Expert Systems With Applications 274 (2025) 126912

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Seamless Detection: Unifying Salient Object Detection and Camouflaged Object Detection

Yi Liu [a], Chengxin Li [a], Xiaohui Dong [a], Lei Li [b], Dingwen Zhang [c,1], Shoukun Xu [a], Jungong Han [d,1]

[a] School of Computer Science and Artificial Intelligence, Changzhou University, Changzhou 213159, China
[b] Science and Technology on Complex System Control and Intelligent Agent Cooperation Laboratory, Beijing, China
[c] School of Automation, Northwestern Polytechnical University, Xi'an 710129, China
[d] Department of Automation, Tsinghua University, Beijing 100084, China

ARTICLE INFO

Keywords:
Salient object detection
Camouflaged object detection
Contrastive learning

ABSTRACT

Achieving joint learning of Salient Object Detection (SOD) and Camouflaged Object Detection (COD) is extremely challenging due to their distinct object characteristics, *i.e.*, saliency and camouflage. The only preliminary research treats them as two contradictory tasks, training models on large-scale labelled data alternately for each task and assessing them independently. However, such task-specific mechanisms fail to meet real-world demands for addressing unknown tasks effectively. To address this issue, in this paper, we pioneer a task-agnostic framework to unify SOD and COD. To this end, inspired by the agreeable nature of binary segmentation for SOD and COD, we propose a **C**ontrastive **D**istillation **P**aradigm (**CDP**) to distil the foreground from the background, facilitating the identification of salient and camouflaged objects amidst their surroundings. To probe into the contribution of our CDP, we design a simple yet effective contextual decoder involving the interval-layer and global context, which achieves an inference speed of **67 fps**. Besides the supervised setting, our CDP can be seamlessly integrated into unsupervised settings, eliminating the reliance on extensive human annotations. Experiments on public SOD and COD datasets demonstrate the superiority of our proposed framework in both supervised and unsupervised settings, compared with existing state-of-the-art approaches. Code is available on https://github.com/liuyi1989/Seamless-Detection.

1. Introduction

Visually Salient Object Detection (SOD) and Camouflaged Object Detection (COD), which are fundamental image segmentation tasks in the computer vision community, have garnered significant research attention. SOD targets detecting attracting-attention objects that are usually stand out from their surroundings, while COD aims to discover objects that are usually concealed in their surroundings (Liu, Zhang, Zhang and Han, 2021). Due to the object characteristics of SOD and COD, they have been embedded separately in large-scale applications, such as SOD for adversarial defense (Li, Hu, Lin, Wu, & Shen, 2023; Yu, Lu, Li, & Zhou, 2021), and COD for traffic classification (Zhang, Wang et al., 2024) and surveillance (Fang, Liu et al., 2023).

In real-world scenarios, objects often exhibit multiple characteristics simultaneously, such as saliency and camouflage. As shown in Fig. 1, a chameleon is salient when it appear in a new scene. However, it will change its body appearance to conceal itself in the surroundings, which makes it camouflaged. These qualities play crucial roles in various practical applications, including autonomous driving where detecting both salient features and camouflaged objects is essential for safety (Lateef, Kas, & Ruichek, 2021; Song et al., 2022), and in remote sensing where identifying saliency and camouflage aids in data analysis (Dehmollaian & Sarabandi, 2006; Diao et al., 2016). Despite significant advancements in separate SOD and COD techniques, they are confined to specific tasks. Since the object characteristics cannot be forewarned in real-world life, the current separate SOD and COD methods will fail in the opposite tasks. For example, it is not reasonable to detect the chameleon in Fig. 1 using the individual salient or camouflaged object detection model. Although Li et al. (2021) makes the joint learning of SOD and COD, its training and testing are still task-specific, which cannot solve the un-forewarned case. Moreover, when presented with an image, it is challenging for ordinary users to discern saliency and camouflage. Rather, their primary goal is to identify the interesting or useful objects within the scene, regardless of whether the scene features saliency or camouflage. Therefore, these two

E-mail addresses: zhangdingwen2006yyy@gmail.com (D. Zhang), jungonghan77@gmail.com (J. Han).
[1] Equally corresponding authors.

https://doi.org/10.1016/j.eswa.2025.126912
Received 20 December 2024; Received in revised form 7 February 2025; Accepted 13 February 2025
Available online 24 February 2025
0957-4174/

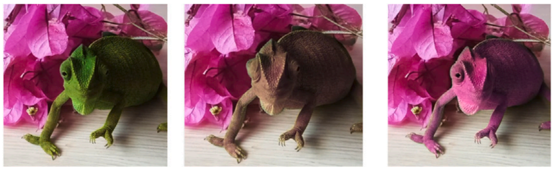

Fig. 1. An easy example for the co-existing saliency and camouflage scene. A chameleon is salient when it appear in a new scene. However, it will change its body appearance to conceal itself in the surroundings, which makes it camouflaged. For the event, it is not reasonable to detect the chameleon using the individual salient or camouflaged object detection model. Inspired by this observation, it is necessary to design a task-agnostic model unifying the abilities of saliency and camouflaged detection.

real-world challenges makes the urgent of unifying SOD and COD in a task-agnostic framework. To solve this urgent problem, in this paper, *we focus on the task-agnostic unified framework of SOD and COD towards the real-world practicability*.

Thanks to the large-scale pixel-level annotated benchmarks, *e.g.*, DUTS (Wang et al., 2017) and COD10K (Fan et al., 2020), deep supervised research in both SOD and COD has advanced significantly in separate avenues. However, the joint learning of SOD and COD is still in its infancy. The work Li et al. (2021), named UJSC, makes specialized study towards this issue via treating SOD and COD as two contradictory tasks, which is implemented by two branches of encoders and decoders that share the structures but not parameters. As shown in Fig. 2, the procedure of UJSC (Li et al., 2021) can be described as (i) Besides the SOD image and COD image in the training stage, extra general object images are required for the contradiction measure to regularize SOD and COD encoders; (ii) UJSC (Li et al., 2021) manually labels SOD as 0 and COD as 1 to train the uncertainty learning separately; (iii) UJSC (Li et al., 2021) trains SOD and COD branches using SOD and COD datasets separately in an alternate manner; (iv) UJSC (Li et al., 2021) infers the SOD and COD task separately. While these steps represent significant progress for UJSC (Li et al., 2021), they make UJSC (Li et al., 2021) a task-specific model, failing to meet the task-agnostic requirement in the real-world unforewarned case.

