CrossCut: Cross-Patch Aware Interactive Segmentation for Remote Sensing Images

Zheng Lin[1], **Nan Zhou**[1*], **Yuhan Wang**[1], **Bojian Zhang**[2]
[1]BNRist, Department of Computer Science and Technology, Tsinghua University
[2]VCIP, College of Computer Science, Nankai University
frazer.linzheng@gmail.com, zhoun1468@gmail.com,
yuhanwang200309@gmail.com, zbjyouxiang2020@163.com

Abstract

Interactive segmentation aims to delineate a user-specified target in an image by leveraging positive and negative clicks. While effective on natural images, existing methods often fail in remote sensing scenarios, where satellite imagery is characterized by ultra-high resolution, sparse object distribution, and significant scale variation. These factors hinder accurate segmentation of fine-grained targets like roads, buildings, and aircraft. To overcome these problems, we propose CrossCut, a novel interactive segmentation framework tailored for remote sensing imagery. Unlike previous approaches that either process the entire image or treat each patch independently, CrossCut enables simultaneous segmentation across multiple patches by propagating user click information to all patches. This design allows the model to fully utilize click guidance regardless of object location, effectively resolving the challenge of inter-patch information isolation. Furthermore, CrossCut supports flexible inference by allowing segmentation results from different patch configurations to be fused, enhancing both accuracy and robustness. Extensive evaluations across multiple remote sensing datasets demonstrate that CrossCut achieves state-of-the-art performance. Quantitative results and visualizations show that CrossCut significantly advances the field of interactive segmentation for remote sensing imagery.

Code — https://github.com/nanzhou02/CrossCut

Introduction

Interactive segmentation (IS) (Ramadan, Lachqar, and Tairi 2020) aims to extract a user-specified object from an image by incorporating sparse user guidance, typically in the form of clicks or scribbles. It serves as a powerful tool across various real-world domains by enabling efficient and precise object delineation with minimal user input. In natural image scenarios, IS is widely used for tasks such as photo editing, content generation, and dataset annotation (Rombach et al. 2022; Zhang, Rao, and Agrawala 2023; Chen et al. 2024). In medical imaging, it assists clinicians in outlining anatomical structures or lesions with high precision for diagnostic and treatment purposes (Wang et al. 2018; Marinov et al. 2024). In industrial settings, IS helps segment objects in quality inspection, defect detection, and robotic manipulation from factory or product imagery (Du et al. 2022; Hao et al. 2022). These domain-specific applications demonstrate the versatility and importance of interactive segmentation in handling varied visual data with high annotation efficiency.

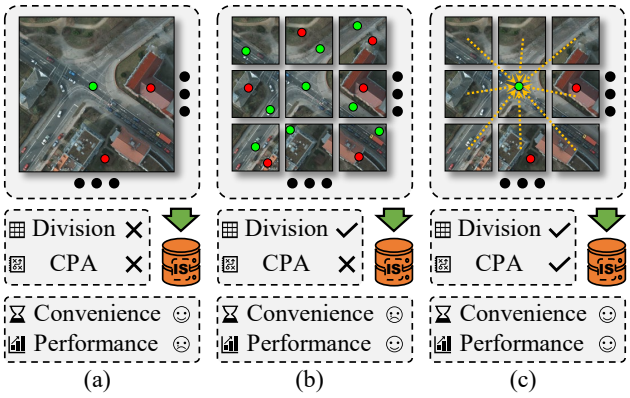


Figure 1: Comparison of IS approaches on remote sensing images. (a) Whole-image segmentation is simple but inaccurate. (b) Patch-by-patch segmentation is precise but inefficient. (c) Our method enables accurate and efficient segmentation via cross-patch aware paradigm. Green and red points denote positive and negative points, respectively.

However, when applied to remote sensing imagery, IS faces significant challenges. While deep learning-based methods have achieved impressive progress in remote sensing image segmentation (Maggiori et al. 2017), IS in this domain remains underexplored. (Liu et al. 2025) recently introduced a dedicated benchmark for remote sensing IS and highlighted the domain-specific challenges such as ultra-high resolution and sparse object distribution. This further motivates the need for specialized solutions beyond conventional IS frameworks. Satellite images are often characterized by ultra-high spatial resolution, complex backgrounds, sparse object distributions, and drastic scale variations. These challenges have been widely recognized in remote sensing literature (Cheng, Han, and Lu 2017; Yuan et al. 2021; Zhu et al. 2017; Audebert, Saux, and Lefèvre 2018), where high intra-class variation and small object sizes often degrade segmentation performance. Moreover,

*Internship at Tsinghua University.

conventional segmentation methods struggle to generalize across diverse remote sensing scenes due to their complex spatial structures and multi-scale characteristics. These factors severely hinder the ability of existing IS models to distinguish fine-grained and small-scale targets such as roads, buildings, and aircraft. Moreover, the large image size and scene diversity introduce additional difficulties in efficiently propagating user guidance across the entire image.

To overcome these limitations, we propose CrossCut, a novel IS framework specifically designed for remote sensing imagery. Unlike traditional methods that treat the image as a whole, CrossCut divides the input image into multiple patches and performs segmentation in a distributed yet collaborative manner. Interactive clicks are encoded and propagated across patches to enable simultaneous segmentation of all subregions. This design not only allows efficient handling of high-content data but also improves the localization of sparse and multi-scale objects.

As illustrated in Figure 1, directly applying interactive segmentation on the entire high-content image is simple but often leads to suboptimal results due to insufficient detail modeling and spatial constraints. On the other hand, manually segmenting each patch individually yields better performance but is extremely time-consuming and impractical for real-world applications. Our method strikes a balance by supporting simultaneous segmentation across all patches while maintaining high accuracy, effectively overcoming the limitations of both global and patch-wise approaches. This enables effective and efficient interaction across large-scale remote sensing images.

