NTClick: Achieving Precise Interactive Segmentation With Noise-tolerant Clicks

Chenyi Zhang[1, 2, 3], Ting Liu[2], Xiaochao Qu[2], Luoqi Liu[2], Yao Zhao[1, 3, 4], Yunchao Wei[1, 3, 4]✉
[1] Institute of Information Science, Beijing Jiaotong University
[2] MT Lab, Meitu Inc.
[3] Visual Intelligence + X International Joint Laboratory of the Ministry of Education
[4] Pengcheng Laboratory, Shenzhen, China
{chenyi22, yzhao, yunchao.wei}@bjtu.edu.cn
{lt, qxc, llq5}@meitu.com

Abstract

Interactive segmentation is a pivotal task in computer vision, focused on predicting precise masks with minimal user input. Although the click has recently become the most prevalent form of interaction due to its flexibility and efficiency, its advantages diminish as the complexity and details of target objects increase because it's time-consuming and user-unfriendly to precisely locate and click on narrow, fine regions. To tackle this problem, we propose NTClick, a powerful click-based interactive segmentation method capable of predicting accurate masks even with imprecise user clicks when dealing with intricate targets. We first introduce a novel interaction form called noise-tolerant click, a type of click that does not require user's precise localization when selecting fine regions. Then, we design a two-stage workflow, consisting of an Explicit Coarse Perception network for initial estimation and a High Resolution Refinement network for final classification. Quantitative results across extensive datasets demonstrate that NTClick not only maintains an efficient and user-friendly interaction mode but also significantly outperforms existing methods in segmentation accuracy.

1. Introduction

Interactive segmentation is a crucial task in computer vision, aimed at predicting accurate masks of target objects with minimal and diverse forms of user inputs. This technique is widely used in various applications such as photo editing (e.g., composition, inpainting ,and color grading), medical image analysis and industrial inspection. More importantly, interactive segmentation is capable of efficiently building and expanding datasets and is playing an increasingly significant role in the pursuit of large vision models, as demonstrated by Segment Anything Model (SAM) [14]. As a tool for data annotation, the pursuit of annotation accuracy is endless. Therefore, recent interactive segmentation methods like SAM-HQ [12] and SegNext [21] use dichotomous segmentation datasets (e.g., DIS5K [26]) and thin object segmentation datasets (e.g., ThinObject5K [18]) as benchmarks because these datasets possess finer scenarios, higher resolution and superior annotation quality.

However, current interactive segmentation methods face two primary challenges when dealing with detailed targets: **1) Inefficient interactions.** Methods proposed recently usually prefer to adopt *click* as the form of interaction due to *click's* unparalleled flexibility. However, as the target objects or regions become more intricate (sometimes only a few pixels wide), the efficiency of users in accurately locating and clicking on narrow target regions decreases rapidly, whether using a computer mouse or a mobile device, as shown in Fig. 1 (a). Therefore, scribble-based methods, like Slim Scissors [6], was proposed. Scribble-based methods allow users to utilize coarse scribbles instead of precise clicks to segment thin objects, as shown in Fig. 1 (b). Moreover, while coarse scribbles do not require precise localizations, its efficiency is also limited when fine regions are dense, leading to an excessive amount of area that needs to be smeared. Thus, there is still a lack of an interaction that is truly efficient and user-friendly when dealing with fine objects. **2) Resolution limitations.** Interactive segmentation methods like SAM [14] and SegNext [21] adopt transformer architecture (e.g. ViT [5]) to build a sufficiently complex feature space to achieve exceptionally high segmentation accuracy and robust zero-shot capability. However, because the computational complexity of self-attention mechanism in ViT increases quadratically with image resolution, these methods must downsample high-resolution images to a fixed resolution (e.g., 1024^2) when processing them. This enforced downsampling inevitably causes significant loss of high-resolution details, limiting the model's ability to predict intricate regions accurately. Thus, even with adequate

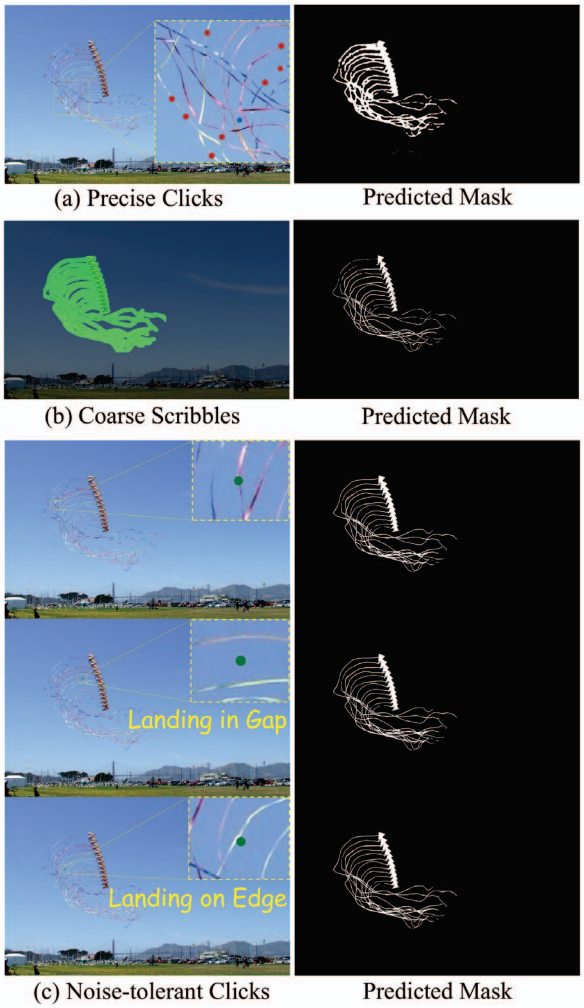

Figure 1. Interaction forms of existing click-based methods, scribble-based methods ,and NTClick. •, and • refer to *foreground click* and *background click*, • refers to *noise-tolerant click*. Zoom in to observe details.

interactions, current methods still fall short of generating high-quality masks, as illustrated in Fig. 1 (a) and (b).

