The Thirty-Ninth AAAI Conference on Artificial Intelligence (AAAI-25)

FOCUS: Towards Universal Foreground Segmentation

Zuyao You[1,2], **Lingyu Kong**[1*], **Lingchen Meng**[1,2*], **Zuxuan Wu**[1,2†],

[1]Shanghai Key Lab of Intell. Info. Processing, School of CS, Fudan University
[2]Shanghai Collaborative Innovation Center of Intelligent Visual Computing
{zyyou23, lykong22}@m.fudan.edu.cn, {lcmeng20, zxwu}@fudan.edu.cn

Abstract

Foreground segmentation is a fundamental task in computer vision, encompassing various subdivision tasks. Previous research has typically designed task-specific architectures for each task, leading to a lack of unification. Moreover, they primarily focus on **recognizing foreground objects** without effectively **distinguishing them from the background**. In this paper, we emphasize the importance of the background and its relationship with the foreground. We introduce **FOCUS**, the **F**oreground **O**bje**C**ts **U**niversal **S**egmentation framework that can handle multiple foreground tasks. We develop a multi-scale semantic network using the edge information of objects to enhance image features. To achieve boundary-aware segmentation, we propose a novel distillation method, integrating the contrastive learning strategy to refine the prediction mask in multi-modal feature space. We conduct extensive experiments on a total of **13 datasets** across **5 tasks**, and the results demonstrate that FOCUS consistently outperforms the state-of-the-art task-specific models on most metrics.

Code — https://github.com/geshang777/FOCUS/

Introduction

Foreground segmentation is a fundamental task in computer vision where the primary goal is to delineate the prominent objects (foreground) from the rest of the image (background), typically referring to salient object detection (SOD) and camouflaged object detection (COD) (Pang et al. 2022a, 2024a). In this paper, the concept of foreground segmentation can be extended to delineating objects that interest you most in the image, where the primary goal is to obtain the Mask of Interest (MoI), *e.g.*, MoI should denote the mask of the camouflaged object in COD. According to this definition, tasks such as shadow detection (SD), defocus blur detection (DBD), forgery detection (FD), *etc.* belong to the category of foreground segmentation, too.

Currently, in the field of generic segmentation, *e.g.* instance segmentation, semantic segmentation, and panoptic segmentation, *etc.*, there are already many sophisticated models (Kirillov et al. 2023; Cheng et al. 2022; Jain et al. 2023; Ding et al. 2023b,a). However, these models often lack targeted training for specific foreground segmentation tasks. For instance, in the COD task, SAM struggles to distinguish camouflaged objects from the background (Hu et al. 2024). Furthermore, without prompt-guided methods, most traditional segmentation algorithms generate multiple masks for one image at the same time (Cheng, Schwing, and Kirillov 2021; Cheng et al. 2022; Jain et al. 2023), but users do not require such many masks in many real-world scenarios, *e.g.* image background removal, MoI is all they need. While foreground segmentation typically produces a single or specific type of mask, making it more in line with user needs.

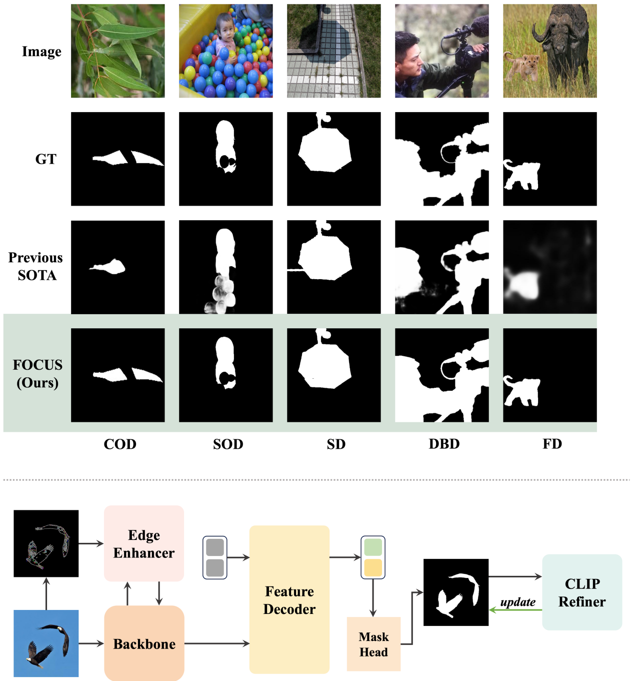

Figure 1: With one unified architecture, FOCUS can handle various foreground segmentation tasks. Our proposed method can generate boundary-aware masks that are smoother and more detailed than the previous state-of-the-art task-specific models. Zoom in for more details.

However, as mentioned earlier, when the concept of fore-

[*]These authors contributed equally.
[†]Corresponding author.

ground is generalized as MoI, the scope of foreground segmentation tasks is very broad. Currently, there is a lack of an excellent and universal framework that can handle all foreground segmentation tasks. Most foreground segmentation models are task-specific (Wang et al. 2022a; Zhao et al. 2021; Zhu et al. 2021; Zheng et al. 2024a; Xie et al. 2022; Wang et al. 2022b) . Some models (Pang et al. 2024a, 2022a) achieve universality in SOD and COD tasks, but given the similarity between COD and SOD tasks, they will not be discussed as universal models here. To the best of our knowledge, the work most closely related to ours is (Liu et al. 2023). However, it still significantly lags behind task-specific models after fine-tuning in the subdivision tasks.

Besides, previous foreground segmentation models primarily focused on recognizing the foreground objects without effectively distinguishing them from the background, neglecting the background and the relationship between the background and the foreground. In fact, background information plays a critical role in computer vision tasks (Li et al. 2023; Meng et al. 2024). Foreground segmentation inherently involves distinguishing the foreground from the background, making both elements and their relationship vital. However, current approaches fail to address the background segmentation separately. Consequently, this oversight impacts the overall performance of foreground segmentation.

The issues above can be summarized as follows: **(1)**How to generally represent the foreground and background of different foreground segmentation tasks? **(2)**How to fully utilize the background information of an image to optimize prediction results? In this paper, we introduce **FOCUS**, a unified multi-modal approach to tackle multiple subdivision tasks of foreground segmentation.

To universally represent the foreground and background, we borrow the object queries concept from DETR (Carion et al. 2020) by introducing ground queries. We apply the multi-scale strategy (Cheng et al. 2022) to extract image features to feed the transformer decoder, using masked attention to enable the ground queries to focus on relevant features corresponding to foreground and background. We utilize the feature map obtained from the backbone to initialize the masked attention, which can serve as a localization prior. During this process, the ground queries adapt to learn the features relevant to the context of different tasks, making them universal features.

To fully leverage the background information in images, we employ contrastive learning strategies. We propose the CLIP refiner, using the powerful multi-modal learning ability from CLIP (Radford et al. 2021) to correct the masks generated by previous modules. We fuse the mask and image and align the fused image and its corresponding text in multi-modal feature space to refine the masks. This not only refines the edges of the mask but also accentuates the distinction between foreground and background. We treat foreground segmentation and background segmentation as two independent tasks, and in the inference stage, the probability map of both foreground and background will jointly determine the boundary of MoI.

