Click Prompt Learning with Optimal Transport for Interactive Segmentation

Jie Liu[1], Haochen Wang[1], Wenzhe Yin[1], Jan-Jakob Sonke[2], Efstratios Gavves[1]

[1]University of Amsterdam [2]The Netherlands Cancer Institute
[1]{j.liu5, h.wang3, w.yin, E.Gavves}@uva.nl [2]j.sonke@nki.nl

Abstract. Click-based interactive segmentation aims to segment target objects conditioned on user-provided clicks. Existing methods typically interpret user intention by learning multiple click prompts to generate corresponding prompt-activated masks, and selecting one from these masks. However, directly matching each prompt to the same visual feature often leads to homogeneous prompt-activated masks, as it pushes the click prompts to converge to one point. To address this problem, we propose Click Prompt Learning with Optimal Transport (CPlot), which leverages optimal transport theory to capture diverse user intentions with multiple click prompts. Specifically, we first introduce a prompt-pixel alignment module (PPAM), which aligns each click prompts with the visual features in the same feature space by plain transformer blocks. In such way, PPAM enables all click prompts to encode more general knowledge about regions of interest, indicating a consistent user intention. To capture diverse user intentions, we further propose the click prompt optimal transport module (CPOT) to match click prompts and visual features. CPOT is designed to learn an optimal mapping between click prompts and visual features. Such unique mapping facilities click prompts to effectively focus on distinct visual regions, which reflect underlying diverse user intentions. Furthermore, CPlot learns click prompts with a two-stage optimization strategy: the inner loop optimizes the optimal transport distance to align visual features with click prompts through the Sinkhorn algorithm, while the outer loop adjusts the click prompts from the supervised data. Extensive experiments on eight interactive segmentation benchmarks demonstrate the superiority of our method for interactive segmentation. Project page: https://jliu4ai.github.io/cplot_projectpage/.

1 Introduction

Interactive image segmentation seeks to generate high-quality masks with user interaction as guidance, leading to various practical applications, such as image editing [5, 19] and medical image analysis [18, 20]. Typically, users can interact with segmentation models through various representations, including clicks [47], scribbles [52], polygons [1], bounding box [54], and their combinations [57]. Among these methods, click-based interactive segmentation is particularly notable for its simplicity and well-established click simulation strategies.

Fig. 1: Comparison between our model without and with Click Prompt Optimal Transport (CPOT). (a) Without CPOT, all click prompts tend to converge to one point, resulting in homogeneous prompt-activated masks and inferior mask prediction. (b) With the proposed CPOT, click prompts are encouraged to focus on distinct visual regions. Consequently, our model with CPOT predicts a more accurate mask by integrating diverse prompt-activated masks.

In click-based interactive segmentation, the users successively place positive/negative clicks to specify the foreground and background. The basic paradigm [47] adopts Gaussian maps or disk maps [56] to represent clicks, then concatenates these maps with the input image, and sends it to the segmentation model to predict the mask. Based on such paradigm, various works have been proposed to simultaneously improve segmentation quality and reduce interaction rounds, such as local refinement [9, 30], click attention [31], pseudo clicks [34], and model adaptation [21, 25, 46]. Despite notable advancements, interpreting user intention with a few clicks remains a significant challenge in interactive segmentation [27, 51]. For instance, when a user clicks on a man's jacket, it remains unclear whether the intention is to select the jacket or the man.

To mitigate the intention ambiguity problem, some works [24, 27, 28] have demonstrated that diverse intermediate masks provide robust priors to infer the real intention of input clicks. These methods usually employ multiple click prompts or subnetworks to predict multiple prompt-activated masks and select one of these mask. For instance, SAM [24] proposes to generate multiple masks and confidence scores from click prompts, and subsequently the mask with highest confidence score is selected as final mask. Despite impressive performance, as illustrated in Fig. 1 (a), directly matching click prompts with visual features often yields homogeneous prompt-activated masks, thus capturing limited diversity of user intention and resulting in suboptimal mask predictions. We suspect that this phenomenon arises from the absence of implicit constraints in the click prompt learning process, which inadvertently pushes click prompts converge toward the same point.

To address the above issue, we propose Click Prompt Learning with Optimal Transport (CPlot), which captures diverse user intentions by applying optimal transport (OT) to regularize the learning of multiple click prompts. Specifically, we first introduce a *prompt-pixel alignment module* (PPAM) to achieve a coarse alignment between visual features and multiple click prompts in the same feature space with plain transformer blocks. Consequently, as shown in Fig. 1 (a), all click prompts are encouraged to encode more knowledge about regions of interest, which roughly indicate a consistent user intention. To capture diverse user

intention with click prompts, we further propose a novel *click prompt optimal transport module* (CPOT). CPOT formulates visual features and click prompts as the samplings of two discrete distributions and adopts OT to facilitate cross-modal matching. In such way, click prompts are pushed to focus on distinct visual regions, generating more diverse prompt-activated masks, as illustrated in Fig. 1 (b). Additionally, to reduce the computational cost and avoid the extra model parameters, we learn the click prompts with a two-stage optimization strategy. In the inner loop, we freeze the visual features and optimize the optimal transport problem by a fast Sinkhorn algorithm [10]. In the outer loop, we fix all parameters of optimal transport and backpropagate the gradient to learn click prompts from pixel-wise annotated data. By incorporating the proposed PPAM and CPOT, our method enable click prompts to generate diverse prompt-activated masks, thus capturing diverse user intentions and providing robust priors to infer the real user intention.

In a nutshell, our main contributions are threefold:

- We propose click prompt learning with optimal transport for interactive segmentation, leveraging the optimal transport theory to capture diverse user intentions with click prompts.
- We introduce two novel modules, prompt-pixel alignment module and click prompt optimal transport module, to enable click prompts to generate diverse prompt-activated masks and make accurate final mask prediction.
- Extensive experiments demonstrate that the proposed method achieves state-of-the-art performance on both nature and medical images datasets.

2 Related Work

Interactive Segmentation. Interactive segmentation aims to achieve high-quality segmentation by incorporating iterative user interaction as guidance. Early works [14, 15, 23, 43] primarily adopt optimization-based graphical models built on low-level image features. Since coarse mask from the last interaction round contains object information, RITM [47] and 99%AccuracyNet [12] feed previous mask to the network to achieve more accurate mask prediction. To better excavate click information, FCANet [31] emphasizes the critical role of the first click and designs a first-click attention for better segmentation, PseudoClick [34] proposes a click-imitation mechanism to generate pseudo next-click to reduce interactive rounds. From the perspective of model adaptation, some work, *e.g.*, BRS [21], f-BRS [46], and CA [25] adopt loss back-propagation or model parameter adaptation to perform mask refinement for testing images. Meanwhile, FocalClick [9] and FocusCut [30] adopt similar local refinement strategies to correct desired local regions for high-quality segmentation. Recently, iSegFormer [33] and SimpleClick [32] apply vision transformer to the interactive segmentation and show promising performance. GPCIS [61] introduces Gaussian Process [45] into the model to perform effective mask refinement. [27, 28] proposes to synthesize plausible segmentation masks and select from them as the

final mask prediction. More recently, SAM [24] and SEEM [63] show remarkable improvements in interactive segmentation by training on large datasets.

