SCOUT: Semi-supervised Camouflaged Object Detection by Utilizing Text and Adaptive Data Selection

Weiqi Yan[1], **Lvhai Chen**[1], **Shengchuan Zhang**[1*], **Yan Zhang**[1] and **Liujuan Cao**[1]

[1]Key Laboratory of Multimedia Trusted Perception and Efficient Computing, Ministry of Education of China, Xiamen University, 361005, P.R. China.
weiqi_yan@outlook.com, lvhaichen2002@gmail.com, zsc_2016@xmu.edu.cn

Abstract

The difficulty of pixel-level annotation has significantly hindered the development of the Camouflaged Object Detection (COD) field. To save on annotation costs, previous works leverage the semi-supervised COD framework that relies on a small number of labeled data and a large volume of unlabeled data. We argue that there is still significant room for improvement in the effective utilization of unlabeled data. To this end, we introduce a **S**emi-supervised **C**amouflaged **O**bject Detection by **U**tilizing **T**ext and Adaptive Data Selection (**SCOUT**). It includes an Adaptive Data Augment and Selection (ADAS) module and a Text Fusion Module (TFM). The ADSA module selects valuable data for annotation through an adversarial augment and sampling strategy. The TFM module further leverages the selected valuable data by combining camouflage-related knowledge and text-visual interaction. To adapt to this work, we build a new dataset, namely RefTextCOD. Extensive experiments show that the proposed method surpasses previous semi-supervised methods in the COD field and achieves state-of-the-art performance. Our code will be released at https://github.com/Heartfirey/UCOD-DPL.

1 Introduction

Camouflaged object detection (COD) [Fan *et al.*, 2020; Fan *et al.*, 2022] aims at segmenting objects that are visually concealed in their surroundings, which has important applications in several fields, such as military [Cannaday *et al.*, 2023], environmental monitoring [Yadav *et al.*, 2018], urban security [Li *et al.*, 2024] etc. Existing COD methods always require large amounts of labeled data to segment camouflaged objects precisely [Mei *et al.*, 2021].

However, camouflaged objects often employ complex camouflage strategies to blend deeply with their background. Such camouflage strategies make pixel-level camouflaged object annotations difficult to obtain, and annotating the entire dataset requires a significantly greater cost. To reduce the annotation costs, semi-supervised COD methods utilize only a small amount of labeled data alongside a large amount of unlabelled data. However, as shown in Figure 1, we identified the following two main issues with existing methods [Lai *et al.*, 2024; Zhang *et al.*, 2024]: 1). Most existing methods rely on random sampling to select a portion of the data as the labeled set, without thoroughly considering the quality of the selected data. This often results in the annotation of some meaningless data. 2). For the selected small amount of labeled data, existing models struggle to fully learn camouflage-related knowledge, resulting in poor segmentation performance.

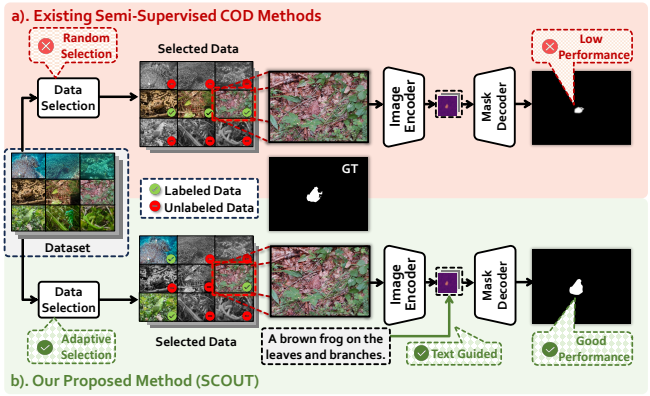

Figure 1: Comparison between the proposed method SCOUT and previous semi-supervised COD methods. The proposed SCOUT introduces an adaptive selection strategy and text guidance strategy and achieves a better segmentation performance.

Therefore, we argue that there is still significant room for improvement in the effectiveness of unlabeled data utilization. As shown in Figure 1, we propose a SCOUT model that introduces an adaptive selection strategy and a text guidance strategy. Specifically, the adaptive selection strategy is implemented by the Adaptive Data Augment and Selection(ADAS) module. The ADAS module adaptively selects valuable training data through the Adaptive Data Augment (ADA) and the Adaptive Data Selection (ADS) component. The text guidance strategy is implemented by the Text Fusion Module (TFM). To help the model fully learn camouflage-related knowledge, TFM introduces referring text assistance

*Corresponding author

and enhances representation ability, thereby improving the model's performance across various camouflage scenarios.

By annotating precise image-level referring text for existing mainstream COD datasets, we develop a new dataset called RefTextCOD. Our comprehensive evaluations on this dataset demonstrate substantial enhancements over existing semi-supervised COD models. Especially, our method improves the Mean Absolute Error (MAE) by 52.0% and the S-Measure by 19.1% compared to previous SOTA semi-supervised COD methods. Additionally, our approach outperforms some supervised COD methods, highlighting its greater practical applicability.

Our main contributions are summarized as follows:

- We proposed an innovative semi-supervised COD model SCOUT, and extensive experiments have demonstrated its high performance and effectiveness.
- To avoid meaningless data annotations, we proposed an ADAS module, which selects valuable data through an adversarial augment and sampling strategy.
- To fully utilize the referring text, we proposed a TFM module, which leverages the selected valuable data by combining camouflage-related knowledge and text-visual interaction.
- We build a new dataset, namely RefTextCOD. We performed image-level referring text annotations on existing mainstream camouflaged object detection (COD) datasets, providing a data foundation for this paper and future explorations of other COD tasks.

2 Related Works

2.1 Camouflaged Object Detection

Camouflaged object detection (COD) has a long-standing history, and deep learning-based COD methods have seen rapid development in recent years. Existing fully-supervised COD methods have already achieved great performance by employing multiple strategies. For example, [Pang *et al.*, 2022; Pang *et al.*, 2023] extract multi-scale features from the backbone and design strategies for fusion. [Fan *et al.*, 2022; Jia *et al.*, 2022; Zhang *et al.*, 2022] further use multi-stage refine. Some methods introduce additional information, *e.g.* boundary guidance [Sun *et al.*, 2022; Ji *et al.*, 2023; Zhai *et al.*, 2021], texture clues [Ji *et al.*, 2023; Zhu *et al.*, 2021; Ren *et al.*, 2023], and other information such as frequency domain and depth [Zhong *et al.*, 2022; Lin *et al.*, 2023].

The fully-supervised COD methods heavily rely on a large amount of labeled data, while pixel-level annotation incurs significant costs. As a result, some unsupervised, weakly-supervised, and semi-supervised methods have emerged. UCOS-DA [Zhang and Wu, 2023] is the first unsupervised COD method that addresses the task as a domain adaptation problem. SCOD [Zhang *et al.*, 2024] is the first weakly supervised COD method, which introduces a novel feature-guided loss and consistency loss with a new scribble learning approach. Semi-supervised COD methods have also received considerable attention from researchers. This paper focuses on investigating the rationality of data selection and the sufficiency of data utilization in semi-supervised COD methods.

2.2 Semi-Supervised Camouflaged Object Detection

In traditional fully-supervised learning, models require extensive labeled data for training to achieve optimal performance. However, obtaining labeled data in practical applications is often costly and time-consuming. Semi-supervised learning (SSL) enhances the model's generalization capability by combining labeled data and unlabeled data [Chen *et al.*, 2023; Grandvalet and Bengio, 2004; Mi *et al.*, 2022], which effectively addresses the challenges of acquiring labeled data. Some previous works [Chen *et al.*, 2021; Sohn *et al.*, 2020; Wang *et al.*, 2022] introduce the pseudo-labeling mechanism, where a teacher model is trained using a small amount of labeled data, and then the teacher model is used to produce the pseudo labels of the unlabeled data for the subsequent training of the student model. Semi-supervised learning has shown significant potential in various fields.