We conduct detailed experiments on 13 datasets across five foreground segmentation tasks and achieve or exceed state-of-the-art on most provided metrics. Fig. 1 shows the outstanding performance of our proposed FOCUS on different sub-tasks of the foreground segmentation.

Our contributions can be summarized as follows:

- We propose a unified framework for foreground segmentation tasks, including SOD, COD, SD, DBD, and FD;
- We propose a novel module, using the contrastive learning strategy to utilize the background information to refine the mask while widening the distance between the foreground and the background;
- We conduct extensive experiments on multiple datasets across multiple tasks, and results demonstrate that our method achieves state-of-the-art performance.

Related Work

Foreground Segmentation

As mentioned earlier, several tasks are crucial in foreground segmentation, including salient object detection(SOD), camouflaged object detection (COD), shadow detection (SD), defocus blur detection (DBD), and forgery detection (FD). SOD aims at segmenting the most visually attractive objects from the input images. COD focuses on disguised objects that blend seamlessly into their surroundings, *e.g.* mimetic animals and body paintings. SD aims to segment shadow regions from natural scenes. DBD aims at separating in-focus and out-of-focus regions, which is caused by the different focal lengths of the cameras, slightly different from SOD. The goal of FD is to identify altered or manipulated areas in images, typically involving addition, replacement, or deletion. Previous models normally designed architectures for specific foreground segmentation task (Wang et al. 2022a; Zhao et al. 2021; Zhu et al. 2021; Zheng et al. 2024a; Xie et al. 2022), and currently, there is a lack of effective methods to handle this foreground segmentation tasks universally.

Universal Segmentation

Universal segmentation has emerged as a significant trend in computer vision. It aims to unify various segmentation tasks within a single framework. This trend started with efforts to unify semantic and instance segmentation through panoptic segmentation (Kirillov et al. 2019) and has since expanded to include a broader range of tasks. Recent works have shifted towards designing universal segmentation models with generalization ability and versatility. Mask2Former (Cheng et al. 2022) utilizes a masked-attention mechanism to unify instance, semantic and panoptic segmentation. OneFormer (Jain et al. 2023) further improves Mask2Former with a multi-task train-once design. More recent approaches like SAM (Kirillov et al. 2023) push the boundaries of universal segmentation with the ability of zero-shot segmentation. In the field of foreground segmentation, the unified architecture most related to ours is EVP (Liu et al. 2023). EVP freezes a pre-trained model and then learns task-specific knowledge using an adapter structure, but its performance falls behind task-specific models. In this work, we aim to find a more effective way to unify the foreground segmentation tasks using one single architecture.

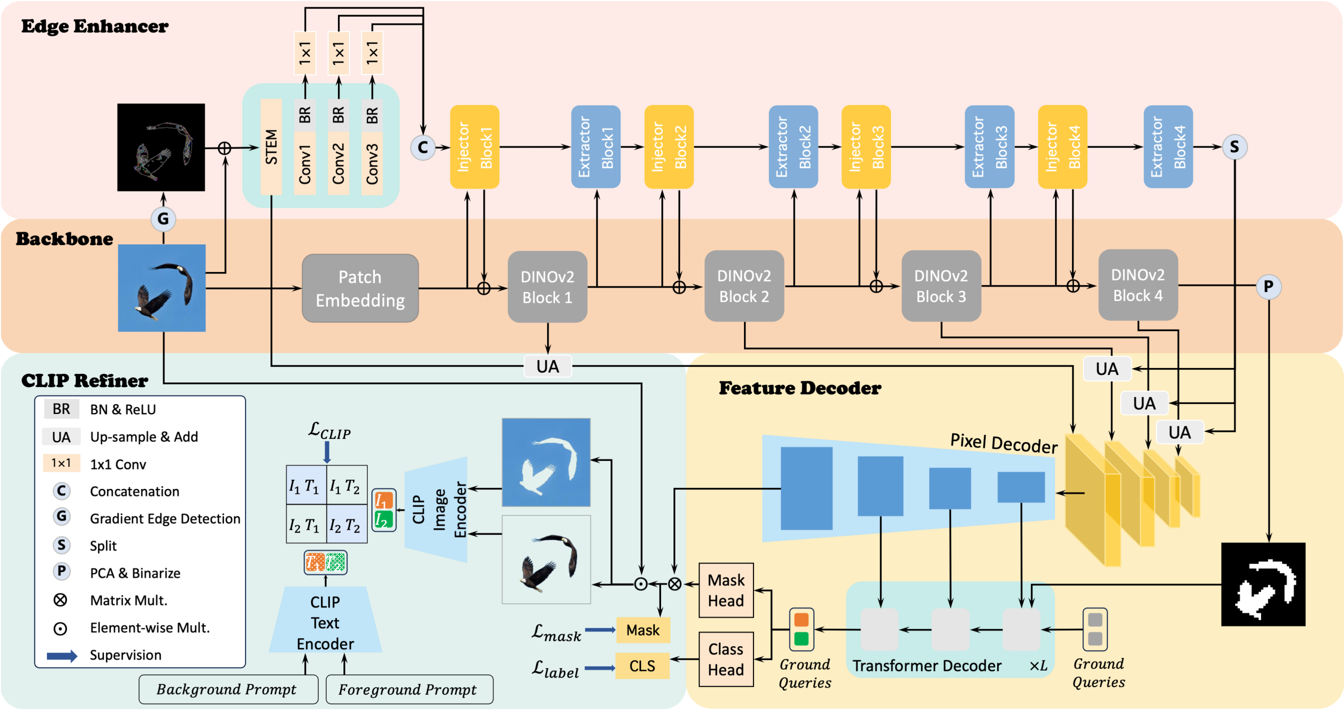

Figure 2: An overview of our proposed FOCUS, a multi-scale and multi-modal semantic framework for universal foreground segmentation, mainly includes the backbone, edge enhancer, feature decoder, and CLIP refiner. Refer to the main text for details.

Methods

Unified Architecture

Previously, there was a lack of unified architecture for handling all foreground segmentation subdivision tasks. Given an image from different foreground segmentation tasks, our goal is to use a unified architecture to predict the corresponding MoI in the task context. The problem can be defined by:

$$U(I, T_i) = MoI$$

T_i refers to different foreground segmentation tasks, $\forall T_i \in \{T_1, \dots, T_n\}$ the unified framework U should infer the corresponding MoI from the images I.