Prompt Learning. Prompt learning emerges as a promising method in Natural Language Processing (NLP) for efficiently adapting pre-trained large language models to downstream tasks. With frozen pre-trained model, prompts provide task-specific guidance to the model [22, 36, 40], making it easier to adapt the pre-trained knowledge to downstream tasks [6]. Recently, learnable textual prompts are introduced into Vision-Language Models (VLM) [42, 60] to perform various vision-language tasks, such as referring image segmentation [50] and zero-shot segmentation [26]. Visual Prompt Tuning (VPT) extends prompt learning techniques to computer vision. It introduces few learnable visual prompts to pre-trained vision models for efficient adaptation to various downstream tasks [48, 55]. In this work, we strive to learn click prompts tailored for interactive segmentation.

Optimal Transport. Optimal Transport (OT) [38] is a mathematical framework to learn the correspondence between two probability distributions. OT theory involves finding the most efficient way to transport one distribution to another while satisfying certain constraints, *e.g.*, preserving the total mass of the distributions. With the great distribution matching [41] property, OT has been applied in various computer vision tasks, including generative models [44, 58], graph matching [53], image matching [35, 59], vision-language generalization [6], and domain adaptation [7]. Among various methods, Sinkhorn algorithm can efficiently solve the OT problem through entropy-regularization [10], and it can be directly applied to deep learning frameworks thanks to the extension of Envelop Theorem [41]. In this work, we introduce optimal transport into the interactive segmentation task to learn diverse user intentions with click prompts.

3 Preliminary

3.1 Problem Statement

Click-based interactive segmentation aims to create a model capable of accurately segmenting objects or regions of interest based on user-provided clicks. In the first interaction round, users place clicks on the input image $I \in \mathbb{R}^{w \times h \times 3}$, where h and w denote the height and width of the image, respectively. In subsequent rounds, additional clicks are provided to improve segmentation performance based on the input image I and previous segmentation feedback $\hat{M}_{i-1} \in \mathbb{R}^{h \times w}$. $\hat{M}_0$ is an empty mask for the first interaction round. The annotated clicks over all interactions can be grouped into positive and negative click sets, which emphasize the region of interest and the background, respectively. The positive and negative click sets are encoded as positive map $M_{pos} \in \mathbb{R}^{h \times w}$ and negative map $M_{neg} \in \mathbb{R}^{h \times w}$ using disk encoding [4]. The primary challenge in this task lies in achieving precise segmentation while minimizing interaction rounds.

3.2 Optimal Transport (OT).

Optimal transport is widely used to minimize the transport distance between two probability distributions. In the context of this work, we focus on optimal transport with discrete distributions. Consider two discrete distribution $\boldsymbol{\mu}$ and $\boldsymbol{v}$ defined in probability space $\mathcal{F}$ and $\mathcal{G}$:

$$\boldsymbol{\mu} = \sum_{m=1}^{M} p_m \delta_{\boldsymbol{f}_m}, \boldsymbol{v} = \sum_{n=1}^{N} q_n \delta_{\boldsymbol{g}_n}, \tag{1}$$

where δ_f and δ_g are Dirac Delta function located at support poinst $\boldsymbol{f}$ and $\boldsymbol{g}$, respectively. p_m and q_n are discrete probability vectors that satisfy $\sum_{m=1}^{M} p_m = 1$ and $\sum_{n=1}^{N} q_n = 1$. Then the discrete optimal transport can be formulated as:

$$\begin{aligned} \boldsymbol{T}^* &= \underset{\boldsymbol{T} \in \mathbb{R}^{M \times N}}{\arg\min} \sum_{m=1}^{M} \sum_{n=1}^{N} \boldsymbol{T}_{mn} \boldsymbol{C}_{mn} \\ \text{s.t.} \quad & \boldsymbol{T}\mathbf{1}^N = \boldsymbol{\mu}, \boldsymbol{T}^\top \mathbf{1}^M = \boldsymbol{v}. \end{aligned} \tag{2}$$

Here, $\boldsymbol{C}$ is the cost matrix which measures the distance between $\boldsymbol{f}_m$ and $\boldsymbol{g}_n$, *e.g.*, cosine distance $\boldsymbol{C}_{mn} = 1 - \frac{\boldsymbol{f}_m \boldsymbol{g}_n^\top}{||\boldsymbol{f}_m||_2 ||\boldsymbol{g}_n||_2}$. $\boldsymbol{T}$ is transport plan that learns to minimize the distance between two distributions. To speed up optimization, most works adopt the Sinkhorn algorithm [10], which is a strictly convex optimization problem. Then we have a fast optimization solution with few iterations:

$$\boldsymbol{T}^* = \text{diag}(\boldsymbol{a}^t) \exp(-\boldsymbol{C}/\lambda) \text{diag}(\boldsymbol{b}^t), \tag{3}$$

where t is the iteration, λ is a hyper-parameter, $\boldsymbol{a}^t = \boldsymbol{\mu}/\exp(-\boldsymbol{C}/\lambda)\boldsymbol{b}^{t-1}$, and $\boldsymbol{b}^t = \boldsymbol{v}/\exp(-\boldsymbol{C}/\lambda)\boldsymbol{a}^t$, with the initialization $\boldsymbol{b}^0 = \mathbf{1}$.

4 Method

In this work, we propose **Click Prompt Learning with Optimal Transport (CPlot)** for interactive image segmentation to capture diverse user intentions with click prompts. Specifically, we will first introduce the overall architecture of CPlot in Sec. 4.1. Then, we respectively detail the two key components in CPlot, i.e., prompt-pixel alignment module and click prompt optimal transport module, in Sec. 4.2 and Sec. 4.3.

4.1 Overall Architecture of CPlot

As illustrated in Fig. 2, our CPlot mainly contains 1) image and click encoder, 2) prompt-pixel alignment module, 3) click prompt optimal transport module and 4) mask decoder. All these components collaborate together to achieve superior interactive segmentation performance, and we will introduce in turn.

