TalNet: Text-based Semi-Supervised Referring Camouflaged Object Detection with Active Learning

Weiqi Yan 31520241154531[1], Jinhong Zhu 31520241154543[1], Ziyi Xu 36920241153268 [2],

[1]Information Class
[2]AI Class

Abstract

The difficulty of pixel-level annotation has significantly hindered the development of the existing Camouflaged Object Detection (COD) field, drawing increased attention to the need for reducing annotation costs among researchers in this domain. To save on annotation costs, previous works focus on the semi-supervised COD framework that leverages a small number of annotated data and a large volume of unlabeled data. We argue that there is still significant room for improvement in the efficiency of sample utilization, primarily focusing on the active selection of samples and the utilization of textual modal information. To this end, this paper introduces the active learning-based semi-supervised and text-based referring COD model, dubbed TalNet. It includes an active data selection and annotation (ADSA) module and a text fusion module (TFM). The ADSA module selects high-quality data for annotation through muti-feature clustering initialization and incremental sampling strategy. The TFM module leverages textual information to enhance the localization and segmentation quality of camouflaged objects. Extensive experiments show that our method surpasses previous semi-supervised methods in the COD field and achieves state-of-the-art performance. Especially, our method achieved an average improvement of **30.55%** in mean absolute error compared to previous methods when trained with only **1%** labeled data, further proving the effectiveness of our proposed method.

1 Introduction

Camouflaged object detection (COD) (Fan et al. 2020a), (Fan et al. 2022) aims at segmenting objects that are visually concealed in their surroundings, which has important applications in several fields (Fan et al. 2020b), (Fan et al. 2020c), (Tabernik et al. 2019), (Le et al. 2020), (Turkoglu and Hanbay 2019). In the field of biology, camouflage is defined as a strategy that animals use to adapt their body's physical appearance (*e.g.* texture or color) to match their surroundings for concealment (Singh, Dhawale, and Misra 2013). Such complex camouflaged strategies pose a huge challenge for COD tasks. Existing methods train models by analyzing and exploiting large amounts of data to gain the ability to locate and segment camouflaged objects (Mei et al. 2021),(Le et al. 2022), (Pei et al. 2022).

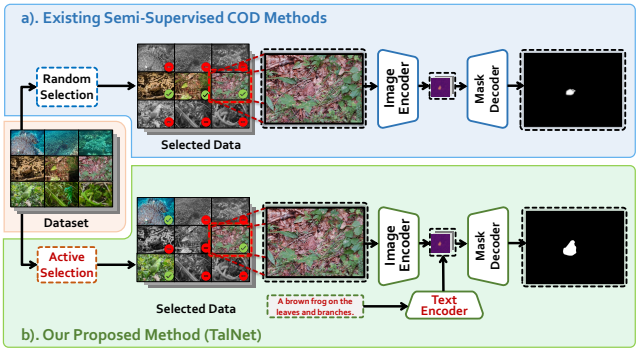

Figure 1: Visual comparison between traditional semi-supervised COD methods and our proposed TalNet. We select high-quality data using a pool-based active learning strategy and add image-level referring text to assist the model in localization and segmentation.

However, pixel-level camouflaged object annotations are difficult to obtain, which has plagued the advancement and application of existing COD methods. To this end, how to reduce annotation costs has become a research focus in COD. To mitigate this issue, semi-supervised COD methods (Lai et al. 2024) emerge as a promising approach by leveraging both labeled and unlabeled COD data. These methods effectively enhance model performance in scenarios with scarce annotations by employing data augmentation techniques specifically designed for the characteristics of camouflaged objects. However, in these methods, the labeled data is obtained through random sampling, which does not fully take into account the quality of the selected data. Additionally, relying solely on raw image information makes it difficult for the model to achieve good performance with extremely limited labeled data.

In order to select truly valuable samples for annotation and incorporate auxiliary referring text information, which is more readily annotated to aid in the detection of camouflaged objects, we introduce a semi-supervised referring COD model based on active learning, termed TalNet. It includes an active data selection and annotation (ADSA) mod-

ule and a text fusion module (TFM). The ADSA module selects high-quality data for annotation through multi-feature clustering initialization and a pool-based active learning sampling strategy so that the valuable data can be objectively selected for annotation and training. To leverage the semantic information and inherent knowledge embedded in referring text, the TFM module uses contrastive language-image pre-training (CLIP) to encode the referring text, and then a hierarchical shared multi-modal cross-attention is used to fuse the image features and referring text features.

Since none of the existing mainstream COD datasets have image-level camouflaged object referring text, this severely hinders the research in this paper. Therefore, we first use the vision language model (VLM) to generate annotations on images through visual guidance and designed prompts with contextual logic. To ensure high-quality annotations, we further adopt manual screening and refinement to check and correct all the referring text annotations. Finally, we have annotated a total of 9,487 images from four datasets (*e.g.* CHAMELEON (Wu, Su, and Huang 2019), CAMO (Yan et al. 2021) (Le et al. 2019), COD10K (Fan et al. 2020a), NC4K(Le et al. 2019)). These annotations can not only be utilized for the research in this paper but also provide data support for more exploration of the COD task.

To validate the effectiveness of our proposed method, we trained models at various split ratios of labeled data and compared their performance against existing semi-supervised COD methods across multiple test datasets. The experimental results conclusively demonstrate that our method significantly outperforms all current semi-supervised COD approaches, achieving state-of-the-art (SOTA) levels. This robustly substantiates the efficacy of our proposed method.

To our knowledge, our main contributions can be summarized as follows:

1. We proposed a semi-supervised referring COD model based on active learning called TalNet, which uses the ADSA module to adaptively select valuable data and uses TFM to integrate the semantic information produced by referring text to enhance the model performance on camouflaged object localization of segmentation.
2. This is the first time that active learning strategies have been introduced to the semi-supervised COD field to select better-quality data for annotation and training.
3. This is the first time that precise referring text has been used to assist in the COD task.
4. We annotated four existing mainstream camouflage object detection datasets with image-level camouflaged object referring text, using a combination of VLM annotation and manual refinement on **9,487** images. These annotations provides a data foundation for the research in this paper and other related tasks.
5. We conducted extensive experimental validation to prove the effectiveness of our proposed method. Especially, our method achieved an average improvement of 30.55% in mean absolute error compared to previous methods when trained with only 1% labeled data and tested on all four test sets. These results show that our method surpasses previous semi-supervised methods in the COD field and achieves SOTA performance.

2 Related Works

In this section, we will focus on several tasks related to our semi-supervised referring camouflaged object detection and introduce the dataset on which our annotation work is based.

2.1 Camouflaged Object Detection

Camouflaged Object Detection (COD) is a challenging task that focuses on segmenting objects that are deliberately designed to blend into their surroundings. It has a long history in the field. Early methods utilized classical image processing approaches and handcrafted features e.g. texture, color, boundary, and intensity features. However, recent advances have seen a shift towards deep learning (DL) approaches, which have significantly improved detection rates.