We propose FOCUS, a unified architecture that can handle multiple foreground segmentation tasks. We borrow the concept of object queries from (Carion et al. 2020) and introduce the ground queries (**GQ**) here. **GQ** are two distinct tensors, designated as the foreground query and background query, we aim to only use these two learned tensors to respectively embed and represent the foreground and the background within the image based on the context of the task. Fig. 2 provides an overview of our approach FOCUS. After obtaining multi-scale edge-enhanced features from the backbone and the edge enhancer, the pixel decoder will generate pixel-level output and these pixel-level features will be fed into the transformer decoder with **GQ**, where **GQ** updated by masked attention (Cheng et al. 2022) to get ground-centric output. It can be formulated as:

$$\mathbf{X}_l = \mathrm{softmax}(\mathcal{M}_{l-1} + \mathbf{GQ}_l \mathbf{K}_l^\top)\mathbf{V}_l + \mathbf{X}_{l-1},$$

Here, $\mathbf{K}_l, \mathbf{V}_l \in \mathbb{R}^{H_l W_l \times C}$ denotes the linearly transformed C-dimensional image feature from l_{th} block of pixel decoder, $\mathbf{X}_l \in \mathbb{R}^{2\times C}$ refers to the query feature from the l_{th} transformer decoder block and $\mathbf{X}_0$ is initialized by the input query feature of transformer decoder. $\mathbf{GQ}_l \in \mathbb{R}^{2\times C}$ is the l_{th} ground queries, and $\mathcal{M}_{l-1}$ is defined by:

$$\mathcal{M}_{l-1}(x,y) = \begin{cases} 0 & \text{if } \mathbf{M}_{l-1}(x,y) = 1 \\ -\infty & \text{otherwise} \end{cases},$$

$\mathbf{M}_{l-1} \in \{0,1\}^{2\times H_l W_l}$ is obtained by decoding $\mathbf{GQ}_{l-1}$ and binarizing, with dimensions resizing consistent with $\mathbf{K}_l$. DINOv2 (Oquab et al. 2023) is a recently proposed model designed for visual representation learning. The visualization of its feature map indicates that DINOv2 has already focused on the prominent objects in the image without supervision, showing richer semantics compared to other foundation models (Wang et al. 2022c; Meng et al. 2022). Therefore, we choose DINOv2 as the backbone for FOCUS, PCA and binarize the feature map of its last block to initialize the attention mask $\mathcal{M}_0$. $\mathcal{M}_0$ is formulated as:

$$\mathcal{M}_0(x,y) = \begin{cases} 0 & \text{if } \mathbf{F}_{DINOv2}(x,y) = 1 \\ -\infty & \text{otherwise} \end{cases}.$$

Here, $\mathbf{F}_{DINOv2}$ refers to the binary feature map from the last backbone block. It is resized to the same resolution of $\mathbf{K}_1$. The adoption of the new initialization method can leverage the localization prior knowledge learned by the DINOv2 on large-scale data.

We use two multi-layer perceptrons, designated as mask head and class head, to decode ground queries and generate mask and class predictions for both the foreground and background. During the inference stage, the foreground and background probability distributions are combined to predict the final MoI.

Edge Enhancer

In order to utilize the edge information of the object, we propose the edge enhancer, an effective module that uses foreground object edge information to correct the image features obtained by the backbone.

Inspired by the recent study that shows convolutions can help transformer understand local spatial information (Chen et al. 2022; Wang et al. 2022c), we use ResNet50 (He et al. 2016) to extract edge features from the image. We convert the image into grayscale to reduce the confusion caused by color, apply Gaussian smoothing (Davies 2004) to reduce noise, and then use an edge detector (Canny 1986) to obtain a gradient map and overlay it on the original image. As shown in Fig. 2, the ResNet can be divided into the STEM and the rest, the STEM serves as the initial feature extractor, comprising a series of convolution, batch normalization, and ReLU activation layers. The output of the rest convolution blocks will be flattened and projected into the same dimension D by 1×1 convolutions and concatenated to obtain feature pyramid $F^1_{\text{edge}} \in \mathbb{R}^{(\frac{HW}{8^2}+\frac{HW}{16^2}+\frac{HW}{32^2})\times D}$, H and W represent the resolution of the input image. Then, we follow ViT-Adapter (Chen et al. 2022), using the structure of the injector-extractor based on cross attention to fuse the image features from the backbone and ResNet. The injector can be formulated as:

$$\hat{F}^i_{\text{DINOv2}} = F^i_{\text{DINOv2}} + \gamma^i \text{MSDA}(F^i_{\text{DINOv2}}, F^i_{\text{edge}}),$$

MSDA refers to multi-scale deformable attention (Zhu et al. 2020), which takes the normalized backbone feature $F^i_{\text{DINOv2}} \in \mathbb{R}^{\frac{HW}{16^2}\times D}$ as the query, and the normalized edge feature $F^i_{\text{edge}} \in \mathbb{R}^{(\frac{HW}{8^2}+\frac{HW}{16^2}+\frac{HW}{32^2})\times D}$ as the key and value. γ^i is a learnable parameter for balancing the backbone feature and the fused feature. Similarly, the extractor can be formulated as:

$$\hat{F}^i_{\text{edge}} = F^i_{\text{edge}} + \text{ConvFFN}(\text{MSDA}(F^i_{\text{edge}}, F^{i+1}_{\text{DINOv2}})).$$

It is another multi-scale deformable attention like injector while taking the normalized edge feature $F^i_{\text{edge}} \in \mathbb{R}^{(\frac{HW}{8^2}+\frac{HW}{16^2}+\frac{HW}{32^2})\times D}$ as the query, and the output feature $F^{i+1}_{\text{DINOv2}} \in \mathbb{R}^{\frac{HW}{16^2}\times D}$ as the key and value. ConvFFN refers to the structure with two fully connected layers and a depthwise separable convolution layer. The $\hat{F}^i_{\text{edge}}$ will serve as the input for the next injector. We upscale the output from different blocks of backbone to resolutions of 1/4, 1/8, 1/16, and 1/32. Besides, we split the output of the last extractor, and restore them to their original size. Then we add the up-scaled backbone features with the corresponding split output from extractor and output from STEM to get the edge-enhanced multi-scale image features. These features will be fed into the pixel decoder, another module based on multi-scale deformable attention, for dense pixel-level predictions.

CLIP Refiner

Since the proposal of CLIP, there have been many works using CLIP for segmentation (Xu et al. 2022; Li et al. 2022; Wang et al. 2022d; Liang et al. 2023), which have proven that CLIP is effective not only at the image level but also at the pixel level. In this paper, we propose CLIP refiner, which uses the powerful multi-modal ability of CLIP to correct the masks of foreground and background.

