BAPNET: BOUNDARY AND AMBIGUITY PRIORS FOR INTERACTIVE CAMOUFLAGED OBJECT SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Camouflaged object segmentation involves strong foreground–background similarity and ambiguous boundaries, making pixel-level annotation costly and time-consuming. Interactive segmentation offers an efficient solution by using limited user interactions, such as clicks, to assist mask generation. However, existing methods are mainly designed for natural images and often struggle in camouflaged scenarios, where ambiguous boundaries and high prediction uncertainty can lead to inaccurate contours and foreground–background confusion. To address these challenges, we propose a Boundary and Ambiguity Priors for Interactive Camouflaged Object Segmentation (BAPNet) that jointly models boundary information and prediction uncertainty. Specifically, a Grassmann Boundary Guidance Module (GBGM) is introduced to explicitly predict object boundaries and guide the model to focus on foreground–background transition regions, while an Ambiguity-Prior Refinement Module (APRM) identifies potential areas with unreliable predictions and enhances local discrimination in ambiguous regions. To the best of our knowledge, BAPNet is the first systematic study of interactive camouflaged object segmentation. Extensive experiments on multiple camouflaged object datasets demonstrate that the proposed method achieves superior segmentation accuracy and interaction efficiency compared with existing approaches.

1 INTRODUCTION

Camouflaged object segmentation (Song et al., 2025; Pang et al., 2024) aims to identify targets that are highly similar to the background in terms of color, texture, and structure. Due to camouflaged objects often exhibit no obvious visual differences from their surroundings, the boundary between foreground and background is usually extremely ambiguous, making this task one of the most challenging problems in computer vision. In recent years, numerous deep learning methods (Wu et al., 2025; Yue et al., 2024) have been proposed to improve camouflaged object segmentation performance, but these methods rely heavily on high-quality pixel-level annotations for training. However, in camouflaged scenes, object boundaries are often difficult to discern, requiring numerous clicks and meticulous manual outlining to obtain accurate pixel-level annotations, which substantially increases the labeling burden. As illustrated in Fig. 1(a), manual pixel-level annotation is considerably more time-consuming than interactive annotation. The manual annotation times are obtained from published literature (Le et al., 2019; Fan et al., 2020a), while interactive annotation times are estimated by multiplying the average number of clicks required to reach 95% IoU by a per-click cost of 4 seconds. This comparison indicates potential time savings of 72.7% on CAMO (Le et al., 2019) and 98.4% on COD10K (Fan et al., 2020a). Therefore, improving labeling efficiency while maintaining annotation quality is important.

To reduce the manual cost of pixel-level annotation, interactive segmentation (Lee et al., 2024) progressively generates object masks by introducing a small amount of user interaction (e.g., clicks, bounding boxes, or scribbles), thereby significantly improving annotation efficiency. In recent years, with the development of deep learning, a large number of interactive segmentation methods (Hu et al., 2024; Zhao et al., 2024) have been proposed and have achieved remarkable progress in natural image annotation tasks. These methods (Chen et al., 2026; Liu et al., 2023) typically exploit

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.

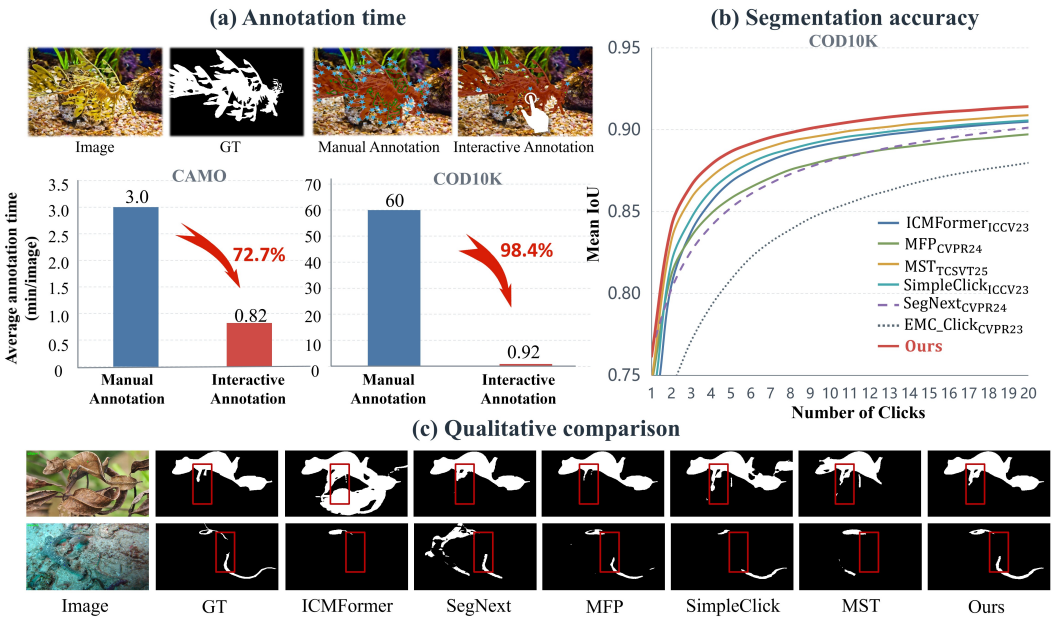

Figure 1: (a) Comparison of time cost for manual pixel-level and interactive annotation on CAMO and COD10K, with 72.7% and 98.4% time reduction using interactive tools. Pixel-level annotation time is from existing literature; interactive time is computed by multiplying average clicks for 95% IoU by per-click cost of 4s. (b) Average IoU versus number of clicks, showing our method achieves higher segmentation precision with fewer clicks. (c) Qualitative comparison of segmentation results from different methods on challenging camouflaged objects.

user-provided prompts to guide the model to localize target regions, and continuously refine the segmentation results through multiple rounds of interaction, enabling annotators to obtain high-quality segmentation masks with fewer operations. As a result, interactive segmentation (Huang et al., 2023) has gradually become an important tool for building large-scale visual datasets and has been widely applied to multimedia data processing tasks such as data annotation and data cleaning.