In this paper, we make an exploration into the task-agnostic unified framework for SOD and COD by leveraging the inherent commonality between them. Traditionally, humans possess the natural ability to distinguish targets from their surroundings, whether the targets are salient or camouflaged. This highlights the shared essence of SOD and COD tasks: identifying the foreground from the background. Inspired by this shared essence, we use such agreeable nature of binary segmentation to design the task-agnostic unified framework. Specifically, we propose a **C**ontrastive **D**istillation **P**aradigm (**CDP**) to distil the target from the background. As shown in our framework of Fig. 2, the SOD and COD tasks share the same encoder and decoder in terms of structures and parameters, involving one input for mixed SOD & COD images. During training, the foreground semantics and background semantics, derived from the decoder and encoder, respectively, are fed into the contrastive distillation framework, which is learned by the contrastive loss. During test, both SOD and COD tasks are inferred using the same encoder and decoder with the same parameters. Compared with UJSC (Li et al., 2021), our CDP has three advantages: (i) No extra general object images avoid the contradiction measure; (ii) Only one input for mixed SOD & COD images ensures the proposed framework task-agnostic in the training and testing stages, making it adaptable to real-world applications; (iii) Our CDP can be plugged in the existing image segmentation models. To discuss the effectiveness of our CDP, we design simple but effective encoder and decoder, which produces a real-time inference with 67 fps. To be concrete, we opt for ResNet-50 (He, Zhang, Ren, & Sun, 2016) as our encoder choice. The decoder is designed using the integration of **I**nterval-layer and **G**lobal **C**ontext (**IGC**) in which the interval layers and the deepest layer of the backbone are involved to be integrated to decode the image semantics.

Furthermore, UJSC (Li et al., 2021) tackles the joint learning of SOD and COD using large-scale pixel-level SOD and COD labelled benchmarks, demanding significant human effort, with each image label consuming approximately 60 min (Fan et al., 2020). Although there have been a lot of attempts for unsupervised SOD (Lin, Wu, Chen, Li, & Yu, 2022; Melas-Kyriazi, Rupprecht, Laina, & Vedaldi, 2022; Nguyen et al., 2019; Shin, Albanie, & Xie, 2022; Wang, Zhang, Wang, Liu and Lu, 2022; Zhang, Han and Zhang, 2017; Zhou, Chen, Yang, Xie and Lai, 2022; Zhou, Qiao, Yang, Lai, & Xie, 2023) and unsupervised COD (Zhang & Wu, 2023) separately, the joint unsupervised learning of SOD and COD is an undeveloped field, which is a demand in real-world life. Due to its versatility, our CDP extends beyond the confines of supervised learning, enabling the joint unsupervised learning of SOD and COD. Concretely, the deep features of DINO (Caron et al., 2021) are parsed to generate the initial pseudo mask for supervision at the first two epochs. In the following, our CDP absorbs the foreground semantics and pseudo background semantics for contrastive learning to distil the foreground objects. To train the model with high-quality labels, pseudo labels are updated for each epoch.

In general, unifying SOD and COD in a task-agnostic framework offers a more practical and user-friendly solution for real-world object detection, overcoming the limitations of existing separate SOD and COD techniques and aligning with human perception capabilities. Specific advantages lie in two folds. (i) This integration overcomes the limitations of existing methods that are confined to specific SOD and COD tasks, addressing the real-world challenge where objects often exhibit multiple characteristics, such as saliency and camouflage. By leveraging the common nature of foreground–background identification shared by SOD and COD, our framework is highly adaptable to complex and dynamic real-world environments, making it a more practical and reliable solution for detecting objects with diverse characteristics. (ii) Our framework aligns with the natural human ability to identify interesting or useful objects in a scene, regardless of whether they are salient or camouflaged. This user-centric approach simplifies the detection process for ordinary users, who may find it challenging to forewarn saliency and camouflage. By focusing on the universal goal of foreground–background segmentation, our method enhances the practicality and usability of object detection in various real-world applications.

Contributions of this paper are listed as follows:

(i) We make the study to unify SOD and COD in a task-agnostic framework via a contrastive distillation paradigm, inspired by their agreeable nature of binary segmentation.

(ii) By unifying SOD and COD on both supervised and unsupervised settings, we alleviate the need for extensive human annotations, thereby reducing the laborious task of large-scale dataset labelling.

(iii) Experiments on public SOD and COD benchmarks demonstrate that our task-agnostic framework achieves the competitive performance in the supervised setting and State-Of-The-Art (SOTA) performance in the unsupervised setting, compared with the previous task-specific methods.

The paper is organized as follows. Section 2 reviews the related work to the proposed method. Section 3 introduces details of the proposed method. Section 4 conducts a serious of experiments to understand the proposed method. Section 5 concludes the paper.

2. Related work

In this section, we will review the related works to our method, including supervised SOD and COD, unsupervised SOD and COD, and contrastive learning.

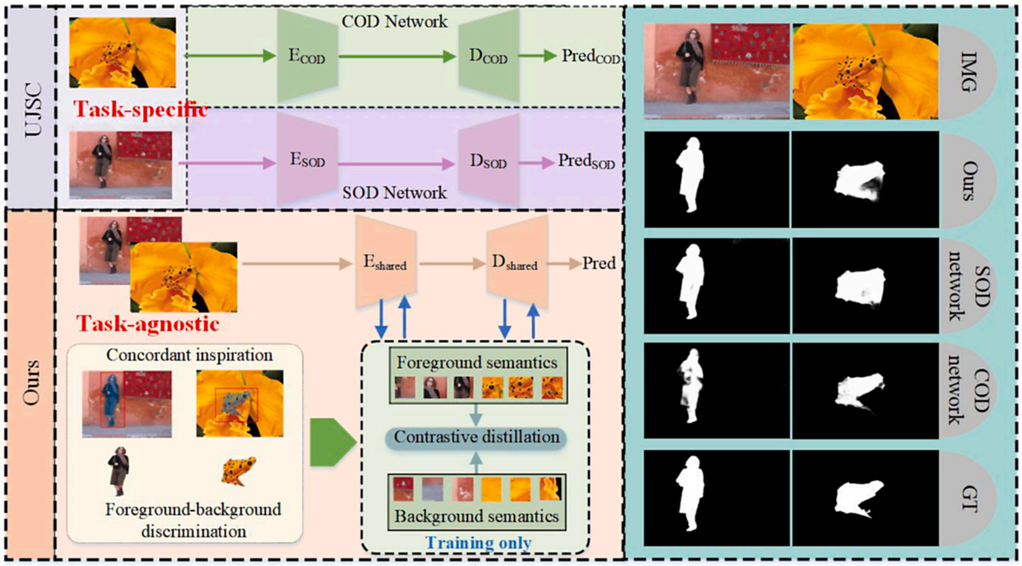

Fig. 2. Motivation statement. The previous UJSC (Li et al., 2021) is task-specific, which must fed salient and camouflaged image into SOD network and COD network correspondingly, otherwise generating poor results. This challenge can well be solved by our task-agnostic framework.

