remote sensing

MDPI

Article

ARE-Net: An Improved Interactive Model for Accurate Building Extraction in High-Resolution Remote Sensing Imagery

Qian Weng [1,2], Qin Wang [1,2], Yifeng Lin [1,2] and Jiawen Lin [1,2,*]

[1] College of Computer and Data Science, Fuzhou University, Fuzhou 350000, China; fzuwq@fzu.edu.cn (Q.W.); 211027008@fzu.edu.cn (Q.W.); 221027076@fzu.edu.cn (Y.L.)
[2] Fujian Key Laboratory of Network Computing and Intelligent Information Processing, Fuzhou University, Fuzhou 350000, China
* Correspondence: ljw@fzu.edu.cn

Abstract: Accurate building extraction for high-resolution remote sensing images is critical for topographic mapping, urban planning, and many other applications. Its main task is to label each pixel point as a building or non-building. Although deep-learning-based algorithms have significantly enhanced the accuracy of building extraction, fully automated methods for building extraction are limited by the requirement for a large number of annotated samples, resulting in a limited generalization ability, easy misclassification in complex remote sensing images, and higher costs due to the need for a large number of annotated samples. To address these challenges, this paper proposes an improved interactive building extraction model, ARE-Net, which adopts a deep interactive segmentation approach. In this paper, we present several key contributions. Firstly, an adaptive-radius encoding (ARE) module was designed to optimize the interaction features of clicks based on the varying shapes and distributions of buildings to provide maximum a priori information for building extraction. Secondly, a two-stage training strategy was proposed to enhance the convergence speed and efficiency of the segmentation process. Finally, some comprehensive experiments using two models of different sizes (HRNet18s+OCR and HRNet32+OCR) were conducted on the Inria and WHU building datasets. The results showed significant improvements over the current state-of-the-art method in terms of NoC_{90}. The proposed method achieved performance enhancements of 7.98% and 13.03% with HRNet18s+OCR and 7.34% and 15.49% with HRNet32+OCR on the WHU and Inria datasets, respectively. Furthermore, the experiments demonstrated that the proposed ARE-Net method significantly reduced the annotation costs while improving the convergence speed and generalization performance.

Keywords: interactive building extraction; adaptive-radius encoding; two-stage training; remote sensing

check for updates

Citation: Weng, Q.; Wang, Q.; Lin, Y.; Lin, J. ARE-Net: An Improved Interactive Model for Accurate Building Extraction in High-Resolution Remote Sensing Imagery. *Remote Sens.* **2023**, *15*, 4457. https://doi.org/10.3390/rs15184457

Academic Editor: Hossein M. Rizeei

Received: 12 August 2023
Revised: 3 September 2023
Accepted: 4 September 2023
Published: 10 September 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

In recent years, deep learning algorithms have significantly advanced the accuracy of building extraction for high-resolution remote sensing images [1–5]. These techniques have found extensive applications in diverse areas such as smart city development and planning [6–8], economic activity distribution [9], urban disaster prevention and mitigation [10–13], and unauthorized construction detection [14]. The major objective of building extraction for high-resolution remote sensing imagery is to ascertain the presence of buildings in an image and assign each pixel as either belonging to a building or not [15]. While supervised fully automated methods have achieved impressive results, they rely heavily on a large number of labeled samples. Inadequate data can result in model overfitting, which subsequently diminishes the model's generalizability.

For building extraction from high-resolution remote sensing images, the issue of limited samples is prevalent. Existing high-quality samples are often restricted to a few specific datasets, hampering the application of building extraction methods to remote

sensing images without samples. On the other hand, traditional interactive building extraction methods for sample-free remote sensing images only rely on manual fine tracing or clicking to complete fine building extraction, which has greater adaptability compared to fully automated methods. However, traditional interactive segmentation methods are based on hand-crafted features, meaning that the pixel-level labeling of natural images takes an average of 10.1 min per image [16–22], while remote sensing image annotation is even more time-consuming and challenging due to various factors, such as water vapor and lighting conditions [23,24]. Hence, it is vital to utilize deep learning techniques to understand objects and semantics and achieve the high-precision extraction of structures at a lower cost.

The primary objective of interactive segmentation in the framework of deep learning is to refine the annotation of image pixels made by deep learning networks through human-provided interaction information, such as mouse clicks. These deep-learning-based interactive applications are superior to traditional interactive applications, significantly improving the accuracy and efficiency of object extraction, and even providing a promising solution for the challenge of extracting buildings from high-resolution remote sensing images without cross-domain samples. Commonly used interactive segmentation methods can be broadly classified into the following categories: (1) Bounding-box-based methods [25–27]. These methods do not effectively provide cues for foreground and background, and in the context of high-resolution remote sensing images, accurately handling overlapping bounding boxes in large images is a challenging problem. (2) Polar-coordinate-based methods [28,29]. These approaches have been proven to be effective but require a click at the corner location of the building, which increases the burden on the user. (3) Doodle-based approaches [30–32]. Although these approaches can provide a substantial amount of interaction information, it is computationally expensive to simulate real graffiti to train the convolutional neural network (CNN). (4) Click-based approaches [33–37]. These methods are simple and effective, allowing the extraction of the target by clicking at any position with low cost. Therefore, this paper focuses on click-based interactive building extraction from high-resolution remote sensing images, and we redesigned an interactive encoding method to improve the rate of convergence (RoC), reduce the number of clicks required to achieve the desired accuracy, and simultaneously ensure the network inference time to continually reduce user interaction latency and time cost in interactive applications. To our knowledge, this is the first work to improve interactive encoding in the interactive building extraction of RS images.

Traditional interactive segmentation methods can be categorized into boundary-based and region-based methods according to the interaction approach [38]. Boundary-based methods include techniques such as Intelligent Scissors [39], Level Set [40], Active Contours [41], and Snakes [42]. The Intelligent Scissors method utilizes dynamic programming to find the lowest-cost path on the user-defined target boundary, which is considered as the segmentation boundary. Level Set, Active Contours, and Snakes process the initial contour provided by the user by minimizing an energy function evolution curve to obtain the final boundary contour. Boundary-based methods rely heavily on high-quality user-supplied clicks or initial contours and are sensitive to image grayscale inhomogeneity and noise, which can affect the segmentation stability.

The first region-based interactive segmentation approaches were based on graph cuts [43], where users label foreground and background pixel points by drawing lines. Similar to graph cuts, Random Walk [44] constructs segmentation models based on graph theory, representing an image as an undirected graph with each pixel as a node and image edges representing the relationship between pixel points. Interaction based on region growing algorithms [45] starts from a labeled pixel point and expands in eight directions, grouping pixel points based on the range of the absolute difference in pixel values. The interaction is gradually iterated to achieve segmentation. Region-based interactive segmentation does not require explicit boundaries but relies on the labeling of foreground and background points. However, these algorithms predict the foreground/background

distribution based on hand-crafted image features, which may not meet the accuracy requirements of segmentation in complex high-resolution remote sensing images.

Interactive segmentation has gained significant attention in recent years, allowing users to accurately select objects of interest by providing inputs such as scribble and bounding boxes. The graph cut method has played a pivotal role in advancing interactive segmentation, leading to the development of numerous algorithms to address this problem. Taking advantage of the remarkable advances in deep neural networks, particularly convolutional neural networks (CNNs), in image classification and detection, researchers have applied them to semantic segmentation tasks. Ning Xu et al. [33] were the first to incorporate deep learning into interactive segmentation. Their approach utilized user inputs of foreground and background clicks to calculate the Euclidean distance as a distance map. This distance map, together with the input image, was fed into an FCN network for interactive semantic segmentation, enhancing both accuracy and efficiency. Based on the click-based input method, the image features extracted by the VGG16 model [46] were incorporated into a segmentation network [34]. Subsequently, it was found that the first click plays an important role in determining the location and subject information of the target object [19]. Edge information from the image along with user clicks as inputs to fine-tune the network was proposed in [37]. A different click encoding method using disk encoding instead of the Euclidean distance map was introduced in [47]. In each iteration epoch, this method incorporated the previous prediction results as prior information along with the encoding. On this basis, an interactive segmentation scheme that progressed from coarse to fine was designed [48]. It performed region segmentation in the click area and refined the segmented area, significantly reducing the computation time and number of model parameters. To further enhance the performance of the segmentation model, dynamic encoding and phased incremental learning strategies were proposed [49]. Interactive building extraction techniques can make full use of user knowledge to guide the extraction process and refine the extraction accuracy. Compared to fully automated and manual extraction methods, deep-learning-based interactive building extraction strikes a balance between accuracy and efficiency.

In our research, we investigated several prominent interactive segmentation methods [20,35,47,48]. However, these methods all rely on a fixed-radius click encoding scheme to assist in the segmentation task. Compared to scribble annotation, the fixed-radius approach provides limited a priori information. Our summary of existing research found that in high-resolution remote sensing images, buildings can have various geometric shapes, such as rectangles; characters (H, L, T, U, Z); circles; and combinations thereof. Simply enlarging the radius of the clicks can result in the wrong areas being covered. In addition, we analyzed the training process of deep-learning-based interactive methods and discovered a difference compared to traditional deep learning network training. In the early stage of deep-learning-based interactive methods, the main purpose of network training is to learn interactive features and maximize the influence of each click to quickly establish building masks. In the later stage, the focus of network training should be shifted to fine-tuning the mask output by the network according to the correction information provided by each interaction, which can also weaken the impact of clicks to prevent convergence (CG) deterioration caused by misclassification or blurring and improve the convergence speed. Some methods, such as DRE-Net [49], employ an incremental learning training strategy to enhance the low rate of convergence (RoC), but this increases the training cost. Furthermore, the interaction features obtained from fixed-radius clicks overlook the fine-tuning effect of correction information and lead to confusion in the building extraction task for high-resolution remote sensing images. To fully exploit the potential of the user's prior knowledge and to adapt to different training stages with different learning goals, we designed a strategy involving adaptive-size disk coding and a two-stage training approach.

In this paper, in order to maximize the extracted semantic information of the buildings in the interactive segmentation features, we designed the adaptive-radius encoding (ARE) algorithm. It can fully exploit the potential of manual clicking based on prior knowledge

and improve the accuracy of building extraction. Furthermore, we proposed a training strategy that divides the network training into two distinct stages, each with different goals. By providing specific supervision at different stages of the training process, this strategy can continuously guide deep networks to learn the correction information of manual clicks, thereby improving the RoC performance in building extraction and refining its segmentation results. The contributions of our work can be summarized as follows:

- We proposed a novel interactive segmentation model, ARE-Net, for building extraction from high-resolution remote sensing images. Compared to state-of-the-art interactive information encoding modules, the proposed ARE module can learn more priori information to support the segmentation task in buildings of various shapes.
- We designed a two-stage training strategy that guides the network to treat clicks at different stages differently in order to more efficiently refine the accuracy of building semantic segmentation.
- We conducted an evaluation of the method on the Wuhan University aerial building dataset (WHU [50]) and the Inria aerial dataset (Inira [51]). The experimental results demonstrated that the proposed method could achieve better performance compared to existing methods while significantly reducing the number of annotations.