However, most existing interactive segmentation methods (Liu et al., 2024; Xu et al., 2025) are designed for natural images with relatively salient foreground–background contrast and often fail to achieve desirable results in camouflaged scenarios. As illustrated in the first row of Fig. 1(c), even the recently proposed MST method (Xu et al., 2025) cannot recover accurate object contours near ambiguous boundaries. This issue stems from the fact that camouflaged objects highly resemble their backgrounds in color, texture and structure, leading to faint visual transitions that hinder the localization of foreground–background boundaries. Consequently, predicted masks tend to suffer from incomplete contours or segmentation regions bleeding into the surrounding background. The second row of Fig. 1(c) further demonstrates that existing methods also struggle to separate foreground from background in visually ambiguous regions, resulting in local segmentation errors on objects. In such regions, similar appearance cues cannot supply sufficient evidence for robust foreground–background discrimination, and model predictions are prone to foreground omissions and background false positives. Without explicitly identifying and refining these error-prone regions, such confusion will persist and raise the user's interactive correction effort. Therefore, enhancing boundary awareness while effectively identifying and handling highly uncertain regions in camouflaged scenes is the key challenge for achieving efficient interactive camouflaged object segmentation.

To address these challenges, we propose an interactive camouflaged object segmentation method for camouflaged scenarios that improves segmentation accuracy and interaction efficiency by explicitly modeling boundary information and prediction uncertainty. Specifically, we design two key modules. First, to tackle the difficulty of recognizing boundary regions in camouflaged scenes, we introduce a Grassmann Boundary Guidance Module (GBGM), which explicitly predicts object

boundaries and incorporates them as guidance signals into the segmentation process, encouraging the model to focus more on the transition regions between foreground and background, thereby improving boundary localization and boundary segmentation quality. Second, to alleviate local prediction uncertainty caused by the high similarity between foreground and background, we propose an Ambiguity-Prior Refinement Module (APRM), which predicts an uncertainty map to identify potentially high-error regions and guides the model to prioritize these hard-to-segment areas, thus enhancing foreground–background discrimination and reducing the impact of erroneous segmentation on interactive correction efficiency. As illustrated in Fig. 1(b), our method achieves superior mean IoU under the same click budget, requiring fewer interactions to reach a given accuracy level. To the best of our knowledge, this work is the first to systematically investigate interactive camouflaged object segmentation, providing a new solution for efficiently constructing camouflaged object datasets.

The main contributions of this paper are summarized as follows:

- To the best of our knowledge, this is the first work exploring interactive camouflaged object segmentation, which aims to assist pixel-level annotation of camouflaged objects with a small amount of user interaction, thereby reducing the cost of constructing high-quality annotated data in camouflaged scenarios. This task offers a new perspective for efficiently building camouflaged object datasets.
- To address blurred boundaries and prediction ambiguity caused by high foreground–background similarity in camouflaged scenes, we propose an interactive segmentation method that jointly models boundary and ambiguity priors. Specifically, we design a Grassmann Boundary Guidance Module (GBGM) to enhance boundary awareness and an Ambiguity-Prior Refinement Module (APRM) to identify and refine regions with unreliable predictions, thereby improving foreground–background discrimination and reducing segmentation errors.
- Extensive experimental results on camouflaged object datasets demonstrate that the proposed method outperforms existing methods in both segmentation accuracy and interaction efficiency, validating its effectiveness and practical value in camouflaged scenarios.

2 Related Work

2.1 Camouflaged Object Segmentation

Camouflaged object segmentation aims to identify and delineate objects which appearance closely resembles their surroundings. It is relevant to applications such as species discovery and agricultural pest detection, and advances visual perception when foreground–background contrast is weak (Mei et al., 2021). In its dense prediction formulation, Camouflaged object segmentation produces pixel-level masks and therefore requires both target localization and accurate contour recovery. Fan et al. (Fan et al., 2020b) introduce the COD10K benchmark and a search-and-identification framework, establishing a foundation for learning and evaluating camouflage representations. Subsequent methods improve discrimination through more targeted reasoning. For example, PFNet (Mei et al., 2021) mines distracting foreground and background regions to refine object predictions.

Fully supervised COD relies on densely annotated data for learning and evaluation. Obtaining such annotations is particularly demanding in camouflaged scenes, where annotators must resolve weak contours and distinguish object structures from similar background textures. He et al. (He et al., 2023) explicitly identify this annotation burden and investigate scribble supervision to reduce the labeling required for training. Constructing accurate dense masks for new images remains a complementary annotation problem. Reducing the human effort involved in this process is therefore important for expanding COD datasets and supporting their practical use.

2.2 Interactive Image Segmentation

Interactive image segmentation offers a means of reducing dense annotation effort by converting sparse user prompts into object masks. An annotator identifies the target and corrects errors, while the model propagates these interactions over the image. RITM (Sofiiuk et al., 2022b) uses iterative training and previous-mask guidance to incorporate corrective clicks. SimpleClick (Liu

et al., 2023) employs a plain vision transformer with symmetric image and interaction embeddings, and MFP (Lee et al., 2024) strengthens the information conveyed by previous probability maps. MST (Xu et al., 2025) designs a multi-scale token adaptation scheme using token similarity to balance the segmentation accuracy of large and small targets. Together, these methods demonstrate how learned representations and prediction feedback can support efficient mask annotation.