Therefore, we propose NTClick in this paper, a powerful interactive segmentation method capable of performing fine-grained perception at high resolution based on imprecise user interactions. First, beyond the traditional *foreground click* and *background click*, we introduce a novel type of interaction click, called *noise-tolerant click*. Different from *foreground click* and *background click*, which explicitly indicate to the model whether the class of a specific region is foreground or background, *noise-tolerant click* informs the model that there are fine structures or objects **in the vicinity**. Thus, by definition, *noise-tolerant click* does not require the user to perform precise localization and is inherently tolerant to the noise of localization. Next, we propose a two-stage workflow to handle such possible imprecise and noisy user interactions. In the first stage, we utilize an Explicit Coarse Perception network to predict which regions in the image are solid foreground, solid background, and regions potentially containing fine structures or objects. This estimation is output as a three class mask, called **F**oreground **B**ackground **U**ncertain map (FBU map). At this stage, we do not predict the specific class of each pixel in the uncertain region, allowing this initial perception to be processed and output at a relatively low resolution. In the second stage, we upsample the FBU map to a high resolution and concatenate it with original RGB image, feeding them together into a **H**igh **R**esolution **R**efinement network (HRR). Guided by the FBU map, HRR finally classify all pixels in the uncertain region. By exploring the construction of long-range attention and short-range attention at high resolution, we achieve a balance between simplicity and performance in HRR.

Consequently, NTClick not only achieves higher segmentation accuracy but also offers better user-friendliness. As shown in Fig. 1 (c), when encountering difficult-to-locate regions, NTClick allows users to provide *noise-tolerant click* to segment the intricate target. Regardless of whether the noise-tolerant click lands accurately on the object, in the gap, or on the edge between the object and the background, NTClick consistently provides stable performance.

We conduct extensive experiments on highly accurate annotated datasets such as DIS5K [26] (including TE1-TE4), HRSOD [36] and ThinObject5K [36]. Under a more flexible and efficient interaction form, NTClick still demonstrates superior performance. On DIS5K-VD, NTClick achieves an improvement of **3.12** in NoC@90 and **3.54** in 5-mIoU. On the most challenging and complex test set, DIS5K-TE4, NTClick achieves a notable improvement of **5.46** in NoC@90 and **8.04** in 5-mIoU. We believe that NTClick has tremendous practical application value due to its exceptional accuracy and user-friendliness.

2. Related Works

2.1. Interactive Segmentation

Interactive segmentation is a fundamental task in computer vision as it serves as a tool for data annotation, providing sufficient datasets for a wide range of tasks. Therefore, even before the rise of deep learning, interactive segmentation methods, such as GrabCut [27] and Snakes [10], had already been proposed. In the subsequent period, interactive segmentation has experienced significant, continuous and diverse development with the advancement of deep learning. DIOS [31] was the first to apply deep convolutional neural networks to interactive segmentation. Click-based methods, represented by ITIS [22], RITM [29], In-

