Catch Me If You Can Describe Me: Open-Vocabulary Camouflaged Instance Segmentation with Diffusion

Tuan-Anh Vu[1,2*], Duc Thanh Nguyen[3*], Qing Guo[4*], Nhat Chung[2,5], Binh-Son Hua[6], Ivor W. Tsang[2], Sai-Kit Yeung[1]

[1]The Hong Kong University of Science and Technology, Hong Kong.
[2]CFAR & IHPC, A*STAR, Singapore.
[3]Deakin University, Australia.
[4]Nankai University, China.
[5]National University of Singapore, Singapore.
[6]Trinity College Dublin, Ireland.

*Corresponding author(s). E-mail(s): tavu@connect.ust.hk; duc.nguyen@deakin.edu.au; tsingqguo@ieee.org;
Contributing authors: nhatcm@u.nus.edu; binhson.hua@tcd.ie; ivor_tsang@a-star.edu.sg; saikit@ust.hk;

Abstract

Text-to-image diffusion techniques have shown exceptional capabilities in producing high-quality, dense visual predictions from open-vocabulary text. This indicates a strong correlation between visual and textual domains in open concepts and that diffusion-based text-to-image models can capture rich and diverse information for computer vision tasks. However, we found that those advantages do not hold for learning of features of camouflaged individuals because of the significant blending between their visual boundaries and their surroundings. In this paper, while leveraging the benefits of diffusion-based techniques and text-image models in open-vocabulary settings, we aim to address a challenging problem in computer vision: open-vocabulary camouflaged instance segmentation (OVCIS). Specifically, we propose a method built upon state-of-the-art diffusion empowered by open-vocabulary to learn multi-scale textual-visual features for camouflaged object representation learning. Such cross-domain representations are desirable in segmenting camouflaged objects where visual cues subtly distinguish the objects from the background, and in segmenting novel object classes which are not seen in training. To enable such powerful representations, we devise complementary modules to effectively fuse cross-domain features, and to engage relevant features towards respective foreground objects. We validate and compare our method with existing ones on several benchmark datasets of camouflaged and generic open-vocabulary instance segmentation. The experimental results confirm the advances of our method over existing ones. We believe that our proposed method would open a new avenue for handling camouflages such as computer vision-based surveillance systems, wildlife monitoring, and military reconnaissance.

Keywords: Camouflaged object detection, camouflaged instance segmentation, instance segmentation, text-to-image diffusion, text-image transfer, open vocabulary segmentation.

1 Introduction

Camouflage is a powerful biological mechanism for avoiding detection and identification. In nature, camouflaged tactics are employed to deceive the sensory and cognitive processes of both prey and predators. Wild animals utilise these tactics in various ways, ranging from blending themselves into the surrounding environment to employing disruptive patterns and colouration (Nguyen et al., 2023). Thus, identifying camouflages is pivotal in many wildlife surveillance applications (Fleming et al., 2014; Yan et al., 2021), as it helps locate hidden individuals for monitoring and conservation.

In fact, localisation of camouflaged objects (Fan et al., 2020; C. He et al., 2023), such as Camouflaged Object Detection (COD) and Camouflaged Instance Segmentation (CIS), has been an important research topic in computer vision, whose main challenge lies in the need to learn discriminative features that for discerning camouflaged target objects from their surroundings. Existing COD techniques can be utilised to roughly identify camouflaged objects at regional scales (*e.g.*, bounding boxes), but they are not designed to distinguish individual instances at finer scales like pixel level. CIS, on the other hand, operates under the assumption that individual instances' features closely resemble one another and aims to provide class-independent segmentation masks (Pei et al., 2022). However, the diversity of camouflages within a single scene can lead to complex intertwining patterns, making the CIS task more challenging in severe environmental conditions, *e.g.*, terrestrial and aquatic environments, under poor imaging quality, *e.g.*, occlusions, image blur, and low-light conditions in underwater applications. These challenges also hinder the collection and annotation of high-quality data for training and testing CIS algorithms.

Meanwhile, while humans can recognise an unlimited number of target categories, and open-vocabulary recognition has been developed to mimic human intelligence with unbounded understanding, current endeavours focus only on generic objects and individuals (Du et al., 2022; Gao et al., 2022; Ghiasi et al., 2022; Kuo et al., 2023; Minderer et al., 2022; J. Xu et al., 2023; Zang et al., 2022). For example, while (J. Xu et al., 2023) suggested that Internet-scale text-to-image diffusion models can be utilised to create a state-of-the-art open-vocabulary segmenter for many concepts, our investigations show that they demonstrate inconsistent segmentation results when working with camouflages, as indicated by their pixel-wise embeddings in Figure 1. Although pretrained generative features offer strong potential for open-vocabulary generalization, our findings highlight their limitations in capturing fine-grained visual ambiguities such as camouflage. Notably, existing open-vocabulary segmentation methods (Ding et al., 2023; J. Xu et al., 2023; X. Xu et al., 2023; H. Zhang et al., 2023; Zou, Dou, et al., 2023; Zou, Yang, et al., 2023) share this limitation, as camouflage detection is not central to their design.

To overcome the aforementioned hurdles, we propose a method that leverages text-to-image diffusion to address the problem of OVCIS. Our method is inspired by the advanced object representation learning capabilities of diffusion techniques and the language-vision transferability of text-image models. Text-to-image diffusion models, *e.g.*, the stable diffusion model by (Robin et al., 2022), are designed to learn essential object features in the presence of noise, making them useful for extracting features relevant to target objects in a noisy and cluttered background. While we observed that features learnt solely from the visual domain are weak to distinguish camouflaged objects from their surroundings, the features learnt by text-image discriminative models, *e.g.*, CLIP (Radford et al., 2021), contain rich information about the real world thanks to the variety of concepts in open-vocabulary training data (J. Wu et al., 2024). We hypothesize that an effective combination of features learnt from both the textual and visual domains would benefit the representation learning of camouflaged objects. We illustrate the effectiveness of textual-visual representations for CIS in Figure 1. To the best of our knowledge, such a cross-domain combination with open-vocabulary for CIS is *novel*, and ours is the *first framework* to localise camouflaged object instances at this scale.

To effectively learn textual-visual representations of camouflaged objects, our method assimilates an input image and a text prompt about objects included in the input image, so the input image and its implicit caption (generated by a captioner) are integrated into a text-to-image diffusion model to extract visual features. These features are processed at multiple scales and fused into a visual feature map, which is then used to generate object masks. Simultaneously, textual features

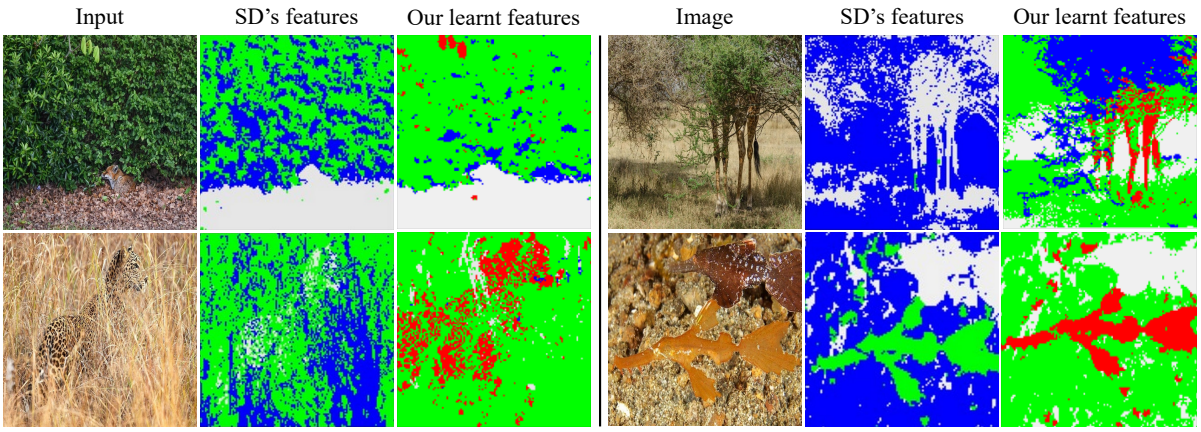

Fig. 1: Illustration of textual-visual features of off-the-shelf Stable Diffusion when dealing with CIS and our learnt features. Given an input image, textual-visual features are extracted and clustered using a K-means clustering algorithm ($K = 4$). As shown, camouflaged animals can be localised based on the clustering results. We leverage these rich features to perform instance segmentation of camouflaged objects. This figure is best viewed in colour.

are extracted from the text prompt using a text encoder. These textual features are enriched from open-vocabulary category labels and proven to improve the discriminative power of camouflaged objects' representations against the background. Our proposed pipeline aggregates textual and visual features in a mask-out manner to recognise the masks of the target objects. The diffusion model utilises a cross-attention mechanism to link textual features with visual features and condition the feature learning process. Hence, the learnt features are likely to be distinct and connected to high/mid-level semantic notions that may be expressed in the language part. While our method somewhat shares a similar approach with the works by (J. Xu et al., 2023; Zhao et al., 2023) at a high-level perspective, our pipeline is more specialised to CIS by designing camouflage-specialised modules.

COD vs. CIS vs. OVCIS. Camouflaged Object Detection (COD) aims to separate camouflaged regions from background and typically produces a *binary* camouflage mask without requiring instance separation. Camouflaged Instance Segmentation (CIS) extends COD by requiring *instance-level* separation of multiple camouflaged objects, but prior CIS formulations are often class-agnostic or focus primarily on instance delineation rather than open-vocabulary semantic generalization (Pei et al., 2022). In contrast, Open-Vocabulary Camouflaged Instance Segmentation (OVCIS) requires *both* (i) robust instance separation under camouflage and (ii) *open-vocabulary* category assignment at inference via textual category prompts, where training categories $\mathcal{C}_{\text{train}}$ and test categories $\mathcal{C}_{\text{test}}$ may be disjoint. While open-vocabulary segmentation has been explored in general-domain settings (Ding et al., 2023; J. Xu et al., 2023; X. Xu et al., 2023; H. Zhang et al., 2023; Zou, Dou, et al., 2023; Zou, Yang, et al., 2023), existing methods are not designed to address the boundary ambiguity and low-contrast appearance intrinsic to camouflage. Accordingly, OVCIS lies at the intersection of camouflage understanding, instance segmentation, and open-vocabulary recognition (Table 1).

In summary, we make the following contributions to our work:

- We address a new and challenging task: open-vocabulary camouflaged instance segmentation (OVCIS), which would enhance the capability of many critical applications such as computer vision-based surveillance systems, wildlife monitoring, and military reconnaissance.
- We propose a method for OVCIS built upon text-to-image diffusion and text-image transfer techniques, advanced with open-vocabulary utilisation.
- We propose an object representation learning paradigm specialised for camouflages. Our

Table 1: Conceptual distinctions among Camouflaged Object Detection (COD), Camouflaged Instance Segmentation (CIS), and Open-Vocabulary Camouflaged Instance Segmentation (OVCIS). OVCIS combines camouflaged instance separation with open-vocabulary category assignment at inference.