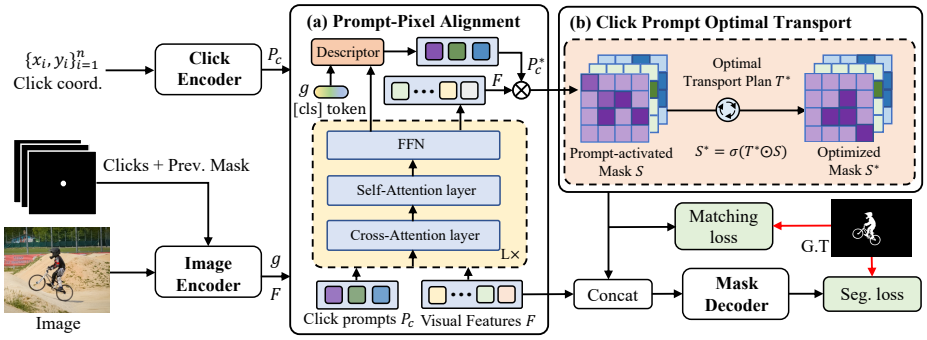


Fig. 2: Overview of the proposed Click Prompt Learning with Optimal Transport (CPlot). Given input image, click disk maps, and previous mask, the **Image Encoder** extracts visual features F. The **Click Encoder** initializes click prompts P_c with click coordinates. (a) **Prompt-Pixel Alignment** aims to align click prompts P_c with the visual features F in the feature space. (b) **Click Prompt Optimal Transport** adopts optimal transport plan to generate optimized mask S^* from vanilla prompt-activated mask S. A lightweight mask decoder is used to implicitly analyze optimized prompt-activated mask with visual features and make mask predictions.

Image and Click Encoder. The image encoder takes the input image I, the previous mask $\hat{M}$, and the positive and negative disk maps M_{pos} and M_{neg}, to produce visual features $F \in \mathbb{R}^{d \times H \times W}$ and class token $g \in \mathbb{R}^{1 \times d}$, where d, H, and W represent the channel, width, and height of the visual features, respectively. Generally, we could employ popular network architectures for image segmentation as the image encoder, such as ResNet series, HRNet, and Vision Transformer (ViT). In our approach, we use ViT-B as the model backbone to extract image features, as in SimpleClick [32]. To capture multi-scale information, we also employ the feature pyramid network [32] to merge the multi-scale features into the final visual features.

The click encoder is designed to transform input clicks into multiple click prompts, which are utilized to capture user intentions. Different from the click prompts in SAM [24] and SEEM [63], we only utilize positive clicks, which indicate regions of interest, to initialize the click prompts. Meanwhile, we leave the positive and negative disk map M_{pos} and M_{neg} to highlight the target in the visual features. Formally, we initialize click prompts as:

$$P_c = P + P_{p.e.} + P_{con}, \tag{4}$$

where $P_c \in \mathbb{R}^{N \times d}$ represent the set of N click prompts, each with a dimension of d. $P_{p.e.}$ and P_{con} respectively denote position prompt and content prompt, which encode position and content information about the target object. P indicates a learnable prompt. Given positive clicks, we first transform them as the disk map M_{pos}, which represents a small circular region around clicks. Then the positional prompt is formulated as:

$$P_{p.e.} = f_t\big(\texttt{MAP}(M_{pos}, K_p)\big), \tag{5}$$

where $f_t(\cdot)$ represents the non-linear transformation function, *e.g.*, an multi-layer perceptron. $\texttt{MAP}$ is the masked average pooling operation, $K_p \in \mathbb{R}^{d\times H\times W}$ is the positional encoding of visual features using sinusoidal encoding [49]. By doing so, we aggregate position information about the target object from a small circular region around positive clicks.

For the content prompt P_{con}, it is designed to encapsulate the semantic and appearance characteristics of the target object from the positive clicks. To this end, the content prompt is defined as:

$$P_{con} = f_t(\texttt{MAP}(M_{pos}, F)). \tag{6}$$

By introducing the click encoder, we initialize click prompts by integrating both position and content information about region of interest. As illustrated in Table 4 of Sec. 5.3, such click prompts initialization strategy leads to superior interactive segmentation results.

Prompt-Pixel Alignment Module. Given a plain click, click prompts encodes limited information about the target of interest. To enable click prompts aggregates as much information about target as possible, we introduce the prompt-pixel alignment module (PPAM). PPAM aligns click prompts P_c and visual features F in the same feature space, resulting in an aligned click prompts P_c^*, which consistently focus on region of interest. As shown in Table 3, removing this part results in severe performance degradation, we will introduce this key module in Sec. 4.2.

Click Prompt Optimal Transport Module. As shown in Fig. 1 (a), the prompt-activate masks resemble each other and cannot reflect diverse user intentions, particularly when the real user intention is very ambiguous. To this end, click prompt optimal transport module (CPOT) is designed to capture diverse user intention with click prompts. CPOT takes input click prompts P_c^* and visual features F, and generate multiple prompt-activated mask S^*, which indicates diverse user intentions. As illustrated in Table 3, removing this part results serve performance degradation, we we introduce this key module in Sec. 4.3.

Mask Decoder. The optimized prompt-activate masks S^* consists of multiple masks, each indicating a underlying user intention. Then we feed it with the visual features F into a mask decoder to predict the target mask:

$$\hat{M} = \Phi(\texttt{Cat}[F, S^*]), \tag{7}$$

where $\hat{M} \in \mathbb{R}^{H\times W}$ is the predicted mask. Φ is a light-weighted mask decoder, which consists of multiple convolutional layers followed by a classifier head.

4.2 Prompt-Pixel Alignment Module

Here we introduce prompt-pixel alignment module (PPAM) in detail. Given that click prompts and visual features represent information from different modalities, PPAM aims to align this cross-modal knowledge with plain transformer network in a unified feature space. Specifically, the click prompts and visual features are fed into a transformer network, which is composed of L transformer blocks. Each

block is sequentially structured with a cross-attention layer `CrossAtten`, a self-attention layer `SelfAtten`, and a feed-forward network `FFN`. For instance, in the l-th transformer block:

$$\begin{aligned} P_c^l &= \texttt{CrossAtten}(P_c^{l-1}, F) + P_c^{l-1}, \\ P_c^l &= \texttt{SelfAtten}(P_c^l) + P_c^l, \\ P_c^l &= \texttt{FFN}(P_c^l) + P_c^l, \end{aligned} \tag{8}$$

where P_c^l and P_c^{l-1} represent the click prompts in the l-th and $\{l-1\}$-th transformer block, respectively. Through mutual interaction, the updated click prompts $\hat{P}_c$ effectively encode local visual details derived from the visual features. To encapsulate global information about the target, we incorporate a descriptor [62] to integrate class token g of input image and the updated click prompts $\hat{P}_c$. Formally, the descriptor is defined as:

$$P_c^* = f_t(\texttt{cat}[\hat{P}_c \odot g, g]), \tag{9}$$

where $P_c^* \in \mathbb{R}^{N \times d}$ is the aligned click prompts, $\texttt{cat}[\cdot]$ is the concatenation operator, and $\odot$ denotes the Hadamard Product. By incorporating both local and global visual information, all click prompts are encouraged to focus on consistent regions of interest. As shown in Fig. 1 (a), the prompt-activated masks filter out most background and generally indicates the final target of interest.

