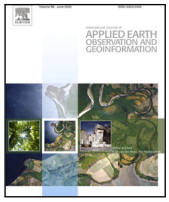
International Journal of Applied Earth Observation and Geoinformation 126 (2024) 103609

Contents lists available at ScienceDirect

International Journal of Applied Earth Observation and Geoinformation

journal homepage: www.elsevier.com/locate/jag



Exploring the user guidance for more accurate building segmentation from high-resolution remote sensing images

Dinghao Yang [b,1], Bin Wang [b,1], Weijia Li [a,*], Conghui He [b]

[a] School of Geospatial Engineering and Science, Sun Yat-Sen University, China
[b] Shanghai Artificial Intelligence Laboratory, China

ARTICLE INFO

Keywords:
User guidance
Building extraction
Semantic segmentation
Boundary correction

ABSTRACT

In recent years, the computer vision domain has witnessed a surge of interest in interactive object segmentation, an area of study that seeks to expedite the annotation process for pixel-wise segmentation tasks through user guidance. Despite this growing focus, existing methods mainly focus on a single type of pre-annotation and neglect the quality of boundary prediction, which significantly influences subsequent manual adjustments to segmentation boundaries. To address these limitations, we introduce a novel end-to-end network to facilitate more precise building segmentation using diverse types of user guidance. In our proposed method, a centroid map is generated to provide foreground prior information crucial to the subsequent segmentation procedure, and the boundary correction module automatically refines the segmentation mask from existing segmentation networks. Extensive experiments on two popular building extraction datasets demonstrate that our method outperforms all existing approaches given various user guidance (bounding boxes, inside-outside points, or extreme points), achieving the IoU scores of over 95% on SpaceNet-Vegas dataset and over 93% on Inria-building dataset. The remarkable performance of our method further demonstrates its immense potential to alleviate the labor-intensive annotation process associated with remote sensing datasets. The code of our proposed method is available at https://github.com/StephenDHYang/UGBS-pytorch.

1. Introduction

Recently, the rapid development of semantic segmentation in both computer vision (Long et al., 2015; Ronneberger et al., 2015) and remote sensing domains (Li et al., 2019) has been facilitated by the incorporation of precise pixel-wise annotation methods and the advent of Fully Convolutional Networks (FCNs). Nevertheless, generating accurate pixel-level segmentation labels remains a labor-intensive and time-consuming process, particularly for remote sensing image annotation, which necessitates the involvement of human experts possessing specialized background knowledge (Wu et al., 2023). To mitigate reliance on high-quality segmentation annotations, there is a burgeoning interest in developing interactive segmentation strategies that strive to produce more accurate segmentation results by providing informative priors (auxiliary pre-annotations) such as bounding boxes (Rother et al., 2004; Xu et al., 2017) and clicks (Liew et al., 2017; Maninis et al., 2018). By employing this approach, human annotators can perform subsequent modifications to the segmentation results to obtain the final pixel-wise annotation. This interactive process typically demands considerably less annotation time compared with traditional methods of conducting pixel-wise annotation on the original input images.

We first meticulously disentangle the entire process of interactive object segmentation methods. As depicted in the top of Fig. 1, current interactive object segmentation methods typically comprise three principal stages. Generally, the inference speed of the second stage is rapid, rendering the required time negligible. Consequently, it is crucial to minimize user interaction time during the first and third stages while simultaneously preserving the quality of the final prediction. However, existing methods rely on simplistic auxiliary pre-annotations to reduce the time needed to provide pre-annotation in the first stage, focusing predominantly on enhancing segmentation quality in the second stage (Papadopoulos et al., 2017; Maninis et al., 2018; Zhang et al., 2020). These approaches neglect to analyze the fundamental reasons behind the disparities in segmentation results derived from different types of pre-annotations. Simultaneously, user correction time in the third stage tends to be considerably longer than that of the first two stages due to the disregard for boundary segmentation accuracy in current methodologies.

In this study, we initially investigate the effect of various types of user guidance on segmentation performance, concluding that the

* Corresponding author.
E-mail address: liweij29@mail.sysu.edu.cn (W. Li).
[1] D. Yang and B. Wang contribute equally to this work.

https://doi.org/10.1016/j.jag.2023.103609
Received 24 August 2023; Received in revised form 7 November 2023; Accepted 6 December 2023
Available online 18 December 2023

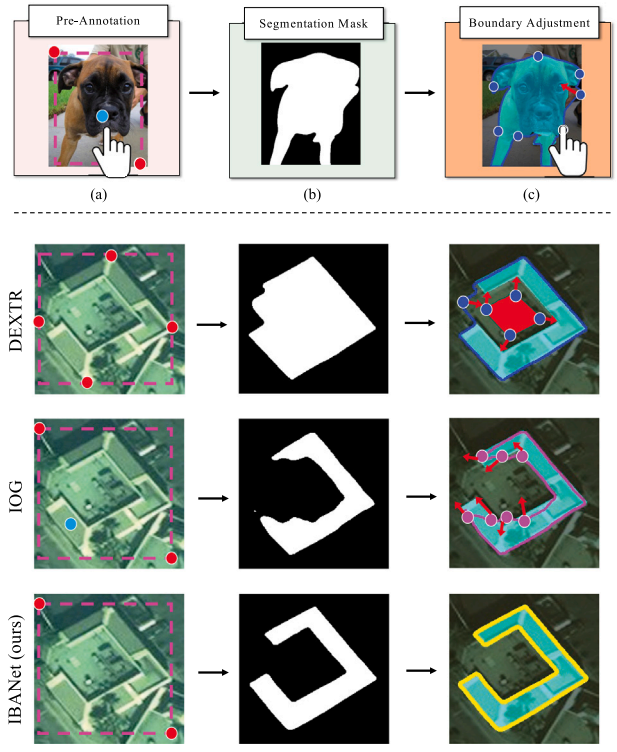


Fig. 1. The general process of the interactive object segmentation, including (a) providing a pre-annotation (with user guidance); (b) predicting the segmentation mask by neural network; (c) adjusting the predicted boundary polygon by user (moving the incorrect vertices, adding the missing regions, etc.) The goal of our method is to reduce user interactive labor.

additional prior information of pre-annotation in the first stage can substantially diminish both the prediction scope and prediction difficulties of the target. However, previous interactive segmentation methods (Papadopoulos et al., 2017; Maninis et al., 2018; Zhang et al., 2020) focus on a single type of pre-annotation, which limits their applications in various practical scenarios. Drawing inspiration from this observation, a centroid map is introduced to the segmentation network to further enhance the segmentation results. Additionally, we investigate the boundary correction capacity of the segmentation model to reduce the human effort required for subsequent boundary adjustments. As a result, we propose a novel segmentation network, that integrates centroid map prediction, segmentation mask prediction, and boundary correction within a unified model, and supports multiple types of user guidance. Experimental results illustrate that our method achieves the IoU score of 95.8% and 93.1% on the SpaceNet-Vegas and the Inria-building datasets with only three or four user clicks, and improves the Boundary F-score by 3% and Boundary IoU by 4% compared with the current state-of-the-art methods. The main contributions of this work can be summarized as follows:

- We propose an end-to-end building segmentation network that supports various types of pre-annotations by user guidance (*i.e.*, bounding boxes, inside-outside points, and extreme points).
- The proposed method introduces two novel modules, i.e., a centroid map prediction module that provides additional foreground prior information for the subsequent segmentation model, and a boundary correction module that further improves the segmentation boundary performance.
- The proposed method is evaluated on two popular building extraction benchmarks, *i.e.*, SpaceNet-Vegas (Van Etten et al., 2018) and Inria-building datasets (Maggiori et al., 2017), achieving much better results in both IoU and boundary segmentation performance (BF-score Perazzi et al., 2016 and BIoU Cheng et al., 2021) compared with all existing state-of-the-art methods.

2. Related work

2.1. Building extraction

Building footprint extraction from high-resolution aerial or satellite images has been extensively studied for decades (Ling et al., 2012; Yu et al., 2022; Li et al., 2022; Liu et al., 2023). With the rapid progress of deep learning methods for remote sensing image analysis, recent building extraction studies are mostly based on pixel-wise semantic and instance segmentation models (Li et al., 2019; Liu et al., 2022). To improve the building segmentation performance, various strategies are proposed and combined with the pixel-wise segmentation network, such as data fusion (Sun et al., 2018; Xie et al., 2023), multi-task learning (Turker and Koc-San, 2015; Li et al., 2021), boundary regularization (Wei et al., 2019), etc. Many studies propose polygonal building extraction methods to obtain the vector building footprints, which either propose post-processing methods to vectorize the pixel-wise segmentation results (Li et al., 2020), or design polygon-based models to directly predict the polygon vertices (Castrejon et al., 2017; Acuna et al., 2018; Ling et al., 2019). In addition, many recently proposed methods concentrate on extracting individual buildings from the cropped images which are sectioned off by bounding boxes of ground truth (GT bboxes) (Li et al., 2023) Among these studies, the active contour model (ACM) is a prevalent technique often utilized for the extraction of individual buildings., such as DSAC (Marcos et al., 2018), DarNet (Cheng et al., 2019), and CVNet (Xu et al., 2022).

