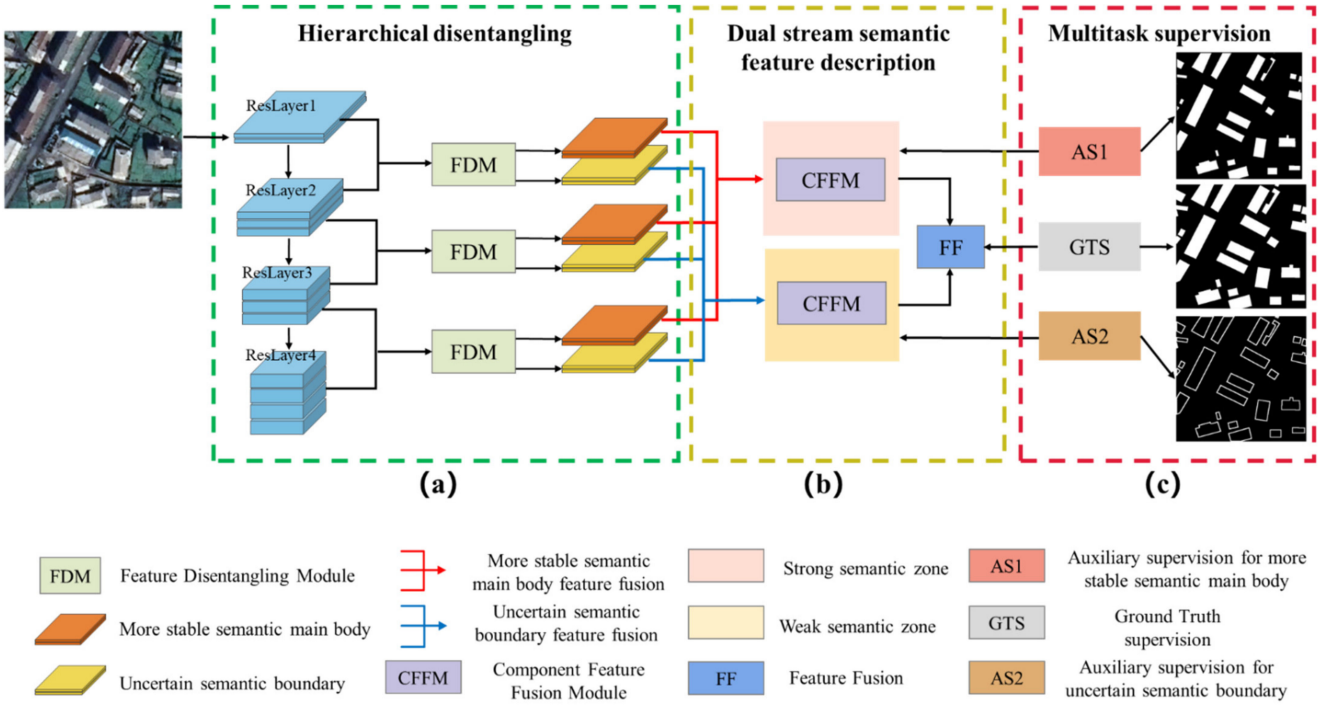


Figure 2. The framework of the proposed HDNet. (**a**) is the hierarchical disentangling strategy; (**b**) is the dual-stream semantic feature description; (**c**) is the multitask supervision.

3.2. Hierarchical Disentangling Strategy

As far as we know, during general semantic feature generalization, the feature information is gradually aggregated through the encoding process, and then a more stable and abstract description is generated in the deep layers. However, at each deeper level, the features in the encoder will inevitably lose fine-scale information (e.g., precise boundaries and clear shapes), resulting in serious feature misalignment among a large number of hierarchical features, creating a semantic gap during hierarchical feature fusion. Thus, it is difficult for previous methods to balance their classification and localization capabilities. As discussed and introduced in Section 1, we expect to decompose the semantic features in the hierarchical feature representation into two parts, which can be defined as two feature sets: the more stable semantic main body and the uncertain semantic boundary. Here,

in each feature layer, the more stable semantic main body can be considered to be the invariant pixel-level description for the foreground (i.e., building) and background, and the uncertain semantic boundary can be seen as the description of the boundary pixels with an uncertain state between the foreground (i.e., buildings) and background. Therefore, an FDM module was designed and applied to two adjacent encoder features to achieve the feature disentangling shown in Figure 3.

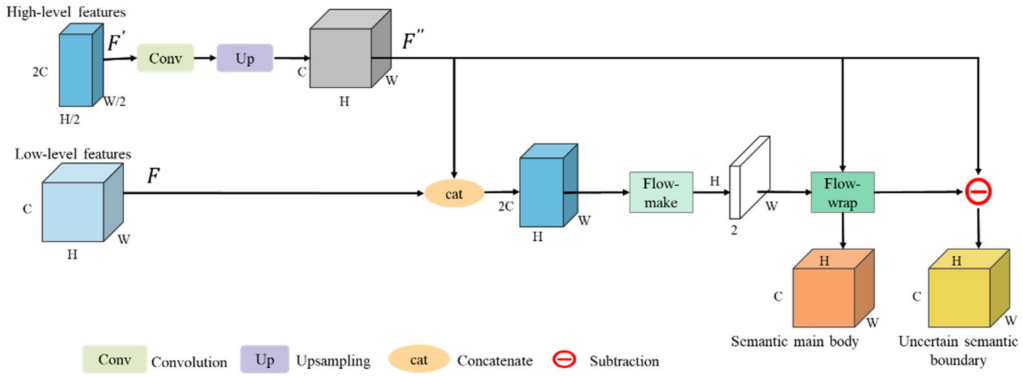


Figure 3. The detailed structure of the designed FDM.

In Figure 3, the inputs $F \in \mathbb{R}^{C\times H\times W}$ and $F' \in \mathbb{R}^{2C\times \frac{H}{2}\times \frac{W}{2}}$ are low- and relatively high-level features obtained by a 1×1 convolution to squeeze channels from the original encoder features, and C, H, and W indicate the channel, height, and width of the low-level feature map F. Here, the relatively high-level feature F' has a smaller spatial size and larger channel number than the low-level feature F. Now, in order to disentangle the semantic features, we need to know which feature points in the relatively high-level feature map are more stable or uncertain. Consecutive convolution operations not only aggregate the deeper semantic feature information to be more stable, but they also lead to the loss of details and feature misalignment, so we considered taking the adjacent low-level features as a reference to align the low- and relatively high-level features to extract more stable semantic main body and uncertain semantic boundary pixels. Therefore, as shown in Figure 3, before feature disentangling, the size of F and F' should be unified. Here, a 1×1 convolution and bilinear interpolation up-sampling are applied for F' to generate $F'' \in \mathbb{R}^{C\times H\times W}$ that is the same size (i.e., C, H and W) as the low-level feature F. To make the network automatically learn the specific feature misalignment, the feature flow field $\delta \in \mathbb{R}^{2\times H\times W}$ from F'' to F can be determined by a flow mechanism, such as the one shown in Figure 3. Specifically, F and F'' are first concatenated together, and then a 3×3 convolution is employed to generate the δ, which has two-dimensional channels that determine the deviation direction of each feature point in the flow field. The δ depicts the relationship to align the F'' with the low-level feature F. Thus, the formulation of the flow mechanism can be expressed as below:

$$\delta = \mathrm{conv}_{3\times 3}(cat(F, F'')) \tag{1}$$

In Equation (1), $cat(\cdot)$ represents the concatenation operation, and $\mathrm{conv}_{3\times 3}(\cdot)$ represents a 3×3 convolution layer to capture the local information of the feature field. Then, through Equation (1), each feature point in the relatively high-level feature map has a corresponding offset for the differentiable bilinear sampling mechanism that can be employed to rectify F'' via a feature flow warp operation, as expressed by Equation (2):

$$F'_{MainBody} = \Phi(F'', \delta) = \sum_{\rho\in\mathbb{N}(\rho)} \omega_\rho F''(\rho) \tag{2}$$

In Equation (2), $\Phi(\cdot)$ denotes the feature warp operation, and ρ is the feature point in a relatively high-level feature F''. $\mathbb{N}(\rho)$ indicates the neighbors of each warped feature point ρ in F'', and ω_p is the corresponding offset weight in the flow field. Thus, using Equations (1)

and (2) and referring to the low-level feature flow guidance, the strong semantic zone of the more stable semantic main body $F'_{MainBody}$ in a high-level feature map can be decomposed by feature warp operations. Next, the weak semantic zone with an uncertain semantic boundary can be produced by Equation (3):

$$F'_{UncertainBoundary} = F'' - F'_{MainBody} \tag{3}$$

Based on Equation (3), we can achieve the feature disentangling in the relatively high-level features that refer to adjacent low-level features. Then, coupled with hierarchical feature representation, the designed FDM can be performed on three groups of adjacent features, as shown in Figure 2. Thus, the proposed hierarchical disentangling strategy would alleviate the semantic gap that exists in the interlayer, setting up a good foundation for subsequent dual-stream semantic feature description.