Furthermore, CrossCut supports flexible inference by enabling users to select different numbers of patches according to their annotation needs. The final segmentation result can be further refined by fusing outputs from multiple patch configurations, effectively leveraging complementary information from varying granularities.

Extensive experiments conducted on three challenging remote sensing segmentation datasets demonstrate that CrossCut achieves state-of-the-art performance. Both quantitative metrics and qualitative visualizations validate its effectiveness and robustness in addressing the unique difficulties of remote sensing imagery. Our approach significantly advances the field of interactive segmentation for large-scale, high-content satellite images.

Our main contributions are summarized as follows:

- We propose CrossCut, an interactive segmentation framework that propagates user-provided guidance across all patches, enabling effective global context modeling within high-content remote sensing imagery.
- Our method supports flexible inference with different patch division strategies during testing. By fusing results from multiple scales, CrossCut effectively captures complementary information across varying granularities.
- Our method achieves state-of-the-art performance on three challenging datasets, significantly surpassing existing interactive segmentation methods.

Related Work

Interactive Segmentation

Interactive image segmentation has seen significant progress over the past decade. Early methods such as GraphCut (Boykov and Funka-Lea 2006; Blake et al. 2004; Rother, Kolmogorov, and Blake 2004; Vicente, Kolmogorov, and Rother 2008; Veksler 2008), Random Walks (Grady 2006; Kim, Lee, and Lee 2008), and Geodesic Distance-based approaches (Gulshan et al. 2010; Bai and Sapiro 2009; Price, Morse, and Cohen 2010) leveraged user strokes or clicks to guide graph-based optimization processes.

With the rise of deep learning, methods like RITM (Sofiiuk, Petrov, and Konushin 2022), FocalClick (Chen et al. 2022), SimpleClick (Liu et al. 2023), and MFP (Lee, Lee, and Kim 2024) have enabled highly accurate segmentation by processing positive and negative clicks through neural networks, achieving strong results on natural images.

Recently, SAM (Kirillov et al. 2023) introduced a foundation model trained on large-scale segmentation data, enabling powerful zero-shot transfer capabilities. These advancements reflect a growing trend toward integrating user interaction, large-scale pretraining, and transformer architectures to develop generalizable and efficient interactive segmentation frameworks.

Remote Sensing Image Segmentation

Remote sensing image segmentation is a core task in geospatial analysis. Traditional semantic segmentation models such as FCNs (Long, Shelhamer, and Darrell 2015), U-Net (Ronneberger, Fischer, and Brox 2015), and DeepLab (Chen et al. 2018) have been widely applied in this domain. However, these methods heavily rely on large-scale pixel-level annotations, which are costly and time-consuming to acquire, especially in ultra-high-resolution satellite imagery.

To reduce annotation burdens, recent efforts have explored weakly-supervised, semi-supervised, and interactive segmentation methods. Nevertheless, remote sensing scenarios introduce unique challenges such as sparse object distribution, large-scale variations, and complex backgrounds, which limit the effectiveness and generalization of conventional interactive techniques.

In response, recent research has begun to tailor interactive segmentation approaches specifically for remote sensing. For example, ROS-SAM (Shan et al. 2025) enhances mask quality for moving objects using LoRA fine-tuning and boundary refinement. AerOSeg (Dutta et al. 2025) incorporates SAM into an open-vocabulary framework with multi-scale attention, achieving strong performance on unseen categories. In addition, RefPrompt (Wang et al. 2024) proposes a reference-guided prompting mechanism that transfers knowledge across images for better segmentation.

Furthermore, broader trends in remote sensing segmentation are summarized in recent surveys (Liu et al. 2025; Bao et al. 2025). Liu et al. discuss the evolution from pixel-level methods to foundation model-driven approaches, while Bao et al. highlight the scalability of Mamba-based architectures for modeling global context in high-resolution imagery.