4.3 Click Prompt Optimal Transport Module

In this subsection, we introduce the details of click prompt optimal transport module (CPOT). CPOT regularizes multiple click prompts to describe distinct visual regions, which reflect diverse underlying user intentions. To reduce the computational cost and avoid the extra model parameters, we learn the click prompts with a two-stage optimization strategy.

Specifically, CPOT takes input the flattened visual features $F \in \mathbb{R}^{HW \times d}$ and the click prompts $P_c^* \in \mathbb{R}^{N \times d}$. In the inner loop, we fix visual features F and learn the transport plan T between F and P_c^* by optimizing the cost matrix $C \in \mathbb{R}^{HW \times N}$ to push P_c^* to F. Here, we define the cost matrix in Eq. (2) as the cosine distance between P_c^* and F:

$$C := 1 - F(P_c^*)^\top. \tag{10}$$

With the fast optimization solution in Eq. (3) provided by the Sinkhorn algorithm, we can derive the optimal transport plan $T^* \in \mathbb{R}^{HW \times N}$.

The optimal transport plan T^* represents a mapping matrix that assigns each pixel in visual features F to corresponding click prompt with minimal cost C. In such way, optimal transport enables click prompts to focus on distinct visual regions, which reflect diverse user intentions. To obtain prompt-activated masks with optimized click prompts, we have:

$$S^* = \sigma(T^* \odot S), \tag{11}$$

where $\sigma(\cdot)$ is the sigmoid function, and $S = F(P_c^*)^\top$ is the coarse prompt-activated mask derived from click prompts P_c^*. Each prompt-activated mask in S^* is optimized to highlight a distinct visual region, such diverse masks provide robust priors to infer real user intention. We further feed the optimized prompt-activate masks S^* into the mask decoder with visual features to implicitly infer real user intention.

In the outer loop, we fix the transport plan T^*, then supervise the learning of click prompts and other model parameters, i.e., image and click encoder, mask decoder, and PPAM, from annotated data. Overall, we adopt the following objective:

$$\mathcal{L} = \mathcal{L}_{\text{fc}}(\hat{M}, M_{gt}) + \alpha \mathcal{L}_{\text{fc}}(\frac{1}{N}\sum_{i=1}^{N}(S_i^*), M_{gt}) + \beta \underset{\mathcal{L}}{argmin}\, \mathcal{L}_{\text{fc}}(S_i^*, M_{gt}), \tag{12}$$

where $\mathcal{L}_{\text{fc}}$ denotes the normalized focal loss, $S_i^* \in \mathbb{R}^{H\times W}$ is the i-th prompt-activated mask indicating one specific visual region, α and β are two hyper-parameters. We adopt the second and third loss terms to ensure that N prompt-activated masks S_i^* contain characteristics of overlapped visual regions. Note that the transport plan T^* in Eq. (3) only contains matrix multiplication and exponential operation, thus the CPOT is fully differentiable, which is easy to implement using an autograd library like PyTorch. To this end, though the optimization strategy of our method is two-stage, we implement the whole training flow in an end-to-end manner.

5 Experiments

Datasets. We conduct model training on either SBD [17] or COCO [29]+LVIS [16] separately. SBD [17] contains 8,498 training images with 20,172 instances and 2,857 validation images with 6,671 instances. We adopt the training set of SBD for model training. COCO [29]+LVIS [16] consists of 118K training images (1.2M instances) and various scene variations. Then we evaluate our model on five natural image datasets, GrabCut [43], Berkeley [37], SBD [17], DAVIS [39], Pascal VOC [11], and three medical datasets, ssTEM [13], BraTS [3], OAIZIB [2]. GrabCut [43] contains 50 images with 50 instances, each with clear foreground and background differences. Berkeley [37] includes 96 images with 100 instances and shares some small object images with GrabCut. SBD [17] has 2,857 validation images with 6,671 instances. DAVIS [39] is composed of 50 videos divided into 345 frames, each frame has high-quality mask. Pascal VOC [11] has 1,449 images with 3,427 instances in the validation set. ssTEM [13] has two image stacks, each with 20 medical images. BraTS [3] consists of 369 slices from magnetic resonance images (MRI) volumes. OAIZIB [2] includes 150 slices with 300 instances from 507 MRI volumes.

Evaluation Metric. We follow the same click simulation strategy as in previous methods [9, 30, 32, 47, 61], where the next click is placed at the center of the largest error region derived from comparing the prediction and ground truth.

Table 1: Comparison with state-of-the-art interactive segmentation methods. We report results on 5 benchmarks: GrabCut [43], Berkeley [37], SBD [17], DAVIS [39], and Pascal VOC [11]. † denotes generalist model, while ‡ denotes results provided by SSEM [63]. Best results are marked in bold. Our CPlot is a consistent top-performer on all benchmarks.

Method	Backbone	Train	GrabCut		Berkeley		SBD		DAVIS		Pascal VOC	
			NoC85	NoC90	NoC85	NoC90	NoC85	NoC90	NoC85	NoC90	NoC85	NoC90
SAM† [24]	ViT-B	SA-1B	-	-	-	-	6.50‡	9.76 ‡	-	-	-	-
SEEM† [63]	DaViT-B	C+L	-	-	-	-	6.67	9.99	-	-	-	-
LD [27] CVPR18	VGG-19	SBD	3.20	4.79	-	-	7.41	10.78	5.05	9.57	-	-
BRS [21] CVPR19	DenseNet	SBD	2.60	3.60	-	5.08	6.59	9.78	5.58	8.24	-	-
f-BRS [46] CVPR20	ResNet-101	SBD	2.30	2.72	-	4.57	4.81	7.73	5.04	7.41	-	-
RITM [47] Preprint21	HRNet-18	SBD	1.76	2.04	1.87	3.22	3.39	5.43	4.94	6.71	2.51	3.03
CDNet [8] ICCV21	ResNet-34	SBD	1.86	2.18	1.95	3.27	5.18	7.89	5.00	6.89	3.61	4.51
PseudoClick [34] ECCV22	HRNet-18	SBD	1.68	2.04	1.85	3.23	3.38	5.40	4.81	6.57	2.34	2.74
FocalClick [9] CVPR22	SegF-B0	SBD	1.66	1.90	-	3.14	4.34	6.51	5.02	7.06	-	-
FocusCut [30] CVPR22	ResNet-101	SBD	1.46	1.64	1.81	3.01	3.40	5.31	4.85	6.22	-	-
GPCIS [61] CVPR23	ResNet-50	SBD	1.64	1.82	1.60	2.60	3.80	5.71	4.37	5.89	-	-
FCFI [51] CVPR23	ResNet-101	SBD	1.64	1.80	-	2.84	3.26	5.35	4.75	6.48	-	-
SimpleClick [32] Preprint23	ViT-B	SBD	1.40	1.54	1.44	2.46	3.28	5.24	4.10	5.48	2.38	2.81
CPlot (Ours)	ViT-B	SBD	**1.34**	**1.48**	**1.40**	**2.18**	**3.05**	**4.95**	**4.00**	**5.29**	**2.23**	**2.62**
RITM [47] Preprint21	HRNet-32	C+L	1.46	1.56	1.43	2.10	3.59	5.71	4.11	5.34	2.19	2.57
CDNet [8] ICCV21	ResNet-34	C+L	1.40	1.52	1.47	2.06	4.30	7.04	4.27	5.56	2.74	3.30
PseudoClick [34] ECCV22	HRNet-32	C+L	1.36	1.50	1.40	2.08	3.46	5.54	3.79	5.11	**1.94**	**2.25**
FocalClick [9] CVPR22	SegF-B0	C+L	1.40	1.66	1.59	2.27	4.56	6.86	4.04	5.49	2.97	3.52
FocalClick [9] CVPR22	SegF-B3	C+L	1.44	1.50	1.55	1.92	3.53	5.59	3.61	4.90	2.46	2.88
FCFI [51] CVPR23	HRNet-18	C+L	1.38	1.46	-	1.96	3.63	5.83	3.97	5.16	-	-
SimpleClick [32] Preprint23	ViT-B	C+L	1.38	1.48	**1.36**	1.97	3.43	5.62	3.66	5.06	2.06	2.38
CPlot (Ours)	ViT-B	C+L	**1.21**	**1.32**	**1.36**	**1.90**	**3.37**	**5.48**	**3.26**	**4.65**	1.98	2.31