2. Materials and Methods

2.1. An Overview of ARE-Net

In our research, interactive building extraction was regarded as a binary segmentation task. The goal is to generate accurate building masks through an iterative process that fully exploits the guidance information provided by user clicks on annotations on high-resolution remote sensing imagery. Each click serves as an iterative step towards achieving a satisfactory result for the user. We focused on optimizing two main aspects of the interactive building extraction process. The first goal was to reduce the number of user clicks required while achieving higher accuracy. By improving the rate of convergence of the model, the number of interactions required to obtain satisfactory segmentation results could be minimized. This optimization would enhance the efficiency of the interactive process and reduce the burden on the user. Secondly, we aimed to reduce the time cost associated with user annotation. The manual annotation of remote sensing images can be time-consuming and labor-intensive. Therefore, we aimed to develop a model that can generalize well across different images and datasets, reducing the need for extensive manual labeling. By ensuring the generalizability of the model, we could minimize the time and effort required for user labeling while still achieving accurate and reliable building extraction results. By addressing these optimization goals, our research aimed to improve the overall efficiency and usability of interactive building extraction, making a more practical and effective approach for high-resolution remote sensing imagery.

In order to fully demonstrate the effectiveness of our approach, we built ARE-Net following the framework of the current state-of-the-art method RITM [47] by inserting the ARE module. ARE-Net can use any semantic segmentation network as a segmentation module. As the segmentation network, we used HRNet18s+OCR and HRNet32+OCR, which are widely used in interactive segmentation. The specific pipeline is shown in Figure 1, which illustrates the training process of ARE-Net. The segmentation model (prediction) is the inference stage, where the original images and interactive features are inferred in the network to obtain a prediction mask. During this process, no adjustments are made to any parameters of the segmentation model, but only the positive/negative points list is updated using the prediction output mask. The segmentation model (training) is the model training stage, where the network parameters of the segmentation model are trained using input raw images and interactive features. During this process, the model parameters are adjusted. First, users place positive/negative clicks randomly in the input image. Clicking on a building is a positive click, and clicking elsewhere is a negative click. These clicks are coded as interaction features by ARE. A zero matrix of the same size as the input image is initialized as the feature map of the third channel. Clicks are randomly

selected on the image, and these clicks are generated as interaction features by ARE. Then, these interaction features are initialized into a zero-initialized matrix and stitched into a feature map with 3 channels. Next, according to the different judging conditions in the two different training stages, these feature maps are fed into the network and trained in two different stages: the first stage and the second stage. The stages initialize the number of iterations and an iteration threshold. The number of iterations is initialized to 0, and the iteration threshold, which is initialized to a larger value N_{iter}, is used to limit the number of iterations. In the first stage, the training strategy of RITM is applied, and the number of iterations and the iteration threshold are initialized. Then, the model fuses the input feature map and the original map for prediction, updates the feature map based on the prediction map until the number of iterations reaches N_{iter}, and inputs it to the network training. In the second stage, one more IoU threshold is initialized. If the number of iterations does not reach the threshold, the intersection over union (IoU) value is calculated for each prediction result. If the IoU value is higher than the initialization threshold, that feature map and the original map are directly utilized for training. Otherwise, the sequence of interacting clicks, the prediction, and the loss function (iteration loss) continue to be computed until the IoU value exceeds the threshold or the number of iterations reaches N_{iter}.

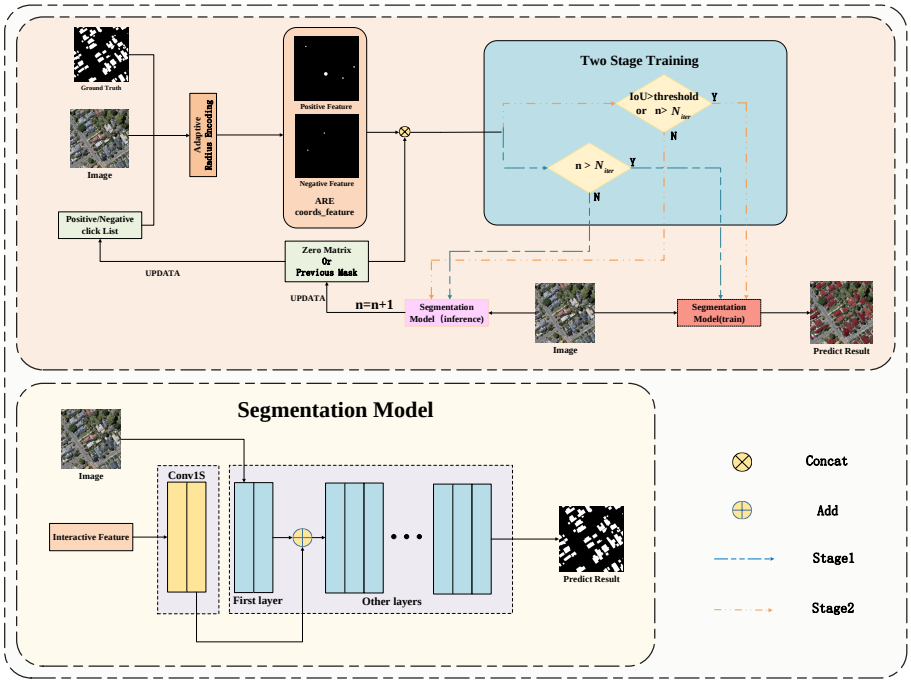


Figure 1. Pipeline of the proposed ARE-Net with the architecture of the segmentation model. The ARE module generates coordinate features from the original image, a positive/negative points list, and the ground truth and connects them with the previous prediction mask as input features. Then, it determines the input features through conditional judgment and sends them to different stages of the segmentation network. The inference mode in the two stages aims to update the positive/negative points list using the prediction mask of the segmentation network. The training mode aims to train the network parameters using input features and output the predicted segmentation mask. In the two-stage training strategy, N_{iter} is the iteration threshold, and n is used as the counting variable. Each time the positive and negative click queue is updated using the mask inferred by the model, an interaction click is added, and n is increased by 1. The number of interactive clicks should not exceed N_{iter}.

2.2. Adaptive-Radius Encoding

The current mainstream interactive information encoding methods transform the coordinates in the interactive sequence into interactive features using fixed-radius binary disk coding. Then, the fusion module fuses these features with the original image features.

Fixed-radius binary disk coding works by using the coordinates in the interaction sequence as the center of a circle. It employs one-hot coding to assign a value of 1 to the region within the fixed radius and 0 to other regions. This local coding approach effectively provides interaction information to the network and mitigates the effects of network confusion to some extent [47]. Therefore, we also adopted this local coding approach to encode interactive clicks and obtain interactive features.

In click-based interaction segmentation, positive and negative clicks have distinct interaction effects. For positive clicks, the binary disk encoding centered on the click coordinates should cover as much of the positive sample area as possible. This ensures that the interaction features contain more semantic information about the building objects. However, for buildings of different shapes, fixed-radius disk encoding will inevitably include information about non-building areas, leading to confusion for the network. On the other hand, negative clicks should randomly appear around buildings and non-building areas. The corresponding binary disk encoding should be as close as possible to the edge areas of the buildings, providing boundary information to constrain the extraction of buildings. As a result, simply enlarging the radius to provide more a priori information is not suitable for extracting buildings with different shapes from high-resolution remote sensing images. Furthermore, the original method samples negative clicks far away from the target, and this kind of sampling will produce ineffective redundancy information in the binary classification problem. Based on these two observations, we designed the adaptive-radius binary disk coding algorithm. For coding positive clicks, the algorithm uses a distance transformation to compute the maximum inner tangent circle radius of the region where the clicks are located in the interaction sequence. This radius is then used as the coding radius for the corresponding clicks. For coding negative clicks, the clicks are located at the edges of the building region, and the minimum distance between the click region and the building is computed using the distance transformation. This distance is then used as the radius of the disk.

Figure 2 shows the class activation mappings (CAMs) generated in the network after encoding the interaction sequences into feature maps using both methods. For buildings of different shapes, the fixed-radius binary disk coding method often produces inaccurate results. As shown in Figure 2, the polygonal area represents a non-building region, indicating that the fixed-radius disk coding scheme included some background regions in the feature maps of the positive clicks. This led to confusion in the final extraction results of the network. To address this issue, the proposed ARE can capture more semantic information about the interaction features from the network model, and it can be seen that the obtained heat zones were all within the building area.

We present the pseudo-code for the ARE algorithm in Algorithm 1. The algorithm operates in two phases: random sampling and corrected sampling. During the random sampling phase, line 1–3, we decompose the binary mask into positive/negative masks using one-hot coding, and then use distance transformation to obtain the corresponding positive/negative distance map. In line 4, by combining the two maps, we can calculate the inscribed circle radius of all pixel closest to the building boundaries. In line 5–9, we convert the clicks into a distance map through distance transformation, calculate the inscribed radius from it, and limit the inscribed radius value of negative clicks to within 5. In line 10, $\varphi(point\_radius, point\_InscribedDistance)$ maps each radius distance to the encoding radius of the corresponding click. In line 11, after the disk encoding, all positive/negative interactive disk mappings are merged to obtain AREMap. Distance transform is a method in computer graphics. The main function of this method is to obtain the distance from each pixel in the mask to the background pixel. In the corrected sampling phase, new clicks are generated based on the prediction results. The same method is used to calculate the distances and update the clicks and distances in the list. Subsequently, the feature maps are generated and returned to the network.

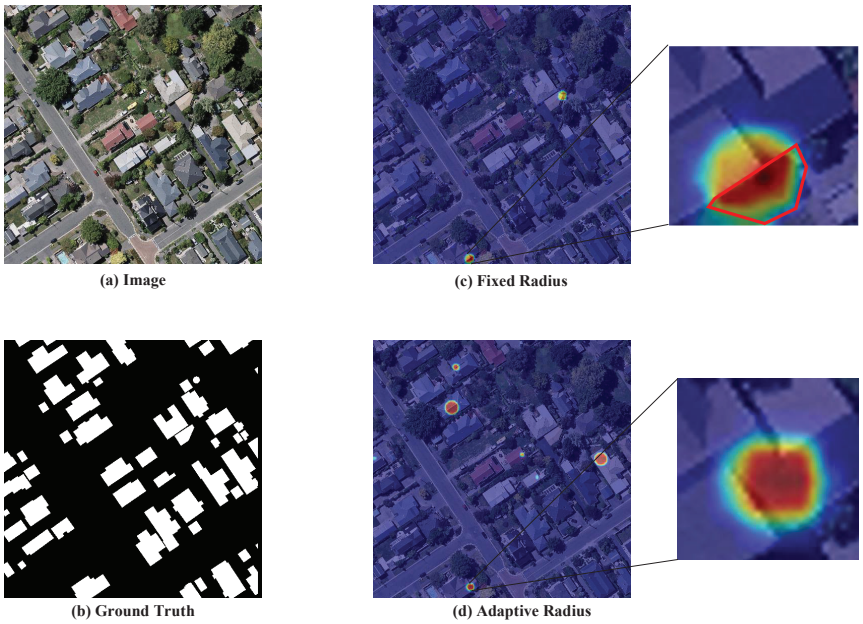


Figure 2. The class activation mappings of fixed-radius binary disk coding and ARE's disk coding. In the fixed-radius binary disk coding, all clicks are encoded using the same radius, while in ARE, the radius of encoding is dynamically adjusted according to the location of the clicks to avoid the problem of network confusion due to anomalous buildings and provide more correct a priori information from the same clicks.