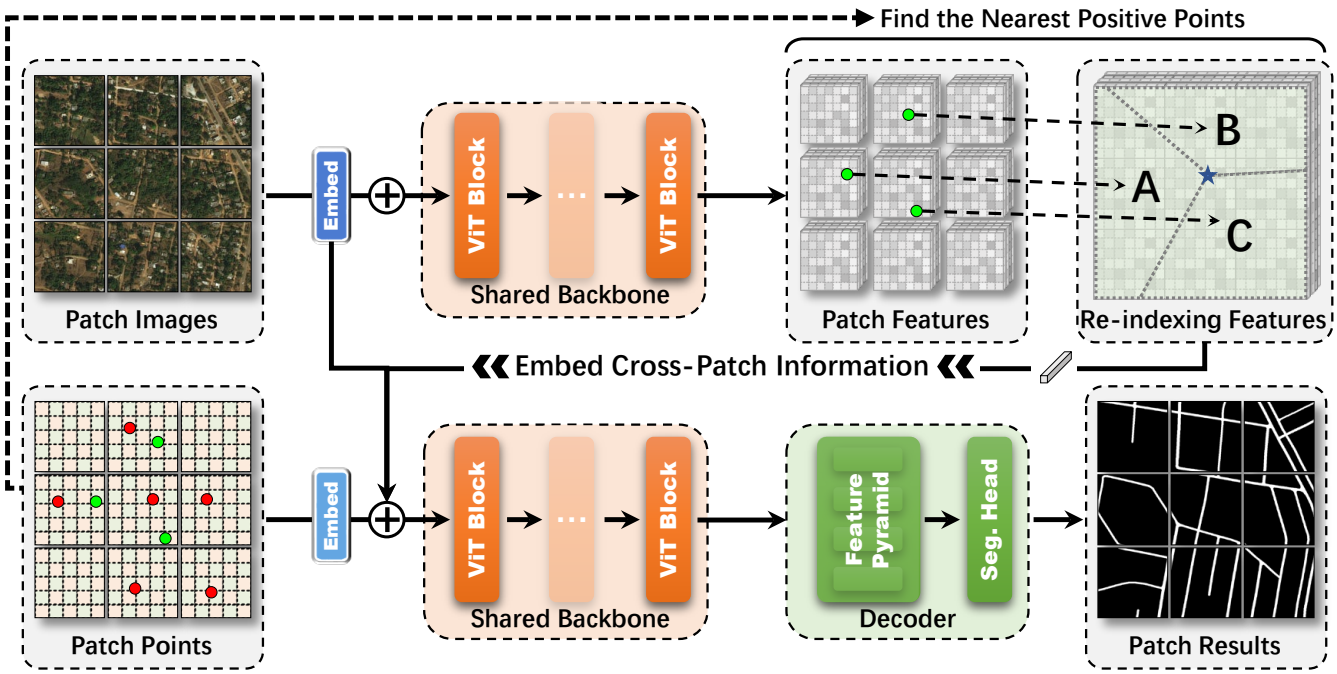


Figure 2: The framework of CrossCut. The lower branch follows a standard interactive segmentation pipeline over image patches. The upper branch generates the Cross-Patch Prompt Embedding, which encodes user clicks and propagates them across all patches, enabling synchronized and context-aware segmentation. The upper branch's Cross-Patch Prompt Embedding guides the lower branch, enabling coherent segmentation across all patches. Final output fuses results from different patch counts.

Proposed Method

Overall Framework

As illustrated in Figure 2, our proposed CrossCut framework builds upon the widely adopted SimpleClick architecture (Liu et al. 2023), which serves as the foundation for numerous recent IS models. Our framework is structured around two main branches, the Target Branch and the Cross-Patch Branch, each serving a distinct role, as elaborated in the following sections.

Target Branch. The target image $\mathbf{I}_t$ is first divided into $n \times n$ patches ($\mathbf{I}_1$, $\mathbf{I}_2$, ..., $\mathbf{I}_{n\times n}$), then they pass through a patch embedding module $\mathtt{E}_{img}$ to obtain image features. Meanwhile, positive and negative clicks are converted into two disk-shaped maps, which are then concatenated with the previous prediction mask to form the auxiliary input $\mathbf{C}_t$. This auxiliary input is also divided into $n \times n$ patches ($\mathbf{C}_1$,$\mathbf{C}_2$, ..., $\mathbf{C}_{n\times n}$), then they are embedded by an extra patch embedding module $\mathtt{E}_{ext}$:

$$\mathbf{F}_i = \mathtt{E}_{img}\left(\mathbf{I}_i\right) + \mathtt{E}_{ext}\left(\mathbf{C}_i\right), \quad i \in \{1, \ldots, n^2\}. \tag{1}$$

The backbone is a Vision Transformer (ViT) (Dosovitskiy 2020), and the decoder consists of a simple feature pyramid followed by a segmentation head, as described in SimpleClick. During training, the predicted segmentation $\mathbf{S}_t$ is supervised using the ground truth mask via a loss function, which is used to optimize the model parameters.

Cross-Patch Branch. The Cross-Patch Branch enhances the segmentation framework by introducing global semantic guidance derived from propagated positive clicks. While the detailed generation of the Cross-Patch Prompt Embedding $\mathbf{P}_c$ is described in the next section, we focus here on how it integrates into the overall model.

Specifically, we modify the feature fusion step. The final fused representation $\mathbf{F}_i$ for each patch is formed by integrating the baseline features with our proposed prompt embedding. This prompt $\mathbf{P}$ provides a specific embedding $\mathbf{P}_i$ tailored for the i-th patch.

These three components are integrated through element-wise addition to form the final fused representation:

$$\mathbf{F}_i = \mathtt{E}_{img}\left(\mathbf{I}_i\right) + \mathtt{E}_{ext}\left(\mathbf{C}_i\right) + \mathbf{P}_i, \quad i \in \{1, \ldots, n^2\}. \tag{2}$$

The fused feature $\mathbf{F}_i$ is then passed through the backbone network for deep feature extraction and subsequently decoded to produce the patch segmentation $\mathbf{S}_i$:

$$\mathbf{S}_i = \mathtt{Decoder}\left(\mathtt{Backbone}\left(\mathbf{F}_i\right)\right), \quad i \in \{1, \ldots, n^2\}. \tag{3}$$

Subsequently, these patch segmentations are concatenated in accordance with the original spatial layout of the patches to produce the segmentation result $\mathbf{S}_t^{(n)}$.

Cross-Patch Prompt Embedding

Unlike existing remote sensing interactive segmentation methods that operate on entire images or isolated patches,

our approach targets the overlooked challenge of achieving consistent segmentation across multiple patches simultaneously. To this end, we propose the generator of Cross-Patch Prompt Embedding (CPE), which extracts semantics around positive clicks and propagates this guidance across all patches, enabling coherent and synchronized interaction-aware segmentation over the full image.

The generation of the CPE begins with the features from the patch embedding layer. The input image is first divided into patches, which are encoded by the same patch embedding module $\mathtt{E}_{img}$ used in the main framework (Equation (2)) to obtain patch features $\{\mathtt{E}_{img}(\mathbf{I}_i)\}_{i=1}^{n^2}$. These patch features are then concatenated according to their original spatial order to reconstruct the full-resolution patch embedding feature map $\mathbf{F}_{img}$. We then resize it to match the size of the binary disk-shaped maps.

Given N positive clicks with coordinates $\{\mathbf{p}_j = (x_j, y_j)\}_{j=1}^{N}$, we generate N binary disk-shaped masks $\{\mathbf{d}_j\}_{j=1}^{N}$, where each $\mathbf{d}_j \in \{0,1\}^{H\times W}$ is a circular mask centered at $\mathbf{p}_j$. The use of disk-shaped masks allows soft spatial selection, making the feature extraction process more robust to slight misclicks and spatial ambiguities.