However, the segmentation boundaries of existing building extraction studies are still far from the actual demand and require substantial efforts for further manual correction. In this study, we explore different types of pre-annotations for interactive segmentation and design new modules, which significantly improve the boundary prediction performance.

2.2. Interactive segmentation

Interactive segmentation aims at efficiently extracting the regions of interest from an image with the prior knowledge from user guidance (Ramadan et al., 2020; Yang et al., 2022), which facilitates the segmentation result improvements and reduces the resource consumption. Xu et al. adopt bounding box and foreground and background clicks (Xu et al., 2017), while Liew et al. (2017) improve their method by local feature extraction. For the polygon-based method, Castrejon et al. (2017), Acuna et al. (2018) and Ling et al. (2019) form the closed polygon annotation by multi-point labeling. For the mask-based method, Papadopoulos et al. (2017) propose an efficient way for object annotation with four extreme points, *i.e.*, the top, bottom, left-most and right-most points. Maninis et al. (2018) use extreme points in interactive segmentation and achieve a significant improvement. For the sake of concise interaction and providing rich priors, Zhang et al. (2020) propose the inside-outside guidance (IOG) with the bounding box and foreground click annotations, and adopt a coarse-to-fine pyramid neural network (Chen et al., 2018a). Several recent studies explore the interactive segmentation methods for remote sensing images of urban scenes (Lenczner et al., 2022; Yang et al., 2023). However, the studies on interactive segmentation for remote sensing images are still at a very early stage.

Moreover, there are few boundary-aware interactive segmentation methods, which is hindered by the utility of interactive information and the complexity of post-processing. First, interactive segmentation takes both image and interactive pre-annotation as input, which contains more foreground and background priors. The predicted boundary can become preciser via making full use of this prior information.

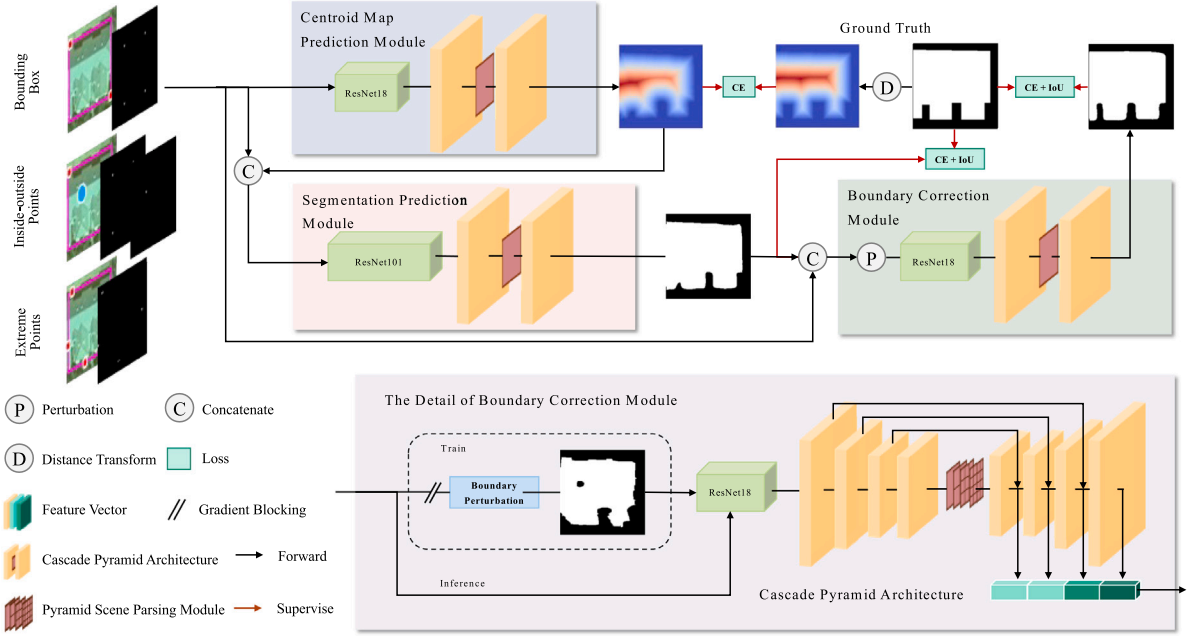


Fig. 2. Framework of our method, which is composed of three consecutive sub-modules: (1) the centroid map prediction module, which concatenates the image and user guidance as input and generates the centroid map prediction; (2) the segmentation prediction module, which concatenates the image, user guidance and centroid map prediction as input and infers the object mask; (3) the boundary correction module, which concatenates the image, user guidance and object mask as input and refines the object boundary. The user guidance is converted into a gray-scale map (1 channel for bounding box, 2 channels for inside-outside points, and 1 channel for extreme points).

Second, many boundary-related methods (Cheng et al., 2020; Guo et al., 2022) are independent post-processing networks, which achieve satisfying performance but are too complex and time-consuming for the interactive scenario.

In summary, the existing interactive segmentation methods focus on a single type of pre-annotation and do not pay much attention to boundary prediction, which are important aspects in practical annotation scenarios, and fixing the low-fidelity boundary will consume a huge amount of time. To address the above deficiencies, we propose an end-to-end interactive segmentation approach that supports various types of pre-annotations and involves novel boundary-aware modules.

3. Method

3.1. User guidance analysis.

Before introducing the three modules of the proposed method, we first analyze the characteristics of different pre-annotation types by user guidance. User guidance can provide the prior information of the foreground and background regions. It also helps to point out which object needs to be segmented in a multi-object image.

Bounding box. The bounding box, provided as the most simple and direct pre-annotation, is widely used in the interactive object segmentation process (Rother et al., 2004; Castrejon et al., 2017; Acuna et al., 2018; Ling et al., 2019). Except for the simplicity of the bounding box annotation, there are two more reasons why many studies use it as pre-annotation. To begin with, given a bounding box around the object, the outside regions can be directly ignored. Besides, the task is converted to foreground/background discrimination at a fixed scale, which reduces the complexity of model representation.

Extreme points. Considering that obtaining a tight and accurate bounding box is cognitively demanding, the pre-annotation of four extreme points (*i.e.*, top, bottom, left-most, right-most) is introduced (Maninis et al., 2018) to provide more object-related information with only a few extra labeling efforts. In fact, extreme points can be converted to bounding boxes directly and provide boundary information on four key positions of the object.

Inside-outside points. The pre-annotation of inside-outside points is proposed by Zhang et al. (2020), which uses a horizontal and a vertical guideline to speed up drawing a bounding box and take corresponding corner points as outside points. In addition, an inside point located around the object center needs to be provided. These two methods enable the improvement of segmentation performance with only a little extra time.

In our method, the choice of pre-annotation is flexible in terms of the practical requirements. The combination method of the pre-annotations and three modules will be introduced in Section 3.5.

3.2. Centroid map prediction module

Essentially, both extreme points and inside-outside points provide critical information, *i.e.*, boundary information and foreground-background correlation information, respectively. Extreme points provide boundary information to guide the network to get better boundary discrimination. Inside-outside points provide both foreground and background priors, which are more conducive to extract object masks. From this perspective, we introduce a centroid map to leverage additional foreground information for segmentation automatically. If we can acquire more prior information from the user guidance, the model can better solve the segmentation problem. Specifically, in the first stage, instead of predicting the segmentation from only an RGB image and the corresponding pre-annotation, we also estimate a coarse foreground probability map for the segmentation object, *i.e.*, centroid map. The ground truth of the centroid map in this paper denotes the distance between each pixel to the object boundary.[2] The farther the distance from the nearest boundary is, the higher the probability it is to be a foreground pixel. Let $\mathcal{M}_{gt}$ denotes the ground truth foreground mask, $\mathcal{L} = \{l_1, \ldots, l_n\}$ denotes the boundary of $\mathcal{M}_{gt}$ with n pixels, the centroid map C is defined as:

$$C_{ij} = \min_{i,j \in \mathcal{M}_{gt}, k \in \mathcal{L}} \|m_{ij} - l_k\|_2, \tag{1}$$

[2] The ground truth of the centroid map in our method is generated via scipy.ndimage.morphology.distance_transform_edt.

where m_{ij} denotes the pixel at position (i, j) of the foreground mask $\mathcal{M}_{gt}$. As shown in Fig. 2, the pixel furthest from the object boundaries has the largest probability values. Although the predicted centroid map is rough, it still provides meaningful information, *i.e.*, regions with very high/low probability values indicate the foreground/background elements, respectively. Furthermore, the object boundary is usually located in areas with ambiguous probability values. Benefiting from the predicted centroid map, the segmentation model achieves soft attention guidance for producing more accurate boundary predictions than using only pre-annotations, which indicates the centroid map is an enhanced prior for human annotation.