Specifically, we decode the ground queries to obtain the masks of the foreground and background, resize them, and overlay them on the image. We use the prompts “It's an image of salient objects without background.” and “It's an image of background with salient objects removed.” to represent foreground and background, respectively. Note that the text can be adjusted according to the task. For example, in shadow detection, prompts can be replaced with “it's an image of shadow without background.” and “it's an image of background without shadow.” to extend CLIP refiner to other foreground segmentation tasks. We borrow the image encoder and text encoder from CLIP to encode the image and text separately. Then, we calculate the contrastive loss ($\mathcal{L}_{\text{clip}}$) between the mask-fused-image and text features.

$$\begin{aligned}\mathcal{L}_{\text{i2t}} = -\frac{1}{2}\Bigg[&\log\frac{\exp(I_f \cdot T_f/\tau)}{\exp(I_f \cdot T_f/\tau) + \exp(I_f \cdot T_b/\tau)} \\ &+ \log\frac{\exp(I_b \cdot T_b/\tau)}{\exp(I_b \cdot T_b/\tau) + \exp(I_b \cdot T_f/\tau)}\Bigg],\end{aligned}$$

$$\begin{aligned}\mathcal{L}_{\text{t2i}} = -\frac{1}{2}\Bigg[&\log\frac{\exp(T_f \cdot I_f/\tau)}{\exp(T_f \cdot I_f/\tau) + \exp(T_f \cdot I_b/\tau)} \\ &+ \log\frac{\exp(T_b \cdot I_b/\tau)}{\exp(T_b \cdot I_b/\tau) + \exp(T_b \cdot I_f/\tau)}\Bigg],\end{aligned}$$

$$\mathcal{L}_{\text{clip}} = \frac{1}{2}(\mathcal{L}_{\text{i2t}} + \mathcal{L}_{\text{t2i}}).$$

Here $I_{\text{f}}, I_{\text{b}}, T_{\text{f}}, I_{\text{b}} \in \mathbb{R}^{2\times S}$ denotes the S-dimensional image feature and text feature of foreground and background obtained by CLIP, τ is temperature parameter used to control the smoothness of the softmax function. The CLIP refiner iteratively refines the edges of masks generated by the preceding module, ensuring that only the appropriate pixels are included in the foreground or background. This process aligns the mask-fused image more closely with the corresponding text in the feature space while distancing it from the mismatched one. This not only makes the mask edges more accurate but also widens the gap between the foreground and background. The CLIP refiner is only used to distill knowledge from CLIP and will be discarded during the inference stage. Additionally, we keep the image and text encoders entirely frozen to fully leverage the multi-modal capabilities of CLIP without the potential performance degradation that might arise from fine-tuning.

	CAMO(250)				COD10K(2,026)				CHAMELEON(76)				NC4K(4,121)			
	$\mathcal{S}_m \uparrow$	$E_\xi \uparrow$	$F_\beta^w \uparrow$	$MAE \downarrow$	$\mathcal{S}_m \uparrow$	$E_\xi \uparrow$	$F_\beta^w \uparrow$	$MAE \downarrow$	$\mathcal{S}_m \uparrow$	$E_\xi \uparrow$	$F_\beta^w \uparrow$	$MAE \downarrow$	$\mathcal{S}_m \uparrow$	$E_\xi \uparrow$	$F_\beta^w \uparrow$	$MAE \downarrow$
SINet$_{20}$	.751	.771	.606	.100	.771	.806	.551	.051	.869	.891	.740	.044	.808	.871	.723	.058
PFNet$_{22}$	.782	.852	.695	.085	.800	.868	.660	.040	.882	.942	.810	.033	.829	.898	.745	.053
ZoomNet$_{22}$	.820	.892	.752	.066	.838	.911	.729	.029	.902	.958	.845	.023	.853	.912	.784	.043
BSA-Net$_{22}$	.794	.867	.717	.079	.818	.901	.699	.034	.895	.957	.841	.027	.842	.907	.771	.048
FSPNet$_{23}$	.856	.899	.799	.050	.851	.895	.735	.026	.908	.965	.851	.023	.879	.915	.816	.035
ZoomNeXt$_{24}$	.889	.945	.857	.041	.898	.956	.827	.018	.924	**.975**	.885	.018	.903	.951	.863	.028
BiRefNet$_{24}$	.904	.954	.890	.030	**.912**	.960	.874	.014	**.932**	-	**.915**	**.015**	.914	.953	.894	.023
FOCUS(Ours)	**.912**	**.963**	**.904**	**.025**	.910	**.974**	**.883**	**.013**	.922	**.975**	.908	.017	**.915**	**.964**	**.906**	**.020**

Table 1: Comparison of FOCUS with recent state-of-the-art COD methods.

	DUTS-TE(5,019)			DUT-OMRON(5,618)			HKU-IS(4,447)			ECSSD(1,000)			PACAL-S(850)		
	$\mathcal{S}_m \uparrow$	$E_\xi \uparrow$	$MAE \downarrow$	$\mathcal{S}_m \uparrow$	$E_\xi \uparrow$	$MAE \downarrow$	$\mathcal{S}_m \uparrow$	$E_\xi \uparrow$	$MAE \downarrow$	$\mathcal{S}_m \uparrow$	$E_\xi \uparrow$	$MAE \downarrow$	$\mathcal{S}_m \uparrow$	$E_\xi \uparrow$	$MAE \downarrow$
VST$_{21}$	.896	.892	.037	.850	.861	.058	.928	.953	.029	.932	.918	.033	.865	.837	.061
BBRF$_{21}$	.908	.927	.025	.855	.887	**.042**	**.935**	.965	.020	.939	.934	.022	.871	.867	.049
EVPv1$_{23}$	.913	.947	.026	.862	.894	.046	.931	.961	.024	.935	.957	.027	.878	.917	.054
EVPv2$_{23}$	.915	.948	.027	.862	.895	.047	.932	.963	.023	.935	.957	.028	.879	.917	.053
MENet$_{23}$	.905	.937	.028	.850	.891	.045	.927	.966	.023	.928	.954	.030	872	.913	.054
SelfReformer$_{23}$	.921	.924	.024	.859	.884	.043	.934	.961	.023	.941	.935	.025	.877	.874	.049
FOCUS(Ours)	**.929**	**.965**	**.019**	**.868**	**.900**	.045	**.935**	**.974**	**.018**	**.943**	**.971**	**.018**	**.898**	**.942**	**.036**

Table 2: Comparison of FOCUS with recent state-of-the-art SOD methods.

	ISTD(540)
	$BER \downarrow$
BDRAR$_{18}$	2.69
DSD$_{19}$	2.17
MTMT$_{20}$	1.72
FDRNet$_{21}$	1.55
EVPv1$_{23}$	1.35
EVPv2$_{23}$	1.35
SILT$_{23}$	1.11
FOCUS(Ours)	**0.98**

(a) SD

	DUT(500)		CUHK(100)	
	$F_\beta \uparrow$	$MAE \downarrow$	$F_\beta \uparrow$	$MAE \downarrow$
DeFusionNet$_{20}$	.823	.118	.818	.117
CENet$_{19}$	.817	.135	.906	.059
DAD$_{21}$	.794	.153	.884	.079
EFENet$_{21}$	.854	.094	.914	.053
DD$_{20}$	.891	.073	.927	.042
EVPv1$_{23}$	.890	.068	.928	.045
EVPv2$_{23}$	.887	.070	.932	.042
FOCUS(Ours)	**.912**	**.048**	**.934**	**.036**

(b) DBD

	CASIA-1.0(921)	
	$F1 \uparrow$	$AUC \uparrow$
ManTra$_{19}$	-	.817
SPAN$_{20}$	.382	.838
PSCCNet$_{22}$	.554	.875
TransForensics$_{21}$	.627	.837
EVPv1$_{23}$	.636	.862
EVPv2$_{23}$	.654	.876
ObjectFormer$_{22}$	.579	.882
FOCUS(Ours)	**.892**	**.940**

(c) FD

Table 3: Comparison of FOCUS with recent state-of-the-art SD, DBD, and FD methods.