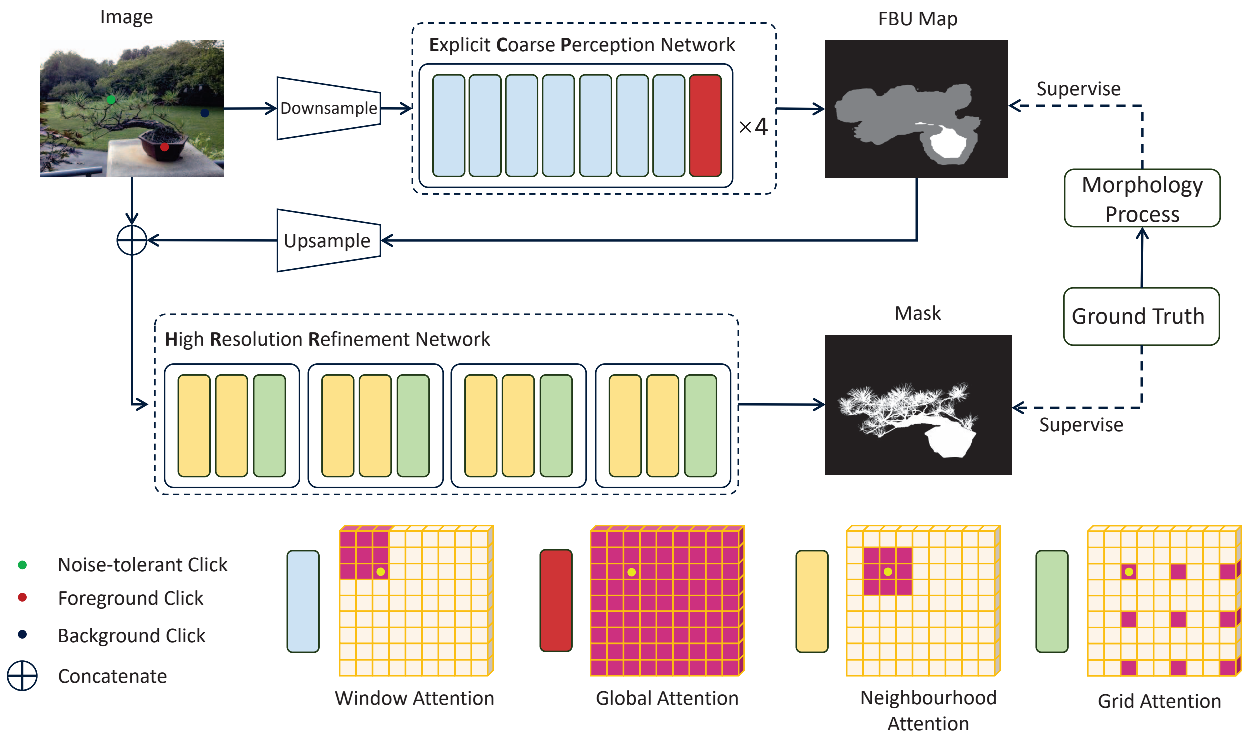

Figure 2. Overall architecture of the proposed NTClick. In the FBU map, black, white, and gray regions denote the foreground, background and uncertain regions, respectively.

terformer [8] and SimpleClick [20], have become the most mainstream and popular methods in recent years. This is because selecting target objects through clicks is not only intuitive but also highly efficient and flexible in practical applications. However, click-based methods also have certain limitations. For example, providing four accurate extreme clicks required by DEXTR [23] can be difficult in certain scenarios, which is why IOG [38] proposed to utilize boxes as an interaction form. Additionally, when dealing with thin or intricate objects, almost all click-based methods become user-unfriendly because precisely locating and clicking in narrow areas is highly inefficient. To address this, Slim Scissors [6] proposes using rough scribbles as an interaction method for fine regions. NTClick proposed in this paper retains the efficiency of click-based methods while incorporating the advantage of scribble-based methods, which do not need precise localization.

2.2. High Quality Segmentation

In recent years, many efforts have been dedicated to higher segmentation accuracy. Approaches such as Pointrend [13], Refinemask [37] and Mask transfiner [11] are designed to directly predict high-precision results. Other methods focus on refining obtained masks, such as CascadePSP [3], SegFix [35] and SegRefiner [30]. These methods refine existing outputs in a decoupled manner. Additionally, there is a subtask of segmentation called image matting [16, 25, 33, 34], which mainly aims to convert a obtained binary mask or a special three-class mask (trimap) into fine-grained predictions. However, matting requires perceiving the transparency of the target, thus creating a gap between it and high quality segmentation. NTClick is inspired by the concept of trimap and incorporates some of the successful designs of matting methods in the refinement stage.

3. Methodology

3.1. Noise-tolerant Click

To address the pain point of requiring frequent precise localization and clicking from users when segmenting intricate targets, we propose the use of *noise-tolerant click*, a third option of *clicks* distinct from *foreground click* and *background click*. *noise-tolerant click* only needs to fall within the approximate range of the fine regions, as it is merely intended to inform the model that there are fine objects or structures in the vicinity.

Take the bonsai in Fig. 2 as an example, for difficult-to-locate regions such as the foliage, users are encouraged to provide a *noise-tolerant click*. Whether the *noise-tolerant click* lands on the foliage, in the gaps between the foliage, or on the edges between the foliage and gaps, NTClick can stably predict accurate results. Meanwhile, for large continuous regions that are easy to locate, like the plant pot, trunk and background, NTClick encourages users to follow the original interaction mode by providing *foreground click* and *background click*.

We have retained as much of the original interaction mode of click-based methods as possible because *click* is inherently an excellent form of interaction, capable of segmenting large dense regions with very sparse user input. Therefore, *noise-tolerant click* is a supplement to the existing interaction mode based on *foreground click* and *background click*. Both *foreground click*, *background click* and *noise-tolerant click* are encoded using disk maps and added with the RGB image as separate channels, following the click encoding used by RITM [29] and SimpleClick [20].

3.2. Explicit Coarse Perception Network

In the first stage, we use an **E**xplicit **C**oarse **P**erception network (ECP) to make an initial explicit estimation, called **F**oreground **B**ackground **U**ncertain (FBU) map. FBU map is essentially a three-class mask, with the three classes being *foreground*, *background* and *uncertain*:

$$\text{FBU map} \in \{\textit{foreground}, \textit{background}, \textit{uncertain}\}^{W\times H} \tag{1}$$

Taking the bonsai in Fig. 2 as an example, *foreground* region is supposed to include simple parts without complex structures, such as the cup and trunk. *background* region is supposed to include large, continuous backgrounds, such as the yard. *uncertain* is supposed to include regions with intricate structures, such as foliage and edges. Pixels within *uncertain* regions are temporarily marked as uncertain, without specific classification.

We choose to use an explicit approach to constrain ECP's learning of the FBU map. This required creating "ground truth FBU map", for which we apply a simple morphology process to convert the fine-grained ground truth masks provided by the dataset into "ground truth FBU maps." This morphological processing mainly consists of erosion and dilation operations, as it is a common practice in image processing to smooth fine details in masks. Take the bonsai in Fig. 2 as an example, applying a few rounds of dilation and erosion transforms the foliage and edge regions into *uncertain* regions. The code for morphology process can be found in the supplementary materials. In both the training and evaluation phases, we use the ground truth FBU map generated by this manner to determine where to sample *foreground click*, *background click*, or *noise-tolerant click*.

