Semantics and Contour Based Interactive Learning Network For Building Footprint Extraction

Xiaoqian Zhu, Xiangrong Zhang, *Senior Member, IEEE,* Tianyang Zhang, Xu Tang, *Member, IEEE,* Puhua Chen, *Member, IEEE,* Huiyu Zhou, Licheng Jiao, *Fellow, IEEE,*

***Abstract*—Building footprint extraction plays an important role in the analysis of remote sensing images and has an extensive range of applications. Obtaining precise boundaries of buildings remains a challenge in existing building extraction methods. Some previous works have made notable efforts to address this concern. However, most of these methods require cumbersome and expensive post-processing steps. Moreover, they ignored the correlation between building semantics and contours, which we believe is crucial for building footprint extraction. To mitigate this issue, our paper presents an intuitive and effective framework that explores semantic and contour cues of buildings and fully excavates their correlation. Specifically, we construct an interactive dual-stream decoder. The Intermediate connections within this decoder interactively transmit features between branches, contributing to learning correlations between semantics and contours. We propose the Semantic Collaboration Module (SCM) to strengthen the connection between the two branches. To further boost performance, we build the Multi-Scale Semantic Context Fusion Module (MSCF) to fuse semantic information from the higher and lower layers of the network, allowing the network to obtain superior feature representations. The experimental results on the WHU, INRIA, and Massachusetts building datasets demonstrate the superior performance of our method. The code will be available at https://github.com/jilaizhizeanzhi/Building-Segmentattion.**

***Index Terms*—Building footprint extraction, deep learning, interactive dual-stream decoder, Semantic and contour collaboration, Multi-Scale feature fusion.**

I. Introduction

BUILDING footprint extraction aims at extracting a building footprint using pixel-based or object-based algorithms [1], which is highly valuable for urban planning, urban change monitoring, 3D modeling of remote sensing image mapping, data acquisition of GIS and updating of urban spatial database, etc. In the past decades, many researchers have made significant efforts to automate building extraction algorithms. Most of the early works are dedicated to defining single or combined criteria [2], [3] that represent the characteristics of a building's appearance, such as uniform color [4], regular shapes [5], and neighboring shadows [6]. However, these manually established criteria are poorly generalized and therefore struggle to handle building footprint extraction in complex scenes.

In recent years, convolutional neural networks (CNN) have been widely practiced in remote sensing image applications. Benefiting from the powerful feature representation capabilities of CNN, many CNN-based building segmentation algorithms [7]–[9] have emerged, treating building extraction as a semantic segmentation challenge. In order to obtain superior segmentation performance, these approaches incorporate attention mechanisms [10]–[13] within the segmentation network. Alternatively, they construct fusion schemes of high and low-level semantic features [14] or design feature memory modules [15] to enhance the feature representation power.

Although CNN-based approaches mentioned above achieve promising performance in building footprint extraction, they fail to deal well with error predictions around buildings. One main group of the follow-up effort attempts to learn the polygonal contours of buildings [13], [16]–[19]. While these methods significantly improve the prediction of building boundaries, they either require laborious and costly post-processing steps or elaborate additional parameters to control the training process of their networks. Another line of work draws on multi-task learning to improve the segmentation of buildings [20]–[22]. Nevertheless, these methods primarily apply boundary-related learning tasks in a straightforward manner to enhance segmentation without delving into deeper correlations. Consequently, the network is not sufficiently efficient, causing limited improvement.

In this paper, we propose an intuitive but effective method to obtain accurate building boundaries, thereby improving the performance of building footprint extraction. We construct an interactive dual-stream decoder consisting of a semantic branch and a contour branch to explore semantic and contour cues and their correlation. The semantic branch focuses on acquiring superior semantic feature representations, while the contour branch captures intricate local details of buildings. The intermediate connections of the decoder (indicated by the purple arrows in Fig. 1) allow these two branches to interactively transmit the learned features to each other to learn the correlation between semantics and contours. Although the intermediate connections facilitates the collaboration of semantic and contour information to some extent, the different task-driven nature of these two branches leads to non-negligible semantic inconsistencies of the features learned by each. Besides, we believe the simple fusion operations (e.g.,

This work was supported in part by the National Natural Science Foundation of China under Grants 62276197, 61871306, and 62171332; the Key Research and Development Program in the Shaanxi Province of China under Grant 2019ZDLGY03-08.

Xiaoqian Zhu,Xiangrong Zhang, Tianyang Zhang, Xu Tang, Puhua Chen, and Licheng Jiao are with the School of Artificial Intelligence, Xidian University.Xi'an 710071, China (e-mail: xrzhang@mail.xidian.edu.cn).

Huiyu Zhou is with the School of Computing and Mathematical Sciences.University of Leicester, University Road, Leicester LE1 7RH, United Kingdom.

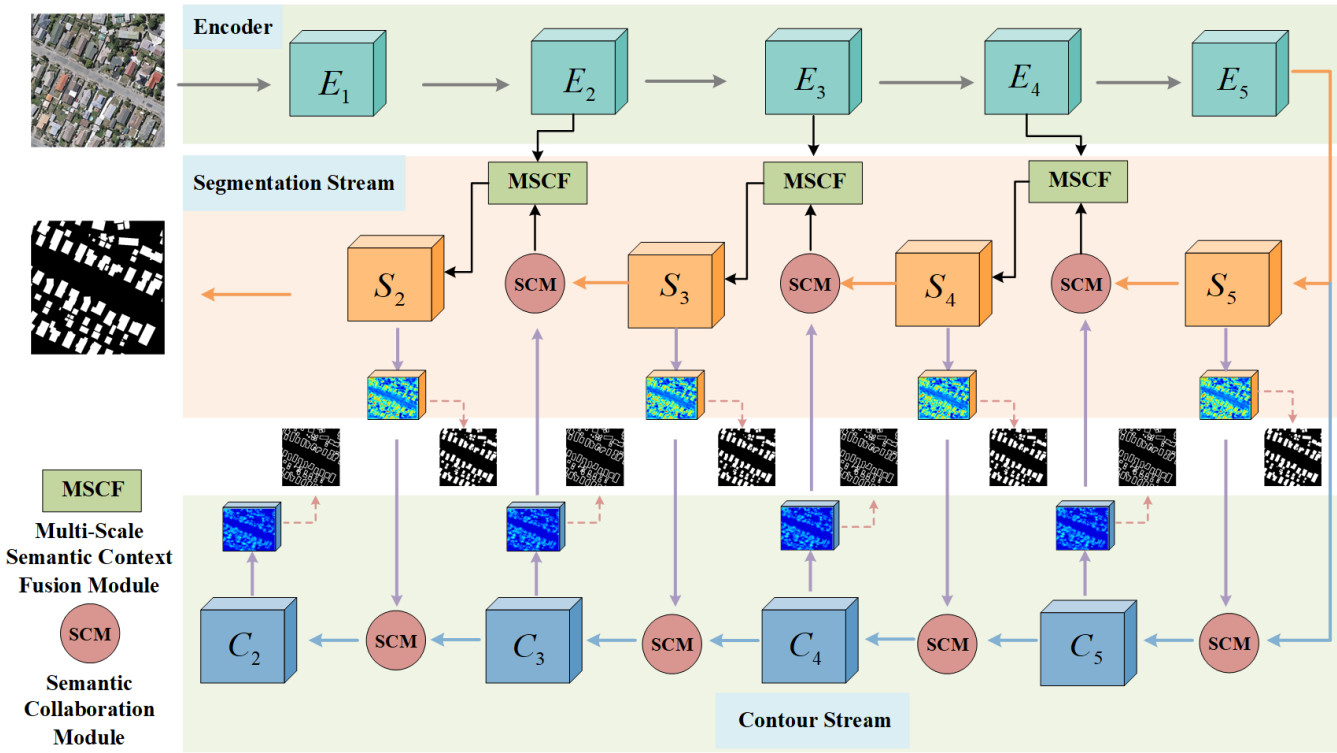


Fig. 1. Illustration of the network structure of our proposed method. Our network employs an encoder-decoder pattern, with the decoder consisting of a semantic branch and a contour branch. The intermediate connections of the decoder (purple arrows) interactively transfer the features learned in one branch to the other branch forcing the network to learn the correlation between semantic and contour cues. The SCM strengthens the connection between the two branches and the MSCF seamlessly integrates multi-scale semantic features.

addition or concatenation) are not optimally leveraging the information from these two branches, and propose the Semantic Collaboration Module (SCM) to strengthen the connection between semantic and contour feature learning. It is widely known that the high-level features tend to have rich semantic information while lacking detailed information, which may lead to incomplete segmentation of the buildings. We thus propose the Multi-Scale Semantic Context Fusion Module (MSCF) to aggregate contextual information of objects at different scales and generate high-resolution features with rich semantic information to alleviate the segmentation errors caused by scale variations of buildings.

The contribution of this paper can be summarised as follows:

1. We propose an interactive dual-stream decoder to explore the correlation between building semantics and contour cues. The intermediate connections of the decoder enable both the semantic and contour branches to interactively transmit their learned features to each other, contributing to the learning of correlations between semantics and contours.

2. We propose the Semantic Collaboration Module (SCM) to build stronger connections between these two task-specific branches and integrate the complementary information among the two branches.

3. We propose the Multi-Scale Semantic Context Fusion Module (MSCF) to seamlessly fuse high-level semantic features with low-level ones, driving the network to obtain high-resolution and strong semantic features to improve performance further.

4. We propose an effective method to improve building boundaries and experiments on the WHU, INRIA, and Massachusetts building datasets demonstrate the effectiveness of our approach, achieving state-of-the-art performance.