For each positive click $\mathbf{p}_i$, we compute the localized feature vector $\mathbf{f}_i \in \mathbb{R}^C$ by performing average pooling over the activated disk region:

$$\mathbf{f}_j = \frac{\sum_{x,y} \mathbf{d}_j(x,y) \cdot \mathbf{F}_{img}[:,x,y]}{\sum_{x,y} \mathbf{d}_j(x,y)}, \quad j \in \{1,\dots,N\}. \tag{4}$$

We then construct a nearest-click index map $\mathbf{M} \in \{1,\dots,N\}^{H\times W}$, which assigns each pixel to its closest positive click based on Euclidean distance:

$$\mathbf{M}(x,y) = \mathtt{argmin}_j \left\|(x,y) - \mathbf{p}_j\right\|_2, j \in \{1,\dots,N\}. \tag{5}$$

The resulting $\mathbf{M}$ partitions the image into Voronoi-like regions, allowing us to construct a re-indexing features map $\mathbf{F_r} \in \mathbb{R}^{C\times H\times W}$ by assigning each pixel the feature vector of its nearest click:

$$\mathbf{F}_r[:,x,y] = \mathbf{f}_{\mathbf{M}(x,y)}, \quad \forall (x,y) \in [1,H]\times[1,W]. \tag{6}$$

Finally, we encode $\mathbf{F_r}$ with a multi-layer perceptron (MLP) to produce the Cross-Patch Prompt Embedding $\mathbf{P}$:

$$\mathbf{P} = \mathtt{MLP}\left(\mathbf{F_r}\right). \tag{7}$$

This global prompt $\mathbf{P}$ is then spatially divided into an $n \times n$ grid of patch-level prompts, $\{\mathbf{P}_i\}_{i=1}^{n^2}$. Each $\mathbf{P}_i$ is subsequently injected into its corresponding i-th patch as described in Equation (2), enabling precise guidance across the entire image.

Multi-Scale Division Fusion

Since our method propagates prompts through cross-image information, it can be directly applied to inputs with varying numbers of patches. This allows us to generate multiple segmentation results for a target image in different slice configurations. By merging these results, we can obtain a more accurate and robust final segmentation $\mathbf{S}_t$:

$$\mathbf{S}_t = \sum_{k=1}^{K} w_k \cdot \mathbf{S}_t^{(k)}, \tag{8}$$

where $\mathbf{S}_t^{(k)}$ is the result from the k-th patch configuration, w_k is the corresponding fusion weight, and the weights satisfy $\sum_{k=1}^{K} w_k = 1$ and $w_k \geq 0$.

To mitigate the limitations of individual slicing strategies (e.g., boundary misalignment or missing contextual information), we apply multiple dividing configurations (e.g., 2×2, 3×3, 4×4 grids) to the same image. Each configuration captures the object at different resolutions and contextual ranges, which complements each other when fused.

Inspired by ensemble learning, Multi-Scale Division Fusion (MDF) enhances robustness by aggregating diverse segmentation hypotheses. In practice, we assign higher weights to configurations that cover larger context or demonstrate better confidence. The fusion weights w_k can be determined heuristically, uniformly, or based on validation performance. In this paper, we set $w_2 = w_3 = w_4 = \frac{1}{3}$.

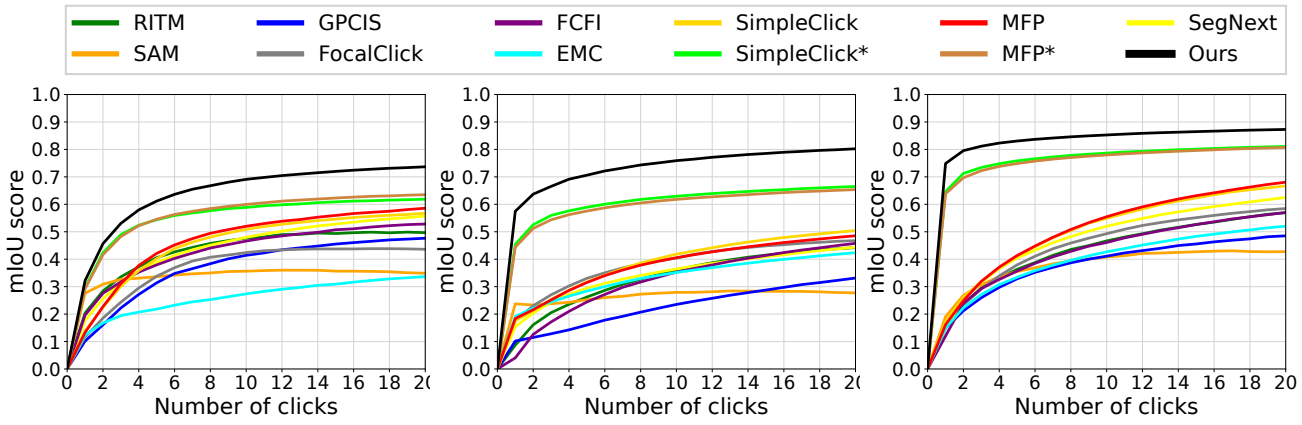
Experiment

Experiments Setting

Dataset. To comprehensively evaluate our method's performance and generalization ability, we conduct experiments on three representative remote sensing datasets: DeepGlobe Road Extraction Dataset (DeepGlobe) (Demir et al. 2018), iSAID (Waqas Zamir et al. 2019; Xia et al. 2018), and Inria Aerial Image Labeling Dataset (Inria) (Maggiori et al. 2017).