We use the Number of Clicks (NOC) as the evaluation metric, which computes the average number of clicks required to achieve a fixed Intersection over Union (IoU). Specifically, we set the target IoU as 85% and 90%, leading to the metrics NoC85 and NoC90, respectively. The maximum number of clicks allowed for each instance is set to 20 by default. We also use the average IoU given a fixed number of clicks (mIoU@k) as an additional evaluation metric to measure the segmentation quality at different interactive stages.

Implementation Details. In our experiments, we follow the training strategies as in [32]: training for 55 epochs on either SBD or COCO+LVIS datasets. We use ViT-B as the backbone, which is initialized by MAE pretraining on ImageNet. Images are resized and cropped to 448×448, and we use the same data augmentation during training as in [32]. Our model is trained with the Adam optimizer with parameters set as $\beta_1 = 0.9, \beta_2 = 0.999$, the initial learning rate is set to 5×10^{-5} and decreases to 5×10^{-6} after 50 epochs. We set the number of click prompts as $N = 5$ in our model and the number of transformer block layers in the prompt-pixel alignment module is set as $L = 2$. The hyper-parameters in Eq. (12) are set as $\alpha = 1$ and $\beta = 1$. We implement our methods in PyTorch and all experiments are conducted on four Tesla A100 GPUs.

5.1 Comparison with State-of-the-art

Table 1 reports qualitative results for interactive segmentation on five natural image benchmarks. We compare our method, CPlot, with two types of methods: two generalist segmentation models, i.e., SAM [24] and SEEM [63], and

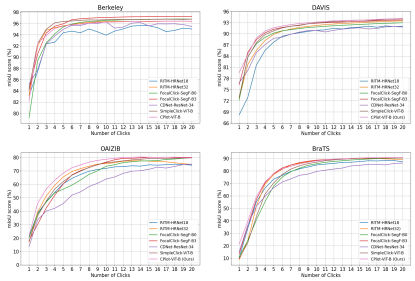


Fig. 3: Convergence analysis of existing interactive segmentation methods and our CPlot. First row: Results on natural image datasets Berkeley [37] and DAVIS [39]. Second row: Results on medical datasets OAIZIB [2] and BraTS [3]. Zoom in for details.

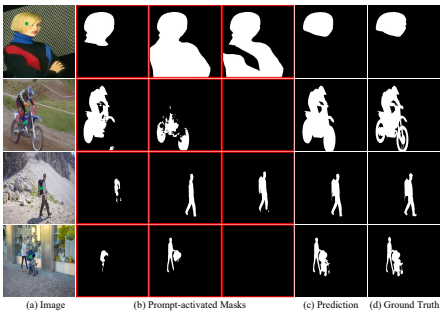


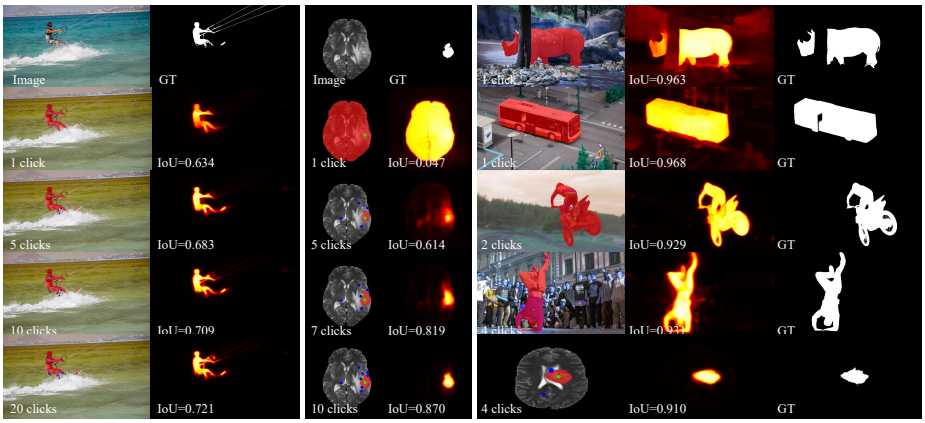
Fig. 4: Visualization of diverse prompt-activated masks given one positive click. (a) Input image with a positive click. (b) Intermediate masks derived from score maps S^*. (c) Final prediction. (d) Ground Truth. Positive click is marked in Green.

typical click-based interactive segmentation methods. Generally, CPlot demonstrates state-of-the-art performance on all benchmarks with model trained either on SBD or COCO+LVIS datasets. Compared with generalist models SAM and SEEM, our CPlot significantly reduces the number of interactive rounds. For instance, with the model pretrained on the SBD, our method reduces the required clicks to achieve 85% IoU to 3.05, which is 3.62 clicks fewer than the generalist model SAM. We suspect that these generalist models are generally optimized to take diverse prompt formats like text and bounding boxes, and neglect capturing diverse user intention, thus leading to inferior performance in click-based interactive segmentation task.