II. Related Work

A. Building extraction

In recent years CNN has been widely applied in building footprint extraction tasks. One commonly adopted network structure is the fully convolutional neural network (FCN). Zuo *et al.* [23] achieved arbitrary scale and variable appearance building footprint extraction by merging contextual information from different convolutional layers. In order to obtain finer building features to improve the segmentation performance, MAP-Net [14] progressively extracts high-level semantic features at a fixed resolution in each stage through multiple parallel paths. MS-GeoNet [10] integrates the multi-scale attention module CBAM [24] nested within the CNN network. GCCINet [11] combines CBAM and the Dilated Convolution [25] to design the feature fusion module in the network to fuse high-level features and low-level features across layers. Another prevalent architecture is the encoder-decoder framework, such as the typical U-Net [26]. Ji *et al.* [27] proposed a weight-sharing two-branch U-Net called Siamese U-Net to combine segmentation prediction of the original image and the corresponding down-sampled counterpart, which improved the classification accuracy of large buildings. Khalel *et al.* [28]

employed a cascaded U-Net structure to enhance the prediction results of the network progressively. CrossGeoNet [12] applies a Siamese U-net network with a cross-geolocation attention module to achieve building segmentation. EMMGResUnet [15] constructed a dynamically updated feature memory that stores positive and negative sample features and enhances the feature representation of the building with a form of non-local attention.

In order to obtain sharper building boundaries, some work is dedicated to extracting the polygonal contours of buildings. Wei *et al.* [18] introduced multi-scale aggregation and two post-processing strategies to obtain a segmentation map in FCN, on which a polygon regularisation algorithm consisting of coarse and fine adjustments was applied to refine the building outlining. TDAC [19] combined convolutional neural networks and Active Contour Models (ACMS) [29] to construct a trainable and end-to-end framework to obtain sharp contours of buildings. Girard *et al.* [13] proposed a post-processing polygonization algorithm to improve building boundaries by leveraging the output of a frame field additionally learned by the network. However, these methods require laborious post-processing steps or additional prior parameter settings. Other works improve the segmentation of building boundaries with the help of multi-task learning. DMBC-Net [20] designed frameworks for semantic segmentation, direction prediction, and distance estimation for three learning tasks to assist the network in segmenting buildings. He *et al.* [21] added a boundary learning task to the segmentation network and improved the building extraction by combining the predicted boundary mask with semantic features. CBR-Net [22] combined the edge prediction to refine the boundaries of buildings in a coarse-to-fine manner progressively and improved the boundaries by perceiving the orientation of pixels to the center of the nearest object in the network. In contrast to these multi-task learning approaches, our method improves the segmentation performance of building boundaries by exploring the correlation between tasks.

B. Boundary for Segmentation

Mis-classification of pixels tends to occur at the boundaries of objects. Many semantic segmentation-related works have achieved remarkable performance by obtaining more accurate object boundaries. BFP [30] learned edges as an additional semantic category to guide the pixel-to-pixel connections of recurrent undirected graph neural networks, increasing the similarity of features between objects of the same category while attenuating the strength of pixel connections between objects of different categories. Gated-SCNN [31] explicitly treated shape information as a separate stream and used a new gating mechanism to connect the intermediate layers. Shape streams collaborate with each other to obtain more precise segmentation predictions. Li *et al.* [32] decomposed the feature map into two parts, body features, and edge features, and dealt with semantic segmentation by explicitly modeling body consistency and edge preservation at the feature level and then jointly optimizing them in a unified framework. SegFix [33] proposed a novel model agnostic post-processing mechanism by replacing the labels of boundary pixels with the labels of the corresponding internal pixels to reduce the boundary errors of the segmentation results. Unlike the above work, in our framework, the constant interaction of contour cues and semantic information during the decoding process leads to stronger and more discriminative feature representations, significantly reducing the misclassification of the background and obtaining finer boundaries.

C. Multi-task Learning

Several existing works have confirmed the effectiveness of jointly complementary multi-task learning [20] [21] [34] [35]. The Dual Super-Resolution Learning [36] (DSRL) framework enhances the detailed information of features by incorporating super-resolution task learning to obtain high-resolution representations and performs well in semantic segmentation and human pose estimation. Wu et al. [37] learned both semantic categories and affinity relations of pixels to produce a stronger feature representation, and the output affinity information can also be applied to refining the original segmentation predictions. In this paper, we combine boundary prediction and semantic segmentation and utilize the relationship between them to improve segmentation. Zhen et al. [38] propose a joint multi-task learning framework by designing the iterative Pyramid Context Module to couple semantic segmentation and semantic boundary detection tasks. MCN [39] performs the three basic tasks of object detection, semantic segmentation, and human pose estimation simultaneously with a single network. OmniDet [40] is aimed at a multitasking visual perception system for circumferential fisheye lenses, integrating multiple tasks of depth estimation, visual odometry, semantic and motion segmentation, object detection, and lens contamination detection. YOLOP [41] performs three tasks simultaneously: traffic object detection, driveable area segmentation, and lane detection.

III. METHODOLOGY

A. Overview

Many existing CNN-based building footprint extraction methods have struggled to obtain accurate boundaries of buildings, and the existing methods for improving building boundaries are either complicated in modeling or insufficient in information utilization. In this paper, we propose simple yet effective methods to alleviate the problem of blurred boundaries. Fig. 1 illustrates the overall structure of our proposed method. Following the CNN-based building segmentation method, our proposed network adopts the encoder-decoder pattern, where ResNet50 [42] is employed as the encoder, and the decoder only uses the last four stages of ResNet to reduce the computational cost. The decoder consists of two branches, a semantic branch aiming at capturing rich semantic information to predict the segmentation map, and a contour branch towards mining out local detail information to predict the contour map. The intermediate connection of the decoder allows interaction between the two cues, prompting the network to learn the correlation between them. The two

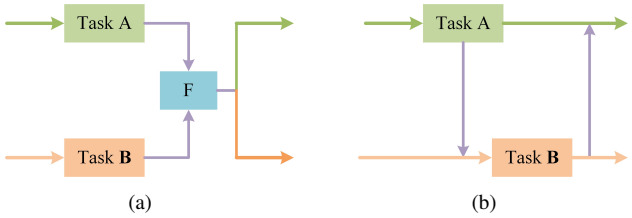


Fig. 2. A comparison of our method with other multi-task learning approaches. Figures (a) and (b) show the learning patterns of the other methods and our method respectively.

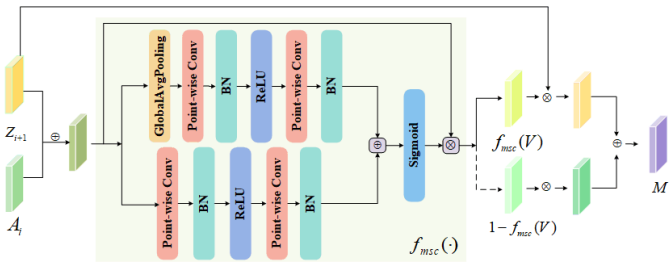


Fig. 3. Illustration of the Multi-Scale Semantic Context Fusion Module structure.

branches focus on learning distinct cues, resulting in the inherent semantic inconsistency between the two learned features. Therefore, we propose the Semantic Collaboration Module (SCM) to enhance the network's collaborative exploitation of semantic and contour information. Semantic features from the shallow layers carry more precise location information but suffer from semantic noise. We further propose the Multi-Scale Semantic Context Fusion Module (MSCF) to facilitate the network fusing high-level semantic features and low-level detail features. The proposed components seamlessly integrate different semantic information in the network, which strikes a common goal of the network to learn stronger discriminative feature representations from semantic and contour cues.

B. Interactive Dual-Stream Decoder

In our decoder, the semantic and contour branches are responsible for learning corresponding features while supervised by two distinct tasks: segmentation prediction and contour detection. Instead of following the classical multi-task learning approach [37] [21], where the two tasks are independently learned and subsequently merged through a fusion module, our decoder promotes interactive cooperation between these tasks. We illustrate the distinctions between our interactive decoder and other multi-task learning approaches, refer to Fig. 2. In our approach, each branch of the decoder continuously transmits its learned features to the other branch, facilitating the interactive integration of information into the learning process of both tasks. The interaction enhances the complementarity between the features learned by each task and promotes the correlation between the semantic and contour cues acquired by the network. In Fig. 1, the purple arrows indicate the interaction of information between the semantic and contour branches.

We begin by describing the workflow of the decoder. Initially, the feature maps from the last four stages of the encoder undergo a 1×1 convolution layer, which reduces the number of channels to 256, resulting in the creation of the feature A_i:

$$A_i = \mathrm{Conv}_{1\times1}\left(E_i\right) \qquad i=2,\ 3,\ 4,\ 5 \tag{1}$$

Here, E_i denotes the feature maps from stage i of the encoder.

To transmit the learned features, the feature S_i of the semantic branch and C_i of the contour branch from stage i are passed through a channel pooling layer, and a convolution layer to generate the intermediate features S_i' and C_i', respectively, as well as to generate the corresponding prediction map:

$$S_i' = f_{s_i}(\mathrm{cp}\left(S_i\right)), P_i^s = f_{s_i'}\left(S_i'\right) \tag{2}$$

$$C_i' = f_{c_i}(\mathrm{cp}\left(C_i\right)), P_i^c = f_{c_i'}\left(C_i'\right) \tag{3}$$

where the f_* denotes the convolution operation, and the P_i^s and the P_i^c denotes the segmentation prediction map and contour prediction map for the current stage i, respectively.

$$\mathrm{cp}(X) = \mathrm{collect}_{j\in[0,m-1]}\left(\max_{k\in\left[0,\frac{n}{m}-1\right]} X^{j\times\frac{n}{m}+k}\right) \tag{4}$$

where j and k are integers and n and m are the number of the input and output channels of the features, respectively. The channel pooling layer divides the n-dimensional channels of the original feature map into m groups and takes the maximum value of each group as the channel value of the output feature map.