3.3. Dual-Stream Semantic Feature Description

After the hierarchical disentangling strategy introduced in Section 3.2, two multi-scale feature sets with a more stable semantic main body and uncertain semantic boundary can be obtained, as shown on the left of Figure 4. Here, unlike previous studies that only fuse the deep- and shallow-layer features to generate a semantic description, the multi-scale characteristics of buildings of various scales are considered, and a dual-stream semantic feature description architecture is set up to integrate these multi-scale semantic features that belong to strong and weak semantic zones in two parallel branches. Therefore, a progressive feature fusion strategy and a CFFM are proposed.

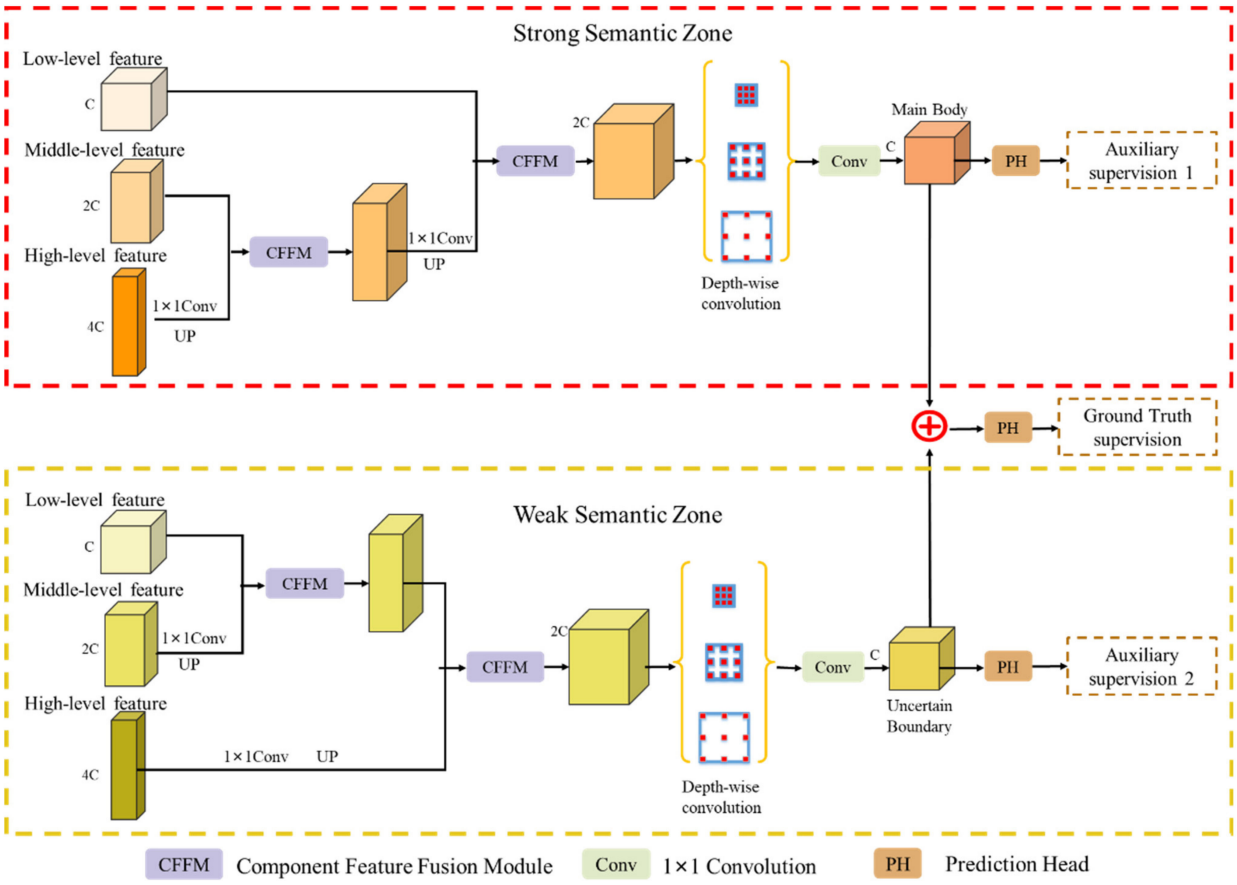
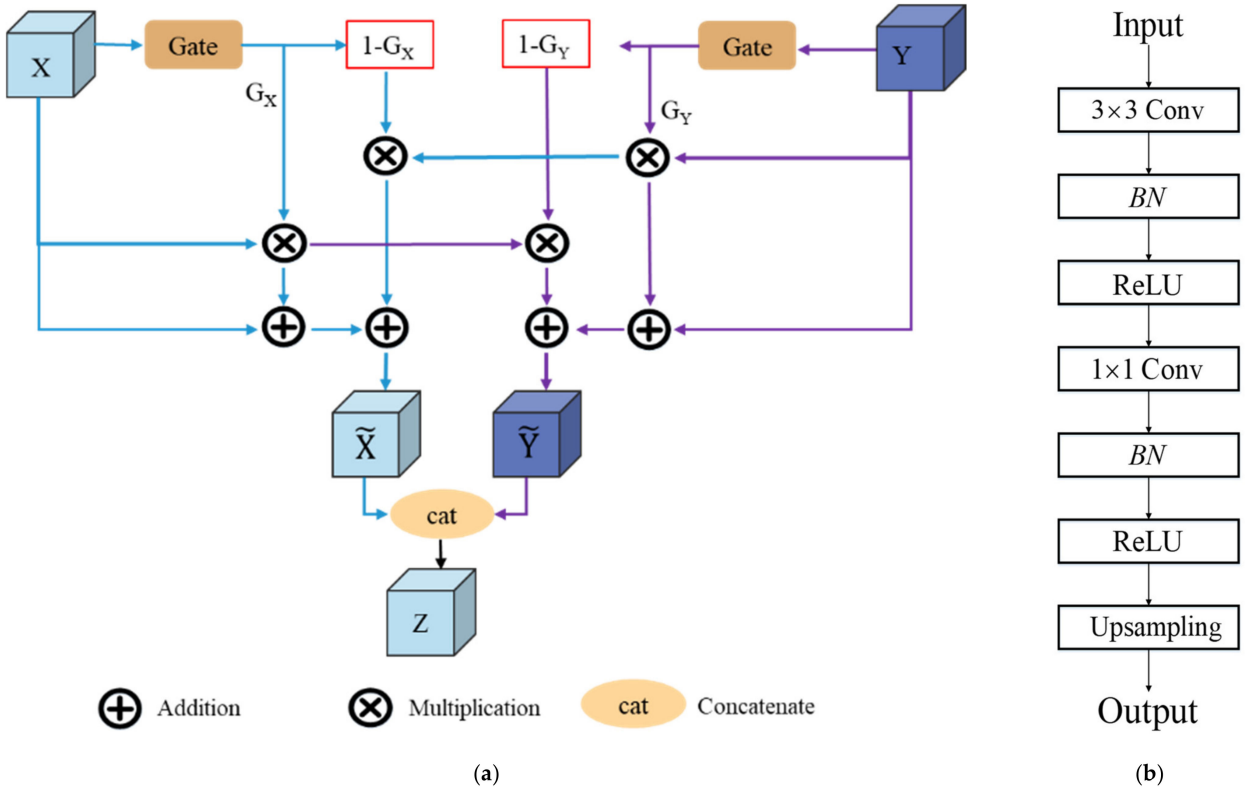


Figure 4. The framework of the dual-stream feature description architecture.

As illustrated in Figure 4, before layer-by-layer feature fusion, the up-sampling, down-sampling, and channel compression operations are employed to regulate these hierarchical

disentangling features, which have different spatial sizes and channels (i.e., C, 2C and 4C), to have the same spatial size and 2C channels at the end. Then, in the strong semantic zone, to ensure a more stable pixel-level prediction ability, a series of more stable semantic main body features are fused in a top-down way. Additionally, in the weak semantic zone, to describe the uncertain state of the boundary information, a series of weak semantic boundary features are fused in a bottom-up way.

For efficient component feature fusion and to alleviate the semantic gap, a CFFM, the purple block in Figure 4, was designed to selectively fuse the valid information layer by layer and was applied in both the strong and weak semantic zones. Additionally, the depth-wise, parallel multi-rate atrous convolutions (e.g., the atrous rates are 1, 2, and 5) and the 1×1 convolution are then individually applied to the strong and weak semantic zones to achieve feature aggregation. Finally, two features from the strong and weak semantic zones are fused by a feature fusion module, which is a pixel-wise addition in this paper, and fed into the prediction head (PH), which is the yellow block in Figure 4 for GT supervision. In addition, the strong and weak semantic zones are also individually fed into the PH for auxiliary supervision. A detailed structure of the CFFM is shown in Figure 5a.



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

Figure 5. The details of the CFFM and PH. (**a**) is the component feature fusion module (CFFM); (**b**) is the prediction head (PH).