In semi-supervised COD methods, CamoTeacher [Lai *et al.*, 2024] applies different augmentation strategies to teacher-student networks and introduces a dual-rotation consistency loss. SCOD-ND [Fu *et al.*, 2024] proposes a window-based voting strategy and an ensemble learning algorithm to eliminate noise from labels. However, these methods typically split the dataset into labeled sets and unlabeled sets by random sampling, without considering the value of the selected data. Moreover, they do not make full use of the labeled data. In this paper, We adaptively select valuable samples for training and introduce referring text to assist the model in learning camouflage-related knowledge from the labeled data.

2.3 Referring Camouflaged Object Detection

The concept of Referring Camouflaged Object Detection (Ref-COD) was first proposed by [Zhang *et al.*, 2023], which leverages a batch of images as the referring information to guide the identification of the specified camouflaged objects. With the development of MLLMs, the rich intrinsic knowledge that MLLMs learned from massive amounts of data can be used to augment a variety of downstream tasks. Recently works [Cheng *et al.*, 2023; Hu *et al.*, 2025] have extended this concept by utilizing MLLMs and designing a series of prompts to assist the COD task. In this paper, we focus on how to use precise referring text to assist the model in learning camouflage-related knowledge from labeled data.

2.4 Active Learning

Active learning in the context of deep learning is a strategy designed to reduce the labeling cost by allowing the model to selectively query the most informative samples from an unlabeled dataset for annotation. It is particularly useful in scenarios where data labeling is expensive, time-consuming, or requires expert knowledge. Active learning approaches can be divided into membership query synthesis [Angluin, 1988; King *et al.*, 2004], stream-based selective sampling [Argamon-Engelson and Dagan, 1999], and pool-based active learning from application scenarios [Settles, 2009]. In this paper, we use an adversarial augmentation and selection strategy to select valuable data that the model can partially understand but still needs further learning.

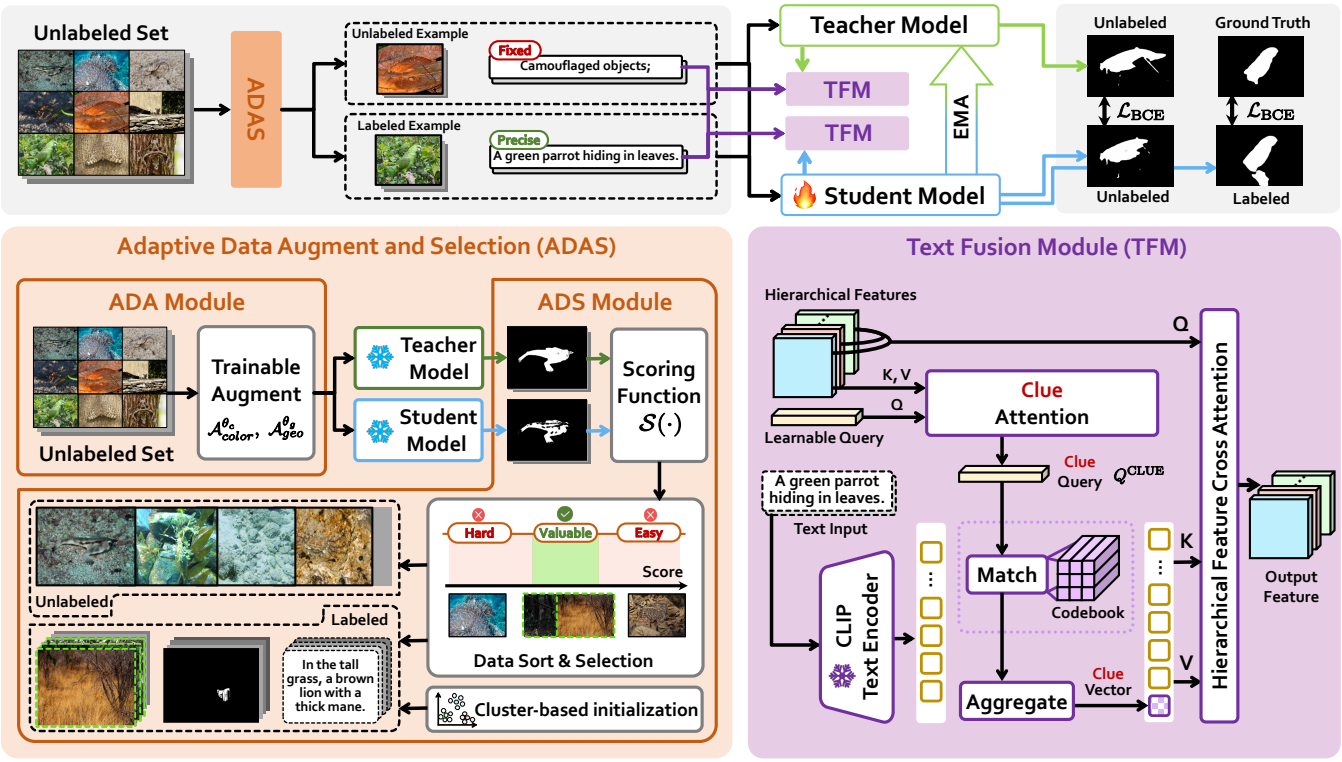

Figure 2: **Overview of the proposed SCOUT.** The SCOUT is an innovative semi-supervised COD model, which consists of an ADAS module and a TFM module.

3 RefTextCOD Dataset

To adapt to this work, we build a new dataset, namely RefTextCOD. The motivation behind this proposed dataset is that existing datasets do not have image-level referring text.

To achieve a fair and effective comparison with existing methods, we expect to be able to construct text-based referring COD experiments in settings that are as similar as possible. Referring to [Chen *et al.*, 2022; Fan *et al.*, 2022], we used the mainstream COD datasets: CHAMELEON ([Wu *et al.*, 2019]), CAMO ([Yan *et al.*, 2021; Le *et al.*, 2019]), COD10K ([Fan *et al.*, 2020]), NC4K ([Le *et al.*, 2019]) as the base image data. This allows us to focus more on annotations without collecting camouflaged images from scratch.

It is time-consuming to annotate camouflaged objects with captions, and manual annotating tends to lead to inconsistent quality. To ensure that the camouflaged object captions contain meaningful information for COD tasks (*i.e.* texture, color, and shape), we used two vision language models (*e.g.* QwenVL [Bai *et al.*, 2023] and GPT4-Vision [OpenAI, 2024]) to generate a summary first. Specifically, we design a series of prompts with contextual logic: 1). First, the VLM model will be guided to locate the object and justify its classes; 2). Then, the model is directed to characterize the physical properties of foreground objects and background. 3). Finally, we request the model to aggregate and streamline all the features to generate complete referring text.

Finally, we employing manual screening and conducted a thorough review and revision of all annotations to ensure their accuracy and quality. As a result, we annotated **9,487** images from four datasets and organized them into a new dataset called RefTextCOD. The prompt used in the above process, the whole pipeline will be provided in the appendix.

4 Methods

The overview of SCOUT is depicted in Figure 2. In Section 4.1, we introduce the Adaptive Data Augment and Selection (ADAS) module. The module selects valuable data for annotation through an Adaptive Data Augment (ADA) component and an Adaptive Data Selection (ADS) component. In Section 4.2, we present the Text Fusion Module (TFM). This module conducts semantic knowledge transfer and text-visual interaction to make full use of the selected valuable data. We provide the details of the total loss and training process in Section 4.3.

4.1 Adaptive Data Augment and Selection

During this phase, our objective is to select valuable data from an unlabeled set $\mathcal{D}_{\mathrm{U}} = \{\mathrm{X}^I_{u,j}\}^M_{j=1}$ for semi-supervised training, represented by the ADAS module. The ADAS module contains two sub-components: an Adaptive Data Augmentation (ADA) component and an Adaptive Data Selection (ADS) component. The ADA is a trainable data augmentation module through an adversarial augment strategy, while the ADS finds valuable data via an adversarial sampling strategy.