Training Objectives

In order to perform foreground and background segmentation jointly, we convert the foreground segmentation dataset into binary form, with the white areas representing the foreground ground truth and the black areas representing the background ground truth. Following (Cheng et al. 2022) , we use the combination of binary cross entropy ($\mathcal{L}_{\text{bce}}$) and dice loss ($\mathcal{L}_{\text{dice}}$) as the loss of the mask, where:

$$\mathcal{L}_{\text{mask}} = \mathcal{L}_{\text{bce}} + \mathcal{L}_{\text{dice}}$$

Recent study (Li et al. 2023) shows that parallel execution of object detection and segmentation can benefit each other. In this paper, we use the rectangular boundary of the ground truth mask as the ground truth bounding box to perform object detection. We use combination of the L1 Regression Loss ($\mathcal{L}_{\text{L1}}$) and generalized IoU loss ($\mathcal{L}_{\text{gIoU}}$) as the loss for $\mathcal{L}_{\text{bbox}}$, which can be formulated as:

$$\mathcal{L}_{\text{bbox}} = \alpha\mathcal{L}_{\text{L1}} + \beta\mathcal{L}_{\text{gIoU}}$$

α and β are set to 5.0 and 2.0 respectively. We use the standard cross entropy loss as the $\mathcal{L}_{\text{label}}$. The final training objective is defined as follows:

$$\mathcal{L} = \lambda_{\text{clip}}\mathcal{L}_{\text{clip}} + \lambda_{\text{label}}\mathcal{L}_{\text{label}} + \lambda_{\text{mask}}\mathcal{L}_{\text{mask}} + \lambda_{\text{bbox}}\mathcal{L}_{\text{bbox}}$$

here, λ_{clip}, λ_{label}, λ_{mask}, λ_{bbox} refer to the weight of corresponding loss, set to 1.0, 1.0, 5.0, 1.0 respectively. To find the allocation with the lowest cost, we use Hungarian matching (Carion et al. 2020; Cheng, Schwing, and Kirillov 2021) between the prediction and the ground truth.

Experiments

Datasets and Evaluation Metrics

For COD, we follow (Fan et al. 2021; Zheng et al. 2024a), training FOCUS on the combination of CAMO-TR (Le et al.

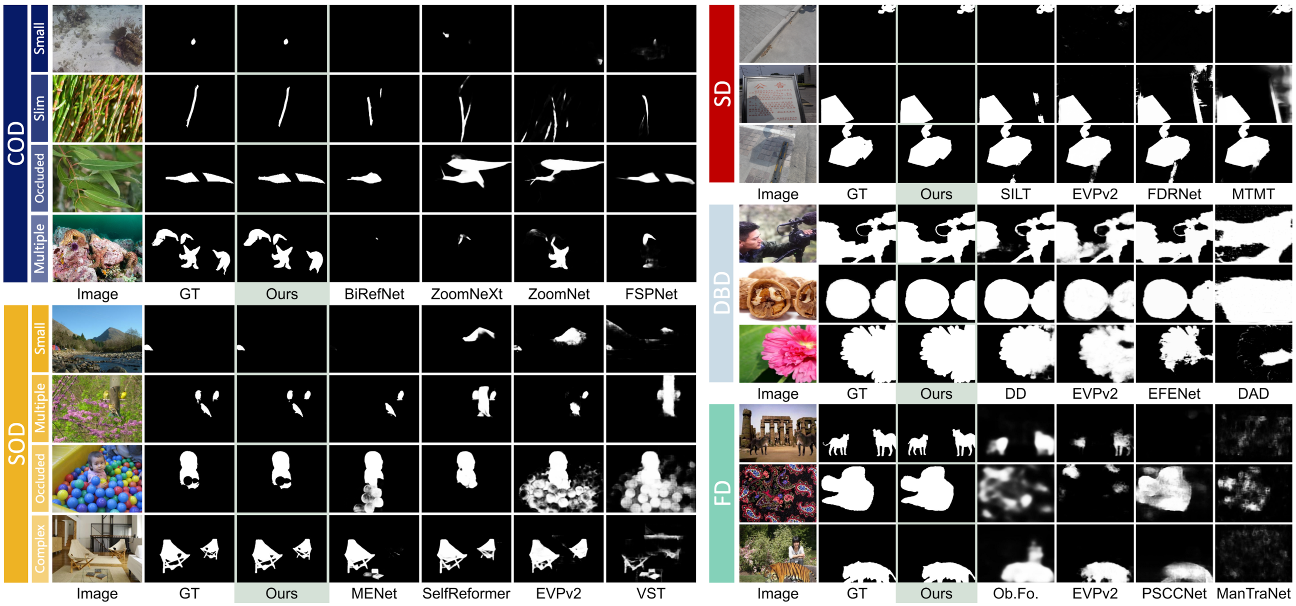

Figure 3: Qualitative comparison of FOCUS and previous methods on COD, SOD, SD, DBD, and FD. Zoom in for more details.

2019) and COD10K-TR (Fan et al. 2020) and evaluating on CAMO-TE, COD10K-TE, CHAMELEON (Skurowski et al. 2018) and NC4K (Lv et al. 2021). We use S-measure ($\mathcal{S}_m$), E-measure (E_ξ), weighted F-measure (F_β^w) and mean absolute error (MAE) to evaluate FOCUS.

For SOD task, we follow (Wang et al. 2023), using DUTS-TR (Wang et al. 2017) as training dataset without extra data, evaluating our model on DUTS-TE, DUT-OMRON (Yang et al. 2013), HKU-IS (Li and Yu 2015), ECSSD (Shi et al. 2015) and PACAL-S (Li et al. 2014) respectively. We use $\mathcal{S}_m$, E_ξ, MAE as evaluation metrics for SOD.

For SD, We use ISTD (Wang, Li, and Yang 2018) as our training and evaluation dataset and use balanced error rate (BCE) as the metric. For DBD, following previous work (Zhao et al. 2018), we use the combination of CUHK (Shi, Xu, and Jia 2014) and DUT (Zhao et al. 2018) as training dataset, and the remaining 100 images in CUHK and 500 images in DUT for testing, and we use F-measure (F_β) and MAE as metrics. Following (Wang et al. 2022a), we use CASIA-2.0 (Dong, Wang, and Tan 2013) as the training dataset and evaluate on CASIA-1.0, using pixel-level $F1$ score and area under the curve (AUC) as evaluation metric.