In Figure 5a, X and Y are the regulated adjacent feature layers from the gradually feature fusion method, and they are the inputs of the designed CFFM. This is different from the normal feature fusion methods that capture the concatenation operation without considering feature information interactive flowing. In CFFM, before feature fusion, we carefully consider the complementarity of the adjacent feature layers and utilize the sigmoid function as a gate to select and guide complementary information fusion, which can not only significantly reduce the invalid information during feature fusion, but also make the feature fusion more reasonable. The whole process from Figure 5a also can be formulated via Equations (4)–(6):

$$\widetilde{X} = (1 + G_X) \cdot X + (1 - G_X) \cdot G_Y \cdot Y \tag{4}$$

$$\widetilde{Y} = (1 + G_Y) \cdot Y + (1 - G_Y) \cdot G_X \cdot X \tag{5}$$

$$Z = cat\left(\widetilde{X}, \widetilde{Y}\right) \tag{6}$$

In Equations (4)–(6), $\widetilde{X}$ and $\widetilde{Y}$ are the optimized feature layers that are prepared for concatenation operations, and Z is the output of the fused features produced by the designed CFFM. Then, $cat(\cdot)$ indicates the concatenation operation, and G_x and G_y are the attention coefficients generated from the gate, as shown in Figure 5a. Here, G_x and G_y are produced by a 1×1 convolution and the sigmoid function and can be used to suppress the invalid information by multiplying it with the corresponding inputs of X and Y. On the other hand, $(1 - \mathrm{Gx})$ and $(1 - \mathrm{Gy})$ represent the feature information that is not noticed in the corresponding input features of X and Y; thus, we applied interaction for X and Y to make them more complementary, as shown in Figure 5a and in Equations (4) and (5). Furthermore, when the hierarchical disentangling features of the strong and weak semantic zones go through gradual feature fusion via the designed CFFM, the fused features have a spatial size that is $\frac{1}{4}$ of the input with the 2C channel as two parts. Here, referring to Deeplab v3 [14] and v3+ [15], the depth-wise and parallel multi-rate atrous convolutions are performed on strong and weak semantic zones, achieving semantic feature generalization, and then a 1×1 convolution is employed to refine feature fusion, as shown in Figure 4. Next, after FF, we can obtain three parts (i.e., the more stable semantic main body, uncertain semantic boundary, and their addition), and then, they are parallel fed into three parallel PH blocks for pixel-level predictions and multitask supervision. The details of the PH block are shown in Figure 5b and consist of a cascading 3×3 convolution, BN, ReLU, a 1×1 convolution, and up-sampling operations.

3.4. Multitask Supervision

As shown in Figures 2 and 4, based on the proposed novel semantic feature description paradigm, there are three supervision tasks that can optimize the results in the training phase, and this joint loss includes the strong semantic zone of the more stable semantic main body, the weak semantic zone of the uncertain semantic boundary, and GT. Thus, a multitask supervision function can be written as Equation (7):

$$L_{total} = \lambda_1 \cdot L_S + \lambda_2 \cdot L_B + \lambda_3 \cdot L_E \tag{7}$$

In Equation (7), L_{total} is the total loss in multitask supervision, and λ_1, λ_2, and λ_3 are the weighted factors which are used to balance the contribution of each task. In our work, λ_1, λ_2, and λ_3 are empirically set as *1*, *1*, and *20*. Next, L_S, L_B, and L_E in Equation (7) correspond to entire segmentation loss, loss in the strong semantic zone of the more stable semantic main body, and loss in the weak semantic zone of the uncertain semantic boundary, respectively. Here, the supervision signal of the more stable semantic main body is considered to be the foreground (i.e., buildings) and background without uncertain boundary information, and, on the contrary, the uncertain semantic boundary is considered as the uncertain state of the building boundary information. Therefore, the GT is used for L_S, and the auxiliary supervision signals of L_B and L_E are manually generated by morphological image operations (e.g., erosion, dilation, and subtraction), which decompose the GT into two parts for L_B and L_E. Next, the binary cross entropy (BCE) loss [72] is chosen for L_S, L_B, and L_E in HDNet training phase, which can be expressed as Equation (8):

$$L_S = L_B = L_E = -\frac{1}{N}\sum_{i=1}^{N}[y_i \cdot \log(p_i) + (1 - y_i) \cdot \log(1 - p_i)] \tag{8}$$

In (8), N is the number of samples; $y_i \in \{0, 1\}$ denotes the label of pixel i, which represents whether pixel i belongs to the building or not; and i is the index. $P_i \in [0, 1]$ is the prediction probability of pixel i. Finally, using Equation (8), the multi-task supervision can leverage the proposed HDNet to rapidly converge and generate more refined building extraction results.

4. Experiment

4.1. Evaluation Indexes and Datesets

To evaluate the building extraction performance, we used pixel-level metrics that included *Recall*, *Precision*, *F1-score*, and *IoU*. The Equations are as follows:

$$Precision = \frac{TP}{TP + FP} \tag{9}$$

$$Recall = \frac{TP}{TP + FN} \tag{10}$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{11}$$

$$IoU = \frac{TP}{FP + TP + FN} \tag{12}$$

In Equations (9)–(12), *TP* denotes the number of pixels that are correctly predicted as the building area, *FP* denotes the number of pixels where the background pixels are incorrectly predicted as buildings, *TN* denotes the number of pixels that are correctly predicted as the background, and *FN* is the number of pixels where building pixels are incorrectly predicted as the background.

The datasets that we used for building extraction in our study are three open accessed building datasets, namely the WHU Aerial imagery dataset [5], the WHU Satellite dataset II dataset [5], and the INRIA aerial image labeling dataset [6]. These datasets have been widely used in published works based on convolutional neural networks, and their data volume can support deep learning research. Therefore, these datasets will be able to provide a good assessment of the generalization ability of our network. Information regarding these three datasets is reported below:

WHU Aerial imagery dataset: This dataset covers an area of 450 square kilometers and has a 0.3 m spatial resolution. The dataset consists of more than 187,000 independent buildings extracted from aerial images from Christchurch, New Zealand. These areas contain countryside, residential, cultural, and industrial areas.

WHU Satellite dataset II dataset: This dataset consists of six neighboring satellite images covering 860 km^2 over East Asia. It is a useful complement to other datasets collected from Europe, America, and New Zealand and supplies regional diversity. The dataset contains images of 34,085 buildings with a 2.7 m ground resolution

INRIA aerial image labeling dataset: This dataset contains 360 aerial RGB images with various building characteristics. The dataset covers different European and American cities, including Chicago, San Francisco, Vienna, Innsbruck, Kitsap, Bloomington, and East/West Tyrol, etc. The images have a spatial resolution of 0.3 m and cover an overall area of 810 square kilometers.

4.2. Implementation Setting

All of the experiments are carried out and analyzed on the published WHU aerial [5] and satellite [5] and INRIA [6] building extraction datasets. The WHU aerial [5] imagery dataset contains 5772 images for training and 2416 images for testing. The WHU satellite dataset [5] has 3135 images for training and 903 images for testing. The image size of the WHU aerial and satellite datasets is 512 × 512. Each image in the INRIA building [6] dataset is 5000 × 5000, and all of the images were spliced into 512 × 512 pixels to finally obtain 6000 images for training and 2500 images for testing. To improve the robustness of the model, we also used several data augmentation operations, including random flipping, random rotation, and color enhancement. Moreover, all of the comparison methods were implemented by Pytorch 1.5.0 and were performed on a TITAN RTX GPU. The proposed HDNet was trained over 200 epochs, and the batch size was set to 8. Here, we adopted a stochastic gradient descent (SGD) optimization strategy, and the initial learning rate

and momentum term were set as 0.01 and 0.9, respectively to dampen the oscillations during optimization.

4.3. Performance Comparison and Analysis

To evaluate the effectiveness of the proposed HDNet model for refined building extraction from VHR optical remote sensing imagery, we selected several remarkable CNNs for semantic segmentation and building extraction as comparative methods. Among these methods, FCN [17], Unet [21], Deeplabv3 [26], and Deeplabv3+ [27] are the classical semantic segmentation architectures. Moreover, some recently determined boundary-aware semantic segmentation algorithms, including Gated-SCNN [62], BACNet [63], BAMTL [65], and SBANet [66], were chosen to verify the superiority of multi-task supervision. MAPNet [45] is a recently published method for building extraction, which proposes a multi-parallel path for extracting multiscale buildings and precise boundaries. Res2-Unet [46], a recently published enhanced version of Unet, focuses on multi-scale learning to deal with easily omitted small buildings and complicated details of background objects. Then, to demonstrate the performance of the compared methods under different scenarios, the satellite and aerial observation datasets from the WHU [5] and INRIA [6] datasets were employed for performance comparison and analysis.

4.3.1. Comparison using the INRIA Dataset

The INRIA dataset covers most types of buildings in five cities, including sparse courtyards, dense residential areas, and large venues. Several comparative experiments were carried out on the INRIA dataset to compare the building extraction performance achieved by the proposed HDNet and the comparative methods. The quantitative evaluation results are reported in Table 1.

Table 1. Quantitative comparison on the INRIA dataset.