Formally, given the unlabeled set $\mathcal{D}_{\mathrm{U}} = \{\mathrm{X}^I_{u,j}\}^M_{j=1}$, the ADA component adopts a set of parameterizable data aug-

menter [Suzuki, 2022] $\{\mathcal{A}^{\theta_c}_{color}, \mathcal{A}^{\theta_g}_{geo}\}$ (*i.e.* , color augmentation and geometric transformation) to augment each unlabeled data $\mathrm{X}^I_{u,j}$,

$$\mathrm{X}^{I,\mathrm{Aug}}_{u,j} = \mathcal{A}^{\theta_c}_{color} \circ \mathcal{A}^{\theta_g}_{geo}(\mathrm{X}^I_{u,j}), \tag{1}$$

where $\mathrm{X}^{I,\mathrm{Aug}}_{u,j}$ denotes the augmented unlabeled data. We feed $\mathrm{X}^{I,\mathrm{Aug}}_{u,j}$ into both teacher and student models, parameterized by θ_T and θ_S, to obtain segmentation masks $\hat{\mathrm{Y}}^T_{u,j}$ and $\hat{\mathrm{Y}}^S_{u,j}$,

$$\hat{\mathrm{Y}}^T_{u,j} = \mathcal{F}(\mathrm{X}^{I,\mathrm{Aug}}_{u,j}; \theta_T),\ \hat{\mathrm{Y}}^S_{u,j} = \mathcal{F}(\mathrm{X}^{I,\mathrm{Aug}}_{u,j}; \theta_S), \tag{2}$$

where $\mathcal{F}(\mathrm{X}^{I,\mathrm{Aug}}_{u,j}; \theta_T), \mathcal{F}(\mathrm{X}^{\mathrm{I,Aug}}_{u,j}; \theta_S)$ denote the predictions of the teacher model and student model, respectively, on augmented unlabeled data $\mathrm{X}^{I,\mathrm{Aug}}_{u,j}$.

After obtaining the segmentation masks $\hat{\mathrm{Y}}^T_{u,j}$ and $\hat{\mathrm{Y}}^S_{u,j}$, the ADS calculates the score of each unlabeled data $\mathrm{X}^I_{u,j}$ and selects valuable data for annotation. Specifically, we first use a scoring function $\mathcal{S}(\cdot)$ to calculate the unlabeled data $\mathrm{X}^I_{u,j}$ scores,

$$\mathcal{S}(\mathrm{X}^I_{u,j}) = \mathrm{SSIM}(\hat{\mathrm{Y}}^T_{u,j}, \hat{\mathrm{Y}}^S_{u,j}) - \mathrm{MAE}(\hat{\mathrm{Y}}^T_{u,j}, \hat{\mathrm{Y}}^S_{u,j}), \tag{3}$$

where $\mathrm{SSIM}(\cdot)$ denotes the structural similarity metrics, and $\mathrm{MAE}(\cdot)$ denotes the mean absolute error. The ADS component uses Kernel Density Estimation (KDE) normalization to normalize all scores and uniformly map them to the $[0, 1]$ interval, then it sorts the unlabeled set $\mathcal{D}_{\mathrm{U}}$ based on these scores. At this point, data with scores close to 0 are considered hard data, while those close to 1 are considered easy data. Among these ordered data, extremely low-scoring hard data are too complex for the model or contain noise that hinders its convergence. On the other hand, extremely high-scoring easy data are of no value to train the model and contribute to overfitting. Therefore, the ADS component selects data with moderate scores that are close to 0.5, which the model can partially understand but still needs further learning. Specifically, the ADS component samples the unlabeled data in unlabeled set $\mathcal{D}_{\mathrm{U}} = \{\mathrm{X}^I_{u,j}\}^M_{j=1}$ with scores close to 0.5. Those selected unlabeled data $\mathrm{X}^I_{l,i}$ will undergo referring text annotation $\mathrm{X}^T_{l,i}$ and segmentation mask annotation $\mathrm{Y}_{l,i}$ to construct the labeled set $\mathcal{D}_{\mathrm{L}} = \{\mathrm{X}^I_{l,i}, \mathrm{X}^T_{l,i}, \mathrm{Y}_{l,i}\}^N_{i=1}$ for next training phase. Those remaining unlabeled data will form a new unlabeled set $\mathcal{D}_{\mathrm{R}} = \{\mathrm{X}^I_{u,o}\}^O_{o=1}$.

To obtain a pre-trained teacher and student model, we initialize the labeled set $\mathcal{D}_{\mathrm{L}}$ by considering the diversity of data in $\mathcal{D}_{\mathrm{U}}$. Specifically, we perform K-Means clustering using the color, texture, and frequency domain features of data in $\mathcal{D}_{\mathrm{U}}$. Then we select the data near the clustering centers to construct the initial labeled set $\mathcal{D}_{\mathrm{L}}$.

The ADA component adopts an adversarial manner for training. Specifically, given the labeled set $\mathcal{D}_{\mathrm{L}}$, we use the pre-trained augmenter $\{\mathcal{A}^{\theta_c}_{color}, \mathcal{A}^{\theta_g}_{geo}\}$ to obtain the augmented input $\mathrm{X}^{I,\mathrm{Aug}}_{l,i}$ by Eq. (1). Then we train the teacher and student models with Binary Cross-Entropy (BCE) loss $\mathcal{L}_{\mathrm{BCE}}$. Following [Suzuki, 2022], the training objective of augmenter loss $\mathcal{L}_{\mathrm{Aug}}$ is to minimize the loss of the teacher model and maximize the loss of the student model,

$$\begin{aligned}\mathcal{L}_{\mathrm{Aug}} =& \mathcal{L}_{\mathrm{BCE}}(\mathcal{F}(\mathrm{X}^{I,\mathrm{Aug}}_{l,i}; \theta_T), \mathrm{Y}_{l,i}) \\ &- \mathcal{L}_{\mathrm{BCE}}(\mathcal{F}(\mathrm{X}^{I,\mathrm{Aug}}_{l,i}; \theta_S), \mathrm{Y}_{l,i}).\end{aligned} \tag{4}$$

4.2 Text Fusion Module

The TFM module utilizes the precise referring text $\mathrm{X}^T_{l,i}$ in the labeled set $\mathcal{D}_{\mathrm{L}} = \{\mathrm{X}^I_{l,i}, \mathrm{X}^T_{l,i}, \mathrm{Y}_{l,i}\}^N_{i=1}$ to further extract the knowledge of camouflage object. Specifically, for input labeled data $\mathrm{X}^I_{l,i}$, with corresponding referring text annotations $\mathrm{X}^T_{l,i}$, the TFM utilizes CLIP-Text Encoder [Radford *et al.*, 2021] $\mathbf{CLIP}_{\mathcal{T}}(\cdot)$ to obtain text feature $\mathrm{F}^T_{l,i}$,

$$\mathrm{F}^T_{l,i} = \mathbf{CLIP}_{\mathcal{T}}(\mathrm{X}^T_{l,i}). \tag{5}$$

For input labeled data $\mathrm{X}^I_{l,i}$, assuming that the hierarchical features obtained by the Image Encoder are $\{\mathrm{F}^I_{l,i,k}\}, k \in [1, N_E]$,where N_E is the number of hierarchical feature levels, we use a clue attention mechanism to generate the clue query Q^{CLUE}, which highlights the camouflage-related region by the guidance of attention score $\mathrm{A}^{\mathrm{CAMO}}_{l,i}$.