Since ECP does not require precise classification of fine-grained regions and only needs to categorize them as *uncertain*, this perception can be accomplished at a relatively low resolution. As shown in the ablation study in Tab. 3, an ECP network with an input resolution of 448 is capable of outputting high-quality FBU map.

In designing the model structure, we aimed to keep it as simple as possible. As shown in Fig. 2, we opt to follow SimpleClick and SegNext [20, 21], using a plain ViT [17] composed of window attention blocks and global attention blocks as the encoder, and a decoder composed of a simple feature pyramid [19] and MLP. However, the input to ECP consists of three types of clicks rather than two, and the output is an FBU map instead of a binary mask. Therefore, the model's learning objective is fundamentally different. Details of specific structure of ECP, training and data processing can be found in Sec. 1 of supplementary materials.

3.3. High Resolution Refinement Network

In the second stage, we utilize a **H**igh **R**esolution **R**efinement network (HRR) to perform fine-grained perception. As shown in Fig. 2, we first upsample the FBU map predicted by ECP from a low resolution to 1024^2 and concatenate it with the RGB image before feeding them into HRR. Since the FBU map has already classified a substantial portion of the image as either *foreground* or *background*, and this information provides a strong prior for the perception, HRR can achieve accurate classification of pixels within the remaining *uncertain* regions. The designing of HRR's structure is not challenging, since using a dense map to assist in fine-grained segmentation is a widely adopted and validated approach, both in fields of mask refinement [30, 35] and image matting [16, 34]. To maintain the simplicity of NTClick, we construct the HRR with a plain ViT and a simple convolution-based decoder, following ViTMatte [33]. We use ground truth masks to directly supervise the training of HRR.

Hereby, a complete NTClick is ready. As shown by the first result in Tab. 4, this version of NTClick has already achieved competitive results in comparison with state-of-the-art methods, demonstrating the feasibility of our workflow: "noise-tolerant click $\rightarrow$ FBU map $\rightarrow$ mask".

But what if we want higher precision? We believe that the key lies in removing the constraints of resolution. As shown in Fig. 3, when performing fine-grained segmentation, limiting the resolution to 1024 leads to significant detail loss. Some details, which are only a few pixels wide, are irretrievably lost due to downsampling. However, we cannot directly increase the input resolution of HRR because both the computational and space complexity of the global attention block in plain ViT increase quadratically with the image resolution. This is also one of the key reasons why popular interactive segmentation methods (e.g., SAM-HQ [12] and SegNext [21]) must limit input resolutions to 1024^2, because their use of ViT leads to running out of memory at higher resolutions (e.g., 2048^2 and 4096^2). Therefore, we replace global attention block with grid attention block proposed in MobileViT [24] for two reasons: 1) Grid attention block has a large perceptive field and ca-

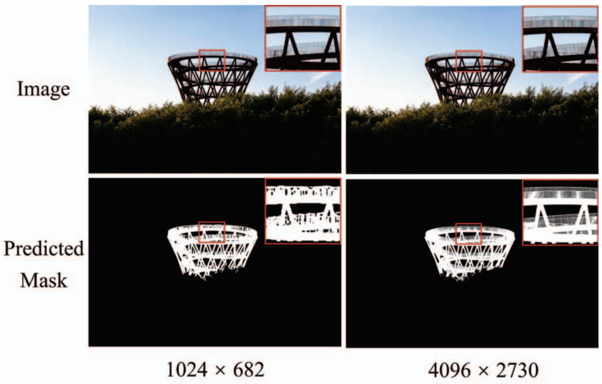

Figure 3. The impact of resolution on qualities of images and predicted masks. Please zoom in to examine the details.

pable of establishing long-range dependencies like global attention block. 2) Because grid attention only computes self-attention for patches spaced by a interval of K, the computational complexity of grid attention is significantly lower. Consider an image of $w \times h$ patches, the computational complexity of global attention and grid attention are:

$$\begin{aligned} \Omega(\text{Global Attention}) &= 4whC^2 + 2(wh)^2C \\ \Omega(\text{Grid Attention}) &= 4whC^2 + 2\frac{(wh)^2}{K^2}C \end{aligned} \tag{2}$$

where C represents the dimension of a patch and K represents the interval of grid attention. For high-resolution images, the computational burden primarily arises from the latter term. Therefore, by setting a relatively large K, we can significantly reduce the computational complexity. Meanwhile, analyzing from the perspective of space complexity, the space complexity of global attention is $O(w^2h^2)$, whereas that of grid attention is only $O(\frac{w^2h^2}{K^2})$. The default setting for K in NTClick is 8.

However, this replacement introduces issues in establishing local dependencies. When the global attention block in plain ViT is replaced with grid attention block, the backbone is essentially composed of window attention block and grid attention block. As demonstrated in Fig. 4, due to the fixed window partitioning mechanism in window attention and the lack of global attention block, two adjacent patches often have no chance to construct direct dependency with each other from start to finish. The exchange of information between them can only be facilitated through intermediary patches. We consider that the lack of direct dependencies for adjacent patches is detrimental for detail-oriented tasks like fine-grained segmentation. To this end, we replace window attention block with neighborhood attention block proposed in NAT [7]. As shown in Fig. 2, neighborhood attention always computes direct attention with patches in its immediate vicinity, rather than computing attention within

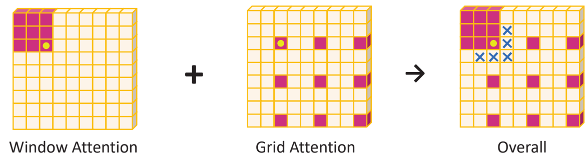

Figure 4. ■ represents all patches that have the opportunity to compute direct attention with the • patch. × represents all patches that are adjacent to the • patch but have no chance to compute direct attention with • patch.

a fixed window region like window attention. The ablation study in Tab. 4 also demonstrates that the use of neighborhood attention yields a significant performance improvement compared to a rough replacement. Finally, we significantly increase the input resolution of HRR with a manageable computational overhead. For example, we can achieve refinement at a resolution of 4096^2 with a consumer-grade GPU (NVIDIA RTX 3090 24GB). Under the same hardware conditions, methods like SAM-HQ [12] and SegNext [21] run out of memory when the input resolution reaches only 2048^2. As shown in Fig. 3, expanding input resolution brings about a substantial performance improvement.