- `DeepGlobe`: It is a high-resolution satellite image dataset from the CVPR 2018 DeepGlobe Challenge, used here for road segmentation. It contains 6266 RGB images (1024×1024, 0.5 m/pixel). Since only training set masks are available, we use 4358 for training and 1868 for testing. Roads are annotated as thin binary masks.
- `iSAID`: It is a large-scale aerial image dataset for instance and semantic segmentation, derived from DOTA. It includes 1411 images for training and 458 images for validation (spatial resolution: 0.3–3 m/pixel), we use validation images for testing. It covers 16 object categories with dense, multi-oriented semantic masks.
- `Inria`: It is a building segmentation dataset with 180 high-resolution RGB images (5000×5000, 0.3 m/pixel), covering urban areas in the US and Europe. Since only training masks are provided, we split the data manually for training and testing. To facilitate training and evaluation, we divide each original image into non-overlapping 1000 × 1000 pixel patches. Models are trained and tested on these patches, 2926 for training and 1246 for testing.

Evaluation Metrics. To quantitatively evaluate the performance of our method, we employ two commonly used metrics: Intersection over Union (IoU) and Number of Clicks (NoC). IoU measures the overlap between the predicted mask and the ground truth, and is defined as the ratio of the intersection area to the union area. A higher IoU indicates more accurate segmentation performance. For interactive segmentation tasks, we further use NoC to assess the efficiency of user interactions. Specifically, NoC represents the number of clicks required to reach a predefined IoU threshold. In our experiments, we report NoC (up to 20



(a) iSAID (b) DeepGlobe (c) Inria

Figure 3: The mIoU-NoC curves on three benchmark datasets.

Method	Backbone	iSAID			DeepGlobe			Inria		
		@70	@75	@80	@70	@75	@80	@70	@75	@80
RITM	HRNet-32	13.97	14.52	15.32	18.55	19.00	19.39	16.28	17.07	17.83
FocalClick	SegF-B3	15.04	15.47	15.98	17.96	18.46	18.91	16.31	17.11	17.88
EMC-Click	HRNet-32	17.09	17.44	17.81	18.71	19.11	19.43	17.28	18.00	18.57
GPCIS	ResNet-50	15.53	16.07	16.61	19.57	19.70	19.81	17.27	17.86	18.44
FCFI	HRNet-18	14.51	15.06	15.62	18.69	19.15	19.49	16.40	17.15	17.84
SAM	ViT-B	15.63	16.24	16.87	18.99	19.43	19.73	18.05	18.70	19.23
SegNext	ViT-B	14.64	15.36	16.13	18.15	18.62	19.07	15.51	16.44	17.41
SimpleClick	ViT-B	14.08	14.71	15.39	17.60	18.12	18.69	15.24	16.39	17.35
MFP	ViT-B	13.77	14.47	15.20	17.69	18.24	18.74	14.64	15.82	16.85
SimpleClick*	ViT-B	11.71	12.75	13.91	13.12	15.04	16.62	4.88	7.90	11.61
MFP*	ViT-B	11.65	12.68	13.77	13.86	15.50	16.91	5.68	8.26	11.91
Ours	ViT-B	**9.19**	**10.21**	**11.63**	**7.11**	**9.71**	**12.60**	**2.14**	**2.97**	**4.68**

Table 1: Comparison of different IS methods on three benchmark datasets. The evaluation metric is NoC, lower NoC indicates higher click efficiency. The **bold** value represents the best result. * indicates using above three datasets for finetuning.

clicks) at IoU thresholds of 70%, 75%, and 80%. A lower NoC value at a given threshold implies that the model can achieve high-quality segmentation with fewer user interactions, indicating better guidance and interaction efficiency. Together, these metrics assess both segmentation accuracy and interactive performance.

Implementation Details. We build CrossCut based on the SimpleClick framework.In the experiments, we mainly use the ViT-B model trained on COCO (Lin et al. 2014)+LVIS (Gupta, Dollar, and Girshick 2019) datasets as our base model, and finetune it using the above three datasets.We conducted our training on four NVIDIA RTX 4090 GPUs. To reduce GPU memory consumption, each original large image was divided into 2×2 patches, and no further cropping was applied. During training, we employed a series of data augmentation strategies, including random scaling, random cropping, and horizontal flipping. The network was optimized using the normalized focal loss (Sofiiuk, Petrov, and Konushin 2022) to better handle class imbalance. We trained the model for 55 epochs on the three benchmark datasets using the Adam (Kingma and Ba 2015) optimizer with an initial learning rate of 5e-5, which was decayed by 0.1 after the 50th epoch to facilitate convergence.

Results & Analysis

Quantitative Results. To thoroughly evaluate the effectiveness of our method, we conduct experiments on three challenging remote sensing benchmarks: iSAID, DeepGlobe, and Inria. We compare our approach against a series of recent interactive segmentation baselines, including SAM (Kirillov et al. 2023), RITM (Sofiiuk, Petrov, and Konushin 2022), GPCIS (Zhou et al. 2023), FocalClick (Chen et al. 2022), EMC-Click (Du et al. 2023), FCFI (Wei, Zhang, and Yong 2023), SegNext (Liu et al. 2024), SimpleClick (Liu et al. 2023), and MFP (Lee, Lee, and Kim 2024). The evaluation is performed under various numbers of user clicks, and results are reported in terms of mIoU (%) as well as the number of clicks required to reach specified IoU thresholds (Figure 3 and Table 1). Notably,