$$\begin{gathered}Q_C = q^{\mathrm{CAMO}} W_Q,\ K_C = \mathrm{F}^I_{l,i,N_E} W_K,\ V_C = \mathrm{F}^I_{l,i,N_E} W_V \\ \mathrm{A}^{\mathrm{CAMO}}_{l,i} = \mathrm{Softmax}\left(\frac{Q_C K_C^\top}{\sqrt{d_h}}\right) \\ Q^{\mathrm{CLUE}} = \mathrm{A}^{\mathrm{CAMO}}_{l,i} V_C,\end{gathered} \tag{6}$$

where q^{CAMO} is a learnable query, and W_Q, W_K, W_V denote the linear projection weight, respectively. d_h represents the attention head dimension. In order to enable the clue attention mechanism to learn how to extract camouflage-related region features, the following loss function is leveraged,

$$\mathcal{L}^{\mathrm{Attn}}_{\mathrm{TFM}} = \mathcal{L}_{\mathrm{BCE}}(\mathrm{A}^{\mathrm{CAMO}}_{l,i}, \mathrm{down}(\mathrm{Y}_{l,i})), \tag{7}$$

where $\mathrm{down}(\cdot)$ denotes a downsampling function. Then we apply Q^{CLUE} to calculate the similarity scores $W^{\mathcal{C}}$ between the codebook $\mathcal{C}^T$ that stores class-related knowledge extracted from the labeled data, and then aggregates to refine coarse referring text, the similarity scores is calculated by

$$W^{\mathcal{C}}_k = \mathrm{COS}(Q^{\mathrm{CLUE}}, \mathcal{V}_k),\ \mathcal{V}_k \in \mathcal{C}^T, \tag{8}$$

where the $\mathrm{COS}(\cdot)$ is the consine similarity function. Next, we use these scores $W^{\mathcal{C}}$ to perform a weighted aggregation of the vectors in the codebook, generating a clue vector,

$$\mathcal{V}^{\mathrm{CLUE}} = \sum_k^{|\mathcal{C}^T|} W^{\mathcal{C}}_k \mathcal{V}_k. \tag{9}$$

We use the class word $\mathrm{X}^{T_{cls}}_{l,i}$ extracted from the referring text $\mathrm{X}^T_{l,i}$ to supervise the clue vector $\mathcal{V}^{\mathrm{CLUE}}$,

$$\mathcal{L}^{\mathrm{CLUE}}_{\mathrm{TFM}} = \mathcal{L}_{\mathrm{MSE}}(\mathcal{V}^{\mathrm{CLUE}}, \mathbf{CLIP}(\mathrm{X}^{T_{cls}}_{l,i})). \tag{10}$$

We combine this clue vector $\mathcal{V}^{\mathrm{CLUE}}$ with the referring text features $\mathrm{F}^T_{l,i}$ to enhance the representation ability. Then, we

CHAMELEON (76)

Methods	1% (41)						5% (202)						10% (404)					
	$\mathcal{S}_m \uparrow$	$\mathcal{F}_\beta^\omega \uparrow$	$\mathcal{F}_\beta^m \uparrow$	$\mathcal{E}_\phi^m \uparrow$	$\mathcal{E}_\phi^x \uparrow$	$\mathcal{M} \downarrow$	$\mathcal{S}_m \uparrow$	$\mathcal{F}_\beta^\omega \uparrow$	$\mathcal{F}_\beta^m \uparrow$	$\mathcal{E}_\phi^m \uparrow$	$\mathcal{E}_\phi^x \uparrow$	$\mathcal{M} \downarrow$	$\mathcal{S}_m \uparrow$	$\mathcal{F}_\beta^\omega \uparrow$	$\mathcal{F}_\beta^m \uparrow$	$\mathcal{E}_\phi^m \uparrow$	$\mathcal{E}_\phi^x \uparrow$	$\mathcal{M} \downarrow$
Mean Teacher [Lai *et al.*, 2024]	.537	.199	.229	.418	.636	.204	.611	.309	.353	.524	.745	.137	.679	.450	.512	.650	.812	.102
CamoTeacher [Lai *et al.*, 2024]	.652	.472	.558	.714	.762	.093	.729	.587	.656	.785	.822	.070	.756	.617	.684	.813	.851	.065
SCOD-ND [Fu *et al.*, 2024]	-	-	-	-	-	-	-	-	-	-	-	-	.850	.773	-	**.928**	-	.036
SCOUT (Ours) †	.846	.773	.806	.885	.896	.039	.877	.821	.850	.929	.941	.028	.874	.815	.838	.916	.925	.027
SCOUT (Ours) ‡	**.847**	**.777**	**.810**	**.887**	**.898**	**.038**	**.880**	**.827**	**.852**	**.935**	**.944**	**.027**	**.876**	**.817**	**.837**	.922	**.932**	**.026**

CAMO (250)

Methods	1% (41)						5% (202)						10% (404)					
	$\mathcal{S}_m \uparrow$	$\mathcal{F}_\beta^\omega \uparrow$	$\mathcal{F}_\beta^m \uparrow$	$\mathcal{E}_\phi^m \uparrow$	$\mathcal{E}_\phi^x \uparrow$	$\mathcal{M} \downarrow$	$\mathcal{S}_m \uparrow$	$\mathcal{F}_\beta^\omega \uparrow$	$\mathcal{F}_\beta^m \uparrow$	$\mathcal{E}_\phi^m \uparrow$	$\mathcal{E}_\phi^x \uparrow$	$\mathcal{M} \downarrow$	$\mathcal{S}_m \uparrow$	$\mathcal{F}_\beta^\omega \uparrow$	$\mathcal{F}_\beta^m \uparrow$	$\mathcal{E}_\phi^m \uparrow$	$\mathcal{E}_\phi^x \uparrow$	$\mathcal{M} \downarrow$
Mean Teacher [Lai *et al.*, 2024]	.518	.207	.227	.399	.620	.226	.575	.286	.322	.482	.708	.184	.625	.397	.454	.578	.773	.150
CamoTeacher [Lai *et al.*, 2024]	.621	.456	.545	.669	.736	.136	.669	.523	.601	.711	.775	.122	.701	.560	.635	.742	.795	.112
SCOD-ND [Fu *et al.*, 2024]	-	-	-	-	-	-	-	-	-	-	-	-	.789	.732	-	.859	-	.077
SCOUT (Ours) †	**.798**	.732	**.782**	.845	.864	**.076**	.847	.802	.839	.897	**.909**	.055	.847	.812	.849	.901	.912	.052
SCOUT (Ours) ‡	.795	**.732**	.780	**.845**	**.864**	.077	**.848**	**.807**	**.841**	**.901**	.908	**.054**	**.859**	**.828**	**.857**	**.919**	**.925**	**.047**

COD10K (2026)

Methods	1% (41)						5% (202)						10% (404)					
	$\mathcal{S}_m \uparrow$	$\mathcal{F}_\beta^\omega \uparrow$	$\mathcal{F}_\beta^m \uparrow$	$\mathcal{E}_\phi^m \uparrow$	$\mathcal{E}_\phi^x \uparrow$	$\mathcal{M} \downarrow$	$\mathcal{S}_m \uparrow$	$\mathcal{F}_\beta^\omega \uparrow$	$\mathcal{F}_\beta^m \uparrow$	$\mathcal{E}_\phi^m \uparrow$	$\mathcal{E}_\phi^x \uparrow$	$\mathcal{M} \downarrow$	$\mathcal{S}_m \uparrow$	$\mathcal{F}_\beta^\omega \uparrow$	$\mathcal{F}_\beta^m \uparrow$	$\mathcal{E}_\phi^m \uparrow$	$\mathcal{E}_\phi^x \uparrow$	$\mathcal{M} \downarrow$
Mean Teacher [Lai *et al.*, 2024]	.546	.168	.226	.441	.633	.161	.621	.272	.343	.555	.732	.107	.683	.404	.482	.666	.799	.078
CamoTeacher [Lai *et al.*, 2024]	.699	.517	.582	.788	.797	.062	.745	.583	.644	.827	.840	.050	.759	.594	.652	.836	.854	.049
SCOD-ND [Fu *et al.*, 2024]	-	-	-	-	-	-	-	-	-	-	-	-	.819	.725	-	.891	-	.033
SCOUT (Ours) †	.833	.733	.770	.891	.899	.031	.855	.768	.801	.913	.924	.027	.858	.783	**.815**	.917	.928	.024
SCOUT (Ours) ‡	**.834**	**.736**	**.774**	**.892**	**.901**	**.031**	**.859**	**.776**	**.804**	**.919**	**.926**	**.026**	**.861**	**.785**	.811	**.925**	**.932**	**.024**