Interactive camouflaged object segmentation aims to reduce annotation effort by generating accurate masks through sparse user corrections. However, identifying the target with clicks does not fully resolve weak boundaries or foreground–background confusion, motivating boundary recovery and feature refinement that adapt to the evolving interaction state.

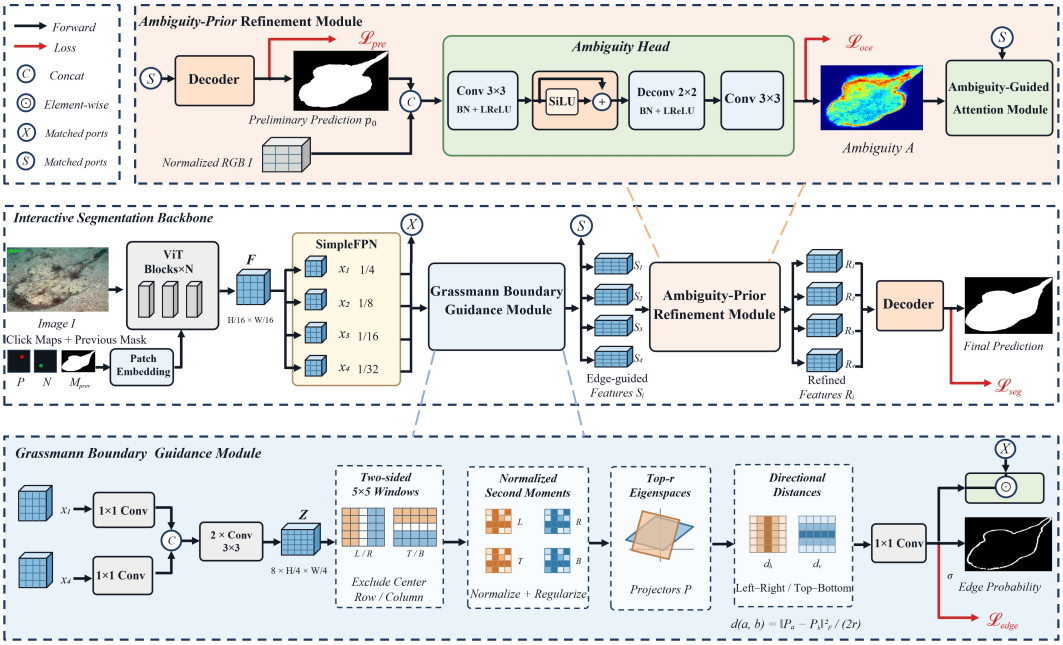

Figure 2: Framework of BAPNet. Multi-scale features are progressively refined by the Grassmann Boundary Guidance Module (GBGM) and the Ambiguity-Prior Refinement Module (APRM). GBGM models structural boundary transitions via Grassmann subspace comparison, while APRM identifies prediction-ambiguous regions for interaction-aware feature refinement.

3 Methodology

3.1 Overall Framework

Given an image $I \in \mathbb{R}^{3\times H\times W}$, accumulated clicks $\mathcal{C}^t$, and the previous-round mask M^{t-1}, BAPNet first encodes the positive and negative click maps together with M^{t-1} through a patch embedding layer. The interaction embeddings are incorporated into the image representation and processed by the ViT (Dosovitskiy et al., 2020) backbone to produce a feature map F at $1/16$ of the input resolution. SimpleFPN transforms F into multi-scale features $\{X_i\}_{i=1}^4$ at resolutions of $1/4$, $1/8$, $1/16$, and $1/32$, respectively. GBGM then estimates a boundary prior and uses it to obtain boundary-guided features $\{S_i\}_{i=1}^4$. These features are passed to APRM, which estimates the ambiguity of the current preliminary prediction and produces refined features $\{R_i\}_{i=1}^4$. Finally, the decoder decodes these features into the segmentation probability mask P^t, which provides mask feedback for the next interaction round.

3.2 Grassmann Boundary Guidance Module

Existing methods such as BTDGNet (Song et al., 2025) use edge information to guide camouflage feature learning, but weak boundary transitions and distracting textures still complicate boundary

estimation. To capture regional structural cues beyond pointwise appearance differences, GBGM compares local feature subspaces across opposing neighborhoods on the Grassmann manifold, producing a boundary prior for multi-scale feature enhancement.

To construct the features for this comparison, we select the finest and coarsest features X_1 and X_4 from SimpleFPN to combine spatial details for boundary localization with semantic context for distinguishing object contours. The two features are projected into a shared embedding space and spatially aligned before concatenation:

$$Z = \phi_Z\left(\mathcal{D}\left([\phi_4(X_4), \phi_1(X_1)]\right)\right) \in \mathbb{R}^{d\times H_1\times W_1}. \tag{1}$$

Here, ϕ_1 and ϕ_4 are 1×1 projection blocks, $[\cdot,\cdot]$ denotes channel-wise concatenation, and $\mathcal{D}$ is a depthwise separable convolution.

Using the embedding Z, we characterize structural differences by comparing feature distributions on opposite sides of each spatial position. For each position p, we construct four directional neighborhoods $s \in \{L, R, T, B\}$ within a $k\times k$ window. The horizontal pair excludes the center column, and the vertical pair excludes the center row, ensuring that each pair of opposing neighborhoods does not overlap. Each neighborhood contains $n = k\lfloor k/2\rfloor$ feature vectors, collected as $Z_s(p) \in \mathbb{R}^{d\times n}$.