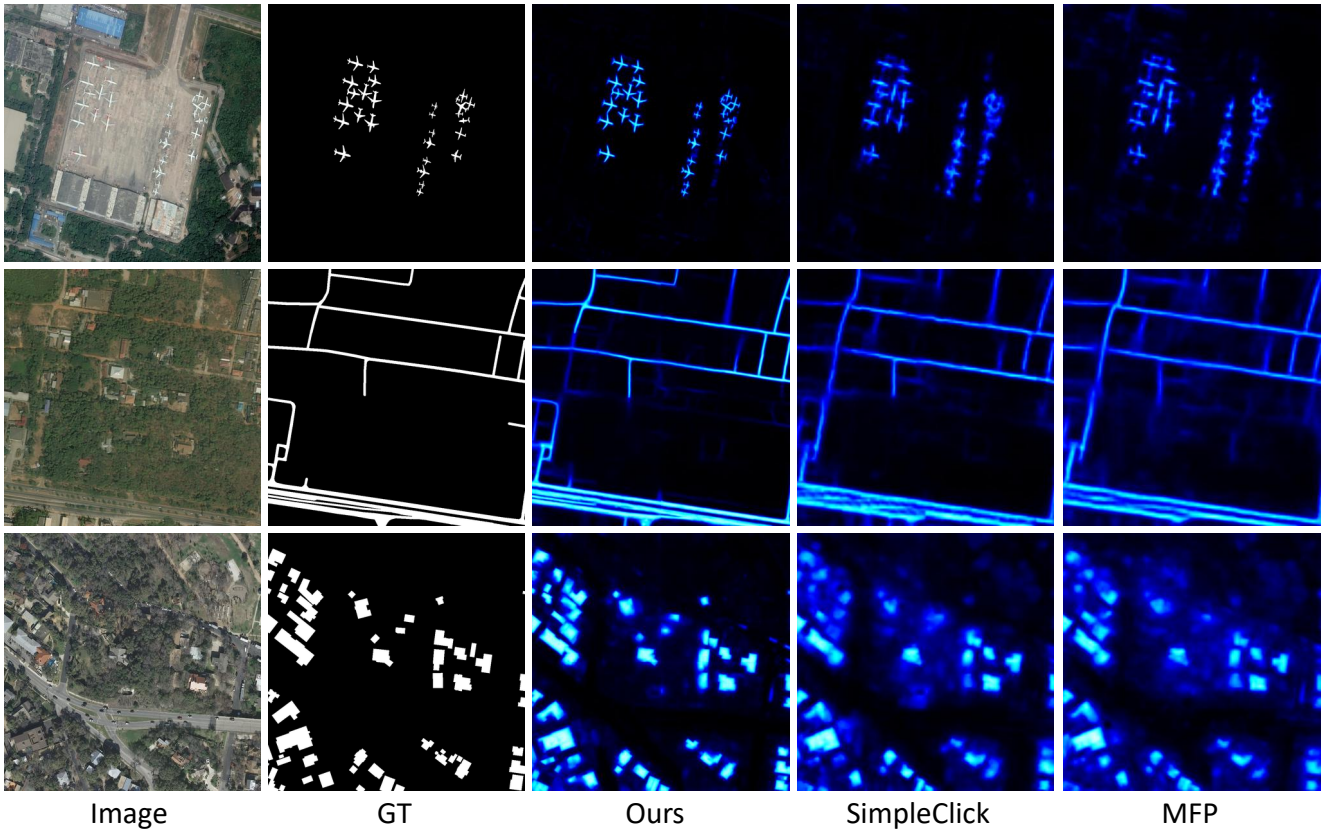


Figure 4: Comparison of visualized results on three datasets with 5 interaction points.

existing IS models often struggle to generalize well to remote sensing imagery, largely due to its distinct characteristics and complexity. To address this issue, we selected two representative methods for finetuning on the remote sensing datasets: the baseline method SimpleClick and the high-performing method MFP.

As shown in Figure 3, our proposed CrossCut consistently outperforms all existing interactive segmentation baselines across all datasets and click counts. This performance advantage is especially prominent in the low-click regime, indicating higher efficiency with minimal user input. The effectiveness is most evident on the Inria dataset with dense urban structures, where CrossCut exhibits a significantly sharper accuracy gain. On iSAID and DeepGlobe, it also consistently outperforms other methods across all click levels, demonstrating robustness to complex layouts, scale variation, and sparse targets. These results validate the effectiveness of our cross-patch guidance mechanism in handling challenging remote sensing segmentation tasks.

Furthermore, Table 1 reports the number of clicks required to reach 70%, 75%, and 80% IoU thresholds. Our method requires significantly fewer interactions to reach each threshold across all datasets. For example, to achieve 80% IoU on Inria, CrossCut needs only 4.68 clicks, whereas MFP and SimpleClick require 11.91 and 11.61 clicks, respectively. This reduction in user effort highlights the efficiency and effectiveness of our Cross-Patch Prompt Embedding and Multi-Scale Division Fusion strategies.

These quantitative results validate the robustness and generalization ability of CrossCut in diverse and complex remote sensing scenarios. By leveraging distributed patch-level segmentation with global interaction, our method demonstrates superior performance in both segmentation quality and interaction efficiency.

Visualization Results. As shown in Figure 4, we visualize the segmentation results on three representative images from different datasets, with segmentation targets being aircraft (top row), roads (middle row), and buildings in remote sensing scenes (bottom row). Each example presents the 5-point interaction results and includes five columns: the image, the ground truth (GT), the segmentation result produced by our method, SimpleClick, and MFP.

It can be clearly observed that our method yields more precise and sharper object boundaries across all three scenarios. For instance, in the aircraft case, our segmentation more accurately captures the contour of each airplane. In the road extraction example, the linear structures and intersections are better preserved, demonstrating stronger alignment with the GT. Similarly, for remote sensing buildings, our method results in more accurate and better-localized masks, reducing boundary blur and over-segmentation.

Method	iSAID			DeepGlobe			Inria		
	&5	&10	&20	&5	&10	&20	&5	&10	&20
Baseline	57.22	66.77	71.82	55.25	67.93	76.12	63.09	80.50	85.11
Baseline + CPE	60.11	67.60	72.17	69.72	74.93	79.17	82.60	84.85	86.84
Baseline + CPE + MDF	**61.17**	**69.10**	**73.67**	**70.61**	**75.91**	**80.23**	**83.06**	**85.26**	**87.28**

Table 2: Effectiveness of CPE and MDF on segmentation accuracy. Results are reported as mIoU on three benchmark datasets. The **bold** value represents the best result.