NC4K (4121)

Methods	1% (41)						5% (202)						10% (404)					
	$\mathcal{S}_m \uparrow$	$\mathcal{F}_\beta^\omega \uparrow$	$\mathcal{F}_\beta^m \uparrow$	$\mathcal{E}_\phi^m \uparrow$	$\mathcal{E}_\phi^x \uparrow$	$\mathcal{M} \downarrow$	$\mathcal{S}_m \uparrow$	$\mathcal{F}_\beta^\omega \uparrow$	$\mathcal{F}_\beta^m \uparrow$	$\mathcal{E}_\phi^m \uparrow$	$\mathcal{E}_\phi^x \uparrow$	$\mathcal{M} \downarrow$	$\mathcal{S}_m \uparrow$	$\mathcal{F}_\beta^\omega \uparrow$	$\mathcal{F}_\beta^m \uparrow$	$\mathcal{E}_\phi^m \uparrow$	$\mathcal{E}_\phi^x \uparrow$	$\mathcal{M} \downarrow$
Mean Teacher [Lai *et al.*, 2024]	.541	.213	.258	.424	.637	.193	.634	.355	.420	.556	.767	.140	.700	.492	.565	.670	.827	.109
CamoTeacher [Lai *et al.*, 2024]	.718	.599	.675	.779	.814	.090	.777	.677	.739	.834	.859	.071	.791	.687	.746	.842	.868	.068
SCOD-ND [Fu *et al.*, 2024]	-	-	-	-	-	-	-	-	-	-	-	-	.838	.787	-	.903	-	.046
SCOUT (Ours) †	.849	.791	.829	.893	.904	.045	.874	.826	.856	.919	.930	.036	.873	.833	.864	.919	.930	.035
SCOUT (Ours) ‡	**.850**	**.793**	**.832**	**.894**	**.905**	**.045**	**.877**	**.832**	**.858**	**.924**	**.931**	**.035**	**.879**	**.839**	**.864**	**.928**	**.935**	**.033**

Table 1: **Quantitative comparison with existing methods on four COD benchmark testing sets including CHAMELEON, CAMO, COD10K and NC4K.** We provide experimental results under two test settings: † denotes that all data use fixed referring text during test time(*i.e.* "*camouflaged objects; concealed objects; hidden objects;*"), and ‡ denotes that all data used precise referring text during test time. **Bold** indicates the best result in group settings, and underline indicates the second-best result.

use a hierarchical feature cross-attention mechanism for text-visual information fusion,

$$\begin{aligned} \mathrm{F}_{l,i}^{T'} &= \mathrm{Concat}(\mathrm{F}_{l,i}^{T}, \mathcal{V}^{\mathrm{CLUE}}) \\ \mathrm{F}_{l,i,k}^{I'} &= \mathrm{CrossAttn}(\mathrm{F}_{l,i,k}^{I}, \mathrm{F}_{l,i}^{T'}), \end{aligned} \tag{11}$$

where $\mathrm{Concat}(\cdot)$ denotes the concatenation operation, and $\mathrm{CrossAttn}(\cdot)$ denotes the cross-attention function. During training and inference of unlabeled data in $\mathcal{D}_{\mathrm{R}}$, only the attention scores $\mathrm{A}_{l,i}^{\mathrm{CAMO}}$ are supervised by the pseudo-labels from the teacher model. The total loss function for the TFM module can be formulated as,

$$\mathcal{L}_{\mathrm{TFM}} = \mathcal{L}_{\mathrm{TFM}}^{\mathrm{Attn}} + \lambda_t \mathcal{V}^{\mathrm{CLUE}}, \tag{12}$$

where $\lambda_t = 1$ for labeled data in $\mathcal{D}_{\mathrm{L}}$ and $\lambda_t = 0$ for unlabeled data in $\mathcal{D}_{\mathrm{R}}$. We use the class words $\mathrm{X}_{l,i}^{T_{cls}}$ extracted from the referring text annotation $\mathrm{X}_{l,i}^{T}$ of the labeled data in $\mathcal{D}_{\mathrm{L}}$ to initialize the codebook $\mathcal{C}^T$.

4.3 Total Loss

The total loss function can be represented as,

$$\mathcal{L}_{tot} = \mathcal{L}_{\mathrm{s}} + \lambda_{\mathrm{u}} \mathcal{L}_{\mathrm{u}}, \tag{13}$$

where $\mathcal{L}_{\mathrm{s}}$ denotes the supervised loss, $\mathcal{L}_{\mathrm{u}}$ denotes the unsupervised loss, and λ_{u} is the weight of unsupervised loss to balance the loss terms. The labeled loss function consists of three parts,

$$\mathcal{L}_{\mathrm{s}} = \mathcal{L}_{\mathrm{Seg}}^{\mathrm{s}} + \mathcal{L}_{\mathrm{Aug}} + \mathcal{L}_{\mathrm{TFM}}, \tag{14}$$

where $\mathcal{L}_{\mathrm{Seg}}^{\mathrm{s}}$ denotes the supervised segmentation loss between the student predictions and groud-truth. The unlabeled loss function consists of two parts,

$$\mathcal{L}_{\mathrm{u}} = \mathcal{L}_{\mathrm{Seg}}^{\mathrm{u}} + \mathcal{L}_{\mathrm{TFM}}, \tag{15}$$

where $\mathcal{L}_{\mathrm{Seg}}^{\mathrm{u}}$ denotes the semi-supervised loss between the student predictions and pseudo labels (produced by the teacher model). Following [Zheng *et al.*, 2024], we use a combination of Binary Cross Entropy loss $\mathcal{L}_{\mathrm{BCE}}$, Intersection Over Union loss $\mathcal{L}_{\mathrm{IOU}}$, and Structure Similarity Index Measure loss $\mathcal{L}_{\mathrm{SSIM}}$ to build the supervised loss $\mathcal{L}_{\mathrm{Seg}}^{\mathrm{s}}$ and semi-supervised loss $\mathcal{L}_{\mathrm{Seg}}^{\mathrm{u}}$. The complete definition of these loss functions can be found in the appendix.