3.3. Segmentation prediction module

In the segmentation module of the proposed method, we utilize a segmentation network for object mask prediction. The network input consists of three parts, *i.e.*, the RGB image, the pre-annotations, and the object centroid map generated from the centroid map prediction module. The input is cropped by the pre-annotation, and several relaxed pixels are adopted for context introduction. Similar to the IOG method (Zhang et al., 2020), considering the usage of multi-level feature, we incorporate a coarse-to-fine CPN structure (Chen et al., 2018a) into ResNet-50/ResNet-101-based DeepLabV3+ (Chen et al., 2018b) architecture. The CPN structure fuses high segmentation information with low-level details, which can further enhance the robustness of features extracted from the popular DeepLabV3+ model. The segmentation module can be easily updated by other segmentation backbones, and we conduct a detailed comparison of different backbones in the ablation experiments.

Algorithm 1: Boundary Perturbation Operation.

Input: Prediction mask $\mathcal{M}_{seg}$.
Output: Perturbed mask $\mathcal{M}_{pert}$.

Get vertices set $V \leftarrow$ *findContours*($\mathcal{M}_{seg}$);
Get vertices subset by random drop $V' \leftarrow$ random.sample(V, sample_ratio=0.9);
Generate sampled mask $\mathcal{M}'_{seg} \leftarrow$ drawContours(V');
$\mathcal{M}_{pert} = \mathcal{M}'_{seg}$;
while $IoU(\mathcal{M}_{pert}, \mathcal{M}_{seg}) > IoU_{target}$ **do**
  $h, w \leftarrow$ the shape of $\mathcal{M}_{seg}$;
  $x \leftarrow$ random.randint(w);
  $y \leftarrow$ random.randint(h);
  $\Delta w \leftarrow$ random.randint($x+1, w+1$);
  $\Delta h \leftarrow$ random.randint($y+1, h+1$);
  if *random.rand() < 0.25* **then**
    Randomly modify the foreground region by replacing 0/1;
    $x' \leftarrow \lfloor (x+\Delta w)/2 \rfloor$;
    $y' \leftarrow \lfloor (y+\Delta h)/2 \rfloor$;
    $\mathcal{M}_{pert}(x', y') \leftarrow$ random.randint(2)$*$ 255;
  if *random.rand() < 0.5* **then**
    Conduct dilation operation with a random kernel size of (3, 10);
    $\mathcal{M}_{pert}(y:\Delta h, x:\Delta w) \leftarrow \mathcal{M}_{pert}(y:\Delta h, x:\Delta w) \oplus K_{dilation}(size = random(3,10))$;
  else
    Conduct erosion operation with a random kernel size of (3, 10);
    $\mathcal{M}_{pert}(y:\Delta h, x:\Delta w) \leftarrow \mathcal{M}_{pert}(y:\Delta h, x:\Delta w) \ominus K_{erosion}(size = random(3,10))$;
return $\mathcal{M}_{pert}$

3.4. Boundary correction module

Robust segmentation results can be obtained from the first two modules. Following the general interactive annotation process, the segmentation mask can be directly converted into an object polygon,[3] which is convenient to be edited and adjusted into a tight and accurate boundary polygon by the human annotators. Fig. 1 shows the details of the boundary adjustment process. Obviously, when the predicted boundary mask is coarse, the annotator needs to conduct a lot of adjustments to get a tight and accurate boundary polygon. From this perspective, a boundary correction module is introduced to alleviate the above issue in the third stage of the proposed method.

Based on the motivation to reduce human labor, we intuitively design a boundary correction module that is simple but effective for boundary refinement. Inaccurate boundary pixels are expected to be corrected via this module. Moreover, we find the buildings with hollows are error-prone, we hope the boundary correction module can also fix some wrong pixel predictions within the building, not only limited to narrow boundary regions. Thus we also set a random modified pixel operation for the perturbation. Consequently, we adopt a segmentation perturbation strategy on the input segmentation mask before feeding it into the boundary correction module. The detailed algorithm description of the boundary perturbation operations is shown in Algorithm 1.

The prediction mask itself contains high-level segmentation information. So we construct a lightweight network that pays more attention to extracting detailed features of the object boundary from the RGB input. The proposed boundary correction module is incorporated into an end-to-end segmentation architecture without using cascade structures, leading it to be more lightweight and efficient. The detail of the boundary correction module is illustrated at the right of Fig. 2.

3.5. Network training

Network architecture. As shown in Fig. 2, we choose ResNet-18 as the backbone of the centroid map prediction module and the boundary correction module. For the first stage, even a coarse centroid map can provide a certain number of accurate foreground and background pixels. Meanwhile, a high-level segmentation mask is incorporated into the input channel in the third stage. So a lightweight network is a reasonable choice taking both effectiveness and efficiency into consideration. While in the segmentation prediction module, we adopt ResNet-101 as the feature extraction backbone for more robust feature representation following IOG (Zhang et al., 2020) and DEXTR (Maninis et al., 2018).

For the boundary correction module, it should be noted that the gradient blocking strategy is applied to the segmentation input with perturbations. As shown in Fig. 2, this operation can prevent the gradient back-propagation from disturbing the previous segmentation network training, and enable the third stage to focus on rough boundary correction.

Cascaded Pyramid Network (CPN). We also employ a Cascaded Pyramid Network (CPN) (Chen et al., 2018a; Zhang et al., 2020) following each stage's corresponding backbone. The architecture of CPN is shown in the right-bottom of Fig. 2, and the details of the coarse-to-fine structure are introduced as follows. Four feature maps are generated from conv2 ~ conv5 block. The Pyramid Pooling Module (PPM) (Zhao et al., 2017) enriches the conv5 feature map by fusing features of four different pyramid scales. Then the first subnetwork (GlobalNet) encodes the segmentation feature with high-resolution details. Finally, the second subnetwork (RefineNet) integrates the abundant information on different scales by upsampling and concatenation.

Input. The input channels of the three modules are different. In the first stage, only RGB image (3 channels) and the gray-scale map of one specific user guidance (1 channel for bounding box, 2 channels for inside-outside points, or 1 channel for extreme points) are available, which are concatenated as the input of the centroid map prediction module. Then the centroid map (1 channel) generated from the first stage is concatenated with the input of the first stage, which can serve as an additional input for the second stage. Ultimately, the coarse segmentation result (1 channel) obtained from the second stage, the RGB image, and the user guidance are concatenated as the input of the final stage.

[3] This can be implemented conveniently via combining findContours and approxPolyDP functions in Opencv.

Loss. We first introduce two basic loss functions adopted in our training stage, *i.e.* cross entropy (CE) loss $\mathcal{L}_{CE}$ and intersection over union (IoU) loss $\mathcal{L}_{IoU}$. The definitions are illustrated in formula (2) and formula (3), respectively.

$$\mathcal{L}_{CE} = -(Glog(P) + (1-G)log(1-P)), \tag{2}$$

$$\mathcal{L}_{IoU} = 1 - \frac{P \cap G}{P \cup G}, \tag{3}$$

where G and P denote the ground truth and prediction, correspondingly.

Then we introduce the loss functions for each module. For the loss $\mathcal{L}_{cen}$ of the centroid map prediction module, we adopt CE loss to supervise the pixel-level regression:

$$\mathcal{L}_{cen} = \mathcal{L}_{CE}. \tag{4}$$

For the loss $\mathcal{L}_{seg}$ of the segmentation prediction module and $\mathcal{L}_{bou}$ of the the boundary correction module, to better constrain the boundary shape prediction, we further adopt a mixed loss:

$$\mathcal{L}_{seg} = \mathcal{L}_{bou} = \sum_{i=1}^{3}\sum_{j=1}^{5} \mathcal{L}_{CE}^{i,j} + \lambda \mathcal{L}_{IoU}^{i,j}, \tag{5}$$

where i and j refer to specific stage-index and layer of the GlobalNet and RefineNet (Zhang et al., 2020), respectively. The hyper-parameter λ balances the two losses.

Finally, we introduce the training strategy for the entire network. All three modules of the entire network are trained simultaneously, and the total loss $\mathcal{L}_{total}$ can be calculated as follows:

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{cen} + \lambda_2 \mathcal{L}_{seg}, \tag{6}$$

where λ_1 and λ_2 are the weights of the centroid map prediction module and the segmentation prediction module, respectively. The two modules are optimized by the total loss. Since the gradient of the boundary correction module is blocked during the training stage, it is optimized by its own mixed loss $\mathcal{L}_{bou}$, separately.