Task	Primary output	Instance separation	Vocabulary regime	Typical supervision
COD	Binary mask (camouflage vs. background)	×	N/A	Binary mask annotation
CIS	Instance masks (often class-agnostic)	✓	Closed or class-agnostic	Instance masks (w/ or w/o categories)
OVCIS (Ours)	Instance masks & open-vocab category labels	✓	Open-vocabulary ($\mathcal{C}_{\text{train}} \rightarrow \mathcal{C}_{\text{test}}$)	Instance masks + category names/text

camouflage-specialised components include a Multi-scale Features Fusion (MSFF) module to encapsulate visual features from diffusion, a Textual-Visual Aggregation (TVA) module to utilise textual information that pronounces visual features, and a Camouflaged Instance Normalisation (CIN) module to adaptively capture textual-visual information that enhances camouflaged object representations.

- We conduct extensive experiments and ablation studies that demonstrate the advantages of our method over existing works.

2 Related Work

We start our review of related work with an overview of deep learning-based advances for camouflaged object understanding. Following it, we delve into contemporary research in text-to-image diffusion, thereby discussing their role in facilitating open-vocabulary computer vision tasks. Then, we review prior research on generative models and their applications to visual segmentation.

2.1 Camouflaged Object Understanding

The main aim of camouflaged object understanding lies in learning object representations that are difficult to dissimilate from their background. Existing research has attempted to address various tasks in camouflaged object understanding from images. For instance, (G. Sun et al., 2023) counted objects that blended seamlessly into backgrounds. Following closely, (Lyu et al., 2021) identified salient image regions of hidden objects that align with the nuances of human perception. COD was studied by (C. He et al., 2023), in which the authors decomposed learnt features into different frequency bands using learnable wavelets to identify the most informative features to differentiate target objects and backgrounds. In addition, an auxiliary reconstruction network was built to boost up further the discriminative power of the foreground's features against the background's ones. In the work by (Fan et al., 2022), a method for segmenting camouflaged objects was proposed to segment obscured objects without pinpointing specific categories for the objects.

CIS was brought forth by (Pei et al., 2022) to emphasise the learning of object-vs-background-discriminative representations, which is different from general instance segmentation (Xie et al., 2021) that aims to maximise inter-object distances. Although this goal is common in existing camouflaged object understanding methods and various attempts have been made to address it in the literature, learning such representations from solely imagery data is challenging as it is the nature of visual camouflages. Our research differs from existing ones by exploring the potential of diffusion-based representations and textual data as additional cues to drive the open-vocabulary learning of CIS, thereby utilising them to make camouflaged object representations adaptive to camouflages that are never seen in training.

Thanks to the variety of concepts, textual features learnt from text prompts about objects included in an input image can help to find visual features relevant to the objects. In addition, an effective combination of both textual and visual features would further enhance the robustness of camouflaged object representations, where visual features solely are not robust enough to distinguish camouflaged objects from their surroundings. To the best of our knowledge, our study is the first of such work.

2.2 Text-to-Image Diffusion

Significant progress has been made in Artificial Intelligence (AI)-empowered picture creation with recent advances in large-scale text-to-image diffusion models, including Stable Diffusion (Robin et al., 2022), DALL-E 2 (Ramesh et al., 2022), and Imagen (Saharia et al., 2022). These models have demonstrated photo-realistic quality image generation by being trained on text-image datasets of substantial scale sourced from the Internet. They also have shown the ability to be conditioned on unrestricted text prompts in order to produce visuals that closely resemble real-life photographs.

The application of text-to-image diffusion models has facilitated the creation and manipulation of visual contents in an ever-easy and convenient manner via language-based interactions (*e.g.*, text prompts). This has enabled a wide spectrum of applications such as content-personalised customisation (Kumari et al., 2023), zero-shot translation (Parmar et al., 2023), content editing (Hertz et al., 2023), and image generation (Gal et al., 2023).

In this paper, we do not apply text-to-image diffusion technique to image creation and/or image manipulation. Instead, we explore its capability of cross-domain feature learning. Most related to our work, (J. Xu et al., 2023) showed that pre-trained representations in diffusion models can be utilised for open-vocabulary segmentation. However, we found that their method performs poorly and inconsistently on camouflaged datasets, due to a lack of ability to identify object boundaries in camouflages. To address this limitation, we devise a feature fusion strategy based on a state-of-the-art text-to-image diffusion architecture to fuse image features with implicit caption features at multiple scales. Our experiments show that such a fusion facilitates the learning of object-vs-background discriminative features, which are crucial for CIS.

2.3 Generative Models for Segmentation

Many studies are related to our work in terms of applying image generative models, such as Generative Adversarial Networks (GANs) (Esser et al., 2021; Karras et al., 2020) or diffusion models (Dhariwal & Nichol, 2021; Ho et al., 2020; J. Song et al., 2021), to semantic segmentation (Baranchuk et al., 2022; D. Li et al., 2022; Rewatbowornwong et al., 2023). For GANs, a straightforward approach is to synthesise images and their corresponding semantic maps to train a segmentation network (D. Li et al., 2022). (Rewatbowornwong et al., 2023), segmentation is performed by training a generative model on datasets with a limited vocabulary. For example, (Baranchuk et al., 2022) proposed a diffusion-based framework, named DDPMSeg, based on the denoising diffusion probabilistic model (DDPM) (Ho et al., 2020) to learn a feature map for an input image. The feature map was then passed to a pixel classifier to perform semantic or part segmentation. A few hand-annotated examples per category are then utilised to classify learnt representations into semantic regions. Similarly, J. Xu et al. (2023) showed that pre-trained representations in diffusion models can be utilised for open-vocabulary segmentation in the wild. Their insights suggest that internal representations learnt by diffusion models can well correlate with high- and mid-level semantic concepts that can be described in language, addressing the lack of spatial and relational understanding in traditional open-vocabulary segmentation. Therefore, their approach introduces a new capacity for generative models, *e.g.*, image generation-driven representation learning. However, while promising as a practical tool, we found that diffusion-based pre-trained representations are not designed to tackle camouflaging effects, even though the intermediate representations of a generative model can be trained to capture high-level semantic concepts (*e.g.*, the presence of an object in an input image) under specific feature constraints.

2.4 Open-Vocabulary Detection and Segmentation

Numerous studies have been proposed to incorporate vision-language models (VLMs) into open-vocabulary detection and segmentation (Du et al., 2022; Gao et al., 2022; Ghiasi et al., 2022; Kuo et al., 2023; Minderer et al., 2022; Rasheed et al., 2022; Zang et al., 2022; Zhong et al., 2022). This has enabled the detection and classification of novel objects from a vast conceptual domain with the help of pre-trained VLMs (J. Wu et al., 2024; J. Zhang et al., 2023). OVR-CNN was the first open-vocabulary object detection introduced by (Zareian et al., 2021), which underwent pre-training with image-caption data in order to learn and identify unknown objects, followed by fine-tuning for zero-shot detection.

Following recent advances in VLMs (Jia et al., 2021; Radford et al., 2021), ViLD (Gu et al., 2022) pioneered the incorporation of extensive representations of pre-trained CLIP (Radford et al., 2021) into an object detector, and many works (Du et al., 2022; Kuo et al., 2023; X. Zhou et al., 2022) have followed the similar framework. (Du et al., 2022) proposed DetPro, a sophisticated automated prompt learning method, to learn the presence of an object in a background via prompt training. F-VLM (Kuo et al., 2023) adopted a frozen VLM to generate new object categories based on cropped CLIP features. (X. Zhou et al., 2022) extended the ability of the well-known object detector, Faster R-CNN (Ren et al., 2015) to newly introduced object categories by replacing the classification weights (in the classification head) by fixed language embeddings learnt from open-vocabulary.

Despite the successes achieved, existing methods have limited capabilities against camouflaged objects due to the utilisation of small closed vocabularies and/or the incorporation of VLMs for generic object classes, which are often distinguishable from the background. It is because the pre-trained representations learnt on general object classes are not designed for discerning object boundaries between camouflaged individuals (Ding et al., 2023; J. Xu et al., 2023; X. Xu et al., 2023; H. Zhang et al., 2023; Zou, Dou, et al., 2023; Zou, Yang, et al., 2023). While exploiting insights and advantages from prior studies, our work stands out in a specifically focused direction: tackling the challenge of open-vocabulary instance segmentation of camouflaged targets, yet without losing much representation localisation capability on general objects. Our proposed method extend towards segmentation of novel object categories with concealed appearances in natural environments using an open-vocabulary set.

3 Proposed Method

3.1 Problem Definition

We aim to build and train an instance segmentation model with a set of pre-defined object categories, referred to as $\mathbf{C}_{\text{train}}$. The instance segmentation model can work on a new domain with $\mathbf{C}_{\text{test}}$ object categories, where $\mathbf{C}_{\text{test}}$ and $\mathbf{C}_{\text{train}}$ may or may not share common object categories. In other words, $\mathbf{C}_{\text{test}}$ may include object categories previously unseen during the training of the instance segmentation model.

Throughout the training process, it is presumed that binary mask annotations for target objects in each training image are available. Moreover, each mask is either associated with a category name or a caption presented in the text form. During the testing phase, however, neither the category label nor the caption is accessible for any test image. Note that, only the names of the test categories in $\mathbf{C}_{\text{test}}$ are provided.

3.2 Overview

3.2.1 Preliminaries

We build our method upon two technical advances: text-to-image diffusion and text-image transfer. We first briefly summarise those techniques and then describe how they can be applied to our method.

Text-to-Image Diffusion facilitates the creation of high-quality images guided by text prompts. A text-to-image diffusion model is trained on a massive corpus of image-text pairs amassed through web crawling, as indicated in the literature (Nichol et al., 2022; Saharia et al., 2022; J. Xu et al., 2023). Text inputs are encoded into embeddings using an established text encoder, e.g., T5 (Raffel et al., 2020). An image is initially perturbed by introducing Gaussian noise at a controlled intensity before being fed into the diffusion network. The network is fine-tuned to reverse the noise

application, utilising noisy images and associated text embeddings to diminish the distortion. In the inference phase, the model synthesises an image from inputs, including pure Gaussian noise shaped to the image's dimensions and a user-provided description's text embedding. Through successive inference iterations, the model iteratively denoises the input and finally results in a photo-realistic image of the user-provided text description.

In our work, we adopt the Stable Diffusion (SD) model developed by (Robin et al., 2022) and pre-trained on the LAION-5B dataset (Schuhmann et al., 2022). SD is chosen for two reasons. First, SD is well known for its ability in effective fusion of textual and visual information, which we found useful for camouflaged instance segmentation where visual features only can be indistinguishable. Second, thanks to the denoising process, SD is able to manage noisy and subtle visual distinctions effectively, making them particularly suitable for camouflage segmentation where visual boundaries blend significantly with the background.

The SD model is composed of a trio of elements: ① a captioner (realised by a pre-trained text encoder) that generates a text embedding (implicit caption) for an input image; ② a pre-trained variational auto-encoder for learning of image representations; and ③ a denoising time-conditional U-Net $\epsilon_\theta(\cdot)$, which applies progressive convolution operations to downsample and upsample feature maps of an input image with skip connections. Within the U-Net, textual-visual interactions are enabled by cross-attention. In detail, the captioner projects a text input y into an embedding, which is then transformed into `Key` and `Value` pairs. At the same time, a feature map of a noisy image undergoes a linear projection to form a `Query`. This design allows for iterative updates of input images conditioned on accompanying text descriptions.