Methods	Recall (%)	Precision (%)	F1-Score (%)	IoU (%)
FCN [17]	70.4	74.3	72.3	56.7
Unet [21]	82.1	83.2	82.6	71.4
Deeplabv3 [26]	82.4	78.0	80.2	66.9
Deeplabv3+ [27]	82.7	85.8	84.3	72.7
MAPNet [45]	83.4	85.0	84.2	72.7
Res2-Unet [46]	80.3	84.4	82.3	69.9
Gated-SCNN [62]	85.9	85.4	85.6	74.9
BCANet [63]	84.4	84.5	84.4	73.1
BAMTL [65]	83.9	84.4	84.2	72.7
SBANet [66]	83.0	85.0	83.9	72.3
HDNet (ours)	**86.7**	**87.7**	**87.2**	**77.3**

For every test region, the highest values for different metrics are highlighted in bold.

It can be seen that the proposed HDNet outperforms other methods on all four metrics. The FCN and Deeplabv3 perform worse than most of the other methods because they lack detailed structural information in shallow layers. Compared to our baseline model deeplabv3 [26], HDNet can improve F1-score by 7.0% and the IoU by 10.4% on the INRIA aerial dataset. Moreover, the proposed HDNet also outperforms Deeplabv3+ [27], and achieving performance improvements in the F1-score by 2.9% and in the IoU by 4.6%, which indicates that the proposed hierarchical disentangling strategy and dual-stream semantic feature description in HDNet are better than simply introducing a shallow feature layer into the final fused semantic feature to make a tradeoff between the pixel-level classification and localization performance. Additionally, compared to the second-best method, Gated-SCNN [62], HDNet improved the F1-score from 85.6% to the highest score of 87.2% and improved the IoU score from 74.9% to 77.3%. These improvements indicate that the proposed novel semantic feature description paradigm of HDNet is robust

enough to cope with building extraction tasks using VHR aerial images taken in different complex conditions.

4.3.2. Comparison of the WHU Aerial Dataset

The WHU aerial dataset covers a large variety of buildings, such as small residential blocks, industrial factories, and commercial buildings. The buildings in residential blocks are distributed very densely and are small in size. Buildings in industrial factories and commercial buildings have various shapes and textures. This means that the model needs to extract buildings from such complex environments as well as possible. Table 2 summarizes the quantitative comparison with different models on the WHU aerial dataset. Compared to the results on the INRIA dataset, the IoU metrics are all higher than 80%, with the exception of FCN [17] and Deeplabv3 [26], indicating that the WHU aerial dataset is of higher quality and is easier to distinguish. Especially, the HDNet achieves the best scores on all four metrics, and the IoU score of HDNet exceeds 90%.

Table 2. Quantitative comparison on the WHU aerial dataset.

Methods	Recall (%)	Precision (%)	F1-Score (%)	IoU (%)
FCN [17]	72.4	74.9	73.7	58.3
Unet [21]	89.4	93.6	93.3	82.1
Deeplabv3 [26]	78.1	77.6	77.9	63.7
Deeplabv3+ [27]	91.2	93.0	92.1	85.4
MAPNet [45]	93.3	94.6	94.0	88.7
Res2-Unet [46]	92.3	91.8	92.0	85.3
Gated-SCNN [62]	93.8	94.8	94.3	89.2
BCANet [63]	93.6	93.3	93.5	87.7
BAMTL [65]	92.5	92.5	92.5	86.1
SBANet [66]	92.3	91.8	92.0	85.3
HDNet (ours)	**94.8**	**95.2**	**95.0**	**90.5**

For every test region, the highest values for different metrics are highlighted in bold.

4.3.3. Comparison on the WHU Satellite Dataset

Table 3 provides a quantitative comparison of the different models on the WHU satellite dataset. Compared to the results of the WHU aerial dataset, all of the metrics are much lower. There are two reasons for this. First, the WHU satellite dataset has a much lower resolution. Second, the scenes in the WHU satellite dataset have low contrast. Specifically, the small-scale buildings that have merged into the surroundings create great difficulties for extraction. Despite this, it is clear that the proposed HDNet outperforms other methods. Compared to recently published work using Res2-Unet [46], the proposed HDNet improved the IoU by 5.2%. Moreover, the HDNet exhibits a considerable advantage over boundary-aware semantic segmentation methods, such as the Gated-SCNN [62], BCANet [63], BAMTL [65], and SBANet [66] methods.

To visually demonstrate how these networks perform building extraction tasks in different scenarios, some examples are shown in Figure 6. The first two columns are input images and GT images, and the following columns are probability images predicted by FCN [17], Deeplabv3 [26], Deeplabv3+ [27], Res2-Unet [46], BAMTL [65], MAPNet [45], and the proposed HDNet method, where white pixels represent buildings, and the black pixels are objects that are not buildings. The red pixels are false alarms that should be non-buildings but that were predicted as building pixels. The green pixels indicate the buildings that were missed during extraction. Then, the first to the sixth rows are examples from the INRIA dataset; rows 7 to 10 are examples from the WHU aerial dataset, and the last five rows are from the WHU satellite dataset. From the INRIA dataset, the first two rows are examples of isolated buildings. It can be seen that the proposed HDNet was the only algorithm that was able to successfully extract the buildings from the surrounding environment. The following four rows are examples of large-scale and sparsely distributed buildings for which the proposed HDNet can provide the best results both in terms of the integrity and

the accuracy of the predicted building boundaries. For the WHU aerial dataset, it can be seen that HDNet presents the least false alarms in the face of densely distributed and very small-scale buildings. The more complete main body and finer boundary extraction of the small-scale buildings prove that the proposed hierarchical disentangling strategy and the effective dual-stream semantic feature description method can provide powerful semantic description ability for buildings in complex environments. The last five rows of Figure 6 show examples of buildings that have been extracted from rural scenes in the WHU satellite dataset. In the first three scenarios, the buildings are randomly scattered in rural settlements, and the buildings and background elements have similar shapes and colors. Many misclassifications were observed at the junction of the buildings and the background when the six comparative methods were used. The last two scenarios are for isolated buildings in different seasons and lighting conditions. It was difficult for the six comparative methods to separate the confusing non-building backgrounds and buildings with a dense distribution with low-contrast backgrounds. Overall, HDNet produces the cleanest and the most accurate compared to the other methods, meaning that the semantic description paradigm of the proposed HDNet is capable of learning the generalized representation of semantic features.

Table 3. Quantitative comparison on the WHU satellite dataset.

Methods	Recall (%)	Precision (%)	F1-Score (%)	IoU (%)
FCN [17]	54.1	68.0	60.2	43.1
Unet [21]	72.1	86.0	78.4	64.5
Deeplabv3 [26]	74.1	69.0	71.4	55.6
Deeplabv3+ [27]	71.3	78.8	74.9	59.8
MAPNet [45]	80.4	85.1	82.7	70.5
Res2-Unet [46]	74.9	85.3	79.8	66.3
Gated-SCNN [62]	73.8	**87.1**	79.9	66.5
BCANet [63]	76.7	86.0	81.1	68.1
BAMTL [65]	79.7	83.1	81.3	68.6
SBANet [66]	80.5	82.4	81.4	68.7
HDNet (ours)	**82.7**	84.8	**83.8**	**72.1**

For every test region, the highest values for different metrics are highlighted in bold.