Algorithm 1 Adaptive-radius encoding

Input: click_List, previous_masks, images
Output: ARE_Maps
PositiveMask, NegativeMask = One-Hot(previous_masks)
PositiveDistanceMap = distanceTransform(PositiveMaskt, images)
NegativeDistanceMap = distanceTransform(NegativeMask, images)
incircle_maps = max(PositiveDistanceMap, NegativeDistanceMap)
dist_maps = distanceTransform(click_List, images)
point_radius, point_InscribedDistance=dist_maps[point], incircle_maps [point]
if point is Negative_clicks then **then**
    encoding_map[point] = min(5, encoding_map)
end if
encoding_map = φ(point_radius, point_InscribedDistance)
ARE_Maps = sum(encoding_map)
return ARE_Maps

2.3. Two-Stage Training Strategy

The current mainstream iterative correction sampling strategy involves the following steps: Firstly, a random maximum iteration threshold, denoted as N, is set. The initial correction sampling is performed using model predictions on randomly sampled clicked images. New clicks are added within the region of maximum error prediction. The remaining $N-1$ sampling iterations are conducted based on the model predictions of the prediction map from the previous round of correction sampling, along with the augmented clicked images, and new clicks are added in a similar manner as before. Finally, the prediction map, along with the final set of clicked images, is sent for training.

After analyzing the IoU of the prediction maps generated at each iteration of the model during training, we observed a fluctuation in the IoU with an increase in clicks during the initial stages of training. However, we noticed that the IoU gradually improved as the number of clicks increased in the later stages, as shown in Figure 3. This indicated that the training objectives differed between the first and second stages of the network. During the early stages of network training, the IoU exhibited erratic fluctuations, indicating that the primary task at this point was for the network to learn the motivating effect of the interaction features. Conversely, in the later stages of training, the IoU demonstrated a consistent upward trend with increasing clicks. This suggested that the focus in the later stages should be on enhancing the network's RoC performance. Our work aimed to achieve higher accuracy with minimal clicks. However, during the later stages of training, the model tended to prioritize fine-tuning the network by adding more clicks while potentially neglecting the comprehensive learning of local features in newly clicked regions. As a result, treating clicks at different stages indiscriminately led to limited improvement in the *IoU* and challenged the network's ability to improve the number of clicks (NoC) metric.

We divided the model training into two stages. The first stage remained the same as the original training stage, aiming to enhance the network's performance for building extraction. The second stage focused on fine-tuning, and we introduced two key improvements: (1) During each correction sampling iteration, we calculated the IoU of the model's prediction results. If the IoU reached a predefined threshold, we terminated the current round of iterative sampling early and proceeded to train the model directly. (2) We introduced the iteration loss function to maximize the improvement of the NoC metric. This loss function is shown in Equation (1).

$$iteration\_loss = \sum_{k=0}^{n}(1 - IoU_k) \tag{1}$$

Considering the number of correction sampling iterations and the IoU achieved in each correction, if there are more correction samplings and the IoU for each correction is lower, the value of this loss will be larger.

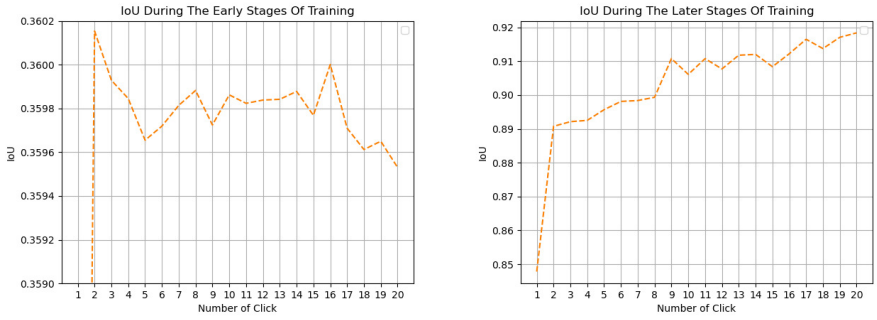


Figure 3. Changes in IoU with increasing clicks at different stages.

We used the normalized focal loss (NFL) [52] to calculate the loss of the model output results and promote the convergence of the model. The final design of our loss function is shown in Equation (2).

$$loss = \frac{1}{2}iteration\_loss + \frac{1}{2}NFL \tag{2}$$

3. Experiments and Results

3.1. Dataset

In our experiments, we assessed the effectiveness of our proposed ARE-Net using two open building extraction datasets: WHU and Inria. Below are presented detailed descriptions of these datasets.

Inria dataset [51] (this dataset can be downloaded from: https://project.inria.fr/aerialimagelabeling/, accessed on 1 January 2022). The Inria dataset comprises five subsets: Austin, Chicago, Kitsap, Tyrol, and Vienna. Each subset contains 36 tiles with a spatial resolution of 0.3 m and a size of 5000 × 5000 pixels. The total area covered by the dataset is 405 km², with a total of 180 tiles. These subsets represent different cities and include buildings with diverse styles, structures, and shapes. Notably, the highest average accuracy achieved on this dataset using fully automated segmentation networks is only 76.21%. Therefore, due to the varying distribution of buildings and the complex environment surrounding them, this dataset poses challenges for interactive building extraction. To ensure fairness, we divided the dataset as follows: the first 25 tiles of each city were used as the training set, tiles 26–30 were allocated to the validation set, and the remaining tiles were designated for testing the effectiveness of our method. Each tile was sequentially cropped into non-overlapping RGB images with a size of 512 × 512 pixels. Consequently, the training set, validation set, and test set of each city consisted of 2500, 500, and 600 images, respectively. Sample images from the Inria dataset are shown in Figure 4.

The WHU building dataset [50] (this dataset can be downlaod from http://gpcv.whu.edu.cn/data/, accessed on 1 January 2022). The WHU dataset is renowned as a challenging dataset for building detection. It comprises three sub-datasets: an aerial image dataset, satellite dataset I, and satellite dataset II. For our experiments, we focused on the aerial image dataset, as shown in Figure 5, which consists of 8188 non-overlapping RGB images with a size of 512 × 512 pixels collected over Christchurch, New Zealand. The training set, validation set, and test set of this dataset contain 4736, 1036, and 2416 images, respectively. We followed the official setup provided by the dataset creators to conduct our experiments.

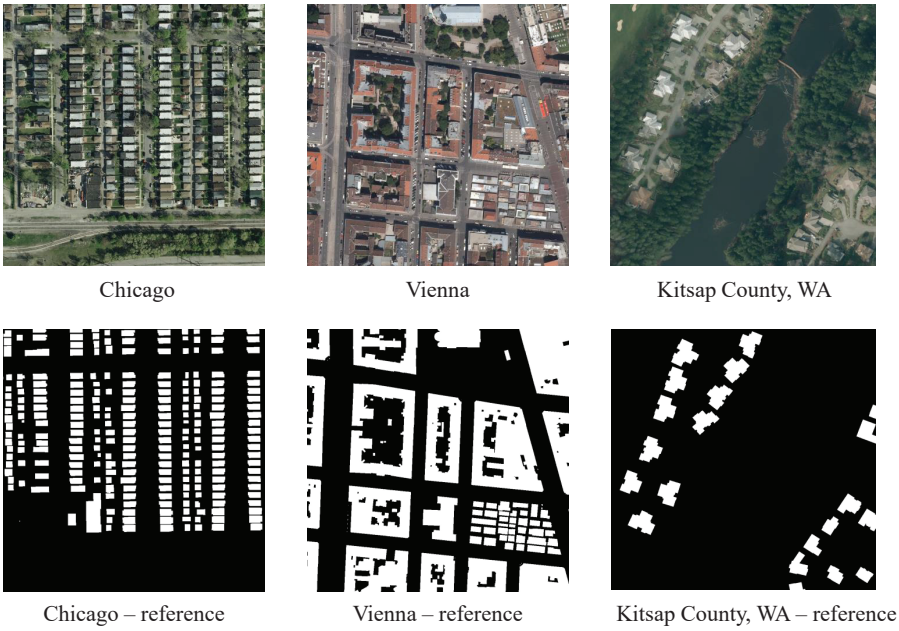
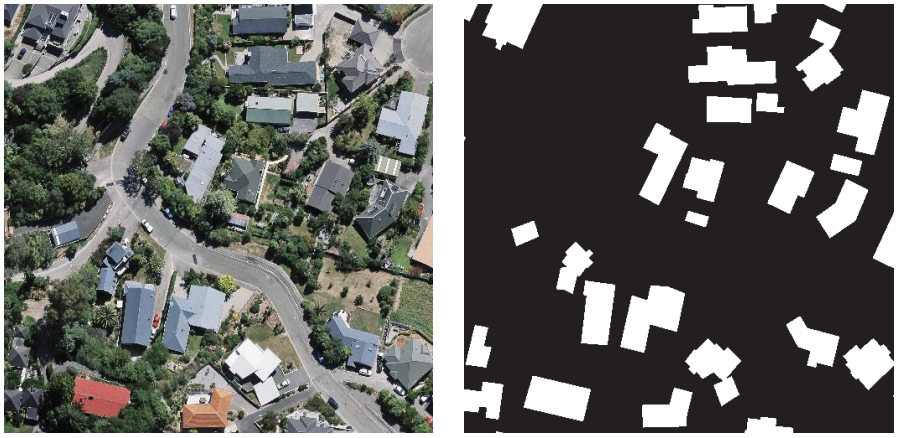


Figure 4. Sample imagery of Inria dataset.



The sample of aerial imagery The sample of aerial imagery—reference

Figure 5. Aerial image dataset sample of the WHU dataset.

By utilizing these datasets, we aimed to address the difficulties posed by varying building distributions and complex environmental factors and evaluate the effectiveness of our interactive building extraction method.

3.2. Evaluation Metrics

To ensure a fair comparison, we employed an extensive sampling strategy [20,47,48] to simulate the number of clicks during the evaluation process. This strategy involved simulating the clicking process using ground-truth masks as references. The first click was placed at the center of the ground-truth mask, while subsequent clicks were placed at the center of the region with the maximum error from the previous mask. This approach allowed us to simulate the interactive process and evaluate the performance of different methods.

Considering the challenges posed by these datasets, we used the $NoC_{80/85/90}$ metric to assess the RoC of different methods. $NoC_{80/85/90}$ represents the average number of clicks required for the segmentation results to achieve the target IoU of 80%, 85%, and 90%, respectively. The formula for calculating $NoC_{80/85/90}$ is shown in Equation (3).

$$NoC_k = \frac{1}{N}\sum_{i=1}^{N} n_{i,k} \tag{3}$$

where N is the number of predicted images, and $n_{i,k}$ denotes the total number of clicks consumed for the ith image to reach $IoU@k$. As in the previous work, we limited the number of clicks to 20 when comparing the RoC.