Existing methods have improved the performance degradation problem of the conventional salient object detection methods in the field of COD by employing multiple strategies, *e.g.* (Pang et al. 2022), (Pang et al. 2023), (Fan et al. 2020a), (Chou, Chen, and Shuai 2022), (Zhuge et al. 2022), (Chen et al. 2022) extracting multi-scale features from backbone and designing strategies for fusion, (Fan et al. 2022), (Jia et al. 2022), (Zhang et al. 2022), (Wang et al. 2022a), (Lin et al. 2017) further use multi-stage refine, some other methods introduce additional information, *e.g.* boundary guidance (Sun et al. 2022a), (Ji et al. 2023a), (Zhai et al. 2021), (Zhu et al. 2022), (Zhou et al. 2022), (Sun, Jiang, and Qi 2023), texture clues (Ji et al. 2023a), (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), (Wang et al. 2023), (Xiang et al. 2021). As stated in (Zhang et al. 2023), it is very difficult to force the network to learn how to extract these high-quality features from limited data, which further brings about the obfuscation problem of camouflaged object segmentation.

2.2 Semi-Supervised Learning

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), (Lee 2013), (Mi et al. 2022), (Oliver et al. 2018), (Berthelot et al. 2022), (Sohn et al. 2020a), (Tarvainen and Valpola 2017), (Wang et al. 2021), which effectively addresses the challenges of acquiring labeled data. Some previous works (Chen et al. 2021), (Sohn et al. 2020b), (Wang et al. 2022b), (Xu et al. 2021), (Yang 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. Recently, some methods (Lai et al. 2024) have applied SSL to

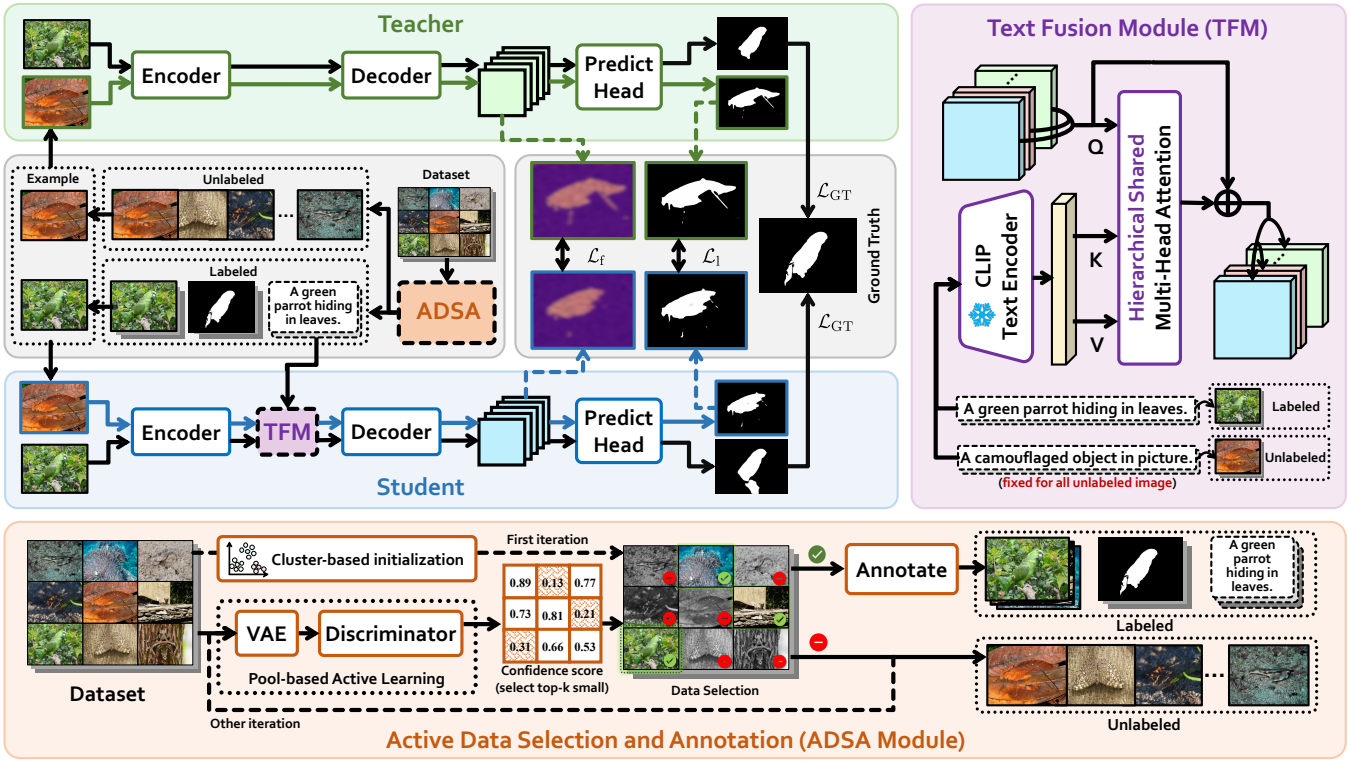

Figure 2: Main framework of the proposed TalNet. It includes an ADSA module and a TFM module. The ADSA module uses pool-based active learning to select high-quality data, and the TFM module employs hierarchical shared multi-modal cross-attention to fuse the image features and referring text features.

the field of COD, significantly improving the performance of the COD models under limited labeled data training scenarios.

2.3 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. AL 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 AL from application scenarios (Settles 2009). In pool-based active learning, (Sinha, Ebrahimi, and Darrell 2019) utilized a variational autoencoder (VAE) and an adversarial network to learn a latent space that distinguishes between unlabeled and labeled data. This adversarial approach effectively learns a low-dimensional latent space in large-scale settings and provides a computationally efficient sampling method.

2.4 Referring Expression Segmentation

Referring Expression Segmentation (RES) aims at labeling the pixels of the given image that represent an object referred to by natural language description. This task is crucial for visual understanding and human-computer interaction because it merges the understanding of visual content with the parsing of linguistic instructions.

Existing methods are categorized into two types: fully supervised and weakly supervised methods. Among them, fully-supervised methods are solved by using joint-embedding (Nagaraja, Morariu, and Davis 2016), (Hu et al. 2016b), (Yu et al. 2016), (Luo and Shakhnarovich 2017), modular network (Hu et al. 2016a), (Liu et al. 2019a), graph-based methods (Wang et al. 2018), (Yang, Li, and Yu 2019), (Liu et al. 2019b), (Yang, Li, and Yu 2020), and pretrained language model. Recently, with the rise of the contrastive language-image pretraining (CLIP) model (Radford et al. 2021), some methods have employed the CLIP model to obtain image features and text embeddings. We use a CLIP text encoder to encode the caption of the camouflaged object as the referring information to assist in localization and segmentation.