The intermediate semantic feature S_i' and contour feature C_i' from stage i are fed into the SCM to generate the primary semantic features Z_i:

$$Z_i = \mathrm{SCM}\left(S_i', C_i'\right) \tag{5}$$

The feature Z_i is fused with the encoder features A_i from the lower layer by the MSCF to obtain the semantic features of the subsequent stage $i-1$:

$$S_{i-1} = \mathrm{MSCF}\left(Z_i, A_{i-1}\right) \tag{6}$$

The intermediate semantic feature S_{i-1}' is fed into the SCM with contour features C_i from the subsequent stage i to learn the required contour features of the stage $i-1$:

$$C_{i-1} = \mathrm{SCM}\left(\left(S_{i-1}'\right), upsample(C_i)\right) \tag{7}$$

We repeat the process described in Eq. (2)-(7) until we obtain the semantic features S_2 that are used to generate the final network prediction.

It is important to note that while the intermediate connections within the decoder utilize convolutional layers to learn the correlation between semantics and contours, it is inappropriate to equate correlation operations with traditional convolution operations. This is because the proposed whole network, including the intermediate connections and the Semantic Collaboration Module, collaboratively facilitates the model to learn and integrate semantic and contour correlation.

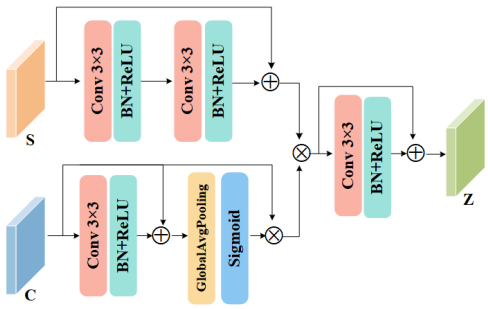


Fig. 4. Illustration of the Semantic Collaboration Module structure.

C. Semantic Collaboration Module

Given that the two branches of the decoder are designed to learn distinct cues, a notable semantic disparity naturally emerges between these two sets of learned features. We believe the information on the two features cannot be properly exploited in a simple and straightforward fusion strategy. In order to fully leverage the two different cues, we propose the Semantic Collaborati on Module (SCM) to integrate the two distinct features and strengthen the correlation between the semantic branches and the contour branch.

In the following section, we begin to describe the workflow of SCM. The semantic features S transform the learned semantic features through successive convolutional layers and apply residual connections to strengthen the correlated semantic information, as shown in the upper branch of Fig. 4, which can be represented by the following formula:

$$S_{co} = \delta(\beta(Conv_{3\times3}(\delta(\beta(Conv_{3\times3}(s)))))) \oplus S \tag{8}$$

where $\oplus$ denotes the element-wise addition and $Conv$ denote convolution operation, and δ and β denote ReLU activation and batch normalization, respectively.

The contour feature C is processed through the lower branch of the SCM module to process the contour information, which is expressed by the following formula:

$$C_{co} = \sigma(\mathrm{GAP}(\delta(\beta(Conv_{3\times3}(C) \oplus C)))) \otimes C \tag{9}$$

where $\otimes$ denotes the element-wise multiplication, GAP denotes Global Average Pooling, and σ denotes the sigmoid activation function.

Finally, the outputs of the two branches of the SCM module are multiplied together and passed through a convolutional layer as well as batch normalization layers and Rectified Linear Unit to obtain the fused features Z:

$$Z = (\delta(\beta(Conv_{3\times3}(S_{co} \otimes C_{co})))) \oplus (S_{co} \otimes C_{co}) \tag{10}$$

D. Multi-Scale Semantic Context Fusion Module

A few networks [26] [43] [44] apply long skip connections to bridging fine-detail features from the lower layers and coarse-resolution high-level semantic features, we follow this idea to help the network obtain high-resolution semantic features. However, long skip connections impose cross-scale semantic information, leading to misalignment between semantic features at different scales. To mitigate the above problem, we constructed the MSCF to promote the full exploitation of high and low-level semantic information by the network.

MSCF first aggregates features from both high and low scales and assigns weighting factors to features at each scale through a multi-scale contextual aggregator, which works through two paths: global contextual aggregation and local contextual aggregation. Global context aggregation provides contextual guidance cues from the overall perspective of features, while local context aggregation obtains semantic contextual information from a local perspective. The contextual guidance obtained from the two paths generates weighting factors to adjust the semantic dependencies between features, thus enabling the fusion of semantic features at different scales. The specific process is as follows:

The feature Z_{i+1} from the higher stage is upsampled and then summed with the encoder feature A_i from the current stage by element-wise addition to obtain the composite feature V_i :

$$V = Z_{i+1} \oplus A_i' \tag{11}$$

The feature V_i is then fed into a multi-scale contextual aggregator via a global contextual aggregation path and a local contextual aggregation path, which can be represented by the following formula:

$$f_{msc}(V) = V \otimes \sigma(L(V) \oplus G(V)) \tag{12}$$

where $G(V)$ is the global contextual aggregator and $L(V)$ is the local contextual aggregator. Specially, $L(V)$ can be formulated as follows:

$$L(V) = \beta\left(f_{PW_2}\left(\delta\left(\beta\left(f_{PW_1}(V)\right)\right)\right)\right) \tag{13}$$

where f_{PW} denotes point-wise convolution to obtain the interaction of point-wise channels at each spatial location. The kernel sizes for f_{PW_1} and f_{PW_2} are $\frac{C}{r} \times C \times 1 \times 1$ and $C \times C \times 1 \times 1$ respectively, where r is the channel scaling rate and C is the number of the input channels.

And $G(V)$ can be formulated as follows:

$$G(V) = \beta\left(f_{PW_2}\left(\delta\left(\beta\left(f_{PW_1}(GAP(V))\right)\right)\right)\right) \tag{14}$$

Finally, we utilize the obtained multi-scale contextual information to adjust the semantic dependencies of the features to obtain the integrated multi-scale semantic features M, represented by the following equation:

$$M = (f_{msc}(V) \otimes Z_{i+1}) \oplus ((1 - f_{msc}(V)) \otimes A_i) \tag{15}$$

Fig. 3 illustrates the structure of MSCF where the dashed lines indicate the $(1 - f_{msc}(V)) \otimes A_i$.

E. Loss Function

Our decoder has two branches, e.g. the segmentation and contour branches, and we use binary cross-entropy loss:

$$L = \sum_{i=2}^{5} L_{BCE}(y_i, \hat{y}_i) + \sum_{i=2}^{5} L_{BCE}(c_i, \hat{c}_i) \tag{16}$$

$$L_{BCE}(x,y) = -\frac{1}{HW}\sum_{k=1}^{HW}[y_k \log(x_k) + (1-y_k)\log(1-x_k)] \quad (17)$$

where $\hat{y}$ and $\hat{c}$ denote the segmentation and contour prediction maps of the network, y and c are the corresponding Ground Truth of the segmentation outcome and the boundary. During the network training process, for each segmentation prediction map P_i^s and contour prediction map P_i^c in Eq.2 and Eq.3, it is supervised by the Ground Truth of the building segmentation map and the boundary Ground Truth of the building.

IV. EXPERIMENT AND ANALYSIS

A. Datasets

Our experiments are conducted on three publicly available building footprint extraction datasets, including the WHU building dataset [27], the INRIA dataset [45], and the Massachusetts buildings dataset [46].

The WHU building dataset includes aerial and satellite subdatasets along with the corresponding shape files and the raster masks. In this experiment, we use the aerial data, including more than 1870000 buildings with a resolution of $3cm$ and covering a ground area of more than 450 km^2. It contains rural, residential, cultural, and industrial areas, with a variety of building types, different colors, and sizes. The aerial image set is seamlessly cropped into 8189 tiles of 512 $\times$ 512 pixels, including 4736 tiles for the training set, 1036 tiles for the validation set, and 2416 tiles for the test set.

The INRIA dataset consists of 360 orthorectified aerial images of size 5000×5000 pixels with a spatial resolution of 0.3 m/pixel. The training set of 180 images covers the Austin, Chicago, Kitsap County, West Tyrol, and Vienna regions, while the test set of 180 images covers Bellingham, Bloomington, Innsbruck, San Francisco, and East Tyrol. The data were obtained on different flights across the country, covering different areas with different lighting characteristics, and landscapes and settlements varying in shape, density, and appearance, so the dataset provides a thorough assessment of the generalization ability of the different models. For a fair comparison, we split the training set according to the guidelines shown in [45] and remove the top five images of each region from the training set for validation. Because the Ground Truth of the test set is agnostic, we only test the algorithm on the validation set.

The Massachusetts buildings data set consists of 151 RGB aerial images of the area at Boston, where each image has 1500×1500 pixels and covers an area of $2.25km^2$ with a ground spatial resolution of 1 m/pixel. The entire dataset covers roughly 340 square kilometers. The dataset was randomly divided into a training set, a validation set, and a test set with 137, 4, and 10 images, respectively.

B. Evaluation Metric

Since using a segmentation-based approach, we use pixel-level evaluation methods such as IoU and F1-score which are more suitable for evaluating our algorithm. The metrics are calculated as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}}$$

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

$$\text{IoU} = \frac{\text{Precision } \times \text{ Recall}}{\text{Precision } + \text{ Recall } - \text{ Precision } \times \text{ Recall}}$$

where TP is the number of positive samples correctly predicted, FP is the number of positive samples incorrectly predicted and FN is the number of negative samples incorrectly predicted. Precision is the proportion of the correctly identified samples among those identified and Recall is the percentage of the correctly identified samples among all the positive samples. F1 is a weighted summation average of Precision and Recall, taking into account the combination of completeness and accuracy. The IoU is the ratio of the intersection of the prediction with the Ground Truth and the union of the prediction with the Ground Truth. To ensure a fair comparison, we followed the established conventions by most previous building extraction methods [11], [18], [20]–[22], [27], [47]–[50], using Precision, Recall, F1, and IoU as metrics to measure model performance. Following [46], we also use the break-even point metric for the Massachusetts data set, where the break-even point is the point on the precision-recall curve where precision is the same as recall.