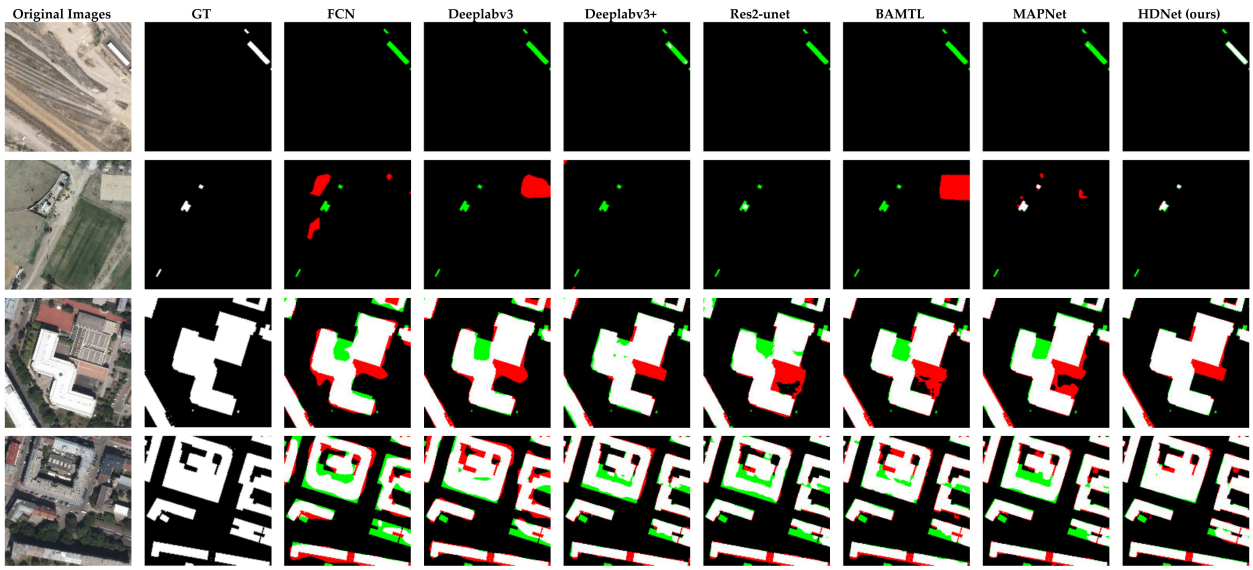


Figure 6. *Cont.*

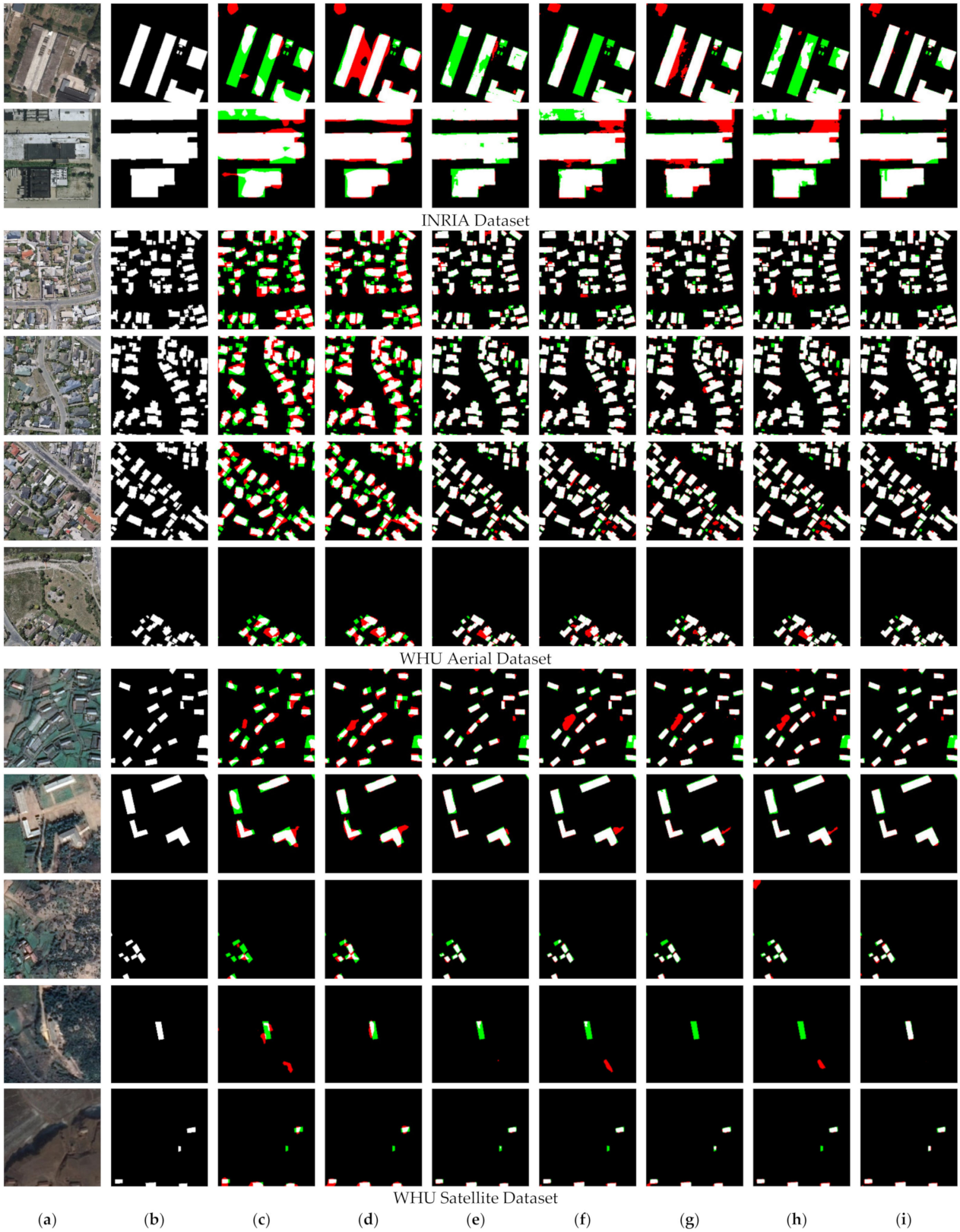


Figure 6. The visualization results of the building extraction results from the WHU building datasets and INRIA dataset. (**a**) original image; (**b**) ground truth (GT); (**c**) FCN [17]; (**d**) Deeplabv3; (**e**) Deeplabv3+; (**f**) Res2-UNet; (**g**) BAMTL; (**h**) MAPNet; (**i**) HDNet (ours). Black, white, red, and green pixels indicate the non-buildings, buildings, false alarm areas, and missed inspection areas, respectively.

5. Discussion

5.1. Ablation Study

In order to better verify the effectiveness of each of the modules that were designed for the proposed HDNet, several ablation studies were carried out on the WHU satellite dataset, and the results are reported in Table 4.

Table 4. Ablation experiments for HDNet.

Methods	FDM			CFFM	Multitask Supervision	F1-Score (%)	IoU (%)
	FDM ^{2}	FDM ^{3}	FDM ^{4}				
Baseline						68.4	52.0
HDNet^1			✓		✓	81.8	69.2
HDNet^2		✓	✓		✓	83.0	70.9
HDNet^3	✓	✓	✓		✓	83.4	71.6
HDNet^{3,*}	✓	✓	✓			82.7	70.5
HDNet (our)	✓	✓	✓	✓	✓	**83.8**	**72.1**

FDM ^{2}, FDM ^{3}, and FDM ^{4} indicate disentangling residual layer2, layer3, and layer4 from ResNet-50 by the designed FDM, respectively; CFFM indicates that CFFM was used for dual-stream semantic feature description. ^{1}, ^{2}, ^{3} in HDNet indicate the number of FDM used in the ablation experiments. * indicates the method without multitask supervision. For every test region, the highest values for different metrics are highlighted in bold.

In Table 4, the baseline method is based on Deeplabv3 [26] with some modifications. Here, for a fairer comparison, similar to the proposed HDNet, ResNet-50 was employed as the encoder in Deeplabv3 [26], and the last layer of ResNet-50 was upsampled by up to four times, with the number of channels being compressed to 128 before being sent to the ASPP module. It should be noted that the original convolutions of the ASPP module are replaced by three parallel depth-wise atrous convolutions with atrous rates of 1, 2, and 5, the same as HDNet. As shown in Table 4, HDNet^1, which only uses the designed FDM for semantic feature disentangling on residual layer 4, can obtain large improvements over the baseline method, examples of which are the increase in the F1-score by 13.4% and the increase in the IoU by 17.2%. Next, as more FDM modules are added to the residual layers for hierarchical disentangling, the performance of the F1-score and IoU from HDNet^1 to HDNet^3 was improved even further. Furthermore, from the comparison of HDNet^3 and HDNet^{3,*}, it can be seen that the multitask semantic boundary supervision can facilitate the description of weak semantic zone and can promote the final extraction performance of the building. Finally, as shown in the last row of Table 4, after the CFFM module is added to the dual-stream semantic feature description module of HDNet^3, the final proposed HDNet is able to provide superior performance compared to the other iterations.

Moreover, the PR and ROC curves were calculated from the WHU satellite dataset and correspond to the ablation study and are shown in Figure 7. From the PR curve, it can be seen that the proposed HDNet can create a better balance between precision and recall and achieves a better overall performance. From the ROC curve, it can be seen that by disentangling the semantic feature into strong and weak semantic zones, the proposed method produces few false alarms.

5.2. Impacts of Supervision Labels Generation

In order to promote feature disentangling in the strong and weak semantic zones to ensure the integrity of the bodies of the buildings and the accuracy of the building boundaries, multitask supervision was performed in the training stage. Two of the auxiliary supervision tasks include the supervision of the main bodies and semantic boundaries of the buildings. The supervision labels in these two tasks are generated manually through image erosion operations on the original building extraction labels. Specifically, the original GT is eroded according to a certain number of pixels to obtain the label of the more stable semantic main body. Since the semantic features of the semantic main body and the semantic boundary are generated by feature disentangling, the supervision label of

the semantic boundary can also be obtained by subtracting the supervision label of the semantic main body from the original GT. Semantic boundary supervision labels with different bandwidths can be obtained according to the number of eroded pixels, as shown in Figure 8.