2.5 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. 2024) have ex-

tended this concept by utilizing MLLMs and designing a series of prompts to assist the COD task. There is no doubt that the usage of MLLMs does bring performance improvement, but they are not designed for COD task (Ji et al. 2023b), if MLLMs thenselves fail to localize the camouflaged object in the image, then the reference information provided introduces noise, which will bring about a performance degradation. We mainly focus on how to use the precise referring text to assist the semi-supervised COD model in more accurately localizing and segmenting camouflaged objects.

3 Proposed Method

3.1 Formulation

Camouflaged Object Detection (COD) is a challenging task that aims to segment objects that are visually hidden in their surroundings. Semi-supervised COD further increases the challenge and difficulties and challenges of this task. It leverage a limited amount of annotated data to train a detector that is still capable of identifying objects that seamlessly blend into their surroundings. Our method aims to select high-quality data adaptively and integrate referring text to assist in localizing and segmenting the camouflaged object. We assume that none of the images in the training set have any annotation(*i.e.* both segmentation mask and referring text), and we will expand it by incrementally selecting data. In particular, all images in original training set S_{train}^{all} are represented as $\mathrm{I}_i^{in} \in \mathbb{R}^{3\times H\times W}, i \in [1, |S_{train}^{all}|]$, where H, W denote the height and width of the image, $|\cdot|$ denote the size of the set. If the *i-th* image is selected for actual training, its corresponding referring text and semantic segmentation masks are denoted as $\mathrm{T}_i^{ref}, \mathrm{M}_i^{seg} \in \mathbb{R}^{1\times H\times W}$. All annotated referring text have several words less than or equal to 64.

3.2 Overall Framework

As shown in Fig. 2, we adopt the teacher-student paradigm as a preliminary to construct the semi-supervised COD framework. The framework consists of two similar models and an Active Data Selection and Annotation (ADSA) module. The ADSA module uses multi-feature clustering strategies to initialize labeled data and then uses active-learning-based methods to select high-quality data incrementally. We expect the teacher model to quickly learn the fundamental knowledge about camouflaged objects and initially possess the ability to locate these objects. Therefore, in the teacher model, we employ a simple encoder-decoder structure that learns only from the selected labeled data and generates pseudo-label predictions for the unlabeled data. For the student model, we expect it to fully utilize the knowledge from the teacher model and, combined with the semantic information in referring text to achieve more accurate localization and segmentation. To this end, we incorporate a designed Text Fusion Module (TFM) to fuse the semantic information with hierarchical image features and use a pseudo-label distillation loss to transfer the knowledge from the teacher model to the student model. The overall loss function $\mathcal{L}_{\mathrm{tot}}$ can be termed as:

$$\mathcal{L}_{\mathrm{tot}} = \mathcal{L}_{\mathrm{s}} + \mathcal{L}_{\mathrm{us}} + \mathcal{L}_{\mathrm{ADSA}} \tag{1}$$

where $\mathcal{L}_s$ and $\mathcal{L}_{\mathrm{us}}$ denote supervised loss and unsupervised loss respectively, $\mathcal{L}_{\mathrm{ADSA}}$ denotes the loss of ADSA module. Following the previous work, we use the combination of binary cross-entropy loss $\mathcal{L}_{\mathrm{BCE}}$, intersection over union loss $\mathcal{L}_{\mathrm{IoU}}$, and structure similarity index measure $\mathcal{L}_{\mathrm{SSIM}}$ in supervised loss:

$$\begin{aligned} \mathcal{L}_{\mathrm{s}} &= \mathcal{L}_{\mathrm{GT}}^{\mathrm{tea}} + \mathcal{L}_{\mathrm{GT}}^{\mathrm{stu}} \\ \mathcal{L}_{\mathrm{GT}} &= \lambda_1\mathcal{L}_{\mathrm{BCE}} + \lambda_2\mathcal{L}_{\mathrm{IoU}} + \lambda_3\mathcal{L}_{\mathrm{SSIM}} \end{aligned} \tag{2}$$

where λ_1, λ_2, and λ_3 are respectively set to 30, 0.5, and 10 following (Zheng et al. 2024) to keep all the losses on the same quantitative level. For unsupervised loss, we used the same loss as $\mathcal{L}_{\mathrm{GT}}$ for hard pseudo-label distillation. Additionally, we perform distillation on the intermediate features of the decoder. Sepcifically, we generate segmentation mask $\hat{y}_{i,j}^{\mathrm{tea}}, \hat{y}_{i,j}^{\mathrm{stu}}$ using the j-th intermediate features of the teacher and student model for the i-th input image. The mask generated by the teacher model $\hat{y}_{i,j}^{\mathrm{tea}}$ is used as a pseudo label to supervise the mask produced by the student model. We use only the $\mathcal{L}_{\mathrm{BCE}}$ for distilling these intermediate features. The final unsupervised loss can be termed as:

$$\begin{aligned} \mathcal{L}_{\mathrm{us}} &= \mathcal{L}_{\mathrm{f}} + \mathcal{L}_{\mathrm{l}} \\ \mathcal{L}_{\mathrm{f}} &= \lambda_1\mathcal{L}_{\mathrm{BCE}}^{\mathrm{f}} \\ \mathcal{L}_{\mathrm{l}} &= \lambda_1\mathcal{L}_{\mathrm{BCE}}^{\mathrm{l}} + \lambda_1\mathcal{L}_{\mathrm{IoU}}^{\mathrm{l}} + \lambda_1\mathcal{L}_{\mathrm{SSIM}}^{\mathrm{l}} \end{aligned} \tag{3}$$

where $\mathcal{L}_{\mathrm{f}}, \mathcal{L}_{\mathrm{l}}$ denotes the distillation loss of student model. The complete definition of loss function $\mathcal{L}_{\mathrm{BCE}}, \mathcal{L}_{\mathrm{IoU}}, \mathcal{L}_{\mathrm{SSIM}}$ can be found at supplementary materials.

3.3 Referring Text Annotations

Since none of the existing datasets have image-level referring text, we annotate four existing mainstream datasets first. Based on a series of pre-existing datasets, we can focus more on the task of annotating the image without having to collect camouflaged images from scratch.

To achieve a fair, easy, and effective comparison with existing methods, we expect to be able to construct text-based referring camouflaged object detection experiments in as similar settings as possible. Referring to (Chen et al. 2022), (Fan et al. 2022) , we decided to use the mainstream COD datasets: CHAMELEON (Wu, Su, and Huang 2019), CAMO (Yan et al. 2021) (Le et al. 2019), COD10K (Fan et al. 2020a), NC4K(Le et al. 2019)) as the base image data. Similarly, the aforementioned dataset does not contain image-level referring text that cannot be applied to our proposed methods. Therefore, we first need to annotate these datasets.

Annotating camouflaged objects with captions is an extremely time-consuming and labor-intensive task, and manual annotating tends to lead to inconsistent labeling quality. To ensure that the camouflaged object captions contain meaningful information for COD task *i.e.* texture, color, and shape, we used a vision language model (VLM) to generate a summary first, then we manually filtered and refined it to obtain an annotation that better meets our expectations.