We summarize each neighborhood using its uncentered second-moment matrix $M_s(p)$ and obtain a normalized, regularized structural matrix $C_s(p)$:

$$M_s(p) = \frac{1}{n} Z_s(p) Z_s(p)^\top, \qquad C_s(p) = \frac{M_s(p)}{\mathrm{tr}(M_s(p)) + \epsilon} + \rho D, \tag{2}$$

where $D = \mathrm{diag}\left(1,\ 1+\frac{1}{d-1},\ \ldots,\ 2\right)$ is shared by all neighborhoods. In Eq. equation 2, the uncentered second moment retains the mean feature direction, while trace normalization reduces sensitivity to overall feature energy. The diagonal perturbation regularizes eigenspace selection in low-energy or nearly degenerate neighborhoods.

From $C_s(p)$, we extract the orthonormal eigenvectors corresponding to its largest r eigenvalues to form $Q_s(p) \in \mathbb{R}^{d\times r}$. The resulting subspace is represented by the projector $\Pi_s(p) = Q_s(p)Q_s(p)^\top$ on the Grassmann manifold $\mathrm{Gr}(r, d)$. To compare the orientations of opposing subspaces independently of the choice of basis, we compute the normalized squared chordal distance in the horizontal direction as

$$\delta_h(p) = \frac{\|\Pi_L(p) - \Pi_R(p)\|_F^2}{2r} = 1 - \frac{\|Q_L(p)^\top Q_R(p)\|_F^2}{r}. \tag{3}$$

The vertical distance $\delta_v(p)$ is computed analogously between the top and bottom subspaces.

Both distances lie in $[0, 1]$ and measure directional disagreement between dominant local subspaces. A learned 1×1 classifier then converts these distances into boundary logits and probabilities:

$$z_E(p) = a_h\delta_h(p) + a_v\delta_v(p) + b_E, \qquad E(p) = \sigma(z_E(p)), \tag{4}$$

where a_h, a_v, and b_E are learnable parameters. The resulting boundary probability map E serves as a structural prior for enhancing boundary representations in multi-scale features.

We incorporate this prior into each pyramid level through residual spatial gating:

$$S_i = X_i \odot (1 + E_i).$$

Here, E_i is spatially aligned with feature X_i, and the residual form preserves the original representation while enhancing predicted boundary regions.

Eigenspace calculations use single precision during mixed-precision training. During the projector backward pass, an eigengap floor of 10^{-4} bounds the inverse-eigengap terms, providing a stabilized first-order approximation near a cutoff degeneracy.

3.3 Ambiguity-Prior Refinement Module

Boundary guidance alone cannot resolve errors within object interiors or visually similar background regions. Inspired by OCENet (Liu et al., 2022a), APRM estimates ambiguity from the image and current preliminary prediction to track regions requiring correction across interaction rounds. It

combines a preliminary decoder, an Ambiguity Head, and an Ambiguity-Guided Attention Module (AGAM) to estimate ambiguity and guide feature refinement.

As user interactions refine the segmentation, previously ambiguous regions may be resolved while errors persist elsewhere. An image-only estimate cannot capture these changes. We therefore condition ambiguity estimation on both the image I and the detached current preliminary prediction P_0, following OCENet (Liu et al., 2022b). The resulting map $A = \sigma(z_A)$ estimates where the current prediction is unreliable, including confident foreground omissions and background false positives. This interaction-dependent guidance enables subsequent refinement to focus on regions that still require correction.

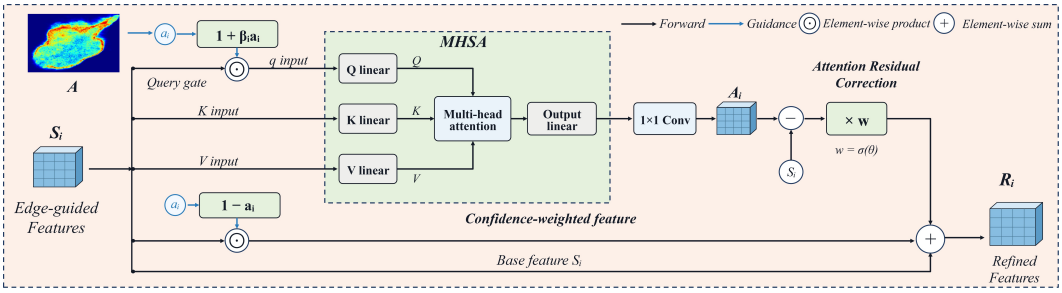

Figure 3: Ambiguity-Guided Attention Module.

To translate the estimated ambiguity into feature refinement, AGAM uses A to modulate contextual retrieval while retaining access to the complete feature map. This avoids directly suppressing ambiguous regions, which may discard useful object evidence. Let $T_i \in \mathbb{R}^{N_i \times C_i}$ and $a_i \in \mathbb{R}^{N_i \times 1}$ denote the flattened boundary-guided features S_i and the spatially aligned ambiguity map A_i, respectively. We first modulate the query input:

$$\widetilde{T}_i = T_i \odot (1 + \beta_i a_i), \tag{5}$$

where β_i is a learnable scale parameter for pyramid level i. For head h, learned projections yield $Q_i^h = \psi_{Q,i}^h(\widetilde{T}_i)$ from the modulated features and $K_i^h = \psi_{K,i}^h(T_i)$, $V_i^h = \psi_{V,i}^h(T_i)$ from the original features. Attention is computed as

$$O_i^h = \mathrm{softmax}\left(\frac{Q_i^h (K_i^h)^\top}{\sqrt{d_h}}\right) V_i^h, \tag{6}$$

where d_h is the head dimension. The multi-head outputs are projected and reshaped into S_i'. Query-only modulation makes contextual retrieval ambiguity-dependent while retaining the original features as keys and values.

To combine this contextual information with reliable local evidence, AGAM fuses the original features, an ambiguity-weighted local term, and a contextual residual:

$$R_i = S_i + S_i \odot (1 - A_i) + w(S_i' - S_i). \tag{7}$$

The second term reinforces reliable local evidence in low-ambiguity regions, while the last term introduces ambiguity-guided contextual refinement. We set the shared scalar w to 0.5 at initialization.