To this end, we have optimized both the short-range and long-range dependencies in plain ViT, achieving fine-grained perception at ultra-high resolution with controlled computational overhead. This optimization fundamentally only alters the mechanism for computing attention between patches without introducing new structures or parameters. In this study, we primarily validate the significance of expanding resolution for fine-grained segmentation with minimal modifications to a plain ViT.

4. Experiments

4.1. Benchmark

Datasets. We conduct evaluations on three datasets: DIS5K [26], ThinObject5K [18] and and HRSOD [36]. DIS5K includes four test sets, TE1, TE2, TE3, and TE4, each containing 500 images with increasing resolution and complexity. Among them, TE4 is the most challenging one, featuring the highest resolution and the most intricate object structures. Additionally, DIS5K includes a validation set that features a comprehensive range of resolutions and scene complexities, making it our first choice for ablation studies. HRSOD [36] is specifically designed for the task of salient object detection at high resolutions and ThinObject5K [18] focuses on the challenging task of segmenting thin and elongated objects. We also test on these two datasets to further enhance the diversity of scenarios in our benchmark. In Tab. 1, we present isoperimetric inequality quotient (IPQ) of each test set, which mainly describes the overall structure complexity of the scenes contained within the dataset. The

Method \ Benchmark	DIS5K-VD IPQ=157		DIS5K-TE1 IPQ=27		DIS5K-TE2 IPQ=51		DIS5K-TE3 IPQ=93		DIS5K-TE4 IPQ=443		HRSOD IPQ=6		ThinObject-5K IPQ=27	
	NoC90	5-mIoU	NoC90	5-mIoU	NoC90	5-mIoU	NoC90	5-mIoU	NoC90	5-mIoU	NoC90	5-mIoU	NoC90	5-mIoU
RITM[29] ICIP23	15.21	50.49	11.86	64.68	14.16	54.88	16.25	50.13	18.55	43.93	4.82	87.80	8.31	74.59
CDNet[1] ICCV21	15.40	63.60	12.90	73.77	15.44	68.20	17.10	63.88	18.82	54.04	5.49	88.60	7.47	85.71
FocalClick[2] CVPR22	12.56	77.97	9.16	85.69	11.39	82.58	14.00	78.66	17.46	69.56	3.48	93.93	4.19	92.51
SAM[14] ICCV23	13.01	71.37	9.06	81.71	11.73	76.31	14.44	69.54	18.14	59.79	3.34	92.96	6.41	88.72
SimpleClick[20] ICCV23	12.89	72.98	9.78	85.09	11.79	78.20	14.46	73.13	17.98	60.83	3.84	93.08	4.38	91.42
FocSAM[9] CVPR24	11.70	76.33	7.73	86.95	9.66	82.76	12.40	76.86	17.21	61.77	3.10	94.03	5.25	90.88
SAM-HQ[12] Neurips23	11.04	80.21	7.35	87.39	8.80	84.49	11.08	79.86	16.18	70.35	3.12	94.65	5.01	91.84
SegNext[21] CVPR24	10.35	85.69	8.19	87.76	8.38	88.34	10.15	87.21	14.61	79.94	3.56	94.24	3.19	94.26
NTClick (Ours)	**7.23**	**89.23**	**6.12**	**90.34**	**5.90**	**90.82**	**6.54**	**89.96**	**9.15**	**87.98**	**2.16**	**95.64**	**2.76**	**95.20**

Table 1. Comparison with click-based methods on various datasets. Due to constraints on the width of the table, we denote NoC@90 as NoC90 here. All the methods compared in this table represent their strongest versions. The best results are highlighted in bold black font.

IPQ values are collected from DIS5K [26], where the IPQs for HRSOD and ThinObject5K represent those of the entire dataset, rather than just the test set. Nevertheless, we report them as a reference. Overall, DIS5K features the most challenging scenarios and the highest annotation quality.

Metrics. Following previous work, we mainly use Number of Click (NoC) and mIoU. NoC indicates the average number of clicks required per image to achieve a predefined IoU. It assesses the interaction cost of the model in practical use. For example, NoC@90 indicates the average number of clicks required to reach an IoU of 90. Additionally, we set the maximum number of clicks at 20 to handle scenarios where the IoU fails to reach the predefined threshold following convention. mIoU represents the average intersection over union (IoU), while 5-mIoU refers to the mIoU after 5 continuous clicks. We use this metric to directly demonstrate the segmentation accuracy of different methods, following SegNext [35].

4.2. Implementation Details

We used HQSeg-44k [12] as our training set. HQSeg-44k contains six datasets, including DIS5K (train set) [26], ThinObject5K (train set) [18], FSS-1000 [15], ECSSD [28], MSRA-10K [4] and DUT-OMRON [32], with a total of 44,320 images. ECP and HRR can actually be trained in a decoupled manner. ECP was trained for 25 epochs using the adam optimizer, with the learning rate set to 5×10^{-5}. HRR was trained for 200k iterations using the AdamW optimizer, with the learning rate set to 5×10^{-4}. More details about training and evaluation can be found in Sec. 2 of supplementary materials.