2.1. Supervised SOD and COD

Before deep learning (Krizhevsky, Sutskever, & Hinton, 2012), SOD and COD move forward relying on the hand-crafted methods (Borji, Cheng, Jiang, & Li, 2015; Jiang et al., 2013; Liu, Han, Zhang and Wang, 2018; Mondal, 2020). The emergence of deep learning has broken the performance bottleneck of these tasks, which has been a popular trend.

SOD. The development of deep SOD began in Han et al. (2014) that used a residual reconstruction network. Thereafter, deep learning has been largely developed for SOD (Wang et al., 2021) thanks to the large-scale pixel-level labels, using Muli-Layer Perceptron (MLP) classifiers (Li & Yu, 2015; Zhao, Ouyang, Li, & Wang, 2015), Fully Convolutional Network (FCN) (Chen, Tan, Wang, & Hu, 2018; Pang, Zhao, Zhang, & Lu, 2020; Siris, Jiao, Tam, Xie, & Lau, 2021; Tian, Zhang, Xiang, & Dai, 2023; Wang, Shen, Dong, & Borji, 2018; Zhang, Liu, Lu, & Shen, 2019; Zhang, Wang, Lu, Wang and Ruan, 2017; Zhao et al., 2019; Zhou et al., 2021), Capsule Networks (CapsNets) (Liu, Zhang, Liu, Xu, & Han, 2022; Liu, Zhang, Zhang, & Han, 2019, 2021c), and Transformer (Liu, Zhang, Wan, Shao and Han, 2021; Zhang, Xie, Barnes, & Li, 2021).

COD. Inspired by the hunting of predators, Fan et al. (2020) designed a search-identification network to detect the camouflaged object, which turns on the research on deep COD. Thereafter, a lot of attempts have been devoted to the development of deep COD. For example, Zhai et al. (2021) used graph learning towards COD. Zhong et al. (2022) detected the camouflaged object using the frequency domain knowledge. Pang, Zhao, Xiang, Zhang, and Lu (2022) introduced the multi-scale detection network for COD. Zhou, Zhou, Gong, Yang and Zhang (2022) introduced a boundary guidance mechanism in their feature aggregation and propagation network for COD, enhancing boundary-aware features to improve detection performance. Huang et al. (2023) designed a feature shrinkage pyramid architecture using Transformer to detect the camouflaged object. He et al. (2023) introduced the learnable wavelets towards the task of COD.

There are efforts towards the joint learning of SOD and COD. Li et al. (2021) implemented the joint learning of SOD and COD via learning the task contradiction and uncertainty. However, their task-specific model cannot solve the real-world unforewarned case, which can be solved well by our task-agnostic framework.

2.2. Unsupervised SOD and COD

Traditional deep SOD and COD usually rely on large-scale pixel-level labels, *e.g.*, DUTS (Wang et al., 2017) and COD10K (Fan et al., 2020), which consume huge labour. To solve this problem, unsupervised learning without human annotations is employed to implement SOD and COD.

Unsupervised SOD. Zhang, Han et al. (2017) opened the research of unsupervised SOD, which generated high-quality pseudo labels via discovering consistency from noisy traditional detectors (Shi, Yan, Xu, & Jia, 2015; Zhang & Sclaroff, 2015; Zhang et al., 2015). Following this route, there are a few works towards unsupervised SOD (Nguyen et al., 2019; Zhang, Zhang, Dai, Harandi, & Hartley, 2018). More recently, unsupervised SOD generates high-quality pseudo labels from high-level deep semantics. For example, Zhou, Chen et al. (2022) activated the multi-level semantics for high-quality labels generation to train the detector. Later on, they developed unsupervised SOD via mining saliency knowledge from easy and hard samples (Zhou et al., 2023). Shin et al. (2022) introduced spectral clustering to generate pseudo labels for unsupervised SOD.

Unsupervised COD. In Zhang and Wu (2023), a source-free unsupervised domain adaptation was introduced to solve the task of unsupervised COD.

There have few efforts to the joint unsupervised learning of SOD and COD. In this paper, we make the study for unified SOD and COD without human annotations.

2.3. Contrastive learning

Recently, contrastive learning, which learns general and robust feature representations by comparing similar and dissimilar pairs, has shown power in a lot of computer vision tasks. For example, Lo et al. (2021) learned better illuminant-dependent features for colour constancy via constructing contrastive pairs. Li et al. (2022) devised a targeted supervised contrastive learning framework to enhance the feature distribution uniformity for further image recognition in the case of long-tail data. Tang, Zhan, Chen, Yu, and Tao (2022) segmented point cloud objects using a contrastive boundary learning framework. Zhao, Vemulapalli et al. (2021) achieved semantic segmentation with limited labels using a contrastive training strategy. Zhang et al. (2023) solved the cross-modal animal pose estimation problem using a contrastive learning paradigm for the language knowledge and animal

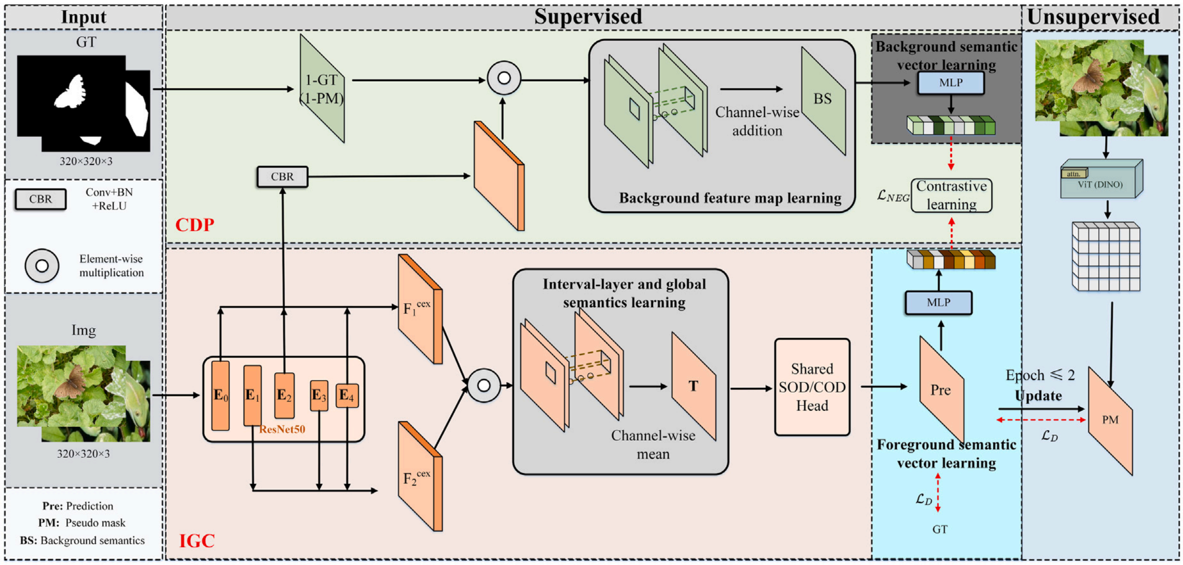

Fig. 3. Overview of the framework. $\mathbf{E}_*$ is the last layer of different blocks in ResNet-50 (He et al., 2016). $\mathcal{L}_D$ and $\mathcal{L}_{NEG}$ denote the training losses of Eqs. (8) and (9), respectively. Under the supervised setting, the foreground map inferred by IGC is supervised by ground truth. Besides, the foreground semantics and background semantics, generated by the decoder and encoder of IGC, respectively, are supervised using the contrastive loss within CDP. Under the unsupervised setting, the deep features of DINO (Caron et al., 2021) are parsed to generate the pseudo masks at the initial two epochs, which will be updated at each epoch. Note that, only IGC is run for inference at the test stage for both supervised and unsupervised settings.

pose images. Meng, Shao, Guo, and Gao (2023) combined the learning of self-contrast, cross-contrast, and ambiguity contrast for multi-object tracking.