The training process of the SD model is outlined as follows. For a given pair $(\mathcal{I}, y)$ in a training dataset, the image $\mathcal{I}$ is encoded into a latent representation z and then subjected to noise, resulting in a noised vector $z^t := \alpha^t z + \sigma^t \epsilon$, where $\epsilon \sim \mathcal{N}(0, 1)$ is a noise variable, and α^t, σ^t are parameters that manage the noise level and the fidelity of each sample. The training aims to fine-tune the time-conditional U-Net $\epsilon_\theta(\cdot)$ to anticipate the noise vector ϵ and to accurately reconstruct the initial latent vector z, while being conditioned on the text input y. The fine-tuning is performed by using a loss function that minimises the mean squared error of noise prediction as follows:

$$\mathcal{L}_{\text{diffusion}} = \mathbb{E}_{z,\epsilon\sim\mathcal{N}(0,1),t,y}\left[||\epsilon - \epsilon_\theta(z^t, t, y)||_2^2\right] \quad (1)$$

where the time variable t is randomly selected from the set $\{1, \ldots, T\}$.

During the inference phase, the SD model synthesises an image by sequentially refining a latent vector $z^T \sim \mathcal{N}(0, I)$, with the process being contingent on a text input y. Specifically, for each time step $t = 1, \ldots, T$ of the denoising sequence, z^{t-1} is derived from the current z^t and the U-Net's noise prediction, which in turn takes z^t and the text prompt y as inputs. After the final denoising stage, the latent vector z^0 is transformed back to produce a final output image $\mathcal{I}'$.

Text-Image Transfer originally aims to learn directly from raw text about images. This technique leverages rich textual representations learnt from the textual domain to scale up representation learning in the visual domain. As shown in the literature, natural language can be used to supervise a wide set of visual concepts through its generality (Desai & Johnson, 2021; Sariyildiz et al., 2020; Y. Zhang et al., 2022). Recently, CLIP proposed by (Radford et al., 2021) offers text-image transferibility in both directions, i.e., text-to-image and image-to-text.

In our work, we adopt a CLIP model, pre-trained on 400 million image-text pairs crawled from the Internet. This model is used to generate text embeddings for implicit captions of input images and text embeddings for text prompts associated with input images. Due to learning from large-scale and diverse training data, we observed that these text embeddings greatly aid in improving camouflaged objects' representation.

3.2.2 Our Pipeline

Figure 2 illustrates the pipeline of our method. At an abstract level, our method takes an image and a text prompt about target objects as inputs and produces instance masks with object categories for the target objects as outputs.

The input image is first passed to the SD model (Robin et al., 2022), which is pre-trained and frozen (no training), to extract latent features. The input image is also fed to the pre-trained and

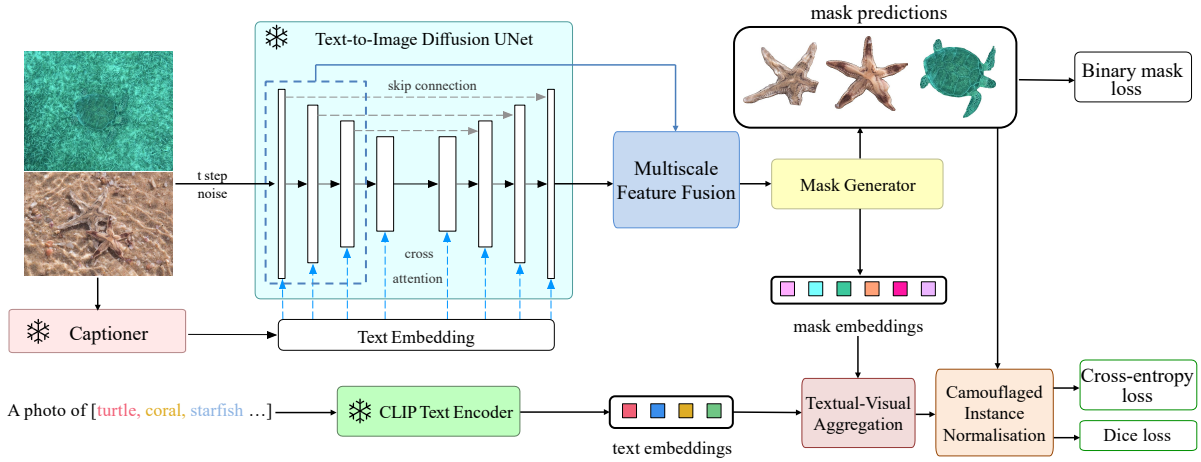

Fig. 2: Pipeline of our proposed method for Open-vocabulary Camouflaged Instance Segmentation (OVCIS). Inputs include an image and a text prompt about target objects in the input image. Outputs include instance masks of the target objects. The target objects can be novel and have never been seen in the training data. We leverage state-of-the-art text-to-image diffusion and text-image transfer techniques to learn textual-visual features that facilitate the object representation learning for segmenting camouflaged objects.

frozen CLIP model (Radford et al., 2021) to calculate its implicit caption embedding. The caption embedding is inserted into the SD model at various scales (layers) and fused with the SD model's last layer to form image-guided features. We call these features "image-guided features" though they somewhat include textual information. This is because the input image drives the textual features from the implicit caption embedding. The image-guided features, coupled with annotated training masks, serve as inputs to train a mask generator capable of producing instance masks for all potential categories within the input image. The instance masks are then used to locate object-relevant features in a mask-out manner. This step results in mask embeddings (i.e., features extracted within masked regions).

The input text prompt is concurrently processed by the CLIP (Radford et al., 2021), independently of the input image, and its corresponding text embeddings are calculated. These text embeddings are transferable to visual features yet extracted from the textual input, hence considered as "text-guided features". The text embeddings (text-guided features) and mask embeddings (image-guided features) are aggregated by a textual-visual aggregation module, which aims to emphasise the learnt features towards foreground objects defined in the input text prompt. This module results in a textual-visual representation for the input image and text prompt.

Next, the textual-visual representation is normalised regarding the instance masks segmented by the mask generator and finally classified by a mask classifier into object categories.

The entire pipeline is trained with object categories in $\mathbf{C}_{\text{train}}$. Note that, since the SD and CLIP models have been pre-trained and frozen, the training of the entire pipeline is equivalent to learning of parameters in modules specialised for CIS (multi-scale feature fusion, mask generator, textual-visual aggregation, camouflaged instance normalisation). Once the training is completed, the inference process carries out open-vocabulary instance segmentation, i.e., the pipeline can perform instance segmentation of object categories in $\mathbf{C}_{\text{test}}$.

To make our pipeline specialised to CIS, we develop several technical components to facilitate camouflaged object representation learning (see Section 3.3) and camouflaged instance normalisation (see Section 3.4).

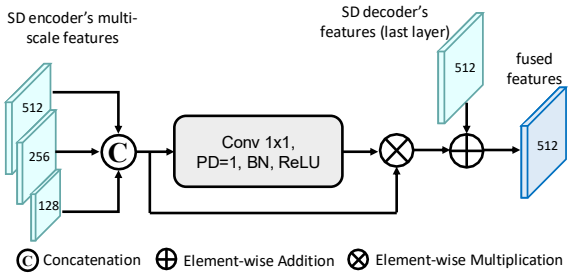

Fig. 3: Architecture of the multi-scale features fusion (MSFF) module.

3.3 Camouflaged Object Representation Learning

Given the features learnt by the SD model from the input image and the text embeddings produced by the CLIP from the input text prompt, we perform camouflaged object representation learning via three modules: multi-scale feature fusion, mask generator, and textual-visual aggregation. These modules are described below.

3.3.1 Multi-scale Features Fusion

The MSFF module fuses the multi-scale features from the encoder part of the SD model and the features from the last layer of the decoder part of the SD model. We present the architecture of the MSFF module in Figure 3.

The fusion process concatenates multi-scale SD encoder features and applies the 1×1 convolution on the concatenated features. The resulting features are then combined with the concatenated features via element-wise multiplication, and the modulated output is added to the SD decoder's final-layer features.

3.3.2 Mask Generator

We adopt the decoder in the mask-attention Transformer, the core component in the Mask2Former architecture (Cheng et al., 2022), to realise our mask generator. The mask generator receives input as a fused feature vector from the MSFF module and produces outputs including N class-agnostic binary masks $\{m_i^{pred}\}_{i=1}^{N}$ and their corresponding N mask embedding features $\{z_i^{pred}\}_{i=1}^{N}$ for all possible objects in the input image. We illustrate the mask generator in Figure 4.

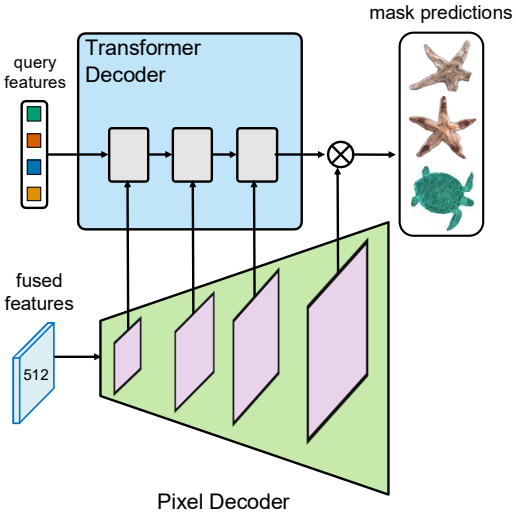

Fig. 4: Architecture of the mask generator.

The mask generator employs a pixel decoder that progressively increases the resolution of the fused features from the MSFF module and generates per-pixel high-resolution embeddings. This pixel decoder is designed meticulously, using multiple layers to capture fine-grained and broad contextual information. Following that, a Transformer's decoder processes the intermediate feature maps in the pixel encoder to handle object queries, which are initialised randomly but then learnt through training. To effectively process the intermediate feature maps in the pixel decoder, the mask generator guides each feature map at a scale to an individual layer in the Transformer's decoder. Consequently, each layer in the Transformer's decoder focuses on a feature map at a specific scale in the range of $\{1/32, 1/16, 1/8\}$. We observed that this strategy significantly enhances the ability of the mask generator to adeptly handle objects in various sizes.

3.3.3 Textual-Visual Aggregation

The TVA module is designed to highlight object-relevant features to drive the object representation learning towards foreground objects, whose architecture is shown in Figure 5. We later show that experimental results validated its effectiveness.

The TVA module in our proposed pipeline operates as follows. Like the Mask R-CNN (K. He et al., 2017), we crop corresponding features from the MSFF module and perform mask pooling for each

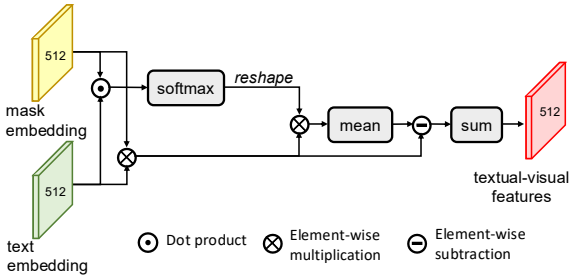

Fig. 5: Architecture of the textual-visual aggregation (TVA) module.