For the selection of VLM models, we used QwenVL (Bai et al. 2023) and GPT4-Vision (OpenAI 2024), both of which

support multiple vision inputs and serve as our annotations models. Since the visual language model is not designed for COD tasks, to avoid the annotation quality being affected by the VLM's camouflaged object localization and comprehension ability, we directly use the ground-truth mask corresponding to the input image to segment the foreground object and guide the model for annotation step by step. 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 camouflaged object captions.

Finally, we employed manual screening and conducted a thorough review and revision of all annotations to ensure their accuracy and high quality. As a result, we annotated a total of **9,487** images from four datasets. The prompt used in the above process, the whole pipeline, and all referring text annotations will be provided in the supplementary materials.

3.4 Active Data Selection and Annotation Module

To solve the issue that previous semi-supervised COD models couldn't actively select high-quality samples, we introduce an ADSA module for active sample selection and annotation simulation. In the initialization phase, we aim to select samples that are as diverse and representative as possible. Considering that camouflaged objects usually exhibit unique characteristics in texture(Ji et al. 2023a), (Zheng et al. 2024), (Ren et al. 2023), (Zhu et al. 2021), edges(Sun, Jiang, and Qi 2023), (Sun et al. 2022b), (Zhai et al. 2021), (Zhu et al. 2022), and the frequency domain(Luo et al. 2023), we extract the gray-level co-occurrence matrix(GLCM) from the original images as the texture feature, use the Sobel operator to extract edge features, and apply Fourier transform to extract frequency domain features. After reducing the dimensionality with the PCA algorithm, we perform K-Means clustering and select the image samples closest to the cluster centers as the initial labeled data.

For the incremental sampling strategy, we adopted the pool-based active learning method (Sinha, Ebrahimi, and Darrell 2019), employing a Variational Autoencoder(VAE) and an adversarial network to learn the latent space, aiming to distinguish between labeled and unlabeled data. The VAE model attempts to deceive the adversarial network into predicting that all data originate from the labeled pool; while the adversarial network learns to differentiate dissimilarities within the latent space. We employ the aforementioned strategy for incremental sampling, expanding from 1% of labeled data to 5% and 10%. The loss function of the ADSA module can be termed as:

$$\begin{aligned} \mathcal{L}_{\mathrm{ADSA}} &= \mathcal{L}_{\mathrm{VAE}} + \mathcal{L}_{\mathrm{D}} \\ \mathcal{L}_{\mathrm{VAE}} &= \mathcal{L}_{\mathrm{VAE}}^{\mathrm{trd}} + \mathcal{L}_{\mathrm{VAE}}^{\mathrm{adv}} \end{aligned} \tag{4}$$

where $\mathcal{L}_{\mathrm{VAE}}$ denotes the total VAE loss, which includes the training loss $\mathcal{L}_{\mathrm{VAE}}^{\mathrm{trd}}$ and adversarial loss $\mathcal{L}_{\mathrm{VAE}}^{\mathrm{adv}}$. For a detailed definition of the above losses, please refer to the supplementary materials.

3.5 Text Fusion Module

To leverage the semantic intelligence and inherent knowledge of MLLMs to enhance the localization and segmentation quality of camouflaged objects, we propose a TFM module to align and aggregate image features with referring text features and produce semantically rich features that can be used in subsequent decoders. We encode the input referring text using CLIP text encoder to obtain textual features $\mathrm{F}_i^{\mathrm{Text}}$:

$$\mathrm{F}_i^{\mathrm{Text}} = \mathbf{CLIP}_{\mathrm{T}}(\mathrm{T}_i^{ref}) \tag{5}$$

where $\mathrm{CLIP}_{\mathrm{T}}$ denotes the CLIP text encoder. Subsequently, we use two linear layers to map these text features into key and value vectors for use in attention calculations. The effectiveness of hierarchical features in COD has been shown in previous works (Huang et al. 2023), (Jia et al. 2022), (Sun, Jiang, and Qi 2023), (Zheng et al. 2024), (Zhu et al. 2022), (Zhu et al. 2021), (Ren et al. 2023). We aim to facilitate information exchange between hierarchical features while engaging in cross-modal feature interactions. To achieve this, we employ a hierarchical shared multi-head cross-attention mechanism. Specifically, we first use 1×1 convolutions to align the channels of hierarchical features. Then, we use a linear layer to obtain the query vector. Finally, the query, key, and value vectors pass through a multi-head attention layer, resulting in semantically rich features $\{\mathrm{F}_{i,j}^{\mathrm{attn}}\}_{j=1}^{\mathcal{M}}$:

$$\mathrm{F}_{i,j}^{\mathrm{attn}} = \mathbf{MultiHeadAttn}(Q, K, V) \tag{6}$$

where Q, K, V denotes the query, key, and value vectors as previously described, and $\mathcal{M}$ denote the number of hierarchical features. The semantically rich features will be utilized in the subsequent decoder to generate the segmentation masks.

4 Experiments

4.1 Experiment Settings

Training Set. To compare with the existing works, following (Luo et al. 2023), (Fan et al. 2020a), we used 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 followed the data partition ratios from previous semi-supervised camouflaged object detection results, training the model with 1%, 5%, and 10% of labeled data. However, diverging from traditional semi-supervised segmentation approaches, we didn't employ a random data sampling strategy. Instead, we utilized the proposed ADSA module to actively select the valuable data and simulate annotating labels(*i.e.* segmentation mask and referring text), while the remaining portion was treated as unlabeled data. For all unlabeled data, the referring text is fixed to a single sentence "A camouflaged object in the picture".

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 tested its performance under two different settings: using fixed referring text (*i.e.* "A camouflaged object in the picture") and using precise image-level referring text on all text images.

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)	0.537	0.199	0.229	0.418	0.636	0.204	0.611	0.309	0.353	0.524	0.745	0.137	0.679	0.450	0.512	0.650	0.812	0.102
CamoTeacher (Lai et al. 2024)	0.652	0.472	0.558	0.714	0.762	0.093	0.729	0.587	0.656	0.785	0.822	0.070	0.756	0.617	0.684	0.813	0.851	0.065
TalNet †	0.716	0.58	0.656	0.731	0.816	0.064	0.817	0.744	0.794	0.878	**0.899**	0.044	0.844	0.762	0.801	0.893	**0.914**	0.040
TalNet ‡	**0.772**	**0.673**	**0.733**	**0.834**	**0.885**	**0.053**	**0.827**	**0.759**	**0.802**	**0.886**	0.896	**0.041**	**0.849**	**0.783**	**0.814**	**0.903**	**0.914**	**0.037**