In this paper, we used contrastive learning to make SOD and COD tasks in a unified paradigm via distilling the target from the background.

2.4. Distillation learning

Distillation learning is a widely used technique in the deep learning community, aimed at transferring knowledge from a teacher model to a student model. Recent studies have applied distillation learning to machine leaning and computer vision tasks. For instance, Xing, Xiao, Qu, Zhu, and Zhao (2022) employed a student–teacher framework and distance-based matching to efficiently transfer knowledge across multi-task time series classification models. Wu, Wu, Lyu, Huang, and Xie (2022) reduced communication costs by 94.89% in federated learning using adaptive mutual knowledge distillation and dynamic gradient compression. Chen et al. (2022) leveraged mutual knowledge distillation between visual and attribute features to enhance zero-shot learning. Guo et al. (2024) proposed pixel distillation for distilling knowledge across different image sizes and network architectures, optimizing resource usage for edge devices. Zhang, Han et al. (2024) introduced MHKD for glomerulus detection in low-resolution kidney images, enhancing feature extraction and accuracy through offline and online distillation. Fang, Wang et al. (2023) introduced a mutual distillation method to enhance medical image segmentation with imperfect annotations by using two models to clean noisy labels and exchange reliable knowledge. Zhao, Han, Yang, Wang and Zhang (2021) improved weakly supervised temporal action localization with background response and self-distillation learning.

Due to the significance of the distillation learning, in this paper, we extend this concept to SOD and COD tasks, which is implemented by incorporating contrastive learning to enhance foreground semantics with the help of background semantic cues.

3. Proposed method

In this section, we will illustrate the proposed method with clear details for technically understanding.

3.1. Overview

Fig. 3 overviews the proposed framework. It includes supervised and unsupervised settings. In the supervised setting, IGC infers the foreground map, which is supervised by ground truth. Besides, the foreground semantics and background semantics, generated by the decoder and encoder of IGC, respectively, are supervised using the contrastive loss within CDP. Under the unsupervised setting, The deep features of DINO (Caron et al., 2021) are parsed to generate the initial pseudo mask for supervision at the first two epochs, which will be updated at each epoch. IGC is run for inference at the test stage for both supervised and unsupervised settings.

3.2. Interval-layer and global contextual network

Context is an important semantic in deep neural networks. To design a simple but superior network to decode the image semantics for further inference, we dig into the context implicitly contained in the deep backbone, including interval-layer context and global context.

Suppose the image denoted as $\mathbf{I} \in \Re^{H\times W\times 3}$, which is fed into ResNet-50 (He et al., 2016) to extract multi-level deep features, denoted as $\mathbf{E}_i \in \Re^{\frac{H}{2^{i+1}}\times\frac{W}{2^{i+1}}\times C_i}, i \in [0,1,2,3,4], C_i \in [64,256,512,1024,2048]$. First, the interval-layer features, *i.e.*, $\mathbf{E}_0$ & $\mathbf{E}_2$ and $\mathbf{E}_1$ & $\mathbf{E}_3$, have different receptive fields with respect to the input image, concretely $2^2\times$ spatial difference. These interval-layer contexts will help to capture different-scale objects. Besides, the deepest-layer features, *i.e.*, $\mathbf{E}_4$ learns the high-level image semantics from the global perspective thanks to the $2^5\times$ receptive field. Based on these two findings, we expect to integrate the interval-layer contexts and the global context for the semantic exploration of the input image. Specifically, All $\mathbf{E}_i$s are transformed into the same-channel versions, *i.e.*,

$$\mathbf{F}_i \in \Re^{\frac{H}{2^{i+1}}\times\frac{W}{2^{i+1}}\times 64} = f_{conv}(\mathbf{E}_i), i = 0,1,2,3, \tag{1}$$

where $f_{conv}(\cdot)$ means the operation of convolution.

On top of that, the interval-layer features, *i.e.*, $\mathbf{F}_0$ & $\mathbf{F}_2$ and $\mathbf{F}_1$ & $\mathbf{F}_3$, and the global features, *i.e.*, $\mathbf{F}_4$ are mixed together to combine the interval-layer context and the global context, *i.e.*,

$$\begin{aligned} \mathbf{F}_1^{cex} &\in \Re^{\frac{H}{2^2}\times\frac{W}{2^2}\times 64} = f_{conv}([\![\mathbf{F}_1, bi(\mathbf{F}_3), bi(\mathbf{F}_4)]\!]), \\ \mathbf{F}_2^{cex} &\in \Re^{\frac{H}{2^1}\times\frac{W}{2^1}\times 64} = f_{conv}([\![\mathbf{F}_0, bi(\mathbf{F}_2), bi(\mathbf{F}_4)]\!]), \end{aligned} \tag{2}$$

where $[\![*]\!]$ means the concatenation operation along the channel dimension. $bi(\cdot)$ denote the upsampling operations using the bilinear interpolation.

In the following, the context-aware features $\mathbf{F}_1^{cex}$ and $\mathbf{F}_2^{cex}$ are integrated further to compute the final context semantics, *i.e.*,

$$\mathbf{F}^{cex} \in \Re^{\frac{H}{2^1}\times\frac{W}{2^1}\times 64} = bi(\mathbf{F}_1^{cex}) \odot \mathbf{F}_2^{cex}, \quad (3)$$

where $\odot$ means the element-wise multiplication.

The target inference is obtained by an average along the channel dimension, *i.e.*,

$$\mathbf{T} \in \Re^{\frac{H}{2^1}\times\frac{W}{2^1}\times 1} = \frac{\sum_{i=1}^{64}\mathbf{F}^{cex}(:,:,i)}{64}. \quad (4)$$

The final inference map can be achieved by activating $\mathbf{T}$ using the Sigmoid function, *i.e.*,

$$\mathbf{T}^{act} \in \Re^{H\times W\times 1} = \frac{1}{1+e^{-(bi(\mathbf{T}))}}. \quad (5)$$

3.3. Contrastive distillation paradigm

To identify the targets from the surroundings, we propose a CDP model to distil the foreground from the background, which suits the saliency and camouflage scenes. The idea is to compute the contrast between the target inference and the background with the purpose of training the model for better inference. To this end, as shown in Fig. 3, CDP is implemented by three components, including foreground semantic vector, background semantic vector, and foreground–background contrastive learning, which will be described in detail in the following.