C. Implementation

Following previous CNN-based building segmentation methods, the backbone used in our method is ResNet-50 for all the experiments, which were pre-trained on ImageNet [51]. For all the experiments, these models were trained with the SGD optimizer and with a "poly" learning rate policy, where the initial learning rate was set to 0.01 and multiplied by $\left(1 - \frac{step}{max-step}\right)^{\text{power}}$ with power = 0.9. We used synchronized SGD over 2 GPUs with a total of 8 images per mini-batch (4 images per GPU), weight decay of 0.0001 and momentum of 0.9. The synchronized batch normalization was used for cross-gpu communication of statistic in the batch normalization layer. During training, the images are randomly transformed in brightness and contrast with a probability of 0.5. The code will be available at https://github.com/jilaizhizeanzhi/Building-Segmentattion.

D. Comparison With Existing Methods

We would like to clarify that, in order to maintain fairness, we have adhered to the convention established by previous work [11], [14], [21], [22], [48]–[50] of solely visualizing results through comparisons with state-of-the-art CNN-based semantic segmentation methods, as shown in Fig. 5, 6, 7, and 8. This is because the source code of most previous building extraction algorithms is not publicly available. In addition, in order to demonstrate the effectiveness of our method

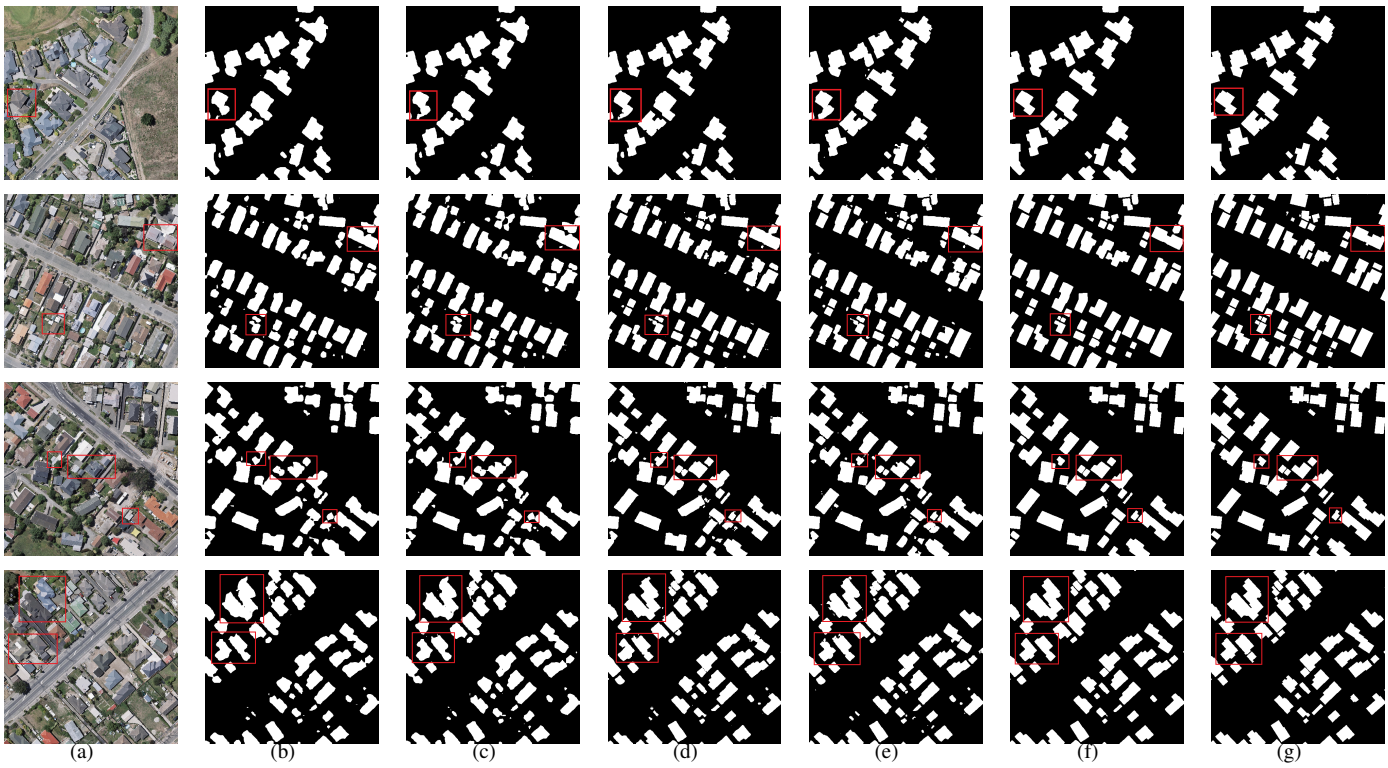


Fig. 5. Exemplar results of PSPNet, Deeplabv3+, HRNet, UNet, and our method for detecting small buildings in the WHU building dataset. (a) Original image, (b) PSPNet, (c) Deeplabv3+, (d) HRNet, (e) UNet, (f) Ours, and (g) Ground Truth.

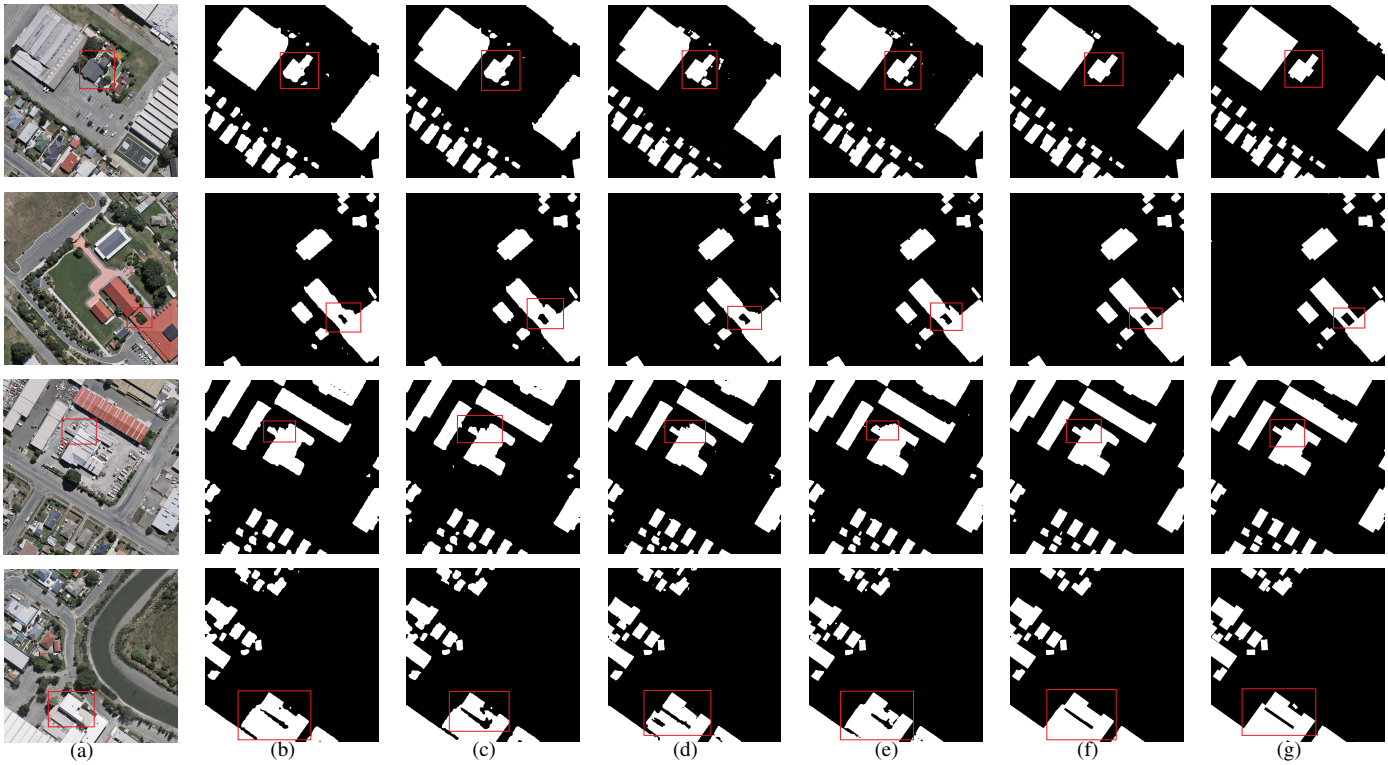


Fig. 6. Exemplar results of PSPNet, Deeplabv3+, HRNet, UNet, and our method for detecting large buildings in the WHU building dataset. (a) Original image, (b) PSPNet, (c) Deeplabv3+, (d) HRNet, (e) UNet, (f) Ours, and (g) Ground Truth.