3.4 Training Objective

Let Y and G denote the foreground and edge annotations, respectively.

Boundary supervision. We supervise the aligned boundary probabilities $\widehat{E}$ using binary cross-entropy and Dice losses:

$$\mathcal{L}_E = \mathrm{BCE}(\widehat{E}, G) + 2\,\mathrm{Dice}(\widehat{E}, G). \tag{8}$$

BCE uses per-image class balancing and detached difficulty weights to handle sparse edges and hard pixels.

Ambiguity supervision. Following OCENet (Liu et al., 2022b), we derive a soft error target $A^\star$ from the detached preliminary mask P_0:

$$A^\star = Y \odot (1 - P_0) + (1 - Y) \odot P_0, \qquad \mathcal{L}_A = \mathrm{BCE}(z_A, A^\star). \tag{9}$$

Here, BCE operates on logits, and $A^\star$ captures foreground omissions and background false positives, including confident errors.

Overall objective. The total loss supervises both segmentation stages and both priors:

$$\mathcal{L} = \mathcal{L}_{\mathrm{seg}}(P, Y) + \mathcal{L}_{\mathrm{seg}}(P_0, Y) + \mathcal{L}_E + \mathcal{L}_A. \tag{10}$$

All four terms have equal outer weights. We adopt normalized focal loss (Sofiiuk et al., 2022a) for $\mathcal{L}_{\mathrm{seg}}$, with $\alpha = 0.5$ and $\gamma = 2$. Separate supervision of P_0 supports its training despite detachment before ambiguity estimation.

4 Experiments

4.1 Experimental Settings

Datasets: We evaluate BAPNet on four standard camouflaged object segmentation datasets, CAMO (Le et al., 2019), with 1,000 training and 250 test camouflaged images; CHAMELEON (Skurowski et al., 2018), containing 76 camouflaged images; COD10K (Fan et al., 2020a), comprising 10,000 images with a standard train/test split; and NC4K (Lv et al., 2021), providing 4,121 images for testing.

Evaluation Metrics: We adopt two metrics for evaluation: Number of Clicks (NoC) and AUC. **NoC** measures the average clicks needed to attain a target IoU. Following prior works (Xu et al., 2025; Lee et al., 2024), we report **NoC@85** and **NoC@90**. To better distinguish high-performance models, we introduce **NoC@95** for a stricter criterion, representing the average clicks required to reach 95% IoU. **AUC** denotes the area under the IoU-click curve, which integrates segmentation performance across all click budgets and offers a measure of interactive segmentation efficiency.

Implementation Details: All experiments are conducted on a server equipped with an NVIDIA RTX 4090 GPU. BAPNet is trained for 55 epochs with input images resized to 448×448 pixels, using the AdamW optimizer with a learning rate of 1×10^{-5}. More implementation details are left supplementary due to space limitation.

Table 1: Comparison of Methods on CAMO, CHAMELEON, COD10K, NC4K. The metrics NOC mean the average number of clicks required for the target IoU.

Method	Backbone	CAMO			CHAMELEON			COD10K			NC4K		
		NOC85	NOC90	NOC95	NOC85	NOC90	NOC95	NOC85	NOC90	NOC95	NOC85	NOC90	NOC95
Unsupervised: Training-Free													
M2n2$_{\text{CVPR25}}$	SD2	9.89	12.78	18.31	8.53	11.91	16.08	9.39	13.10	17.89	6.99	10.46	16.50
Supervised: Trained on Ground Truth Labels													
EMC-Click$_{\text{CVPR23}}$	HRnet-18	5.54	8.45	15.21	5.39	7.68	13.87	7.88	11.22	16.44	5.01	7.74	13.78
FCFI$_{\text{CVPR23}}$	ResNet-101	5.69	8.78	15.57	5.75	7.74	12.84	6.70	9.91	15.62	4.92	7.52	13.36
SimpleClick$_{\text{ICCV23}}$	ViT-B	3.67	5.92	13.36	5.00	6.86	12.46	5.70	8.86	14.81	3.71	6.22	12.24
InterFormer$_{\text{ICCV23}}$	ViT-B	6.40	8.73	14.87	5.55	7.34	12.64	7.19	10.32	15.63	5.64	8.31	13.98
iCMFormer$_{\text{ICCV23}}$	ViT-B	3.88	6.16	13.50	5.09	7.16	12.37	5.91	9.09	14.84	3.79	6.29	12.20
SegNext$_{\text{CVPR24}}$	ViT-B	4.88	7.68	14.63	4.75	6.67	12.18	6.05	9.36	15.28	4.74	7.40	13.65
MFP$_{\text{CVPR24}}$	ViT-B	3.99	6.54	13.81	4.95	6.92	12.59	6.22	9.42	15.20	4.21	6.83	13.04
MST$_{\text{TCSVT25}}$	ViT-B	3.13	5.36	12.43	4.63	6.78	12.20	5.37	8.54	14.55	3.46	5.81	11.75
Ours	**ViT-B**	**2.92**	**5.08**	**12.32**	**4.22**	**6.54**	**11.41**	**4.95**	**8.01**	**14.21**	**3.12**	**5.44**	**11.26**

4.2 Comparison with Previous Methods

We compare BAPNet with the training-free method M2N2 (Karmann & Urfalioglu, 2025) and eight supervised interactive segmentation methods: EMC-Click (Du et al., 2023), FCFI (Wei et al., 2023), SimpleClick (Liu et al., 2023), InterFormer (Huang et al., 2023), iCMFormer (Li et al., 2023), SegNext (Liu et al., 2024), MFP (Lee et al., 2024), and MST (Xu et al., 2025). The supervised methods are trained on the combined training splits of COD10K and CAMO.