4. Experiments

4.1. Datasets and evaluation protocols

Datasets. Our method is evaluated by two public datasets: SpaceNet-Vegas (Van Etten et al., 2018) and Inria-building dataset (Maggiori et al., 2017). Specifically, the SpaceNet building dataset is processed into instance-level and we select the samples of Las Vegas following Li et al. (2023). The dataset comprises 3851 images and approximately 1,080,000 instances of buildings. These are randomly split into 3081 images for training, 385 for validation, and another 385 for testing purposes following Li et al. (2023). The Inria-building dataset is cropped based on single-building object in the center of the image, and sampled from Inria Aerial Image Labeling Dataset (Maggiori et al., 2017) following Xu et al. (2022). The dataset contains 18,952 building images, which are split into train/val/test sets with a ratio of 60%/20%/20% following Xu et al. (2022).

Evaluation protocols. We report our performance with three main metrics: IoU, Boundary F-score (Perazzi et al., 2016) and Boundary IoU (Cheng et al., 2021). Existing interactive methods mainly focus on IoU with less consideration of boundary pixels. In contrast, we introduce two additional metrics to evaluate the boundary segmentation quality of our method. In addition, to conduct a fair comparison with other methods on the Inria-building dataset, we also introduce the Weighted Coverage (WCov) and Dice metrics following Xu et al. (2022).

4.2. Implementation details

Pre-annotation simulation. In the training and testing stage of the model, we generate the user guidance from the ground truth mask by user simulation following the same way as previous interactive segmentation studies (Xu et al., 2017; Papadopoulos et al., 2017; Maninis et al., 2018; Zhang et al., 2020). The three types of user guidance are generated as follows:

- For the bounding box, we first generate a precise bounding box from the ground truth mask by acquiring the top, bottom, left, and right coordinates. Then we relax the box with 10 pixels to simulate human behavior, so that the user do not have to conduct very precise interaction to get accurate prediction.
- For the inside-outside points, we generate the outside points from the top-left and bottom-right points of the simulated bounding box obtained above, and we sample the inside points from the center of the ground truth mask. To simulate human behavior, the inside points are randomly perturbed with [0, 5] pixels within the mask.
- For the extreme points, we first get the precise top-most, bottom-most, left-most, and right-most points of the ground truth mask, and then we perturb the four points with a random pixel of [0, 5] to simulate human behavior.

Training and Testing details. We use the ImageNet (Russakovsky et al., 2015) pre-trained ResNet (He et al., 2016) as the backbone. For the sake of better refinement capability, we perturb the output of the segmentation prediction module by modifying its boundary pixels randomly (Cheng et al., 2020). Our model is trained for 100 epochs on SpaceNet-Vegas and the Inria-building datasets, with SGD as the optimizer, a learning rate of $2.5 \times e^{-3}$, a momentum of 0.9, a weight decay of $5 \times e^{-4}$, and the *poly* policy adopted. We set the hyper-parameter λ for balancing the two losses as 0.4, and 1 for both λ_1 and λ_2. The model is trained on 2 NVIDIA Tesla V100, with a batch size of 24 for SpaceNet-Vegas and 64 for the Inria-building datasets. The IoU target is set as 0.9 for boundary perturbation in the training stage, while it is omitted in the testing stage.

4.3. Comparison with state-of-the-art methods

First, to evaluate the performance of our method, we compare it with the state-of-the-art interactive segmentation methods on the SpaceNet-Vegas building dataset. We evaluate the performance of these methods using their released code. For Polygon-RNN++ (Acuna et al., 2018), previous studies (Li et al., 2023) proved that the evaluator adopting beam search strategy can improve the prediction results, and the strategies of reinforcement learning and upscaling with a GGNN can deteriorate the building segmentation performance. Thus we apply the evaluator module and remove the latter two strategies for the Polygon-RNN++ method. For Curve-GCN (Ling et al., 2019), since the buildings are polygon-type objects, we select the Polygon-GCN model rather than Spline-GCN model. Taking the real distribution of vertex numbers into consideration, its total number is set as 20. In addition, we employ the point matching loss (with K set to 1280 following (Ling et al., 2019)) since it achieves superior performance when compared to the differentiable accuracy loss in our specific task.

For the interactive methods with other types of pre-annotations, we use the default setting of DEXTR (Maninis et al., 2018) and IOG (Zhang et al., 2020) (*i.e.*, extreme points for DEXTR, inside-outside points for IOG). Considering the fairness of the comparison, we set the number of clicks as 5 and use the backbone of HRNet-32 for FocalClick (Chen et al., 2022).

Furthermore, to fairly evaluate the gain of our method that comes from not only pre-annotations but also architecture designs, we compare the proposed method with the state-of-the-art ACM-based models

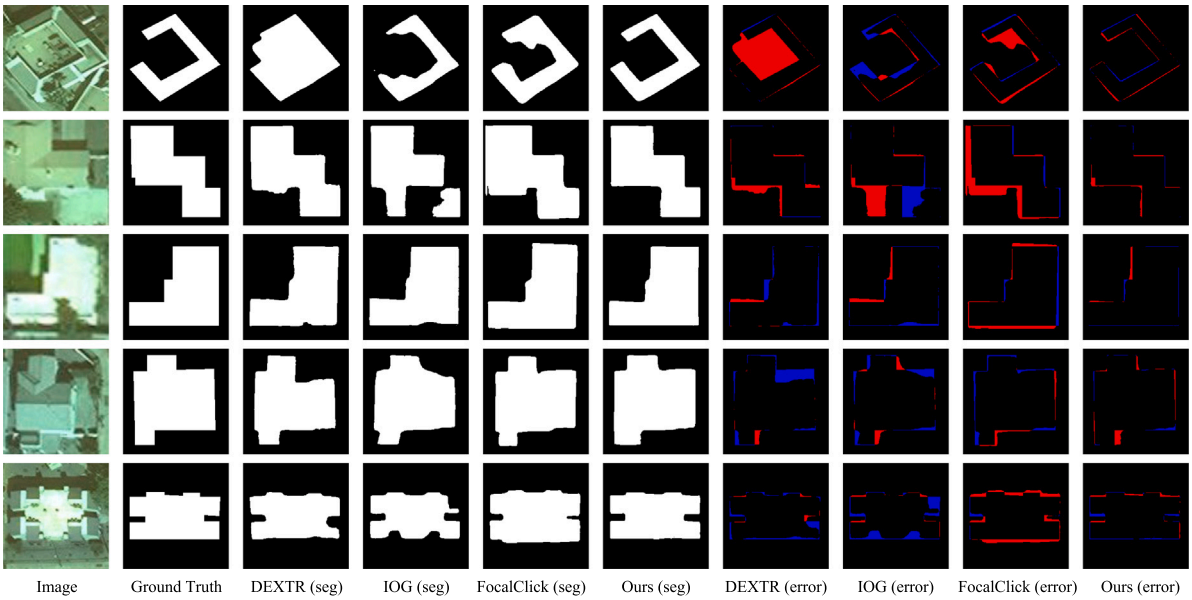


Fig. 3. Visual comparison on the SpaceNet-Vegas dataset. The *seg* and *error* are the abbreviations of the segmentation result and error map, respectively. For the last three columns, the red region denotes False Positive (FP), while the blue region denotes False Negative (FN). Results show that our method produces more accurate results with more precise boundaries compared with other methods.

Table 1
Comparison with the state-of-the-art methods on SpaceNet-Vegas dataset. All the metrics are scaled by 10^2.

Methods	Pre-annotation	IoU	BF-score	BIoU
Polygon-RNN++ (Acuna et al., 2018)	Bounding Box	86.1	67.2	26.6
Polygon-GCN (Ling et al., 2019)	Bounding Box	88.7	73.1	30.1
DEXTR (Maninis et al., 2018)	Extreme Points	94.9	85.6	45.6
IOG (Zhang et al., 2020)	Inside-outside Points	95.0	85.4	45.3
FocalClick (Chen et al., 2022)	Iterative Clicks	95.1	84.7	43.2
Ours	Bounding Box	94.0	81.6	42.1
	Extreme Points	95.8	87.6	**53.4**
	Inside-outside Points	**95.8**	**88.4**	51.0

Table 2
Comparison with the state-of-the-art methods on the Inria-building dataset. All the metrics are scaled by 10^2.

Methods	Pre-annotation	IoU	WCov	BF-score	Dice
DSAC (Marcos et al., 2018)	Bounding Box	35.1	37.8	5.9	51.2
DARNet (Cheng et al., 2019)	Bounding Box	65.8	60.5	33.0	77.2
CVNet (Xu et al., 2022)	Bounding Box	77.6	75.6	42.2	86.7
DEXTR (Maninis et al., 2018)	Extreme Points	92.4	92.4	84.3	96.0
IOG (Zhang et al., 2020)	Inside-outside Points	92.2	92.3	84.0	95.9
FocalClick (Chen et al., 2022)	Iterative Clicks	88.0	88.4	64.5	93.6
Ours	Bounding Box	92.1	92.2	84.3	95.9
	Extreme Points	**93.1**	**93.1**	**86.9**	**96.4**
	Inside-outside Points	92.8	92.9	86.0	96.3

designed for single building extraction (*i.e.*, DSAC (Marcos et al., 2018), DARNet (Cheng et al., 2019), and CVNet (Xu et al., 2022)), using the default experimental settings of these methods on the Inria-building dataset.