Compared with typical interactive segmentation models such as SimpleClick [32], our method CPlot achieves superior interactive segmentation performance. Specifically, in terms of NOC85 on complex DAVIS dataset, CPlot trained on COCO+LVIS dataset reduces the number of clicks from 3.66 (SimpleClick result) to 3.26. Meanwhile, we also report convergence analysis in Fig. 3. As shown in the first row, with the number of clicks increasing, the proposed CPlot consistently improves the segmentation performance and generally achieves better results than its competitors across all click counts. This further highlights the superiority of CPlot for the click-based interactive segmentation task. Fig. 5 (c) provides some vi-

Table 2: Out-of-domain evaluation on three medical benchmarks: ssTEM [13], BraTS [3], and OAIZIB [2]. mIoU@10 denotes mIoU score after 10 clicks. Our model generalize well to the medical domain even without fine-tuning on medical data.

Model	ssTEM mIoU@10	BraTS mIoU@10	OAIZIB mIoU@10
RITM-H18 [47]	93.15	87.05	71.04
CDN-RN34 [8]	66.72	58.34	38.07
RITM-H32 [47]	94.11	88.34	75.27
CDN-RN34 [8]	88.46	80.24	63.19
FC-SF-B0 [9]	92.62	86.02	74.08
FC-SF-B3 [9]	93.61	88.62	75.77
SimpleClick-ViT-B [32]	93.72	86.98	76.05
CPlot-ViT-B (Ours)	**95.33**	**88.96**	**78.45**



(a) Challenge case on natural image (b) Challenge case on medical image (c) Normal cases

Fig. 5: Segmentation results on DAVIS [39] and BraTS [39]. We show (a) a challenge case on the natural image, (b) a challenge case on the medical image, and (c) five normal cases. The segmentation probability maps are shown in golden; the segmentation maps are overlaid in red on the original images. Positive and negative clicks are marked with green and blue dots on the image, respectively.

sualization of our method, demonstrating that our method could make more accurate mask prediction with few clicks.

5.2 Out-of-domain Evaluation

Our interactive segmentation model, typically domain-agnostic, is further evaluated for out-of-domain generalization on three medical datasets: ssTEM [13], BraTS [3], and OAIZIB [2]. Specifically, we directly evaluate interactive segmentation performance on a medical dataset with the pre-trained model on the COCO+LVIS dataset. As shown in Table 2, our model exhibits good generalization to the medical domain even without any fine-tuning on medical data. For instance, our model achieves 78.61 mIoU after 10 clicks on the OAIZIB dataset, surpassing the SimpleClick model by 2.61. Furthermore, we also showcase the convergence analysis of our model with medical data in Fig. 3. As shown in the second row, our method CPlot achieves overall better performance, especially with fewer clicks. Such results highlight the strong generalization capabilities of our model across various domains.

5.3 Ablation study

We conduct extensive ablations on five natural image datasets to demonstrate the effectiveness of our method. We take SimpleClick [32] as our baseline model, perform training on the SBD dataset, and report NoC90 as the evaluation metric.

Table 3: Component analysis of CPlot. PPAM and CPOT denote Prompt-Pixel Alignment Module and Click Prompt Optimal Transport Solver, respectively. With both PPAM and CPOT, our model achieves consistent best performance on all five benchmarks.

PPAM	CPOT	GrabCut NoC90	Berkeley NoC90	SBD NoC90	DAVIS NoC90	VOC NoC90
-	-	1.54	2.46	5.24	5.48	2.81
✓	-	1.54	2.37	5.12	5.34	2.67
-	✓	1.56	2.32	5.18	5.42	2.70
✓	✓	**1.48**	**2.18**	**4.95**	**5.29**	**2.62**

Table 4: Effects of initialization formats of click prompt. P denote learnable prompt, $P_{p.e.}$ and P_{con} represent position and content prompts derived from positive clicks, respectively. Our initialization strategy from Eq. (4) achieves superior results than its variants.

P	$P_{p.e.}$	P_{con}	GrabCut NoC90	Berkeley No90	SBD NoC90	DAVIS NoC90	VOC NoC90
✓	-	-	1.54	2.47	5.17	5.89	2.77
✓	✓	-	1.50	2.25	5.12	5.46	2.63
✓	-	✓	1.55	2.38	5.24	5.53	2.68
✓	✓	✓	**1.48**	**2.18**	**4.95**	**5.29**	**2.62**

Component analysis. To analyze the effects of different components in our model, we conduct a component-wise analysis. The proposed method has two core components: prompt-pixel alignment module (PPAM in Sec. 4.2) and click prompt optimal transport module (CPOT in Sec. 4.3). As shown in Table 3, removing PPAM from our model results in more click numbers to achieve 90% IoU. This result demonstrates the significance of the PPAM in aligning click prompts and visual features. Similarly, removing CPOT from our model also leads to the increase of click number on all datasets, demonstrating the importance of CPOT in our model. Additionally, Fig. 4 showcases the diversity of prompt-activated masks derived from click prompts in Eq. (11). With the proposed CPOT, our model generates a more diverse set of prompt-activated masks, providing robust priors to better understand the intention of input clicks and consequently leading to more accurate mask predictions.

Effects of initialization formats of click prompt. To analyze the effectiveness of the proposed initialization strategies from Eq. (5) for click prompts, we compare it with three initialization variants: (1) randomly initialized as learnable prompts P, (2) learnable prompt and position prompt $P + P_{p.e.}$, and (3) learnable prompt and content prompt $P + P_c$. We report the experimental results in Table 4. Generally, randomly initialized prompt P exhibits the worst results as it directly learns from the input image without explicit information about clicks. Both position prompt $P_{p.e.}$ and content prompt P_{con} achieve better performance than learnable prompt P, as position and content information about region of interest are provided, respectively. With the proposed initialization strategy, our model achieves the best performance across all five benchmarks. This is reasonable as such initialization incorporates both position and content priors about the target object into click prompts, facilitating effective optimization of prompts.

Effect of click prompts number N. To evaluate the effects of click prompt number N in our model, we vary it in $[1, 3, 5, 7]$ and report corresponding results in Table 5. Notably, with only one click prompt, our model exhibits inferior performance, *e.g.*, a NoC90 of 2.69 on the Berkeley dataset. As we increase the number of clicks, our model consistently demonstrates much better performance than that achieved with just one prompt, such as achieving a NoC90 of 2.18 on the Berkeley dataset with 5 click prompts. These results demonstrate that

capturing diverse click-activated masks helps to infer real user intention, leading to better interactive segmentation results.

Table 5: Benefits of the number of click prompts N. Our model with five prompts achieves best performance.

N Prompts	GrabCut NoC90	Berkeley NoC90	SBD NoC90	DAVIS NoC90	VOC NoC90
1	1.62	2.69	5.47	6.04	3.25
3	1.54	2.29	4.96	5.57	**2.55**
5	**1.48**	**2.18**	4.95	**5.29**	2.62
7	1.52	2.46	**4.89**	5.52	2.66

Table 6: Benefits of the number of transformer blocks L**.** Our model with two transformer blocks performs best.