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)	0.518	0.207	0.227	0.399	0.620	0.226	0.575	0.286	0.322	0.482	0.708	0.184	0.625	0.397	0.454	0.578	0.773	0.150
CamoTeacher (Lai et al. 2024)	0.621	0.456	0.545	0.669	0.736	0.136	0.669	0.523	0.601	0.711	0.775	0.122	0.701	0.560	0.635	0.742	0.795	0.112
TalNet †	0.608	0.432	0.511	0.604	0.695	0.130	0.695	0.578	0.650	0.716	0.750	0.108	0.775	0.692	0.750	0.831	**0.855**	0.086
TalNet ‡	**0.684**	**0.557**	**0.633**	**0.720**	**0.784**	**0.113**	**0.723**	**0.622**	**0.688**	**0.752**	**0.771**	**0.098**	**0.782**	**0.718**	**0.768**	**0.841**	0.853	**0.080**

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)	0.546	0.168	0.226	0.441	0.633	0.161	0.621	0.272	0.343	0.555	0.732	0.107	0.683	0.404	0.482	0.666	0.799	0.078
CamoTeacher (Lai et al. 2024)	0.699	0.517	0.582	0.788	0.797	0.062	0.745	0.583	0.644	0.827	0.840	0.050	0.759	0.594	0.652	0.836	0.854	0.049
TalNet †	0.693	0.511	0.587	0.715	0.805	0.050	0.790	0.689	0.745	0.851	**0.885**	0.036	0.820	0.712	0.759	0.885	**0.901**	0.034
TalNet ‡	**0.764**	**0.623**	**0.676**	**0.835**	**0.861**	**0.042**	**0.801**	**0.706**	**0.757**	**0.860**	0.883	**0.033**	**0.825**	**0.737**	**0.777**	**0.888**	**0.901**	**0.030**

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)	0.541	0.213	0.258	0.424	0.637	0.193	0.634	0.355	0.420	0.556	0.767	0.140	0.700	0.492	0.565	0.670	0.827	0.109
CamoTeacher (Lai et al. 2024)	0.718	0.599	0.675	0.779	0.814	0.090	0.777	0.677	0.739	0.834	0.859	0.071	0.791	0.687	0.746	0.842	0.868	0.068
TalNet †	0.726	0.616	0.690	0.766	0.819	0.076	0.809	0.744	0.797	0.861	0.883	0.055	0.834	0.759	0.801	0.884	0.899	0.051
TalNet ‡	**0.786**	**0.696**	**0.746**	**0.848**	**0.866**	**0.063**	**0.822**	**0.763**	**0.809**	**0.875**	**0.890**	**0.050**	**0.840**	**0.784**	**0.819**	**0.901**	**0.901**	**0.046**

Table 1: Quantitative comparison with existing methods on four COD benchmark testing sets includes CHAMELEON, CAMO, COD10K and NC4K. We provide experimental results under two test settings: † indicates that all test images used fixed referring text(*i.e.* "A camouflaged object in the picture."), and ‡ indicates that all images used precise referring text.

Evaluation Protocol. For a fair and comprehensive evaluation, we employed the S-measure ($\mathcal{S}_m$) (Fan et al. 2017), mean and weighted F-measure ($\mathcal{F}_\beta^m$, $\mathcal{F}_\beta^\omega$) (Margolin, Zelnik-Manor, and Tal 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).

4.2 Implementation Details

All images are resized to 1024×1024 for training and testing. The output segmentation masks are resized to the original size of the corresponding ground-truth masks by bilinear interpolation. We employ the Pyramid Vision Transformer(PVT) 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 we use 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. All experiments are implemented with PyTorch 2.1 and are run on a machine with Intel(R) Xeon(R) Silver 4214R CPU @ 2.40GHz, 256GiB RAM, and two NVIDIA Titan A100-40G GPUs. The batch size is set to 2 for each GPU during training. All experiments use the same random seed. More implementation details will be provided in the supplementary materials.

4.3 Qualitative Analysis

In Fig. 3, we present a visual comparison of our TalNet trained on different labeled data split ratios (*i.e.* 1%, 5%, 10%) . We select various typical and challenging camouflaged images and arrange them in order of camouflaged object size, from smallest to largest. With the assistance of referring text, the model can locate highly camouflaged objects even when trained with only 1% labeled training data. As the amount of labeled training data increases, the edges of these camouflaged objects becomes progressively clearer.

4.4 Quantitative Analysis

In Tab. 1, we compared the proposed TalNet with two existing semi-supervised camouflaged object detection methods. To more comprehensively demonstrate the performance of our model, we used two different settings during testing: using fixed referring text (*i.e.* "A camouflaged object in the picture") and using precise referring text(*i.e.* image-level referring text annotation). As presented in Tab. 1, benefiting from the high-quality samples selected by the ADSA module and the auxiliary role of referring text during training, our method has reached a new state-of-the-art level. Even when using fixed referring text during testing, our method surpasses previous methods across all metrics on all testing sets. If precise referring text is further incorporated during testing, our method can achieve even greater performance

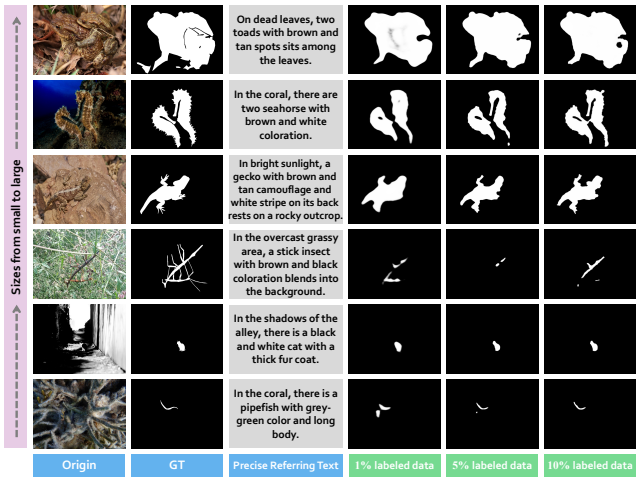

Figure 3: Visual comparisons of the proposed TalNet trained with different labeled data split ratios: 1%, 5%, and 10%. We used precise referring text for evaluation, and all unlabeled data used fixed referring text during the training phase.

improvements. For instance, when using precise referring text during testing as opposed to fixed referring text, our method shows an improvement of 0.017 in the mean absolute error($\mathcal{M}$) on the CAMO testing set and 0.013 on the NC4K testing set, even when trained with only 1% labeled data.

4.5 Ablation Study

Ablation of the proposed modules. We demonstrated the effectiveness of three modules: semi-supervised learning framework, ADSA module, and TFM. For all experimental groups that didn't use the ADSA module, we randomly selected 5% of the training set as labeled data and added annotations. For referring text, all experimental groups using TFM applied fixed referring text for unlabeled data during training and all data during testing. As shown in Tab. 2, we have proved the superiority of our methods. The proposed ADSA module effectively enhances the model's performance by selecting high-quality data, while the TFM fully utilizes the additional information contained in the text to assist the model in more accurately locating and segmenting camouflaged objects.

Ablation of the referring text. We further study the impact of referring text on model performance. We retrained the model in two different settings (using fixed referring text and using precise referring text) during the training and testing phases. As shown in Tab. 3, almost all performance metrics showed improvements under the setting of using precise referring text, which further proves the referring text could help the model better locate and segment the camouflaged object.