We also employed the same evaluation metric as in [48], $NoF^{100}_{85/90}$ (number of failures: the number of images whose IoU did not reach 85/90 in a maximum of 100 clicks), to evaluate the generalizability. The formula is shown as follows in Equation (4):

$$NoF_k = \sum_{i=1}^{N} m_{i,k} \tag{4}$$

where N is the number of predicted images, $m_{i,k} = \{0, 1\}$, and $m_{i,k} = 0$ means that the ith image can reach the target $IoU@k$ in a maximum of 100 clicks before the maximum number of clicks.

Considering the labor cost and the possibility of using the model on mobile devices and websites, we compared the time to process the entire data set ($Time$) and seconds per click (SPC), whose formula is shown in Equation (5).

$$SPC = \frac{Time}{all\_clicks} \tag{5}$$

where *Time* is the time taken to process the whole dataset, and all_clicks is the total number of clicks consumed in the case of a maximum of 20 clicks.

In addition, we evaluated the convergence based on the average IoU@k curves of different methods at 20 clicks [20] and the generalization ability of different methods based on their NoF metric at a maximum of 100 clicks [47].

3.3. Implementation Details

To ensure the fairness of the experiments, we followed a consistent procedure to train the models. The images were randomly resized within the range of (0.75, 1.4), maintaining their aspect ratio. Both the Inria and WHU datasets had an original image input size of 512×512 pixels. We also applied image transformations such as inversion and random variations in brightness, contrast, and RGB values. During the random sampling and correction sampling stages, we limited the total number of positive and negative clicks to not exceed 24. By consistently applying these preprocessing steps and limitations across the training process, we aimed to establish a fair and comparable experimental setup for evaluating the performance of our proposed method.

In our experiments, we employed the Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$, which are hyperparameters used to control the exponential decay rates for the first and second moments of the gradients, respectively. We trained two models with different sizes: HRNet18s+OCR and HRNet32+OCR [53,54]. The batch size was set to 16 for HRNet18s+OCR and 8 for HRNet32+OCR. Accounting for the variations in building size distributions across different sub-datasets, we set the min_object parameter to 0. This meant that we did not remove any buildings with small targets during the training process. This choice allowed us to retain all buildings in the dataset, regardless of their size. The training process consisted of a total number of epochs, a learning rate, and a learning rate decay strategy that were consistent with the RITM method [47]. Meanwhile, we divided the total number of epochs into two segments. The first 150 epochs followed the original training strategy, while the remaining epochs incorporated our iterative training strategy. By utilizing these training configurations, we aimed to compare the performance of our models against the baseline and evaluate the effectiveness of our proposed iterative training strategy.

3.4. Results

Our method utilized CNN-based models to adapt and accelerate inference on mobile devices and websites. We collaborated with several mainstream methods based on CNN models: RITM [47], RGB-BRS, f-BRS [20], BRS [35], and FocalClick [48]. To ensure the fairness of the experiment, we implemented all of them using the RITM framework.

3.4.1. RoC and CG Analysis

In order to draw a fair comparison regarding the convergence speed of different methods for high-resolution image building extraction, we compared the NoC_{80}, NoC_{85}, and NoC_{90} among the six methods using the same segmentation models (HRNet18s+OCR and HRNet32+OCR), as shown in Tables 1 and 2. It was evident that our method achieved a significant improvement on all sub-datasets of the Inria and WHU datasets when using the same model. In the HRNet18s+OCR model, we observed improvements of 17.92%, 15.14%, and 7.98% for NoC_{80}, NoC_{85}, and NoC_{90}, respectively, on the Inria sub-dataset. Simultaneously, on the WHU dataset, we achieved improvements of 10.53%, 13.38%, and 13.03% for NoC_{80}, NoC_{85}, and NoC_{90}, respectively.

Table 1. Comparison of the NoC_{80}, NoC_{85}, and NoC_{90} for HRNet18s+OCR (the best results are in bold).

Data		NoC	RITM	RGB-BRS	BRS	f-BRS	FocalClick	ARE-Net
Inria	Austin	NoC_{80}	4	4.01	4.16	4.58	12.39	**3.75**
		NoC_{85}	9.75	9.9	10.09	11.03	14.92	**9.49**
		NoC_{90}	15.81	15.95	16.02	16.15	16.89	**15.55**
	Chicago	NoC_{80}	5.13	5.43	5.46	5.58	7.23	**4.17**
		NoC_{85}	7.9	8.52	8.53	8.77	9.74	**6.8**
		NoC_{90}	11.99	13.07	13.01	13.07	13.43	**10.75**
	Kitsap	NoC_{80}	10.3	11.06	11.49	11.43	14.19	**9.89**
		NoC_{85}	14.05	14.55	14.99	15.28	16.58	**13.55**
		NoC_{90}	17.89	18.05	18.09	18.44	18.45	**17.38**
	Tyrol	NoC_{80}	5.16	5.36	5.46	5.47	10.16	**3.81**
		NoC_{85}	11.19	11.78	12.14	11.85	13.77	**9.15**
		NoC_{90}	16.73	17.06	16.89	16.87	17.08	**15.35**
	Vienna	NoC_{80}	9.17	9.22	9.47	9.65	10.71	**7.7**
		NoC_{85}	13.12	13.36	13.69	13.73	13.73	**11.07**
		NoC_{90}	17.5	17.53	17.72	17.27	17.27	**16.23**
	Average	NoC_{80}	6.752	7.016	7.208	7.342	10.936	**5.864**
		NoC_{85}	11.202	11.622	11.888	12.12	13.748	**10.012**
		NoC_{90}	15.984	16.332	16.346	16.456	16.624	**15.052**
WHU		NoC_{80}	1.71	1.66	1.66	1.67	11.62	**1.53**
		NoC_{85}	2.69	2.7	2.75	2.28	13.62	**2.33**
		NoC_{90}	9.98	10.22	10.38	10.9	15.69	**8.68**

Table 2. Comparison of the NoC_{80}, NoC_{85}, and NoC_{90} for HRNet32+OCR (the best results are in bold).

Data		NoC	RITM	RGB-BRS	BRS	f-BRS	FocalClick	ARE-Net
Inria	Austin	NoC_{80}	2.77	2.65	2.95	2.77	12.79	**2.31**
		NoC_{85}	6.95	6.92	7.64	7.62	14.99	**6.05**
		NoC_{90}	14.2	14.21	14.67	14.57	17	**13.54**
	Chicago	NoC_{80}	4.35	4.66	4.91	4.58	7.83	**3.36**
		NoC_{85}	6.71	7.41	7.73	7.22	10.17	**5.63**
		NoC_{90}	10.82	11.77	12.2	11.53	13.67	**9.32**
	Kitsap	NoC_{80}	9.67	10.37	10.87	10.64	14.73	**9.35**
		NoC_{85}	13.32	13.84	14.32	14.27	16.76	**13.05**
		NoC_{90}	17.48	17.64	17.78	18.01	18.55	**16.97**
	Tyrol	NoC_{80}	4.53	4.69	4.73	4.67	10.63	**3.4**
		NoC_{85}	10.63	11.26	11.31	11.05	14.41	**8.4**
		NoC_{90}	16.28	16.96	16.79	16.65	17.27	**14.92**
	Vienna	NoC_{80}	7.55	7.7	7.92	7.9	10.08	**6.11**
		NoC_{85}	10.74	11.11	11.49	11.36	13.85	**9.29**
		NoC_{90}	15.89	16.14	16.52	16.57	17.18	**14.44**
	Average	NoC_{80}	5.774	6.014	6.296	6.112	11.212	**4.906**
		NoC_{85}	9.67	10.108	10.498	10.31	14.036	**8.484**
		NoC_{90}	14.934	15.344	15.592	15.466	16.734	**13.838**
WHU		NoC_{80}	1.59	1.52	1.52	1.53	11.13	**1.43**
		NoC_{85}	2.19	2.1	2.2	2.19	13.13	**1.88**
		NoC_{90}	5.94	6.12	6.71	6.87	15.39	**5.02**

For HRNet32+OCR, we observed improvements of 15.03%, 12.26%, and 7.34% for NoC_{80}, NoC_{85}, and NoC_{90}, respectively, on the Inria sub-dataset, and 10.06%, 14.16%, and 15.49% on the WHU dataset. Additionally, the average recall error (ARE) demonstrated performance improvements across all sub-datasets, indicating that our approach was not limited to a specific geographic building style but was applicable to various geographic building categories. Furthermore, our method exhibited a noticeable improvement in large-scale buildings or sub-datasets with a dense building distribution (such as Chicago, Tyrol, and Vienna), which provided evidence that the method effectively leveraged a priori information to enhance building extraction.

Figures 6 and 7 show the mean IoU@K curves for different methods using the HRNet18s+OCR and HRNet32+OCR models to investigate the changes with an increasing click count. The increase in the number of clicks not only provides more guidance information, but also increases noise, leading to network confusion. To reduce the noise impact caused by input changes, RITM, FocalClick, and ARE-Net add previous masks, while BRS and RGB BRS use multiple forward and back passes through the entire model. F-BRS is a lightweight solution for BRS and RGB-BRS. Considering the time cost for the multiple propagation of BRS and RGB-BRS, only auxiliary parameters are used to optimize the network. As shown in the figure, although this method could effectively reduce time costs, it resulted in the network being sensitive to noise and decreasing when the number of clicks increased to eight, while other methods had a stable upward trend. According to the figure, ARE-Net had a significant advantage when the number of clicks was less than five on the Inria dataset with fewer samples. For the HRNet18S+OCR model, ARE-Net could achieve an 85% IoU with 10 clicks. For the HRNet32+OCR model, an 85% IoU could be achieved with just seven clicks. Compared to the other methods on the WHU dataset with a large number of samples, the advantage was not significant when the number of clicks was low, but the upward trend of ARE-NET was stable and quick, reaching a 90% IoU faster. ARE-Net achieved a 90% IoU with five clicks for HRNet18S+OCR, while the other methods required nine clicks. For HRNet32+OCR, we achieved a 90% IoU with three clicks, while the other methods required five clicks. This indicated that compared to existing methods, our proposed method could provide faster, more effective convergence and lower interaction cost requirements.

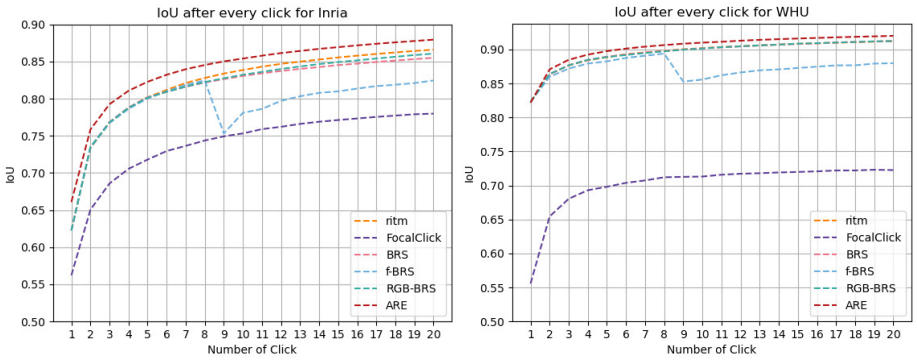


Figure 6. Mean IoU@k curves for the WHU dataset and Inria dataset in HRNet18s+OCR with an increasing click number.