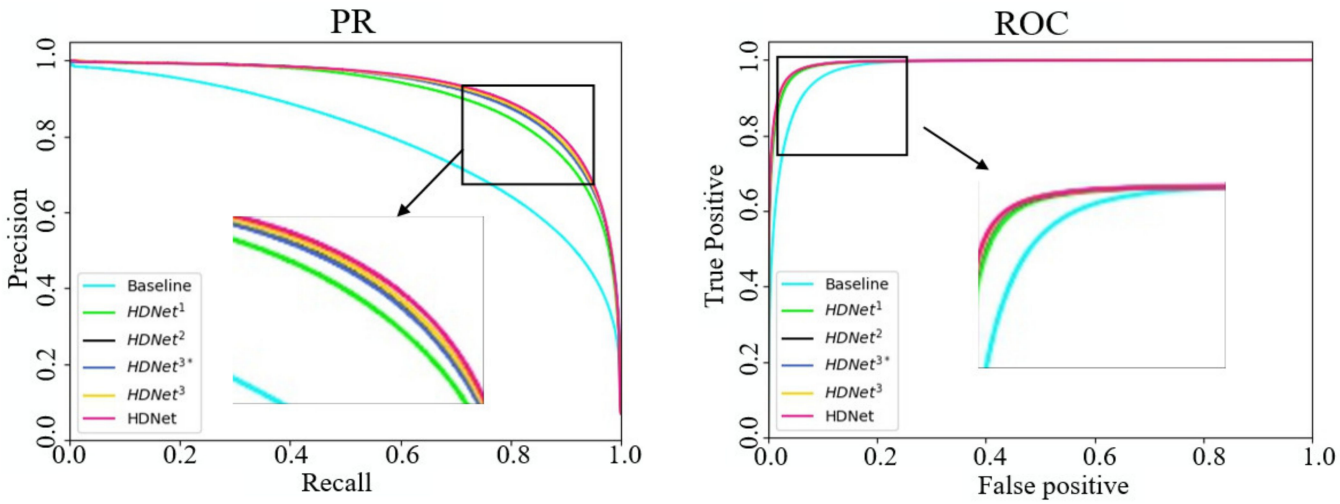


Figure 7. Analysis on the WHU satellite dataset. (**a**) PR curve; (**b**) ROC curve. [1], [2], [3] in HDNet indicate the number of FDM used in the methods. * indicates the method without multitask supervision.

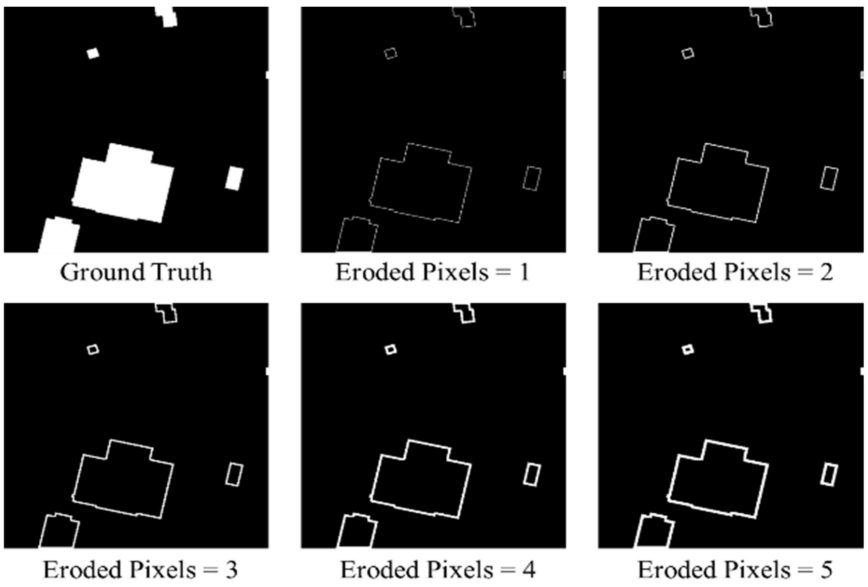


Figure 8. Building boundaries under different eroded pixel sizes. The image first is the ground truth. The next five images are the different-sized building boundaries according to the size of the eroded pixels.

Considering the possibility of feature misalignment between the semantic main body features and the semantic boundary features, it is reasonable to discuss the impact of the overlap between the semantic main body label and semantic boundary label. As such, two types of experiments were conducted to explore the effect of the number of eroded pixels on the overall performance depending on whether the boundary label and main body label overlap. Within each set, a comparison of the different bandwidths was performed. The evaluation results of the two sets of experiments are shown in Tables 5 and 6.

Table 5. Experiments with overlapping semantic boundary and semantic main body labelling.

Methods	Recall (%)	Precision (%)	F1-Score (%)	IoU (%)
Edge 1	82.5	84.0	83.3	71.3
Edge 2	82.6	83.9	83.2	71.3
Edge 3	**82.7**	**84.8**	**83.8**	**72.1**
Edge 4	82.4	84.3	83.3	71.4
Edge 5	82.4	84.1	83.3	71.3

For every test region, the highest values for different metrics are highlighted in bold.

Table 6. Experiments with non-overlapping semantic boundary and semantic main body labelling.

Methods	Recall (%)	Precision (%)	F1-Score (%)	IoU (%)
Edge 1†	82.3	84.2	83.3	71.5
Edge 2†	83.0	83.7	83.3	71.5
Edge 3†	83.0	83.8	83.4	71.5
Edge 4†	82.7	**84.4**	**83.5**	**71.7**
Edge 5†	**84.1**	82.4	83.3	71.3

For every test region, the highest values for different metrics are highlighted in bold. † represents the methods with non-overlapping semantic boundary and semantic main body supervision.

In Table 5, methods Edge 1 to Edge 5 represents the experiments in which the semantic boundary supervision labels were obtained when the number of eroded pixels ranged from one to five pixels. Note that the supervision labels of the semantic main body are fixed and obtained by eroding the original GT by one pixel. Therefore, there are different degrees of overlap between the semantic main body label and semantic boundary labels. As seen in Table 5, the best building extraction performance is obtained when the number of eroded pixels is three.

Furthermore, as reported in Table 6, in methods Edge1† to Edge 5†, each semantic boundary and semantic main body supervision label was obtained when the number of eroded pixels ranged from one to five pixels, and they are complementary without overlapping. It can be seen that four pixels are a better setting to define the width of the uncertain semantic boundary for multi-task supervision on this dataset. Additionally, the performance of the overlapping supervision strategy is better than that of the non-overlapping supervision strategy. This is because the overlapping supervision strategy can reduce occurrences of missed extraction, and the overlapping part can play a complementary role for the more stable semantic body and uncertain semantic boundary in building extraction tasks. This experiment also reflects the problem that semantic feature description for buildings using feature disentangling is affected by the manually generated supervision labels to some extent. Additionally, the influence will differ for buildings of different scales, distribution densities, and surrounding interferences. Appropriate multi-task supervision labels can promote feature extraction in the strong and weak semantic zones and can facilitate the semantic description of uncertainty and reduce false alarms.

5.3. Generalization Ability of HDNet

Moreover, we also selected typical building extraction images from the WHU satellite dataset to further evaluate the generalization performance of HDNet in different complex scenarios. The selected sample images can be classified into three subsets according to the statistics in Figure 1 in the introduction section. In order to verify the generalization ability of HDNet for buildings with different scales, subset 1 selects typical images containing large-scale and small-scale buildings. Additionally, to verify the generalization ability of different spatial distributions, subset 2 selects typical images containing densely distributed buildings and isolated buildings. Then, to verify the generalization ability of different surrounding interferences, subset 3 selects typical images with occlusions, shadows, and low contrast ratios. Additionally, the comparison results of the four SOTA methods (i.e., Gated-SCNN [62], Res2-Unet [46], BAMTL [65], MAPNet [45]) are shown in

Figure 9. It turns out that the proposed HDNet outperforms the other SOTA methods in all three subsets.

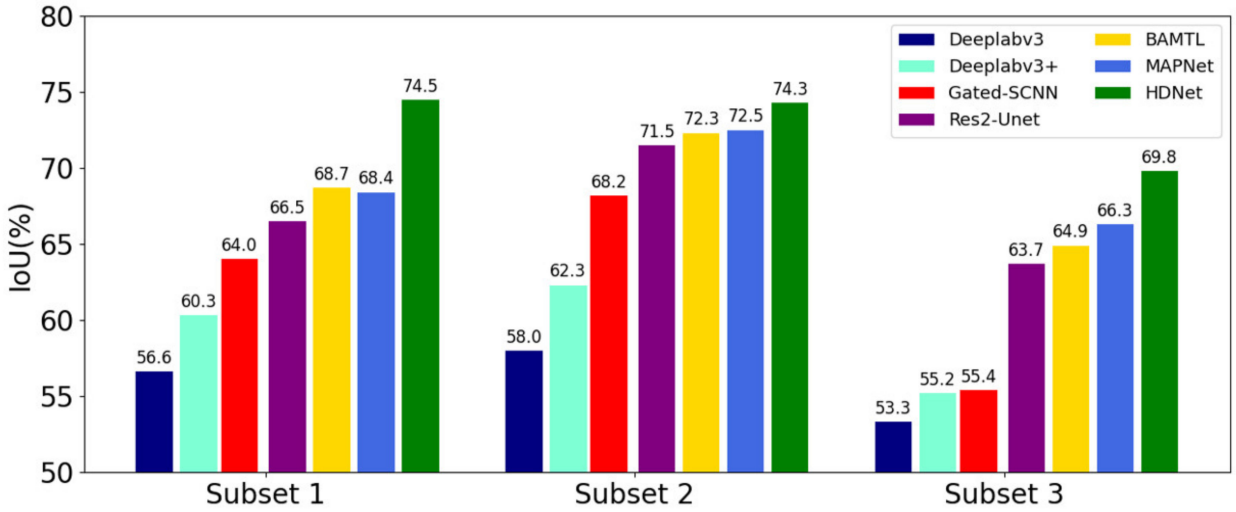


Figure 9. Analysis of building extraction experiments in three subsets on the WHU satellite dataset.