Given a batch of n samples, denoted as $\mathbf{I}_{\{1:n\}}$, the foreground semantic vector $\mathbf{v}^f_{\{1:n\}}$ is learned via a fully-connected layer on the inference $\mathbf{T}$, *i.e.*,

$$\mathbf{v}^f_{\{1:n\}} \in \Re^{64} = \mathbf{W}_b\mathbf{T} + \mathbf{b}_b \quad (6)$$

where $\mathbf{W}_b$ and $\mathbf{b}_b$ are the learned weights and bias in the fully-connected layer.

The background semantic vector $\mathbf{v}^b_{\{1:n\}}$ is learned using the background mask $(1-\mathbf{GT})$ and the intermediate-layer backbone features, which is determined as the third-layer features $\mathbf{E}_2$ of the input image due to their suitable receptive fields. Concretely,

$$\mathbf{v}^b_{\{1:n\}} \in \Re^{64} = \mathbf{W}_f\left(\mathbf{E}_{2\{1:n\}} \odot \left(1 - dw\left(\mathbf{GT}_{\{1:n\}}\right)\right)\right) + \mathbf{b}_f, \quad (7)$$

where $dw(\cdot)$ and $\mathbf{GT}$ mean the downsampling operation and ground truth, respectively. $\mathbf{W}_f$ and $\mathbf{b}_f$ are the learned weights and bias in the fully-connected layer.

On top of the foreground semantic vector $\mathbf{v}^f_{\{1:n\}}$ and the background semantic vector $\mathbf{v}^b_{\{1:n\}}$ within the batch, the model can be trained using the contrastive loss as

$$\mathcal{L}_{NEG} = -\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\log\left(1 - \frac{\left\langle \mathbf{v}_i^b, \mathbf{v}_i^f\right\rangle}{\left\|\mathbf{v}_i^b\right\|\left\|\mathbf{v}_i^f\right\|}\right), \quad (8)$$

where $\langle * \rangle$ and $\|*\|$ indicate the inner product and matrix modulus, respectively.

Difference to ContrastMask (Wang, Zhao et al., 2022). ContrastMask (Wang, Zhao et al., 2022) migrates from base to novel by sharing query vectors, in which the foreground query and background query are obtained by averaging features from partitions within a batch of object proposals accordingly. Differently, we generate the foreground vector and background vector using pseudo-labels within an image, which tend to be more semantically specific.

3.4. Model training

3.4.1. Supervised training

To train the model, besides the contrastive learning loss $\mathcal{L}_{NEG}$ between foreground and background, extra loss between the inference and ground truth is involved. To generate inferences closer to the ground truth in terms of three aspects, including minimum loss, structure similarity, and intersection, we take three loss functions, *i.e.*, binary cross entropy loss ($\mathcal{L}_B$), structural similarity loss ($\mathcal{L}_S$) (Wang, Bovik, Sheikh, & Simoncelli, 2004), and intersection-over-union loss ($\mathcal{L}_I$), to train the model as

$$\mathcal{L}_D = \mathcal{L}_B(\mathbf{T}^{act},\mathbf{GT}) + \mathcal{L}_S(\mathbf{T}^{act},\mathbf{GT}) + \mathcal{L}_I(\mathbf{T}^{act},\mathbf{GT}). \quad (9)$$

The overall loss $\mathcal{L}$ for the model learning is formulated as the combination of $\mathcal{L}_{NEG}$ and $\mathcal{L}_D$, *i.e.*,

$$\mathcal{L} = \mathcal{L}_{NEG} + \mathcal{L}_D. \quad (10)$$

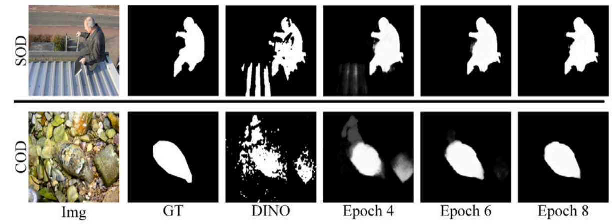

Fig. 4. Visualizations for pseudo masks update.

3.4.2. Unsupervised training

Initial pseudo mask. Recently, DINO (Caron et al., 2021), a new baseline for self-supervised semantic segmentation, has been proved successful for many segmentation tasks. Inspired by its power, it is adopted to produce the initial coarse mask in our framework. Specifically, as shown in Fig. 3, the input image is fed into the pretrained DINO (Caron et al., 2021) to generate one class token and patch tokens. The latter are shapely transformed into dense features. On top of that, one convolution with a kernel of 1 × 1 is carried out to achieve the desired pseudo masks $\mathbf{PM} \in \Re^{H\times W\times 1}$.

Model training. The loss functions of $\mathcal{L}_{NEG}$ and $\mathcal{L}_D$ are both employed to train the model. However, the unsupervised setting is learned by the pseudo mask instead of ground truth, there is some minor modifications. Firstly, $\mathcal{L}_D$ computes the contrast between the current-epoch target inference and the previous-epoch background. To this end, the background semantic vector is computed by using the previous-epoch pseudo mask $\mathbf{PM}$, *i.e.*,

$$\mathbf{v}^b_{\{1:n\}} \in \Re^{64} = fc\left(\mathbf{E}_{2\{1:n\}} \odot \left(1 - down\left(\mathbf{PM}_{\{1:n\}}\right)\right)\right). \quad (11)$$

Secondly, $\mathcal{L}_D$ computes the error between the inference and the pseudo mask as

$$\mathcal{L}_D = \mathcal{L}_B(\mathbf{T}^{act},\mathbf{PM}) + \mathcal{L}_S(\mathbf{T}^{act},\mathbf{PM}) + \mathcal{L}_I(\mathbf{T}^{act},\mathbf{PM}). \quad (12)$$

Pseuod mask update. To train the model well, pseudo labels are updated during the training epochs using the following moving average strategy

$$\mathbf{PM}^i = \begin{cases} \mathbf{PM}^0, & \text{if } i \le 2, \\ \lambda\mathbf{PM}^i + (1-\lambda)\mathbf{T}^{act^i}, & \text{if } i > 2, \end{cases} \quad (13)$$

where the superscript means the training epoch. $\mathbf{PM}^0$ is the pseudo mask $\mathbf{PM}$ generated by parsing deep features of DINO (Caron et al., 2021). λ is experimentally set to 0.4 in this paper. As shown in Fig. 4, the pseudo-mask quality improves along the training epochs.

