RETHINKING INTERACTIVE IMAGE SEGMENTATION: A PROTOTYPE SELECTION VIEW

Anonymous authors
Paper under double-blind review

ABSTRACT

Interactive image segmentation aims to extract target objects based on the user guidance, and recent click-based approaches gain popularity for their simplicity and efficiency. However, existing deep interactive methods often struggle to retain sparse spatial prompts, capture intra-class variance, and leverage label semantics beyond predefined channels. In this paper, we rethink interactive segmentation from a prototype learning perspective, where each user click defines a prototype comprising both spatial and label attributes. We propose a novel framework that learns spatial prototypes to guide feature flow and progressive label prototypes to model object characteristics, enhancing the embedding of user intention and enabling parallel segmentation querying without fixed label constraints. Instead of direct selecting an object, the joint representation of multiple prototypes helps to capture the intra-instance variance. Additionally, we introduce difficulty-aware and similarity-driven training mechanisms to enhance focus on challenging regions and improve inter-class discrimination. Vast experiments demonstrate that our method outperforms current state-of-the-art methods across diverse benchmarks.

1 INTRODUCTION

Interactive image segmentation is a task that aims to segment specific objects of interest based on user interactions Boykov & Jolly (2001). It provides a pixel-level intelligent interaction system that follows human intent, so it is useful in many applications such as image editing, controllable image generation, and medical pathological analysis.

Early interactive segmentation methods typically formulated the task as optimizing hand-crafted energy functions Boykov & Jolly (2001); Grady (2006). However, their reliance on low-level visual features made them vulnerable to challenges such as similar appearances, complex texture, and varying illumination Wang et al. (2019). The advent of deep learning has markedly advanced this field: leveraging powerful feature representation, user interactions have been streamlined to just a few clicks Xu et al. (2016). Due to its simplicity and well-established protocols, click-based interactive segmentation has emerged as a research focus.

Xu et al. Xu et al. (2016) established the first deep interactive segmentation framework by jointly feeding the image and user input into the fully convolutional network (FCN) Long et al. (2015), enabling end-to-end learning of interaction patterns and segmentation intent. Subsequent studies improved network architectures Liu et al. (2023); Chen et al. (2022); Wang et al. (2023), interaction representations Majumder & Yao (2019); Du et al. (2023); Lee et al. (2024), and iterative training strategies Sofiiuk et al. (2022); Sun et al. (2024); Lin et al. (2025), while SAM Kirillov et al. (2023) enhanced generalization using large-scale training data. *The essence of deep learning-based methods lies in perceiving specific targets from the spatial location and label attributes of user interactions.* Although significant improvements have been made compared to the conventional energy-based methods, existing deep learning-based methods still face the following challenges. First, the spatial position clicked is generally utilized to select a target of interest, and the sparse click information is easily lost in the long-path network. Though many efforts have been made Lin et al. (2020); Jang & Kim (2019); Sofiiuk et al. (2020) to strengthen the attention to the clicked points, these spatial prompts cannot fundamentally reflect the characteristics of the target (shown in Figure 1(a)). Second, regardless of how many clicks are provided to describe an object, usually only one single weight

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

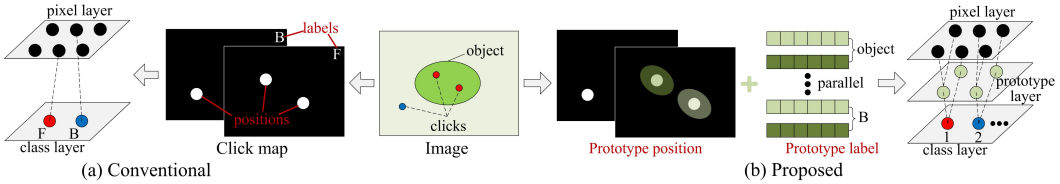

Figure 1: The basic framework of interactive segmentation. F and B represent the foreground and background, respectively.

is learned to estimate sigmoid distributions for dense prediction, insufficient to capture rich intra-instance variance. Moreover, the attribute of the clicked label is generally reflected in distinguishable image channels (such as the 2-channel click map in Figure 1), resulting in existing models often being limited by pre-defined fixed label spaces and lacking in-depth exploration of category priors. Last, the commonly used iterative training strategy mainly focuses on the selection of click samples during continuous corrections, ignoring the supervision of click-aware hard regions.

On the other hand, prototype-based classification has been studied for a long time. Classes are represented by prototypes that are then used for classification, and observations can be directly compared with representative examples due to its exemplar-driven nature and intuitive interpretation Duda & Hart (1973). The recent work Zhou et al. (2022) unifies FCN and attention-based semantic segmentation models from a prototype perspective. Inspired by the above, in this paper we rethink the interactive segmentation task from the perspective of prototype learning: each user interaction aims to define a prototype (including prototype position and prototype label) instead of selecting a target in existing methods.

To address the above problems, we propose a click-based deep interactive segmentation method based on prototype learning in this paper. The prototype space can be freely defined by the user and each click is used to create a prototype (shown in Figure 1(b)). Prototype positions aim to guide where the target is and prototype labels aim to represent the characteristics of an object in parallel, thereby promoting the description of what the object looks like in the deep network. Meanwhile, each object can be abstracted by a set of prototypes that aligns with the fact multiple clicks are generally used to select an object, which helps to capture intra-instance variance. In our method, prototypes of different targets can be utilized to query the dense embeddings simultaneously (similar to Cheng et al. (2022)), without being limited by fixed label channels, while also effectively capturing label priors. The detailed contributions of this paper are as follows:

- Rethinking interactive image segmentation from a prototype view and modeling user interaction as a prototype selection.
- Developing a spatial prototype learning method to control the feature flow from encoder to decoder following the guidance of clicks, enhancing the embedding of user segmentation intention into the network.
- Designing a progressive label learning method from the initial click-driven prototypes to capture object characteristics and intra-variance, compensating for the sparse spatial prompts and achieving parallel querying of the segmentation regardless of the number of objects.
- Measuring the segmentation difficulty of temporal click-aware regions and pixel-level similarity within and between prototypes during training, strengthening attention to hard regions and improving inter-class discrimination.

2 Related Work

Conventional interactive methods: Early methods optimized energy functions, including graph cuts Boykov & Jolly (2001), level sets Li et al. (2010), and random walks Grady (2006). Handcrafted features limited their performance.

Deep learning-based interactive methods: Xu et al. Xu et al. (2016) first used convolutional neural networks (CNN) for interactive segmentation by transforming user clicks into click maps via

distance transform and fusing them with the image as network input. Works in Sofiiuk et al. (2022); Sun et al. (2024); Chen et al. (2021) adopted stronger backbones such as ResNet He et al. (2016), HRNet Sun et al. (2019), and Transformer Vaswani et al. (2017) to enhance segmentation accuracy. These methods improved click representations to better encode user intentions through Gaussian disks Chen et al. (2021); Benenson et al. (2019), superpixels Majumder & Yao (2019), previous masks Hao et al. (2021); Liu et al. (2024), and probability maps Lee et al. (2024). To avoid the loss of sparse click information, these methods Lin et al. (2020); Jang & Kim (2019); Sofiiuk et al. (2020) imposed constraints on user interaction by enhancing the attention of clicked pixels. Iterative training models temporal corrections Sofiiuk et al. (2022), while later work introduced granularity-controllable learning Zhao et al. (2024) and self/cross-attention for global feature diffusion Rana et al. (2023); Li et al. (2024). This paper instead applies pixel queries to prototype selection.

Segment anything model: SAM Kirillov et al. (2023) learns rich object priors from the large-scale SA-1B dataset and predicts multiple candidate masks simultaneously to address interaction ambiguity, demonstrating strong generalization and broad transferability to domains such as medicine Wong et al. (2024). However, its network structure and interaction encoding still suffer from these limitations, and it lags behind specialized models in high-quality segmentation tasks Liu et al. (2024). Subsequent variants including MobileSAM Zhang et al. (2023), HQ-SAM Ke et al. (2023), and extended prompts for diverse scenarios Liu et al. (2024); Wong et al. (2024) have improved efficiency and adaptability.