Quantitative Results. Table 1 reports the average number of clicks required to reach 85%, 90%, and 95% IoU on CAMO, CHAMELEON, COD10K, and NC4K. Across all methods, the click requirement increases substantially at higher IoU thresholds, reflecting the difficulty of correcting the remaining segmentation errors. BAPNet requires the fewest clicks in all 12 settings. Compared with MST, it reduces NoC@90 by 5.2%, 3.5%, 6.2%, and 6.4% on the four benchmarks, respectively. Its advantage also holds at 95% IoU, with NoC@95 reductions of 6.5% on CHAMELEON and 4.2% on NC4K. These results show that BAPNet reduces interaction effort not only for coarse object localization but also for obtaining more accurate masks.

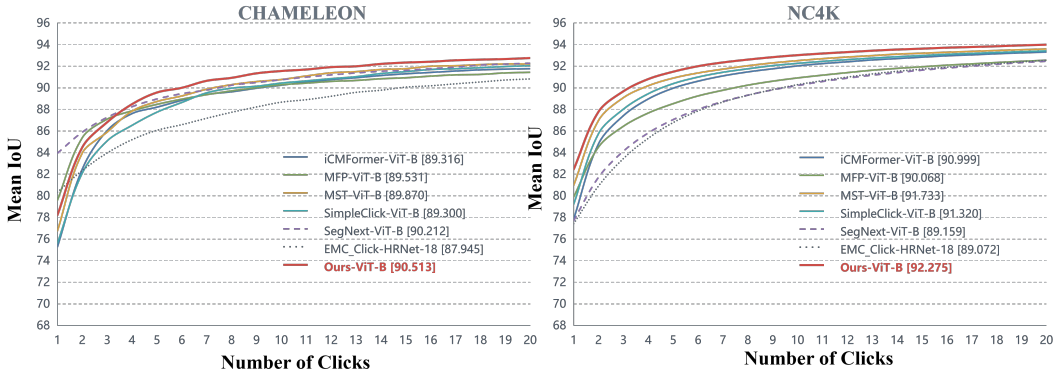

Figure 4: The average Intersection over Union (IoU) varies with the number of clicks , indicating that the proposed method can achieve higher precision with fewer clicks.

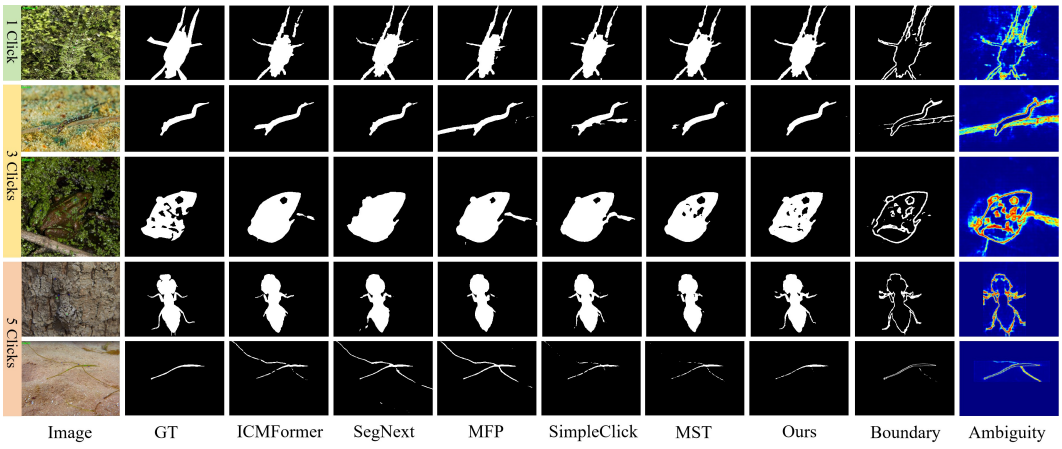

Figure 5: Visual comparison of segmentation masks generated by different interactive segmentation methods on challenging camouflaged samples under 1, 3 and 5 clicks.

Comparison of IoU and AUC. To investigate how segmentation accuracy changes with user interaction, Fig. 4 compares mean IoU across different click budgets. BAPNet generally produces more accurate masks given the same number of clicks. The AUC scores in the figure legends summarize these curves, with BAPNet obtaining the highest scores on the reported datasets. Together with the NoC results, this shows that its improvement extends across click budgets rather than being restricted to a particular accuracy threshold.

Qualitative Results. Fig. 5 compares the predicted masks under 1, 3, and 5 clicks. The differences are particularly visible in thin structures and regions with weak foreground–background contrast, where competing methods omit object parts or include surrounding background. Even MST leaves some fine structures incompletely segmented, whereas BAPNet recovers masks that more closely follow the ground-truth contours. These examples illustrate the practical benefit of combin-

Table 2: Ablation experiments on four publicly available camouflaged object datasets. The metrics NOC mean the average number of clicks required for the target IoU.

Method	CAMO			CHAMELEON			COD10K			NC4K		
	NOC85↓	NOC90↓	NOC95↓	NOC85↓	NOC90↓	NOC95↓	NOC85↓	NOC90↓	NOC95↓	NOC85↓	NOC90↓	NOC95↓
Baseline	3.14	5.31	12.48	4.58	6.88	11.86	5.28	8.34	14.46	3.36	5.71	11.58
B+GBGM	3.08	5.08	12.32	4.38	6.57	11.45	5.04	8.06	14.28	3.21	5.54	11.38
B+APRM	2.93	5.09	12.38	4.38	6.54	11.63	5.09	8.11	14.30	3.17	5.53	11.42
BAPNet(Full)	**2.92**	**5.08**	**12.32**	**4.22**	**6.54**	**11.41**	**4.95**	**8.01**	**14.21**	**3.12**	**5.44**	**11.26**

Table 3: Ablation study of boundary estimation strategies.The metrics NOC mean the average number of clicks required for the target IoU.