4. Experiment and analysis

In this section, we will conduct a serious of ablations and experiments to understand the contributions and performance of the proposed method.

Table 1
Training datasets and settings for models using datasets DUTS-TR (Wang et al., 2017), CAMO (Le et al., 2019), COD10K (Fan et al., 2020), and MSRA-B (Liu et al., 2010) with supervised and unsupervised training. See Tables 2, 3 and 8 for details.

Model	Training datasets					Setting	Details
Ours_{D-C}	DUTS-TR (Wang et al., 2017) 10 553 images	&	CAMO (Le et al., 2019) 1000 images	&	COD10K (Fan et al., 2020) 3040 images	Supervised & Unsupervised	See Tables 2 & 3
Ours_{M-C}	MSRA-B (Liu et al., 2010) 2500 images	&	CAMO (Le et al., 2019) 1000 images	&	COD10K (Fan et al., 2020) 3040 images	Unsupervised	See Table 3
Ours_sod	DUTS-TR (Wang et al., 2017) 10 553 images					Supervised & Unsupervised	See Table 8
Ours_cod	CAMO (Le et al., 2019) 1000 images	&	COD10K (Fan et al., 2020) 3040 images			Supervised & Unsupervised	See Table 8

4.1. Implementation details

4.1.1. Training details

We use the SGD optimizer (Bottou, 2012) with an initial learning rate of 0.005 to train our model with the batchsize of 20 using one RTX 3090 Ti GPU. The input is resized to 320 × 320. The training convergence occurs for 50 epochs. The training datasets include SOD training datasets (DUTS-TR training subset Wang et al., 2017 and MSRA-B training subset Liu et al., 2010) and COD training datasets (CAMO Le, Nguyen, Nie, Tran, & Sugimoto, 2019 and COD10K Fan et al., 2020).

4.1.2. Benchmarks

The benchmarks for evaluation includes SOD and COD datasets.

ECSSD (Shi et al., 2015) contains 1000 images with complicated structures, which are collected from the Internet.

HKU-IS (Li & Yu, 2015) consists of 3000 training images and 1447 test images, which are with multiple disconnected objects.

DUTS (Wang et al., 2017) contains 10,533 training images and 5019 test images, which are with different scenes and various sizes.

DUT-OMRON (Yang et al., 2013) has 5168 images with different sizes and complex structures.

In terms of HKU-IS (Li & Yu, 2015) and DUTS (Wang et al., 2017), only the test images are used for evaluations in our experiments.

CHAMELEON (Skurowski et al., 2018) is an unpublished dataset that has only 76 images collected from the Internet via the Google search engine using "camouflaged animal" as a keyword.

CPD1K (Li et al., 2014) is the earliest dataset for camouflaged people detection, which contains 1000 images covering two scene types, namely woodland and snowfield. The test subset has 400 images.

COD10K (Fan et al., 2020), which is collected from multiple photography websites, contains 10,000 images, including 5066 camouflaged images, 3000 background images, and 1934 non-camouflaged images. The test subset includes 2026 images.

CAMO (Le et al., 2019) has 1250 images, which are divided into 1000 training images and 250 testing images.

NC4K (Lv et al., 2021) is a large-scale COD testing dataset, comprising 4121 images.

4.1.3. Training dataset setting

Training datasets and settings for models are listed in the Table 1. Concretely, Table 1 lists the training details of "Ours_{D-C}", "Ours_{M-C}", "Ours_sod" and "Ours_cod" in Tables 2, 3 and 8, respectively.

Ours_{D-C} is a task-agnostic model, trained on mixed DUTS-TR (Wang et al., 2017), CAMO (Le et al., 2019), and COD10K (Fan et al., 2020) datasets. This setting applies to both supervised and unsupervised training, as demonstrated in Tables 2 and 3, respectively.

Ours_{M-C} is also a task-agnostic model, trained on the MSRA-B (Liu et al., 2010), CAMO (Le et al., 2019), and COD10K (Fan et al., 2020) datasets. This configuration is used for unsupervised training, as presented in Table 3.

Ours_sod is a task-specific model, trained on the SOD training dataset, DUTS-TR (Wang et al., 2017), which includes 10,553 images for training. This setup is used for both supervised and unsupervised training, with the results shown in Table 8.

Ours_cod is a task-specific model, trained on the mixed COD training datasets, CAMO (Le et al., 2019) and COD10K (Fan et al., 2020), which collectively include 4040 images. This configuration is also applied for both supervised and unsupervised training, with the results provided in Table 8.

4.1.4. Metrics

We evaluate the performance of our model as well as other state-of-the-art methods from both visual and quantitative perspectives. The quantitative metrics include weighted F-measure (F_β) (Achanta, Hemami, Estrada, & Susstrunk, 2009), Mean Absolute Error (MAE) (Achanta et al., 2009), S-measure (S_m) (Fan, Cheng, Liu, Li, & Borji, 2017), and E-measure (E_m) (Fan et al., 2018). Given a continuous saliency map, a binary mask $\hat{B}$ is achieved by thresholding the saliency map B. Precision is defined as $Precision = \left|\hat{B} \cap G\right| \Big/ \left|\hat{B}\right|$, and recall is defined as $Recall = \left|\hat{B} \cap G\right| / |G|$. Then, the PR curve is plotted under different thresholds.

F-measure is an overall performance indicator, which is computed by

$$F_\beta = \frac{\left(1+\beta^2\right) Precision \times Recall}{\beta^2 Precision + Recall}. \tag{14}$$

As suggested in Achanta et al. (2009), $\beta^2 = 0.3$.

MAE is defined as

$$MAE = \frac{1}{\hat{W} \times \hat{H}} \sum_i |B(i) - G(i)|, \tag{15}$$

where $\hat{W}$ and $\hat{H}$ are the width and height of the image, respectively.

S-measure (S_m) (Fan et al., 2017) is computed by

$$S_m = \alpha S_o + (1-\alpha) S_r, \tag{16}$$

where S_o and S_r represent the object-aware and region-aware structure similarities between the prediction and the ground truth, respectively. α is set to 0.5 (Fan et al., 2017).

E-measure (E_m) (Fan et al., 2018) combines local pixel values with the image-level mean value to jointly evaluate the similarity between the prediction and the ground truth.

4.2. Supervised learning

4.2.1. SOTAs

The SOTA methods for comparison include one supervised joint learning method (UJSC Li et al., 2021), six supervised SOD methods

Table 2
Performance (%) of different methods on SOD and COD benchmarks under the supervised setting. "–" represents the authors have not released the results.