Prototype learning: Prototype learning has a long history in machine learning, primarily used for nonparametric classification Garcia et al. (2012). Recently, integrating deep learning with prototype learning Snell et al. (2017) has shown promise in few-shot, supervised, and unsupervised classification, inspiring few-shot segmentation methods Dong & Xing (2018); Vinyals et al. (2016). Notably, Zhou et al. (2022) unified FCN and attention-based semantic segmentation models from a prototype perspective, constructing a nonparametric, non-learnable prototype-based segmentation framework that enhances efficiency and generalization for large-vocabulary tasks. Inspired by these works, this paper explores prototype learning from the perspective of user interaction (including prototype position and prototype label perception) for interactive segmentation.

3 Interactive Segmentation based on Prototype Learning

User interaction generally contains two types of information: target location and user-defined labels. Existing deep learning-based methods widely adopt sparse distance maps (Euclidean Xu et al. (2016) or Gaussian Chen et al. (2021)) to represent location clues in the network. Label attributes are reflected in distinguishable channels, where one map is used as foreground and the other as background. Unlike existing models, in this paper, we are rethinking interactive segmentation at a prototype view, aiming at achieving sufficient characterization of varying objects with multiple prototypes corresponding to multiple user interactions.

3.1 Overall Framework

Given an image, the user selects targets through clicks, each of which defines a prototype. As shown in Figure 2, the network takes the image and a three-channel position map as input. In prototype position guidance, we only distinguish between object things and background stuff (same as existing methods in one-object selection), since compared with regular foreground objects, the background usually does not have a fixed shape. In multi-object settings, the thing and previous-mask channels remain label-free across instances and serve only as spatial priors. The previous mask also helps connect temporal segmentation corrections Sofiiuk et al. (2022). The network structures involved mainly include ResNet He et al. (2016), HR-Net Sun et al. (2019) and Transformer Vaswani et al. (2017).

The framework contains the spatial and label prototype learning modules. The former uses regularly sampled spatial prototypes to transmit label information from the encoder to the decoder through skip connections; the latter progressively learns queryable label prototypes from multi-level decoder features. After mask merging, each class is queried using its label prototypes and the decoder output.

Both modules are initialized by user clicks. Given a click c in the image space $I \in \mathbb{R}^{H\times W\times 3}$, its mapped position c' in a feature space F is enclosed by four discrete neighbors $\{z_1, z_2, z_3, z_4\}$. The

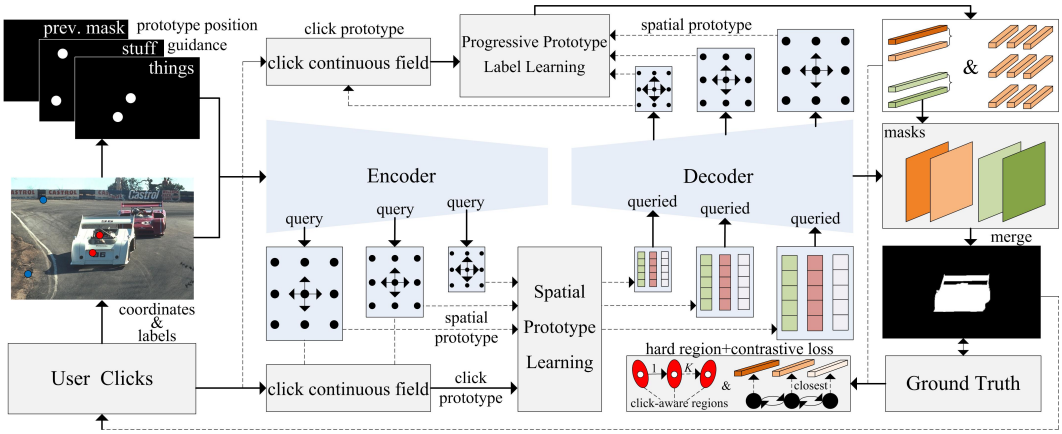

Figure 2: Overall architecture of the proposed model.

initial click prototype is generated based on spatial continuous field Hu et al. (2022):

$$P_c = \sum_{m\in\{1,2,3,4\}} \frac{s_m}{\sum_{m\in\{1,2,3,4\}} s_m} F(z_m) \tag{1}$$

where $F(z_m)$ represents the feature vector at point z_m, s_m corresponds to the enclosed area with z_m and c'. Click continuous field can help eliminate the difference between continuous interaction space and discrete feature space. Click prototype will serve as an initial guide for subsequent prototype learning.

3.2 Spatial Prototype Learning

As shown in Figure 3, multi-scale features $\{E_i\}_{i=1}^3$ and $\{D_i\}_{i=1}^3$ can be extracted from the encoder and decoder, respectively, and features of adjacent levels have a resolution difference of 2 times. In existing methods, the skip connections are utilized to transmit low-level information from encoding layers to the corresponding decoding layers. However, this paper Wang et al. (2022) points out that skip connections are not always effective for the segmentation. For the interactive segmentation task, the feature learning should be guided by the user interactions. Therefore, we explore introducing user intents into the skip connections to control the information transmission from the encoder to decoder. Spatial prototypes are used as efficient communication carriers in this step. In this paper, we define spatial prototype as the clustering center of a local region. Image patches are selected to uniformly sample continuous local regions of the feature space. Positions of spatial prototypes are determined as the centers of patches, and characteristics of spatial prototypes are determined by all feature points within patches, where the global average pooling (GAP) operation is adopted to extract prototype vectors in this paper. We first unify the dimensions of features E_i to C (setting to 256 in this paper) with the 1×1 convolutions. Patches with fixed sizes $\{d, d/2, d/4\}$ are applied for each E_i, respectively, to generate spatial prototypes P_s with size $N_{sp} \times C$, where N_{sp} represents the number of spatial prototypes. Based on Equation 1, we can also obtain the initial click prototypes P_c with size $N_c \times C$, where N_c represents the number of clicks. It is worth noting here that additional N_b learnable vectors are expanded into P_c to model the background in the absence of background clicks. Then the positional encoding Vaswani et al. (2017) is imposed for each spatial prototype and each click ($C/2$ dimension for each horizontal and vertical axes), and is added to P_s and P_c respectively, to incorporate spatial prior guidance of prototypes.

Each P_s extracts spatial structure clues in different semantic levels of the encoder, and each P_c contains the descriptions of labels in their respective feature levels. We design learning strategies between each P_s and P_c under the guidance of user intents in the skip connections. Referring to existing Transformer-like structures Cheng et al. (2022); Vaswani et al. (2017), multi-layer learning is constructed and each Transformer layer $t \in \{1, \ldots, T\}$ consists of a cross-attention module, a self-attention module and a feedforward network. The similarities between spatial prototypes and click prototypes are measured first, and P_s is updated as:

$$P_s^t = \mathrm{softmax}\left(P_s^t P_c^\top\right) P_c + P_s^{t-1} \tag{2}$$

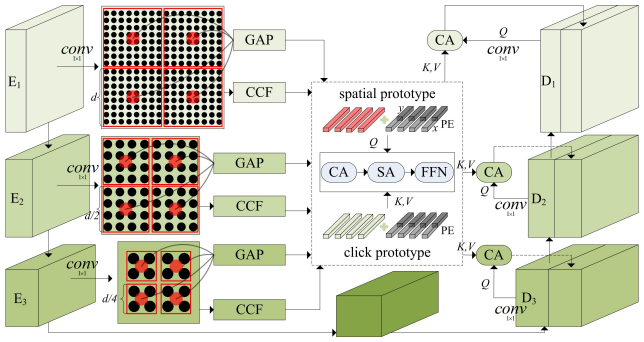

Figure 3: Detailed structure of spatial prototype learning, where CCF denotes the click continuous field, CA and SA denote cross- and self-attention, PE denotes positional encoding and FFN denotes a feedforward network.

where P_s^t denotes the spatial prototypes at layer t. Then the self-diffusion is utilized to capture the context among all spatial prototypes:

$$P_s^t = \mathrm{softmax}\big(P_s^t (P_s^t)^\top\big)\, P_s^t + P_s^t \tag{3}$$

Here, the 1×1 convolutions before cross- and self-attention are omitted. The operations in each layer are completed after a feedforward network with a residual connection Vaswani et al. (2017).

Let P_s^T denote the final feature of spatial prototypes after all Transformer layers. We transmit P_s^T to the decoder feature D_i as:

$$D_i = \mathrm{concat}\big(D_i, \mathrm{softmax}\big(\mathrm{conv}_{1\times1}(D_i)(P_s^T)^\top\big)\, P_s^T\big) \tag{4}$$

where $\mathrm{conv}_{1\times1}$ projects D_i to C channels and concat denotes concatenation. Thus, prototype learning transmits click-label-guided features from the encoder to the decoder through skip connections, complementing low- and high-level features according to user intent.