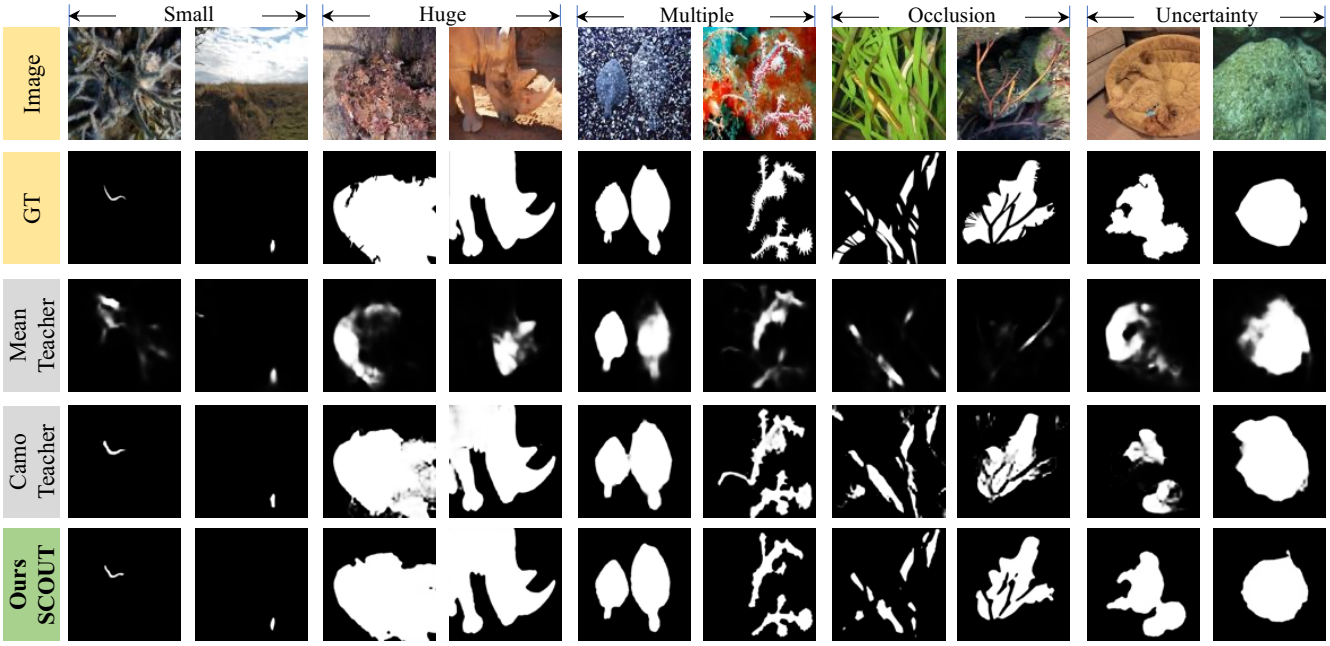

Figure 3: **Visual comparison of our method with other existing methods in challenging scenarios.** Our method has clearer and more precise segmentation boundaries and correctly recognizes depth-camoflaged objects.

5 Experiments

5.1 Experiment Settings

Training Set. To compare with the existing works, following [Luo *et al.*, 2023; Fan *et al.*, 2020], we use 1000 images from the CAMO trainset and 3040 images from the COD10K trainset as the training set for our experiments. During the training process, we follow the data partition ratios from previous semi-supervised COD results [Lai *et al.*, 2024], training the model with 1%, 5%, and 10% of labeled data. However, diverging from traditional semi-supervised segmentation approaches, we do not employ a random data sampling strategy. Instead, we utilize the proposed ADSA module to actively select the valuable data with annotating labels (*i.e.* segmentation mask and referring text), while the remaining portion is treated as unlabeled data. For all unlabeled data, the referring text is fixed to a single sentence "*Camouflaged objects; hidden objects; concealed objects*".

Testing Sets. We test the model's performance on four mainstream COD benchmark testing sets, CHAMELEON with 76 test images, CAMO with 250 test images, COD10K with 2026 test images, and NC4K with 4121 test images. To comprehensively evaluate the model, we test its performance under two different settings: using fixed referring text (*i.e.* "*Camouflaged objects; hidden objects; concealed objects*") and using precise image-level referring text during test.

Evaluation Protocol. For a fair and comprehensive evaluation, we employ the S-measure ($\mathcal{S}_m$) [Fan *et al.*, 2017], mean and weighted F-measure ($\mathcal{F}_\beta^m$, $\mathcal{F}_\beta^\omega$) [Margolin *et al.*, 2014], max and mean E-measure ($\mathcal{E}_\epsilon^x$, $\mathcal{E}_\epsilon^m$) [Fan *et al.*, 2018], mean absolute error ($\mathcal{M}$) [Perazzi *et al.*, 2012].

Implementation Details. All images are resized to 640×640 for training and testing. We employ the ImageNet pretrained Swin-base [Liu *et al.*, 2021] as our image encoder, use recently developed BiRefBlock [Zheng *et al.*, 2024] from High-Resolution Dichotomous Image Segmentation (HR-DIS) fields to build the decoder, and utilize CLIP-ViT-Large as our text encoder. The parameters of the CLIP text encoder are frozen during the training process, while all others are trainable. The batch size is set to 6 for each GPU during training, Adam is used as the optimizer, and the learning rate is initialized to 1e-4 and use multi-step decay strategy with 30 training epochs. All experiments are implemented with PyTorch 2.1 and a machine with Intel(R) Xeon(R) Silver 4214R CPU @ 2.40GHz, 256GiB RAM, and 8 NVIDIA Titan A800-80G GPUs.

5.2 Main Results

Qualitative Analysis. We show visualizations of a series of camouflaged object segmentation masks predicted by our method and related methods in some challenging scenarios. As shown in Figure 3, we notice that our method achieves higher segmentation accuracy than existing semi-supervised COD approaches, with clearer edges and a more detailed representation of the camouflaged object.

Quantitative Analysis. We compare the proposed SCOUT's performance with existing semi-supervised models on four COD test datasets. Since these methods are not open-sourced and are difficult to reproduce, we only report the experimental results published in their original papers. As shown in Table 1, SCOUT has already surpassed all previous methods across all metrics on all datasets when tested with fixed text (*e.g.* "*Camouflaged objects; hidden objects; concealed objects; CLUE_Token*"). Thanks to the assistance of the

Settings			COD10K (2026)					
ADAS-ADA	ADAS-ADS	TFM	$\mathcal{S}_m \uparrow$	$\mathcal{F}_\beta^\omega \uparrow$	$\mathcal{F}_\beta^m \uparrow$	$\mathcal{E}_\phi^m \uparrow$	$\mathcal{E}_\phi^x \uparrow$	$\mathcal{M} \downarrow$
			.805	.685	.729	.856	.871	.038
✓			.824	.730	.765	.878	.900	.034
		✓	.822	.720	.772	.869	.914	.033
✓		✓	.830	.745	.795	.885	.913	.029
✓	✓		.849	.763	.797	.905	.913	.028
✓	✓	✓	**.855**	**.768**	**.801**	**.913**	**.924**	**.027**

Table 2: **Ablation study to evaluate the proposed modules.** We retrain our model with different settings under 5% labeled data settings, and evaluation on the COD10K-Test sets.

Settings				COD10K (2026)					
Rand-Color	Rand-Geo	Ada-Color	Ada-Geo	$\mathcal{S}_m \uparrow$	$\mathcal{F}_\beta^\omega \uparrow$	$\mathcal{F}_\beta^m \uparrow$	$\mathcal{E}_\phi^m \uparrow$	$\mathcal{E}_\phi^x \uparrow$	$\mathcal{M} \downarrow$
				.822	.720	.772	.869	.914	.033
✓				.833	.744	.791	.886	.915	.032
	✓			.837	.737	.766	.899	.907	.031
✓	✓			.846	.757	.796	.901	.918	.030
		✓		.852	.768	**.806**	.908	**.927**	.027
			✓	.841	.750	.787	.901	.913	.029
		✓	✓	**.855**	**.768**	.801	**.913**	.924	**.027**

Table 3: **Ablation study on different data augmentations.** "Rand" denotes the random augmentation, while "Ada" denotes the learnable augmentations. "Color" represents the color augment operation, and "Geo" represents the geometry augment operation. We retrain each model on different combinations of the augmenters under 5% labeled data settings.

clue vector in TFM, SCOUT effectively improves fixed text, achieving results nearly identical to those inferences with precise texts. Compared to the previous best results, our method achieves an improvement of 52.0% in $\mathcal{M}$, 19.1% in $\mathcal{S}_m$, 16.3% in $\mathcal{E}_\phi^m$, and 29.1% in $\mathcal{F}_\beta^m$, effectively demonstrating the high performance of the proposed SCOUT.

5.3 Ablation Study

Module Ablations. To validate the effectiveness of the proposed module, we first perform ablations on the ADA and ADS components in the ADAS module and the TFM module. When the ADS component is not used, the labeled set is obtained through random sampling. When the TFM module is not used, the referring text does not participate in model training. We retrain the model with 5% labeled data setting and test it on COD10K-Test set. The results are shown in Table 2, which indicates a significant performance drop when neither module is used. Compared to the baseline, our method achieves an improvement of 28.9% in $\mathcal{M}$ and 6.21% in $\mathcal{S}_m$.