Table 1 provides the comparison of different methods on SpaceNet-Vegas dataset. Our method achieves the best performance on all metrics, demonstrating a significant improvement on boundary metrics. Since the IoU of the SpaceNet-Vegas building dataset is already higher than 95%, the BF-score and BIoU metrics (related to boundary region) are more sensitive to reflect the segmentation performance. Moreover, compared with recently proposed interactive segmentation methods (*i.e.*, DEXTR (Maninis et al., 2018), IOG (Zhang et al., 2020) and FocalClick (Chen et al., 2022)), our method gains the state-of-the-art results using extreme points and inside-outside points. Even with only bounding box pre-annotation, our method demonstrates competitive performance and significantly improves the performance compared with Polygon-RNN++ (Acuna et al., 2018) and Polygon-GCN (Ling et al., 2019).

Table 2 shows the comparison of different methods on Inria-building dataset. Our method achieves significant improvement compared with previous methods on all metrics. For the methods with only bounding box provided, our method demonstrates a huge performance improvement (14.5% for IoU, 16.6% for WCov, 42.1% for BF-score and 9.2% for Dice) compared with previous state-of-the-art (*i.e.*, CVNet (Xu et al., 2022)). For the methods with other pre-annotation types, our method makes obvious improvement with both extreme points and inside-outside points compared with DEXTR and IOG, especially for boundary metrics (2.6% improvement for DEXTR and 2.0% improvement for IOG). Figs. 3 and 4 illustrate some qualitative results of SpaceNet-Vegas and Inria datasets. Results demonstrate that our method significantly improves the building extraction results compared with the previous state-of-the-art methods, achieving more accurate building boundaries with much fewer error regions.

4.4. Ablation study

In our ablation study, we first perform an extensive experiment to explore how the pre-annotation type can affect the prediction performance. Then we evaluate the effectiveness of each module of the proposed method.

Backbone ablation. To prove the effectiveness of the DeepLabV3-based CPN backbone, we make a comparison with other common backbones for segmentation, *i.e.* U-Net (Ronneberger et al., 2015), HRNet (Wang et al., 2020) and SegFormer (Xie et al., 2021). The additional three backbones are substituted in our framework with a suitable size combination. The U-Net backbone adopts the normal size of its paper (Ronneberger et al., 2015) for the segmentation prediction module, and $\frac{1}{2}$ size for the centroid map prediction and boundary correction module. The HRNet backbone adopts HRNet-w32 for the segmentation prediction module, and HRNet-w18 for the centroid map prediction and boundary correction module. The SegFormer backbone adopts SegFormer-B3 for the segmentation prediction module, and

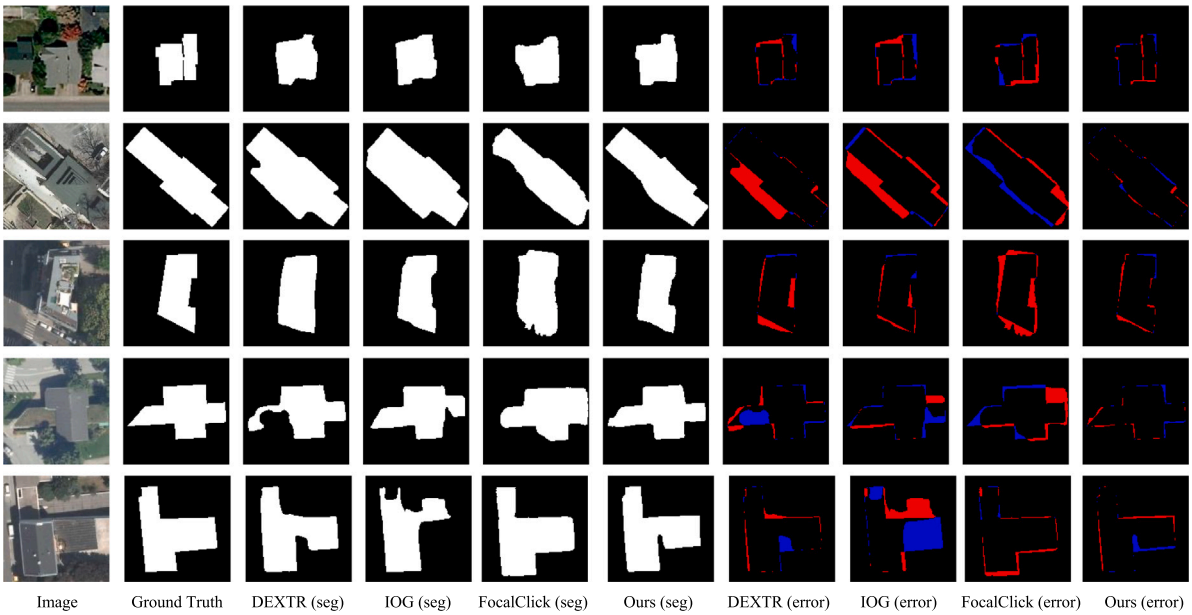


Fig. 4. Visual comparison on the Inria-building dataset. The *seg* and *error* are the abbreviations of the segmentation result and error map, respectively. For the last three columns, and the red region denotes False Positive (FP), while the blue region denotes False Negative (FN).

Table 3
Backbone ablation on the SpaceNet-Vegas dataset. All the methods utilize inside-outside points as user guidance, and set the same batch size on the same device. The size of the model and GPU memory usage is measured in megabytes (MB).

Backbones	Params	Memory usage	IoU	BF-score	BIoU
U-Net (Ronneberger et al., 2015)	32.0	2021.5	91.2	81.3	47.3
HRNet (Wang et al., 2020)	48.8	2564.2	94.9	84.7	50.1
SegFormer (Xie et al., 2021)	54.7	2431.4	95.2	87.4	53.1
CPN (Zhang et al., 2020)	79.4	2315.2	**95.8**	**87.6**	**53.4**

Table 4
Distance metric ablation of the centroid prediction module on the SpaceNet-Vegas dataset. All the methods utilize inside-outside points as user guidance.

Distance functions	IoU	BF-score	BIoU
Taxicab	95.6	87.2	52.9
Chessboard	95.5	87.0	53.1
Euclidean	**95.8**	**87.6**	**53.4**

SegFormer-B0 for the centroid map prediction and boundary correction module.

Meanwhile, the number of parameters cannot accurately demonstrate the overhead of training and inference of the models since it does not consider the H and W dimension of activation maps, and the FLOPs calculated by current tools are not accurate enough for ViT architecture. Thus we set the same batch size 24 for all four backbones on the same devices (2 NVIDIA Tesla V100, 32 GB) and calculate the GPU memory usage per device for a fair comparison. As shown in Table 3, the DeepLabV3-based CPN achieves the best performance under all the segmentation metrics and handles an efficient GPU overhead. Although our backbone has the largest number of parameters, it is still a computation-efficient architecture. As mentioned before, CPN adopts an encoder–decoder architecture with multi-level feature fusion and skip connection, which is not a direct encoder architecture of DeepLabV3. Most activation maps are processed under lower resolution, leading to better prediction for low-level edge information and high-level body information with less computation cost.

Distance function ablation. The centroid map is dedicated to reflecting the probability of the foreground region under 2D Euclidean space. We conduct the distance function ablation experiments to compare the three common measurements (*i.e.* taxicab distance, chessboard distance, and Euclidean distance). Suppose there are two points $A(x_1, y_1)$, $B(x_2, y_2)$, the three distance function can be formulated as follows:

$$D_{taxicab} = |x_2 - x_1| + |y_2 - y_1|, \tag{7}$$

$$D_{chessboard} = \max(|x_2 - x_1|, |y_2 - y_1|), \tag{8}$$

$$D_{Euclidean} = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}. \tag{9}$$

As shown in Table 4, the Euclidean distance is more suitable for the measure of point-to-edge distance directly, which outperforms the other two measurements in terms of all evaluation metrics.