3.3 Parallel Segmentation Querying via Progressive Label Learning

Fixed-channel methods use clicks primarily as spatial selectors and therefore insufficiently model object characteristics and intra-instance variance. We instead progressively refine coarse click prototypes into label prototypes through global interactions with spatial prototypes, enabling multi-component object representation and reducing sensitivity to sparse prompts. As shown in Figure 4, let D_0 denote the full-resolution decoder output in addition to $\{D_i\}_{i=1}^3$. We project $\{D_i\}_{i=0}^3$ to C channels and apply the patch sampling, continuous-field interpolation, and positional encoding described above to $\{D_i\}_{i=1}^3$, obtaining spatial prototypes $\bar{P}_s$ and click prototypes $\bar{P}_c$. At each level, we concatenate them as $\bar{P} = [\bar{P}_s; \bar{P}_c]$ and append N_b learnable background vectors at the starting level $i = 3$. Multiple Transformer layers are constructed for the global learning, and each layer $t \in \{1, \ldots, T\}$ consists of a self-attention module and a feedforward network:

$$\bar{P}^t = \mathrm{softmax}\big(\bar{P}^t (\bar{P}^t)^\top\big)\, \bar{P}^t + \bar{P}^{t-1} \tag{5}$$

$$\bar{P}^t = \mathrm{FFN}\big(\bar{P}^t\big) + \bar{P}^t \tag{6}$$

Here, the 1×1 convolution before self-attention is omitted. FFN denotes a feedforward network Vaswani et al. (2017). From the simple self-attention step, these two prototypes can learn from each other, where click prototypes are gradually refined towards the abstracted labels based on the global interaction with spatial prototypes, and spatial prototypes are also updated layer-by-layer under the label guidance.

Between adjacent feature levels, the progressive constraint is also imposed:

$$\bar{P}_i = \bar{P}_i + \bar{P}_{i+1}^T \tag{7}$$

where $\bar{P}_{i+1}^T$ denotes the output of $i+1$ level. After the above cascade learning across multi-scale features in the decoder, we can obtain the final click prototypes $\widehat{P}_c \in \mathbb{R}^{(N_c+N_b)\times C}$ and the learned spatial prototypes $\widehat{P}_s \in \mathbb{R}^{N_{sp}\times C}$.

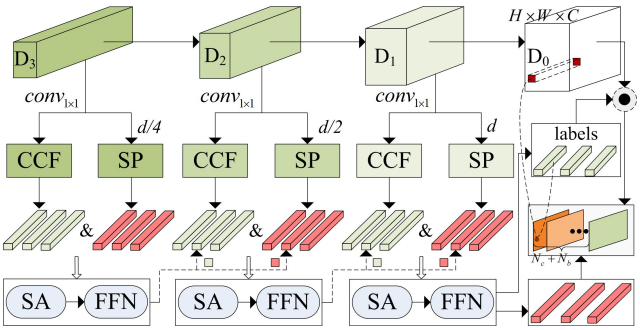

Figure 4: Detailed structure of label prototype learning, where CCF denotes the click continuous field, and SP denotes the spatial prototype.

Each $\widehat{P}_c^j \in \mathbb{R}^{1\times C}$ represents an object component with label $y_j = \ell \in \mathbb{L}$. All vectors with the same label $\mathcal{P}_\ell = \{\widehat{P}_c^j \mid y_j = \ell\}$ comprise a complete object together, effectively capturing the intra-class variance. Unlike the fixed-channel segmentation heads in existing methods, we can query segmentation labels between $\widehat{P}_c$ and D_0 in parallel regardless of the number of objects. For a pixel $p \in I$, the probability of belonging to object ℓ is:

$$\mathbb{P}_{p\ell} = \max_{\widehat{P}_c^j \in \mathcal{P}_\ell} \text{sigmoid}\Big(D_0(p) \cdot \widehat{P}_c^j\Big) \tag{8}$$

where $D_0(p) \in \mathbb{R}^{1\times C}$ is the feature of p. It can be noticed $\mathbb{P}_{p\ell}$ is determined as the similarity with the closest component in the object, and the final label is $\arg\max_\ell \mathbb{P}_{p\ell}$.

3.4 Click-aware Supervision via Temporal Hard-region Measurement

Interactive segmentation iteratively corrects errors through user clicks. Existing methods simulate this process with iterative training Sofiiuk et al. (2022). Normalized focus loss (NFL) Sofiiuk et al. (2019) is commonly used for its stability and convergence, and recent work improves its iterative and adaptive behavior Sun et al. (2024); Lin et al. (2025).Different from these works, from the observation that click prompts actually also imply hard segmentation regions during the iterative training, we develop a click-aware loss function based on temporal hard-region measurement.

We focus on the maximum mis-segmentation area in each iteration during training. Let K be the number of simulated corrections before backpropagation, producing maximum error regions $\{\mathcal{O}_k\}_{k=1}^K$ with ground-truth labels y_k. The segmentation difficulty of each region is jointly determined by the following two aspects: temporal recalcitrance and error magnitude, where the former means that regions still being erroneous despite an increasing number of corrective clicks attest to greater segmentation difficulty, and the latter indicates that regions exhibiting a higher inherent predictive error attest to greater difficulty. Therefore, we measure the level of difficulty for each region $\mathcal{O}_k$ as follows:

$$\omega_k = \left(1 + \lambda_t \frac{k}{K}\right)(1 + \lambda_e E_k) \tag{9}$$

$$E_k = \frac{1}{|\mathcal{O}_k|} \sum_{p\in\mathcal{O}_k} (1 - \mathbb{P}_{py_k}) \tag{10}$$

Here, $\lambda_t = 1$ and $\lambda_e = 2$ balance the two factors, $\mathbb{P}_{py_k}$ is the probability that p belongs to y_k, and $|\mathcal{O}_k|$ is the region size. Thus, ω_k increases for later or more erroneous regions, and the temporal hard-region loss is:

$$\mathcal{L}_{THR} = -\sum_{k=1}^{K} \omega_k \frac{1}{|\mathcal{O}_k|} \sum_{p\in\mathcal{O}_k} \log \mathbb{P}_{py_k} \tag{11}$$

This loss emphasizes click-aware regions with greater segmentation difficulty.

To further enhance the compactness of intra-class representations and improve inter-class discrimination, we introduce a prototype-guided contrastive learning strategy. Instead of directly optimizing

arbitrary pairwise pixel relationships, we construct positive and negative samples according to the learned prototypes $\widehat{P}_c$ and $\widehat{P}_s$. Specifically, for each pixel embedding z_p, we first identify its nearest click prototype among those belonging to its ground-truth label in $\widehat{P}_c$. The positive set $\mathcal{S}_p^+$ then consists of pixels with the same ground-truth label and the same nearest click prototype as z_p, together with spatial prototypes in $\widehat{P}_s$ whose nearest click prototype is identical to that of z_p. All remaining pixel embeddings and spatial prototypes are regarded as the negative set $\mathcal{S}_p^-$. Based on these prototype-guided sample sets, the contrastive loss is formulated as:

$$\mathcal{L}_C = \sum_p \mathrm{InfoNCE}\big(z_p, \mathcal{S}_p^+, \mathcal{S}_p^-\big) \tag{12}$$

where $\mathrm{InfoNCE}(\cdot)$ denotes the standard InfoNCE objective Chen et al. (2020). In summary, the loss can be formulated as follows:

$$\mathcal{L} = \mathcal{L}_{NFL} + \alpha\mathcal{L}_{THR} + \beta\mathcal{L}_C \tag{13}$$

where $\mathcal{L}_{NFL}$ is the normalized focus loss and α, β are balancing parameters.

4 Experimental Results

Datasets: We evaluate on GrabCut Rother et al. (2004) (50 single-object images), Berkeley Martin et al. (2001) (96 images and 100 instance masks), DAVIS Perazzi et al. (2016) (345 sampled frames from 50 videos, following Sofiiuk et al. (2022)), and the SBD validation split Hariharan et al. (2011) (8,498/2,857 training/validation images and 20,172/6,671 instances). Training uses COCO+LVIS Sofiiuk et al. (2022), with roughly 118,000 COCO training images ($\approx$ 1.2M instances) and denser, fine-grained LVIS masks.

