Pattern Recognition 168 (2025) 111826

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

COD-SAM: Camouflage object detection using SAM☆

Dongyang Gao [a], Yichao Zhou [a], Hui Yan [a], Chen Chen [b], Xiyuan Hu [c,*]

[a] School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, 210014, China
[b] State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, 100038, China
[c] College of Computer Science, Beijing University of Technology, Beijing, 100124, China

ARTICLE INFO

Keywords:
SAM
Transformer
COD
Feature cross fusion

ABSTRACT

The recently introduced Segment Anything Model(SAM) signifies a major advancement in segmentation, offering robust zero-shot generalization and interactive prompts that herald a new paradigm for segmentation models. Despite training on 1.1 billion masks, SAM's predictive performance is limited in tasks with complex structures or backgrounds. We introduce COD-SAM, which retains the original prompt design and zeroshot generalization capabilities of SAM, while incorporating the minimal number of additional trainable parameters to enhance its performance in downstream tasks such as camouflaged object detection. First, we freeze SAM's encoder and introduce a integrated learnable module, named "Fuse Anything as Prompts"(FAPs) comprising: (1) COD-Adapter, a global gradient weak prompt embedding to enhance adaptability for downstream tasks; (2) Corner Prompts, a strong prompt embedding for edge corners. For SAM's decoder, we introduce COD-Head, a learnable feature refinement module aimed at enhancing the quality of predicted masks. Subsequently, we introduce a task-adaptive loss function named sIoU loss. Comprehensive experiments not only verify the effectiveness of our method's in camouflage object detection but also assess its applicability to other downstream tasks. With less than a 0.1% increase in parameters, our approach achieves a notable 5.6% improvement in performance, demonstrating its superior efficacy.

1. Introduction

Precise segmentation of diverse objects constitutes a fundamental task in computer vision. These segmentation tasks find applications across numerous domains [1,2], such as image/video editing, robotic manipulation, and medical image analysis. The recently introduced Segment Anything Model (SAM) [3] serves as a foundational visual model in the field of segmentation. It has been trained on tens of millions of images, annotated with 1.1 billion masks. SAM exhibits robust zero-shot segmentation capabilities, enabling segmentation of diverse objects via input prompts such as points, bounding boxes, and rough masks [4]. The capacity to segment anything, combined with the innovative use of prompts, establishes a new paradigm for segmentation models. Collectively, these attributes underscore SAM's significant application potential.

Although SAM has demonstrated remarkable performance, its segmentation outcomes remain unsatisfactory, particularly in specific downstream tasks [5–7]. Camouflage object detection (COD) [8], an exceptionally challenging task, seeks to identify and locate objects in images that closely blend with their backgrounds. This task is characterized by backgrounds that seamlessly merge with the subject, variations in scale and perspective, and the necessity for dynamic adaptation to changing environments. Collectively, these factors demand high robustness and recognition capabilities from detection models. Successful algorithms must discern subtle feature differences, accurately identify objects of various sizes and shapes against complex backgrounds, which may include environmental colors and textures, and adapt to dynamic conditions such as shifting lighting and weather [9].

While SAM demonstrates impressive capabilities in general-purpose segmentation, it exhibits notable limitations in specialized low-level segmentation tasks, particularly in camouflaged object detection and shadow detection, when not properly adapted or provided with task-specific prompts. The model often struggles to precisely delineate fine object boundaries and may generate inaccurate or irrelevant segmentation outputs. These limitations can be attributed primarily to SAM's insufficient incorporation of task-specific prior knowledge, including but not limited to texture patterns, frequency domain information, and illumination-related cues. Such deficiencies substantially constrain the model's ability to generalize effectively across these specialized segmentation scenarios.

☆ This work was supported by the National Natural Science Foundation of China (62172227) and National Key R&D Program of China (2021YFF0602101).
* Corresponding author.
E-mail address: huxiyuan@bjut.ude.cn (X. Hu).

https://doi.org/10.1016/j.patcog.2025.111826
Received 15 November 2024; Received in revised form 7 April 2025; Accepted 6 May 2025
Available online 22 May 2025

In the advancement of deep learning, feature fusion is pivotal for tailoring models to specific downstream tasks [10–14]. In CNN-based architectures, prevalent feature fusion methods include the Feature Pyramid Network (FPN) and residual connections. Conversely, Transformer-based models primarily utilize attention mechanisms for this purpose [15–17]. These techniques significantly enhance the models' expressive capabilities and accuracy by integrating features across various scales and dynamically adjusting the weights of fused features. However, they often introduce complex structures that increase the models' complexity and computational demands. Despite these challenges, feature fusion remains an exceptionally effective strategy, requiring meticulous design optimization.

In both foundational models and their variants [3,18–21], researchers have developed advanced methods that primarily focus on adaptation. For example, extensive experiments across various domains and different backbone architectures, as detailed in [22], have explored the universality and feasibility of visual prompts for diverse recognition tasks. Furthermore, as discussed in [23], visual adapters significantly enhance fine-tuning performance by re-modulating high-frequency information extracted from images, thereby improving visual prompts. However, this high-frequency information is susceptible to noise, which can degrade performance in downstream tasks.

After conducting a comprehensive review of current research, we identified two critical limitations in SAM: first, its limited adaptability to downstream tasks, which hampers the understanding of crucial task-specific information; second, the frequent under utilization or omission of details, often due to imprecise mask boundaries, which overlooks small object structures.

Building on these challenges, we integrated the "fusion" concept from traditional deep learning models with the "adapt" principle from foundational model variants to develop a novel prompt learning approach and a corresponding loss function. This method consists of two main components: local strong prompts and global gradient weak prompts. Local strong prompts underscore learnable features of specific targets, emphasizing subtle yet highly discriminative attributes. In contrast, global gradient weak prompts focus on broad characteristics relevant to the task at hand, such as essential foreground-background dynamics in camouflage object detection (COD). The loss function's design philosophy mirrors that of the global gradient weak prompts, aiming for a comprehensive understanding of task-specific characteristics. Our methodology begins by enhancing the foundational model's effectiveness and robustness in downstream tasks, specifically within the challenging domain of COD, thus preserving the model's inherent strengths in zero-shot segmentation and adaptability. To maintain efficiency and zero-shot performance, we minimized the increase in model complexity, adding less than 0.1% in parameters to improve segmentation capabilities.