L Blocks	GrabCut NoC90	Berkeley NoC90	SBD NoC90	DAVIS NoC90	VOC NoC90
1	1.58	2.62	5.09	5.48	2.72
2	**1.48**	**2.18**	**4.95**	**5.29**	2.62
3	1.58	2.46	5.06	5.69	2.68
4	1.64	2.57	5.20	5.65	**2.57**

Effect of Transformer Block number L**.** Our PPAM consists of L transformer blocks. To analyze model performance with different number of transformer blocks (Sec. 4.2), we conduct experiments with our model using different number of transformer blocks, the experimental results are shown in Table 6. Our model achieves best performance when PPAM consists of L transformer blocks, while more transformer blocks also lead to performance degradation. This analysis highlights the effectiveness and efficiency of the proposed PPAM.

5.4 Limitations

Although our method achieves state-of-the-art performance on both nature image and medical image datasets, it does have certain limitations, as illustrated in Figure 5 (a) and (b). One of these limitations is the difficulty in accurately segmenting thin structures such as ropes and bicycles. To address this issue, incorporating local refinement techniques [9, 30] could be beneficial. Additionally, in the case of complex structures found in medical images, our method may require a higher number of clicks to achieve satisfactory performance. This challenge can be partially mitigated by fine-tuning the model on medical datasets. We leave these two limitation for further research.

6 Conclusion

In this work, we propose click prompt learning with optimal transport (CPlot) to capture diverse user intentions from click prompts. We first propose a prompt-pixel alignment module (PPAM) to align click prompts and visual features in the same embedding space. By doing so, click prompts are pushed to focus on consistent foreground regions of input image. To learn diverse user intention with click prompts, we propose a click prompt optimal transport module (CPOT). By incorporating the optimal transport, click prompts are encouraged to focus on distinct visual regions, which reflect diverse user intentions. Furthermore, we formulate learning of click prompts with a two-stage optimisation strategy. Extensive experiments on both natural images and medical datasets demonstrate that our method achieves state-of-the-art performance in the click-based interactive segmentation task.

Acknowledgements

This work was partially funded by Elekta Oncology Systems AB and a RVO public-private partnership grant (PPS2102).

References

1. Acuna, D., Ling, H., Kar, A., Fidler, S.: Efficient interactive annotation of segmentation datasets with Polygon-RNN++. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 859–868 (2018) 1
2. Ambellan, F., Tack, A., Ehlke, M., Zachow, S.: Automated segmentation of knee bone and cartilage combining statistical shape knowledge and convolutional neural networks: Data from the osteoarthritis initiative. Medical image analysis **52**, 109–118 (2019) 9, 11, 12
3. Baid, U., Ghodasara, S., Mohan, S., Bilello, M., Calabrese, E., Colak, E., Farahani, K., Kalpathy-Cramer, J., Kitamura, F.C., Pati, S., et al.: The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. arXiv preprint arXiv:2107.02314 (2021) 9, 11, 12
4. Benenson, R., Popov, S., Ferrari, V.: 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) 4
5. Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18392–18402 (2023) 1
6. Chen, G., Yao, W., Song, X., Li, X., Rao, Y., Zhang, K.: Prompt learning with optimal transport for vision-language models. arXiv preprint arXiv:2210.01253 (2022) 4
7. Chen, L., Gan, Z., Cheng, Y., Li, L., Carin, L., Liu, J.: Graph optimal transport for cross-domain alignment. In: International Conference on Machine Learning. pp. 1542–1553. PMLR (2020) 4
8. Chen, X., Zhao, Z., Yu, F., Zhang, Y., Duan, M.: Conditional diffusion for interactive segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7345–7354 (2021) 10, 11
9. Chen, X., Zhao, Z., Zhang, Y., Duan, M., Qi, D., Zhao, H.: FocalClick: Towards practical interactive image segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1300–1309 (2022) 2, 3, 9, 10, 11, 14
10. Cuturi, M.: Sinkhorn distances: Lightspeed computation of optimal transport. Advances in neural information processing systems **26** (2013) 3, 4, 5
11. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The PASCAL visual object classes (VOC) challenge. International Journal of Computer Vision **88**(2), 303–338 (2010) 9, 10
12. Forte, M., Price, B., Cohen, S., Xu, N., Pitié, F.: Getting to 99% accuracy in interactive segmentation. arXiv preprint arXiv:2003.07932 (2020) 3
13. Gerhard, S., Funke, J., Martel, J., Cardona, A., Fetter, R.: Segmented anisotropic sstem dataset of neural tissue. figshare pp. 0–0 (2013) 9, 11, 12
14. Grady, L.: Random walks for image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence **28**(11), 1768–1783 (2006) 3

15. Gulshan, V., Rother, C., Criminisi, A., Blake, A., Zisserman, A.: Geodesic star convexity for interactive image segmentation. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition. pp. 3129–3136 (2010) 3
16. Gupta, A., Dollar, P., Girshick, R.: Lvis: A dataset for large vocabulary instance segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5356–5364 (2019) 9
17. Hariharan, B., Arbeláez, P., Bourdev, L., Maji, S., Malik, J.: Semantic contours from inverse detectors. In: Proceedings of IEEE International Conference on Computer Vision. pp. 991–998 (2011) 9, 10
18. Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H.R., Xu, D.: Unetr: Transformers for 3d medical image segmentation. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 574–584 (2022) 1
19. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-Or, D.: Prompt-to-prompt image editing with cross attention control. arXiv preprint arXiv:2208.01626 (2022) 1
20. Huang, Y., Yang, X., Liu, L., Zhou, H., Chang, A., Zhou, X., Chen, R., Yu, J., Chen, J., Chen, C., et al.: Segment anything model for medical images? arXiv preprint arXiv:2304.14660 (2023) 1
21. Jang, W.D., Kim, C.S.: Interactive image segmentation via backpropagating refinement scheme. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5297–5306 (2019) 2, 3, 10
22. Jiang, Z., Xu, F.F., Araki, J., Neubig, G.: How can we know what language models know? Transactions of the Association for Computational Linguistics **8**, 423–438 (2020) 4
23. Kim, T.H., Lee, K.M., Lee, S.U.: Nonparametric higher-order learning for interactive segmentation. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition. pp. 3201–3208 (2010) 3
24. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., Girshick, R.: Segment anything. arXiv:2304.02643 (2023) 2, 4, 6, 10
25. Kontogianni, T., Gygli, M., Uijlings, J., Ferrari, V.: Continuous adaptation for interactive object segmentation by learning from corrections. In: Proceedings of the European Conference on Computer Vision. pp. 579–596 (2020) 2, 3
26. Li, B., Weinberger, K.Q., Belongie, S., Koltun, V., Ranftl, R.: Language-driven semantic segmentation. arXiv preprint arXiv:2201.03546 (2022) 4
27. Li, Z., Chen, Q., Koltun, V.: Interactive image segmentation with latent diversity. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 577–585 (2018) 2, 3, 10
28. Liew, J.H., Cohen, S., Price, B., Mai, L., Ong, S.H., Feng, J.: Multiseg: Semantically meaningful, scale-diverse segmentations from minimal user input. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 662–670 (2019) 2, 3
29. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. pp. 740–755. Springer (2014) 9
30. Lin, Z., Duan, Z.P., Zhang, Z., Guo, C.L., Cheng, M.M.: FocusCut: Diving into a focus view in interactive segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2637–2646 (2022) 2, 3, 9, 10, 14