Evaluation Metrics: For single-object segmentation, following previous works Du et al. (2023); Lee et al. (2024); Sofiiuk et al. (2022), the required number of clicks (NoC) to achieve a specific segmentation accuracy (such as the intersection-over-union, IoU) is used to evaluate the algorithm performance. For example, NoC@90 represents the NoC index to achieve 0.90 IoU accuracy. For multi-object segmentation, following Rana et al. (2023) the Normalized Clicks per Image (NCI) metric is adopted for evaluation using an IoU threshold of 85%. Based on NoC, NCI aggregates all clicks at the image level and normalizes by the number of instances, offering a consistent measure of multi-object segmentation efficiency.

Implementation Details: The size of patch d for prototype sampling is set to 8, the number of layers T for prototype learning is set to 3, and the number of learnable background vectors N_b is set to 3. Training is performed end-to-end on COCO + LVIS datasets Sofiiuk et al. (2022) using the AdamW optimizer with a 448×448 input resolution on a single NVIDIA RTX 4090 GPU. The coefficients α and β in Eq. equation 13 are set to 0.1 and 0.01. We follow the iterative training protocol of Sofiiuk et al. (2022), allowing up to three refinement iterations. The maximum number of clicks during testing is capped at 20 for all compared methods.

The proposed method is compared with state-of-the-art deep interactive approaches CDNet Chen et al. (2021), RITM Sofiiuk et al. (2022), EdgeFlow Hao et al. (2021), EMC-Click Du et al. (2023), FCF Wei et al. (2023), SimpleClick Liu et al. (2023), SAM Kirillov et al. (2023), InterFormer Huang et al. (2023), CCF-BC Lin et al. (2023), FocalClick Chen et al. (2022) and DynaMITe Rana et al. (2023). Among them, RITM is a leading CNN-based method known for its strong segmentation performance, while SimpleClick and SAM represent Transformer-based approaches, with SAM offering high generalization through large-scale training, and DynaMITe is a representative method of interactive multi-object segmentation. We retrain DynaMITe under the same experimental settings as ours and all the methods are trained on the same COCO + LVIS datasets for a fair comparison. For interactive single-object segmentation, Table 1 lists the average NoC values of the compared methods on GrabCut, Berkeley, DAVIS and SBD datasets. Based on ResNet and HRNet backbones, our method achieves significant performance improvement compared with the CNN baseline RITM, effectively verifying the role of prototype selection in interactive segmentation task. Compared with other CNN-based approaches, our method obtains competitive results on small-scale datasets GrabCut and Berkeley, while exhibiting the best performance on the challenging large-scale datasets DAVIS and SBD, especially for the high-quality segmentation (NoC@90). Based on

Table 1: NoC and NCI values of the state-of-the-art approaches for interactive single-object and multi-object segmentation tasks, where all methods are trained on the same COCO + LVIS datasets.

Single-object	Backbone	GrabCut	Berkeley	DAVIS		SBD	
Segmentation		NoC@90	NoC@90	NoC@85	NoC@90	NoC@85	NoC@90
RITM	ResNet-34	2.54	4.58	5.69	9.07	4.16	6.57
CDNet	ResNet-34	**1.52**	**2.06**	4.27	5.56	4.30	7.04
Ours	ResNet-34	1.85	2.13	**3.96**	**5.40**	**3.60**	**5.74**
RITM	HRNet-18	1.54	2.26	4.36	5.74	3.80	6.06
EdgeFlow	HRNet-18	1.72	2.40	4.54	5.77	–	–
EMC-Click	HRNet-18	1.50	2.30	4.34	5.59	3.69	5.93
FCF	HRNet-18	**1.46**	**1.96**	4.05	5.58	3.63	5.83
Ours	HRNet-18	1.98	2.41	**3.84**	**5.37**	**3.55**	**5.65**
SimpleClick	ViT-B	1.48	1.97	3.66	5.06	3.43	5.62
SAM	ViT-B	2.72	2.96	6.13	7.88	7.22	11.05
InterFormer	ViT-B	1.50	3.14	4.10	6.19	3.78	6.34
CCF-BC	ViT-B	**1.43**	1.89	3.58	**4.88**	**3.14**	5.57
Ours	ViT-B	1.64	**1.86**	**3.43**	4.95	3.33	**5.35**

Multi-object	Backbone	SBD		COCO		DAVIS	
Segmentation		NCI ↓	IoU ↑	NCI ↓	IoU ↑	NCI ↓	IoU ↑
FocalClick	Segformer-B0	4.81	83.40	7.96	59.30	**5.20**	**82.40**
DynaMITe	Swin-T	4.79	87.15	**7.59**	78.27	6.06	80.28
Ours	ViT-B	**4.30**	**87.75**	7.82	**80.79**	5.26	81.63

ViT-B backbone, our method still obtains significant improvement compared with the Transformer baseline SimpleClick. From all the compared methods, the proposed method produces the lowest NoC values in most cases, demonstrating stable interactive segmentation performance with lower interaction burden. For interactive multi-object segmentation, Table 1 lists the average NCI and image level IoU values achieved after segmenting all the objects in an image on SBD, COCO and DAVIS datasets. Our method outperforms FocalClick and DynaMITe by a large margin on the large-scale SBD dataset and achieves the best segmentation performance overall across all datasets.

Figure 5 compares visual segmentations and IoU values under the same number of clicks (increasing from 1 to 3), where different methods start from the same first click selected as the object center, and subsequent clicks are automatically determined as the center of the current largest error area. We compare our method with RITM, EMC-Click and SimpleClick across network architectures. From the results in the third column, our method demonstrates stronger user intention perception and can accurately capture the complete object contour with just one click. However, sparse spatial guidance cannot help other methods obtain good results under extremely limited interactions. As the number of clicks increases, prototype learning can help to capture object characteristics and intra-variance, enabling our method to have excellent continuous correction capability. Comparatively, other methods are generally limited by challenges such as object occlusion, overlapping, and poor lighting, and cannot obtain satisfactory results based on limited user interaction. Our method produces high-quality segmentations with only three clicks regardless of the impact of these challenges, demonstrating better interactive performance.

We carry out plenty of ablation studies to verify the effectiveness of each component on the GrabCut, Berkeley, DAVIS and SBD datasets. Table 2 reports ablations of the proposed spatial prototype learning (SPL), label prototype learning (LPL), and click-aware supervision (CAS) strategies on the ResNet-34 backbone. By comparison, the NoC index values show a significant increase without all components (row 1) and the best performance (the lowest NoC index) is obtained with all the combined components (row 5) on all four datasets. The results from row 2 to row 4 indicate that the segmentation performance decreases to a certain extent when any of these components are missing, verifying the effectiveness of the proposed SPL, LPL and CAS modules.