Consequently, we introduce the COD-SAM architecture, which integrates seamlessly with the established SAM structure to uphold exemplary zero-shot performance. This study's contributions are summarized as follows:

- We designed a flexible and highly integrated FAPs module that incorporates various sophisticated prompting methods, enabling comprehensive utilization of diverse feature information.
- We conducted a comprehensive analysis of the feature distribution within the dataset and developed a task-adaptive sIoU loss function, which enhances the model's focus on the pertinent task.
- Given its exceptional performance in camouflage detection, we extended its application to shadow detection, achieving favorable outcomes. This underscores COD-SAM's versatility in managing diverse downstream tasks.

2. Related work

2.1. Semantic segmentation

The Segment Anything Model (SAM), introduced by Kirillov et al. (2023), has significantly advanced the field of segmentation and contributed substantially to the development of foundational models in computer vision. SAM effectively integrates prompt learning techniques from natural language processing into its architecture. By employing an image engine with interactive annotations, SAM efficiently performs instance analysis, edge detection, object proposals, text-to-mask conversion, and various other techniques [24].

SAM is specifically engineered to tackle the challenges associated with segmenting diverse objects within complex visual environments. Unlike traditional methods that concentrate on segmenting specific object categories, SAM endeavors to segment any object, offering a universal solution applicable across a broad spectrum of challenging contexts. Numerous studies currently employ SAM (He et al. 2023 [25]; Zhang et al. 2023 [18]) as a foundational tool for subsequent visual tasks, with applications spanning medical imaging, video analysis, and data annotation (Brown et al. 2023 [5]).

2.2. Prompts

The concept of visual prompt fine-tuning, initially proposed in the field of natural language processing (NLP) [5], was exemplified by GPT-3 [26]. GPT-3 demonstrated robust generalization capabilities in downstream transfer learning tasks, including zero-shot and few-shot settings, through the use of manually selected prompts [27]. This concept was subsequently adapted for visual tasks, introducing memory tokens [28], learnable embedding vectors assigned to each transformer layer. Visual Prompt Tuning (VPT) [22] further explored the universality and feasibility of visual prompts across multiple domains and backbone architectures, encompassing a range of recognition tasks through extensive experimentation.

SAM pioneered the integration of dense prompts into the Transformer architecture, enhancing the model's capability to interpret and utilize diverse prompt information effectively. This innovation improved object localization and recognition in image segmentation tasks significantly. By converting prompt information into high-dimensional representations, the model could more efficiently exploit this data to improve segmentation results. This approach has produced exceptional outcomes in dense prediction tasks, prompting the incorporation of this embedding technique in various SAM variants and subsequent studies that build on SAM's enhancements.[1]

In contrast, our research aims to identify the optimal visual content for low-level structural segmentation, deviating from the primary focus on recognition tasks in the previously mentioned methods.

2.3. Camouflage object detection

Detecting disguised objects poses significant challenges as foreground objects often exhibit visual patterns that closely mimic those of the background [29]. Initial studies differentiated foreground from background using basic promtps such as texture, brightness, and color [30]. Recently, deep learning techniques have demonstrated substantial effectiveness in recognizing complex disguised objects [31]. Le et al. [32] proposed a pioneering end-to-end network for disguised object detection, incorporating classification and segmentation branches. Li et al. [33] introduced an innovative search-recognize network and compiled the largest dataset for disguised object detection to date. PFNet [34] employs a biomimetic strategy that mimics the locational and recognition processes found in predatory behaviors. FBNet [35] presents a method to separate frequency modeling and enhance critical frequency components.

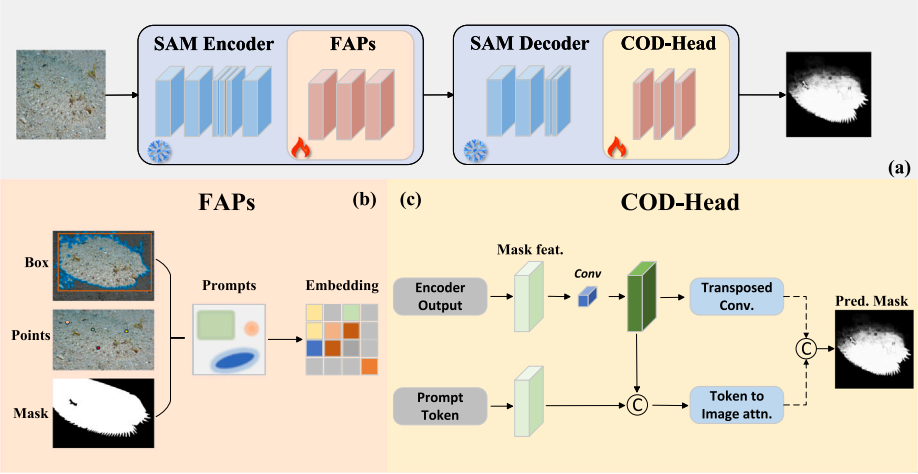

Fig. 1. (a)The Overview of the proposed COD-SAM architecture. (b)The **F**use **A**nythin as **P**rompts module (FAPs). (c)The COD-Head module in SAM Decoder. We utilized a frozen SAM encoder–decoder architecture, effectively integrating various forms of prompts including points, boxes, and masks.

3. The proposed approach

We introduce COD-SAM, illustrated in Fig. 1, as an enhancement to the SAM model, specifically designed for high-quality segmentation tasks. COD-SAM features a lightweight design and achieves significant improvements in two critical aspects of the SAM model. Section 3.1 succinctly reviews the architecture of the foundational SAM model, the basis for COD-SAM. Section 3.2 describes the innovative prompt methods implemented in the encoder, segmented into two parts. Section 3.3 explains the refinement module, named COD-Head, which is integrated into the decoder to extract local features from the original image, thus improving the mask decoder's prediction accuracy. Section 3.4 presents the task-adaptive loss, named sIoU loss, which increases the robustness of the model's adaptation to downstream tasks.

3.1. Use SAM as backbone

SAM was trained on the expansive SA-1B dataset, comprising over 1 billion automatically generated masks and 11 million images. This dataset's size significantly exceeds that of other segmentation datasets. SAM demonstrates robust zero-shot generalization capabilities, obviating the need for additional training.

SAM comprises three primary modules: (a) Image Encoder, constructed using the ViT backbone [36], this module extracts image features and generates spatial embeddings of size 64 × 64. (b) Prompt Encoder, this module encodes interactive positional information derived from input points, boxes, or masks, supplying input to the mask decoder. (c) Mask Decoder, a two-layer Transformer-based decoder that integrates extracted image embeddings and prompt tokens to predict the final mask.