Pre-annotation ablation. As illustrated in Table 5, we perform the pre-annotation ablation experiments on SpaceNet-Vegas dataset, and the qualitative comparison results are illustrated in Fig. 5. Among the three types of pre-annotations, the bounding box provides the least foreground/background priors since the samples are instance-level and this type only provides clicks on background pixels, which leads to the worst performance compared with the other two types. It is suitable for rectangular buildings without hollows, as bounding box take the least human labor with 2 clicks and can acquire a precise result. The 2nd row of Fig. 5 is a similar example. The extreme points type achieves the best quantitative performance on average, which includes most priors since each click provides both foreground and background information. As illustrated in the 1st/2nd/4th rows of Fig. 5, the features of extreme points make it suitable for most regular polygon buildings with simple and prominent shapes, even overlapped by trees or shadows. However, as shown in the last two rows of Fig. 5, the extreme points type fails for some building samples with concave or hollow shapes. On the contrary, as shown in the 3rd/5th/6th rows, the inside-outside points alleviates the above weakness via providing an inside click. In conclusion, the choice of pre-annotations depends on different actual cases, and our proposed method is flexible for user guidance, making it more convenient in practical data annotation compared with other methods. For most cases without hollows, the extreme points type is better in terms of the qualitative results. The choice of bounding box and extreme points should consider a trade-off between efficiency and

Table 5
Ablation study on SpaceNet-Vegas dataset. Module 1/2/3 represents centroid map prediction, segmentation prediction, and boundary correction, respectively. Our method can handle all three types of pre-annotation to achieve the best result.

Module			Bounding box			Inside-outside points			Extreme points		
1	2	3	IoU(%)	BF-score(%)	BIoU(%)	IoU(%)	BF-score(%)	BIoU(%)	IoU(%)	BF-score(%)	BIoU(%)
	✓		93.7	80.3	41.3	95.2	85.4	47.4	95.1	86.3	47.5
✓	✓		93.8	80.9	41.5	95.4	86.7	51.8	95.3	87.8	49.6
	✓	✓	93.8	81.2	41.9	95.5	86.7	49.6	95.5	88.0	50.1
✓	✓	✓	**94.0**	**81.6**	**42.1**	**95.8**	**87.6**	**53.4**	**95.8**	**88.4**	**51.0**

Table 6
Comparison with IOG on different pre-annotations, BF is the abbreviation of BF-score, and all the metrics are scaled by 10^2.

Interactive	Methods	Vegas			Inria		
		IoU	BF	BIoU	IoU	BF	BIoU
Bounding box	IOG (Zhang et al., 2020)	93.7	80.3	41.3	91.4	82.1	41.9
	ours	**94.0**	**81.6**	**42.1**	**92.1**	**84.3**	**44.5**
Inside-outside points	IOG	95.2	85.4	47.4	92.3	84.0	43.7
	ours	**95.8**	**87.6**	**53.4**	**92.8**	**86.0**	**45.9**
Extreme points	IOG	95.1	86.3	47.5	92.7	85.3	44.7
	ours	**95.6**	**88.4**	**51.0**	**93.1**	**87.0**	**47.0**

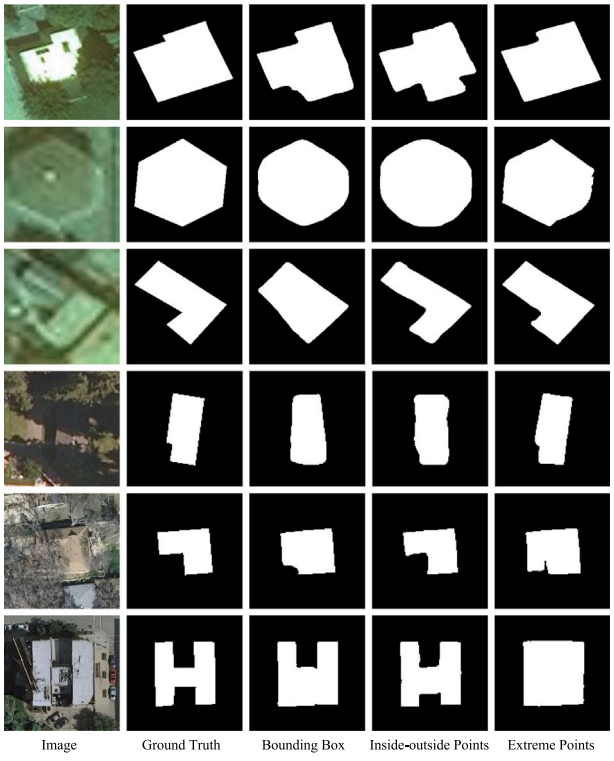


Fig. 5. Comparison of the building extraction results obtained from different pre-annotation types on the SpaceNet-Vegas and the Inria-building dataset.

precision sample-by-sample. For the situations that need more inside prior, the inside-outside points type is a more suitable choice.

We further make a comprehensive comparison between our method and IOG using three types of pre-annotations, as illustrated in Table 6. We can find that our method outperforms IOG under every type of pre-annotations, with over 2.0% boundary metrics improvements. Moreover, on the Inria-building dataset, the boundary F-score and boundary IoU of our method with bounding box are 84.3% and 44.5%, respectively, which are even higher than IOG with both inside and outside points input (84.0% and 43.7%). In other words, our method achieves **better boundary prediction** with **simpler pre-annotation**. As shown in Fig. 4, compared to IOG, our method achieves significant performance gains on the segmentation boundary.

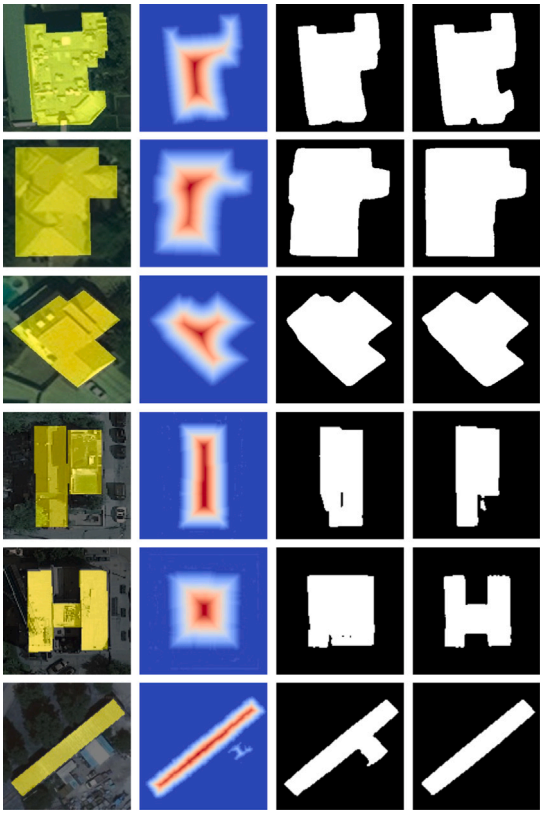


Fig. 6. Visualization of each sub-module of our method. The $1st$/ column shows the GT. The 2nd/3rd/4th column shows the result from the centroid map prediction module, the segmentation prediction module, and the boundary correction module, respectively.

Table 7
Boundary perturbation ablation on the SpaceNet-Vegas dataset. All the methods utilize inside-outside points as user guidance.

Mask perturbation	IoU(%)	BF-score(%)	BIoU(%)
–	95.5	86.9	51.2
✓	**95.8**	**87.6**	**53.4**

Module ablation. To understand how each stage of our method facilitates the segmentation performance, we evaluate each module independently. As illustrated in Table 5, the results indicate that both

Table 8
Module effectiveness validation on the SpaceNet-Vegas dataset. All the methods utilize inside-outside points interactive input. R is the abbreviation of ResNet, and the size of the model is measured in megabytes (MB).

Module			Backbone	Params	IoU(%)	BF-score(%)	BIoU(%)
1	2	3					
	✓		R-101	52.1	95.2	85.4	47.4
✓	✓		R-18+50	**46.8**	**95.6**	**86.3**	**51.2**
	✓	✓	R-50+18	**46.8**	95.6	85.9	49.9

Table 9
Optimization strategy ablation on the SpaceNet-Vegas dataset. All the methods utilize inside-outside points as interactive input. Module 1/2/3 represents centroid map prediction, segmentation prediction, and boundary correction, respectively.

Simultaneously optimized module			Gradient blocking	Total epochs	IoU(%)	BF-score(%)	BIoU(%)
Module1	Module2	Module3					
				220	95.4	87.1	52.8
✓	✓			160	**95.9**	**87.6**	53.2
	✓	✓		160	94.8	86.5	51.9
✓	✓	✓		100	95.2	86.9	52.4
✓	✓		✓	100	95.8	**87.6**	**53.4**

centroid map prediction module and boundary correction module can benefit segmentation individually, and the combination of all modules can achieve the best performance, which indicates the three modules mutually benefit the segmentation performance. Furthermore, we conduct additional ablation studies to verify the effectiveness of the segmentation perturbation strategy, and the result is shown in Table 7. We also validate the performance gain is due to the module design instead of the deepening of the network architecture. As shown in Table 8, the backbone of the centroid map prediction module and boundary correction module is ResNet-18, while we use ResNet-50 for module 2. We also show the results of module 2-only with ResNet-101 backbone for reference. Results show that both the two combinations (*i.e.*, module 1 & 2, module 2 & 3) outperform module 2-only with ResNet-101, which demonstrates that the proposed method can **achieve better segmentation performance** with **simpler architecture**. Fig. 6 visualizes some samples to intuitively show the procedures of our method. With the enhancement of each module, the segmentation predictions are gradually improved.