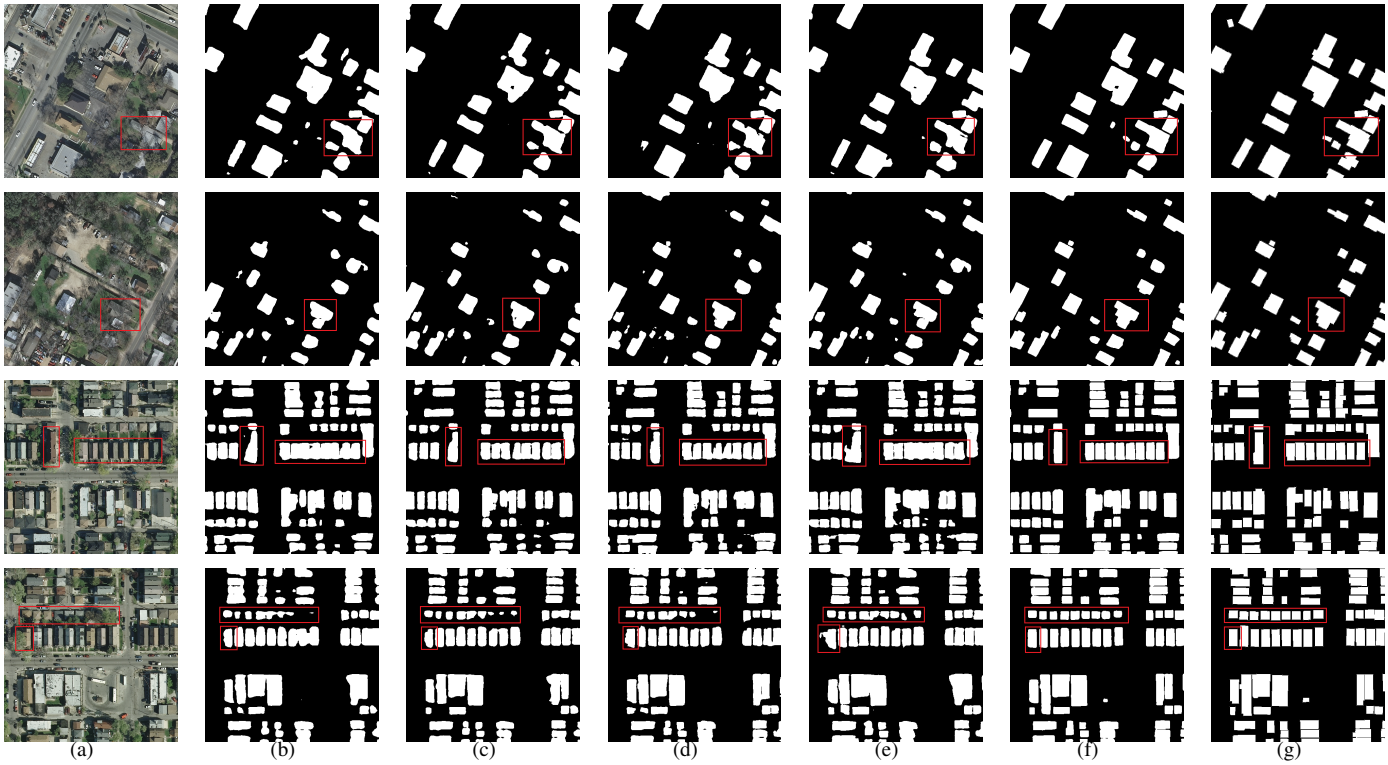


Fig. 7. Exemplar results of PSPNet, Deeplabv3+, HRNet, UNet, and our method for the INRIA building dataset. (a) Original image, (b) PSPNet, (c) Deeplabv3+, (d) HRNet, (e) UNet, (f) Ours, and (g) Ground Truth.

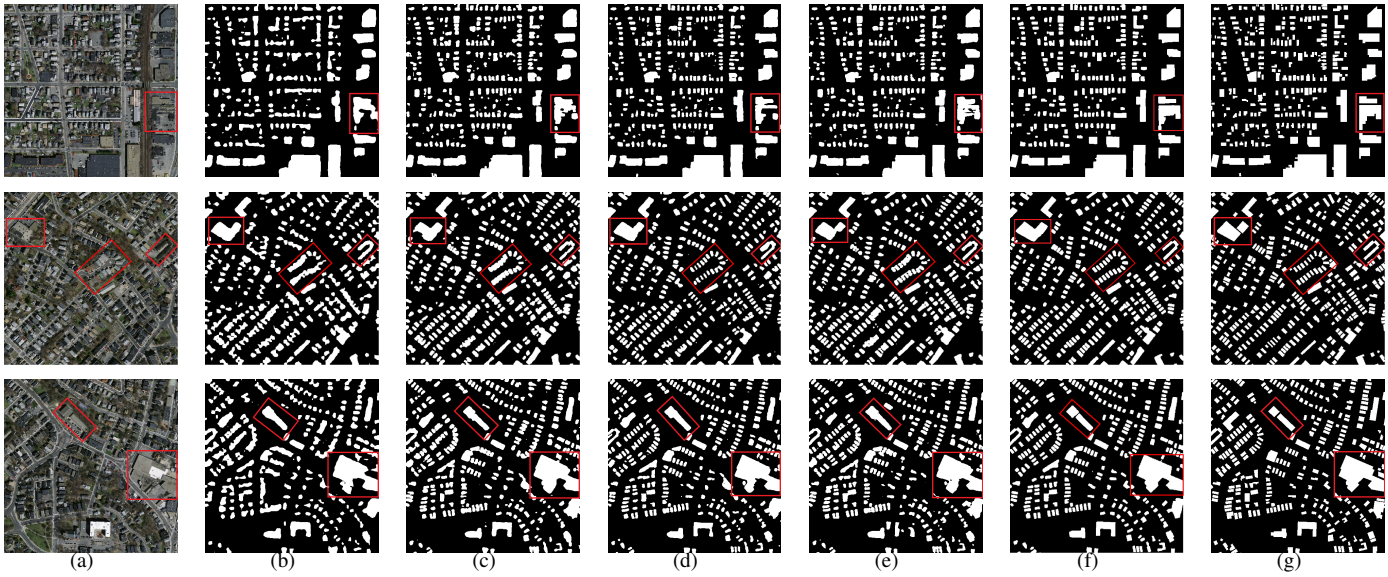


Fig. 8. Exemplar results of PSPNet, Deeplabv3+, HRNet, UNet and our method for the Massachusetts building dataset. (a) Original image, (b) PSPNet, (c) Deeplabv3+, (d) HRNet, (e) UNet, (f) Ours and (g) Ground Truth.

Table I
COMPARISON OF THE STATE-OF-THE-ART METHODS AND OURS ON THE WHU DATA SET

Method	IoU(%)	Precision(%)	Recall(%)	F1-score(%)
PSPNet	88.13	94.21	93.18	93.69
DeepLabV3+	89.06	94.82	93.61	94.21
HRNet48	89.31	94.34	94.37	94.35
CU-Net	87.10	94.60	91.70	93.13
SiU-Net	88.40	93.80	93.90	93.85
SRI-Net	89.23	95.67	93.69	94.51
DE-Net	90.12	95.00	94.60	94.80
EU-Net	90.56	94.98	95.10	95.04
MA-FCN	90.70	95.20	95.10	95.15
MAP-Net	90.86	95.62	94.81	95.21
He *et al*	90.5	95.1	94.9	95.0
GCCINet	90.83	95.16	95.23	95.20
CBR-Net	91.4	95.31	95.7	95.51
DMBC-Net	91.66	96.15	95.16	95.65
UNet(Baseline)	88.62	93.97	93.96	93.96
Ours	**91.94**	**96.24**	**95.37**	**95.80**

Table II
COMPARISON OF THE STATE-OF-THE-ART METHODS AND OURS ON THE INRIA DATASET.

Method	IoU(%)	Precision(%)	Recall(%)	F1-score(%)
PSPNet	77.33	88.57	85.91	87.22
DeepLabV3+	77.19	89.19	85.15	87.13
HRNet48	77.98	88.47	86.80	87.63
SRI-Net	71.76	85.77	81.46	83.56
Sheng et al	77.2	83.5	**91.1**	87.1
EU-Net	80.50	90.28	88.14	89.20
MFCNN	79.35	88.58	87.91	88.38
DS-Net	80.73	-	-	-
DMBC-Net	80.74	89.94	88.77	89.35
GCCINet	78.88	89.09	87.31	88.19
CBR-Net	81.1	89.93	89.2	89.56
UNet(Baseline)	77.05	88.00	86.10	87.04
Ours	**82.48**	**91.86**	88.99	**90.40**

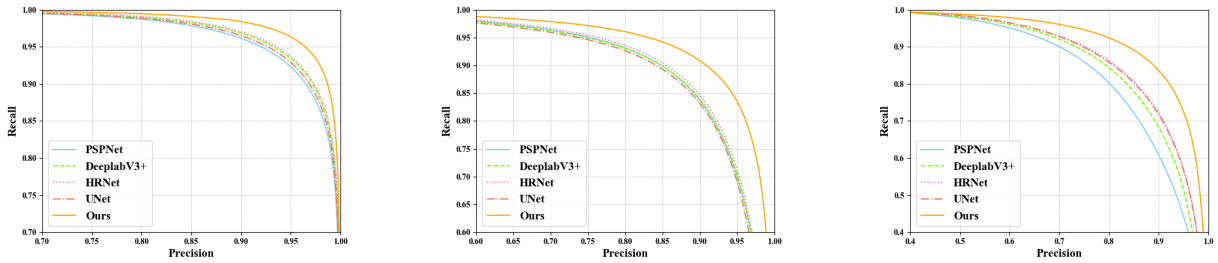
more clearly, we also provide a comparison of the precision-recall curves of these semantic segmentation methods and our method on different datasets, as shown in Fig. 9. However, we have conducted a comprehensive comparison of the numerical outcomes.

1) Results on the WHU Buildings Dataset: Table I shows the quantitative results of the WHU buildings dataset. UNet is the baseline of our approach, which employs a symmetrical encoder-decoder structure. PSPNet provides effective global contextual priors by fusing pooling features at different scales through the Pyramid Pooling Module (PPM), effectively expanding the network's field of attention. HRNet obtains high resolution semantic feature maps by concatenating feature maps of different resolutions and interacting between them. DeepLabV3+ utilizes an Atrous Spatial Pyramid Pooling (ASPP) module to obtain contextual information and merge high and low-level features to obtain a discriminative feature representation. It is clear from these tables that the performance of our approach surpasses these advanced semantic segmentation methods by a wide margin. This observation can be attributed to the fact that these advanced semantic segmentation methods are not designed for building characteristics. In contrast, our approach takes into account both the semantic and contour information of buildings and fully leverages these two distinct characteristics. On the WHU dataset, the IoU of our method exceeds UNet, PSPnet, Deeplabv3+, and HRNet by 3.32%, 3.81%, 2.88%, 2.63%, and F1 scores by 1.84%, 2.11%, 1.59%, 1.45% respectively.