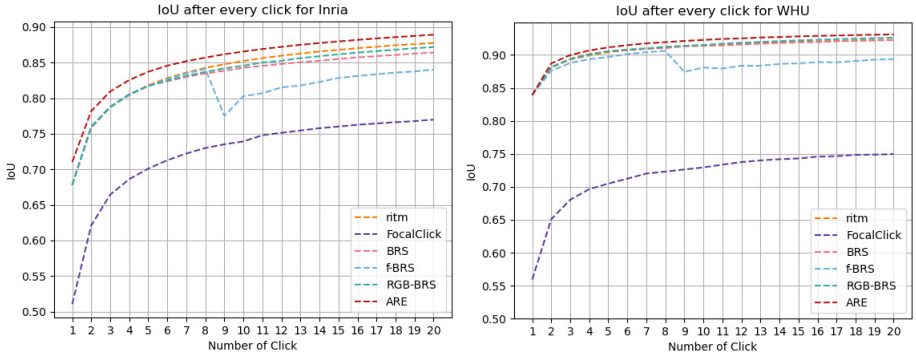


Figure 7. Mean IoU@k curves for the WHU dataset and Inria dataset in HRNet32+OCR with an increasing click number.

3.4.2. Generalizability Analysis

Generalizability is a crucial aspect of interactive segmentation algorithms. To assess the generalization performance, we conducted computer simulations on both the Inria and WHU datasets, considering up to 100 clicks using two different models. The results are summarized in Table 3. Notably, FocalClick exhibited a limited generalization capability, as it failed to reach the set IoU for more than 1000 test sets out of a total of 3000 test sets in the Inria and WHU datasets. This degradation in generalization could be attributed to FocalClick's focus on fine-tuning segmentation by zooming in on regions with segmentation errors, which neglects the segmentation of other targets in complex multi-target scenarios. Both BRS and f-BRS optimized segmentation by tuning the network and exhibited good generalization performance on the WHU dataset. However, their performance on the more challenging Inria dataset fell short. In contrast, RGB-BRS and RITM demonstrated favorable results on both datasets. Importantly, our method outperformed RGB-BRS and RITM in terms of generalization. Even though some images did not achieve the set accuracy using our method, the number of such images was significantly lower than for the other methods. This finding indicated that our method offered a higher generalization performance, serving as strong evidence of its effectiveness. The robust generalization capability of our method underscores its suitability for diverse and complex scenarios, reinforcing its potential for real-world applications.

Table 3. Generalizability analysis on Inria and WHU datasets. NoF_{IoU}^{100} indicates the number of images that could not reach the specified IoU within 100 clicks for both models (the best results are in bold).

Model	Data	NoF	RITM	RGB-BRS	BRS	f-BRS	FocalClick	ARE-Net
HRNet18s +OCR	Inria	NoF_{85}^{100}	253	314	468	789	1067	**158**
		NoF_{90}^{100}	787	807	1092	1377	1360	**603**
	WHU	NoF_{85}^{100}	43	11	12	25	1023	**10**
		NoF_{90}^{100}	131	85	225	635	11980	**51**
HRNet32 +OCR	Inria	NoF_{85}^{100}	169	147	438	600	1020	**99**
		NoF_{90}^{100}	590	649	1019	1212	1309	**419**
	WHU	NoF_{85}^{100}	36	9	14	14	922	**8**
		NoF_{90}^{100}	78	27	136	248	1122	**25**

3.4.3. Labeling Cost Analysis

Reducing the time cost of annotation is another crucial aspect of interactive segmentation algorithms. To evaluate our method against previous work in terms of time cost, we analyzed the seconds per click (SPC) and the total time consumed (time) based on the rules described in Section 3.2. All experiments for this metric were conducted on a 5 vCPU Intel(R) Xeon(R) Platinum 8338C CPU @ 2.60GHz, without any external interference. Table 4 presents the results of our analysis. RGB-BRS and BRS required constant back-and-forth

passes in the network, resulting in a large computational burden and significantly increased the time consumed per click. Consequently, the labeling time cost was substantially higher for these methods. Although f-BRS optimized only the network parameters, its inference speed was better than that of the previous two methods. However, due to the increased number of clicks and the resulting decline in accuracy, the labeling time cost remained relatively high. FocalClick utilized the ZoomIn method, which reduced the number of parameters in the network and achieved a relatively good SPC index. However, the method did not effectively address multi-target segmentation tasks, leading to increased time consumption and an unsatisfactory annotation time cost. In contrast, RITM and our method (ARE) required the least annotation time cost. Our method performed significantly better than RGB-BRS, BRS, and f-BRS in terms of annotation time cost for different models and datasets. This result demonstrated that our method had minimal impact on the time cost for annotation, ensuring efficient and effective interactive segmentation. By reducing the time required for annotation, our method enhances the user experience and facilitates practical applications that involve large-scale or time-sensitive segmentation tasks.

Table 4. Average time per click and total time cost for WHU and Inria in HRNet18s+OCR and HRNet32+OCR (the best results are in bold, and the second-best results are underlined).

Model	Data	Time Cost	RITM	RGB-BRS	BRS	f-BRS	FocalClick	ARE-Net
HRNet18s +OCR	Inria	SPC, s	**0.0298**	1.3492	0.9994	0.0622	<u>0.0334</u>	0.034
		Time, H:M:S	**0:18:56**	14:00:28	10:43:48	0:40:47	0:21:59	<u>0:19:40</u>
	WHU	SPC, s	**0.024**	0.885	0.06	0.035	<u>0.033</u>	<u>0.033</u>
		Time, H:M:S	**0:06:51**	6:11:53	4:25:08	0:16:53	0:15:45	<u>0:08:19</u>
HRNet32 +OCR	Inria	SPC, s	**0.0566**	1.2884	0.8944	0.0994	0.053	<u>0.0574</u>
		Time, H:M:S	<u>0:32:44</u>	13:00:40	9:06:25	1:00:26	0:35:01	**0:31:59**
	WHU	SPC, s	<u>0.06</u>	1.338	0.939	0.096	0.053	**0.059**
		Time, H:M:S	<u>0:10:21</u>	3:56:11	3:01:04	0:19:03	0:23:36	**0:08:37**

3.4.4. Comprehensive Comparison

In Table 5, we rank the methods based on the metrics of RoC, CG, generalizability, and labeling costs for the two different models. RGB-BRS and BRS prioritized performance at the expense of increased labeling costs. This resulted in higher annotation time costs compared to the other methods. On the other hand, FocalClick sacrificed performance in order to reduce labeling costs but failed to achieve a balanced performance. Our method exhibited slightly higher annotation costs on the small model compared to RITM but outperformed RITM on the large model, indicating its higher potential. Overall, our method achieved the top ranking, demonstrating superior performance across all metrics. The comprehensive analysis indicated that our method struck a good balance between performance and labeling costs, offering the best overall results among the compared methods.

Table 5. Comprehensive ranking of the different methods.

Model	Metrics	First	Second	Third	Fourth	Fifth	Sixth
HRNet18s +OCR	ROC	ARE	RTIM	RGB-BRS	BRS	f-BRS	FocalClick
	CG	ARE	RTIM	RGB-BRS	BRS	f-BRS	FocalClick
	Generalizability	ARE	RGB-BRS	RITM	BRS	f-BRS	FocalClick
	Labeling costs	RITM	ARE	FocalClick	f-BRS	BRS	RGB-BRS
HRNet32 +OCR	ROC	ARE	RTIM	RGB-BRS	BRS	f-BRS	FocalClick
	CG	ARE	RTIM	RGB-BRS	BRS	f-BRS	FocalClick
	Generalizability	ARE	RGB-BRS	RITM	BRS	f-BRS	FocalClick
	Labeling costs	ARE	RITM	FocalClick	f-BRS	BRS	RGB-BRS

4. Discussion

Interactive semantic segmentation differs from general semantic segmentation in that it incorporates human clicks to guide the network's segmentation task. These clicks are converted into feature maps using binary disk coding and then fed into the network. As a result, the feature maps are transformed differently based on the size of the clicks. In high-resolution remote sensing images, buildings exhibit diverse shapes and characteristics, making the click radius an important factor. Expanding the click radius of a fixed-size may introduce non-target areas into the feature maps and degrade network training, while using a smaller click radius may weaken the influence of human clicks in guiding the network. The network is trained by learning from RGB images and click-transformed feature maps, aiming to achieve click-guided network segmentation. The network's accuracy fluctuates during the early stage of training as the number of clicks increases but gradually improves as more clicks are added in the later stage. This discrepancy reflects the differences in training objectives between the pre-training and post-training phases. In the initial stage, the network learns to handle the increasing click information, while in the later stage, the focus is on expanding the loss cost associated with additional clicks to enhance network convergence. This is why ARE-Net improved in terms of all performance metrics. To demonstrate the effectiveness of our method, we visualized the interaction process and conducted ablation experiments, as presented in the following sub-sections.

4.1. Comparison with Fully Supervised Classification Methods

To demonstrate the low cost and effectiveness of ARE-Net, we chose HRNet18+OCR and HRNet32+OCR as the benchmark models for fully supervised classification and compared their classification results with those of ARE. The results are shown in Tables 6 and 7. According to Table 6, under fully supervised classification, the IoU of HRNet18s+OCR for the Inria and WHU datasets reached 73% and 87.86%, respectively, while ARE exceeded the fully supervised building extraction results with only two to three clicks. Similarly, from Table 7, it can be observed that the IoU of HRNet32+OCR for the Inria and WHU datasets reached 74.61% and 89.06%, respectively. ARE also exceeded the fully supervised building extraction results with a maximum of three clicks. This indicated that the interaction extraction method proposed in this article could significantly improve the accuracy of building extraction with a small amount of human–computer interaction.

Table 6. Comparison of extraction accuracy between HRNet18s+OCR and ARE with different click counts.

Dataset	HRNet18s+OCR	ARE (with 1 Click)	ARE (with 2 Clicks)	ARE (with 3 Clicks)	ARE (with 5 Clicks)	ARE (with 10 Clicks)	ARE (with 20 Clicks)
Inria	73%	66.86%	75.84%	79.25%	82.27%	85.39%	87.94%
WHU	87.86%	82.15%	87.07%	88.48%	89.75%	91%	92.02%

Table 7. Comparison of extraction accuracy between HRNet32+OCR and ARE with different click counts.

Dataset	HRNet32+OCR	ARE (with 1 Click)	ARE (with 2 Clicks)	ARE (with 3 Clicks)	ARE (with 5 Clicks)	ARE (with 10 Clicks)	ARE (with 20 Clicks)
Inria	74.61%	71.02%	78.17%	80.95%	83.71%	86.55%	88.92%
WHU	89.06%	83.89%	88.72%	89.99%	91.15%	92.28%	93.11%

4.2. Visualization Analysis

Interactive progress visualizations for six interactive segmentation methods in building extraction are shown in Figures 8 and 9. Figure 8 shows the extraction results of buildings after 1, 3, 4, and 6 clicks, while Figure 9 shows the extraction results after 7, 9, 10, and 11 clicks. The corresponding IoU values are listed under each image. The first row shows the original image, and the second row shows the ground truth. The methods in each column from left to right are FocalClick, RGB-BRS, BRS, f-BRS, Ritm, and the proposed ARE-Net method. The red areas on the graphs are building masks. Due to the use of adaptive-radius binary encoding in ARE-Net, the size of the green dots representing positive clicks and red dots representing negative clicks adapted to the size of the building and non-building areas, as shown in the sixth column of the figure. The green and red dots in columns one to five are fixed in size. According to these visualization figures, our method demonstrated superior performance compared to the others. Each click was supervised effectively during the two-stage training strategy of ARE-Net.