3.2. Fuse anything as prompts

The robustness of a camouflage object detection model hinges on the algorithm's capacity to accurately discern subtle feature distinctions and reliably identify targets within complex backgrounds. This process unfolds in two stages: initially, the model comprehends the task's characteristics and integrates low-level semantic information comprehensively; subsequently, it concentrates on detecting subtle feature differences and effectively utilizes high-level semantic information.

We introduce a highly adaptable and cohesive prompt module, primarily consisting of the COD-Adapter as the global gradient-weak prompt module and the Corner Prompts module for localized, strong prompts at edge corners.

(1) COD-Adapter: Global Gradient Weak Prompts. The COD-Adapter, introduced in this section as a method employing globally weak gradient prompts, is designed to understand the task characteristics pertinent to the initial-stage mission.

In the domain of computer vision tasks, high-frequency information is defined as segments of an image characterized by rapid changes [37,38]. The strategic use of high-frequency information has demonstrated significant effectiveness in specific explicit visual tasks. However, its efficacy in more nuanced tasks, such as detecting camouflaged objects, is still suboptimal.

Furthermore, it has been observed that foundational models such as SAM and its variants often fail to distinguish between foreground and background elements. Specifically, in downstream tasks like camouflage detection, these models tend to overlook the transitions between these elements. This limitation seems inherent to Transformer models [39–41], which, despite their strong global perception capabilities, may not effectively utilize local information compared to CNN-based approaches [42,43].

To facilitate lightweight processing, the design of our COD-Adapter, inspired by [44], incorporates task-specific insights. The architecture of the COD-Adapter consists of two main components: the image embedding module and the gradient information fusion module, as illustrated in Fig. 2(left).

$$Adapter = MLP_{up}(G(MLP^{i}_{tune}((F_{pe} + F_{edge})))), \quad (1)$$

G means GELU [45], the linear layer MLP^{i}_{tune} operates within COD-Adapter, and MLP_{up} acts as a shared layer between all COD-Adapter, consistent with each layer of the SAM transformer in dimension. F_{pe} and F_{edge} are patch embeddings.

The image embedding module aims to refine the pre-trained patch embeddings. In the pre-trained SAM model, patches are mapped onto corresponding dimensional features. We freeze this projection and introduce an adjustable linear layer to map the original embedding onto dimensional features, denoted as F_{pe}.

The gradient information fusion module, denoted as F_{edge}, is specifically designed to obtain edge information of features in order to fuse low-level semantic information, which are paramount in the context of camouflage object detection. Unlike typical regions of interest (ROIs), edge information is deemed more vital due to its propensity for identifying key points and its rich geometric structural data. The integration of the Adapter module, coupled with the use of visual prompts, enhances the utility of this information. Initially, a flexible gradient detection algorithm(specifically the Sobel edge detection operator), is applied to the input image x, resulting in an edge-detected image x'.

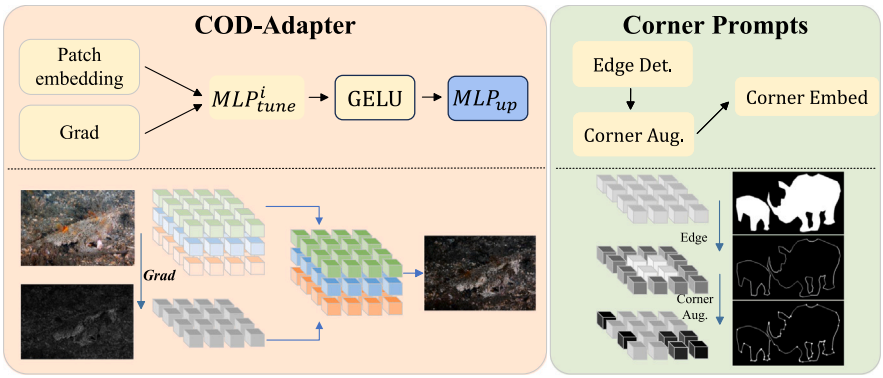

Fig. 2. The architecture of Fuse Anything as Prompts module (FAPs). FAPs mainly consists of two parts: COD-Adapter, global gradient weak prompts; Corner Prompts, local strong prompts.

Given that pure edge detection is susceptible to noise, occlusion, and edge blurring, an element-wise multiplication is performed between the edge-detected image x' and the original image x to produce an edge-enhanced feature map.

Although this direct multiplication may reduce gradient information, it simultaneously enriches the feature map with the global context of the original image. The adapter's design facilitates the processing of feature pairs with diminished gradient information (F_{edge} and its embeddings) through a linear layer, MLP^i_{tune}, specifically tuned for each adapter, and then through the GELU activation function. Concurrently, MLP_{up} acts as an upper projection layer, modifying the feature dimensions within the transformer and enabling feature transfer across transformer layers in the encoder. This process effectively introduces global weak gradient prompts.

(2) Corner Prompts: Local Strong Prompts. In this section, we introduce the Corner Prompts, a local strong prompts designed for the second-stage task, aimed at accentuating the learnable features of specific objectives, particularly those subtle yet highly discriminative characteristics, as depicted in Fig. 2(right).

Existing literature predominantly assumes the utilization of dense prompt embedding in SAM [46,47]. Our findings suggest that dense prompts largely depend on the reuse of masks, incorporating conventional techniques such as down-sampling, feature extraction, normalization, and activation. Although these processes are crucial for facilitating mask-guided attention or feature fusion in semantic segmentation models, they overlook the potential information loss, especially in scenarios involving small targets or detailed textures. Moreover, dense embedding of prompts are associated with additional computational costs.

To mitigate this issue, we abandon traditional prompt embedding methods in favor of a novel approach named "corner embedding". Initially, we processed the original masked image similarly to the dense prompt, but with the addition of edge detection to mask background elements, thereby generating an edge map of the mask. This technique allows the model to focus on easily overlooked or less conspicuous foreground details. Subsequently, we applied corner detection algorithms to highlight areas of significant feature changes on the edge map, thereby extracting more advanced semantic information from the foreground. This refinement enables the model to concentrate on the more subtle details in the foreground. The method was then integrated with image embeddings of the same dimension, as illustrated in Fig. 2(right).

By converting masks into corner representations, our approach more effectively preserves the spatial information of target areas in the foreground, eliminating the need to process background information. This enables the model to interpret and manage intricate image details more accurately. In complex scenarios, such as camouflage detection, corner representations clearly delineate areas of interest, enhancing the model's attention mechanism and significantly reducing the information loss associated with traditional down-sampling. Additionally, corner embedding provides greater flexibility and scalability, for example, by allowing adjustments to embedding points or the incorporation of additional features like color or texture.