Boundary estimation	COD10K			NC4K		
	NoC@85↓	NoC@90↓	NoC@95↓	NoC@85↓	NoC@90↓	NoC@95↓
Conv. head	5.18	8.14	14.37	3.34	5.67	11.51
Mean contrast	5.21	8.19	14.42	3.43	5.65	11.52
Second-moment contrast	5.13	8.16	14.34	3.38	5.59	11.43
Grassmann contrast (Ours)	**5.04**	**8.06**	**14.28**	**3.21**	**5.54**	**11.38**

ing boundary guidance with ambiguity-aware refinement: better contour recovery and fewer local foreground–background errors under limited interaction.

4.3 Ablation Study

Effectiveness of GBGM. Table 2 shows that adding GBGM to the baseline reduces NoC across all four datasets and all three IoU thresholds. For example, GBGM lowers NoC@90 by 4.3% on CAMO and 4.5% on CHAMELEON. These consistent improvements support the effectiveness of the proposed boundary guidance in reducing the corrective interactions required for accurate segmentation.

Effectiveness of APRM. Adding APRM alone also improves every evaluated metric, reducing NoC@85 by 6.7% on CAMO and 5.7% on NC4K relative to the baseline. This demonstrates the benefit of using prediction-conditioned ambiguity to guide feature refinement. Combining APRM with GBGM yields a 7.1% reduction in NoC@85 on NC4K over the baseline. The full model achieves the best or tied-best results in all 12 settings, supporting the complementary contributions of the two modules.

To evaluate boundary estimation based on comparing local feature subspaces across opposing neighborhoods, we compare Grassmann contrast with a convolutional head, mean contrast, and second-moment contrast, keeping the remaining components and training settings unchanged. As shown in Table 3, Grassmann contrast achieves the lowest NoC at all three thresholds on both datasets. On NC4K, it reduces NoC@85 by 3.9%, 6.4%, and 5.0% relative to the three alternatives, respectively. These results support the effectiveness of comparing local feature subspaces for boundary estimation, beyond using mean features or second-moment matrices directly.

5 Conclusion

We propose BAPNet for interactive camouflaged object segmentation, combining Grassmann-based boundary guidance with prediction-conditioned ambiguity refinement. Extensive experiments on four benchmarks demonstrate state-of-the-art segmentation accuracy with fewer corrective clicks than the compared methods, including at 95% IoU. These results highlight the complementary benefits of boundary and ambiguity priors for efficient camouflage annotation. To the best of our knowledge, BAPNet is the first systematic study of interactive camouflaged object segmentation.

AI USE STATEMENT

In this work, we used generative AI tools to assist with brainstorming and refining research ideas, and assisting with the implementation and debugging of research code. Generative AI tools were also used for language editing, improving the readability and organization of the manuscript, and providing suggestions for the presentation and design of scientific figures. In addition, these tools were occasionally used to assist with literature search and summarization. All research hypotheses, methodological choices, experimental settings, and reported results were ultimately determined and verified by the authors. AI-assisted code was manually inspected, integrated, and tested by the authors, and all experimental results reported in this paper were obtained from our own implementations and experiments. AI-assisted literature suggestions were manually checked against the original publications, and all citations were verified by the authors. AI-generated or AI-edited text and figure suggestions were reviewed and revised before inclusion in the manuscript. Generative AI tools were not used to fabricate experimental data, generate unverified experimental results, or autonomously draw scientific conclusions. We take full responsibility for the final content of this work, including all text, claims, code, figures, and experimental results produced with the aid of generative AI.

ETHICS STATEMENT

This paper focuses on interactive camouflaged object segmentation. The datasets used are existing publicly released computer vision benchmarks, and the research does not involve human subjects, privacy risks or harmful real-world applications. No ethical concerns are identified for this work.

REPRODUCIBILITY STATEMENT

We have made efforts to facilitate the reproducibility of our results. Section 4.1 describes the datasets, evaluation metrics, training configuration, and major implementation settings used in our experiments. The model architecture and the formulations of the Grassmann Boundary Guidance Module (GBGM) and Ambiguity-Prior Refinement Module (APRM) are described in detail in Sections 3.2 and 3.3. Additional implementation details, including the interactive click simulation protocol, training settings, module hyperparameters, and experimental configurations, are provided in the supplementary material. The ablation studies report the effects of the major components and design choices under consistent experimental settings. Code and configuration files necessary to reproduce the main experiments will be released upon publication.

REFERENCES

Junjie Chen, Junwei Lin, Ren Hong, Shengjie Liu, Yuming Fang, Feng Qian, and Yifan Zuo. Learning and aligning click-aware shape prior for interactive amodal instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20478–20487, 2026.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, abs/2010.11929, 2020. URL https://arxiv.org/abs/2010.11929.

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.

Deng-Ping Fan, Ge-Peng Ji, Guolei Sun, Ming-Ming Cheng, Jianbing Shen, and Ling Shao. Camouflaged object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2777–2787, 2020a.

Deng-Ping Fan, Ge-Peng Ji, Guolei Sun, Ming-Ming Cheng, Jianbing Shen, and Ling Shao. Camouflaged object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2777–2787, June 2020b. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Fan_Camouflaged_Object_Detection_CVPR_2020_paper.html.

Ruozhen He, Qihua Dong, Jiaying Lin, and Rynson W. H. Lau. Weakly-supervised camouflaged object detection with scribble annotations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 781–789, 2023. doi: 10.1609/aaai.v37i1.25156. URL https://ojs.aaai.org/index.php/AAAI/article/view/25156.