Effectiveness of Data Augment. We conduct comparative experiments between random augmentation and the trainable augmenters used in this paper. We retrain the model on different settings with 5% labeled data and test on the COD10K-Test set. As shown in Table 3, when using the learnable augmentation, we find that color augmentation has a significant effect. This is related to the complex colors typically found in camouflaged objects.

Effectiveness of Data Scoring. We conduct an ablation experiment on the data sampling strategy used in the ADS component in the ADAS module. Specifically, we resample 5% of the data by using Top-K easy, Top-K hard, and different score selection centers, and train the model. We then test it on the COD10K-Test set, and the results are shown in Table 4. When only easy and hard samples are used, the model struggles to achieve optimal performance. However, when selecting data around 0.5, the selected data not only suits the model's learning but also balances difficulty, resulting in the best performance.

Settings	COD10K (2026)					
	$\mathcal{S}_m \uparrow$	$\mathcal{F}_\beta^\omega \uparrow$	$\mathcal{F}_\beta^m \uparrow$	$\mathcal{E}_\phi^m \uparrow$	$\mathcal{E}_\phi^x \uparrow$	$\mathcal{M} \downarrow$
Top-K Easy (Close to 1)	.836	.715	.756	.879	.909	.035
Top-K Hard (Close to 0)	.845	.747	.770	.912	.920	.028
Center-0.75	.820	.725	.767	.880	.910	.030
Center-0.25	.834	.751	.797	.886	.915	.027
Center-0.5	**.855**	**.768**	**.801**	**.913**	**.924**	**.027**

Table 4: **Ablation study on different data selection strategies.** "Top-K Easy/Hard" denotes the selection of data by sampling in ascending/descending order. "Center-x" denotes the selection method where data is sampled symmetrically around a center, with uniform sampling both before and after the center x. We retrain the model under 5% labeled data settings and test on the COD10K-Test set.

Train		Test		COD10K (2026)					
Fixed.	Precise.	Fixed.	Precise.	$\mathcal{S}_m \uparrow$	$\mathcal{F}_\beta^\omega \uparrow$	$\mathcal{F}_\beta^m \uparrow$	$\mathcal{E}_\phi^m \uparrow$	$\mathcal{E}_\phi^x \uparrow$	$\mathcal{M} \downarrow$
✓		✓		.845	.765	.803	.901	.917	.027
✓			✓	.843	.761	**.806**	.899	.926	.028
	✓	✓		.855	.768	.801	.913	.924	.027
	✓		✓	**.859**	**.776**	.804	**.919**	**.926**	**.026**

Table 5: **Ablation study to evaluate the different settings on referring text.** We retrain the model under 5% labeled data setting with fixed and precise referring text for labeled data, then test on COD10K-Test dataset with fixed and precise referring text.

Effectiveness of Referring Text. We conduct further exploration into the role of referring text. We compare the performance of using precise and fixed referring text during both training and testing on the unlabeled set. The results are shown in Table 5. We find that precise referring text indeed helps the model learn camouflage-related knowledge more effectively, and the TFM is able to enhance fixed text, achieving performance comparable to that of precise text.

6 Conclusions

In this paper, we address the shortcomings of existing semi-supervised COD methods, which fail to adaptively select and utilize high-quality data, resulting in poor performance. We propose an innovative semi-supervised COD model SCOUT. Specifically, we propose the ADAS module to avoid meaningless data annotations by selecting valuable data through an adversarial augment and sampling strategy. Additionally, we propose the TFM to fully utilize the referring text by combining camouflage-related knowledge and text-visual interaction. Furthermore, we propose a RefTextCOD dataset, which contains a large number of image-level referring text annotations. Extensive experiments demonstrate the effectiveness of the proposed framework and modules.

Ethical Statement

There are no ethical issues.

Acknowledgements

This work was supported by the National Science Fund for Distinguished Young Scholars (No.62025603), the National Natural Science Foundation of China (No. U21B2037, No. U22B2051, No. U23A20383, No. 62176222, No. 62176223, No. 62176226, No. 62072386, No. 62072387, No. 62072389, No. 62002305 and No. 62272401), and the Natural Science Foundation of Fujian Province of China (No. 2021J06003, No. 2022J06001).

References

[Angluin, 1988] Dana Angluin. Queries and concept learning. *ML*, 2(4):319–342, apr 1988.

[Argamon-Engelson and Dagan, 1999] S. Argamon-Engelson and I. Dagan. Committee-based sample selection for probabilistic classifiers. *JAIR*, 11:335–360, November 1999.

[Bai *et al.*, 2023] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond, 2023.

[Cannaday *et al.*, 2023] Alan B. Cannaday, Curt H. Davis, and Trevor M. Bajkowski. Detection of camouflage-covered military objects using high-resolution multispectral satellite imagery. In *IGARSS 2023*, pages 5766–5769, Pasadena, CA, USA, 2023. IEEE.

[Chen *et al.*, 2021] Xiaokang Chen, Yuhui Yuan, Gang Zeng, and Jingdong Wang. Semi-supervised semantic segmentation with cross pseudo supervision. In *CVPR 2021*, pages 2613–2622, 2021.

[Chen *et al.*, 2022] Geng Chen, Si-Jie Liu, Yu-Jia Sun, Ge-Peng Ji, Ya-Feng Wu, and Tao Zhou. Camouflaged object detection via context-aware cross-level fusion. *IEEE TCSVT*, 32:1–1, 10 2022.

[Chen *et al.*, 2023] Hao Chen, Ran Tao, Yue Fan, Yidong Wang, Jindong Wang, Bernt Schiele, Xing Xie, Bhiksha Raj, and Marios Savvides. Softmatch: Addressing the quantity-quality trade-off in semi-supervised learning, 2023.

[Cheng *et al.*, 2023] Shupeng Cheng, Ge-Peng Ji, Pengda Qin, Deng-Ping Fan, Bowen Zhou, and Peng Xu. Large model based referring camouflaged object detection, 2023.

[Fan *et al.*, 2017] Deng-Ping Fan, Ming-Ming Cheng, Yun Liu, Tao Li, and Ali Borji. Structure-measure: A new way to evaluate foreground maps, 2017.

[Fan *et al.*, 2018] Deng-Ping Fan, Cheng Gong, Yang Cao, Bo Ren, Ming-Ming Cheng, and Ali Borji. Enhanced-alignment measure for binary foreground map evaluation, 2018.

[Fan *et al.*, 2020] Deng-Ping Fan, Ge-Peng Ji, Guolei Sun, Ming-Ming Cheng, Jianbing Shen, and Ling Shao. Camouflaged object detection. In *CVPR 2020*, pages 2774–2784, 2020.

[Fan *et al.*, 2022] Deng-Ping Fan, Ge-Peng Ji, Ming-Ming Cheng, and Ling Shao. Concealed object detection. *IEEE TPAMI*, 44(10):6024–6042, 2022.

[Fu *et al.*, 2024] Yuanbin Fu, Jie Ying, Houlei Lv, and Xiaojie Guo. Semi-supervised camouflaged object detection from noisy data. In *ACMM MM 2024*, page 4766–4775, New York, NY, USA, 2024. Association for Computing Machinery.

[Grandvalet and Bengio, 2004] Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. In *NIPS 2004*, page 529–536, 2004.

[Hu *et al.*, 2025] Jian Hu, Jiayi Lin, Shaogang Gong, and Weitong Cai. Relax image-specific prompt requirement in sam: A single generic prompt for segmenting camouflaged objects. In *AAAI 2025*, volume 38, pages 12511–12518, 2025.

[Ji *et al.*, 2023] Ge-Peng Ji, Deng-Ping Fan, Yu-Cheng Chou, Dengxin Dai, Alexander Liniger, and Luc Van Gool. Deep gradient learning for efficient camouflaged object detection. *MIR*, 20:92–108, 2023.

[Jia *et al.*, 2022] Qi Jia, Shuilian Yao, Yu Liu, Xin Fan, Risheng Liu, and Zhongxuan Luo. Segment, magnify and reiterate: Detecting camouflaged objects the hard way. In *CVPR 2022*, pages 4703–4712, 2022.