Methods	Interactions	DIS5K-VD	ThinObject-5K
Slim Scissors [6]	Click&Scribble	82.57	91.10
NTClick (Ours)	Click	**89.23**	**95.20**

Table 2. Comparison with scribble-based method.

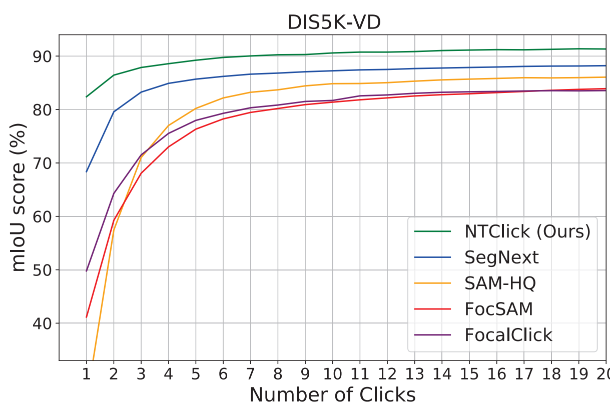

Figure 5. IoU curves of NTClick and other methods.

4.3. Quantitative Comparisons

We first compare NTClick with various click-based methods, including interactive segmentation methods and the methods derived from SAM [14]. As shown in Tab. 1, NTClick achieves performance superior to state-of-the-art methods, demonstrating that our method is capable of performing fine-grained perception using imprecise user interactions and the feasibility of our two-stage workflow. NTClick achieves optimal performance across all seven test sets. Particularly on the most challenging test set, DIS5K-

Backbone		Resolution		NoC@90↓	5-mIoU↑
ViT-Base	ViT-Huge	448^2	1024^2		
✓		✓		8.58	87.32
✓			✓	7.94	88.45
	✓	✓		**7.23**	**89.23**

Table 3. Ablation study of ECP network. A large backbone with a small input resolution can achieve better coarse perception.

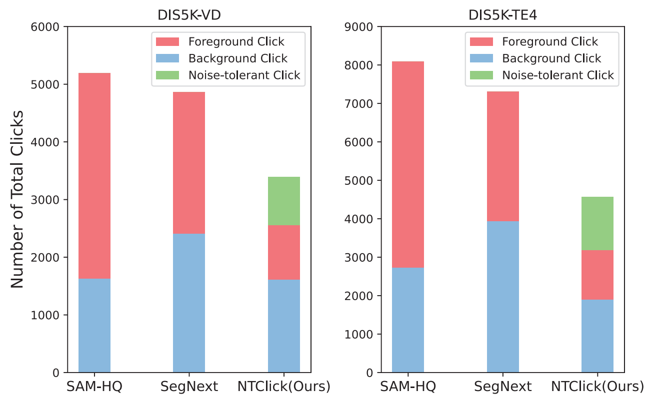

Figure 6. Statistic of different types of clicks. The data was collected by recording the number of clicks required for different methods to achieve a 90% mIoU on DIS5K-VD (470 images) and DIS5K-TE4 (500 images). Following common practice, the maximum click limit was set to 20.

TE4, NTClick delivers a remarkable improvement of 5.46 in NoC@90 and 8.04 in 5-mIoU. IoU curve of NTClick is shown in Fig. 5.

In addition, we recorded the number of each type of clicks sampled during the evaluation process, as illustrated in Fig. 6. To achieve the same level of accuracy, NTClick not only requires fewer clicks but also allows users to employ *noise-tolerant clicks* in intricate regions where foreground and background are difficult to distinguish, without the need for precise localization, making it comparatively more efficient and user-friendly. Moreover, this advantage becomes more pronounced as the complexity of scenes and objects increases.

In Tab. 2, we also compare NTClick with Slim Scissors [6], a powerful method capable of performing fine segmentation based on users' rough scribbles. As both methods do not require precise localization and inputs, NTClick achieves higher segmentation accuracy with sparser input. Overall, NTClick delivers superior segmentation accuracy with a more user-friendly interactive form, both in comparison with existing click-based and scribble-based methods.

4.4. Ablation Study

Explicit Coarse Perception. ECP essentially performs an initial coarse perception, without the need to predict the final mask of intricate regions. Therefore, ECP does not require a high input resolution. In Tab. 3, we explore the performance of different input resolutions and backbone scales. The results indicate that using a large backbone achieves the best even at low resolutions. Meanwhile, ECP with a small scale backbone and low input resolution can also achieve performance superior to existing models.

Local Attention	Long-range Attention	NoC@90↓	5-mIoU↑
window	global	9.54	86.98
window	×	17.31	72.17
window	grid	7.74	88.90
neighbourhood	grid	**7.23**	**89.23**

Table 4. Ablation study of HRR network. Using neighbourhood attention to establish local dependencies and grid attention to enable long-range dependencies achieves better performance.

Method	Resolution Limitations	NoC@90↓	5-mIoU↑	Latency/s
SAM-HQ	1024	11.04	80.21	0.43
SegNext	1024	10.35	85.69	0.15
NTClick	1024	9.89	86.39	0.40
	2048	7.70	88.99	0.52
	4096	7.23	89.23	0.81

Table 5. Comparison of the first-click latency with other methods.