3.3. COD-head: Feature refinement

We developed an efficient and learnable module, termed the COD-Head, to improve mask quality. COD-Head acts on the decoder before the Transformer prediction layer. As depicted in Fig. 1, the original mask decoder in SAM utilizes an output token, akin to the object queries in DETR [48], for mask prediction. This token generates dynamic MLP weights, which are then applied element-wise to the mask features. To augment mask quality in COD-SAM without significant alterations, as seen in HQ-SAM, we adhered to a lightweight design philosophy by integrating COD-Head. This strategy diverges from SAM by eschewing the direct input of coarse masks.

In the mask decoder, feeding the original image x directly into the Transformer's prediction layer for mask generation, though simple, neglects the local features of x. The COD-Head, serving as a learnable component, primarily comprises a 3×3 convolutional layer followed by a ReLU activation function. This convolutional operation enhances local feature detection, thereby capturing the image's local characteristics. Moreover, COD-Head's adaptability to various tasks and datasets is facilitated by modifying the convolution kernel's size, stride, or padding, and incorporating additional convolutional layers. Considering the prevalence of camouflage objects in natural settings and the dataset's inclusion of complex foreground-background interactions, COD-Head marginally improves the model's generalization ability.

3.4. sIoU: Task-adaptive loss

The IoU [49] loss is commonly used in segmentation tasks to quantify the extent of overlap between segmentation masks and ground truth. Although IoU offers a reliable metric for segmenting objects of various sizes, it can engender a "hole" issue by exclusively emphasizing the overall object integrity, thereby potentially compromising segmentation quality. This challenge becomes particularly evident in abnormal tasks. Furthermore, in instances where there is no overlap, IoU reduces to zero, potentially resulting in gradient vanishing and impeding the learning process for the model. We posit that an effective approach to leveraging SAM for downstream tasks involves enabling the model to prioritize the processing of features possessing task-specific characteristics.

We re-conceptualize the loss first by considering data statistical analysis, with the objective of discerning a more representative distribution of data statistics. This ensures that the loss directs greater attention to this concern during the handling of downstream tasks. An analysis of the aspect ratio within the COCO [50] and COD10K dataset

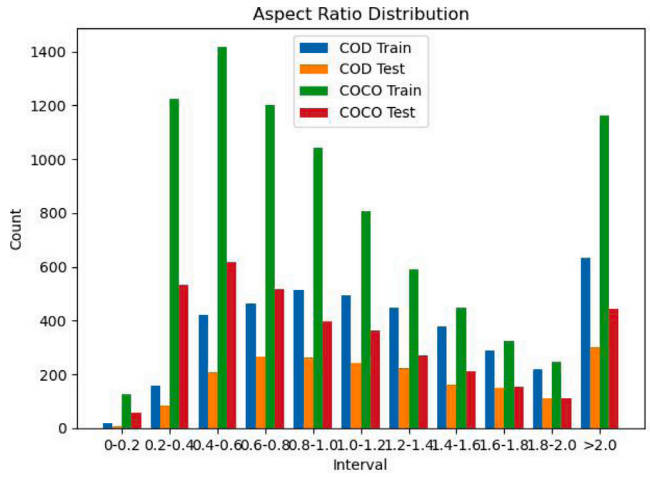

Fig. 3. Statistical analysis of sIoU. We use 1% COCO training set and 10% COCO test set.

reveals that the most concentrated range is (0.6–1.0), as illustrated in Fig. 3. The inclusion of parameters such as aspect ratio, center distance, overlapping area, and diagonal distance renders the work by [49] comparatively advanced in the field of loss research. Upon reference and examination of [49], we identified specific design flaws, particularly an undue emphasis on regressing the penalty term to 0. Our data analysis indicates that this regression process poses significant challenges. When the x-value approaches 0, it implies that the target's width-height ratio is 0. However, such occurrences are exceedingly rare or non-existent, rendering the enforcement of loss regression to 0 unreasonable. Notably, within the most concentrated aspect ratio interval, the convergence effect of the sigmoid function outperforms other activation functions. This observation aligns cohesively with the overarching design philosophy of downstream tasks. Consequently, we advocate for an enhanced loss design, denoted as sIoU,

$$L_{sIoU} = 1 - IoU + \frac{\rho^2(b, b^{gt})}{c^2} + \frac{v}{(1 - IoU) + v} \cdot v, \tag{2}$$

$$v = \frac{4}{\pi^2}\left[\left(1 + e^{-w^{gt}/h^{gt}}\right)^{-1} - \left(1 + e^{-w/h}\right)^{-1}\right]^2, \tag{3}$$

where gt is ground truth, IoU denotes the ratio of the overlapping area between the prediction and gt to the sum of areas, $\rho^2(b, b^{gt})$ is the square of the distance between the centers of the prediction and gt, c is the square of the diagonal distance between the prediction and gt, v is a parameter describing the consistency of the aspect ratio.

4. Experiments

SAM incurs a notably high training cost, employing a batch size of 256 images. Distributed training of SAM, based on ViT-H, for 2 epochs on SA-1B necessitates 256 GPUs. However, we have addressed this issue by freezing SAM's encoder, thereby reducing the training cost and accelerating the training process.

4.1. Metric and dataset

The primary evaluation metrics used in this paper include F_β^ω (weighted F_β-measure), MAE (Mean Absolute Error), E-measure (Enhanced Alignment Measure), and S-measure (Structure measure).

The source training set, SA-1B, was mainly constructed by collecting from natural environments. In this work, we identified two types of downstream segmentation tasks, some of which feature a drastic distribution shift from SA-1B. In the experiments, we utilize the COD10K, CAMO, CHAMELEON and ISTD [51] datasets. The COD10K dataset consists of 3040 training images and 2026 test images. The CAMO dataset comprises 1000 training images and 250 test images. Following the training protocol outlined in [52], we trained the model using the combined COD10K+CAMO dataset and tested it on their respective test sets. We utilized the ISTD dataset for shadow detection, which comprises three sub-datasets: Train-A, Train-B, Train-C. In our experiment, only the Train-A, Train-B and their corresponding test sets were used.

4.2. Implementation details

In our experiments, we employed SAM's ViT-H model, utilized the AdamW optimizer, set the initial learning rate to 0.0002, and implemented a cosine decay learning strategy. Balanced BCE loss is used for shadow detection. BCE loss and IOU loss are used for camouflaged object detection by default. The training spanned 20 epochs and was performed using Tesla A800 GPU. The software environment, including PyTorch, remained consistent with [52].