[King *et al.*, 2004] Ross D. King, Ken E. Whelan, Ffion Mair Jones, Philip G. K. Reiser, Christopher H. Bryant, Stephen H. Muggleton, Douglas B. Kell, and Stephen G. Oliver. Functional genomic hypothesis generation and experimentation by a robot scientist. *Nature*, 427:247–252, 2004.

[Lai *et al.*, 2024] Xunfa Lai, Zhiyu Yang, Jie Hu, Shengchuan Zhang, Liujuan Cao, Guannan Jiang, Zhiyu Wang, Songgan Zhang, and Rongrong Ji. Camoteacher: Dual-rotation consistency learning for semi-supervised camouflaged object detection. In *ECCV 2024*, 2024.

[Le *et al.*, 2019] Trung-Nghia Le, Tam V. Nguyen, Zhongliang Nie, Minh-Triet Tran, and Akihiro Sugimoto. Anabranch network for camouflaged object segmentation. *CVIU*, 184:45–56, July 2019.

[Li *et al.*, 2024] Bing Li, Rongqian Zhou, Lu Yang, Qiwen Wang, and Huang Chen. Mildetr: Detection transformer for military camouflaged target detection. *IEEE Access*, 12:26163–26174, 2024.

[Lin *et al.*, 2023] Jiaying Lin, Xin Tan, Ke Xu, Lizhuang Ma, and Rynson W. H. Lau. Frequency-aware camouflaged object detection. *TOMM*, 19(2), March 2023.

[Liu *et al.*, 2021] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV 2021*, pages 9992–10002, 2021.

[Luo *et al.*, 2023] Naisong Luo, Yuwen Pan, Rui Sun, Tianzhu Zhang, Zhiwei Xiong, and Feng Wu. Camouflaged instance segmentation via explicit de-camouflaging. In *CVPR 2023*, pages 17918–17927, 2023.

[Margolin *et al.*, 2014] Ran Margolin, Lihi Zelnik-Manor, and Ayellet Tal. How to evaluate foreground maps. In *CVPR 2014*, pages 248–255, 2014.

[Mei *et al.*, 2021] Haiyang Mei, Ge-Peng Ji, Ziqi Wei, Xin Yang, Xiaopeng Wei, and Deng-Ping Fan. Camouflaged object segmentation with distraction mining. In *CVPR 2021*, 2021.

[Mi *et al.*, 2022] Peng Mi, Jianghang Lin, Yiyi Zhou, Yunhang Shen, Gen Luo, Xiaoshuai Sun, Liujuan Cao, Rongrong Fu, Qiang Xu, and Rongrong Ji. Active teacher for semi-supervised object detection. In *CVPR 2022*, pages 14462–14471, 2022.

[OpenAI, 2024] OpenAI. Gpt-4 technical report, 2024.

[Pang *et al.*, 2022] Youwei Pang, Xiaoqi Zhao, Tian-Zhu Xiang, Lihe Zhang, and Huchuan Lu. Zoom in and out: A mixed-scale triplet network for camouflaged object detection. In *CVPR 2022*, 2022.

[Pang *et al.*, 2023] Youwei Pang, Xiaoqi Zhao, Tian-Zhu Xiang, Lihe Zhang, and Huchuan Lu. Zoomnext: A unified collaborative pyramid network for camouflaged object detection, 2023.

[Perazzi *et al.*, 2012] Federico Perazzi, Philipp Krähenbühl, Yael Pritch, and Alexander Hornung. Saliency filters: Contrast based filtering for salient region detection. In *CVPR 2012*, pages 733–740, 2012.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

[Ren *et al.*, 2023] Jingjing Ren, Xiaowei Hu, Lei Zhu, Xuemiao Xu, Yangyang Xu, Weiming Wang, Zijun Deng, and Pheng-Ann Heng. Deep texture-aware features for camouflaged object detection. *IEEE TCSVT*, 33(3):1157–1167, 2023.

[Settles, 2009] Burr Settles. Active learning literature survey. In *None*, 2009.

[Sohn *et al.*, 2020] Kihyuk Sohn, Zizhao Zhang, Chun-Liang Li, Han Zhang, Chen-Yu Lee, and Tomas Pfister. A simple semi-supervised learning framework for object detection. In *arXiv:2005.04757*, 2020.

[Sun *et al.*, 2022] Yujia Sun, Shuo Wang, Chenglizhao Chen, and Tian-Zhu Xiang. Boundary-guided camouflaged object detection. In Lud De Raedt, editor, *IJCAI 2022*, pages 1335–1341, 7 2022. Main Track.

[Suzuki, 2022] Teppei Suzuki. Teachaugment: Data augmentation optimization using teacher knowledge. In *CVPR 2022*, pages 10894–10904, New Orleans, LA, USA, 2022. IEEE.

[Wang *et al.*, 2022] Yuchao Wang, Haochen Wang, Yujun Shen, Jingjing Fei, Wei Li, Guoqiang Jin, Liwei Wu, Rui Zhao, and Xinyi Le. Semi-supervised semantic segmentation using unreliable pseudo-labels. In *CVPR 2022*, pages 4238–4247, 2022.

[Wu *et al.*, 2019] Zhe Wu, Li Su, and Qingming Huang. Cascaded partial decoder for fast and accurate salient object detection, 2019.

[Yadav *et al.*, 2018] Deepti Yadav, Kailash Tiwari, Manoj Arora, and Jayanta Ghosh. Detection and identification of camouflaged targets using hyperspectral and lidar data. *Defence science journal*, 10 2018.

[Yan *et al.*, 2021] Jinnan Yan, Trung-Nghia Le, Khanh-Duy Nguyen, Minh-Triet Tran, Thanh-Toan Do, and Tam V. Nguyen. Mirrornet: Bio-inspired camouflaged object segmentation. *IEEE Access*, 9:43290–43300, 2021.

[Zhai *et al.*, 2021] Qiang Zhai, Xin Li, Fan Yang, Chenglizhao Chen, Hong Cheng, and Deng-Ping Fan. Mutual graph learning for camouflaged object detection. In *CVPR 2021*, 2021.

[Zhang and Wu, 2023] Yi Zhang and Chengyi Wu. Unsupervised camouflaged object segmentation as domain adaptation. In *ICCV 2023 Workshops*, pages 4334–4344, October 2023.

[Zhang *et al.*, 2022] Miao Zhang, Shuang Xu, Yongri Piao, Dongxiang Shi, Shusen Lin, and Huchuan Lu. Preynet: Preying on camouflaged objects. *ACM MM 2022*, 2022.

[Zhang *et al.*, 2023] Xuying Zhang, Bowen Yin, Zheng Lin, Qibin Hou, Deng-Ping Fan, and Ming-Ming Cheng. Referring camouflaged object detection, 2023.

[Zhang *et al.*, 2024] Jin Zhang, Ruiheng Zhang, Yanjiao Shi, Zhe Cao, Nian Liu, and Fahad Shahbaz Khan. Learning camouflaged object detection from noisy pseudo label. In *ECCV 2024*, pages 158–174, 2024.

[Zheng *et al.*, 2024] Peng Zheng, Dehong Gao, Deng-Ping Fan, Li Liu, Jorma Laaksonen, Wanli Ouyang, and Nicu Sebe. Bilateral reference for high-resolution dichotomous image segmentation. *CAAI 2024 Artificial Intelligence Research*, 2024.

[Zhong *et al.*, 2022] Yijie Zhong, Bo Li, Lv Tang, Senyun Kuang, Shuang Wu, and Shouhong Ding. Detecting camouflaged object in frequency domain. In *CVPR 2022*, pages 4494–4503, 2022.

[Zhu *et al.*, 2021] Jinchao Zhu, Xiaoyu Zhang, Shuo Zhang, and Junnan Liu. Inferring camouflaged objects by texture-aware interactive guidance network. In *AAAI 2021*, 2021.