			SOD methods					COD methods				Joint learning methods	
			PiCANet (Liu, Han and Yang, 2018)	CPD (Wu, Su, & Huang, 2019)	BASNet (Qin et al., 2019)	ITSD (Zhou, Xie, Lai, Chen, & Yang, 2020)	MINet (Pang et al., 2020)	SINet (Fan et al., 2020)	PFNet (Mei et al., 2021)	FEDER (He et al., 2023)	ZoomNet (Pang et al., 2022)	UJSC (Li et al., 2021)	Ours$_{D-C}$
			Task-specific	Task-specific	Task-specific	Task-specific	Task-specific	Task-specific	Task-specific	Task-specific	Task-specific	Task-specific	Task-agnostic
SOD	ECSSD (Shi et al., 2015)	$MAE\downarrow$	4.64	4.02	3.70	4.01	3.62	3.58	3.26	3.22	**2.73**	3.00	3.39
		$F_\beta\uparrow$	88.67	91.15	91.68	91.01	91.87	91.75	92.52	92.48	93.33	**93.50**	92.59
		$S_m\uparrow$	91.38	91.02	91.62	91.42	91.91	92.39	92.60	92.27	**93.47**	93.30	92.24
		$E_m\uparrow$	92.33	93.77	94.32	93.75	94.32	94.61	95.08	95.18	95.79	**96.00**	94.81
	DUTS (Wang et al., 2017)	$MAE\downarrow$	5.41	4.29	4.76	4.23	3.94	4.10	3.82	3.92	3.27	3.20	**3.16**
		$F_\beta\uparrow$	78.20	82.44	82.24	83.23	83.49	82.83	84.31	84.55	86.63	86.60	**86.65**
		$S_m\uparrow$	86.70	86.66	86.56	87.71	87.49	87.87	88.17	87.89	**90.00**	89.90	89.40
		$E_m\uparrow$	87.24	90.20	89.54	90.56	90.67	91.78	91.74	91.72	92.96	**93.70**	93.07
	DUT-O (Yang, Zhang, Lu, Ruan, & Yang, 2013)	$MAE\downarrow$	6.79	5.67	5.65	6.32	5.69	5.63	5.54	5.63	5.27	5.10	**4.92**
		$F_\beta\uparrow$	72.24	73.85	76.68	75.24	74.04	75.16	75.99	77.11	77.12	**78.20**	77.43
		$S_m\uparrow$	82.64	81.77	83.62	82.88	82.18	83.24	83.25	83.68	84.09	**85.00**	83.86
		$E_m\uparrow$	83.28	84.50	86.50	85.28	84.58	85.81	86.24	87.18	86.61	**88.40**	86.85
	PASCAL-S (Li, Hou, Koch, Rehg, & Yuille, 2014)	$MAE\downarrow$	7.83	7.21	7.58	6.81	6.39	6.65	6.48	6.68	**5.46**	–	5.76
		$F_\beta\uparrow$	80.02	82.30	81.77	83.05	83.03	82.78	83.33	82.75	**85.05**	–	84.57
		$S_m\uparrow$	84.77	84.46	83.80	85.63	85.54	85.89	85.62	85.10	**87.15**	–	86.40
		$E_m\uparrow$	86.86	88.25	87.86	89.15	89.36	89.22	89.92	89.26	**91.25**	–	90.63
	HKU-IS (Li & Yu, 2015)	$MAE\downarrow$	4.15	3.32	3.29	3.46	3.03	3.21	2.92	2.92	**2.34**	2.60	2.59
		$F_\beta\uparrow$	87.08	89.58	90.36	89.40	90.55	89.82	90.47	90.50	92.33	**92.40**	91.87
		$S_m\uparrow$	90.54	90.45	90.77	90.68	91.39	91.44	91.44	91.12	93.08	**93.10**	92.16
		$E_m\uparrow$	92.26	94.24	94.30	93.95	94.65	94.53	94.94	94.96	**96.14**	86.70	95.58
COD	CAMO (Le et al., 2019)	$MAE\downarrow$	12.50	11.29	15.90	10.16	9.03	9.15	8.49	7.12	**6.59**	7.30	7.17
		$F_\beta\uparrow$	57.26	61.77	47.53	66.29	69.12	70.20	74.61	78.09	**79.38**	77.17	76.70
		$S_m\uparrow$	70.13	71.61	61.82	74.99	74.80	74.54	78.23	80.21	**81.97**	80.01	79.78
		$E_m\uparrow$	71.57	72.27	66.12	77.99	79.18	80.35	84.15	86.66	**87.75**	85.87	84.44
	CHAMELEON (Skurowski et al., 2018)	$MAE\downarrow$	8.52	4.80	11.79	5.73	3.58	3.41	3.25	2.96	**2.29**	2.96	3.32
		$F_\beta\uparrow$	61.83	77.05	52.80	70.46	80.24	82.67	82.80	85.13	**86.35**	84.75	81.88
		$S_m\uparrow$	76.46	85.65	68.74	81.35	85.48	87.20	88.19	88.67	**90.17**	89.13	86.45
		$E_m\uparrow$	77.70	87.36	72.13	84.39	91.42	93.63	93.08	**94.64**	94.29	94.52	92.30
	COD10K (Fan et al., 2020)	$MAE\downarrow$	8.06	5.29	10.54	5.11	4.17	4.26	3.96	3.16	**2.89**	3.53	3.62
		$F_\beta\uparrow$	48.87	59.53	41.68	61.51	65.69	67.93	70.11	75.12	**76.56**	72.05	70.93
		$S_m\uparrow$	69.62	75.01	63.43	76.68	76.97	77.64	79.98	82.23	**83.84**	80.89	79.12
		$E_m\uparrow$	71.17	77.63	67.82	80.83	83.24	86.42	87.73	**89.95**	88.80	88.41	86.29
	NC4K (Lv et al., 2021)	$MAE\downarrow$	8.84	7.20	–	6.39	5.55	5.76	5.27	4.43	**4.34**	4.65	5.05
		$F_\beta\uparrow$	63.96	70.53	–	72.88	76.42	76.86	78.44	**82.41**	81.75	80.62	79.88
		$S_m\uparrow$	75.75	78.74	–	81.08	81.22	80.80	82.90	84.70	85.28	84.15	**88.01**
		$E_m\uparrow$	77.27	80.81	–	84.49	86.23	87.13	88.77	**90.70**	89.60	89.84	90.23

(PiCANet Liu, Han, Yang, 2018, CPD Wu et al., 2019, BASNet Qin et al., 2019, ITSD Zhou et al., 2020, MINet Pang et al., 2020, and PFSN Wu et al., 2023), and four supervised COD methods (SINet Fan et al., 2020, PFNet Mei et al., 2021, FEDER He et al., 2023, and ZoomNet Pang et al., 2022).

4.2.2. Quantitative comparison

Table 2 lists the performance of different methods on SOD and COD benchmarks under the supervised setting. To achieve both SOD and COD results, we re-train the SOD/COD methods using COD/SOD training datasets to achieve their COD/SOD results. As listed in Table 2, we conclude five findings: (i) SOD methods trained using COD datasets cannot get consistently good performance for COD because camouflaged objects are hard than the salient objects for identification; (ii) COD methods trained using SOD datasets can get good SOD performance because salient objects are easy samples for COD; (iii) The only joint learning UJSC (Li et al., 2021) gets consistently good performance for both SOD and COD, but its task-specificity cannot solve the unforewarnable case; (iv) As the only task-agnostic method, our method performs well for SOD and COD, and competitively with the previous task-specific methods.