To validate the effectiveness of our method, we also present a comparison of our method with the existing building footprint extraction methods on the WHU dataset, including CU-Net [47], SiU-Net [27], SRI-Net [48], DE-Net [49], EU-Net [50], MA-FCN [18], He *et al*. [21], CBR-Net [22], GCCINet [11], DMBC-Net [20] and MAP-Net [14]. CU-Net introduces the supervision of GT in the middle layer of the decoder of UNet to constrain and optimize the network parameters. SiU-Net employs a weight-sharing two-branch U-Net to combine segmentation prediction of the original image and the corresponding down-sampled counterpart. SRI-Net designs a spatial residual inception module to progressively fuse features from multiple layers to produce multi-scale contexts. DE-Net also utilizes the encoder-decoder architecture and introduces some segmentation techniques to improve the segmentation results. EU-Net designed the dense spatial pyramid pooling (DSPP) module to acquire multi-scale features and optimize the network with a focal loss. MA-FCN [18] first uses CNN to segment the buildings and then converts the segmentation maps into structured individual building polygons using an empirical polygon regularisation. GCCINet [11] combines the CBAM and the Dilated Convolution [25] to design the feature fusion module fuse features across layers. He *et al*. [21] introduced a boundary learning task by performing a spatial variation operation on the segmentation map to assist the network in maintaining the building boundaries. MAP-Net extracted high-level semantic features with a fixed resolution at each stage step-by-step through multiple parallel paths. CBR-Net [22] combined the edge prediction to progressively refine the boundaries of buildings in a coarse-to-fine manner.

It is clear from Table I that our approach achieves an IoU of 91.94% and an F1 score of 95.80%, outperforming the other state-of-the-art algorithms by a large margin. It is worth noting that He *et al*. and DMBC-Net [20] also used boundary learning as an auxiliary task to improve the boundaries, but their method does not dig into and exploit the correlation between building semantics and contours, therefore the IoU of our method outperforms the IoU of He *et al*. and DMBC-Net [20] 1.44% and 0.28%, respectively. Our method also surpasses these Uet-based network structures, such as CU-Net, SiU-Net, and DE-Net.

Fig. 5 vividly illustrates the effectiveness of our method in detecting the boundaries of small buildings in the WHU dataset. In the second row of Fig. 5, it becomes apparent that for some tiny and densely clustered buildings, other algorithms struggle to extract buildings with well-defined contours. In contrast, our method excels by yielding distinctly separated buildings with clear and continuous boundaries. In the third and fourth rows of Fig. 5, whereas other segmentation algorithms obtain uneven borders for the buildings with more curved contours and abundant corners, our method can obtain sharp boundaries due to the complementary information of contour details. In Fig. 6, we show the performance of different algorithms for large buildings in WHU. Boundary acquisition

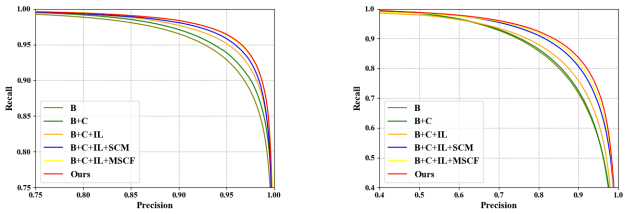


(a) WHU building dataset (b) INRIA building dataset (c) Massachusetts building dataset

Fig. 9. Figures a, b, and c depict the precision-recall curve comparisons of PSPNet, Deeplabv3+, HRNet, UNet, and our proposed method on the WHU, INRIA, and Massachusetts buildings datasets.

Table III
COMPARISON OF THE STATE-OF-THE-ART METHODS AND OURS ON THE MASSACHUSETTS DATASET.

Method	Breakeven ($\rho = 3$)	F1 ($\rho = 3$)	Breakeven ($\rho = 0$)	F1 ($\rho = 0$)	IoU(%)
PSPNet	0.9492	0.9438	0.8009	0.7970	66.25
DeepLabV3+	0.9501	0.9495	0.8201	0.8200	69.49
HRNet48	0.9558	0.9530	0.8309	0.8271	70.52
Mnih *et al.*	0.9292	0.9092	0.7632	0.7407	-
Saito *et al.*	0.9528	0.9441	0.8082	0.7919	-
HF-FCN	0.9643	0.9620	0.8479	0.8373	-
Building-A-Nets	0.9677	0.9656	0.8503	0.8478	-
RA-FCN	0.9608	0.9605	0.8362	0.8340	-
DS-Net	0.9690	0.9672	0.8569	0.8549	-
UNet(Baseline)	0.9492	0.9540	0.8010	0.8256	70.30
Ours	**0.9704**	**0.9704**	**0.8661**	**0.8694**	**76.89**



(a) (b)

Fig. 10. Figures (a) and (b) correspond to the precision-recall curves for the WHU dataset and the Massachusetts building dataset under different component configurations, respectively.

for large buildings is relatively more accessible due to their size. However, the outlined boundaries are blurred due to the occlusion of trees, shadows, and interference from the appearance of similar objects. We observe from Fig. 6 that the contour information can better assist the network in solving these problems.

2) Results on the INRIA Buildings Dataset: Methods of using the INRIA datasets include SRI-Net [48], EU-Net [50], MFCNN [52], DS-Net [53], CBR-Net [22], He *et al.* and DMBC-Net [20]. MFCNN obtains a pixel-level segmentation map using a fully convolutional neural network and then uses morphological filtering to refine the building boundaries. DS-Net captures local and long-range information using a two-branch UNet structure with a shared encoder. DMBC-Net introduces two auxiliary tasks, boundary prediction and distance estimation, and designs two consistency losses to alleviate the boundary ambiguity. As shown in Table II, our approach achieves an IoU of 82.48% and an F1 score of 90.40% on the INRA dataset, with an IoU metric of 3.13% higher and an F1 score of 2.02% higher than that of MFCNN using morphological filtering to refine the boundaries. Although DMBC-Net uses two auxiliary tasks to optimize the boundaries, it does not take full advantage of the correlation between the tasks, and our approach achieves much better performance than this due to a more complementary learning approach. Compared with the CNN-based semantic segmentation methods, the IoU of our method exceeds UNet, PSPnet, Deeplabv3+, and HRNet by 5.43%, 5.15%, 5.29%, 4.5%, and F1 scores by 3.36%, 3.18%, 3.27%, 2.77% respectively.

Fig. 7 shows an example of the results from the INRIA dataset. In the INRIA dataset, buildings are heavily obscured by vegetation (the fourth row of Fig. 7) and some densely arranged buildings (the third row of Fig. 7), making it challenging to detect building boundaries. While the other methods fail to extract buildings under such challenging conditions, our method can still outline fairly complete buildings due to the incorporation of building contour information. Notably, when dealing with densely arranged buildings, our method successfully separates individual buildings and delineates clear and distinct boundaries, a feat that other methods fail to achieve.

3) Results on the Massachusetts buildings dataset: We followed the common evaluation metrics of the Massachusetts building dataset. For computing the Precision-Recall breakeven point and F1 score, we employed a relaxed version of precision and recall [46]. The relaxed precision is defined as the proportion of predicted positive pixels that are within ρ pixels of the Ground Truth positive pixels, whereas the relaxed recall is defined as the proportion of Ground Truth positive pixels that are within ρ pixels of predicted positive pixels. The relaxed parameter ρ is typically set to 3. These methods of using the Massachusetts dataset include HF-FCN [23], Building-A-Nets [54], RA-FCN [55], DS-Net [53], MTMS [56], GAN-SCA [57], Mnih *et al.* [46] and Saito *et al* [58]. As can be seen from Table III, our method achieves the best performance at the strictest setting ($\rho = 0$), with F1 reaching 0.8694 and Break-even reaching 0.8661. With the setting ($\rho = 3$), our algorithm also performs better than

Table IV
EVALUATION RESULTS FOR OUR METHOD WITH DIFFERENT COMPONENTS ON THE WHU DATASET AND MASSACHUSETTS DATASET.

Component				WHU		MASSACHUSETTS	
C	IL	SCM	MSCF	IoU(%)	F1-score(%)	IoU(%)	F1-score(%)
				88.62	93.96	70.3	82.56
✓				89.26	94.33	70.89	82.97
✓	✓			90.66	95.10	72.24	83.88
✓	✓	✓		91.15	95.37	74.99	85.17
✓	✓		✓	91.81	95.73	75.90	86.30
✓	✓	✓	✓	**91.94**	**95.80**	**76.89**	**86.94**

the other algorithms. Compared with state-of-the-art CNN-based semantic segmentation methods, the IoU of our method exceeds UNet, PSPnet, Deeplabv3+, and HRNet by 6.59%, 10.64%, 7.40%, and 6.37% respectively.