31. Lin, Z., Zhang, Z., Chen, L.Z., Cheng, M.M., Lu, S.P.: Interactive image segmentation with first click attention. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13339–13348 (2020) 2, 3
32. Liu, Q., Xu, Z., Bertasius, G., Niethammer, M.: Simpleclick: Interactive image segmentation with simple vision transformers. arXiv preprint arXiv:2210.11006 (2022) 3, 6, 9, 10, 11, 12
33. Liu, Q., Xu, Z., Jiao, Y., Niethammer, M.: isegformer: Interactive segmentation via transformers with application to 3d knee mr images. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part V. pp. 464–474. Springer (2022) 3
34. Liu, Q., Zheng, M., Planche, B., Karanam, S., Chen, T., Niethammer, M., Wu, Z.: Pseudoclick: Interactive image segmentation with click imitation. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VI. pp. 728–745. Springer (2022) 2, 3, 10
35. Liu, W., Zhang, C., Ding, H., Hung, T.Y., Lin, G.: Few-shot segmentation with optimal transport matching and message flow. arXiv preprint arXiv:2108.08518 (2021) 4
36. Liu, X., Zheng, Y., Du, Z., Ding, M., Qian, Y., Yang, Z., Tang, J.: Gpt understands, too. arXiv preprint arXiv:2103.10385 (2021) 4
37. McGuinness, K., O'connor, N.E.: A comparative evaluation of interactive segmentation algorithms. Pattern Recognition **43**(2), 434–444 (2010) 9, 10, 11
38. Monge, G.: Mémoire sur la théorie des déblais et des remblais. Mem. Math. Phys. Acad. Royale Sci. pp. 666–704 (1781) 4
39. Perazzi, F., Pont-Tuset, J., McWilliams, B., Van Gool, L., Gross, M., Sorkine-Hornung, A.: A benchmark dataset and evaluation methodology for video object segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 724–732 (2016) 9, 10, 11, 12
40. Petroni, F., Rocktäschel, T., Lewis, P., Bakhtin, A., Wu, Y., Miller, A.H., Riedel, S.: Language models as knowledge bases? arXiv preprint arXiv:1909.01066 (2019) 4
41. Peyré, G., Cuturi, M., et al.: Computational optimal transport. Center for Research in Economics and Statistics Working Papers (2017-86) (2017) 4
42. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021) 4
43. Rother, C., Kolmogorov, V., Blake, A.: GrabCut: Interactive foreground extraction using iterated graph cuts. ACM Transactions on Graphics **23**(3), 309–314 (2004) 3, 9, 10
44. Salimans, T., Zhang, H., Radford, A., Metaxas, D.: Improving gans using optimal transport. arXiv preprint arXiv:1803.05573 (2018) 4
45. Seeger, M.: Gaussian processes for machine learning. International journal of neural systems **14**(02), 69–106 (2004) 3
46. Sofiiuk, K., Petrov, I., Barinova, O., Konushin, A.: f-BRS: Rethinking backpropagating refinement for interactive segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8623–8632 (2020) 2, 3, 10
47. Sofiiuk, K., Petrov, I.A., Konushin, A.: Reviving iterative training with mask guidance for interactive segmentation. arXiv preprint arXiv:2102.06583 (2021) 1, 2, 3, 9, 10, 11

48. Sohn, K., Hao, Y., Lezama, J., Polania, L., Chang, H., Zhang, H., Essa, I., Jiang, L.: Visual prompt tuning for generative transfer learning. arXiv preprint arXiv:2210.00990 (2022) 4
49. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017) 7
50. Wang, Z., Lu, Y., Li, Q., Tao, X., Guo, Y., Gong, M., Liu, T.: Cris: Clip-driven referring image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11686–11695 (2022) 4
51. Wei, Q., Zhang, H., Yong, J.H.: Focused and collaborative feedback integration for interactive image segmentation. arXiv preprint arXiv:2303.11880 (2023) 2, 10
52. Wu, J., Zhao, Y., Zhu, J.Y., Luo, S., Tu, Z.: MILCut: A sweeping line multiple instance learning paradigm for interactive image segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 256–263 (2014) 1
53. Xu, H., Luo, D., Zha, H., Duke, L.C.: Gromov-wasserstein learning for graph matching and node embedding. In: International conference on machine learning. pp. 6932–6941. PMLR (2019) 4
54. Xu, N., Price, B., Cohen, S., Yang, J., Huang, T.: Deep interactive object selection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 373–381 (2016) 1
55. Zang, Y., Li, W., Zhou, K., Huang, C., Loy, C.C.: Unified vision and language prompt learning. arXiv preprint arXiv:2210.07225 (2022) 4
56. Zhang, R., Zhu, J.Y., Isola, P., Geng, X., Lin, A.S., Yu, T., Efros, A.A.: Real-time user-guided image colorization with learned deep priors. arXiv preprint arXiv:1705.02999 (2017) 2
57. Zhang, S., Liew, J.H., Wei, Y., Wei, S., Zhao, Y.: Interactive object segmentation with inside-outside guidance. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12234–12244 (2020) 1
58. Zhao, H., Phung, D., Huynh, V., Le, T., Buntine, W.: Neural topic model via optimal transport. arXiv preprint arXiv:2008.13537 (2020) 4
59. Zhao, W., Rao, Y., Wang, Z., Lu, J., Zhou, J.: Towards interpretable deep metric learning with structural matching. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9887–9896 (2021) 4
60. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. International Journal of Computer Vision **130**(9), 2337–2348 (2022) 4
61. Zhou, M., Wang, H., Zhao, Q., Li, Y., Huang, Y., Meng, D., Zheng, Y.: Interactive segmentation as gaussian process classification. arXiv preprint arXiv:2302.14578 (2023) 3, 9, 10
62. Zhou, Z., Lei, Y., Zhang, B., Liu, L., Liu, Y.: Zegclip: Towards adapting clip for zero-shot semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11175–11185 (2023) 8
63. Zou, X., Yang, J., Zhang, H., Li, F., Li, L., Gao, J., Lee, Y.J.: Segment everything everywhere all at once (2023) 4, 6, 10