Method	iSAID			DeepGlobe			Inria		
	@70	@75	@80	@70	@75	@80	@70	@75	@80
Baseline	10.13	11.21	12.59	12.02	13.68	15.51	7.06	8.27	10.39
Baseline + CPE	9.61	10.69	12.16	7.68	10.30	13.30	2.17	3.08	5.00
Baseline + CPE + MDF	**9.19**	**10.21**	**11.63**	**7.11**	**9.71**	**12.60**	**2.14**	**2.97**	**4.68**

Table 3: Effectiveness of CPE and MDF on interaction efficiency. Results are reported as NoC on three benchmark datasets. The **bold** value represents the best result.

These visual comparisons further verify that our approach achieves not only higher numerical performance but also superior visual quality, especially in terms of boundary fidelity and structural consistency.

Ablation Study

Cross-Patch Prompt Embedding. The CPE module aims to address the challenge of guiding segmentation across multiple non-overlapping patches by effectively propagating user clicks. Compared with the baseline, adding CPE yields significant improvements across all datasets. On DeepGlobe, mIoU@5 improves from 55.25% to 69.72%, and NoC@80 drops from 15.51 to 13.30. A similar trend is observed on Inria and iSAID, where CPE enhances performance by enabling each patch to receive context-aware guidance based on localized click semantics. These results demonstrate that CPE is crucial for enabling accurate segmentation in our patch-based interaction framework.

Multi-Scale Division Fusion. Building on CPE, we further introduce MDF to integrate information from multiple patch granularities. This fusion allows the model to leverage both local details and global context. The addition of MDF consistently boosts performance: on iSAID, mIoU@20 increases from 72.17% to 73.67%, and NoC@80 decreases from 12.16 to 11.63. The improvements are more pronounced on DeepGlobe and Inria, showing that MDF effectively mitigates scale variance in remote sensing scenes. This confirms that fusing multi-scale division outputs provides complementary cues, enhancing robustness and segmentation precision.

Patch Configuration Analysis. To assess the impact of patch granularity and the effectiveness of MDF strategy, we evaluate different patch configurations on DeepGlobe dataset. As shown in Table 4, the 3×3 division yields the highest performance among the single-scale settings, achieving 79.17% mIoU@20. Building on this, fusing segmentation results from multiple patch scales through MDF consistently improves performance across all click levels. Specifically, MDF achieves mIoU scores of 70.61%, 75.91%, and 80.23% at 5, 10, and 20 clicks, respectively. These results highlight the advantage of leveraging complementary spatial information across scales to enhance segmentation accuracy and robustness.

Configuration	&5	&10	&20
2×2 patches	69.41	73.95	77.72
3×3 patches	69.72	74.93	79.17
4×4 patches	68.07	73.76	78.74
MDF	**70.61**	**75.91**	**80.23**

Table 4: Comparision of the impact of patch number and MDF. Results are reported as mIoU on DeepGlobe dataset. The **bold** value represents the best result.

Conclusion

In this paper, we present CrossCut, a novel interactive segmentation framework tailored for remote sensing imagery. Unlike existing methods that typically process image patches independently or apply segmentation on the entire image, CrossCut enables simultaneous segmentation across all patches by propagating click information globally through our proposed Cross-Patch Prompt Embedding module. This design effectively addresses the challenges of large-scale, sparse, and multi-scale objects inherent in high-content satellite images. Extensive experiments on three remote sensing datasets demonstrate that CrossCut consistently outperforms state-of-the-art methods in both segmentation accuracy and annotation efficiency. We believe our framework provides a new direction for scalable, effective interactive segmentation in the remote sensing domain.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (project No. 62495060), the Research Grant of Tsinghua-Tencent Joint Laboratory for Internet Innovation Technology, the China Postdoctoral Science Foundation(Grant No. 2025T180422, Grant No. 2024M761682, Grant No. GZB20240357) and Shui Mu Tsinghua Scholar (Grant No. 2024SM079).

References

Audebert, N.; Saux, B. L.; and Lefèvre, S. 2018. Beyond RGB: Very high resolution urban remote sensing with multimodal deep networks. *ISPRS P&RS*.

Bai, X.; and Sapiro, G. 2009. Geodesic matting: A framework for fast interactive image and video segmentation and matting. *IJCV*.

Bao, M.; Lyu, S.; Xu, Z.; Zhou, H.; et al. 2025. Vision Mamba in Remote Sensing: A Comprehensive Survey of Techniques, Applications and Outlook. *arXiv preprint arXiv:2505.00630*.

Blake, A.; Rother, C.; Brown, M.; Perez, P.; and Torr, P. 2004. Interactive image segmentation using an adaptive GMMRF model. In *ECCV*.

Boykov, Y.; and Funka-Lea, G. 2006. Graph cuts and efficient N-D image segmentation. *IJCV*.

Chen, L.-C.; Zhu, Y.; Papandreou, G.; Schroff, F.; and Adam, H. 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*.

Chen, X.; Huang, L.; Liu, Y.; Shen, Y.; Zhao, D.; and Zhao, H. 2024. Anydoor: Zero-shot object-level image customization. In *CVPR*.

Chen, X.; Zhao, Z.; Zhang, Y.; Duan, M.; Qi, D.; and Zhao, H. 2022. FocalClick: Towards practical interactive image segmentation. In *CVPR*.

Cheng, G.; Han, J.; and Lu, X. 2017. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*.

Demir, I.; Koperski, K.; Lindenbaum, D.; Pang, G.; Huang, J.; Basu, S.; Hughes, F.; Tuia, D.; and Raska, R. 2018. DeepGlobe 2018: A Challenge to Parse the Earth through Satellite Images. In *CVPR*.