For more ablation study details and results, please refer to the supplementary materials.

Components			COD10K (2026)					
SSL	ADSA	TFM	$\mathcal{S}_m \uparrow$	$\mathcal{F}_\beta^\omega \uparrow$	$\mathcal{F}_\beta^m \uparrow$	$\mathcal{E}_\phi^m \uparrow$	$\mathcal{E}_\phi^x \uparrow$	$\mathcal{M} \downarrow$
✓			0.785	0.671	0.703	0.844	0.852	0.042
	✓		0.789	0.682	0.731	0.856	0.872	0.040
		✓	0.789	0.681	0.734	0.853	0.877	0.039
✓	✓		0.786	0.670	0.719	0.848	0.864	0.040
	✓	✓	0.793	0.686	0.735	0.851	0.875	0.038
✓	✓	✓	**0.790**	**0.689**	**0.745**	**0.851**	**0.885**	**0.036**

Table 2: Ablation study to evaluate the effectiveness of different components in our proposed method, including Semi-Supervised Learning framework (SSL), Active Data Selection and Annotation (ADSA), and Text Fusion Module (TFM). We use 5% labeled data to train the model and test on COD10K-testset with fixed referring text.

Train		Test		COD10K (2026)					
Fixed.	Precise.	Fixed.	Precise.	$\mathcal{S}_m \uparrow$	$\mathcal{F}_\beta^\omega \uparrow$	$\mathcal{F}_\beta^m \uparrow$	$\mathcal{E}_\phi^m \uparrow$	$\mathcal{E}_\phi^x \uparrow$	$\mathcal{M} \downarrow$
✓		✓		0.790	0.689	0.745	0.851	0.885	0.036
✓			✓	0.801	0.706	0.757	0.860	0.883	**0.033**
	✓	✓		0.771	0.667	0.736	0.814	0.847	0.038
	✓		✓	**0.808**	**0.718**	**0.769**	**0.870**	**0.980**	**0.033**

Table 3: Ablation study to evaluate the different settings on referring text. We retrained the model on 5% data with fixed and precise referring text, then tested on COD10K-test set with fixed and precise referring text.

5 Conclusions

In this paper, we address the shortcomings of existing semi-supervised COD methods, which fail to actively select and utilize high-quality data, resulting in poor performance. We introduce a semi-supervised and text-based referring COD model with active learning, TalNet, to address this issue. This model incorporates an active learning module that uses multi-feature clustering initialization and incremental sampling strategies to adaptively select high-quality samples for annotation. Additionally, we aim to leverage the semantic intelligence and inherent knowledge of MLLMs to assist in accurately localizing and segmenting camouflaged objects, thus maximizing the usage of annotated data. To this end, we further introduce a text fusion module that fuses semantic information from referring text and image features through a hierarchical shared multi-modal cross-attention mechanism.

Extensive experiments show that our approach outperforms previous semi-supervised methods in the COD field. Specifically, our method achieved an average improvement of **30.55%** in mean absolute error (MAE) compared to previous methods when trained with only **1%** labeled data, a **31.17%** MAE improvement when trained with **5%** labeled data, and a **35.69%** MAE improvement when trained with **10%** labeled data. These experimental results demonstrate that our strategies significantly enhance the detection accuracy of COD models. Additionally, our precise image-level referring text annotations for existing mainstream datasets(*i.e.* CHAMELEON, CAMO, COD10K, and NC4K) could provide a solid data foundation for subsequent COD-related research.

References

Angluin, D. 1988. Queries and Concept Learning. *Mach. Learn.*, 2(4): 319–342.

Argamon-Engelson, S.; and Dagan, I. 1999. Committee-Based Sample Selection for Probabilistic Classifiers. *Journal of Artificial Intelligence Research*, 11: 335–360.

Bai, J.; Bai, S.; Yang, S.; Wang, S.; Tan, S.; Wang, P.; Lin, J.; Zhou, C.; and Zhou, J. 2023. Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond. arXiv:2308.12966.

Berthelot, D.; Roelofs, R.; Sohn, K.; Carlini, N.; and Kurakin, A. 2022. AdaMatch: A Unified Approach to Semi-Supervised Learning and Domain Adaptation. arXiv:2106.04732.

Chen, G.; Liu, S.-J.; Sun, Y.-J.; Ji, G.-P.; Wu, Y.-F.; and Zhou, T. 2022. Camouflaged Object Detection via Context-Aware Cross-Level Fusion. *IEEE TCSVT*, 32: 1–1.

Chen, H.; Tao, R.; Fan, Y.; Wang, Y.; Wang, J.; Schiele, B.; Xie, X.; Raj, B.; and Savvides, M. 2023. SoftMatch: Addressing the Quantity-Quality Trade-off in Semi-supervised Learning. arXiv:2301.10921.

Chen, X.; Yuan, Y.; Zeng, G.; and Wang, J. 2021. Semi-Supervised Semantic Segmentation with Cross Pseudo Supervision. In *2021 CVPR*, 2613–2622.

Cheng, S.; Ji, G.-P.; Qin, P.; Fan, D.-P.; Zhou, B.; and Xu, P. 2023. Large Model Based Referring Camouflaged Object Detection. arXiv:2311.17122.

Chou, M.-C.; Chen, H.-J.; and Shuai, H.-H. 2022. Finding the Achilles Heel: Progressive Identification Network for Camouflaged Object Detection. In *2022 IEEE ICME*, 1–6.

Fan, D.-P.; Cheng, M.-M.; Liu, Y.; Li, T.; and Borji, A. 2017. Structure-measure: A New Way to Evaluate Foreground Maps. arXiv:1708.00786.

Fan, D.-P.; Gong, C.; Cao, Y.; Ren, B.; Cheng, M.-M.; and Borji, A. 2018. Enhanced-alignment Measure for Binary Foreground Map Evaluation. arXiv:1805.10421.

Fan, D.-P.; Ji, G.-P.; Cheng, M.-M.; and Shao, L. 2022. Concealed Object Detection. *IEEE TPAMI*, 44(10): 6024–6042.

Fan, D.-P.; Ji, G.-P.; Sun, G.; Cheng, M.-M.; Shen, J.; and Shao, L. 2020a. Camouflaged Object Detection. In *2020 CVPR*, 2774–2784.

Fan, D.-P.; Ji, G.-P.; Zhou, T.; Chen, G.; Fu, H.; Shen, J.; and Shao, L. 2020b. PraNet: Parallel Reverse Attention Network for Polyp Segmentation. arXiv:2006.11392.

Fan, D.-P.; Zhou, T.; Ji, G.-P.; Zhou, Y.; Chen, G.; Fu, H.; Shen, J.; and Shao, L. 2020c. Inf-Net: Automatic COVID-19 Lung Infection Segmentation From CT Images. *IEEE Transactions on Medical Imaging*, 39(8): 2626–2637.