Implementation Details

We use batch size 8 for all experiments and 2 NVIDIA A6000 GPUs with 48G memory. The FOCUS is trained on each training dataset with the size of 512 × 512 for 20,000 iterations on average with AdamW optimizer (Loshchilov and Hutter 2017). The initial learning rate is set to 10^{-5} with a weight decay of 0.05 to regularize the model. The L2 norm is used for gradient clipping, and the maximum allowed value for gradients is set to 0.01. We use DINOv2-G (Oquab et al. 2023) pre-trained on ADE20K (Zhou et al. 2017) as the backbone for our SoTA model. Our framework is implemented using PyTorch 2.1.1 (Paszke et al. 2019).

Main Results

Comparison of the state-of-the-art task-specific methods. We compare our proposed FOCUS with recent models for COD including SINet (Fan et al. 2020) , PFNet (Mei et al. 2021) , ZoomNet (Pang et al. 2022b) , BSA-Net (Zhu et al. 2022) , FSPNet (Huang et al. 2023) , ZoomNeXt (Pang et al. 2024b) and BiRefNet (Zheng et al. 2024b) , models for SOD including MENet (Wang et al. 2023) , SelfReformer (Yun and Lin 2023) , BBRF (Ma et al. 2021) , and VST (Liu et al. 2021) , models for SD task including BDRAR (Zhu et al. 2018) , DSD (Zheng et al. 2019) , MTMT (Chen et al. 2020), FDRNet (Zhu et al. 2021) and SILT (Yang et al. 2023), models for DBD including DeFusionNet (Tang et al. 2020) , CENet (Zhao et al. 2019) , DAD (Zhao, Shang, and Lu 2021) , EFENet (Zhao et al. 2021) and DD (Cun and Pun 2020) , and models for FD including ManTra (Wu, AbdAlmageed, and Natarajan 2019) , SPAN (Hu et al. 2020) , PSCCNet (Liu et al. 2022) , TransForensics (Hao et al. 2021) and ObjectFormer (Wang et al. 2022a). FOCUS outperforms these SoTA models on most metrics across 13 datasets covering 5 tasks. Table. 1-3 shows quantitative comparisons between our proposed FOCUS with the previous SoTA models. Qualitative comparisons are in Fig . 3.

In the most challenging foreground segmentation task, COD, which requires the model to recognize the object blending in its surroundings, FOCUS outperforms the existing SoTA methods on most metrics across four mainstream datasets. For SOD tasks, FOCUS exceeds the task-specific models on almost all metrics, especially increasing in terms of E_ξ by an average of 1.8%. In SD tasks, FOCUS dramatically outperforms the previous SoTA on the ISTD dataset, with a 10.3% decrease in BER. In the DBD task, FOCUS surpasses the previous SoTA by a 2.1% increase on F_β on

id	Variants	Backbone	Trainable Param.	Module/Method				COD				SOD			
				JP	CR	EE	PR	$\mathcal{S}_m \uparrow$	$E_\xi \uparrow$	$F_\beta^w \uparrow$	$MAE \downarrow$	$\mathcal{S}_m \uparrow$	$E_\xi \uparrow$	$F_\beta \uparrow$	$MAE \downarrow$
0	**Baseline**	DINOv2-L	0.3G					.853	.931	.825	.043	.851	.919	.849	.054
1	**FOCUS**	DINOv2-L	0.3G	✓				.854	.931	.827	.042	.853	.923	.847	.054
2	**FOCUS**	DINOv2-L	0.3G	✓	✓			.861	.938	.836	.041	.855	.926	.851	.052
3	**FOCUS**	DINOv2-L	0.3G	✓	✓	✓		.872	.937	.848	.041	.870	.922	.864	.051
4	**FOCUS**	DINOv2-G ◇	0.1G	✓	✓	✓		.905	.956	.897	.027	.893	.936	.889	.039
5	**FOCUS**	DINOv2-G	1.2G	✓	✓	✓		.909	.962	.901	.026	.897	.942	.896	.037
6	**FOCUS**	DINOv2-G	1.2G	✓	✓	✓	✓	**.909**	**.963**	**.903**	**.025**	**.898**	**.943**	**.894**	**.037**

Table 4: Ablation study results of the proposed modules or methods of FOCUS, including CLIP Refiner (CR), Jointly Prediction (JP), Edge Enhancer (EE), and Pretrain (PR). ◇ means training with the DINOv2 backbone frozen.

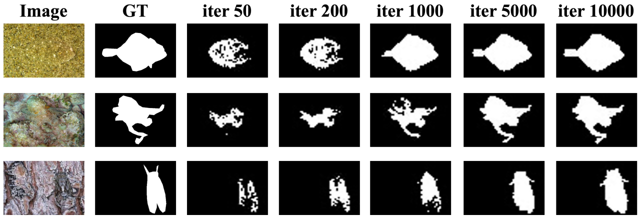

Figure 4: The visualization of PCA-based dimensionality reduction on the feature maps across different iterations.

DUT. In FD tasks, FOCUS also significantly surpasses previous SoTA models, with a 23.8% increase on $F1$ and a 3.8% increase on AUC.

Comparison of the state-of-the-art unified methods. As previously mentioned, there is a lack of unified architecture to handle all foreground tasks. To the best of our knowledge, EVPv1 and EVPv2 (Liu et al. 2023) are the most comparable works to our FOCUS in unifying foreground tasks. To demonstrate the superiority of FOCUS as a unified framework, we conducted extensive experiments comparing it with EVPv1 and EVPv2 across multiple datasets. Our results show that FOCUS consistently outperforms EVPv1 and EVPv2 in all metrics. This highlights the capability of FOCUS to handle a wide range of foreground segmentation tasks effectively, proving it can serve as a versatile and powerful model compared to existing unified methods.

Ablation Study

In this section, we conduct ablation experiments to analyze the properties of FOCUS. We use Mask2Former equipped with the DINOv2-L backbone as a robust baseline and choose the most representative foreground segmentation tasks, COD and SOD, as the ablation tasks. We select the mainstream dataset CAMO and PASCAL-S for COD and SOD respectively. To ensure consistency, all experiments were conducted using the same training recipe, with a batch size of 2. The training iterations are set to 10,000 for COD and 20,000 for SOD. Quantitative results related to each module or method are shown in Table. 4.

As shown in the table, variants of FOCUS with the CLIP refiner perform better than those without it, thanks to the multi-modal knowledge distilled from CLIP. We set the variants with joint prediction to perform foreground segmentation and background segmentation jointly, the comparison with the baseline shows that it can slightly improve the performance of FOCUS. Additionally, with the help of the edge enhancer to inject edge information of the object into the backbone image feature, the performance of variants of DINOv2 significantly improves in the provided metrics. We also evaluate the effectiveness of pretraining on ADE20K, which demonstrates modest improvements.