In addition to the accuracy of the evaluation indexes, we further analyzed the number of parameters, the FLOPs of the model, and the time used to infer each image during testing to better explore the bounds of the proposed HDNet, as shown in Table 7. It can be seen that while HDNet achieves the highest IoU on all of the selected sample images of different complex scenarios, it has the minimum number of parameters, only slightly larger than MAPNet [45]. The proposed HDNet method is an improved version of Deeplabv3 [26] and Deeplabv3+ [27]. The total number of Deeplabv3+ parameters is 43.10 M, while the total number of HDNet parameters is only 27.89 M because the proposed HDNet reduces the number of feature channels. Moreover, the proposed HDNet is not only able to further improve semantic feature description and generalization capabilities through the proposed hierarchical disentangling strategy and dual-stream semantic feature description, it also has acceptable parameters and inference times.

Table 7. Quantitative comparison of efficiency on the WHU satellite dataset.

Methods	IoU (%)	FLOPs (G)	Parameters (M)	Inference (s)
FCN [17]	48.7	80.5	15.31	0.032
Unet [21]	66.1	322.25	39.40	0.077
Deeplabv3 [26]	59.7	25.52	41.68	0.028
Deeplabv3+ [27]	62.3	25.66	43.10	0.029
MAPNet [45]	70.7	94.56	24.84	0.044
Res2-Unet [46]	68.6	67.99	43.68	0.046
Gated-SCNN [62]	66.5	755.43	137.27	0.136
BCANet [63]	69.4	192.76	44.95	0.056
BAMTL [65]	70.5	45.86	43.53	0.032
SBANet [66]	70.8	44.18	42.91	0.034
HDNet (ours)	73.3	33.48	27.89	0.053

In summary, the experimental results demonstrate the effectiveness and generalization ability of the proposed hierarchical disentanglement strategy and dual-stream semantic feature description in HDNet. It indicates that such a paradigm can avoid the risk of interference by introducing additional shallow features into the final fused semantic features to balance the classification and localization performance during building extraction, something that is common in many encoder–decoder architectures. It can also elegantly prevent the semantic gaps in multi-scale semantic features fusion from reducing the performance of

fine-grained building extraction for multi-scale and arbitrarily distributed buildings. However, apart from a few annotation errors in the dataset, in extremely complex scenes with dense building distribution, very small-scale buildings, and very low contrast, the range of weak semantic regions where the semantic description of the buildings and backgrounds intersect is difficult to define. In these cases, the proposed HDNet is slightly inadequate for achieving a precise description of semantic uncertainty.

6. Conclusions

In this article, we consider the existing challenges in refined building extraction and rethink the current semantic description paradigm. A novel hierarchical disentangling network called HDNet is proposed for refined building extraction. Different from the previous encoder–decoder algorithms that trade off the classification and localization performance by fusing specific deep- and shallow-layer features, the proposed HDNet sets up semantic description using a hierarchical disentangling strategy and dual-stream semantic feature description that focus on describing the strong and weak semantic zones of buildings individually and on achieving effective feature fusion for refined building extraction. Extensive experiments were conducted on the WHU satellite/aerial and INRIA building extraction datasets that further highlighted the advantages of HDNet. In summary, the proposed model achieved the highest accuracy metrics on three datasets. In detail, the proposed HDNet achieved an F1-score of 87.2% and an IoU of 77.3% on the INRIA dataset; an F1-score of 95.0% and an IoU of 90.5% on the WHU aerial dataset; and an F1-score of 83.8% and an IoU of 72.1% on the WHU satellite dataset. Note that the proposed hierarchical disentangling strategy was able to improve the IoU by about 20% compared to the baseline Deeplabv3 method in the experiments on the WHU satellite dataset. The outstanding performance of HDNet suggests that describing the stable semantic main body and uncertain semantic boundary separately is a better way to generate a more effective sematic feature description than traditional deep- and shallow-layer feature fusion methods. Considering the annotation errors in the datasets and the inadequate performance of the uncertain semantic description in extremely complex environments, in future work, we will continue to explore better modeling paradigms for stable and uncertain semantic feature description in supervised or semi-supervised building extraction tasks.

Author Contributions: Conceptualization, Y.Z.; methodology, Y.Z. and S.D.; software and validation, J.L. and S.D.; formal analysis, P.G., H.D. and Y.Z.; investigation, Y.Z., J.L. and S.D.; resources, H.C., L.C. and L.L.; data curation, P.G. and H.D.; writing—original draft preparation, Y.Z., J.L. and S.D.; writing—review and editing, Y.Z., P.G. and H.D.; visualization, J.L.; supervision, H.C. and L.L.; project administration, L.C.; funding acquisition, Y.Z., H.D. and L.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Chang Jiang Scholars Program under grant T2012122 and in part by the Civil Aviation Program under grant B0201, the Space-based on orbit real-time processing technology program under grant 2018-JCJQ-ZQ-046, and in part by the National Science Foundation for Young Scientists of China under Grant 62101046 as well as by the National Natural Science Foundation of China (No. 62136001).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets presented in this study are openly available in References [5,6]. Our work is available at https://github.com/1lee1338/HDNet (accessed on 22 February 2022).

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

References

1. 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]
2. 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]
3. 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]
4. Rathore, M.M.; Ahmad, A.; Paul, A.; Rho, S. Urban Planning and Building Smart Cities Based on the Internet of Things Using Big Data Analytics. *Comput. Netw.* **2016**, *101*, 63–80. [CrossRef]
5. 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]
6. 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), Fort Worth, TX, USA, 23–28 July 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 3226–3229.
7. Mnih, V. *Machine Learning for Aerial Image Labeling*; University of Toronto: Toronto, ON, Canada, 2013.
8. 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. [CrossRef]
9. Zhang, L.; Huang, X.; Huang, B.; Li, P. A Pixel Shape Index Coupled with Spectral Information for Classification of High Spatial Resolution Remotely Sensed Imagery. *IEEE Trans. Geosci. Remote Sens.* **2006**, *44*, 2950–2961. [CrossRef]
10. Sirmacek, B.; Unsalan, C. Building detection from aerial images using invariant color features and shadow information. In Proceedings of the 23rd International Symposium on Computer and Information Sciences, Istanbul, Turkey, 27–29 October 2008; IEEE: Piscataway, NJ, USA, 2008; pp. 1–5.
11. Zhang, T.; Huang, X.; Wen, D.; Li, J. Urban Building Density Estimation from High-Resolution Imagery Using Multiple Features and Support Vector Regression. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2017**, *10*, 3265–3280. [CrossRef]
12. Wang, J.; Yang, X.; Qin, X.; Ye, X.; Qin, Q. An Efficient Approach for Automatic Rectangular Building Extraction from Very High Resolution Optical Satellite Imagery. *IEEE Geosci. Remote Sens. Lett.* **2015**, *12*, 487–491. [CrossRef]
13. Du, J.; Chen, D.; Wang, R.; Peethambaran, J.; Mathiopoulos, P.T.; Xie, L.; Yun, T. A Novel Framework for 2.5-D Building Contouring from Large-Scale Residential Scenes. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 4121–4145. [CrossRef]
14. Chen, D.; Shang, S.; Wu, C. Shadow-Based Building Detection and Segmentation in High-Resolution Remote Sensing Image. *JMM* **2014**, *9*, 181–188. [CrossRef]
15. Gao, X.; Wang, M.; Yang, Y.; Li, G. Building Extraction from RGB VHR Images Using Shifted Shadow Algorithm. *IEEE Access* **2018**, *6*, 22034–22045. [CrossRef]
16. 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]
17. Shelhamer, E.; Long, J.; Darrell, T. Fully Convolutional Networks for Semantic Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 640–651. [CrossRef]
18. 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]
19. Peng, C.; Zhang, X.; Yu, G.; Luo, G.; Sun, J. Large kernel matters—Improve semantic segmentation by global convolutional network. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 1743–1751.
20. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 6230–6239.
21. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. *arXiv* **2015**, arXiv:1505.04597.
22. Zhou, Z.; Siddiquee, M.M.R.; Tajbakhsh, N.; Liang, J. UNet++: A Nested U-Net Architecture for Medical Image Segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*; Lecture Notes in Computer Science; Stoyanov, D., Taylor, Z., Carneiro, G., Syeda-Mahmood, T., Martel, A., Maier-Hein, L., Tavares, J.M.R.S., Bradley, A., Papa, J.P., Belagiannis, V., et al., Eds.; Springer International Publishing: Cham, Switzerland, 2018; Volume 11045, pp. 3–11. ISBN 978-3-030-00888-8.
23. Huang, H.; Lin, L.; Tong, R.; Hu, H.; Zhang, Q.; Iwamoto, Y.; Han, X.; Chen, Y.-W.; Wu, J. UNet 3+: A full-scale connected UNet for medical image segmentation. In Proceedings of the ICASSP 2020—IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–9 May 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 1055–1059.
24. Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs. *arXiv* **2016**, arXiv:1412.7062.
25. 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]