In Figure 8, the third and fourth rows demonstrate the advantages of the ARE module. After the first click, most methods only extracted a portion of the building in the upper right corner of the image (marked by a yellow box). ARE-Net covered a more complete range of buildings. This was because the interaction information provided by the fixed-radius click method was limited, and multiple clicks were required for correction. The ARE module maximized the extraction of useful information from positive clicks through the adaptive radius. Compared with other methods for identifying missing building information in the upper right corner of the image, ARE-Net better identified building information in that area. In the third row, only FocalClick and ARE-Net could achieve an IOU of over 80% after the first click, and the method in this article achieved a value 4% higher than that of FocalClick. After another positive and negative click, i.e., after three clicks, ARE-Net extracted the elongated buildings in the central area of the image that had not yet been extracted after the first click, while the other methods could only extract partial area information of the building. Compared to the other methods, ARE-Net had the highest IOU value after three simulated click samples, at 89.58%, which was 3% 6% higher than the result of the other methods. As shown in the blue-framed images in rows five and six, the building footprints in the yellow box area could be completely extracted by other methods through two clicks, while ARE-Net only needed one click to achieve the complete extraction of building information, which demonstrated that ARE-Net had lower interaction costs and a faster RoC. Meanwhile, ARE-Net was the only method that achieved a 90% IOU after four clicks, which was significantly superior to the other methods.

The extraction visualization images after the 7th, 9th, and 11th clicks are shown in Figure 9. Although FocalClick achieved a better IoU with fewer clicks in Figure 8, the mask accuracy actually decreased as the number of clicks increased, as shown in the first column with yellow borders in Figure 9. This was because building extraction is a multi-objective segmentation task, and FocalClick adopted an amplification strategy, which caused the network to focus on the same area (the yellow boundary area in the upper left corner of the image), only increasing the number of clicks in that area, without improving the recognition accuracy of other building targets. When RGB BRS, f-BRS, and BRS were applied to obscured targets such as adding clicks in the occluded area for the 9th, 10th, and 11th clicks (images with green borders in the fifth and sixth rows), the IoU decreased. Typically, when the IoU of f-BRS approached 90% (rows four and six in the figure), additional clicks could cause a sharp decrease in mask accuracy.

According to the visual graphs in Figures 8 and 9, ARE-Net showed a faster rate of convergence, lower interaction costs, and a steadily improved accuracy with the increase in the number of clicks in the interactive process of building extraction.

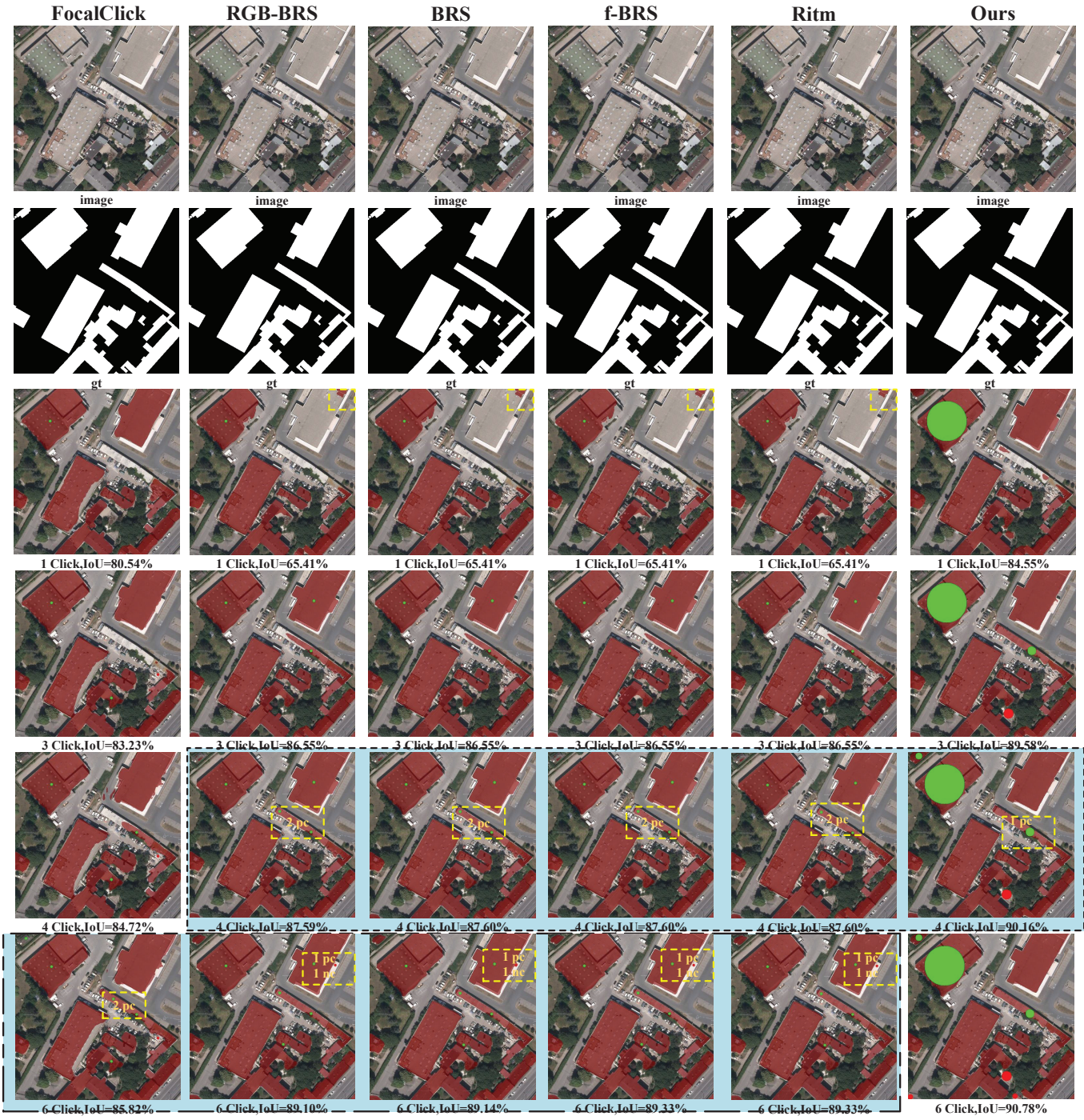


Figure 8. Visualization of the interaction process for all methods at 1–6 clicks. Green and red dots represent positive and negative clicks, respectively. The blue background contains the shortcomings of the other methods compared to our method.

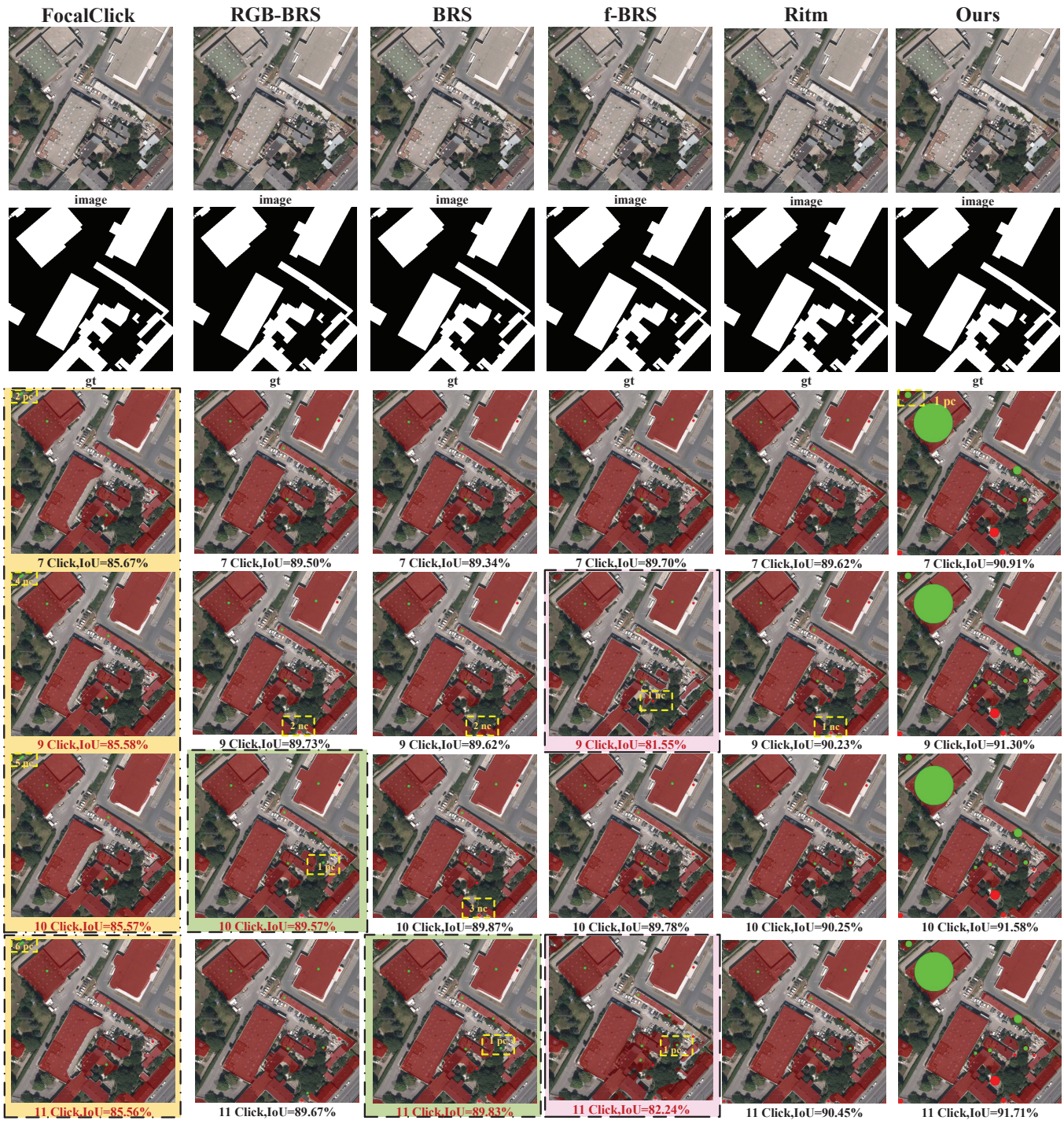


Figure 9. Visualization of the interaction process for all methods at 7–11 clicks. Green and red dots represent positive and negative clicks, respectively. The yellow, green, and pink backgrounds represent the shortcomings of the other methods.