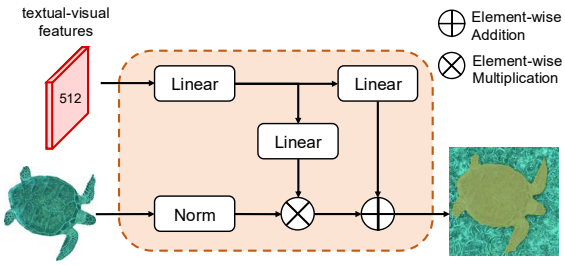

Fig. 6: Architecture of the camouflaged instance normalisation (CIN) module.

object mask returned by the mask generator. This step results in mask embeddings (i.e., embeddings are determined by masks). We then compute the interactions between these mask embeddings and the text embeddings produced by the CLIP. Nevertheless, instead of directly using a dot product to calculate the interaction between two embeddings as in CLIP (Radford et al., 2021), we apply a softmax operator to the dot product of the embeddings to weight features, then apply mean-normalisation to remove irrelevant features before aggregating them by a channel-wise summation. Removing irrelevant features helps to mitigate the problem of noisy activations, making the learning process lean towards features relevant to the object categories specified in the input text prompt.

Figure 1 visualises learnt textual features by our method on several challenging cases. As shown, the learnt textual-visual features on camouflaged objects can be well identified and located, although the objects blend into cluttered backgrounds. This is evident in the ability of our method to learn distinguished object-vs-background features.

3.4 Camouflaged Instance Normalisation

Inspired by the adaptive instance selection network (X. Huang & Belongie, 2017; Pei et al., 2022), we develop a CIN module to achieve final masks for the target objects. We present the architecture of the CIN module in Figure 6.

The CIN module takes inputs as a textual-visual feature map from the TVA module and an object mask from the mask generator. A linear layer first projects the textual-visual feature map into a higher-dimensional space. Next, affine weights and biases are attained by applying two subsequent linear layers to the result of the first linear layer. The affine weights and biases are then combined, together with the input mask from the mask generator, to predict a final instance mask for the object specified in the input mask. Since the CIS task is category-agnostic, we use a confidence score for the existence of a camouflaged object at each location, rather than a classification score in generic instance segmentation.

3.5 Training

We train the entire pipeline of our method by optimising the loss functions (binary mask, cross-entropy, dice losses) used in the mask generator and the CIN module in a supervised fashion.

Specifically, we adopt a binary cross-entropy loss as our binary mask loss $\mathcal{L}_{bce}$ and a dice loss $\mathcal{L}_{dice}$ (Milletari et al., 2016) for supervising binary mask predictions in the mask generator. The dice loss is used to remedy class imbalance.

We carry out the training of the CIN module using the conventional close-vocabulary training approach. Suppose that we can access to the ground-truth category label for each object mask during the training phase. For each mask embedding z_i^{pred} produced by the mask generator, let $y_i^{cate} \in \mathbf{C}_{\text{train}}$ be the corresponding ground-truth category of z_i^{pred}. We invoke the text encoder $\mathcal{T}$ in the pre-trained CLIP model to encode the names of all categories in $\mathbf{C}_{\text{train}}$. This results in a set of text embeddings $\mathcal{T}(\mathbf{C}_{\text{train}}) = \{\mathcal{T}(c_1), \ldots, \mathcal{T}(c_{|\mathbf{C}_{\text{train}}|})\}$ where $c_k \in \mathbf{C}_{\text{train}}$ represents a category name.

The loss for embedding classification (i.e., associating mask embeddings m_i^{pred} with their

categories y_i^{cate}) is calculated as:

$$\mathcal{L}_{\text{ce}} = \frac{1}{N}\sum_{i=1}^{N} \text{CE}\left(\text{Softmax}\left(\frac{z_i^{pred}\mathcal{T}\left(\mathbf{C}_{\text{train}}\right)}{\tau}\right), y_i^{cate}\right) \tag{2}$$

where τ is a learnable temperature parameter and CE is the cross-entropy loss for the classification of each training embedding.

The total loss for the training of our pipeline is finally defined as,

$$\mathcal{L} = \alpha\mathcal{L}_{\text{bce}} + \mathcal{L}_{\text{dice}} + \mathcal{L}_{\text{ce}} \tag{3}$$

where α is a hyper-parameter, we empirically set to 0.4.

In line with the work by (Cheng et al., 2022), we apply the Hungarian matching algorithm (Kuhn, 1955) to match predicted masks with ground-truth masks and compute the loss between the matching pairs.

4 Experiments

4.1 Datasets

Following previous studies (Ding et al., 2023; J. Xu et al., 2023; Zhao et al., 2023; Zheng et al., 2021), we used the instance segmentation part of the MS-COCO dataset (Lin et al., 2014) with 80 object categories to pre-train our model. We then fine-tuned the model on 3,040 images from the training set of the COD10K-v3 dataset (Fan et al., 2022). Pre-training the model on the MS-COCO dataset aims to learn general knowledge about objects in the wild, while fine-tuning the model on the COD10K-v3 dataset adapts the model to camouflaged objects. We empirically found that this strategy significantly boosts up the performance of our method.

We tested our method and others on two benchmark camouflaged object datasets: the test set of the COD10K-v3 (including 2,026 images) and the NC4K (Lyu et al., 2021) (including 4,121 images). The NC4K dataset contains only test images. The training sets (for both pre-training and fine-tuning) and the test sets (for both the COD10K-v3 and NC4K) share only six common object categories (out of 80 and 69 object categories from the MS-COCO and COD10K-v3/NC4K, respectively). This setting, i.e., cross-dataset training-testing, has been used widely in the evaluation of the generalisation ability of CIS models. It reflects the practicality of CIS, thus ensuring the reliability of evaluations.

We also evaluated our method on generic open-vocabulary datasets, including the ADE20K (B. Zhou et al., 2019) and Cityscapes (Cordts et al., 2016). For the ADE20K dataset, we used the validation set of the short version (B. Zhou et al., 2017) covering 150 object categories and 2,000 images. The Cityscapes dataset contains a total of 19 classes, which are divided into 11 "stuff" and 8 "thing" classes. We conducted evaluations on the validation set of the Cityscapes, including 500 images. Note that we pre-trained our method on the MS-COCO dataset and then directly evaluated the method on these open-vocabulary datasets without fine-tuning.

4.2 Implementation Details

We implemented our method in Pytorch and built it on the Detectron2 framework (Y. Wu et al., 2019). We trained our method for 90k iterations with a batch size of 64 on 4 NVIDIA A40 GPUs. All training images were resized to 512×512-pixels. Random jitters in the range $[0.1, 2.0]$ were applied to the training images. We froze both the SD (v1.3) and CLIP models during training. We adopted the Adam optimiser (Loshchilov & Hutter, 2019) with the learning rate γ set to 10^{-4} and weight decay of 0.05. We used a step learning rate scheduler and reduced the learning rate by a factor of 10 at 81k and 86k iterations.

The training took around 4.3 days to complete. Due to class imbalance in the COD10K-v3 dataset, we manually removed some extremely rare classes, *e.g.*, classes with less than five instances. In addition, we applied the `RepeatFactorTrainingSampler` from the Detectron2 framework, to allow a sample to appear more times than others based on its repeat factor.

4.3 Results

We evaluated our method and existing CIS methods using the average precision (AP) values measured at different intersection-over-union (IOU) thresholds. In particular, we calculated the overall

Table 2: Comparison of our method with existing instance segmentation methods on the test set of the COD10K-v3 and the NC4K datasets. Methods of the "closed-set supervised learning approach" are trained on the training set of the COD10K-v3 dataset. Methods of the "open-vocab text-to-image approach" are pre-trained on the MS-COCO dataset. We denote "Ours" and "Ours (task-specific)" for two variants of our method without and with fine-tuning on the training set of the COD10K-v3 dataset. "Params" denotes the number of *trainable* / total parameters. The best results are **bold**, and the second best results are underline.

	Method	COD10K-v3 Test			NC4K			Params (Millions)
		AP	AP50	AP75	AP	AP50	AP75	
closed-set supervised learning	Mask R-CNN (K. He et al., 2017)	25.0	55.5	20.4	27.7	58.6	22.7	43.9/43.9
	MS R-CNN (Z. Huang et al., 2019)	30.1	57.2	28.7	31	58.7	29.4	60.0/60.6
	Cascade R-CNN (Cai & Nuno, 2019)	25.3	56.1	21.3	29.5	60.8	24.8	71.7/71.7
	HTC (K. Chen et al., 2019)	28.1	56.3	25.1	29.8	59.0	26.6	76.9/76.9
	YOLACT (Bolya et al., 2019)	24.3	53.3	19.7	32.1	65.3	27.9	35.3/35.3
	BlendMask (H. Chen et al., 2020)	28.2	56.4	25.2	27.7	56.7	24.2	35.8/35.8
	SOLOv2 (X. Wang et al., 2020)	32.5	63.2	29.9	34.4	65.9	31.9	46.2/46.2
	Condlnst (Tian et al., 2020)	30.6	63.6	26.1	33.4	67.4	29.4	34.1/34.1
	Querylnst (Fang et al., 2021)	28.5	60.1	23.1	33.0	66.7	29.4	172.5/172.5
	SOTR (Guo et al., 2021)	27.9	58.7	24.1	29.3	61.0	25.6	63.1/63.1
	MaskFormer (Cheng et al., 2021)	38.2	65.1	37.9	44.6	71.9	45.8	45.0/45.0
	Mask2Former (Cheng et al., 2022)	39.4	67.7	38.5	45.8	73.6	47.5	43.9/43.9
	Mask Transfiner (Ke et al., 2022)	28.7	56.3	26.4	29.4	56.7	27.2	44.3/44.3
	OSFormer (Pei et al., 2022)	41.0	71.1	40.8	42.5	72.5	42.3	46.6/46.6
	DCNet (Luo et al., 2023)	**45.3**	70.7	**47.5**	52.8	**77.1**	**56.5**	53.4/53.4
	MSPNet (C. Li et al., 2024)	39.7	69.8	39.8	41.8	71.8	42.3	48.1/48.1
	UQFormer (Dong et al., 2024)	45.2	71.6	46.6	47.2	74.2	49.2	37.5/37.5
	CamoFA (M.-Q. Le et al., 2025)	43.5	**74.9**	42.7	45.0	75.7	44.3	-
	Ours (task-specific)	45.1	71.1	47.4	**52.9**	76.8	55.9	28.7/1522.7
open-vocab VLM (*w/o finetuning*)	MaskCLIP (Ding et al., 2023)	3.3	5.9	4.1	6.3	5.6	6.5	542.0/542.0
	MasQCLIP (X. Xu et al., 2023)	4.1	7.7	5.8	8.0	7.6	8.4	375.2/357.2
	X-Decoder (Zou, Dou, et al., 2023)	7.7	12.9	7.5	3.9	8.1	3.4	38.3/38.3
	SEEM (Zou, Yang, et al., 2023)	6.6	10.8	6.5	9.2	12.7	9.9	415.3/415.3
	OpenSeeD (H. Zhang et al., 2023)	6.1	10.4	5.9	9.3	14.5	9.8	116.2/116.2
	TPNet (Z. He et al., 2024)	18.3	41.8	14.3	21.4	**48.3**	16.6	71.78/71.78
open-vocab T2I (*w/o finetuning*)	ODISE (J. Xu et al., 2023)	21.1	37.8	20.5	22.9	37.2	21.4	28.1/1522.1
	Ours	**23.9**	**44.3**	**23.1**	**24.8**	44.2	**23.9**	28.7/1522.7

AP in the range $[50\%, 95\%]$ for the IOU thresholds (i.e., for a threshold within the above range, a predicted instance is considered as true positive if there exists a true instance in the ground-truth such that their IOU is equal or greater than that threshold). We also measured detailed AP for the IOU thresholds of 50% (AP50) and 75% (AP75).