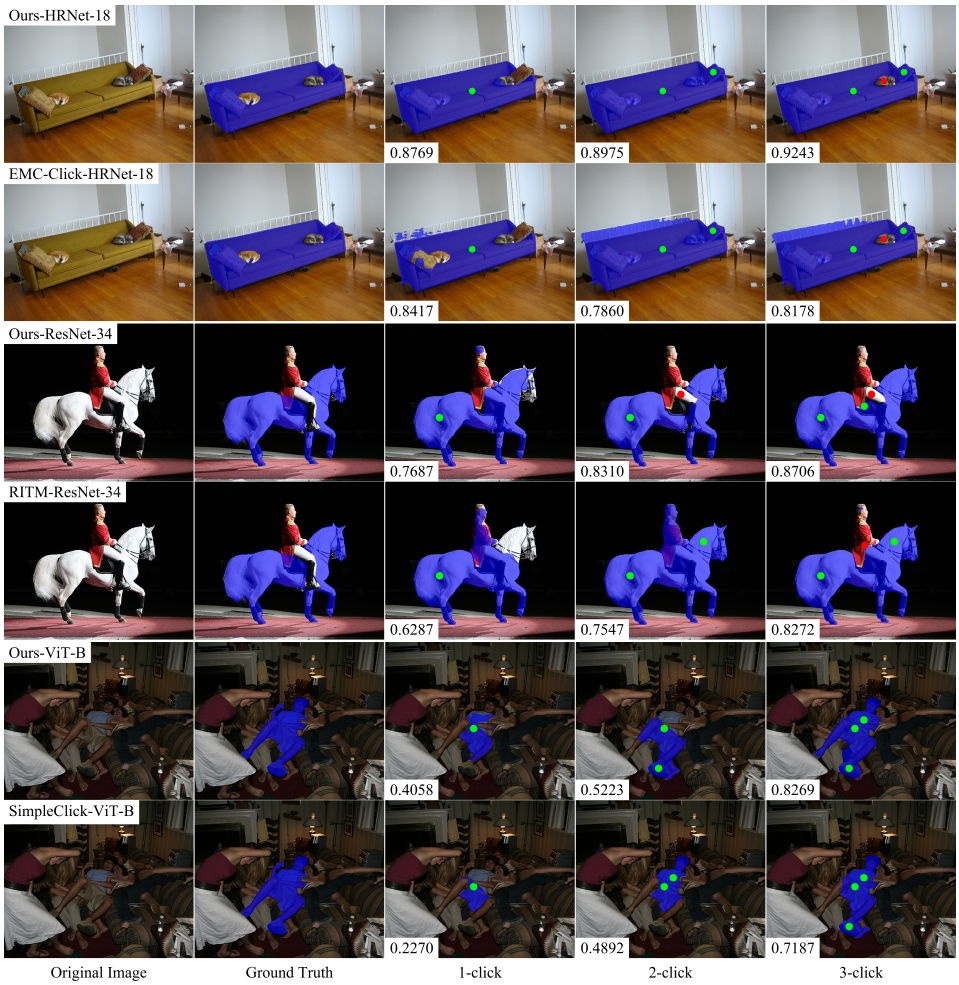

Figure 5: Comparison of continuous correction with gradually increasing clicks, where the first click position of different methods is the same and subsequent clicks are automatically determined as the center of the largest error area in the previous segmentation.

Table 2: Ablation studies for the proposed spatial prototype learning (SPL), label prototype learning (LPL), and click-aware supervision (CAS) on ResNet-34.

SPL	LPL	CAS	GrabCut		Berkeley		DAVIS		SBD	
			NoC@85	NoC@90	NoC@85	NoC@90	NoC@85	NoC@90	NoC@85	NoC@90
			2.06	2.54	2.61	4.58	5.69	9.07	4.16	6.57
	✓	✓	1.90	2.15	1.66	2.31	4.88	6.01	3.91	6.05
✓		✓	1.81	2.09	1.59	2.24	4.50	5.83	3.84	6.14
✓	✓		1.80	1.94	1.61	2.43	4.10	5.66	3.69	5.81
✓	✓	✓	1.72	1.85	1.49	2.13	3.96	5.40	3.60	5.74

5 Conclusion

We present a prototype learning framework for interactive segmentation that reinterprets user clicks as the construction of multiple prototype selections, each capturing both spatial and semantic cues to collectively represent an object. We introduce spatial prototype learning for intention-aware feature propagation and label learning to enhance object characterization and support parallel object querying. A difficulty-aware training strategy further supervises hard regions via click-aware constraints and inter-prototype relationships. Vast experiments verify the superior performance of our method.

AI USE STATEMENT

In preparing this manuscript, generative AI tools were used only to aid or polish the writing. The authors reviewed all AI-assisted text and take responsibility for the final content.

REFERENCES

Rodrigo Benenson, Stefan Popov, and Vittorio Ferrari. Large-scale interactive object segmentation with human annotators. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11700–11709, 2019.

Yuri Y Boykov and M-P Jolly. Interactive graph cuts for optimal boundary & region segmentation of objects in nd images. In *Proceedings of IEEE International Conference on Computer Vision*, pp. 105–112, 2001.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the International Conference on Machine Learning*, pp. 1597–1607, 2020.

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*, pp. 7345–7354, 2021.

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*, pp. 1300–1309, 2022.

Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1290–1299, 2022.

Nanqing Dong and Eric P Xing. Few-shot semantic segmentation with prototype learning. In *Proceedings of the British Machine Vision Conference*, pp. 4, 2018.

Fei Du, Jianlong Yuan, Zhibin Wang, and Fan Wang. Efficient mask correction for click-based interactive image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22773–22782, 2023.

Richard O Duda and Peter E Hart. *Pattern Classification and Scene Analysis*. A Wiley-Interscience Publication, 1973.

Salvador Garcia, Joaquin Derrac, Jose Cano, and Francisco Herrera. Prototype selection for nearest neighbor classification: Taxonomy and empirical study. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(3):417–435, 2012.

Leo Grady. Random walks for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(11):1768–1783, 2006.

Yuying Hao, Yi Liu, Zewu Wu, Lin Han, Yizhou Chen, Guowei Chen, Lutao Chu, Shiyu Tang, Zhiliang Yu, Zeyu Chen, et al. Edgeflow: Achieving practical interactive segmentation with edge-guided flow. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1551–1560, 2021.

Bharath Hariharan, Pablo Arbeláez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 991–998, 2011.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Hanzhe Hu, Yinbo Chen, Jiarui Xu, Shubhankar Borse, Hong Cai, Fatih Porikli, and Xiaolong Wang. Learning implicit feature alignment function for semantic segmentation. In *Proceedings of the European Conference on Computer Vision*, pp. 487–505, 2022.

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*, pp. 22301–22311, 2023.

Won-Dong Jang and Chang-Su Kim. Interactive image segmentation via backpropagating refinement scheme. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5297–5306, 2019.

Lei Ke, Mingqiao Ye, Martin Danelljan, Yu-Wing Tai, Chi-Keung Tang, Fisher Yu, et al. Segment anything in high quality. In *Proceedings of the Advances in Neural Information Processing Systems*, pp. 29914–29934, 2023.

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*, pp. 4015–4026, 2023.

Chaewon Lee, Seon-Ho Lee, and Chang-Su Kim. Mfp: Making full use of probability maps for interactive image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4051–4059, 2024.

Chunming Li, Chenyang Xu, Changfeng Gui, and Martin D Fox. Distance regularized level set evolution and its application to image segmentation. *IEEE Transactions on Image Processing*, 19 (12):3243–3254, 2010.

Kun Li, Hao Cheng, George Vosselman, and Michael Ying Yang. Learning from exemplars for interactive image segmentation. *arXiv preprint arXiv:2406.11472*, 2024.

Jiacheng Lin, Zhiqiang Xiao, Xiaohui Wei, Puhong Duan, Xuan He, Renwei Dian, Zhiyong Li, and Shutao Li. Click-pixel cognition fusion network with balanced cut for interactive image segmentation. *IEEE Transactions on Image Processing*, 33:177–190, 2023.

Jiacheng Lin, Jiajun Chen, Kailun Yang, Alina Roitberg, Siyu Li, Zhiyong Li, and Shutao Li. Adaptiveclick: Click-aware transformer with adaptive focal loss for interactive image segmentation. *IEEE Transactions on Neural Networks and Learning Systems*, 36(3):5759–5773, 2025.

Zheng Lin, Zhao Zhang, Lin-Zhuo Chen, Ming-Ming Cheng, and Shao-Ping Lu. Interactive image segmentation with first click attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 13339–13348, 2020.

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*, pp. 22290–22300, 2023.

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*, pp. 3773–3782, 2024.

Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3431–3440, 2015.

Soumajit Majumder and Angela Yao. Content-aware multi-level guidance for interactive instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11602–11611, 2019.

David Martin, Charless Fowlkes, Doron Tal, and Jitendra Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 416–423, 2001.

Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 724–732, 2016.

Amit Kumar Rana, Sabarinath Mahadevan, Alexander Hermans, and Bastian Leibe. Dynamite: Dynamic query bootstrapping for multi-object interactive segmentation transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1043–1052, 2023.

Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. " grabcut" interactive foreground extraction using iterated graph cuts. *ACM Transactions on Graphics*, 23(3):309–314, 2004.

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Proceedings of the Advances in Neural Information Processing Systems*, pp. 30, 2017.

Konstantin Sofiiuk, Olga Barinova, and Anton Konushin. Adaptis: Adaptive instance selection network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7355–7356, 2019.

Konstantin Sofiiuk, Ilia Petrov, Olga Barinova, and Anton Konushin. f-brs: Rethinking backpropagating refinement for interactive segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8623–8632, 2020.

Konstantin Sofiiuk, Ilya A Petrov, and Anton Konushin. Reviving iterative training with mask guidance for interactive segmentation. In *Proceedings of the IEEE International Conference on Image Processing*, pp. 3141–3145, 2022.

Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5693–5703, 2019.

Shoukun Sun, Min Xian, Fei Xu, Luca Capriotti, and Tiankai Yao. Cfr-icl: Cascade-forward refinement with iterative click loss for interactive image segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 5017–5024, 2024.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 30, 2017.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 29, 2016.

Haonan Wang, Peng Cao, Jiaqi Wang, and Osmar R Zaiane. Uctransnet: rethinking the skip connections in u-net from a channel-wise perspective with transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 2441–2449, 2022.