26. Chen, L.-C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking Atrous Convolution for Semantic Image Segmentation. *arXiv* **2017**, arXiv:1706.05587.
27. Chen, L.-C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. *arXiv* **2018**, arXiv:1802.02611.
28. Boonpook, W.; Tan, Y.; Ye, Y.; Torteeka, P.; Torsri, K.; Dong, S. A Deep Learning Approach on Building Detection from Unmanned Aerial Vehicle-Based Images in Riverbank Monitoring. *Sensors* **2018**, *18*, 3921. [CrossRef] [PubMed]
29. Liu, J.; Wang, S.; Hou, X.; Song, W. A Deep Residual Learning Serial Segmentation Network for Extracting Buildings from Remote Sensing Imagery. *Int. J. Remote Sens.* **2020**, *41*, 5573–5587. [CrossRef]
30. 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]
31. Liu, Y.; Gross, L.; Li, Z.; Li, X.; Fan, X.; Qi, W. Automatic Building Extraction on High-Resolution Remote Sensing Imagery Using Deep Convolutional Encoder-Decoder with Spatial Pyramid Pooling. *IEEE Access* **2019**, *7*, 128774–128786. [CrossRef]
32. 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]
33. Liu, H.; Luo, J.; Huang, B.; Hu, X.; Sun, Y.; Yang, Y.; Xu, N.; Zhou, N. DE-Net: Deep Encoding Network for Building Extraction from High-Resolution Remote Sensing Imagery. *Remote Sens.* **2019**, *11*, 2380. [CrossRef]
34. 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]
35. Cai, J.; Chen, Y. MHA-Net: Multipath Hybrid Attention Network for Building Footprint Extraction from High-Resolution Remote Sensing Imagery. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 5807–5817. [CrossRef]
36. 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.* **2021**, *18*, 895–899. [CrossRef]
37. Hu, Q.; Zhen, L.; Mao, Y.; Zhou, X.; Zhou, G. Automated Building Extraction Using Satellite Remote Sensing Imagery. *Autom. Constr.* **2021**, *123*, 103509. [CrossRef]
38. Abdollahi, A.; Pradhan, B. Integrating Semantic Edges and Segmentation Information for Building Extraction from Aerial Images Using UNet. *Mach. Learn. Appl.* **2021**, *6*, 100194. [CrossRef]
39. Ye, H.; Liu, S.; Jin, K.; Cheng, H. CT-UNet: An improved neural network based on U-Net for building segmentation in remote sensing images. In Proceedings of the 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 10–15 January 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 166–172.
40. Hamaguchi, R.; Hikosaka, S. Building detection from satellite imagery using ensemble of size-specific detectors. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Salt Lake City, UT, USA, 18–22 June 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 223–2234.
41. Guo, H.; Su, X.; Tang, S.; Du, B.; Zhang, L. Scale-Robust Deep-Supervision Network for Mapping Building Footprints from High-Resolution Remote Sensing Images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 10091–10100. [CrossRef]
42. Liao, Y.; Zhang, H.; Yang, G.; Zhang, L. Learning discriminative global and local features for building extraction from aerial images. In Proceedings of the IGARSS 2020—IEEE International Geoscience and Remote Sensing Symposium, Waikoloa, HI, USA, 26 September–2 October 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 1821–1824.
43. Liu, P.; Liu, X.; Liu, M.; Shi, Q.; Yang, J.; Xu, X.; Zhang, Y. Building Footprint Extraction from High-Resolution Images via Spatial Residual Inception Convolutional Neural Network. *Remote Sens.* **2019**, *11*, 830. [CrossRef]
44. 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]
45. 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]
46. Chen, F.; Wang, N.; Yu, B.; Wang, L. Res2-Unet, a New Deep Architecture for Building Detection from High Spatial Resolution Images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 1494–1501. [CrossRef]
47. Xiong, Y.; Chen, Q.; Zhu, M.; Zhang, Y.; Huang, K. Accurate detection of historical buildings using aerial photographs and deep transfer learning. In Proceedings of the IGARSS 2020—IEEE International Geoscience and Remote Sensing Symposium, Waikoloa, HI, USA, 26 September–2 October 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 1592–1595.
48. Liu, Y.; Chen, D.; Ma, A.; Zhong, Y.; Fang, F.; Xu, K. Multiscale U-Shaped CNN Building Instance Extraction Framework with Edge Constraint for High-Spatial-Resolution Remote Sensing Imagery. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 6106–6120. [CrossRef]
49. 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.* **2021**, *59*, 4287–4306. [CrossRef]
50. 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]
51. Abdollahi, A.; Pradhan, B.; Gite, S.; Alamri, A. Building Footprint Extraction from High Resolution Aerial Images Using Generative Adversarial Network (GAN) Architecture. *IEEE Access* **2020**, *8*, 209517–209527. [CrossRef]
52. 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]

53. Girard, N.; Smirnov, D.; Solomon, J.; Tarabalka, Y. Polygonal building extraction by frame field learning. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 5887–5896.
54. Li, W.; Zhao, W.; Zhong, H.; He, C.; Lin, D. Joint semantic-geometric learning for polygonal building segmentation. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 2–9 February 2021.
55. Yin, H.; Zhang, C.; Han, Y.; Qian, Y.; Xu, T.; Zhang, Z.; Kong, A. Improved Semantic Segmentation Method Using Edge Features for Winter Wheat Spatial Distribution Extraction from Gaofen-2 Images. *J. Appl. Rem. Sens.* **2021**, *15*, 028501. [CrossRef]
56. Liao, C.; Hu, H.; Li, H.; Ge, X.; Chen, M.; Li, C.; Zhu, Q. Joint Learning of Contour and Structure for Boundary-Preserved Building Extraction. *Remote Sens.* **2021**, *13*, 1049. [CrossRef]
57. Zhu, Y.; Liang, Z.; Yan, J.; Chen, G.; Wang, X. E-D-Net: Automatic Building Extraction from High-Resolution Aerial Images With Boundary Information. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 4595–4606. [CrossRef]
58. Jiang, X.; Zhang, X.; Xin, Q.; Xi, X.; Zhang, P. Arbitrary-Shaped Building Boundary-Aware Detection with Pixel Aggregation Network. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 2699–2710. [CrossRef]
59. Lee, K.; Kim, J.H.; Lee, H.; Park, J.; Choi, J.P.; Hwang, J.Y. Boundary-Oriented Binary Building Segmentation Model with Two Scheme Learning for Aerial Images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–17. [CrossRef]
60. Jiwani, A.; Ganguly, S.; Ding, C.; Zhou, N.; Chan, D.M. A Semantic Segmentation Network for Urban-Scale Building Footprint Extraction Using RGB Satellite Imagery. *arXiv* **2021**, arXiv:2104.01263.
61. 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]
62. Takikawa, T.; Acuna, D.; Jampani, V.; Fidler, S. Gated-SCNN: Gated Shape CNNs for Semantic Segmentation. *arXiv* **2019**, arXiv:1907.05740.
63. Ma, H.; Yang, H.; Huang, D. Boundary Guided Context Aggregation for Semantic Segmentation. *arXiv* **2021**, arXiv:2110.14587.
64. He, H.; Li, X.; Yang, Y.; Cheng, G.; Tong, Y.; Weng, L.; Lin, Z.; Xiang, S. BoundarySqueeze: Image Segmentation as Boundary Squeezing. *arXiv* **2021**, arXiv:2105.11668.
65. Wang, Y.; Ding, W.; Zhang, R.; Li, H. Boundary-Aware Multitask Learning for Remote Sensing Imagery. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 951–963. [CrossRef]
66. Li, A.; Jiao, L.; Zhu, H.; Li, L.; Liu, F. Multitask Semantic Boundary Awareness Network for Remote Sensing Image Segmentation. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–14. [CrossRef]
67. Jin, Y.; Xu, W.; Zhang, C.; Luo, X.; Jia, H. Boundary-Aware Refined Network for Automatic Building Extraction in Very High-Resolution Urban Aerial Images. *Remote Sens.* **2021**, *13*, 692. [CrossRef]
68. Peng, G.; Jiang, Z.; You, H.; Lu, P.; Hoi, S.; Wang, X.; Li, H. Dynamic Fusion with Intra- and Inter- Modality Attention Flow for Visual Question Answering. *arXiv* **2019**, arXiv:1812.05252.
69. Zhang, T.; Zhuang, Y.; Wang, G.; Dong, S.; Chen, H.; Li, L. Multiscale Semantic Fusion-Guided Fractal Convolutional Object Detection Network for Optical Remote Sensing Imagery. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–20. [CrossRef]
70. Huang, J.; Zhang, X.; Sun, Y.; Xin, Q. Attention-Guided Label Refinement Network for Semantic Segmentation of Very High Resolution Aerial Orthoimages. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 4490–4503. [CrossRef]
71. Gao, P.; Zheng, M.; Wang, X.; Dai, J.; Li, H. Fast Convergence of DETR with Spatially Modulated Co-Attention. *arXiv* **2021**, arXiv:2101.07448.
72. Yi-de, M.; Qing, L.; Zhi-bai, Q. Automated image segmentation using improved PCNN model based on cross-entropy. In Proceedings of the 2004 International Symposium on Intelligent Multimedia, Video and Speech Processing, Hong Kong, China, 20–22 October 2004; IEEE: Piscataway, NJ, USA, 2004; pp. 743–746.