In Section 3.4, we discussed the data analysis of the COCO and COD10K datasets. COCO, a widely utilized dataset, comprises data for tasks such as object detection and segmentation. Our analysis of COCO focused on its object detection dataset. For COD10K, which contains irregular binary mask images for segmentation tasks, we implemented additional preprocessing to facilitate an analysis analogous to that of COCO. Specifically, for irregular masks, we identified the outermost white mask pixels in each of the four cardinal directions. These pixels were used as one edge of a bounding box, oriented perpendicular to the respective direction, to form a closed rectangular box. This preprocessing enabled further analysis of the dataset's characteristics.

4.3. Experimental results

4.3.1. Camouflaged object detection

The evaluation of SAM was conducted across the CHAMELEON, CAMO, and COD10K tasks. Our experiments reveal SAM's subpar performance across these tasks. SAM exhibits shortcomings in detecting certain concealed objects, as depicted in Fig. 4. This deficiency is further corroborated by the quantitative findings outlined in Table 1. Across all metrics considered, SAM's performance notably lags behind that of the currently available state-of-the-art methods.

In Table 1, we compare our method with recent advanced works, including SAM, SINet, and FBNet, etc, revealing our comprehensive outperformance, sometimes exceeding 20%. EVP represents an innovative integration of the Adapter concept into the Transformer architecture models, facilitating enhanced focus on high-frequency information in images. This adaptation leads to state-of-the-art performance in tasks such as camouflage and shadow detection, with minimal parameter augmentation. Despite observing a minor disparity compared to EVP in the CAMO dataset concerning the E-measure and MAE metrics, our method showcases superior performance, reflecting limitations in the CAMO dataset's validation size and generalization capability, as indicated by performance on the CHAMELEON and COD10K datasets. Our method exhibits leads of up to 7% and 9.3% on these two datasets, respectively.

In Fig. 4, visualization trials were conducted using the SAM-Online approach. Three sets of results were obtained using the Everything prompt mode, two sets using the Box prompt mode, and two sets using the Point prompt mode.

Initially, it is crucial to identify the objects for detection. In the Everything mode, the targets are a crab hidden in a sand pile, a seahorse concealed in front of colorful coral reefs, and a fish hidden within brown coral reefs. Although SAM inevitably displays objects other than the targets, it does not hinder our focus on the camouflaged objects. From SAM's detection results, while accurately identifying the main bodies of the crab and seahorse in the first two sets of images, it fails to segment specific components, such as a small pincer of the

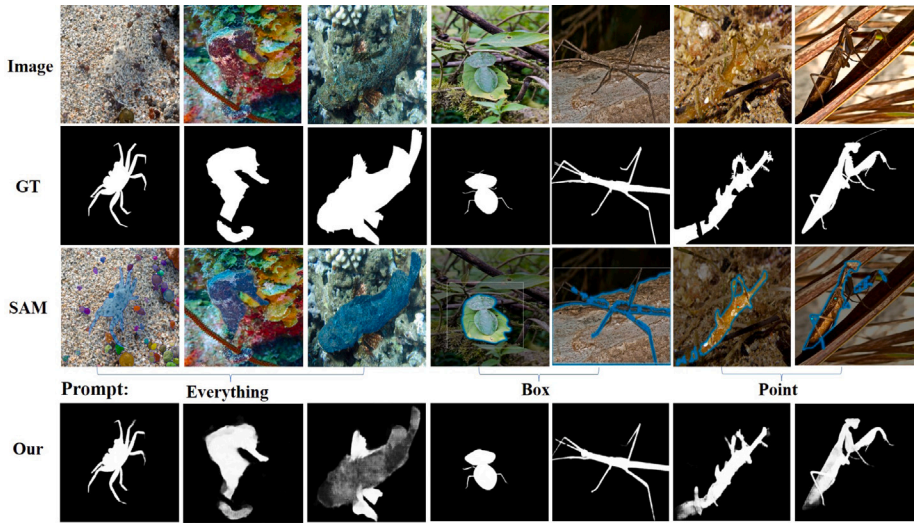

Fig. 4. Visualization results of camouflage detection. As illustrated in the figure, the SAM failed to perceive those animals that are visually hidden/concealed in their natural surroundings. Our approach can significantly elevate the performance of object segmentation with SAM. The samples are from the COD-10K dataset.

Table 1
Quantitative result for camouflage detection.

Method	COD10K [8]				CAMO [32]				CHAMELEON [53]			
	S_α	E_ϕ	F_β^ω	mae↓	S_α	E_ϕ	F_β^ω	mae↓	S_α	E_ϕ	F_β^ω	mae↓
SINet [54]	0.771	0.806	0.551	0.051	0.751	0.771	0.606	0.100	0.869	0.891	0.740	0.440
RankNet [31]	0.767	0.861	0.611	0.045	0.712	0.791	0.583	0.104	0.846	0.913	0.767	0.045
JCOD [55]	0.800	0.872	–	0.041	0.792	0.839	–	0.82	0.870	0.924	–	0.039
PFNet [56]	0.800	0.868	0.660	0.040	0.782	0.852	0.695	0.085	0.882	0.942	0.810	0.330
ZoomNeXT [57]	0.861	0.925	0.768	0.026	0.833	**0.891**	**0.774**	**0.065**	**0.908**	**0.963**	**0.858**	**0.021**
SAM [3]	0.783	0.798	0.701	0.050	0.684	0.687	0.606	0.132	0.727	0.734	0.639	0.081
SAM-Adapter [58]	**0.883**	**0.918**	**0.801**	**0.025**	**0.847**	0.873	0.765	0.070	0.896	0.919	0.824	0.033
Our	**0.899**	**0.941**	**0.832**	**0.021**	**0.870**	**0.906**	**0.796**	**0.055**	**0.924**	0.956	**0.880**	**0.021**

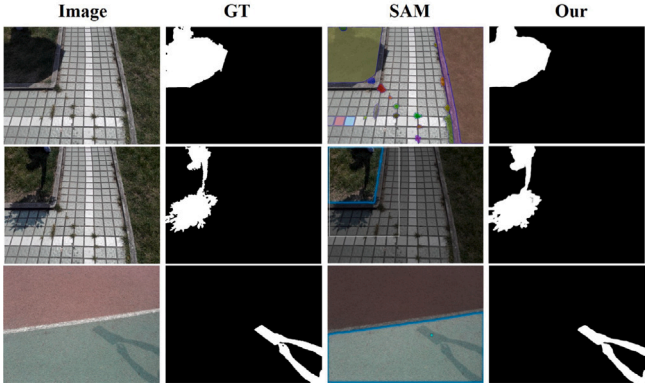

Fig. 5. Visualization results of shadow detection. SAM: Everthing, Box and Click(Top-down).