Tao Wang, Jian Yang, Zexuan Ji, and Quansen Sun. Probabilistic diffusion for interactive image segmentation. *IEEE Transactions on Image Processing*, 28(1):330–342, 2019.

Tao Wang, Haochen Li, Yuhui Zheng, and Quansen Sun. One-click-based perception for interactive image segmentation. *IEEE Transactions on Neural Networks and Learning Systems*, 35(10): 13975–13989, 2023.

Qiaoqiao Wei, Hui Zhang, and Jun-Hai Yong. Focused and collaborative feedback integration for interactive image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18643–18652, 2023.

Hallee E Wong, Marianne Rakic, John Guttag, and Adrian V Dalca. Scribbleprompt: fast and flexible interactive segmentation for any biomedical image. In *Proceedings of the European Conference on Computer Vision*, pp. 207–229, 2024.

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*, pp. 373–381, 2016.

Chaoning Zhang, Dongshen Han, Yu Qiao, Jung Uk Kim, Sung-Ho Bae, Seungkyu Lee, and Choong Seon Hong. Faster segment anything: Towards lightweight sam for mobile applications. *arXiv preprint arXiv:2306.14289*, 2023.

Yian Zhao, Kehan Li, Zesen Cheng, Pengchong Qiao, Xiawu Zheng, Rongrong Ji, Chang Liu, Li Yuan, and Jie Chen. Graco: Granularity-controllable interactive segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3501–3510, 2024.

Tianfei Zhou, Wenguan Wang, Ender Konukoglu, and Luc Van Gool. Rethinking semantic segmentation: A prototype view. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2582–2593, 2022.

APPENDIX

A EXPERIMENTS ON SMALLER SBD DATASET

We further evaluate our method on a smaller SBD dataset training setting to examine its robustness under limited data. As shown in Table 3, our approach consistently requires fewer clicks than both CNN and Transformer baselines, demonstrating more efficient interactive segmentation. This improvement can be attributed to our prototype-driven framework: spatial prototypes effectively guide feature propagation according to user clicks, while progressive label prototypes capture object characteristics and intra-class variance. In addition, the click-aware supervision emphasizes hard regions, enabling the network to focus on challenging areas even with sparse interactions. Together, these components allow our method to achieve stable, high-quality segmentation with minimal user interaction, highlighting the strength of prototype learning in low-data regimes.

Table 3: Average NoC values of the state-of-the-art approaches on GrabCut, Berkeley, DAVIS, and SBD datasets, where all methods are trained on the same SBD dataset.

Method	Backbone	GrabCut	Berkeley	DAVIS		SBD	
		NoC@90	NoC@90	NoC@85	NoC@90	NoC@85	NoC@90
RITM	ResNet-34	2.57	4.69	5.71	9.52	4.12	6.46
CDNet	ResNet-34	2.18	3.27	5.00	6.89	5.18	7.89
Ours	ResNet-34	**1.88**	**2.87**	**4.97**	**6.75**	**3.47**	**5.46**
RITM	HRNet-18	2.04	3.22	4.94	6.71	**3.39**	5.43
EMC-Click	HRNet-18	**1.84**	3.30	5.05	6.71	3.38	5.51
Ours	HRNet-18	2.20	**2.82**	**4.73**	**6.06**	3.42	**5.36**
SimpleClick	ViT-B	**1.52**	3.17	4.57	6.77	3.78	5.97
CCF-BC	ViT-B	1.88	2.91	4.75	6.02	3.42	5.38
Ours	ViT-B	1.68	**2.37**	**4.15**	**5.73**	**3.31**	**5.31**

B VISUALIZATION ANALYSIS OF SPATIAL AND LABEL PROTOTYPE LEARNING

To further analyze the behaviors of the proposed prototype learning framework, we visualize the probability maps obtained by computing the inner product between prototypes and the dense feature map (D_0).

For spatial prototype learning, as shown in Figure 6, the probability maps generated by spatial prototypes exhibit highly consistent activation patterns and largely coincide with the object mask. This observation aligns with the design of spatial prototype learning, where spatial prototypes are constructed by aggregating local features and guided by click prototypes to propagate user intent through skip connections. As a result, spatial prototypes mainly encode object-level spatial cues rather than component-level appearance variations, and their responses highlight the overall object region, providing stable spatial guidance for the decoder during segmentation.

In contrast, label prototype learning produces more diverse response patterns. By computing the inner product between the learned label prototypes (LP) and the feature map (D_0), we obtain probability maps that correspond to different object components. Segmentation prediction is achieved by measuring the similarity between pixel features and label prototypes. As shown in Figure 7, different label prototypes focus on distinct object parts rather than the entire object. These complementary responses jointly cover the object structure, demonstrating that progressive label prototype learning effectively captures.

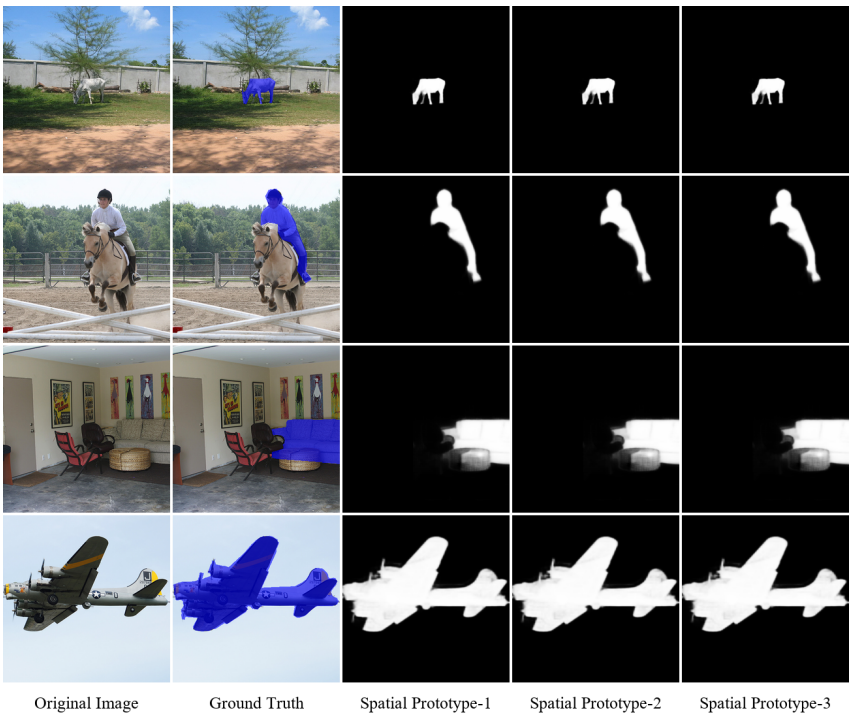

Figure 6: Visualization of probability maps generated by randomly selected spatial prototypes. From left to right: original image, ground truth, and probability maps generated by three spatial prototypes.

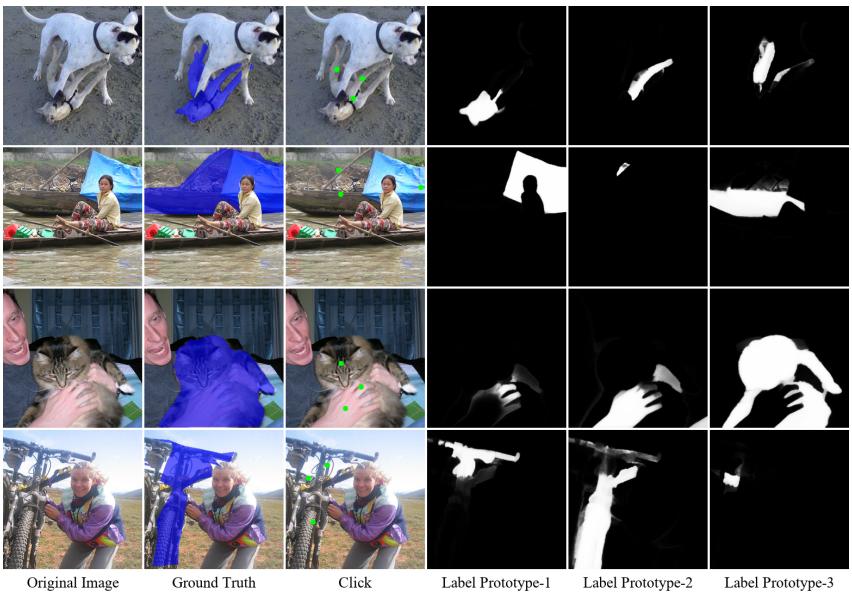

Figure 7: Visualization of probability maps generated by different label prototypes. From left to right: original image, ground truth, user clicks, and the probability maps produced by three label prototypes.