Training strategy ablation. To validate our training strategy for the three modules-based network, we do an optimization ablation study. In experiments, we train the centroid map prediction module and the boundary correction module for 60 epochs, and the segmentation prediction for 100 epochs to get convergence results. The results are shown in Table 9, from which we can conclude that:

- The centroid map prediction module should be optimized with the segmentation prediction module to get the best performance since the foreground prior is directly used for foreground mask prediction.
- The function of the boundary correction module is independent of the main segmentation task, it is dedicated to refining any coarse boundary.

Considering the trade-off between segmentation performance and training cost, we finally select the gradient-blocking strategy for the boundary correction module, so that all modules can be trained simultaneously within 100 epochs and get the approximate best results.

4.5. Cross-domain evaluation

To further demonstrate the advantages of our method in practical application scenarios, we compare it with DEXTR and IOG for cross-domain experiments.

Generalization on distribution-similar scenarios. For the generalization evaluation, we chose the SpaceNet-Vegas and the Inria-building dataset. As illustrated in the 2nd and 3rd rows of Table 10, our method outperforms DEXTR and IOG with more than 3% improvement on boundary metrics, which demonstrates that our method has better class-agnostic and generalization ability compared with DEXTR and IOG.

Fig. 7. Qualitative results of the cross-domain evaluation. Our model is trained on ADE20K training set, and tested on the test set of the SpaceNet-Vegas and the Inria-building datasets.

Cross-domain on different scenarios. To explore the capability of each method for generalization from general vision to remote sensing scenarios, we choose the SpaceNet-Vegas, the Inria-building dataset, PASCAL (Everingham et al., 2010) and ADE20K (Zhou et al., 2017) for the cross-domain experiments. The 5th–8th rows of Table 10 and Fig. 7 show the quantitative and qualitative results of the cross-domain experiments on different scenarios. We can find that our method outperforms DEXTR and IOG on natural to remote sensing scenario on all metrics. The competitive performance of our method (with an IoU of over 80%) also vindicates that training on natural images is capable enough for the building extraction tasks on remote sensing images, which prospects a new application of leveraging the large-scale well-annotated general vision datasets to facilitate the annotation process of remote sensing images.

The cross-domain performance in distribution-similar and different scenarios verifies the potential of our method to facilitate efficient annotation in real-world scenarios.

4.6. Limitation analysis

To comprehensively demonstrate the performance of our method, we analyze some representative failure cases on the SpaceNet-Vegas and the Inria-building dataset.

Dataset mismatch annotation. As shown in Fig. 5, there are indeed some mismatch annotations in the SpaceNet-Vegas and the

Table 10
Comparison with DEXTR and IOG in terms of the cross-domain ability. All the metrics (IoU, BF and BIoU) are scaled by 10^2.

Train	Test	DEXTR			IOG			Ours		
		IoU	BF-score	BIoU	IoU	BF-score	BIoU	IoU	BF-score	BIoU
Vegas	Vegas	94.9	85.6	45.6	95.0	85.4	45.3	**95.4**	**86.3**	**49.4**
	Inria	88.4	76.3	37.1	87.9	75.5	35.9	**89.8**	**78.5**	**38.7**
Inria	Vegas	73.9	26.8	9.7	73.7	25.7	8.4	**74.5**	**28.7**	**11.2**
	Inria	92.4	84.3	43.7	92.2	84.0	42.7	**93.1**	**86.9**	**47.0**
PASCAL	Vegas	77.5	45.2	15.7	79.6	47.6	16.9	**82.4**	**49.6**	**18.7**
	Inria	77.6	63.4	25.5	80.3	65.3	27.8	**83.8**	**67.2**	**30.6**
ADE20K	Vegas	78.8	49.9	15.2	82.1	51.0	16.2	**86.6**	**53.0**	**18.0**
	Inria	79.5	68.4	29.8	81.3	69.6	30.1	**85.4**	**71.2**	**32.1**

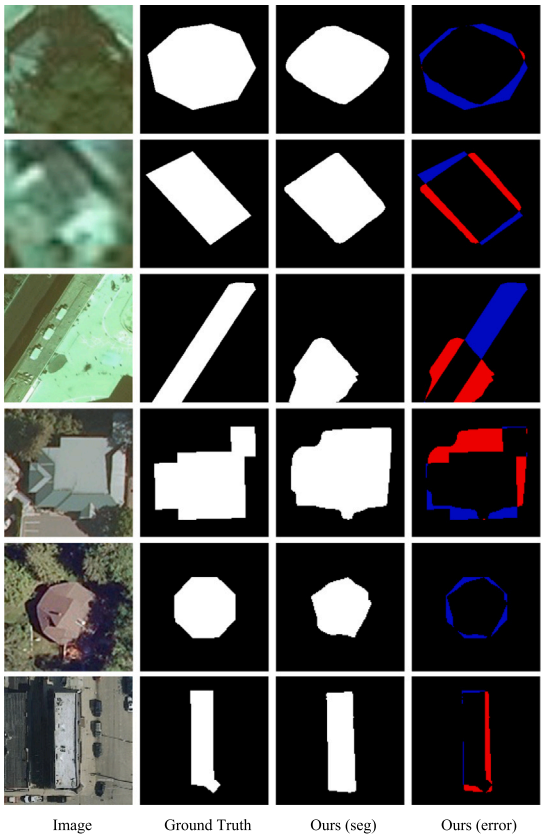


Fig. 8. Failure cases of our method on the SpaceNet-Vegas and the Inria-building dataset. The red region denotes False Positive (FP), while the blue region denotes False Negative (FN).

Inria-building datasets, which is hard to avoid in existing public building segmentation datasets. Nevertheless, our interactive segmentation-based method can somewhat solve this complex case. On the one hand, as shown in Fig. 1, all three pre-annotation types (*i.e.* bounding box, inside-outside points, and extreme points) provide background prior by clicks, from which we generate a soft bounding box with a 20-pixel relax to crop the image before feeding to the network, and this relax improves the robustness of the network for slight mismatch annotations. On the other hand, the foreground and background user guidance is converted into a gray-scale map and concatenated with the cropped image, and our prior-based training enables the network to distinguish the target building more precisely.

Error-prone scenarios. In Fig. 8, we show some error-prone failure cases. There are three main aspects for the poor prediction results of these images. As shown in the first two rows of Fig. 8, the first aspect is the low resolution due to the resizing of each instance, resulting in difficulties for the model to identify semantic information. The second aspect is the confusion between foreground and background. As shown in the fourth and fifth rows, the foreground (building) overlaps with the background (trees and shadows) or has a similar color and texture to the background, resulting in poor segmentation performance. The last aspect is the limitation of click pre-annotation type. As shown in the third and last rows, if the buildings are long and narrow, the boundary prior could not be inferred by several clicks. In such cases, the scribble-like pre-annotation types can better handle this situation. In general, the interactive segmentation is a pixel-level binary classification problem of foreground and background, and the poor result is caused by the unclear boundary due to the image semantic factors and unsuitable pre-annotation priors.

5. Conclusion

In this work, we proposed an end-to-end interactive network to improve the accuracy and quality of building segmentation from remote sensing images, which further facilitates the efficient pixel-wise annotation of building extraction datasets. The innovation of the method lies in providing important prior information with the centroid map, as well as combining the disturbance labeling and segmentation prediction to correct the boundary. Moreover, our method supports multiple types of pre-annotations by user guidance (bounding boxes, inside-outside points, and extreme points), and it can be easily updated if new segmentation structures are used. Finally, quantitative and qualitative experiments verify that our method can achieve start-of-the-art performance on all evaluation metrics for the building extraction task. The extensive ablation study also validates the effectiveness of each pre-annotation type and network module in our proposed method. We believe that our method has significant potential and application values for improving the time-consuming annotation process of remote sensing datasets. The generalization of image domain and user guidance are two remaining challenges for the interactive segmentation-based building extraction task. In our future work, we will improve our method in more complex scenarios with more types of user guidance. We also plan to extend our method to other applications in remote sensing domain.

CRediT authorship contribution statement

Dinghao Yang: Conceptualization, Methodology, Software, Formal analysis, Validation, Data curation, Writing – original draft, Visualization. **Bin Wang:** Conceptualization, Methodology, Software, Formal analysis, Investigation, Writing – original draft. **Weijia Li:** Conceptualization, Methodology, Formal analysis, Investigation, Writing – original draft, Writing – reviewing & editing, Supervision, Project administration, Funding acquisition. **Conghui He:** Investigation, Resources, Writing – reviewing & editing, Supervision, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

We have shared the link to our data and code in the manuscript.

Acknowledgments

This research was supported in part by the National Natural Science Foundation of China (Grant No. 42201358).