Table 2
Comparison and ablation study for shadow detection. EVPv2(ViT/Seg) means EVPv2 with different backbone. By default, IoU is used.

Model					BER↓
	Head	Adapt	Corner	sIoU	
SAM	–	–	–	–	40.51
DSC	–	–	–	–	3.42
DSD	–	–	–	–	2.17
BDRAR	–	–	–	–	2.69
MTMT	–	–	–	–	1.72
FDRNet	–	–	–	–	1.55
EVPv2(ViT) [52]	–	–	–	–	1.80
EVPv2(Seg) [52]	–	–	–	–	**1.35**
	–	–	✓	–	1.79
	–	✓	–	–	1.74
Our	✓	–	–	–	1.54
	✓	✓	✓	–	1.51
	✓	✓	✓	✓	**1.50**

crab or the lower part of the seahorse obscured by coral polyps. In the third set of images, it furthermore fails to segment the fish fins. Direct comparison with the ground truth (GT) demonstrates the effective mitigation of SAM's challenges by our method.

In Box mode, our targets comprise insects resting on green foliage and perched on brown branches. Utilizing bounding boxes, SAM prioritizes objects and their surroundings, substantially boosting detection efficiency. Despite this support, SAM struggles with accurately delineating foreground-background separation, frequently segmenting parts of the object or even the entire object along with its surroundings. In contrast, our method adeptly distinguishes between foreground and background without relying on bounding boxes, accurately segmenting even the slender tentacles of the insects.

In Click mode, our targets comprise bamboo worms and praying mantises hidden among intertwined branches. For SAM, we consistently utilized two-point prompts: one for capturing detailed features and the other for delineating the main body. In this mode, SAM's inability to segment intricate objects is accentuated, whereas our approach is not constrained by this limitation.

Based on the visual results, SAM exhibits notable limitations in segmenting objects with complex backgrounds and detailed parts, intermittently failing to accurately differentiate between foreground and background. COD-SAM demonstrates superior segmentation accuracy

Table 3
Ablation results. 'B': based on Vit-Base model, 'L': based on Vit-Large model. By default, IoU is used.

Method					COD10K				CAMO				CHAMELEON			
	Head	Adapt	Corner	sIoU	S_α	E_ϕ	F_β^ω	mae↓	S_α	E_ϕ	F_β^ω	mae↓	S_α	E_ϕ	F_β^ω	mae↓
SAM	–	–	–	–	0.783	0.798	0.701	0.050	0.684	0.687	0.606	0.132	0.727	0.734	0.639	0.081
SAM-Adapter	–	–	–	–	0.883	0.918	0.801	0.025	0.847	0.873	0.765	0.070	0.896	0.919	0.824	0.033
JCOD	–	–	–	–	0.800	0.872	–	0.041	0.792	0.839	–	0.820	0.870	0.924	–	0.039
Our-B	✓	✓	✓	–	0.798	0.849	0.642	0.047	0.785	0.813	0.665	0.096	0.828	0.862	0.693	0.054
Our-L	✓	✓	✓	–	0.873	0.916	0.784	0.028	0.844	0.880	0.764	0.067	0.902	0.932	0.835	0.030
	–	–	✓	–	0.888	0.931	0.818	0.024	0.861	0.896	0.793	0.061	0.907	0.937	0.841	0.027
	–	✓	–	–	0.879	0.912	0.801	0.024	0.838	0.869	0.766	0.070	0.895	0.923	0.827	0.031
	✓	–	–	–	0.879	0.920	0.811	0.024	0.846	0.880	0.778	0.066	0.898	0.925	0.834	0.031
	✓	–	✓	–	0.888	0.931	0.808	0.024	0.856	0.891	0.777	0.063	0.913	0.945	0.844	0.026
Our-H	–	✓	✓	–	0.890	0.927	0.815	0.022	0.852	0.887	0.777	0.063	0.910	0.940	0.847	0.026
	✓	✓	–	–	0.893	0.934	0.818	0.023	0.860	0.897	0.780	0.059	0.917	0.949	0.863	0.024
	✓	✓	✓	–	0.899	0.937	**0.835**	**0.020**	0.859	0.897	0.795	0.060	0.920	0.949	0.865	0.024
	✓	✓	✓	✓	**0.899**	**0.941**	0.832	0.021	**0.870**	**0.906**	**0.796**	**0.055**	**0.924**	**0.956**	**0.880**	**0.021**

and completeness. In instances where SAM fails to identify objects, COD-SAM not only accurately identifies the camouflaged objects but also offers more detailed segmentation of them.

4.3.2. Shadow detection

Three images were selected from the test set for visual analysis, as shown in Fig. 5. The results displayed include the original image, ground truth (GT), SAM-Online, and COD-SAM. Interactive prompts in Everything, Box, and Click modes were used for the SAM-Online experiment. SAM consistently fails to accurately segment objects, exhibiting significant discrepancies such as object misidentification and incomplete segmentation. However, our segmentation results closely approximate the ground truth.

Table 2 displays the outcomes of straightforward ablation studies focusing on the application of three techniques, compared with contemporary shadow detection methods using the ISTD dataset. Our approach consistently surpasses most of the related studies, whether implementing single techniques or a holistic strategy. Although slightly trailing behind EVPv2, this indirectly confirms the efficacy of our method and its applicability to other downstream tasks. It is worth noting, however, that EVPv2 is fine-tuned using a different backbone. Although it falls behind methods fine-tuned with SegFormer, our approach still exhibits certain advantages over methods also fine-tuned with ViT.

4.4. Ablation study

In the aforementioned experiments, we presented only the optimal outcomes. This section details the results of ablation studies and provides a comprehensive analysis.

The superior performance indicated in Table 1 demonstrates that our method temporarily surpassed competing approaches. Consequently, we will extend our experimental analysis in Table 3.

Table 3 presents comparisons with three representative works: SAM, SAM-Adapter, and JCOD. These works are notable for their segmentation tasks within the new paradigm, advanced adapter design concepts, and focus on COD tasks. The data in the table show that COD-SAM surpasses SAM on the COD10K, CHAMELEON, and CAMO validation sets, with improvements of up to 14.3% (E_ϕ metric), 24.1% (F_β^ω), and 21.9% (E_ϕ), respectively. Compared to SAM-Adapter, COD-SAM shows gains of 3.1%, 5.6%, and 3.3%. Against JCOD, COD-SAM achieves leads of up to 9.9% (S_α metric), 5.4% (S_α), and 6.8% (S_α), respectively. In conclusion, our method significantly outperforms the aforementioned three representative methods.