4.3.1 Camouflaged Object Datasets

We report the performance of our method on camouflaged object datasets (COD10K-v3 and NC4K) in Table 2 (last row). Recall that, following the conventional setting in CIS, e.g., (Ding et al., 2023; J. Xu et al., 2023; Zhao et al., 2023; Zheng et al., 2021), we pre-trained our model on the MS-COCO dataset and then fine-tuned it on the training set of

the COD10K-v3 dataset. To show the effectiveness of this strategy, we experimented with a variant of our method by skipping the fine-tuning phase. In particular, we pre-trained our method on the MS-COCO dataset and then evaluated it directly on the test set of the COD10K-v3 and the NC4K datasets. We show the performance of this strategy in the last row, denoted as "Ours", in Table 2. Experimental results show that fine-tuning the method on a camouflaged object dataset, denoted as "Ours (task-specific)", significantly improves its performance on all evaluation metrics.

We compare our method with existing instance segmentation methods on the CIS task in Table 2. We group existing methods into two groups. We name the first group "closed-set supervised learning approach". The methods of this approach follow the traditional fashion, which supervises an instance segmentation model on a training set and tests the model on a test set. This approach's training and test sets are in the same domain and include imagery data only. Most existing instance segmentation methods in the field can be customised to enable CIS using this approach. In our experiments, the methods of the first group are trained on the training set of the COD10K-v3 dataset. The second group, called the "open-vocab approach," includes methods using the vision and language model (VLM) and text-to-image diffusion techniques with open-vocabulary.

As shown in Table 2, our method with full setting (pre-training and fine-tuning), denoted as "Ours (task-specific)", significantly outperforms ODISE on all evaluation metrics, making a new state-of-the-art for OVCIS. Our method also performs on par with recent methods (DCNet (Luo et al., 2023), MSPNet (C. Li et al., 2024), UQFormer (Dong et al., 2024), and CamoFA (M.-Q. Le et al., 2025)). Nevertheless, compared with recent methods, our method requires much fewer *trainable* parameters (see the last column in Table 2). Table 2 also compares all the methods in terms of the number of parameters used.

In summary, with regard to both segmentation accuracy and memory usage, our method is more advanced, compared with existing ones. Recall that only six object categories are shared between the MS-COCO dataset (with 80 object categories) and the COD10K-v3/NC4K dataset (with 69 object categories). This challenge shows the ability of our method in handling open-vocabulary tasks.

We visualise several results of our methods and existing ones in Figure 7. As shown, our method excels at pixel-level instance segmentation, accurately delineating camouflaged objects along their blurry boundaries in cluttered backgrounds. The results also demonstrate our proficiency in segmenting multiple instances.

In addition, Figure 8 illustrates failure cases of our method. We found that our method would be ineffective in distinguishing and separating an object that shares very similar characteristics with others or consists of fragmented parts. However, such circumstances would also be challenging for human beings as well.

4.3.2 Generic Open-Vocabulary Datasets

To showcase the versatility and generality of our method in various application domains (other than camouflaged objects), we evaluated our method on the ADE20K (B. Zhou et al., 2019) and Cityscapes datasets (Cordts et al., 2016), two widely used open-vocabulary benchmark datasets. Note that these datasets are not designed for camouflaged object detection and segmentation. We summarise the performance of our method and existing open-vocabulary instance segmentation methods on these two datasets in Table 3.

Our method ranks second on both the ADE20K and Cityscapes datasets. Nevertheless, compared with the first ranked method, i.e., OpenSeeD (H. Zhang et al., 2023), our method uses approximately four times fewer trainable parameters than OpenSeeD, while scarifying less than 1% and 8% of the overall AP on the ADE20K and Cityscapes datasets, respectively.

4.4 Ablation Studies

In this section, we present ablation studies to validate different aspects of the design and implementation of our method. First, we investigated the impact of prompt engineering and prompt templates on OVCIS tasks. Second, we validated the technical modules developed in our method to make it specialised to CIS tasks.

4.4.1 Prompt Engineering for OVCIS

For open-vocabulary-based studies, an object category can be specified by multiple alternative text

Table 3: Comparison of our method with existing open-vocabulary instance segmentation methods on the ADE20K and Cityscapes datasets. We measure the accuracy of the segmentation using the AP. "-" denotes no-report performance. The best results are **bold**, and the second best results are underline. In the last two columns, we also report the number of trainable and total parameters used in the methods.

Method	ADE20K	Cityscapes	Trainable Params (M)	Total Params (M)
MaskCLIP (Ding et al., 2023)	6.1	-	354.1 (100%)	354.1
ODISE (J. Xu et al., 2023)	13.9	-	28.1 (1.85%)	1522.1
X-Decoder (Zou, Dou, et al., 2023)	13.1	24.9	38.3 (100%)	38.3
OpenSeeD (H. Zhang et al., 2023)	**15.0**	**33.2**	116.2 (100%)	116.2
Ours	14.1	25.6	28.7 (1.88%)	1522.7

Table 4: Ablation study on applying **prompt engineering** to improve OVCIS. Results are tested on the COD10K-v3 dataset.

Prompt	AP	AP50	AP75
✗	22.8	43.1	22.1
✓	23.4 +0.6	43.8 +0.7	22.6 +0.5

descriptions. For instance, the "cat" category can be described as "cat", "cats", "kitty", or "kitties". To improve the diversity of open-vocabulary in text prompts, we applied the identical prompt engineering method introduced by (Ghiasi et al., 2022) to assemble a list of synonyms, subcategories, and plurals for the categories. Given a text prompt, the category is chosen as the one with the highest probability from an ensembling list of multiple alternative queries. We observed that the prompt engineering technique is simple yet effective in improving the segmentation accuracy of our method. Table 4 shows the impact of applying prompt engineering to CIS.

4.4.2 Prompt templates for OVCIS

Inspired from Pang et al. (2024), we apply the prompt template set, which considers task attributes and shows better performance in Table 5. We can see that using the prompt template can affect the influence of different templates on semantic embedding, which inspires further explorations for more effective prompt engineering.

4.4.3 CIS-Specialised Modules

We developed several technical modules in our method to make it specialised to CIS. We refer the reader to Figure 2 for a recall on how the modules are configured in our pipeline. To confirm the importance of those modules, we experimented with different variants of our method, each variant is made by alteration and/or omission of a module. We pre-trained the variants on the MS-COCO dataset for 30k iterations, then tested them on the test set of the COD10K-v3 dataset. We present the results of this ablation study in Table 6 and visualise the impact of the different modules in Figure 10.

We validated the importance of the use of text in our method (in the 1st row of Table 6). This was implemented by setting the text embeddings used in the method to zeros. We observed a significant drop in the performance of this variant, resulting in the lowest AP (12.2). This indicates that text embeddings play a crucial role as they provide essential contextual or semantic information that helps to identify camouflages.

We propose the MSFF module to fuse image-guided features learnt by the diffusion model at multiple scales. We validated the design of this module by comparing it with the standard fusion approach that concatenates the multiscale features from the encoder with the last layer of the decoder of the diffusion U-Net. The experimental results (in the 2nd and 3rd row of Table 6) show that the standard fusion approach incurs a performance loss. Moreover, compared with the full setting, which fuses all layers of both the encoder and decoder of the diffusion U-Net, the last layer of the diffusion U-Net appears to carry substantial information for the CIS task.

We develop the CIN module to further enhance the representations of camouflaged objects, such as prediction and classification. To validate the

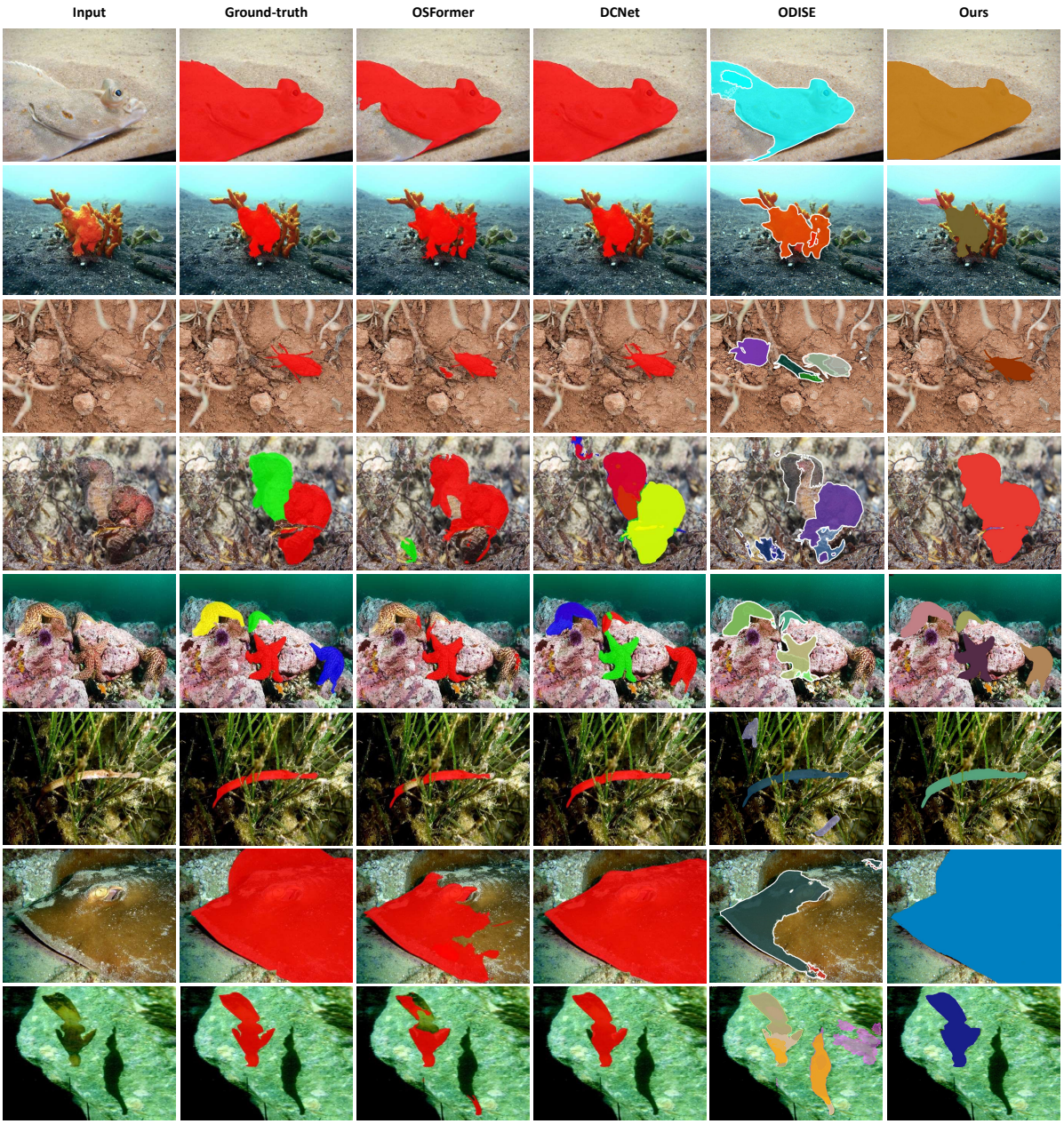

Fig. 7: Qualitative comparison of our method with existing methods on the COD10K-v3 and NC4K datasets. This figure is best viewed in colour.