High Resolution Refinement. We first design a version of HRR composed of a plain ViT, with an input resolution limit of 1024, as reported in the first result in Tab. 4. However, as shown in Fig. 3, the resolution of 1024 obviously limits the accuracy of fine-grained segmentation. Thus, we opts to increase the input resolution during the refinement stage. The challenge here lies in the fact that the computational complexity of global self-attention, used to build long-range dependencies, increases quadratically with image resolution. However, not constructing long-range dependencies would result in a significant drop in performance, as indicated by the second result in Tab. 4. Therefore, we propose using a grid attention instead of global attention to build long-range dependencies, and employing neighborhood attention instead of window attention to compensate for the shortcomings of grid attention in constructing local dependencies. As shown in Tab. 4, this version of HRR provide the superior performance.

Input Resolution Limitation. NTClick is essentially capable of inferring at a dynamic resolution, allowing users to achieve different levels of accuracy and latency by adjusting the resolution limit of the input image, as shown in Tab. 5. Generally, a higher resolution provides greater accuracy but also introduces higher latency. However, higher accuracy reduces the number of interactions required, which is crucial for interactive models. For instance, NTClick achieves an accuracy with just 4 clicks that other methods require 20 clicks to reach, as shown in Fig. 5. Therefore, NTClick's

Figure 7. Visual comparison results of NTClick, SegNext and Slim Scissors. •, and • refer to *foreground click* and *background click*, • refers to *noise-tolerant click*. Zoom in to observe details.

random seed	100	200	300	400	500	Mean Value
NoC@90	7.00	6.85	6.90	6.86	6.84	6.89
5-mIoU	89.99	90.11	90.13	90.18	90.21	90.12

Table 6. Robustness of clicks. The results with slight variations demonstrate that NTClick is robust to localization noise across various click patterns.

time cost is expected to be lower in practical applications when accounting for user response and operation time.

4.5. Robustness of Different Clicks

The core advantage of NTClick is that it does not require precise localization when selecting fine objects or segmenting intricate regions. Fig. 1 and 7 demonstrate this advantage visually, and here we further provide quantitative results in Tab. 6. During this testing, we allow clicks to be randomly sampled within the error regions and collect results with five different random seeds (100, 200, 300, 400, 500). The results indicate that NTClick provides stable performance for various types of clicks.

4.6. Visualization Results

In Fig. 7, we provide several visual comparisons between NTClick, SegNext and SAM-HQ. Notably, *noise-tolerant clicks* of NTClick often fall between the gaps and edges of intricate objects, yet this does not affect NTClick's ability to predict stable and accurate masks. More visualizations can be found in Sec. 3 of supplementary materials.

4.7. Limitations

Although NTClick achieves significant improvements in fine-grained segmentation, there remains a noticeable gap compared to the ground truth. This gap is challenging to eliminate by providing continuous clicks, a common dilemma observed in all existing methods. Once the number of provided clicks reaches a certain threshold (approximately 10), the benefit of subsequently clicks diminishes considerably, as shown in Fig. 5.

5. Conclusion

In this paper, we propose NTClick, a powerful interactive segmentation method capable of predicting high-quality masks with imprecise user clicks. NTClick introduces a new form of interaction that effectively addresses the challenge of requiring frequent precise localization in click-based methods for fine-grained segmentation, significantly enhancing user-friendliness. Furthermore, by employing a two-stage workflow, NTClick offers significant improvements with a more flexible interaction mode. We believe that the user-friendliness of NTClick will lead to substantial practical value.

Acknowledgement. This work was supported by the Fundamental Research Funds for the Central Universities (2024XKRC082), the National Natural Science Foundation of China (No.U23A20314, No.92470203), Beijing Natural Science Foundation (No.L242022).

References

[1] Xi Chen, Zhiyan Zhao, Feiwu Yu, Yilei Zhang, and Manni Duan. Conditional diffusion for interactive segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7345–7354, 2021. 6

[2] Xi Chen, Zhiyan Zhao, Yilei Zhang, Manni Duan, Donglian Qi, and Hengshuang Zhao. Focalclick: Towards practical interactive image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1300–1309, 2022. 6

[3] Ho Kei Cheng, Jihoon Chung, Yu-Wing Tai, and Chi-Keung Tang. Cascadepsp: Toward class-agnostic and very high-resolution segmentation via global and local refinement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8890–8899, 2020. 3

[4] Ming-Ming Cheng, Niloy J Mitra, Xiaolei Huang, Philip HS Torr, and Shi-Min Hu. Global contrast based salient region detection. *IEEE transactions on pattern analysis and machine intelligence*, 37(3):569–582, 2014. 6

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021. 1

[6] Kunyang Han, Jun Hao Liew, Jiashi Feng, Huawei Tian, Yao Zhao, and Yunchao Wei. Slim scissors: Segmenting thin object from synthetic background. In *European Conference on Computer Vision*, pages 379–395. Springer, 2022. 1, 3, 6, 7

[7] Ali Hassani, Steven Walton, Jiachen Li, Shen Li, and Humphrey Shi. Neighborhood attention transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6185–6194, 2023. 5

[8] You Huang, Hao Yang, Ke Sun, Shengchuan Zhang, Liujuan Cao, Guannan Jiang, and Rongrong Ji. Interformer: Real-time interactive image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22301–22311, 2023. 3

[9] You Huang, Zongyu Lan, Liujuan Cao, Xianming Lin, Shengchuan Zhang, Guannan Jiang, and Rongrong Ji. Focsam: Delving deeply into focused objects in segmenting anything. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3120–3130, 2024. 6