Ablation studies were performed, selectively evaluating the effects of integrating COD-Head, COD-Adapter, and Corner. Even when solely substituting traditional dense prompts with Corner, our approach exhibited enhanced performance. The inclusion of COD-Head and COD-Adapter further improved the model's performance, yielding a more distinct advantage. Moreover, our algorithm was validated using Vit-B and Vit-L pretrained models. The findings indicate that COD-SAM significantly enhances performance, with the Base model at times outperforming SAM, and the performance of the L model comparable to that of SAM-Adapter (Vit-H) and JCOD.

Furthermore, we conducted further comparisons with SAM-Adapter regarding the complexity of the model. By replicating this study, we discovered that a mere addition of 0.59M parameters (SAM-Adapter has 641.3M parameters, our method has 641.9M parameters, representing a marginal increase of 0.09%, less than 0.1% increase) resulted in a performance enhancement of 5.6%. These findings strongly demonstrate the efficiency of our method.

5. Conclusion

In this study, we enhance the robust universal segmentation model, SAM, to address both camouflage and shadow detection tasks, introducing COD-SAM. Leveraging the innovative fusion prompt algorithm design concept, we introduce a global gradient weak prompt, a local strong prompt, task-adaptive loss, and a convolutional head specifically designed for optimizing local features. By integrating these novel approaches, we achieve significant improvements in three datasets for camouflage detection and demonstrate considerable promise in shadow detection. These results highlight the effectiveness of our adaptive adjustment techniques in camouflage detection and emphasize the versatility of robust universal segmentation models for a range of downstream tasks.

CRediT authorship contribution statement

Dongyang Gao: Writing – review & editing, Writing – original draft. **Yichao Zhou:** Data curation. **Hui Yan:** Formal analysis. **Chen Chen:** Methodology. **Xiyuan Hu:** Methodology, Formal analysis.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

References

[1] R. Bommasani, D.A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M.S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, et al., On the opportunities and risks of foundation models, 2021, arXiv preprint arXiv:2108.07258.

[2] K. Zhou, Z. Qiu, D. Fu, Multi-scale contrastive adaptor learning for segmenting anything in underperformed scenes, Neurocomputing 606 (2024) 128395.

[3] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A.C. Berg, W.-Y. Lo, et al., Segment anything, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 4015–4026.

[4] S. Li, Y. Ren, Y. Yu, Q. Jiang, X. He, H. Li, A survey of deep learning algorithms for colorectal polyp segmentation, Neurocomputing (2024) 128767.

[5] T.B. Brown, Language models are few-shot learners, 2020, arXiv preprint arXiv:2005.14165.

[6] S. Zhang, C. Chen, X. Hu, S. Peng, Balanced knowledge distillation for long-tailed learning, Neurocomputing 527 (2023) 36–46.

[7] F. Huang, P. Wu, X. Li, J. Li, R. Zhao, Adaptive event-triggered pseudolinear consensus filter for multi-UAVs bearings-only target tracking, Neurocomputing 571 (2024) 127127.

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

[9] T.W. Pike, Quantifying camouflage and conspicuousness using visual salience, Methods Ecol. Evol. 9 (8) (2018) 1883–1895.

[10] X. Hu, Z. Zhang, Z. Jiang, S. Chaudhuri, Z. Yang, R. Nevatia, SPAN: Spatial pyramid attention network for image manipulation localization, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16, Springer, 2020, pp. 312–328.

[11] A. Islam, C. Long, A. Basharat, A. Hoogs, DOA-GAN: Dual-order attentive generative adversarial network for image copy-move forgery detection and localization, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 4676–4685.

[12] R. Salloum, Y. Ren, C.-C.J. Kuo, Image splicing localization using a multi-task fully convolutional network (MFCN), J. Vis. Commun. Image Represent. 51 (2018) 201–209.

[13] Y. Wu, W. Abd-Almageed, P. Natarajan, Deep matching and validation network: An end-to-end solution to constrained image splicing localization and detection, in: Proceedings of the 25th ACM International Conference on Multimedia, 2017, pp. 1480–1502.

[14] Y. Wu, P. Zhang, M. Gu, J. Zheng, X. Bai, Embodied navigation with multi-modal information: A survey from tasks to methodology, Inf. Fusion (2024) 102532.

[15] A. Gupta, P. Dollar, R. Girshick, Lvis: A dataset for large vocabulary instance segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 5356–5364.

[16] A. Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Malloci, A. Kolesnikov, et al., The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale, Int. J. Comput. Vis. 128 (7) (2020) 1956–1981.

[17] X. Bai, P. Zhang, X. Yu, J. Zheng, E.R. Hancock, J. Zhou, L. Gu, Learning from human attention for attribute-assisted visual recognition, IEEE Trans. Pattern Anal. Mach. Intell. (2024).

[18] C. Zhang, D. Han, Y. Qiao, J.U. Kim, S.-H. Bae, S. Lee, C.S. Hong, Faster segment anything: Towards lightweight sam for mobile applications, 2023, arXiv preprint arXiv:2306.14289.

[19] Z. Zhao, Enhancing autonomous driving with grounded-segment anything model: Limitations and mitigations, in: 2023 IEEE 3rd International Conference on Data Science and Computer Application, ICDSCA, IEEE, 2023, pp. 1258–1265.

[20] T. Li, G. Pang, X. Bai, J. Zheng, L. Zhou, X. Ning, Learning adversarial semantic embeddings for zero-shot recognition in open worlds, Pattern Recognit. 149 (2024) 110258.

[21] J. Zhang, L. Huang, X. Bai, J. Zheng, L. Gu, E. Hancock, Exploring the usage of pre-trained features for stereo matching, Int. J. Comput. Vis. (2024) 1–22.

[22] M. Jia, L. Tang, B.-C. Chen, C. Cardie, S. Belongie, B. Hariharan, S.-N. Lim, Visual prompt tuning, in: European Conference on Computer Vision, Springer, 2022, pp. 709–727.

[23] M. Sandler, A. Zhmoginov, M. Vladymyrov, A. Jackson, Fine-tuning image transformers using learnable memory, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 12155–12164.

[24] H.K. Cheng, J. Chung, Y.-W. Tai, C.-K. Tang, Cascadepsp: Toward class-agnostic and very high-resolution segmentation via global and local refinement, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 8890–8899.

[25] A. Kirillov, Y. Wu, K. He, R. Girshick, Pointrend: Image segmentation as rendering, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 9799–9808.