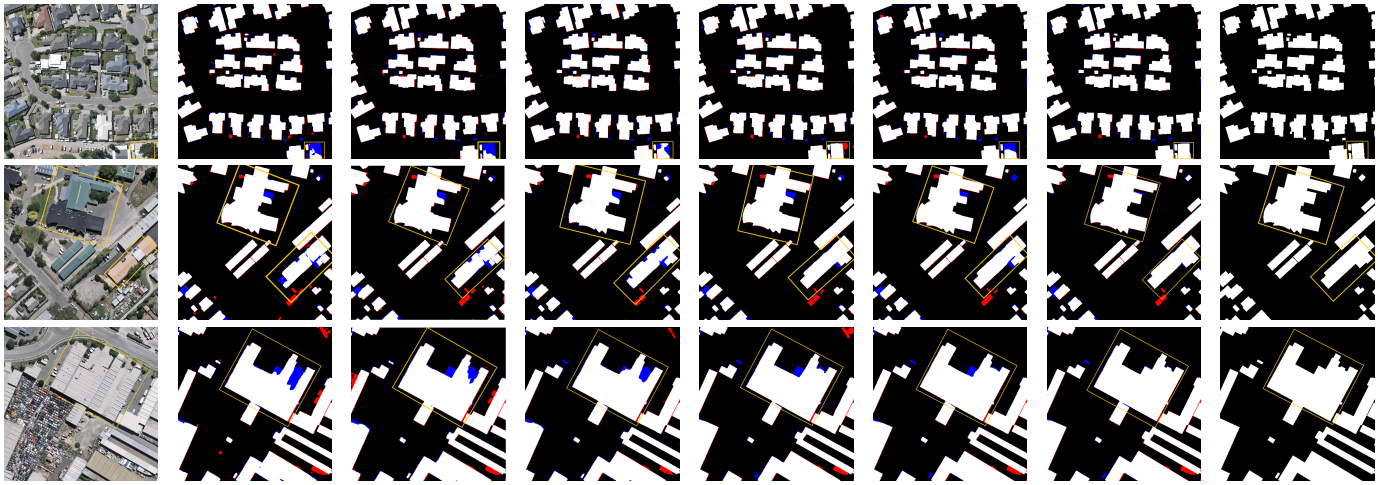
It is seen from Figs. 5-8 that the building boundaries extracted by our approach are much sharper compared to other models, due to the full integration and utilization of building contours and semantic information. Our approach not only improves the relatively effortless boundaries of large buildings but also obtains finer boundaries of smaller buildings. The corners of the building contours are of details, resulting in a more appealing visual appearance. In addition, the discriminative feature representation learned by the network helps to reduce false positives due to the full integration of multiple cues.

E. Ablation Study

To explore the effectiveness of different components of our proposed method, we randomly selected the WHU Buildings dataset and the Massachusetts Buildings dataset for ablation experiments. These components include dual stream decoders for exploring semantics and contours, DSF, and SCM. The quantitative results of the ablation experiments are shown in Table IV. The baseline is the U-net. ”C” means that the decoder of the network structure contains a semantic branch and a contour branch, but there is no information interaction between these two branches. As can be seen from the table, simply incorporating the extraction of contour information brings minimal improvement. To fully exploit both contour and semantic information, we propose to interactively transfer the learned features from both branches via an intermediate connection of the decoder to explore and utilize the correlation between the semantic and contour cues. As the second row of Table IV shows, the performance of the network improved by 2.04% and 1.94% of the IOU and 1.14% and 1.31% of the F1 score on the WHU dataset and the Massachusetts buildings dataset respectively, with ”IL” denoting the intermediate connection of the two branches. We propose the SCM to strengthen the connection between semantic and contour feature learning, hence the performance of the network is improved by 2.53% and 4.69% of the IoU and 1.41% and the 2.61% of the F1 score on the WHU dataset and the Massachusetts buildings dataset respectively compared to the baseline. To enhance the communication of semantic information between high and low layers in the network, we propose the MSCF and the performance of the network is improved by 3.15% and 5.6% of the IoU and 1.45% and 3.74% of the F1 score on the WHU dataset and the Massachusetts buildings dataset respectively compared to baseline. Finally, by applying all the components proposed, the performance of the network improved by 3.32% and 6.59 % of the IoU and 1.84% and 4.38% of the F1 score on the WHU dataset and the Massachusetts buildings dataset respectively, demonstrating the effectiveness of the network model in learning the semantic and contour information of the objects, as well as the full utilization of the correlation between them.

Fig. 11 shows the results of the ablation experiment, where blue indicates false negatives and red indicates false positives. The contour information learned by the network led to an initial improvement in the building boundaries. It helped the model obtain richer discrimination information and thus reduce the number of false negatives and false positives, which can be seen in the fourth column of Fig. 11. Nevertheless, such improvements are limited, therefore the SCM is introduced to facilitate the communication and integration of semantic and contour information to improve the segmentation of the network further. As can be seen in the fifth column of the Fig. 11, the boundaries of the building segmentation prediction map outlined by the yellow rectangle are neater than in the third and fourth columns, demonstrating that the SCM incorporates contour and semantic information more effectively. The sixth column of Fig. 11 shows the network predictions for the interactive dual-stream decoder and MSCF components. False negatives and positives are reduced compared to the baseline, but some predictions are still unsatisfactory. Finally, combining all the proposed components drives the model to seamlessly integrate semantic and contour information and exploit the correlation between them to obtain a more favorable performance, which can be seen in the penultimate column of Fig. 11. To show comparisons of performance under different configurations of the components of the model more clearly, Fig. 10 demonstrates the magnitude variation of the proposed components on the network performance. It is also noticeable from Fig. 10 that each of the components, whether precision or recall, has considerably improved performance of the baseline algorithm.

We compared the model computational complexity and parameter quantity of different methods, as shown in Fig. 12. On the WHU dataset, our approach exhibits only a minimal magnitude increase in model quantity compared to the baseline model, but the performance is improved by a large margin. When compared to state-of-the-art semantic segmentation algorithms, our method demonstrates significant improvements in both model parameter quantity and IoU. Although multi-task learning leads to an increase in the number of model parameters, we have mitigated this effect by reducing the channel dimension of each decoder stage to 256. Additionally, we introduced channel pooling layers to reduce the number of model parameters further. To provide a comprehensive evaluation, we also calculated the floating-point operations (FLOPs) for different methods. FLOPs can be used to assess the computational complexity of a model. From Fig .12, it can



(a) Original image (b) B (c) B+C (d) B+C+IL (e) B+C+IL+SCM (f) B+C+IL+MSCF (g) Ours (h) Ground Truth

Fig. 11. Exemplar results of the different components of the proposed method for the WHU building dataset, where B denotes "Baseline".

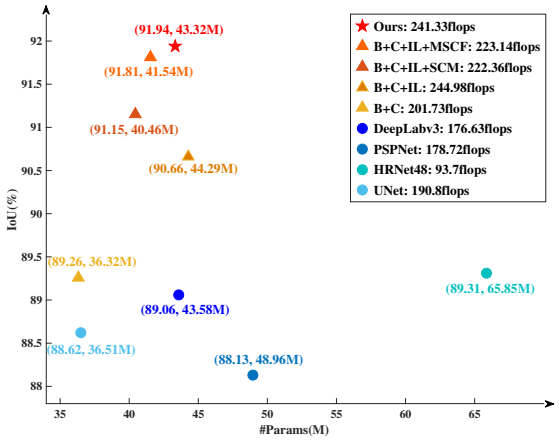


Fig. 12. Computational complexity and parameter quantity of different methods.

be observed that due to the involvement of multitasking, our method does not exhibit a significant advantage over other advanced semantic segmentation methods in this aspect. In future work, we will focus on reducing the model's computational complexity while preserving accuracy as much as possible.

V. Conclusion

In order to extract precise buildings, this paper has presented a simple and intuitive yet effective framework to explore semantic and contour cues and the correlations between them. We proposed an interactive dual-stream decoder consisting of a semantic stream and a contour stream, where the semantic stream learns advanced semantic information of objects, while the contour stream captures more detailed boundary information of the object, and intermediate connections between the two branches interactively learns correlations between semantics and contours. In order to strengthen the connection between the two branches, we propose the SCM exploit more thoroughly the correlation between the semantic and contour cues. To further improve the building segmentation performance, we propose the MSCF to fuse the contextual information of buildings at different scales. Through the above design, our model can efficiently learn and exploit semantic and contour cues to obtain a stronger discriminative feature representation and achieve accurate segmentation of buildings. Experiments on the WHU, INRIA, and Massachusetts building datasets have demonstrated that our approach outperformed other state-of-the-art building segmentation methods by a large margin. However, due to the involvement of multi-task computation in our method, it does not have an advantage in computational complexity. In future work, we will strive to address this issue.

References

[1] I. Khosravi, M. Momeni, and M. Rahnemoonfar, "Performance evaluation of object-based and pixel-based building detection algorithms from very high spatial resolution imagery," *Photogrammetric Engineering & Remote Sensing*, vol. 80, no. 6, pp. 519–528, 2014.

[2] E. Li, J. Femiani, S. Xu, X. Zhang, and P. Wonka, "Robust rooftop extraction from visible band images using higher order crf," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 8, pp. 4483–4495, 2015.

[3] S. Xu, X. Pan, E. Li, B. Wu, S. Bu, W. Dong, S. Xiang, and X. Zhang, "Automatic building rooftop extraction from aerial images via hierarchical rgb-d priors," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 12, pp. 7369–7387, 2018.

[4] M. Cote and P. Saeedi, "Automatic rooftop extraction in nadir aerial imagery of suburban regions using corners and variational level set evolution," *IEEE transactions on geoscience and remote sensing*, vol. 51, no. 1, pp. 313–328, 2012.

[5] J. Inglada, "Automatic recognition of man-made objects in high resolution optical remote sensing images by svm classification of geometric image features," *ISPRS journal of photogrammetry and remote sensing*, vol. 62, no. 3, pp. 236–248, 2007.

[6] J. Femiani, E. Li, A. Razdan, and P. Wonka, "Shadow-based rooftop segmentation in visible band images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 8, no. 5, pp. 2063–2077, 2014.

[7] S. Ji, S. Wei, and M. Lu, "A scale robust convolutional neural network for automatic building extraction from aerial and satellite imagery," *International journal of remote sensing*, vol. 40, no. 9, pp. 3308–3322, 2019.

[8] K. Bittner, F. Adam, S. Cui, M. Körner, and P. Reinartz, "Building footprint extraction from vhr remote sensing images combined with normalized dsms using fused fully convolutional networks," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 11, no. 8, pp. 2615–2629, 2018.