[10] Michael Kass, Andrew Witkin, and Demetri Terzopoulos. Snakes: Active contour models. *International journal of computer vision*, 1(4):321–331, 1988. 2

[11] Lei Ke, Martin Danelljan, Xia Li, Yu-Wing Tai, Chi-Keung Tang, and Fisher Yu. Mask transfiner for high-quality instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4412–4421, 2022. 3

[12] Lei Ke, Mingqiao Ye, Martin Danelljan, Yu-Wing Tai, Chi-Keung Tang, Fisher Yu, et al. Segment anything in high quality. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 4, 5, 6

[13] Alexander Kirillov, Yuxin Wu, Kaiming He, and Ross Girshick. Pointrend: Image segmentation as rendering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9799–9808, 2020. 3

[14] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 1, 6

[15] Xiang Li, Tianhan Wei, Yau Pun Chen, Yu-Wing Tai, and Chi-Keung Tang. Fss-1000: A 1000-class dataset for few-shot segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2869–2878, 2020. 6

[16] Yaoyi Li and Hongtao Lu. Natural image matting via guided contextual attention. In *Proceedings of the AAAI conference on artificial intelligence*, pages 11450–11457, 2020. 3, 4

[17] Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. In *European conference on computer vision*, pages 280–296. Springer, 2022. 4

[18] Jun Hao Liew, Scott Cohen, Brian Price, Long Mai, and Jiashi Feng. Deep interactive thin object selection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 305–314, 2021. 1, 5, 6

[19] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 4

[20] Qin Liu, Zhenlin Xu, Gedas Bertasius, and Marc Niethammer. Simpleclick: Interactive image segmentation with simple vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22290–22300, 2023. 3, 4, 6

[21] Qin Liu, Jaemin Cho, Mohit Bansal, and Marc Niethammer. Rethinking interactive image segmentation with low latency high quality and diverse prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3773–3782, 2024. 1, 4, 5, 6

[22] Sabarinath Mahadevan, Paul Voigtlaender, and Bastian Leibe. Iteratively trained interactive segmentation. *arXiv preprint arXiv:1805.04398*, 2018. 2

[23] Kevis-Kokitsi Maninis, Sergi Caelles, Jordi Pont-Tuset, and Luc Van Gool. Deep extreme cut: From extreme points to object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 616–625, 2018. 3

[24] Sachin Mehta and Mohammad Rastegari. Mobilevit: light-weight, general-purpose, and mobile-friendly vision transformer. *arXiv preprint arXiv:2110.02178*, 2021. 4

[25] GyuTae Park, SungJoon Son, JaeYoung Yoo, SeHo Kim, and Nojun Kwak. Matteformer: Transformer-based image matting via prior-tokens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11696–11706, 2022. 3

[26] Xuebin Qin, Hang Dai, Xiaobin Hu, Deng-Ping Fan, Ling Shao, and Luc Van Gool. Highly accurate dichotomous image segmentation. In *European Conference on Computer Vision*, pages 38–56. Springer, 2022. 1, 2, 5, 6

[27] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. " grabcut" interactive foreground extraction using iterated graph cuts. *ACM transactions on graphics (TOG)*, 23(3): 309–314, 2004. 2

[28] Jianping Shi, Qiong Yan, Li Xu, and Jiaya Jia. Hierarchical image saliency detection on extended cssd. *IEEE transactions on pattern analysis and machine intelligence*, 38(4): 717–729, 2015. 6

[29] Konstantin Sofiiuk, Ilya A Petrov, and Anton Konushin. Reviving iterative training with mask guidance for interactive segmentation. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 3141–3145. IEEE, 2022. 2, 4, 6

[30] Mengyu Wang, Henghui Ding, Jun Hao Liew, Jiajun Liu, Yao Zhao, and Yunchao Wei. Segrefiner: Towards model-agnostic segmentation refinement with discrete diffusion process. *arXiv preprint arXiv:2312.12425*, 2023. 3, 4

[31] Ning Xu, Brian Price, Scott Cohen, Jimei Yang, and Thomas S Huang. Deep interactive object selection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 373–381, 2016. 2

[32] Chuan Yang, Lihe Zhang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. Saliency detection via graph-based manifold ranking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3166–3173, 2013. 6

[33] Jingfeng Yao, Xinggang Wang, Shusheng Yang, and Baoyuan Wang. Vitmatte: Boosting image matting with pretrained plain vision transformers. *Information Fusion*, 103: 102091, 2024. 3, 4

[34] Qihang Yu, Jianming Zhang, He Zhang, Yilin Wang, Zhe Lin, Ning Xu, Yutong Bai, and Alan Yuille. Mask guided matting via progressive refinement network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1154–1163, 2021. 3, 4

[35] Yuhui Yuan, Jingyi Xie, Xilin Chen, and Jingdong Wang. Segfix: Model-agnostic boundary refinement for segmentation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*, pages 489–506. Springer, 2020. 3, 4, 6

[36] Yi Zeng, Pingping Zhang, Jianming Zhang, Zhe Lin, and Huchuan Lu. Towards high-resolution salient object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7234–7243, 2019. 2, 5

[37] Gang Zhang, Xin Lu, Jingru Tan, Jianmin Li, Zhaoxiang Zhang, Quanquan Li, and Xiaolin Hu. Refinemask: Towards high-quality instance segmentation with fine-grained features. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6861–6869, 2021. 3

[38] Shiyin Zhang, Jun Hao Liew, Yunchao Wei, Shikui Wei, and Yao Zhao. Interactive object segmentation with inside-outside guidance. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12234–12244, 2020. 3