CIN module, we removed it from our pipeline by directly passing the output from the TVA module to mask prediction and classification. We found that, by omitting the CIN module, the AP of the pipeline decreases dramatically (from 19.3 to 17.6), as shown in the 4th row of Table 6.

We devise the TVA module to aggregate textual and visual features in an instance-oriented manner, i.e., textual and visual features are aggregated alongside instance masks and consolidated against the background via feature weighting. To validate this module, we simplified its operation by applying an element-wise dot product on the input mask

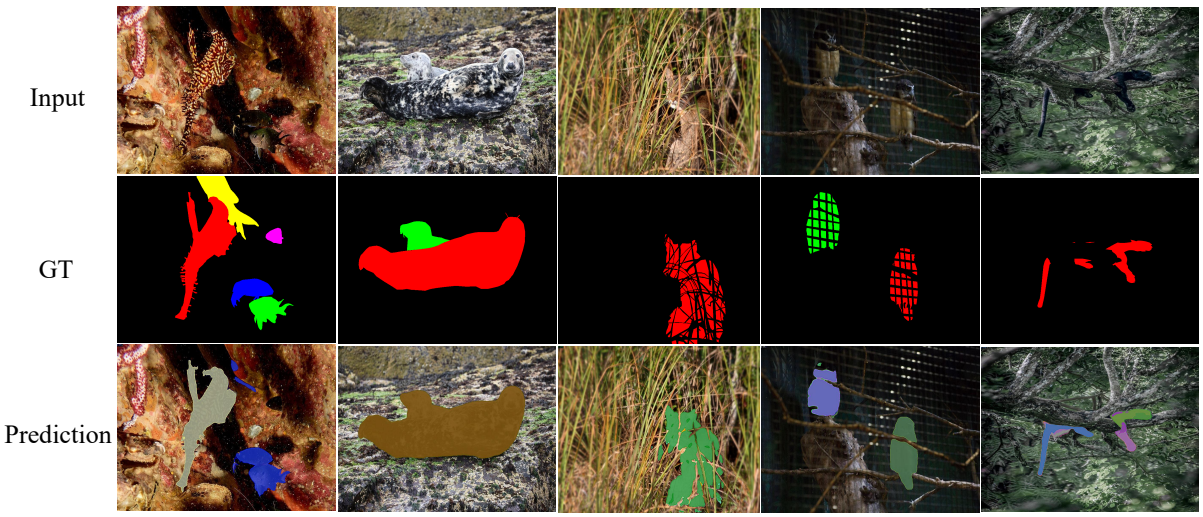

Fig. 8: Failure cases of our method on the COD10K-v3 dataset. In the first and second columns, our method fails to separate instances of nearby and similar objects, such as the yellow fish and two sea lions. Our method can detect and segment camouflaged objects in the third and fourth columns but with slightly less accurate boundaries. In the last column, our method struggles with the significant spatial separation of the black panther's body parts, leading to misclassification of the entire object. This figure is best viewed in colour.

Table 5: Ablation study on applying **prompt templates** to improve OVCIS. Results are tested on the COD10K-v3 dataset.

Task-related templates	**AP**
"A photo of <class>."	23.4
Using multiple templates: ☞ "A photo of the camouflaged <class>." ☞ "A photo of the concealed <class>." ☞ "A photo of the <class> camouflaged in the background." ☞ "A photo of the <class> concealed in the background." ☞ "A photo of the <class> camouflaged to blend in with its surroundings." ☞ "A photo of the <class> concealed to blend in with its surroundings."	23.9 +0.5

embeddings and text embeddings. We observed that, compared with other modules, the TVA module is less critical, which is evident by the low performance drop when simplification is applied to its architecture (see the 5th row of Table 6).

4.5 Additional Analysis

In our method, we utilised CLIP (Radford et al., 2021) (text and image encoders) to extract textual and visual features. We showcase CLIP's capability in Table 7, where we evaluate CLIP in performing zero-shot classification of camouflaged objects on different datasets (COD10K-v3, NC4K, CAMO (T.-N. Le et al., 2019)). Specifically, we applied the NLTK's WordNet to extract the animal type from each image's caption generated by ClipCap (Mokady et al., 2021) and check if the animal type and corresponding ground-truth category share the same hierarchical semantic relation (depth of the hypernym = 10) (Fu et al., 2014). In detail, the depth value helps in understanding the

Table 6: Ablation study on the effectiveness of the proposed technical modules to CIS. Results are tested on the COD10K-v3 dataset by using the AP metric.

Variant	Ours	Ours (task-specific)
no text (text embeddings = 0)	12.2 -7.1	31.4 -13.5
skip MSFF module (only the last layer of the diffusion U-Net is used)	18.4 -0.9	40.5 -4.4
skip MSFF module (concatenation of all multiscale features)	18.1 -1.2	39.8 -5.1
skip CIN module (directly use the TVA's output for instance classification)	17.6 -1.7	37.7 -7.2
skip TVA module (element-wise dot product of mask embedding and text embedding)	18.8 -0.5	42.7 -2.2
Full setting	**19.3**	**44.9**

Table 7: Zero-shot image classification using CLIP on camouflaged datasets.

	COD10K-v3	NC4K	CAMO
Accuracy (%)	48.06	45.69	46.48

```
Entity (Depth 0)
└─ Physical Entity (Depth 1)
    └─ Object (Depth 2)
        └─ Organism (Depth 3)
            └─ Animal (Depth 4)
                └─ Vertebrate (Depth 5)
                    └─ Fish (Depth 6)
                        └─ Ray-Finned Fish (Depth 7)
                            └─ Flatfish (Depth 8)
                                └─ Flounder (Depth 9)
                                    └─ Flat Flounder (Depth 10)
```

Fig. 9: Sample of hierarchical structure of "Summer Flounder" with the hypernym's depth = 10.

position and specificity of a concept within a hierarchical structure (the higher the depth, the more specific the concept). The example of the "Summer Flounder" is shown in Figure 9.

In addition, we conducted a **"prompt coarsening" ablation** by systematically replacing fine-grained category names with their WordNet hypernyms, same as in our CLIP analysis above (*e.g.*, cat → feline → mammal → animal) and reporting performance degradation. Because "animal" is semantically related but less discriminative, we expect decreased class separability among subclasses, which this stress test will quantify. To make this evaluation meaningful under hierarchical substitutions, we will report both standard AP (exact-label) and a semantics-aware metric (e.g., Open AP (H. Zhou et al., 2025)), which explicitly accounts for semantic similarity between predicted and ground-truth names as advocated in prior work on open-vocabulary evaluation. We reported the results in Table 8. As shown, the standard AP exhibits a significant decrease, whereas Open AP indicates only a slight decline. This is mainly due to the stability of localization in Open AP, while fine-grained naming necessitates specific prompts. This behavior aligns with the expected performance of a text-conditioned open-vocabulary system.

We also evaluated our work in the COD setting, where only binary masks are considered. Specifically, we experimented with our method on benchmark COD datasets, including CAMO, Chameleon, and COD10K-v2. We report the results of this experiment in Table 9. The results demonstrate the superiority of our method over existing COD baselines. In detail, while Camouflous (Khan et al., 2024) reports low MAE values on CAMO (0.043 MAE) and COD10K-v2 (0.021 MAE), our model achieves the highest F, S, and E on these datasets, including leading F scores on CAMO (0.847 F) and COD10K-v2 (0.807 F). Compared with the models C2F-Net (Y. Sun et al., 2021) and BCNet (Xiao et al., 2023), which emphasize global context fusion and boundary-aware refinement, our method strikes a balance between semantic precision and contextual depth, thereby producing more robust segmentation results. Notably, on Chameleon, our method slightly trails Camouflous in MAE (0.119 vs. 0.021) but still delivers the highest E (0.959), indicating stronger overall object integrity. These results validate the effectiveness of our task-specific design in COD.

Table 8: Results on OVCIS evaluated on "coarser prompts" by vanilla and Open AP on the test set of the COD10K-v3 and the NC4K datasets.

Coarser Prompts	Metric	COD10K-v3 Test			NC4K		
		AP	AP50	AP75	AP	AP50	AP75
✗	Vanilla AP	23.9	44.3	23.1	24.8	44.2	23.9
✓	Vanilla AP	22.7	42.1	21.9	23.6	41.9	22.7
✓	Open AP	23.3	43.2	22.5	24.2	43.1	23.3

Table 9: Comparison of our method with existing closed-set supervised learning camouflaged detection (binary segmentation) methods on the test set of the CAMO, Chameleon, and COD10K-v2 datasets. We adopt the results from (Jamali et al., 2025).

Method	CAMO				Chameleon				COD10K-v2			
	MAE ↓	F ↑	S ↑	E ↑	MAE ↓	F ↑	S ↑	E ↑	MAE ↓	F ↑	S ↑	E ↑
C2F-Net (Y. Sun et al., 2021)	0.091	0.647	0.796	0.828	0.034	0.782	0.828	0.950	0.043	0.629	0.775	0.872
DiCANet (Ike et al., 2024)	0.068	0.790	0.830	0.886	0.028	0.776	0.853	0.914	0.032	0.676	0.802	0.890
PCFNet (Z. Song et al., 2023)	0.053	0.840	0.844	0.913	0.023	0.876	0.912	0.957	0.027	0.751	0.838	0.924
BCNet (Xiao et al., 2023)	0.069	0.761	0.802	0.865	0.029	0.802	0.839	0.944	0.033	0.704	0.827	0.894
CamoMFCF (Wen et al., 2024)	0.080	0.727	0.796	0.854	0.032	0.805	0.838	0.935	0.036	0.686	0.813	0.890
Camouflous (Khan et al., 2024)	**0.043**	0.842	0.873	0.926	0.021	**0.866**	**0.902**	0.958	**0.021**	0.802	**0.873**	**0.935**
Ours (task-specific)	0.044	**0.847**	**0.878**	**0.931**	**0.119**	0.865	0.901	**0.959**	0.020	**0.807**	**0.873**	0.933

5 Conclusion

This work advances the computer vision research for open-vocabulary camouflaged instance segmentation (OVCIS) by leveraging text-to-image diffusion and text-image transfer techniques. To this end, we propose a method that effectively integrates textual information learnt from open-vocabulary into the visual domain to enrich the representations of camouflaged objects. We evaluate our method and compare it with existing methods in both CIS and generic open-vocabulary segmentation on benchmark datasets. Experimental results show the effectiveness and advantages of our method over existing baselines in both tasks.

Limitations and Future Works. Despite proven strengths, the proposed method has limitations. While the learnt knowledge from natural language can effectively distinguish an object from its background when visual cues are insufficient due to camouflages, it may not be helpful to separate touching/overlapping instances. Additionally, the method struggles with segmenting occluded objects. Under severe occlusions, a camouflaged object can be over-segmented into non-semantic fragments, leading to misclassification of the object. Enhancing object representations with background-aware features from open-vocabulary (i.e., by using text prompts including both foreground and background information, e.g., "a lizard is on a tree") may help to address the aforementioned issues. We consider this research direction to be our future work.

Broader Impact.