Kangpeng Hu, Quansen Sun, Yinghui Sun, and Tao Wang. Interactive segmentation by considering first-click intentional ambiguity. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 4823–4831, 2024.

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.

Markus Karmann and Onay Urfalioglu. Repurposing stable diffusion attention for training-free unsupervised interactive segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 24518–24528, 2025.

Trung-Nghia Le, Tam V Nguyen, Zhongliang Nie, Minh-Triet Tran, and Akihiro Sugimoto. Anabranch network for camouflaged object segmentation. *Computer vision and image understanding*, 184:45–56, 2019.

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

Kun Li, George Vosselman, and Michael Ying Yang. Interactive image segmentation with cross-modality vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 762–772, 2023.

Jiawei Liu, Jing Zhang, and Nick Barnes. Modeling aleatoric uncertainty for camouflaged object detection. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2613–2622, 2022a. doi: 10.1109/WACV51458.2022.00267.

Jiawei Liu, Jing Zhang, and Nick Barnes. Modeling aleatoric uncertainty for camouflaged object detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1445–1454, January 2022b. URL https://openaccess.thecvf.com/content/WACV2022/html/Liu_Modeling_Aleatoric_Uncertainty_for_Camouflaged_Object_Detection_WACV_2022_paper.html.

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 (CVPR)*, 2024. URL https://arxiv.org/abs/2404.00741.

Yunqiu Lv, Jing Zhang, Yuchao Dai, Aixuan Li, Bowen Liu, Nick Barnes, and Deng-Ping Fan. Simultaneously localize, segment and rank the camouflaged objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11591–11601, 2021.

Haiyang Mei, Ge-Peng Ji, Ziqi Wei, Xin Yang, Xiaopeng Wei, and Deng-Ping Fan. Camouflaged object segmentation with distraction mining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8772–8781, June 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Mei_Camouflaged_Object_Segmentation_With_Distraction_Mining_CVPR_2021_paper.html.

Youwei Pang, Xiaoqi Zhao, Jiaming Zuo, Lihe Zhang, and Huchuan Lu. Open-vocabulary camouflaged object segmentation. In *European Conference on Computer Vision*, pp. 476–495. Springer, 2024.

Przemysław Skurowski, Hassan Abdulameer, Jakub Błaszczyk, Tomasz Depta, Adam Kornacki, and Przemysław Kozieł. Animal camouflage analysis: Chameleon database. *Unpublished manuscript*, 2(6):7, 2018.

Konstantin Sofiiuk, Ilya A. Petrov, and Anton Konushin. Reviving Iterative Training with Mask Guidance for Interactive Segmentation. In *2022 IEEE International Conference on Image Processing (ICIP)*, pp. 3141–3145, 2022a. doi: 10.1109/icip46576.2022.9897365. URL https://doi.org/10.1109/icip46576.2022.9897365.

Konstantin Sofiiuk, Ilya A. Petrov, and Anton Konushin. Reviving iterative training with mask guidance for interactive segmentation. In *2022 IEEE International Conference on Image Processing (ICIP)*, pp. 3141–3145. IEEE, 2022b. doi: 10.1109/ICIP46576.2022.9897365. URL https://doi.org/10.1109/ICIP46576.2022.9897365.

Xiaogang Song, Pengfei Zhang, Xiaochang Li, Xinhong Hei, and Rongrong Liu. Btdgnet: A dual-guided camouflaged object detection network leveraging boundary and texture information. *IEEE Transactions on Multimedia*, 2025.

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.

Ranwan Wu, Tian-Zhu Xiang, Guo-Sen Xie, Rongrong Gao, Xiangbo Shu, Fang Zhao, and Ling Shao. Uncertainty-aware transformer for referring camouflaged object detection. *IEEE Transactions on Image Processing*, 2025.

Long Xu, Yongquan Chen, Shanghong Li, and Jun Luo. MST: Adaptive Multi-Scale Tokens Guided Interactive Segmentation. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(8):8148–8160, 2025. doi: 10.1109/tcsvt.2025.3546634. URL https://doi.org/10.1109/tcsvt.2025.3546634.

Guanghui Yue, Shangjie Wu, Tianwei Zhou, Gang Li, Jie Du, Yu Luo, and Qiuping Jiang. Progressive region-to-boundary exploration network for camouflaged object detection. *IEEE Transactions on Multimedia*, 27:236–248, 2024.

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.

A Additional Implementation Details

We implement BAPNet in PyTorch using ViT-B as the backbone, initialized from pretrained weights. SimpleFPN produces four feature levels with 128, 256, 512, and 1024 channels at $1/4$, $1/8$, $1/16$, and $1/32$ resolutions, respectively. All input images are resized and cropped to 448×448. The model is trained for 55 epochs using AdamW with an initial learning rate of 1×10^{-5}. During training, standard data augmentation, including random scaling, horizontal flipping, random cropping, brightness/contrast perturbation, and RGB shifting, is applied. Positive and negative clicks are encoded as disk maps with a normalized radius of 5, and the previous-round prediction is provided as an additional interaction input. We allow at most 24 interaction points and simulate up to three additional corrective clicks during iterative training.

For GBGM, we use a 5×5 local window, an 8-dimensional embedding space, and a Grassmann subspace rank of $r = 2$. Both the regularization coefficient and the eigengap floor are set to 10^{-4}. For APRM, the ambiguity estimator takes the normalized RGB image together with the detached current preliminary prediction as input and uses 16 intermediate channels. The predicted ambiguity map is used to refine all four FPN levels through ambiguity-guided multi-head attention. Both the preliminary and final segmentation predictions are supervised using normalized focal loss with $\alpha = 0.5$ and $\gamma = 2$, while the boundary and ambiguity branches are supervised by their corresponding losses. All loss terms are equally weighted unless otherwise specified.