C MORE QUALITATIVE RESULTS

We also provide more segmentation results of our method on the 4 datasets. Figure 8 shows the mean IoU (mIoU) curves with the varying of the number of clicks on DAVIS and SBD datasets. It can be seen that better performance can be obtained by all the methods as the number of clicks increases. Comparing to SimpleClick, RITM, EMC-Click and FCF, the curves of our method have faster ascent speeds and higher convergence values. Our method obtains higher mIoU ratios with the same number of clicks than the conventional algorithms. Figure 9 shows the histogram analysis of quantity ratio (%) with different IoU accuracies under a fixed number of clicks $n \in \{1, 3, 20\}$ (denoted by IoU@n) on DAVIS and SBD datasets. As the number of clicks increases, the proportion of correctly segmented images (with higher IoU) in each method gradually increases. By comparison, our methods achieve better performance with fewer failure cases than other methods in all n values.

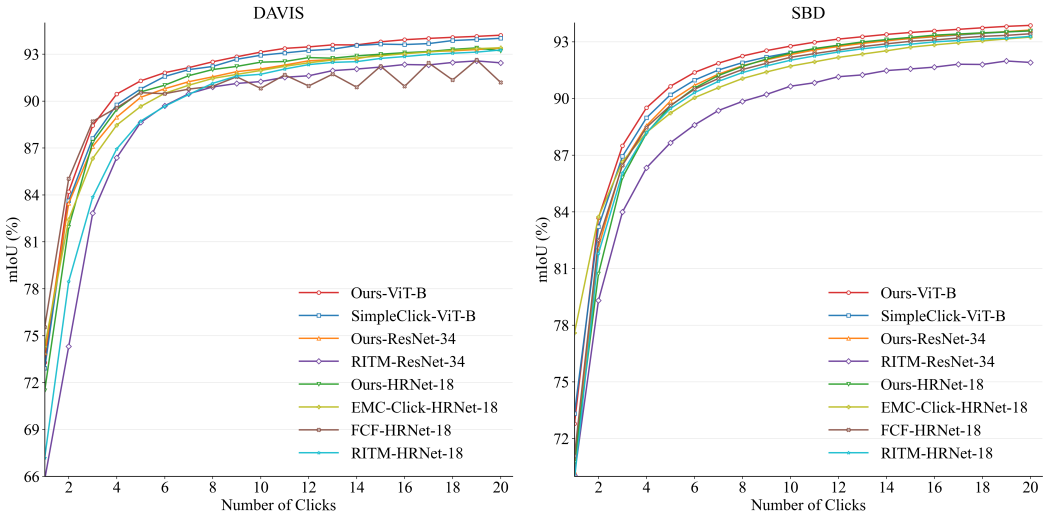

Figure 8: NoC vs. mean IoU (mIoU) accuracy of the compared methods.

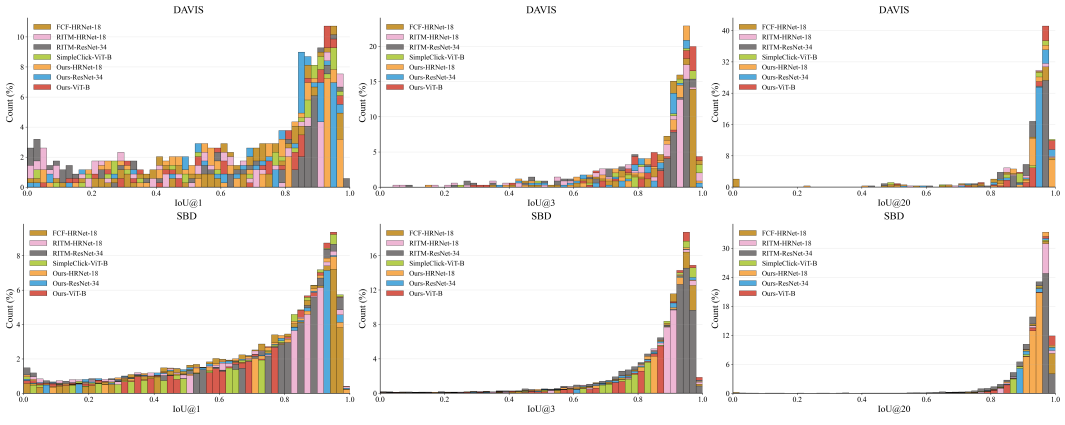

Figure 9: Histogram analysis (quantity ratio) of IoU under a fixed number of clicks $n \in \{1, 3, 20\}$ (denoted by IoU@n) on DAVIS and SBD datasets.

Figure 10 shows more comparison of visual segmentations and IoU values among different methods with up to three clicks. Figure 11 shows the common cases from GrabCut and Berkeley, and Figure 12 represents common cases from SBD and DAVIS. As shown in Figure 11, on the GrabCut and Berkeley datasets, our method achieves high-quality segmentation for single objects with a single click, often exceeding 90% accuracy. The predicted masks closely align with the ground truth, even in challenging backgrounds, demonstrating robustness and precise boundary localization. In

contrast, Figure 12 illustrates results on the more complex SBD and DAVIS datasets. For images with multiple objects, occlusions, or intricate boundaries, initial one-click predictions are less accurate, but additional clicks quickly improve performance, often reaching over 90% with only a few interactions.

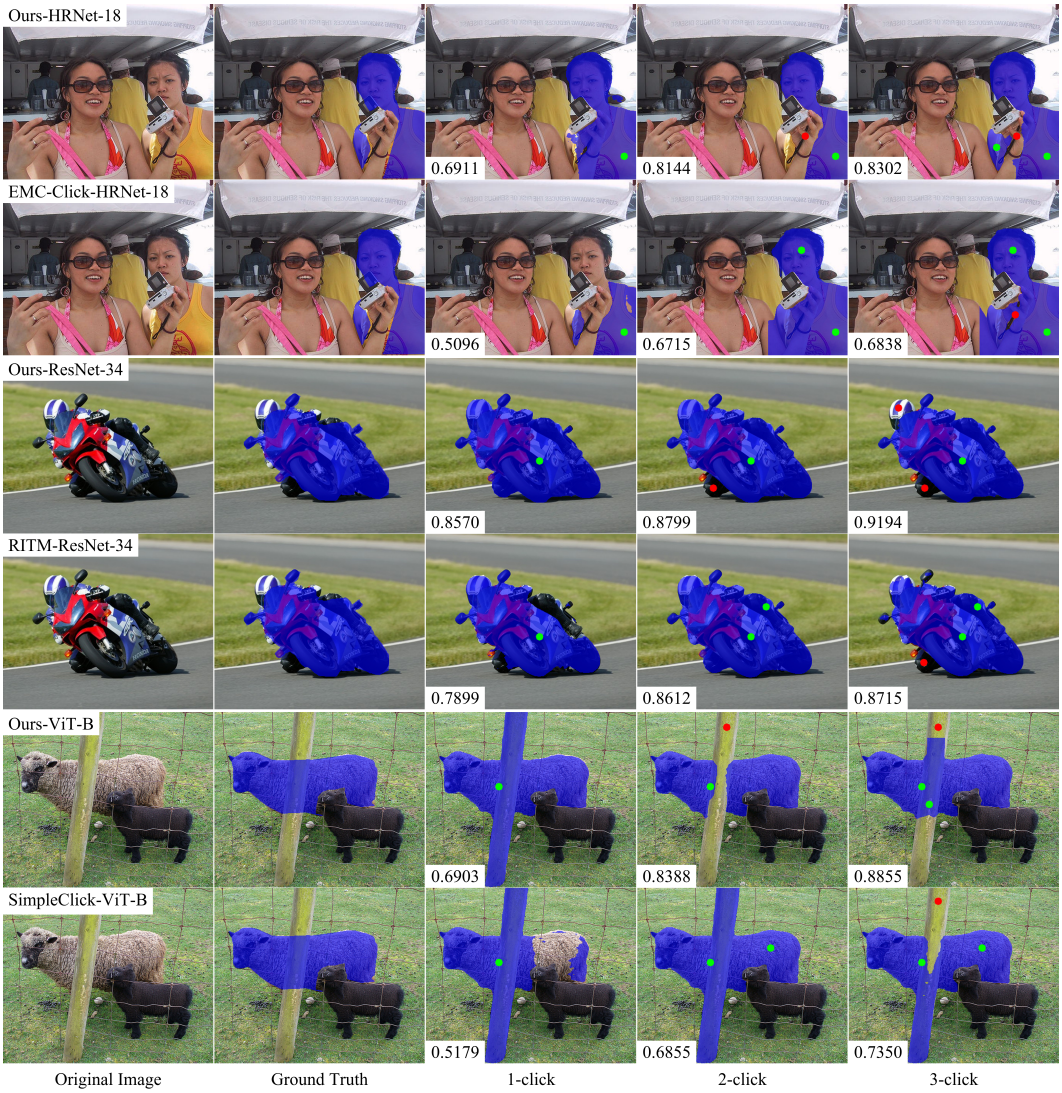

Figure 10: More comparison of continuous correction with gradually increasing clicks among different methods.