Grandvalet, Y.; and Bengio, Y. 2004. Semi-supervised learning by entropy minimization. In *2004 NIPS*, 529–536.

Hu, J.; Lin, J.; Gong, S.; and Cai, W. 2024. Relax Image-Specific Prompt Requirement in SAM: A Single Generic Prompt for Segmenting Camouflaged Objects. In *AAAI 2025*, volume 38, 12511–12518.

Hu, R.; Rohrbach, M.; Andreas, J.; Darrell, T.; and Saenko, K. 2016a. Modeling Relationships in Referential Expressions with Compositional Modular Networks. arXiv:1611.09978.

Hu, R.; Xu, H.; Rohrbach, M.; Feng, J.; Saenko, K.; and Darrell, T. 2016b. Natural Language Object Retrieval. arXiv:1511.04164.

Huang, Z.; Dai, H.; Xiang, T.-Z.; Wang, S.; Chen, H.-X.; Qin, J.; and Xiong, H. 2023. Feature Shrinkage Pyramid for Camouflaged Object Detection with Transformers. arXiv:2303.14816.

Ji, G.-P.; Fan, D.-P.; Chou, Y.-C.; Dai, D.; Liniger, A.; and Van Gool, L. 2023a. Deep Gradient Learning for Efficient Camouflaged Object Detection. *Machine Intelligence Research*, 20: 92–108.

Ji, G.-P.; Fan, D.-P.; Xu, P.; Zhou, B.; Cheng, M.-M.; and Van Gool, L. 2023b. SAM struggles in concealed scenes — empirical study on "Segment Anything". *Science China Information Sciences*, 66(12).

Jia, Q.; Yao, S.; Liu, Y.; Fan, X.; Liu, R.; and Luo, Z. 2022. Segment, Magnify and Reiterate: Detecting Camouflaged Objects the Hard Way. In *2022 CVPR*, 4703–4712.

King, R. D.; Whelan, K. E.; Jones, F. M.; Reiser, P. G. K.; Bryant, C. H.; Muggleton, S. H.; Kell, D. B.; and Oliver, S. G. 2004. Functional genomic hypothesis generation and experimentation by a robot scientist. *Nature*, 427: 247–252.

Lai, X.; Yang, Z.; Hu, J.; Zhang, S.; Cao, L.; Jiang, G.; Wang, Z.; Zhang, S.; and Ji, R. 2024. CamoTeacher: Dual-Rotation Consistency Learning for Semi-Supervised Camouflaged Object Detection. In *2024 ECCV*.

Le, T.-N.; Cao, Y.; Nguyen, T.-C.; Le, M.-Q.; Nguyen, K.-D.; Do, T.-T.; Tran, M.-T.; and Nguyen, T. V. 2022. Camouflaged Instance Segmentation In-the-Wild: Dataset, Method, and Benchmark Suite. *IEEE Transactions on Image Processing*, 31: 287–300.

Le, T.-N.; Nguyen, T. V.; Nie, Z.; Tran, M.-T.; and Sugimoto, A. 2019. Anabranch network for camouflaged object segmentation. *Computer Vision and Image Understanding*, 184: 45–56.

Le, X.; Mei, J.; Zhang, H.; Zhou, B.; and Xi, J. 2020. A learning-based approach for surface defect detection using small image datasets. *Neurocomputing*, 408: 112–120.

Lee, D.-H. 2013. Pseudo-Label : The Simple and Efficient Semi-Supervised Learning Method for Deep Neural Networks. In *Workshop, ICML*.

Lin, J.; Tan, X.; Xu, K.; Ma, L.; and Lau, R. W. H. 2023. Frequency-aware Camouflaged Object Detection. *ACM Trans. Multimedia Comput. Commun. Appl.*, 19(2).

Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; and Belongie, S. 2017. Feature Pyramid Networks for Object Detection. In *2017 CVPR*, 936–944.

Liu, X.; Wang, Z.; Shao, J.; Wang, X.; and Li, H. 2019a. Improving Referring Expression Grounding with Cross-modal Attention-guided Erasing. arXiv:1903.00839.

Liu, Y.; Wan, B.; Zhu, X.; and He, X. 2019b. Learning Cross-modal Context Graph for Visual Grounding. arXiv:1911.09042.

Luo, N.; Pan, Y.; Sun, R.; Zhang, T.; Xiong, Z.; and Wu, F. 2023. Camouflaged Instance Segmentation via Explicit De-Camouflaging. In *2023 CVPR*, 17918–17927.

Luo, R.; and Shakhnarovich, G. 2017. Comprehension-guided referring expressions. arXiv:1701.03439.

Margolin, R.; Zelnik-Manor, L.; and Tal, A. 2014. How to Evaluate Foreground Maps. In *2014 CVPR*, 248–255.

Mei, H.; Ji, G.-P.; Wei, Z.; Yang, X.; Wei, X.; and Fan, D.-P. 2021. Camouflaged Object Segmentation with Distraction Mining. In *CVPR*.

Mi, P.; Lin, J.; Zhou, Y.; Shen, Y.; Luo, G.; Sun, X.; Cao, L.; Fu, R.; Xu, Q.; and Ji, R. 2022. Active Teacher for Semi-Supervised Object Detection. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 14462–14471.

Nagaraja, V. K.; Morariu, V. I.; and Davis, L. S. 2016. Modeling Context Between Objects for Referring Expression Understanding. arXiv:1608.00525.

Oliver, A.; Odena, A.; Raffel, C.; Cubuk, E. D.; and Goodfellow, I. J. 2018. Realistic evaluation of deep semi-supervised learning algorithms. In *Advances in neural information processing systems 31 (2018)*, 3239–3250.

OpenAI. 2024. GPT-4 Technical Report. arXiv:2303.08774.

Pang, Y.; Zhao, X.; Xiang, T.-Z.; Zhang, L.; and Lu, H. 2022. Zoom In and Out: A Mixed-scale Triplet Network for Camouflaged Object Detection. In *CVPR*.

Pang, Y.; Zhao, X.; Xiang, T.-Z.; Zhang, L.; and Lu, H. 2023. ZoomNeXt: A Unified Collaborative Pyramid Network for Camouflaged Object Detection. arXiv:2310.20208.

Pei, J.; Cheng, T.; Fan, D.-P.; Tang, H.; Chen, C.; and Gool, L. V. 2022. OSFormer: One-Stage Camouflaged Instance Segmentation with Transformers. arXiv:2207.02255.

Perazzi, F.; Krähenbühl, P.; Pritch, Y.; and Hornung, A. 2012. Saliency filters: Contrast based filtering for salient region detection. In *2012 CVPR*, 733–740.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. arXiv:2103.00020.

Ren, J.; Hu, X.; Zhu, L.; Xu, X.; Xu, Y.; Wang, W.; Deng, Z.; and Heng, P.-A. 2023. Deep Texture-Aware Features for Camouflaged Object Detection. *IEEE TCSVT*, 33(3): 1157–1167.

Settles, B. 2009. Active Learning Literature Survey.