References

Acuna, D., Ling, H., Kar, A., Fidler, S., 2018. Efficient interactive annotation of segmentation datasets with polygon-rnn++. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. CVPR, pp. 859–868.

Castrejon, L., Kundu, K., Urtasun, R., Fidler, S., 2017. Annotating object instances with a polygon-rnn. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. CVPR, pp. 5230–5238.

Chen, Y., Wang, Z., Peng, Y., Zhang, Z., Yu, G., Sun, J., 2018a. Cascaded pyramid network for multi-person pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. CVPR, pp. 7103–7112.

Chen, X., Zhao, Z., Zhang, Y., Duan, M., Qi, D., Zhao, H., 2022. FocalClick: Towards practical interactive image segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. CVPR, pp. 1300–1309.

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., 2018b. Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European Conference on Computer Vision. ECCV, pp. 801–818.

Cheng, H.K., Chung, J., Tai, Y.-W., Tang, C.-K., 2020. CascadePSP: Toward class-agnostic and very high-resolution segmentation via global and local refinement. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. CVPR, pp. 8890–8899.

Cheng, B., Girshick, R., Dollár, P., Berg, A.C., Kirillov, A., 2021. Boundary iou: Improving object-centric image segmentation evaluation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. CVPR, pp. 15334–15342.

Cheng, D., Liao, R., Fidler, S., Urtasun, R., 2019. Darnet: Deep active ray network for building segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. CVPR, pp. 7431–7439.

Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A., 2010. The pascal visual object classes (voc) challenge. Int. J. Comput. Vis. (IJCV) 88 (2), 303–338.

Guo, H., Du, B., Zhang, L., Su, X., 2022. A coarse-to-fine boundary refinement network for building footprint extraction from remote sensing imagery. ISPRS J. Photogramm. Remote Sens. 183, 240–252.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. CVPR, pp. 770–778.

Lenczner, G., Chan-Hon-Tong, A., Le Saux, B., Luminari, N., Le Besnerais, G., 2022. Dial: Deep interactive and active learning for semantic segmentation in remote sensing. IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens. 15, 3376–3389.

Li, W., He, C., Fang, J., Zheng, J., Fu, H., Yu, L., 2019. Semantic segmentation-based building footprint extraction using very high-resolution satellite images and multi-source GIS data. Remote Sens. 11 (4), 403.

Li, M., Lafarge, F., Marlet, R., 2020. Approximating shapes in images with low-complexity polygons. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. CVPR, pp. 8633–8641.

Li, Q., Mou, L., Hua, Y., Shi, Y., Zhu, X.X., 2022. Crossgeonet: A framework for building footprint generation of label-scarce geographical regions. Int. J. Appl. Earth Obs. Geoinf. 111, 102824.

Li, W., Zhao, W., Yu, J., Zheng, J., He, C., Fu, H., Lin, D., 2023. Joint semantic–geometric learning for polygonal building segmentation from high-resolution remote sensing images. ISPRS J. Photogramm. Remote Sens. 201, 26–37.

Li, W., Zhao, W., Zhong, H., He, C., Lin, D., 2021. Joint semantic–geometric learning for polygonal building segmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). vol. 35, (no. 3), pp. 1958–1965.

Liew, J., Wei, Y., Xiong, W., Ong, S.-H., Feng, J., 2017. Regional interactive image segmentation networks. In: 2017 IEEE International Conference on Computer Vision. ICCV, IEEE Computer Society, pp. 2746–2754.

Ling, H., Gao, J., Kar, A., Chen, W., Fidler, S., 2019. Fast interactive object annotation with curve-gcn. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. CVPR, pp. 5257–5266.

Ling, F., Li, X., Xiao, F., Fang, S., Du, Y., 2012. Object-based sub-pixel mapping of buildings incorporating the prior shape information from remotely sensed imagery. Int. J. Appl. Earth Obs. Geoinf. 18, 283–292.

Liu, X., Chen, Y., Wang, C., Tan, K., Li, J., 2023. A lightweight building instance extraction method based on adaptive optimization of mask contour. Int. J. Appl. Earth Obs. Geoinf. 122, 103420.

Liu, T., Yao, L., Qin, J., Lu, N., Jiang, H., Zhang, F., Zhou, C., 2022. Multi-scale attention integrated hierarchical networks for high-resolution building footprint extraction. Int. J. Appl. Earth Obs. Geoinf. 109, 102768.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. CVPR, pp. 3431–3440.

Maggiori, E., Tarabalka, Y., Charpiat, G., Alliez, P., 2017. 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, pp. 3226–3229.

Maninis, K.-K., Caelles, S., Pont-Tuset, J., Van Gool, L., 2018. Deep extreme cut: From extreme points to object segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. CVPR, pp. 616–625.

Marcos, D., Tuia, D., Kellenberger, B., Zhang, L., Bai, M., Liao, R., Urtasun, R., 2018. Learning deep structured active contours end-to-end. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. CVPR, pp. 8877–8885.

Papadopoulos, D.P., Uijlings, J.R., Keller, F., Ferrari, V., 2017. Extreme clicking for efficient object annotation. In: Proceedings of the IEEE International Conference on Computer Vision. ICCV, pp. 4930–4939.

Perazzi, F., Pont-Tuset, J., McWilliams, B., Van Gool, L., Gross, M., Sorkine-Hornung, A., 2016. A benchmark dataset and evaluation methodology for video object segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. CVPR, pp. 724–732.

Ramadan, H., Lachqar, C., Tairi, H., 2020. A survey of recent interactive image segmentation methods. Comput. Vis. Media 1–30.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. MICCAI, Springer, pp. 234–241.

Rother, C., Kolmogorov, V., Blake, A., 2004. "GrabCut" interactive foreground extraction using iterated graph cuts. ACM Trans. Graph. 23 (3), 309–314.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al., 2015. Imagenet large scale visual recognition challenge. Int. J. Comput. Vis. (IJCV) 115 (3), 211–252.

Sun, Y., Zhang, X., Zhao, X., Xin, Q., 2018. Extracting building boundaries from high resolution optical images and LiDAR data by integrating the convolutional neural network and the active contour model. Remote Sens. 10 (9), 1459.

Turker, M., Koc-San, D., 2015. Building extraction from high-resolution optical space-borne images using the integration of support vector machine (SVM) classification, hough transformation and perceptual grouping. Int. J. Appl. Earth Obs. Geoinf. 34, 58–69.

Van Etten, A., Lindenbaum, D., Bacastow, T.M., 2018. Spacenet: A remote sensing dataset and challenge series. arXiv preprint arXiv:1807.01232.

Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., et al., 2020. Deep high-resolution representation learning for visual recognition. IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI) 43 (10), 3349–3364.

Wei, S., Ji, S., Lu, M., 2019. Toward automatic building footprint delineation from aerial images using CNN and regularization. IEEE Trans. Geosci. Remote Sens..

Wu, Z., Hou, B., Guo, X., Ren, B., Li, Z., Wang, S., Jiao, L., 2023. CCNR: Cross-regional context and noise regularization for SAR image segmentation. Int. J. Appl. Earth Obs. Geoinf. 121, 103363.

Xie, Y., Tian, J., Zhu, X.X., 2023. A co-learning method to utilize optical images and photogrammetric point clouds for building extraction. Int. J. Appl. Earth Obs. Geoinf. 116, 103165.

Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P., 2021. SegFormer: Simple and efficient design for semantic segmentation with transformers. Adv. Neural Inf. Process. Syst. (NeurIPS) 34, 12077–12090.

Xu, N., Price, B., Cohen, S., Yang, J., Huang, T., 2017. Deep grabcut for object selection. In: 28th British Machine Vision Conference. BMVC, BMVA Press.

Xu, Z., Xu, C., Cui, Z., Zheng, X., Yang, J., 2022. CVNet: Contour vibration network for building extraction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. CVPR, pp. 1383–1391.

Yang, S.D., Wang, B., Li, W., Lin, Y., He, C., 2022. Unified interactive image matting. arXiv preprint arXiv:2205.08324.

Yang, L., Zi, W., Chen, H., Peng, S., 2023. DRE-Net: A dynamic radius-encoding neural network with an incremental training strategy for interactive segmentation of remote sensing images. Remote Sens. 15 (3), 801.

Yu, B., Yang, A., Chen, F., Wang, N., Wang, L., 2022. SNNFD, spiking neural segmentation network in frequency domain using high spatial resolution images for building extraction. Int. J. Appl. Earth Obs. Geoinf. 112, 102930.

Zhang, S., Liew, J.H., Wei, Y., Wei, S., Zhao, Y., 2020. Interactive object segmentation with inside-outside guidance. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. CVPR, pp. 12234–12244.

Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J., 2017. Pyramid scene parsing network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. CVPR, pp. 2881–2890.

Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A., 2017. Scene parsing through ade20k dataset. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. CVPR, pp. 633–641.