We use DINOv2-G as the backbone for our SoTA models, which inevitably results in a large number of parameters. To ensure a fair comparison, we freeze the DINOv2-G backbone, limiting the number of trainable parameters in our model to 0.1G. The results indicate a slight decrease in performance compared to the fully fine-tuned version. However, when compared to models like BiRefNet (215M) and SelfReformer (˜220M), the frozen-backbone FOCUS still matches or surpasses previous state-of-the-art performance, despite having fewer trainable parameters.

We initialize the first layer of the transformer decoder with PCA-reduced feature maps from the backbone in our paper. As shown in Fig. 4, these PCA-reduced feature maps begin to exhibit strong semantic features in the early training stages. As training progresses, we are pleasantly surprised to find that even without further forward propagation, the patch-level feature maps, simply reduced by PCA, are able to approach the ground truth quality. Using them for initialization, compared to random initialization, provides a valuable spatial prior for subsequent mask attention.

Conclusion

In this paper, we propose FOCUS, a unified multi-modal approach to tackle multiple subdivision tasks of foreground segmentation. We leverage the concept of object queries to handle foreground segmentation tasks and develop a multi-scale semantic network that simultaneously performs foreground and background segmentation, fully utilizing the background information of the image to optimize prediction. We also introduced a novel distillation method integrating the contrastive learning strategy to enhance boundary-aware foreground segmentation. Theoretically, our model can be extended to any foreground segmentation task. Extensive experiments conducted on diverse datasets demonstrate the effectiveness of our proposed framework.

Acknowledgments

This project was supported by NSFC under Grant No. 62102092.

References

Canny, J. F. 1986. A computational approach to edge detection. *TPAMI.*

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-end object detection with transformers. In *ECCV*.

Chen, Z.; Duan, Y.; Wang, W.; He, J.; Lu, T.; Dai, J.; and Qiao, Y. 2022. Vision transformer adapter for dense predictions. *arXiv preprint arXiv:2205.08534*.

Chen, Z.; Zhu, L.; Wan, L.; Wang, S.; Feng, W.; and Heng, P.-A. 2020. A multi-task mean teacher for semi-supervised shadow detection. In *CVPR*.

Cheng, B.; Misra, I.; Schwing, A. G.; Kirillov, A.; and Girdhar, R. 2022. Masked-attention mask transformer for universal image segmentation. In *CVPR*.

Cheng, B.; Schwing, A.; and Kirillov, A. 2021. Per-pixel classification is not all you need for semantic segmentation. *NeurIPS*.

Cun, X.; and Pun, C.-M. 2020. Defocus blur detection via depth distillation. In *ECCV*.

Davies, E. R. 2004. *Machine vision: theory, algorithms, practicalities*. Elsevier.

Ding, H.; Liu, C.; He, S.; Jiang, X.; and Loy, C. C. 2023a. MeViS: A Large-scale Benchmark for Video Segmentation with Motion Expressions. In *ICCV*.

Ding, H.; Liu, C.; He, S.; Jiang, X.; Torr, P. H.; and Bai, S. 2023b. MOSE: A new dataset for video object segmentation in complex scenes. In *ICCV*.

Dong, J.; Wang, W.; and Tan, T. 2013. Casia image tampering detection evaluation database. In *ChinaSIP*.

Fan, D.-P.; Ji, G.-P.; Cheng, M.-M.; and Shao, L. 2021. Concealed object detection. *TPAMI.*

Fan, D.-P.; Ji, G.-P.; Sun, G.; Cheng, M.-M.; Shen, J.; and Shao, L. 2020. Camouflaged object detection. In *CVPR*.

Hao, J.; Zhang, Z.; Yang, S.; Xie, D.; and Pu, S. 2021. Transforensics: image forgery localization with dense self-attention. In *ICCV*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*.

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

Hu, X.; Zhang, Z.; Jiang, Z.; Chaudhuri, S.; Yang, Z.; and Nevatia, R. 2020. SPAN: Spatial pyramid attention network for image manipulation localization. In *ECCV*.

Huang, Z.; Dai, H.; Xiang, T.-Z.; Wang, S.; Chen, H.-X.; Qin, J.; and Xiong, H. 2023. Feature shrinkage pyramid for camouflaged object detection with transformers. In *CVPR*.

Jain, J.; Li, J.; Chiu, M.; Hassani, A.; Orlov, N.; and Shi, H. 2023. OneFormer: One Transformer to Rule Universal Image Segmentation. In *CVPR*.

Kirillov, A.; He, K.; Girshick, R.; Rother, C.; and Dollár, P. 2019. Panoptic Segmentation. In *CVPR*.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; Dollar, P.; and Girshick, R. 2023. Segment Anything. In *ICCV*.

Le, T.-N.; Nguyen, T. V.; Nie, Z.; Tran, M.-T.; and Sugimoto, A. 2019. Anabranch network for camouflaged object segmentation. *CVIU*.

Li, B.; Weinberger, K. Q.; Belongie, S.; Koltun, V.; and Ranftl, R. 2022. Language-driven semantic segmentation. *arXiv preprint arXiv:2201.03546*.

Li, F.; Zhang, H.; Xu, H.; Liu, S.; Zhang, L.; Ni, L. M.; and Shum, H.-Y. 2023. Mask dino: Towards a unified transformer-based framework for object detection and segmentation. In *CVPR*.

Li, G.; and Yu, Y. 2015. Visual saliency based on multiscale deep features. In *CVPR*.

Li, Y.; Hou, X.; Koch, C.; Rehg, J. M.; and Yuille, A. L. 2014. The secrets of salient object segmentation. In *CVPR*.

Liang, F.; Wu, B.; Dai, X.; Li, K.; Zhao, Y.; Zhang, H.; Zhang, P.; Vajda, P.; and Marculescu, D. 2023. Open-vocabulary semantic segmentation with mask-adapted clip. In *CVPR*.

Liu, N.; Zhang, N.; Wan, K.; Shao, L.; and Han, J. 2021. Visual saliency transformer. In *ICCV*.

Liu, W.; Shen, X.; Pun, C.-M.; and Cun, X. 2023. Explicit visual prompting for low-level structure segmentations. In *CVPR*.

Liu, X.; Liu, Y.; Chen, J.; and Liu, X. 2022. PSCC-Net: Progressive spatio-channel correlation network for image manipulation detection and localization. *TCSVT*.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Lv, Y.; Zhang, J.; Dai, Y.; Li, A.; Liu, B.; Barnes, N.; and Fan, D.-P. 2021. Simultaneously localize, segment and rank the camouflaged objects. In *CVPR*.

Ma, M.; Xia, C.; Xie, C.; Chen, X.; and Li, J. 2021. Receptive field broadening and boosting for salient object detection. *arXiv preprint arXiv:2110.07859*.

Mei, H.; Ji, G.-P.; Wei, Z.; Yang, X.; Wei, X.; and Fan, D.-P. 2021. Camouflaged object segmentation with distraction mining. In *CVPR*.