[26] Z. Yang, Z. Gan, J. Wang, X. Hu, Y. Lu, Z. Liu, L. Wang, An empirical study of gpt-3 for few-shot knowledge-based vqa, in: Proceedings of the AAAI Conference on Artificial Intelligence, 36, (3) 2022, pp. 3081–3089.

[27] L. Ke, M. Danelljan, H. Ding, Y.-W. Tai, C.-K. Tang, F. Yu, Mask-free video instance segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 22857–22866.

[28] K. Zhou, J. Yang, C.C. Loy, Z. Liu, Learning to prompt for vision-language models, Int. J. Comput. Vis. 130 (9) (2022) 2337–2348.

[29] P. Sengottuvelan, A. Wahi, A. Shanmugam, Performance of decamouflaging through exploratory image analysis, in: 2008 First International Conference on Emerging Trends in Engineering and Technology, IEEE, 2008, pp. 6–10.

[30] J.Y.Y.H.W. Hou, J. Li, Detection of the mobile object with camouflage color under dynamic background based on optical flow, Procedia Eng. 15 (2011) 2201–2205.

[31] Y. Lv, J. Zhang, Y. Dai, A. Li, B. Liu, N. Barnes, D.-P. Fan, Simultaneously localize, segment and rank the camouflaged objects, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 11591–11601.

[32] T.-N. Le, T.V. Nguyen, Z. Nie, M.-T. Tran, A. Sugimoto, Anabranch network for camouflaged object segmentation, Comput. Vis. Image Underst. 184 (2019) 45–56.

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

[34] J. Zhang, J. Shao, J. Chen, D. Yang, B. Liang, R. Liang, PFNet: an unsupervised deep network for polarization image fusion, Opt. Lett. 45 (6) (2020) 1507–1510.

[35] J. Lin, X. Tan, K. Xu, L. Ma, R.W. Lau, Frequency-aware camouflaged object detection, ACM Trans. Multimed. Comput. Commun. Appl. 19 (2) (2023) 1–16.

[36] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 10012–10022.

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

[38] K. Xu, X. Hu, X. Zhou, X. Xu, L. Qi, C. Chen, RLGC: Reconstruction learning fusing gradient and content features for efficient deepfake detection, IEEE Trans. Consum. Electron. (2024).

[39] R. Cong, M. Sun, S. Zhang, X. Zhou, W. Zhang, Y. Zhao, Frequency perception network for camouflaged object detection, in: Proceedings of the 31st ACM International Conference on Multimedia, 2023, pp. 1179–1189.

[40] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J.M. Alvarez, P. Luo, SegFormer: Simple and efficient design for semantic segmentation with transformers, Adv. Neural Inf. Process. Syst. 34 (2021) 12077–12090.

[41] A. Vaswani, Attention is all you need, Adv. Neural Inf. Process. Syst. (2017).

[42] W. Zhao, F. Zhao, D. Wang, H. Lu, Defocus blur detection via multi-stream bottom-top-bottom fully convolutional network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 3080–3088.

[43] T.F.Y. Vicente, L. Hou, C.-P. Yu, M. Hoai, D. Samaras, Large-scale training of shadow detectors with noisily-annotated shadow examples, in: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, the Netherlands, October 11-14, 2016, Proceedings, Part VI 14, Springer, 2016, pp. 816–832.

[44] T. Chen, A. Lu, L. Zhu, C. Ding, C. Yu, D. Ji, Z. Li, L. Sun, P. Mao, Y. Zang, Sam2-adapter: Evaluating & adapting segment anything 2 in downstream tasks: Camouflage, shadow, medical image segmentation, and more, 2024, arXiv preprint arXiv:2408.04579.

[45] D. Hendrycks, K. Gimpel, Gaussian error linear units (gelus), 2016, arXiv preprint arXiv:1606.08415.

[46] Y. Xiong, B. Varadarajan, L. Wu, X. Xiang, F. Xiao, C. Zhu, X. Dai, D. Wang, F. Sun, F. Iandola, et al., Efficientsam: Leveraged masked image pretraining for efficient segment anything, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 16111–16121.

[47] Z. Tang, H. Fang, S. Zhou, T. Yang, Z. Zhong, T. Hu, K. Kirchhoff, G. Karypis, AutoGluon-multimodal (AutoMM): Supercharging multimodal AutoML with foundation models, 2024, arXiv preprint arXiv:2404.16233.

[48] X. Dai, Y. Chen, J. Yang, P. Zhang, L. Yuan, L. Zhang, Dynamic detr: End-to-end object detection with dynamic attention, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 2988–2997.

[49] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, D. Ren, Distance-iou loss: Faster and better learning for bounding box regression, in: Proceedings of the AAAI Conference on Artificial Intelligence, 34, (07) 2020, pp. 12993–13000.

[50] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft coco: Common objects in context, in: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13, Springer, 2014, pp. 740–755.

[51] J. Wang, X. Li, J. Yang, Stacked conditional generative adversarial networks for jointly learning shadow detection and shadow removal, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1788–1797.

[52] W. Liu, X. Shen, C.-M. Pun, X. Cun, Explicit visual prompting for low-level structure segmentations, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 19434–19445.

[53] P. Skurowski, H. Abdulameer, J. Błaszczyk, T. Depta, A. Kornacki, P. Kozieł, Animal camouflage analysis: Chameleon database, Unpubl. Manuscr. 2 (6) (2018) 7.

[54] Y. Sun, S. Wang, C. Chen, T.-Z. Xiang, Boundary-guided camouflaged object detection, 2022, arXiv preprint arXiv:2207.00794.
[55] A. Li, J. Zhang, Y. Lv, B. Liu, T. Zhang, Y. Dai, Uncertainty-aware joint salient object and camouflaged object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 10071–10081.
[56] H. Mei, G.-P. Ji, Z. Wei, X. Yang, X. Wei, D.-P. Fan, Camouflaged object segmentation with distraction mining, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 8772–8781.
[57] Y. Pang, X. Zhao, T.-Z. Xiang, L. Zhang, H. Lu, ZoomNeXt: A unified collaborative pyramid network for camouflaged object detection, IEEE Trans. Pattern Anal. Mach. Intell. (2024) http://dx.doi.org/10.1109/TPAMI.2024.3417329.
[58] T. Chen, L. Zhu, C. Ding, R. Cao, Y. Wang, Z. Li, L. Sun, P. Mao, Y. Zang, SAM fails to segment anything?–SAM-adapter: Adapting SAM in underperformed scenes: Camouflage, shadow, medical image segmentation, and more, 2023, arXiv preprint arXiv:2304.09148.