[9] J. Huang, X. Zhang, Q. Xin, Y. Sun, and P. Zhang, "Automatic building extraction from high-resolution aerial images and lidar data using gated residual refinement network," *ISPRS journal of photogrammetry and remote sensing*, vol. 151, pp. 91–105, 2019.

[10] T. Liu, L. Yao, J. Qin, N. Lu, H. Jiang, F. Zhang, and C. Zhou, "Multi-scale attention integrated hierarchical networks for high-resolution building footprint extraction," *International Journal of Applied Earth Observation and Geoinformation*, vol. 109, p. 102768, 2022.

[11] D. Feng, H. Chen, Y. Xie, Z. Liu, Z. Liao, J. Zhu, and H. Zhang, "Gccinet: Global feature capture and cross-layer information interaction network for building extraction from remote sensing imagery," *International Journal of Applied Earth Observation and Geoinformation*, vol. 114, p. 103046, 2022.

[12] Q. Li, L. Mou, Y. Hua, Y. Shi, and X. X. Zhu, "Crossgeonet: A framework for building footprint generation of label-scarce geographical regions," *International Journal of Applied Earth Observation and Geoinformation*, vol. 111, p. 102824, 2022.

[13] N. Girard, D. Smirnov, J. Solomon, and Y. Tarabalka, "Polygonal building segmentation by frame field learning," *arXiv preprint arXiv:2004.14875*, 2020.

[14] Y. Liu, D. Chen, A. Ma, Y. Zhong, F. Fang, and K. Xu, "Multiscale u-shaped cnn building instance extraction framework with edge constraint for high-spatial-resolution remote sensing imagery," *IEEE Transactions on Geoscience and Remote Sensing*, 2020.

[15] W. Jing, J. Lin, H. Lu, G. Chen, and H. Song, "Learning holistic and discriminative features via an efficient external memory module for building extraction in remote sensing images," *Building and Environment*, vol. 222, p. 109332, 2022.

[16] Y. Shi, Q. Li, and X. X. Zhu, "Building segmentation through a gated graph convolutional neural network with deep structured feature embedding," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 159, pp. 184–197, 2020.

[17] Y. Liu, D. Chen, A. Ma, Y. Zhong, F. Fang, and K. Xu, "Multiscale u-shaped cnn building instance extraction framework with edge constraint for high-spatial-resolution remote sensing imagery," *IEEE Transactions on Geoscience and Remote Sensing*, 2020.

[18] S. Wei, S. Ji, and M. Lu, "Toward automatic building footprint delineation from aerial images using cnn and regularization," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 3, pp. 2178–2189, 2019.

[19] A. Hatamizadeh, D. Sengupta, and D. Terzopoulos, "End-to-end trainable deep active contour models for automated image segmentation: Delineating buildings in aerial imagery," in *European Conference on Computer Vision*. Springer, 2020, pp. 730–746.

[20] F. Shi and T. Zhang, "A multi-task network with distance–mask–boundary consistency constraints for building extraction from aerial images," *Remote Sensing*, vol. 13, no. 14, p. 2656, 2021.

[21] S. He and W. Jiang, "Boundary-assisted learning for building extraction from optical remote sensing imagery," *Remote Sensing*, vol. 13, no. 4, p. 760, 2021.

[22] H. Guo, B. Du, L. Zhang, and X. Su, "A coarse-to-fine boundary refinement network for building footprint extraction from remote sensing imagery," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 183, pp. 240–252, 2022.

[23] T. Zuo, J. Feng, and X. Chen, "Hf-fcn: Hierarchically fused fully convolutional network for robust building extraction," in *Asian Conference on Computer Vision*. Springer, 2016, pp. 291–302.

[24] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

[25] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *arXiv preprint arXiv:1511.07122*, 2015.

[26] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[27] S. Ji, S. Wei, and M. Lu, "Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 1, pp. 574–586, 2018.

[28] A. Khalel and M. El-Saban, "Automatic pixelwise object labeling for aerial imagery using stacked u-nets," *arXiv preprint arXiv:1803.04953*, 2018.

[29] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active contour models," *International journal of computer vision*, vol. 1, no. 4, pp. 321–331, 1988.

[30] H. Ding, X. Jiang, A. Q. Liu, N. M. Thalmann, and G. Wang, "Boundary-aware feature propagation for scene segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 6819–6829.

[31] T. Takikawa, D. Acuna, V. Jampani, and S. Fidler, "Gated-scnn: Gated shape cnns for semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 5229–5238.

[32] X. Li, X. Li, L. Zhang, G. Cheng, J. Shi, Z. Lin, S. Tan, and Y. Tong, "Improving semantic segmentation via decoupled body and edge supervision," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16*. Springer, 2020, pp. 435–452.

[33] Y. Yuan, J. Xie, X. Chen, and J. Wang, "Segfix: Model-agnostic boundary refinement for segmentation," in *European Conference on Computer Vision*. Springer, 2020, pp. 489–506.

[34] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[35] Z. Hu, M. Zhen, X. Bai, H. Fu, and C.-l. Tai, "Jsenet: Joint semantic segmentation and edge detection network for 3d point clouds," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*. Springer, 2020, pp. 222–239.

[36] L. Wang, D. Li, Y. Zhu, L. Tian, and Y. Shan, "Dual super-resolution learning for semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3774–3783.

[37] B. Wu, S. Zhao, W. Chu, Z. Yang, and D. Cai, "Improving semantic segmentation via dilated affinity," *arXiv preprint arXiv:1907.07011*, 2019.

[38] M. Zhen, J. Wang, L. Zhou, S. Li, T. Shen, J. Shang, T. Fang, and L. Quan, "Joint semantic segmentation and boundary detection using iterative pyramid contexts," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 13 666–13 675.

[39] F. Heuer, S. Mantowsky, S. Bukhari, and G. Schneider, "Multitask-centernet (mcn): Efficient and diverse multitask learning using an anchor free approach," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 997–1005.

[40] V. R. Kumar, S. Yogamani, H. Rashed, G. Sitsu, C. Witt, I. Leang, S. Milz, and P. Mäder, "Omnidet: Surround view cameras based multi-task visual perception network for autonomous driving," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 2830–2837, 2021.

[41] D. Wu, M.-W. Liao, W.-T. Zhang, X.-G. Wang, X. Bai, W.-Q. Cheng, and W.-Y. Liu, "Yolop: You only look once for panoptic driving perception," *Machine Intelligence Research*, pp. 1–13, 2022.

[42] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[43] S. Anwar and N. Barnes, "Real image denoising with feature attention," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 3155–3164.

[44] H. Son and S. Lee, "Fast non-blind deconvolution via regularized residual networks with long/short skip-connections," in *2017 IEEE International Conference on Computational Photography (ICCP)*. IEEE, 2017, pp. 1–10.

[45] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, "Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark," in *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*. IEEE, 2017, pp. 3226–3229.

[46] V. Mnih, *Machine learning for aerial image labeling*. University of Toronto (Canada), 2013.

[47] G. Wu, X. Shao, Z. Guo, Q. Chen, W. Yuan, X. Shi, Y. Xu, and R. Shibasaki, "Automatic building segmentation of aerial imagery using multi-constraint fully convolutional networks," *Remote Sensing*, vol. 10, no. 3, p. 407, 2018.

[48] P. Liu, X. Liu, M. Liu, Q. Shi, J. Yang, X. Xu, and Y. Zhang, "Building footprint extraction from high-resolution images via spatial residual inception convolutional neural network," *Remote Sensing*, vol. 11, no. 7, p. 830, 2019.

[49] H. Liu, J. Luo, B. Huang, X. Hu, Y. Sun, Y. Yang, N. Xu, and N. Zhou, "De-net: Deep encoding network for building extraction from high-resolution remote sensing imagery," *Remote Sensing*, vol. 11, no. 20, p. 2380, 2019.

[50] W. Kang, Y. Xiang, F. Wang, and H. You, "Eu-net: An efficient fully convolutional network for building extraction from optical remote sensing images," *Remote Sensing*, vol. 11, no. 23, p. 2813, 2019.

[51] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[52] Y. Xie, J. Zhu, Y. Cao, D. Feng, M. Hu, W. Li, Y. Zhang, and L. Fu, "Refined extraction of building outlines from high-resolution remote sensing imagery based on a multifeature convolutional neural network and morphological filtering," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 1842–1855, 2020.

[53] H. Zhang, Y. Liao, H. Yang, G. Yang, and L. Zhang, "A local-global dual-stream network for building extraction from very-high-resolution remote sensing images," *IEEE Transactions on Neural Networks and Learning Systems*, 2020.

[54] X. Li, X. Yao, and Y. Fang, "Building-a-nets: Robust building extraction from high-resolution remote sensing images with adversarial networks," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 11, no. 10, pp. 3680–3687, 2018.

[55] L. Mou, Y. Hua, and X. X. Zhu, "A relation-augmented fully convolutional network for semantic segmentation in aerial scenes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12 416–12 425.

[56] A. Marcu, D. Costea, E. Slusanschi, and M. Leordeanu, "A multi-stage multi-task neural network for aerial scene interpretation and geolocalization," *arXiv preprint arXiv:1804.01322*, 2018.

[57] X. Pan, F. Yang, L. Gao, Z. Chen, B. Zhang, H. Fan, and J. Ren, "Building extraction from high-resolution aerial imagery using a generative adversarial network with spatial and channel attention mechanisms," *Remote Sensing*, vol. 11, no. 8, p. 917, 2019.

[58] S. Saito, T. Yamashita, and Y. Aoki, "Multiple object extraction from aerial imagery with convolutional neural networks," *Electronic Imaging*, vol. 2016, no. 10, pp. 1–9, 2016.