Singh, S. K.; Dhawale, C. A.; and Misra, S. 2013. Survey of Object Detection Methods in Camouflaged Image. *IERI Procedia*, 4: 351–357. 2013 International Conference on Electronic Engineering and Computer Science (EECS 2013).

Sinha, S.; Ebrahimi, S.; and Darrell, T. 2019. Variational Adversarial Active Learning. arXiv:1904.00370.

Sohn, K.; Berthelot, D.; Li, C.-L.; Zhang, Z.; Carlini, N.; Cubuk, E. D.; Kurakin, A.; Zhang, H.; and Raffel, C. 2020a. FixMatch: simplifying semi-supervised learning with consistency and confidence. In *2020 NIPS*. ISBN 9781713829546.

Sohn, K.; Zhang, Z.; Li, C.-L.; Zhang, H.; Lee, C.-Y.; and Pfister, T. 2020b. A Simple Semi-Supervised Learning Framework for Object Detection. In *arXiv:2005.04757*.

Sun, D.; Jiang, S.; and Qi, L. 2023. Edge-Aware Mirror Network for Camouflaged Object Detection. arXiv:2307.03932.

Sun, Y.; Wang, S.; Chen, C.; and Xiang, T.-Z. 2022a. Boundary-Guided Camouflaged Object Detection. In Raedt, L. D., ed., *IJCAI-22*, 1335–1341. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Sun, Y.; Wang, S.; Chen, C.; and Xiang, T.-Z. 2022b. Boundary-Guided Camouflaged Object Detection. arXiv:2207.00794.

Tabernik, D.; Šela, S.; Skvarč, J.; and Skočaj, D. 2019. Segmentation-based deep-learning approach for surface-defect detection. *Journal of Intelligent Manufacturing*, 31(3): 759–776.

Tarvainen, A.; and Valpola, H. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *2017 NIPS*, 1195–1204. Curran Associates Inc.

Turkoglu, M.; and Hanbay, D. 2019. Plant disease and pest detection using deep learning-based features. *Turkish J. Electr. Eng. Comput. Sci.*, 27: 1636–1651.

Wang, K.; Bi, H.; Zhang, Y.; Zhang, C.; Liu, Z.; and Zheng, S. 2022a. D^2C-Net: A Dual-Branch, Dual-Guidance and Cross-Refine Network for Camouflaged Object Detection. *IEEE Transactions on Industrial Electronics*, 69(5): 5364–5374.

Wang, P.; Wu, Q.; Cao, J.; Shen, C.; Gao, L.; and van den Hengel, A. 2018. Neighbourhood Watch: Referring Expression Comprehension via Language-guided Graph Attention Networks. arXiv:1812.04794.

Wang, Q.; Yang, J.; Yu, X.; Wang, F.; Chen, P.; and Zheng, F. 2023. Depth-aided Camouflaged Object Detection. MM '23, 3297–3306. ISBN 9798400701085.

Wang, Y.; Wang, H.; Shen, Y.; Fei, J.; Li, W.; Jin, G.; Wu, L.; Zhao, R.; and Le, X. 2022b. Semi-Supervised Semantic Segmentation Using Unreliable Pseudo-Labels. In *2022 CVPR*, 4238–4247.

Wang, Z.; Li, Y.; Guo, Y.; Fang, L.; and Wang, S. 2021. Data-uncertainty guided multi-phase learning for semi-supervised object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Wu, Z.; Su, L.; and Huang, Q. 2019. Cascaded Partial Decoder for Fast and Accurate Salient Object Detection. arXiv:1904.08739.

Xiang, M.; Zhang, J.; Lv, Y.-Q.; Li, A.; Zhong, Y.; and Dai, Y. 2021. Exploring Depth Contribution for Camouflaged Object Detection.

Xu, M.; Zhang, Z.; Hu, H.; Wang, J.; Wang, L.; Wei, F.; Bai, X.; and Liu, Z. 2021. End-to-End Semi-Supervised Object Detection with Soft Teacher. *2021 ICCV*.

Yan, J.; Le, T.-N.; Nguyen, K.-D.; Tran, M.-T.; Do, T.-T.; and Nguyen, T. V. 2021. MirrorNet: Bio-Inspired Camouflaged Object Segmentation. *IEEE Access*, 9: 43290–43300.

Yang, L.; Zhuo, W.; Qi, L.; Shi, Y.; and Gao, Y. 2022. ST++: Make Self-trainingWork Better for Semi-supervised Semantic Segmentation. In *2022 CVPR*, 4258–4267.

Yang, S.; Li, G.; and Yu, Y. 2019. Dynamic Graph Attention for Referring Expression Comprehension. arXiv:1909.08164.

Yang, S.; Li, G.; and Yu, Y. 2020. Relationship-Embedded Representation Learning for Grounding Referring Expressions. arXiv:1906.04464.

Yu, L.; Poirson, P.; Yang, S.; Berg, A. C.; and Berg, T. L. 2016. Modeling Context in Referring Expressions. arXiv:1608.00272.

Zhai, Q.; Li, X.; Yang, F.; Chen, C.; Cheng, H.; and Fan, D.-P. 2021. Mutual Graph Learning for Camouflaged Object Detection. In *2021 CVPR*.

Zhang, M.; Xu, S.; Piao, Y.; Shi, D.; Lin, S.; and Lu, H. 2022. PreyNet: Preying on Camouflaged Objects. *2022 ACM MM*.

Zhang, X.; Yin, B.; Lin, Z.; Hou, Q.; Fan, D.-P.; and Cheng, M.-M. 2023. Referring Camouflaged Object Detection. arXiv:2306.07532.

Zheng, P.; Gao, D.; Fan, D.-P.; Liu, L.; Laaksonen, J.; Ouyang, W.; and Sebe, N. 2024. Bilateral Reference for High-Resolution Dichotomous Image Segmentation. *CAAI Artificial Intelligence Research*.

Zhong, Y.; Li, B.; Tang, L.; Kuang, S.; Wu, S.; and Ding, S. 2022. Detecting Camouflaged Object in Frequency Domain. In *2022 CVPR*, 4494–4503.

Zhou, T.; Zhou, Y.; Gong, C.; Yang, J.; and Zhang, Y. 2022. Feature Aggregation and Propagation Network for Camouflaged Object Detection.

Zhu, H.; Li, P.; Xie, H.; Yan, X.; Liang, D.; Chen, D.; Wei, M.; and Qin, J. 2022. I Can Find You! Boundary-Guided Separated Attention Network for Camouflaged Object Detection. In *2022 AAAI*.

Zhu, J.; Zhang, X.; Zhang, S.; and Liu, J. 2021. Inferring Camouflaged Objects by Texture-Aware Interactive Guidance Network. In *2021 AAAI*.

Zhuge, M.; Lu, X.; Guo, Y.; Cai, Z.; and Chen, S. 2022. CubeNet: X-shape connection for camouflaged object detection. *Pattern Recognition*, 127: 108644.