Our study directly contributes to advance research on wildlife monitoring, ecological interactions, and evolutionary understanding related to camouflage in nature (Beery et al., 2018; Norouzzadeh et al., 2018; Simões et al., 2023; Troscianko et al., 2017). To the best of our knowledge, our work is the first open-vocabulary approach to camouflaged instance segmentation, offering advanced features such as zero-shot performance ability

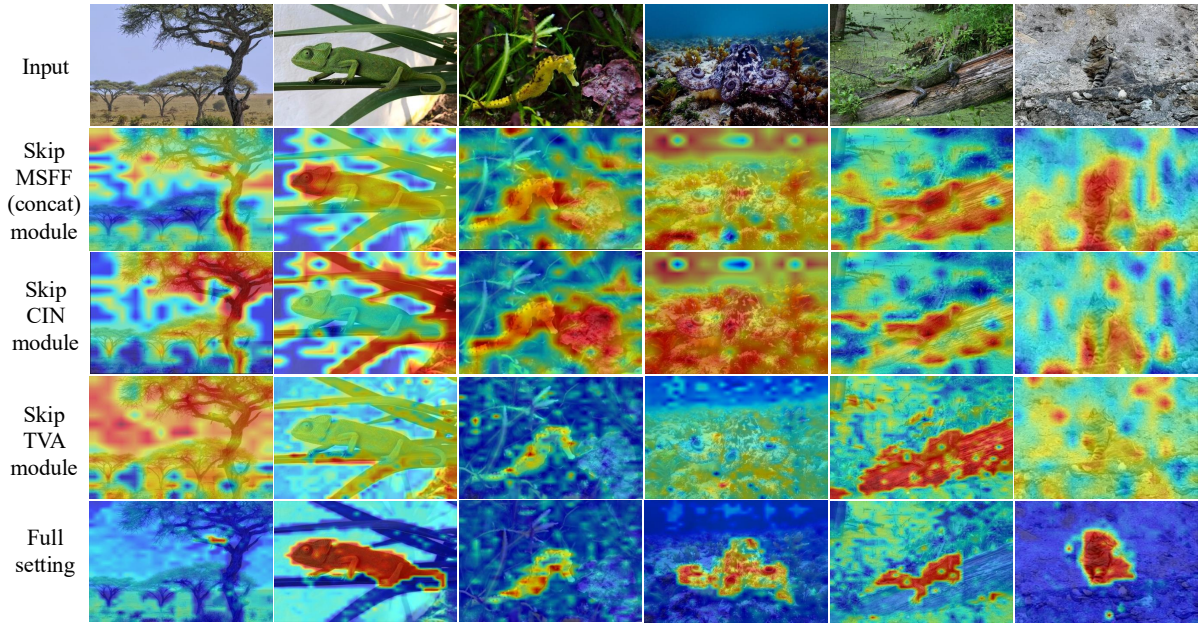

Fig. 10: **Qualitative intermediate outputs for module ablations.** The attention map (interim result) is a heat map of an object instance where foreground pixels are highlighted in red and background pixels are represented in blue. These intermediate outputs explain the quantitative gains in Table 6: the skip MSFF module (concat), the skip CIN module, the skip TVA module, and the full setting. This figure is best viewed in colour.

and multimodal enabling, improving the practicality of computer vision-based ecological studies. In addition, our work can significantly influence future developments in other fields, including, for instance, safety and security applications (e.g., military reconnaissance (Liu & Di, 2023)) and medical diagnostics (e.g., camouflaged colon polyp segmentation (H. Wang et al., 2024)).

Acknowledgement

This research is supported by an internal grant from HKUST (R9429), the National Research Foundation, Singapore, and DSO National Laboratories under the AI Singapore Programme (AISG Award No: AISG2-GC-2023-008), Career Development Fund (CDF) of Agency for Science, Technology and Research (A*STAR) (No.: C233312028), National Research Foundation, Singapore and Infocomm Media Development Authority under its Trust Tech Funding Initiative (No. DTC-RGC-04), a MAAP Discovery funding (2022-2025) from Deakin University and the Science Foundation Ireland under the SFI Frontiers for the Future Programme (22/FFP-P/11522). This work is partially done during Tuan-Anh Vu's research attachment at CFAR & IHPC, A*STAR, Singapore.

Availability of data and materials. All datasets (MS-COCO dataset (Lin et al., 2014), COD10K-v3 (Fan et al., 2022), NC4K (Lyu et al., 2021), CAMO (T.-N. Le et al., 2019), ADE20K (B. Zhou et al., 2019), and Cityscapes (Cordts et al., 2016)) used in our manuscript are available online on their websites. All related materials (models, codes, *etc.*) will be available online upon acceptance.

References

Baranchuk, D., Voynov, A., Rubachev, I., Khrulkov, V., Babenko, A. (2022). Label-efficient semantic segmentation with diffusion models. *Proceedings of the International Conference on Learning Representations.*

Beery, S., Van Horn, G., Perona, P. (2018).

Recognition in terra incognita. *Eccv* (pp. 456–473).

Bolya, D., Zhou, C., Xiao, F., Lee, Y.J. (2019). Yolact: Real-time instance segmentation. *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 9157–9166).

Cai, Z., & Nuno, V. (2019). Cascade r-cnn: High quality object detection and instance segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *43*(5), 1483–1498,

Chen, H., Sun, K., Tian, Z., Shen, C., Huang, Y., Yan, Y. (2020). Blendmask: Top-down meets bottom-up for instance segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 8573–8581).

Chen, K., Pang, J., Wang, J., Xiong, Y., Li, X., Sun, S., ... others (2019). Hybrid task cascade for instance segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 4974–4983).

Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R. (2022). Masked-attention mask transformer for universal image segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 1290–1299).

Cheng, B., Schwing, A., Kirillov, A. (2021). Per-pixel classification is not all you need for semantic segmentation. *Advances in Neural Information Processing Systems*, *34*, 17864–17875,

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., ... Schiele, B. (2016). The cityscapes dataset for semantic urban scene understanding. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 3213–3223).

Desai, K., & Johnson, J. (2021). VirTex: Learning visual representations from textual annotations. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 11162–11173).

Dhariwal, P., & Nichol, A.Q. (2021). Diffusion models beat gans on image synthesis. *Proceedings of the Advances in Neural Information Processing Systems* (pp. 8780–8794).

Ding, Z., Wang, J., Tu, Z. (2023). Open-vocabulary universal image segmentation with maskclip. *Proceedings of the International Conference on Machine Learning.*

Dong, B., Pei, J., Gao, R., Xiang, T.-Z., Wang, S., Xiong, H. (2024). A unified query-based paradigm for camouflaged instance segmentation. *Proceedings of the acm international conference on multimedia* (pp. 2131–2138).

Du, Y., Wei, F., Zhang, Z., Shi, M., Gao, Y., Li, G. (2022). Learning to prompt for open-vocabulary object detection with vision-language model. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 14064–14073).

Esser, P., Rombach, R., Ommer, B. (2021). Taming transformers for high-resolution image synthesis. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 12873–12883).

Fan, D.-P., Ji, G.-P., Cheng, M.-M., Shao, L. (2022). Concealed object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6024—6042,

Fan, D.-P., Ji, G.-P., Sun, G., Cheng, M.-M., Shen, J., Shao, L. (2020). Camouflaged object detection. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 2777–2787).

Fang, Y., Yang, S., Wang, X., Li, Y., Fang, C., Shan, Y., ... Liu, W. (2021). Instances as queries. *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 6910–6919).

Fleming, P.J.S., Meek, P.D., Ballard, G., Banks, P.B., Claridge, A.W., Sanderson, J.G., Swann, D.E. (2014). *Camera trapping: Wildlife management and research.* CSIRO Publishing.

Fu, R., Guo, J., Qin, B., Che, W., Wang, H., Liu, T. (2014). Learning semantic hierarchies via word embeddings. *Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 1199–1209).

Gal, R., Arar, M., Atzmon, Y., Bermano, A.H., Chechik, G., Cohen-Or, D. (2023). Encoder-based domain tuning for fast personalization of text-to-image models. *ACM Transactions on Graphics*, *42*(4), 1–13,

Gao, M., Xing, C., Niebles, J.C., Li, J., Xu, R., Liu, W., Xiong, C. (2022). Open vocabulary object detection with pseudo bounding-box labels. *Proceedings of the European Conference on Computer Vision* (pp. 266–282).

Ghiasi, G., Gu, X., Cui, Y., Lin, T. (2022). Scaling open-vocabulary image segmentation with image-level labels. *Proceedings of the European Conference on Computer Vision* (pp. 540–557).

Gu, X., Lin, T., Kuo, W., Cui, Y. (2022). Open-vocabulary object detection via vision and language knowledge distillation. *Proceedings of the International Conference on Learning Representations.*

Guo, R., Niu, D., Qu, L., Li, Z. (2021). Sotr: Segmenting objects with transformers. *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 7157–7166).

He, C., Li, K., Zhang, Y., Tang, L., Zhang, Y., Guo, Z., Li, X. (2023). Camouflaged object detection with feature decomposition and edge reconstruction. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 22046–22055).

He, K., Gkioxari, G., Dollár, P., Girshick, R.B. (2017). Mask R-CNN. *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 2980–2988).

He, Z., Xia, C., Qiao, S., Li, J. (2024). Text-prompt camouflaged instance segmentation with graduated camouflage learning. *Proceedings of the acm international conference on multimedia* (pp. 5584–5593).

Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-Or, D. (2023). Prompt-to-prompt image editing with cross attention control. *Proceedings of the International Conference on Learning Representations.*

Ho, J., Jain, A., Abbeel, P. (2020). Denoising diffusion probabilistic models. *Proceedings of the Advances in Neural Information Processing Systems* (pp. 6840–6851).

Huang, X., & Belongie, S.J. (2017). Arbitrary style transfer in real-time with adaptive instance normalization. *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 1510–1519).

Huang, Z., Huang, L., Gong, Y., Huang, C., Wang, X. (2019). Mask scoring r-cnn. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 6409–6418).

Ike, C.S., Muhammad, N., Bibi, N., Alhazmi, S., Eoghan, F. (2024). Discriminative context-aware network for camouflaged object detection. *Frontiers in Artificial Intelligence*, , https://doi.org/10.3389/frai.2024.1347898

Jamali, M., Davidsson, P., Khoshkangini, R., Ljungqvist, M.G., Mihailescu, R.-C. (2025). Context in object detection: a systematic literature review. *Artificial Intelligence Review*, ,

Jia, C., Yang, Y., Xia, Y., Chen, Y., Parekh, Z., Pham, H., ... Duerig, T. (2021). Scaling up visual and vision-language representation

learning with noisy text supervision. *Proceedings of the International Conference on Machine Learning* (pp. 4904–4916).

Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T. (2020). Analyzing and improving the image quality of stylegan. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 8107–8116).

Ke, L., Danelljan, M., Li, X., Tai, Y.-W., Tang, C.-K., Yu, F. (2022). Mask transfiner for high-quality instance segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 4412–4421).

Khan, A., Khan, M., Gueaieb, W., El Saddik, A., De Masi, G., Karray, F. (2024, January). Camofocus: Enhancing camouflage object detection with split-feature focal modulation and context refinement. *Proceedings of the ieee/cvf winter conference on applications of computer vision (WACV)* (pp. 1434–1443).

Kuhn, H.W. (1955, March). The Hungarian Method for the Assignment Problem. *Naval Research Logistics Quarterly*, *2*(1), 83–97,

Kumari, N., Zhang, B., Zhang, R., Shechtman, E., Zhu, J.-Y. (2023). Multi-concept customization of text-to-image diffusion. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 1931–1941).