4.3. Qualitative Result

To demonstrate the generalizability of the proposed method across various building shapes, scales, distributions, and other scenarios, we added two datasets with images of buildings with different shapes and distributions in different cities for qualitative experimental analysis. The results are shown in Figure 10. Figure 10a,b,d,e include images from the Chicago, Vienna, and Austin sub-datasets of Inria, and Figure 10c includes images from the WHU dataset. In addition, Figure 10a–d, respectively, include images representing

cases of circular buildings, polygonal buildings, an imbalanced distribution of buildings, and inconsistent building sizes. In the figure, the first column represents the original image, while columns two to four demonstrate the annotation process from scratch. ARE could quickly obtain high-quality predictions with a few clicks. Columns five to six show the results after clicking on the demonstration image to improve the accuracy, and the last column is the ground truth. From Figure 10a–d, it can be seen that ARE-Net could effectively extract buildings with different shapes and distributions, but in Figure 10e with more complex terrain categories, ARE could not effectively extract such slender buildings that were similar to the surrounding environment.

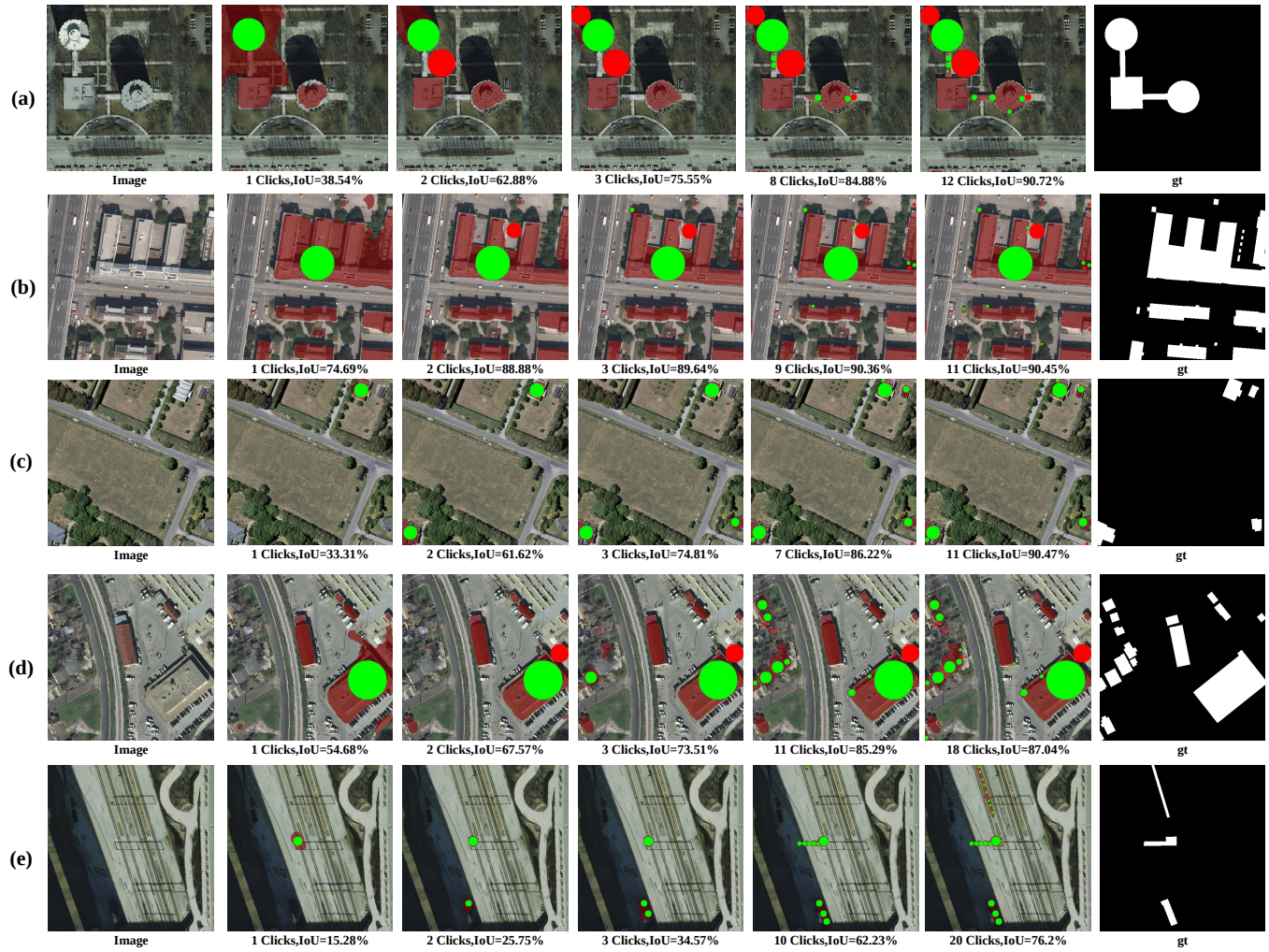


Figure 10. Visualization of the interaction process for ARE-Net with different numbers of clicks on images of buildings with different shapes and distributions in different cities from the two datasets. Green dots and red dots represent positive and negative clicks, respectively. (**a**) Chicago, Inria; (**b**) Vienna, Inria; (**c**) WHU; (**d**) Austin, Inria; and (**e**) Chicago, Inria. The first column shows the original image; columns two to four show the results for 1, 2, and 3 clicks; columns five and six show the results for multiple clicks to improve accuracy; and the last column shows the ground truth.

4.4. Ablation Experiments

The ablation experiments conducted on the Vienna sub-dataset of Inria compared the two-stage training strategy and the ARE module based on the RITM framework. Tables 8 and 9 present the results of these experiments. From the values of NoC_{80}, NoC_{85}, and NoC_{90}, it could be seen that the ARE module and the two-stage training strategy could effectively reduce the value of the NoC, which meant that a higher building extraction

accuracy could be achieved through fewer interactions. At the same time, it could also be observed that compared to the independent use of the ARE module and the two-stage training strategy, their joint use achieved the greatest reduction in NoC_{90}, obtaining the required accuracy with fewer interactive clicks.

On the other hand, the ARE module enhanced the network's potential for building extraction by incorporating a priori information through adaptive-size encoding. This led to notable improvements in the NoC, demonstrating the effectiveness of ARE in leveraging click-guided network segmentation. Combining Table 9, Figures 6 and 7, the average IoU@k curves in the figure and the number of images that did not reach IoU@k at the maximum of 100 clicks in the table reflect the generalization performance of our method. It is worth noting that no additional weights were introduced, which further affirmed the effectiveness of our approach. Overall, the ablation experiments provide strong evidence in favor of the effectiveness of the ARE module compared to the two-stage strategy and the caseline, in terms of improvements in both the NoC and generalization capabilities.

Table 8. RoC evaluation for each fractional ablation experiment on the Vienna sub-dataset.

Baseline	ARE Module	Two-Stage Training Strategy	NoC_{80}	NoC_{85}	NoC_{90}
✓			9.17	13.12	17.5
✓	✓		8	11.67	16.49
✓		✓	8.56	12.27	17.5
✓	✓	✓	7.7	11.07	16.23

Table 9. Generalizability evaluation for each fractional ablation experiment on the Vienna sub-dataset.

Baseline	ARE Module	Two-Stage Training Strategy	NoF_{85}^{100}	NoF_{90}^{100}
✓			163	270
✓	✓		122	218
✓		✓	138	245
✓	✓	✓	113	212

5. Conclusions

In this paper, we presented an efficient method for the interactive extraction of buildings from high-resolution remote sensing images. Our approach leveraged an adaptive-radius encoder and a two-stage training strategy to enhance the network's ability to extract buildings and improve annotation efficiency. Firstly, the adaptive-radius encoding method played a crucial role in transforming each click into a feature map with an adaptive radius size. This approach provided more detailed interaction information to guide the network in accurately extracting buildings. By adapting the click radius, we struck a balance between capturing detailed information and maintaining the effectiveness of the click guidance.

Additionally, our proposed two-stage training strategy effectively addressed the training tasks at different stages. We distinguished the models in the pre-late stage to ensure they focused on specific training objectives. In the later stage, we emphasized the impact of increasing clicks on the loss results, leading to faster network convergence. Extensive experiments were conducted to evaluate our method on various datasets and models. The results demonstrated that our method outperformed existing approaches in terms of both accuracy and efficiency. Notably, on the Inria dataset using the HRNet18s+OCR model, we achieved significant improvements in terms of NoC_{80}, NoC_{85}, and NoC_{90}, as well as NoF_{85}^{100} and NoF_{90}^{100}. Similar improvements were observed on the WHU dataset and the HRNet32+OCR model. Furthermore, our method exhibited optimal convergence and labeling time costs. It struck a balance between annotation efficiency and performance, making it a favorable choice for interactive building extraction. Overall, our method showed state-of-the-art performance in all performance aspects, surpassing previous approaches. The combination of adaptive-radius encoding, two-stage training, and efficient annotation made our method

an effective and practical solution for interactive building extraction in high-resolution remote sensing images. However, the additional computation required for obtaining the adaptive radius size in the coding process introduced some additional time and space consumption costs. Therefore, segmentation methods that are more lightweight and enable faster inference deserve further study. In addition, further generalization verification will be conducted on a wider range of datasets in the future, such as the SZTAKI-INRIA building detection dataset.

Author Contributions: Conceptualization, Q.W. (Qian Weng), Q.W. (Qin Wang) and J.L.; methodology, Q.W. (Qian Weng), Q.W. (Qin Wang) and J.L.; software, Q.W. (Qian Weng), Q.W. (Qin Wang) and Y.L.; validation, Q.W. (Qian Weng), Q.W. (Qin Wang) and J.L.; formal analysis, Q.W. (Qian Weng) and Q.W. (Qin Wang); investigation, Q.W. (Qian Weng), Q.W. (Qin Wang) and J.L.; resources, Q.W. (Qian Weng) and Q.W. (Qin Wang); data curation, Q.W. (Qian Weng), Q.W. (Qin Wang) and Y.L.; writing—original draft preparation, Q.W. (Qian Weng), Q.W. (Qin Wang) and J.L.; writing—review and editing, Q.W. (Qian Weng) and Q.W. (Qin Wang); visualization, Q.W. (Qian Weng) and Q.W. (Qin Wang); supervision, Q.W. (Qian Weng) and Q.W. (Qin Wang); project administration, Q.W. (Qian Weng) and Q.W. (Qin Wang); funding acquisition, Q.W. (Qian Weng) and J.L. All authors have read and agreed to the published version of the manuscript.

Funding: The Natural Science Foundation of Fujian Province under Grant 2023J01432; Fuzhou-Xiamen-Quanzhou National Independent Innovation Demonstration Zone Collaborative Innovation Platform under Grant: 2022FX5; The National Natural Science Foundation of China under the Grant 41801324; The Natural Science Foundation of Fujian Province under Grant 2019J01244.

Data Availability Statement: Our code will soon be available on https://github.com/QinWang-wq/ARE-Net.

Acknowledgments: The authors would like to thank INRIA Sophia Antipolis - Méditerranée for providing the Inria dataset and the Group of Photogrammetry and Computer Vision (GPCV) at Wuhan University for providing the WHU building dataset.