Meng, L.; Dai, X.; Yang, J.; Chen, D.; Chen, Y.; Liu, M.; Chen, Y.-L.; Wu, Z.; Yuan, L.; and Jiang, Y.-G. 2024. Learning from rich semantics and coarse locations for long-tailed object detection. *NeurIPS*, 36.

Meng, L.; Li, H.; Chen, B.-C.; Lan, S.; Wu, Z.; Jiang, Y.-G.; and Lim, S.-N. 2022. Adavit: Adaptive vision transformers for efficient image recognition. In *CVPR*, 12309–12318.

Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H. V.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; Howes, R.; Huang, P.-Y.; Xu, H.; Sharma, V.; Li, S.-W.; Galuba, W.; Rabbat, M.; Assran, M.; Ballas, N.; Synnaeve, G.; Misra, I.; Jegou, H.; Mairal, J.;

Labatut, P.; Joulin, A.; and Bojanowski, P. 2023. DINOv2: Learning Robust Visual Features without Supervision.

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

Pang, Y.; Zhao, X.; Xiang, T.-Z.; Zhang, L.; and Lu, H. 2022b. Zoom in and out: A mixed-scale triplet network for camouflaged object detection. In *CVPR*.

Pang, Y.; Zhao, X.; Xiang, T.-Z.; Zhang, L.; and Lu, H. 2024a. ZoomNeXt: A Unified Collaborative Pyramid Network for Camouflaged Object Detection. *TPAMI*.

Pang, Y.; Zhao, X.; Xiang, T.-Z.; Zhang, L.; and Lu, H. 2024b. Zoomnext: A unified collaborative pyramid network for camouflaged object detection. *TPAMI*.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *NeurIPS*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *ICCV*.

Shi, J.; Xu, L.; and Jia, J. 2014. Discriminative blur detection features. In *CVPR*.

Shi, J.; Yan, Q.; Xu, L.; and Jia, J. 2015. Hierarchical image saliency detection on extended CSSD. *TPAMI*.

Skurowski, P.; Abdulameer, H.; Błaszczyk, J.; Depta, T.; Kornacki, A.; and Kozieł, P. 2018. Animal camouflage analysis: Chameleon database. *Unpublished manuscript*.

Tang, C.; Liu, X.; Zheng, X.; Li, W.; Xiong, J.; Wang, L.; Zomaya, A. Y.; and Longo, A. 2020. DeFusionNET: Defocus blur detection via recurrently fusing and refining discriminative multi-scale deep features. *TPAMI*.

Wang, J.; Li, X.; and Yang, J. 2018. Stacked conditional generative adversarial networks for jointly learning shadow detection and shadow removal. In *CVPR*.

Wang, J.; Wu, Z.; Chen, J.; Han, X.; Shrivastava, A.; Lim, S.-N.; and Jiang, Y.-G. 2022a. Objectformer for image manipulation detection and localization. In *CVPR*.

Wang, J.; Wu, Z.; Ouyang, W.; Han, X.; Chen, J.; Jiang, Y.-G.; and Li, S.-N. 2022b. M2tr: Multi-modal multi-scale transformers for deepfake detection. In *ICMR*.

Wang, L.; Lu, H.; Wang, Y.; Feng, M.; Wang, D.; Yin, B.; and Ruan, X. 2017. Learning to detect salient objects with image-level supervision. In *CVPR*.

Wang, W.; Xie, E.; Li, X.; Fan, D.-P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; and Shao, L. 2022c. Pvt v2: Improved baselines with pyramid vision transformer. *CVM*.

Wang, Y.; Wang, R.; Fan, X.; Wang, T.; and He, X. 2023. Pixels, Regions, and Objects: Multiple Enhancement for Salient Object Detection. In *CVPR*.

Wang, Z.; Lu, Y.; Li, Q.; Tao, X.; Guo, Y.; Gong, M.; and Liu, T. 2022d. Cris: Clip-driven referring image segmentation. In *CVPR*.

Wu, Y.; AbdAlmageed, W.; and Natarajan, P. 2019. Mantranet: Manipulation tracing network for detection and localization of image forgeries with anomalous features. In *CVPR*.

Xie, C.; Xia, C.; Ma, M.; Zhao, Z.; Chen, X.; and Li, J. 2022. Pyramid Grafting Network for One-Stage High Resolution Saliency Detection. In *CVPR*.

Xu, J.; De Mello, S.; Liu, S.; Byeon, W.; Breuel, T.; Kautz, J.; and Wang, X. 2022. Groupvit: Semantic segmentation emerges from text supervision. In *CVPR*.

Yang, C.; Zhang, L.; Lu, H.; Ruan, X.; and Yang, M.-H. 2013. Saliency detection via graph-based manifold ranking. In *CVPR*.

Yang, H.; Wang, T.; Hu, X.; and Fu, C.-W. 2023. Silt: Shadow-aware iterative label tuning for learning to detect shadows from noisy labels. In *ICCV*.

Yun, Y. K.; and Lin, W. 2023. Towards a Complete and Detail-Preserved Salient Object Detection. *TMM*.

Zhao, W.; Hou, X.; He, Y.; and Lu, H. 2021. Defocus blur detection via boosting diversity of deep ensemble networks. *TIP*.

Zhao, W.; Shang, C.; and Lu, H. 2021. Self-generated defocus blur detection via dual adversarial discriminators. In *CVPR*.

Zhao, W.; Zhao, F.; Wang, D.; and Lu, H. 2018. Defocus blur detection via multi-stream bottom-top-bottom fully convolutional network. In *CVPR*.

Zhao, W.; Zheng, B.; Lin, Q.; and Lu, H. 2019. Enhancing diversity of defocus blur detectors via cross-ensemble network. In *CVPR*.

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

Zheng, P.; Gao, D.; Fan, D.-P.; Liu, L.; Laaksonen, J.; Ouyang, W.; and Sebe, N. 2024b. Bilateral Reference for High-Resolution Dichotomous Image Segmentation. *arXiv preprint arXiv:2401.03407*.

Zheng, Q.; Qiao, X.; Cao, Y.; and Lau, R. W. 2019. Distraction-aware shadow detection. In *CVPR*.

Zhou, B.; Zhao, H.; Puig, X.; Fidler, S.; Barriuso, A.; and Torralba, A. 2017. Scene Parsing through ADE20K Dataset. In *CVPR*.

Zhu, H.; Li, P.; Xie, H.; Yan, X.; Liang, D.; Chen, D.; Wei, M.; and Qin, J. 2022. I can find you! boundary-guided separated attention network for camouflaged object detection. In *AAAI*.

Zhu, L.; Deng, Z.; Hu, X.; Fu, C.-W.; Xu, X.; Qin, J.; and Heng, P.-A. 2018. Bidirectional feature pyramid network with recurrent attention residual modules for shadow detection. In *ECCV*.

Zhu, L.; Xu, K.; Ke, Z.; and Lau, R. W. 2021. Mitigating intensity bias in shadow detection via feature decomposition and reweighting. In *ICCV*.

Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; and Dai, J. 2020. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*.