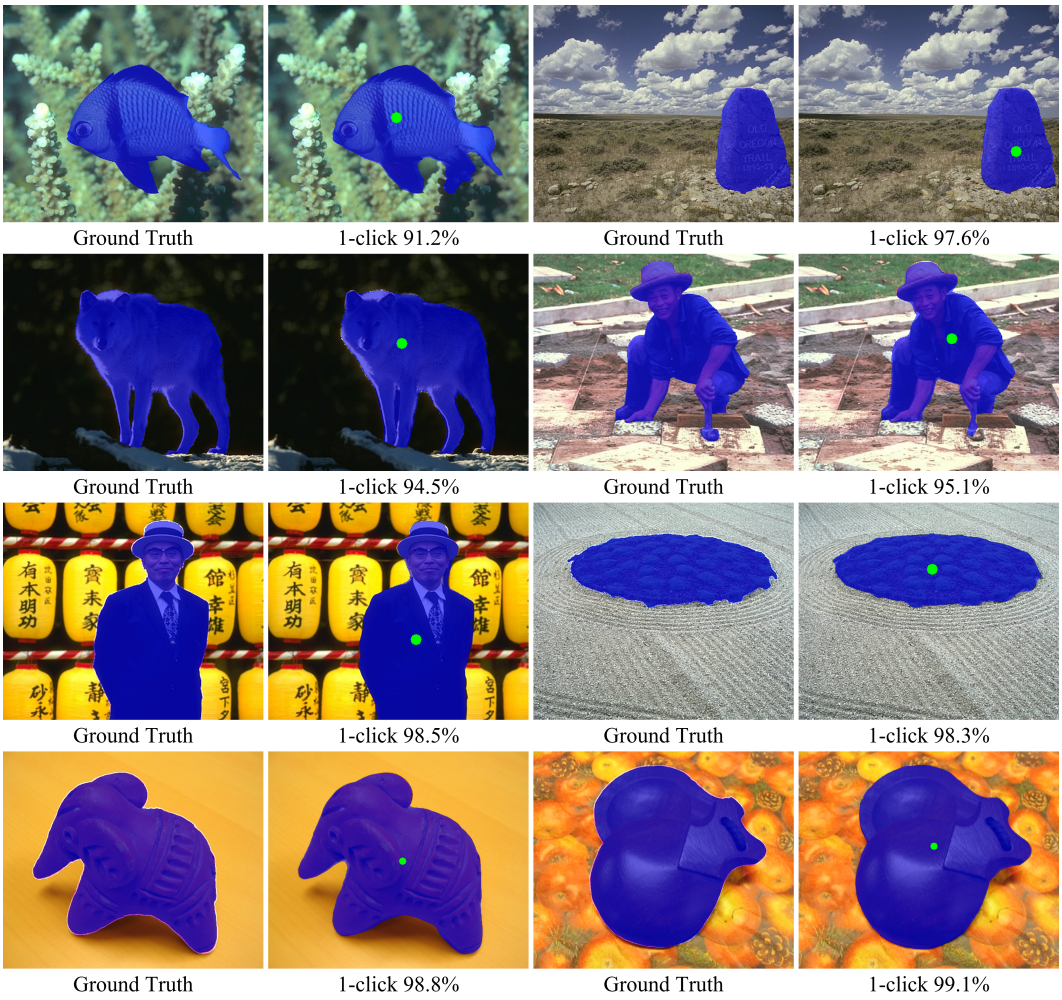

Figure 11: More visualizations of the segmentation results from Berkeley (Row 1-2) and GrabCut (Row 3-5).

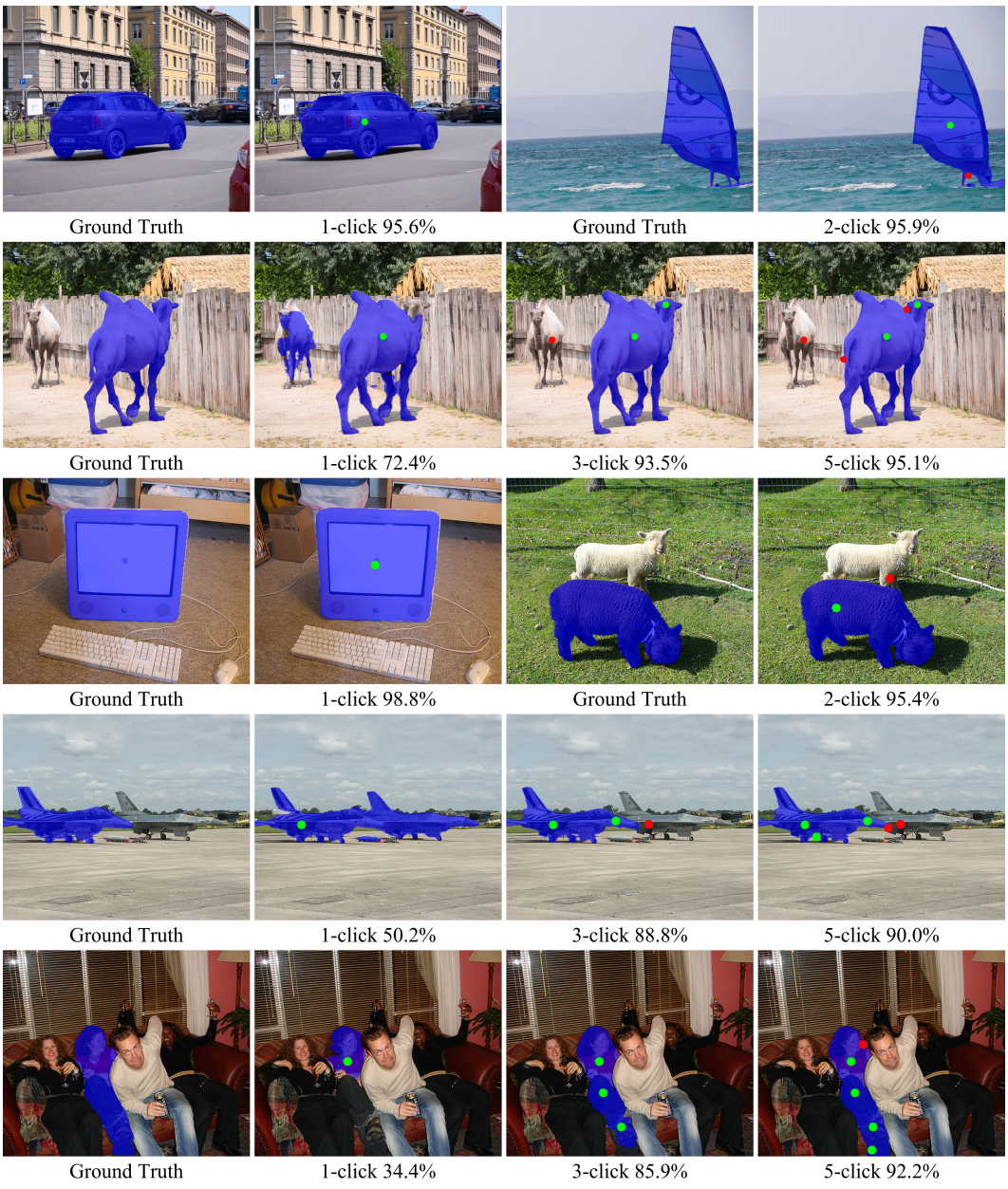

Figure 12: More visualizations of the segmentation results from DAVIS (Row 1-2) and SBD (Row 3-5).