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

References

1. Nikzad, M.; Gao, Y.; Zhou, J. An attention-based lattice network for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5526215. [CrossRef]
2. Farooq, A.; Jia, X.; Hu, J.; Zhou, J. Transferable convolutional neural network for weed mapping with multisensor imagery. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 4404816. [CrossRef]
3. Han, Z.; Hong, D.; Gao, L.; Yao, J.; Zhang, B.; Chanussot, J. Multimodal hyperspectral unmixing: Insights from attention networks. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5524913. [CrossRef]
4. Li, X.; Xu, F.; Liu, F.; Lyu, X.; Tong, Y.; Xu, Z.; Zhou, J. A Synergistical Attention Model for Semantic Segmentation of Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5400916. [CrossRef]
5. Weng, Q.; Chen, H.; Chen, H.; Guo, W.; Mao, Z. A Multisensor Data Fusion Model for Semantic Segmentation in Aerial Images. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 6511905. [CrossRef]
6. Bo, Z.; Chao, W.; Hong, Z.; Fan, W. A review on building extraction and Reconstruction from SAR image. *Remote Sens. Technol. Appl.* **2012**, *27*, 496–503.
7. Feng, T.; Zhao, J. Review and comparison: Building extraction methods using high-resolution images. In Proceedings of the 2009 Second International Symposium on Information Science and Engineering, Shanghai, China, 26–28 December 2009; pp. 419–422.
8. Benedek, C.; Descombes, X.; Zerubia, J. Building development monitoring in multitemporal remotely sensed image pairs with stochastic birth-death dynamics. *IEEE Trans. Pattern Anal. Mach. Intell.* **2011**, *34*, 33–50. [CrossRef]
9. Mishra, A.; Pandey, A.; Baghel, A.S. Building detection and extraction techniques: A review. In Proceedings of the 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, India, 16–18 March 2016; pp. 3816–3821.
10. Yu, Y.; Fu, L.; Cheng, Y.; Ye, Q. Multi-view distance metric learning via independent and shared feature subspace with applications to face and forest fire recognition, and remote sensing classification. *Knowl.-Based Syst.* **2022**, *243*, 108350. [CrossRef]
11. Jozdani, S.; Chen, D. On the versatility of popular and recently proposed supervised evaluation metrics for segmentation quality of remotely sensed images: An experimental case study of building extraction. *ISPRS J. Photogramm. Remote Sens.* **2020**, *160*, 275–290. [CrossRef]
12. Gong, M.; Liu, T.; Zhang, M.; Zhang, Q.; Lu, D.; Zheng, H.; Jiang, F. Context-content collaborative network for building extraction from high-resolution imagery. *Knowl.-Based Syst.* **2023**, *263*, 110283. [CrossRef]

13. Grinias, I.; Panagiotakis, C.; Tziritas, G. MRF-based segmentation and unsupervised classification for building and road detection in peri-urban areas of high-resolution satellite images. *ISPRS J. Photogramm. Remote Sens.* **2016**, *122*, 145–166. [CrossRef]
14. Chen, K.; Zou, Z.; Shi, Z. Building extraction from remote sensing images with sparse token transformers. *Remote Sens.* **2021**, *13*, 4441. [CrossRef]
15. Luo, L.; Li, P.; Yan, X. Deep learning-based building extraction from remote sensing images: A comprehensive review. *Energies* **2021**, *14*, 7982. [CrossRef]
16. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: Proceedings of the 13th European Conference, Zurich, Switzerland, 6–12 September 2014*; Proceedings, Part V 13; Springer: Berlin/Heidelberg, Germany, 2014; pp. 740–755.
17. Cheng, M.M.; Hou, Q.B.; Zhang, S.H.; Rosin, P.L. Intelligent visual media processing: When graphics meets vision. *J. Comput. Sci. Technol.* **2017**, *32*, 110–121. [CrossRef]
18. Cheng, M.M.; Zhang, F.L.; Mitra, N.J.; Huang, X.; Hu, S.M. Repfinder: Finding approximately repeated scene elements for image editing. *ACM Trans. Graph. TOG* **2010**, *29*, 83.
19. Lin, Z.; Zhang, Z.; Chen, L.Z.; Cheng, M.M.; Lu, S.P. Interactive image segmentation with first click attention. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 13339–13348.
20. Sofiiuk, K.; Petrov, I.; Barinova, O.; Konushin, A. f-brs: Rethinking backpropagating refinement for interactive segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 8623–8632.
21. Dupont, C.; Ouakrim, Y.; Pham, Q.C. UCP-net: Unstructured contour points for instance segmentation. In Proceedings of the 2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Melbourne, Australia, 17–20 October 2021; pp. 3373–3379.
22. Wang, G.; Zuluaga, M.A.; Li, W.; Pratt, R.; Patel, P.A.; Aertsen, M.; Doel, T.; David, A.L.; Deprest, J.; Ourselin, S.; et al. DeepIGeoS: A deep interactive geodesic framework for medical image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *41*, 1559–1572. [CrossRef]
23. Reichstein, M.; Camps-Valls, G.; Stevens, B.; Jung, M.; Denzler, J.; Carvalhais, N. Deep learning and process understanding for data-driven Earth system science. *Nature* **2019**, *566*, 195–204. [CrossRef]
24. Li, Z.; Zhang, X.; Xiao, P.; Zheng, Z. On the effectiveness of weakly supervised semantic segmentation for building extraction from high-resolution remote sensing imagery. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 3266–3281. [CrossRef]
25. Rother, C.; Kolmogorov, V.; Blake, A. "GrabCut" interactive foreground extraction using iterated graph cuts. *ACM Trans. Graph. TOG* **2004**, *23*, 309–314. [CrossRef]
26. Cheng, M.M.; Prisacariu, V.A.; Zheng, S.; Torr, P.H.; Rother, C. Densecut: Densely connected crfs for realtime grabcut. In *Computer Graphics Forum*; Wiley Online Library: Hoboken, NJ, USA, 2015; Volume 34, pp. 193–201.
27. Wu, J.; Zhao, Y.; Zhu, J.Y.; Luo, S.; Tu, Z. Milcut: A sweeping line multiple instance learning paradigm for interactive image segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 256–263.
28. Maninis, K.K.; Caelles, S.; Pont-Tuset, J.; Van Gool, L. Deep extreme cut: From extreme points to object segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 616–625.
29. Papadopoulos, D.P.; Uijlings, J.R.; Keller, F.; Ferrari, V. Extreme clicking for efficient object annotation. In Proceedings of the IEEE international conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 4930–4939.
30. Bai, J.; Wu, X. Error-tolerant scribbles based interactive image segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 392–399.
31. Freedman, D.; Zhang, T. Interactive graph cut based segmentation with shape priors. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 20–25 June 2005; Volume 1, pp. 755–762.
32. Gulshan, V.; Rother, C.; Criminisi, A.; Blake, A.; Zisserman, A. Geodesic star convexity for interactive image segmentation. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010; pp. 3129–3136.
33. Xu, N.; Price, B.; Cohen, S.; Yang, J.; Huang, T.S. Deep interactive object selection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 373–381.
34. Li, Z.; Chen, Q.; Koltun, V. Interactive image segmentation with latent diversity. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 577–585.
35. Jang, W.D.; Kim, C.S. Interactive image segmentation via backpropagating refinement scheme. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 5297–5306.
36. Forte, M.; Price, B.; Cohen, S.; Xu, N.; Pitié, F. Interactive training and architecture for deep object selection. In Proceedings of the 2020 IEEE International Conference on Multimedia and Expo (ICME), London, UK, 6–10 July 2020; pp. 1–6.
37. Hao, Y.; Liu, Y.; Wu, Z.; Han, L.; Chen, Y.; Chen, G.; Chu, L.; Tang, S.; Yu, Z.; Chen, Z.; et al. Edgeflow: Achieving practical interactive segmentation with edge-guided flow. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 1551–1560.

38. Zhao, F.; Xie, X. An overview of interactive medical image segmentation. *Ann. BMVA* **2013**, *2013*, 1–22.
39. Mortensen, E.N.; Barrett, W.A. Interactive segmentation with intelligent scissors. *Graph. Model. Image Process.* **1998**, *60*, 349–384. [CrossRef]
40. Cremers, D.; Rousson, M.; Deriche, R. A review of statistical approaches to level set segmentation: Integrating color, texture, motion and shape. *Int. J. Comput. Vis.* **2007**, *72*, 195–215. [CrossRef]
41. Chan, T.F.; Vese, L.A. Active contours without edges. *IEEE Trans. Image Process.* **2001**, *10*, 266–277. [CrossRef] [PubMed]
42. Kass, M.; Witkin, A.; Terzopoulos, D. Snakes: Active contour models. *Int. J. Comput. Vis.* **1988**, *1*, 321–331. [CrossRef]
43. Boykov, Y.Y.; Jolly, M.P. Interactive graph cuts for optimal boundary & region segmentation of objects in ND images. In Proceedings of the Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001, Vancouver, BC, Canada, 7–14 July 2001; Volume 1, pp. 105–112.
44. Grady, L. Random walks for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2006**, *28*, 1768–1783. [CrossRef]
45. Adams, R.; Bischof, L. Seeded region growing. *IEEE Trans. Pattern Anal. Mach. Intell.* **1994**, *16*, 641–647. [CrossRef]
46. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. In Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, 7–9 May 2015; Bengio, Y., Yann, L., Eds.; Conference Track Proceeding; Conference Publishing Services (CPS): Los Angeles, CA, USA, 2015.
47. Sofiiuk, K.; Petrov, I.A.; Konushin, A. Reviving iterative training with mask guidance for interactive segmentation. In Proceedings of the 2022 IEEE International Conference on Image Processing (ICIP), Anchorage, AK, USA, 19–22 September 2022; pp. 3141–3145.
48. Chen, X.; Zhao, Z.; Zhang, Y.; Duan, M.; Qi, D.; Zhao, H. FocalClick: Towards practical interactive image segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 1300–1309.
49. Yang, L.; Zi, W.; Chen, H.; Peng, S. DRE-Net: A Dynamic Radius-Encoding Neural Network with an Incremental Training Strategy for Interactive Segmentation of Remote Sensing Images. *Remote Sens.* **2023**, *15*, 801. [CrossRef]
50. Ji, S.; Wei, S.; Lu, M. Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set. *IEEE Trans. Geosci. Remote Sens.* **2018**, *57*, 574–586. [CrossRef]
51. Maggiori, E.; Tarabalka, Y.; Charpiat, G.; Alliez, P. Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark. In Proceedings of the 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Fort Worth, TX, USA, 23–28 July 2017; pp. 3226–3229.
52. Sofiiuk, K.; Barinova, O.; Konushin, A. Adaptis: Adaptive instance selection network. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 7355–7363.
53. Wang, J.; Sun, K.; Cheng, T.; Jiang, B.; Deng, C.; Zhao, Y.; Liu, D.; Mu, Y.; Tan, M.; Wang, X.; et al. Deep high-resolution representation learning for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *43*, 3349–3364. [CrossRef]
54. Yuan, Y.; Chen, X.; Wang, J. Object-contextual representations for semantic segmentation. In *Computer Vision–ECCV 2020: Proceedings of the 16th European Conference, Glasgow, UK, 23–28 August 2020*; Proceedings, Part VI 16; Springer: Berlin/Heidelberg, Germany, 2020; pp. 173–190.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.