Kuo, W., Cui, Y., Gu, X., Piergiovanni, A., Angelova, A. (2023). F-vlm: Open-vocabulary object detection upon frozen vision and language models. *Proceedings of the International Conference on Learning Representations.*

Le, M.-Q., Tran, M.-T., Le, T.-N., Nguyen, T.V., Do, T.-T. (2025). CamoFA: A Learnable Fourier-Based Augmentation for Camouflage Segmentation . *2025 ieee/cvf winter conference on applications of computer vision (wacv)* (pp. 3427–3436).

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

Li, C., Jiao, G., Yue, G., He, R., Huang, J. (2024). Multi-scale pooling learning for camouflaged instance segmentation. *Applied Intelligence*, *54*(5), 4062–4076,

Li, D., Ling, H., Kim, S.W., Kreis, K., Fidler, S., Torralba, A. (2022). Bigdatasetgan: Synthesizing imagenet with pixel-wise annotations. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 21298–21308).

Lin, T., Maire, M., Belongie, S.J., Hays, J., Perona, P., Ramanan, D., ... Zitnick, C.L. (2014). Microsoft COCO: common objects in context. *Proceedings of the European Conference on Computer Vision* (pp. 740–755).

Liu, M., & Di, X. (2023). Extraordinary MHNet: Military high-level camouflage object detection network and dataset. *Neurocomputing*, *549*, 126466,

Loshchilov, I., & Hutter, F. (2019). Decoupled weight decay regularization. *Proceedings of the International Conference on Learning Representations.*

Luo, N., Pan, Y., Sun, R., Zhang, T., Xiong, Z., Wu, F. (2023). Camouflaged instance segmentation via explicit de-camouflaging. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 17918–17927).

Lyu, Y., Zhang, J., Dai, Y., Li, A., Liu, B., Barnes, N., Fan, D.-P. (2021). Simultaneously localize, segment and rank the camouflaged objects. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 11586–11596).

Milletari, F., Navab, N., Ahmadi, S.-A. (2016). V-Net: Fully convolutional neural networks for volumetric medical image segmentation. *Proceedings of the International Conference on 3D Vision* (pp. 565–571).

Minderer, M., Gritsenko, A.A., Stone, A., Neumann, M., Weissenborn, D., Dosovitskiy, A., ... Houlsby, N. (2022). Simple open-vocabulary object detection with vision transformers. *Proceedings of the European Conference on Computer Vision* (pp. 728–755).

Mokady, R., Hertz, A., Bermano, A.H. (2021). Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*, ,

Nguyen, T.T.T., Eichholtzer, A.C., Driscoll, D.A., Semianiw, N.I., Corva, D.M., Kouzani, A.Z., ... Nguyen, D.T. (2023). Sawit: A small-sized animal wild image dataset with annotations. *Multimedia Tools and Applications*, 1–26,

Nichol, A.Q., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., ... Chen, M. (2022). GLIDE: towards photorealistic image generation and editing with text-guided diffusion models. *Proceedings of the International Conference on Machine Learning* (pp. 16784–16804).

Norouzzadeh, M.S., Nguyen, A., Kosmala, M., Swanson, A., Palmer, M.S., Packer, C., Clune, J. (2018). Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. *PNAS*, *115*(25), E5716–E5725,

Pang, Y., Zhao, X., Zuo, J., Zhang, L., Lu, H. (2024). Open-vocabulary camouflaged object segmentation. *Proceedings of the European Conference on Computer Vision (eccv).*

Parmar, G., Kumar Singh, K., Zhang, R., Li, Y., Lu, J., Zhu, J.-Y. (2023). Zero-shot image-to-image translation. *Proceedings of the ACM SIGGRAPH* (pp. 1–11).

Pei, J., Cheng, T., Fan, D.-P., Tang, H., Chen, C., Van Gool, L. (2022). Osformer: One-stage camouflaged instance segmentation with transformers. *Proceedings of the European Conference on Computer Vision* (pp. 19–37).

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... Sutskever, I. (2021). Learning transferable visual models from natural language supervision. *Proceedings of the International Conference on Machine Learning* (pp. 8748–8763).

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... Liu, P.J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, *21*, 1–67,

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M. (2022). Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1–27,

Rasheed, H.A., Maaz, M., Khattak, M.U., Khan, S.H., Khan, F.S. (2022). Bridging the gap between object and image-level representations for open-vocabulary detection. *Proceedings of the Advances in Neural Information Processing Systems* (pp. 33781–33794).

Ren, S., He, K., Girshick, R.B., Sun, J. (2015). Faster R-CNN: towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems* (pp. 91–99).

Rewatbowornwong, P., Tritrong, N., Suwajanakorn, S. (2023). Repurposing gans for one-shot semantic part segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(4), 5114–5125,

Robin, R., Andreas, B., Dominik, L., Patrick, E., Björn, O. (2022). High-resolution image synthesis with latent diffusion models. *Proceedings of the IEEE/CVF Conference on*

Computer Vision and Pattern Recognition (pp. 10674–10685).

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., ... others (2022). Photorealistic text-to-image diffusion models with deep language understanding. *Proceedings of the Advances in Neural Information Processing Systems* (pp. 36479–36494).

Sariyildiz, M.B., Perez, J., Larlus, D. (2020). Learning visual representations with caption annotations. *Proceedings of the European Conference on Computer Vision (eccv)* (pp. 1–17).

Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., ... others (2022). LAION-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 25278–25294,

Simões, F., Bouveyron, C., Precioso, F. (2023). Deepwild: Wildlife identification, localisation and estimation on camera trap videos using deep learning. *Ecological Informatics*, *75*, 102095,

Song, J., Meng, C., Ermon, S. (2021). Denoising diffusion implicit models. *Proceedings of the International Conference on Learning Representations.*

Song, Z., Kang, X., Wei, X., Li, S. (2023). Pixel-centric context perception network for camouflaged object detection. *IEEE Transactions on Neural Networks and Learning Systems*, ,

Sun, G., An, Z., Liu, Y., Liu, C., Sakaridis, C., Fan, D.-P., Van Gool, L. (2023). Indiscernible object counting in underwater scenes. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 13791–13801).

Sun, Y., Chen, G., Zhou, T., Zhang, Y., Liu, N. (2021). Context-aware cross-level fusion network for camouflaged object detection. *Ijcai* (pp. 1025–1031).

Tian, Z., Shen, C., Chen, H. (2020). Conditional convolutions for instance segmentation. *Proceedings of the European Conference on Computer Vision* (pp. 282–298).

Troscianko, J., Skelhorn, J., Stevens, M. (2017). Quantifying camouflage: how to predict detectability from appearance. *BMC Evolutionary Biology*, *17*, 1–13,

Wang, H., Hu, T., Zhang, Y., Zhang, H., Qi, Y., Wang, L., ... Du, M. (2024). Unveiling camouflaged and partially occluded colorectal polyps: Introducing CPSNet for accurate colon polyp segmentation. *Computers in Biology and Medicine*, *171*, 108186,

Wang, X., Zhang, R., Kong, T., Li, L., Shen, C. (2020). Solov2: Dynamic and fast instance segmentation. *Advances in Neural Information Processing Systems*, *33*, 17721–17732,

Wen, Y., Ke, W., Sheng, H. (2024). Camouflaged object detection based on deep learning with attention-guided edge detection and multi-scale context fusion. *Applied Sciences*, , https://doi.org/10.3390/app14062494

Wu, J., Li, X., Xu, S., Yuan, H., Ding, H., Yang, Y., ... Tao, D. (2024, jul). Towards open vocabulary learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *46*(07), 5092–5113,

Wu, Y., Kirillov, A., Massa, F., Lo, W.-Y., Girshick, R. (2019). *Detectron2*. https://github.com/facebookresearch/detectron2.

Xiao, J., Chen, T., Hu, X., Zhang, G., Wang, S. (2023). Boundary-guided context-aware network for camouflaged object detection. *Neural Computing and Applications*, ,

Xie, E., Wang, W., Wang, W., Sun, P., Xu, H., Liang, D., Luo, P. (2021). Segmenting transparent objects in the wild with transformer. *Proceedings of the International Joint Conferences on Artificial Intelligence* (pp. 1194–1200).

Xu, J., Liu, S., Vahdat, A., Byeon, W., Wang, X., Mello, S.D. (2023). Open-vocabulary panoptic segmentation with text-to-image diffusion models. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 2955–2966).

Xu, X., Xiong, T., Ding, Z., Tu, Z. (2023, October). Masqclip for open-vocabulary universal image segmentation. *Proceedings of the ieee/cvf international conference on computer vision (iccv)* (pp. 887–898).

Yan, J., Le, T., Nguyen, K., Tran, M., Do, T., Nguyen, T.V. (2021). Mirrornet: Bio-inspired camouflaged object segmentation. *IEEE Access*, *9*, 43290–43300,

Zang, Y., Li, W., Zhou, K., Huang, C., Loy, C.C. (2022). Open-vocabulary detr with conditional matching. *Proceedings of the European Conference on Computer Vision* (pp. 106–122).

Zareian, A., Rosa, K.D., Hu, D.H., Chang, S. (2021). Open-vocabulary object detection using captions. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 14393–14402).

Zhang, H., Li, F., Zou, X., Liu, S., Li, C., Yang, J., Zhang, L. (2023, October). A simple framework for open-vocabulary segmentation and detection. *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 1020–1031).

Zhang, J., Huang, J., Jin, S., Lu, S. (2023). Vision-language models for vision tasks: A survey. *arXiv preprint arXiv:2304.00685*, 1–23,

Zhang, Y., Jiang, H., Miura, Y., Manning, C.D., Langlotz, C.P. (2022). Contrastive learning of medical visual representations from paired images and text. *Proceedings of Machine Learning Research*, *182*, 1–24,

Zhao, W., Rao, Y., Liu, Z., Liu, B., Zhou, J., Lu, J. (2023). Unleashing text-to-image diffusion models for visual perception. *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 5729–5739).

Zheng, Y., Wu, J., Qin, Y., Zhang, F., Cui, L. (2021). Zero-shot instance segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 2593–2602).

Zhong, Y., Yang, J., Zhang, P., Li, C., Codella, N., Li, L.H., ... Gao, J. (2022). Regionclip: Region-based language-image pretraining. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 16772–16782).

Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A. (2017). Scene parsing through ade20k dataset. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 5122–5130).

Zhou, B., Zhao, H., Puig, X., Xiao, T., Fidler, S., Barriuso, A., Torralba, A. (2019). Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, *127*(3), 302–321,

Zhou, H., Qi, L., Shen, T., Huang, H., Yang, X., Li, X., Yang, M.-H. (2025). Rethinking evaluation metrics of open-vocabulary segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, ,

Zhou, X., Girdhar, R., Joulin, A., Krähenbühl, P., Misra, I. (2022). Detecting twenty-thousand classes using image-level supervision. *Proceedings of the European Conference on Computer Vision* (pp. 350–368).

Zou, X., Dou, Z.-Y., Yang, J., Gan, Z., Li, L., Li, C., ... Gao, J. (2023, June). Generalized decoding for pixel, image, and language. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 15116–15127).

Zou, X., Yang, J., Zhang, H., Li, F., Li, L., Wang, J., ... Lee, Y.J. (2023). Segment everything everywhere all at once. *Thirty-seventh conference on neural information processing systems.*