Dosovitskiy, A. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Du, F.; Yuan, J.; Wang, Z.; and Wang, F. 2023. Efficient Mask Correction for Click-Based Interactive Image Segmentation. In *CVPR*.

Du, W.; Shen, H.; Ge, Z.; Yao, X.; and Fu, J. 2022. Interactive defect segmentation in X-Ray images based on deep learning. *ESWA*.

Dutta, S.; Vasim, A.; Gole, S.; Rezatofighi, H.; and Banerjee, B. 2025. AerOSeg: Harnessing SAM for Open-Vocabulary Segmentation in Remote Sensing Images. *arXiv preprint arXiv:2504.09203*.

Grady, L. 2006. Random walks for image segmentation. *IEEE TPAMI*.

Gulshan, V.; Rother, C.; Criminisi, A.; Blake, A.; and Zisserman, A. 2010. Geodesic star convexity for interactive image segmentation. In *CVPR*.

Gupta, A.; Dollar, P.; and Girshick, R. 2019. Lvis: A dataset for large vocabulary instance segmentation. In *CVPR*.

Hao, Y.; Liu, Y.; Chen, Y.; Han, L.; Peng, J.; Tang, S.; et al. 2022. EISeg: An Efficient Interactive SEGmentation Tool based on PaddlePaddle. arXiv preprint arXiv:2210.08788.

Kim, T. H.; Lee, K. M.; and Lee, S. U. 2008. Generative image segmentation using random walks with restart. In *ECCV*.

Kingma, D. P.; and Ba, J. 2015. Adam: A method for stochastic optimization. In *ICLR*.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. In *ICCV*.

Lee, C.; Lee, S.-H.; and Kim, C.-S. 2024. MFP: Making Full Use of Probability Maps for Interactive Image Segmentation. In *CVPR*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft COCO: Common objects in context. In *ECCV*.

Liu, Q.; Cho, J.; Bansal, M.; and Niethammer, M. 2024. Rethinking Interactive Image Segmentation with Low Latency High Quality and Diverse Prompts. In *CVPR*.

Liu, Q.; Xu, Z.; Bertasius, G.; and Niethammer, M. 2023. Simpleclick: Interactive image segmentation with simple vision transformers. In *ICCV*.

Liu, T.; Zhang, Y.; Chen, W.; Li, X.; and Wang, Z. 2025. From Pixels to Images: Deep Learning Advances in Remote Sensing Image Semantic Segmentation. *arXiv preprint arXiv:2505.15147*.

Long, J.; Shelhamer, E.; and Darrell, T. 2015. Fully convolutional networks for semantic segmentation. In *CVPR*.

Maggiori, E.; Tarabalka, Y.; Charpiat, G.; and Alliez, P. 2017. Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark. In *IGARSS*.

Marinov, Z.; Jäger, P. F.; Egger, J.; Kleesiek, J.; and Stiefelhagen, R. 2024. Deep interactive segmentation of medical images: A systematic review and taxonomy. *IEEE TPAMI*.

Price, B. L.; Morse, B.; and Cohen, S. 2010. Geodesic graph cut for interactive image segmentation. In *CVPR*.

Ramadan, H.; Lachqar, C.; and Tairi, H. 2020. A survey of recent interactive image segmentation methods. *CVMJ*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *CVPR*.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-Net: Convolutional networks for biomedical image segmentation. In *MICCAI*.

Rother, C.; Kolmogorov, V.; and Blake, A. 2004. GrabCut: Interactive foreground extraction using iterated graph cuts. *ACM TOG*.

Shan, Z.; Liu, Y.; Zhou, L.; Yan, C.; et al. 2025. ROS-SAM: High-Quality Interactive Segmentation for Remote Sensing Moving Object. In *arXiv preprint arXiv:2503.12006*.

Sofiiuk, K.; Petrov, I. A.; and Konushin, A. 2022. Reviving iterative training with mask guidance for interactive segmentation. In *ICIP*.

Veksler, O. 2008. Star shape prior for graph-cut image segmentation. In *ECCV*.

Vicente, S.; Kolmogorov, V.; and Rother, C. 2008. Graph cut based image segmentation with connectivity priors. In *CVPR*.

Wang, C.; Zhang, Y.; Zhang, H.; Zhao, L.; Li, Z.; Lu, H.; Zha, Z.-J.; and Li, H. 2024. RefPrompt: Reference-based Prompting for Click-based Interactive Segmentation. In *CVPR*.

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. 2018. DeepIGeoS: A deep interactive geodesic framework for medical image segmentation. *IEEE TPAMI*.

Waqas Zamir, S.; Arora, A.; Gupta, A.; Khan, S.; Sun, G.; Shahbaz Khan, F.; Zhu, F.; Shao, L.; Xia, G.-S.; and Bai, X. 2019. iSAID: A Large-scale Dataset for Instance Segmentation in Aerial Images. In *CVPR*.

Wei, Q.; Zhang, H.; and Yong, J.-H. 2023. Focused and collaborative feedback integration for interactive image segmentation. In *CVPR*.

Xia, G.-S.; Bai, X.; Ding, J.; Zhu, Z.; Belongie, S.; Luo, J.; Datcu, M.; Pelillo, M.; and Zhang, L. 2018. DOTA: A Large-Scale Dataset for Object Detection in Aerial Images. In *CVPR*.

Yuan, Q.; Shen, H.; Li, T.; and Zhang, L. 2021. A review of deep learning methods for semantic segmentation of remote sensing imagery. *ESWA*.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *ICCV*.

Zhou, M.; Wang, H.; Zhao, Q.; Li, Y.; Huang, Y.; Meng, D.; and Zheng, Y. 2023. Interactive Segmentation as Gaussion Process Classification. In *CVPR*.

Zhu, X. X.; Tuia, D.; Mou, L.; Xia, G.-S.; Zhang, L.; Xu, F.; and Fraundorfer, F. 2017. Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE GRSM*.