It is reasonable that our model is slightly inferior to UJSC (Li et al., 2021) and ZoomNet (Pang et al., 2022) because they are task-specific. More specifically, ZoomNet (Pang et al., 2022) is trained using COD datasets. UJSC (Li et al., 2021) is trained using SOD and COD datasets for SOD decoder and COD decoder separately. However, our model is trained using the joint SOD and COD datasets, which is task-agnostic but introduces some domain bias.

4.2.3. Visual comparison

For better displaying the joint learning of SOD and COD, Fig. 5 shows visual detections for SOD and COD in the supervised setting, respectively. As displayed in Fig. 5, using the unified framework, we get detection results more closer to ground truth in terms of object wholeness and uniformity in the supervised circumstance.

4.3. Unsupervised learning

4.3.1. SOTAs

The chosen SOTA methods for comparison include five unsupervised SOD methods (SBF Zhang, Han et al., 2017, USPS Nguyen et al., 2019, UMNet Wang, Zhang et al., 2022, A2S Zhou, Chen et al., 2022, and A2S v2 Zhou et al., 2023), (only) one unsupervised COD method

Table 3
Performance (%) of different methods on SOD and COD benchmarks under the unsupervised setting. "–" represents the authors have not released the results.

			SOD methods					COD method	Universal method	Joint learning methods	
			SBF (Zhang, Han et al., 2017)	USPS (Nguyen et al., 2019)	UMNet (Wang, Zhang et al., 2022)	A2S (Zhou, Chen et al., 2022)	A2Sv2 (Zhou et al., 2023)	UCOS (Zhang & Wu, 2023)	FOUND (Siméoni, Sekkat, Puy, Vobecký, Zablocki, & Pérez, 2023)	Ours$_{M-C}$	Ours$_{D-C}$
			Task-specific	Task-specific	Task-specific	Task-specific	Task-specific	Task-specific	Task-agnostic	Task-agnostic	
SOD	ECSSD (Shi et al., 2015)	$MAE \downarrow$	8.80	6.11	6.36	6.40	4.41	4.87	5.11	4.36	**4.22**
		$F_\beta \uparrow$	79.84	87.00	87.74	88.87	91.43	88.83	89.41	91.33	**91.80**
		$S_m \uparrow$	83.23	85.66	86.77	86.65	89.35	87.83	87.53	89.82	**89.97**
		$E_m \uparrow$	85.01	88.75	89.89	90.91	93.66	93.13	93.03	94.05	**94.23**
	DUTS (Wang et al., 2017)	$MAE \downarrow$	10.69	6.57	6.67	6.46	**4.68**	–	6.07	5.32	4.77
		$F_\beta \uparrow$	62.70	71.98	74.99	75.91	**81.47**	–	74.97	78.24	80.71
		$S_m \uparrow$	68.61	77.17	80.27	81.11	84.24	–	80.33	83.34	**84.55**
		$E_m \uparrow$	71.54	80.11	84.48	86.51	90.19	–	86.60	88.69	**90.25**
	DUT-O (Yang et al., 2013)	$MAE \downarrow$	10.76	5.70	6.31	6.88	**6.09**	–	8.83	6.85	6.65
		$F_\beta \uparrow$	61.20	72.51	73.67	72.76	**74.98**	–	66.30	72.01	73.16
		$S_m \uparrow$	74.73	79.03	80.47	79.50	**81.22**	–	74.69	79.76	80.33
		$E_m \uparrow$	76.32	81.17	83.29	84.50	**86.35**	–	80.24	84.39	84.83
	PASCAL-S (Li et al., 2014)	$MAE \downarrow$	13.09	10.54	–	10.35	7.25	–	7.86	7.47	**6.90**
		$F_\beta \uparrow$	69.51	74.47	–	78.07	82.10	–	79.96	81.45	**82.41**
		$S_m \uparrow$	75.79	76.54	–	78.70	83.00	–	81.09	82.70	**83.50**
		$E_m \uparrow$	77.77	79.47	–	83.69	88.72	–	87.45	88.12	**89.12**
	HKU-IS (Li & Yu, 2015)	$MAE \downarrow$	7.53	4.21	4.12	4.20	3.65	4.09	4.20	3.34	**3.23**
		$F_\beta \uparrow$	80.50	87.45	88.41	88.78	90.14	86.95	87.47	90.16	**90.57**
		$S_m \uparrow$	82.91	86.68	88.65	88.23	88.99	87.13	86.93	89.71	**89.99**
		$E_m \uparrow$	89.33	90.64	92.67	93.53	94.24	93.46	93.62	94.94	**95.24**
COD	CAMO (Le et al., 2019)	$MAE \downarrow$	–	–	–	13.43	17.32	12.70	12.89	11.94	**11.35**
		$F_\beta \uparrow$	–	–	–	62.80	21.65	64.56	63.31	66.39	**67.83**
		$S_m \uparrow$	–	–	–	67.08	44.57	70.04	68.54	71.57	**72.29**
		$E_m \uparrow$	–	–	–	74.93	38.54	78.41	78.20	78.79	**79.77**
	CHAMELEON (Skurowski et al., 2018)	$MAE \downarrow$	–	–	–	8.78	13.42	9.53	9.51	8.34	**8.14**
		$F_\beta \uparrow$	–	–	–	63.61	17.97	62.90	58.96	65.33	**65.39**
		$S_m \uparrow$	–	–	–	70.41	45.43	71.49	68.43	**73.20**	72.91
		$E_m \uparrow$	–	–	–	81.31	38.11	80.18	81.01	82.80	**83.32**
	COD10K (Fan et al., 2020)	$MAE \downarrow$	–	–	–	8.55	8.45	8.62	8.50	7.83	**7.37**
		$F_\beta \uparrow$	–	–	–	50.53	29.22	54.61	51.95	54.84	**56.46**
		$S_m \uparrow$	–	–	–	66.28	51.96	68.89	67.03	69.52	**70.36**
		$E_m \uparrow$	–	–	–	73.76	52.67	73.95	75.07	76.19	**78.00**
	NC4K (Lv et al., 2021)	$MAE \downarrow$	–	–	–	9.38	13.59	8.52	8.39	7.89	**7.57**
		$F_\beta \uparrow$	–	–	–	65.95	36.36	68.93	67.40	70.56	**71.66**
		$S_m \uparrow$	–	–	–	72.14	51.18	75.45	74.12	76.35	**76.87**
		$E_m \uparrow$	–	–	–	80.27	48.95	81.92	82.44	83.28	**84.18**

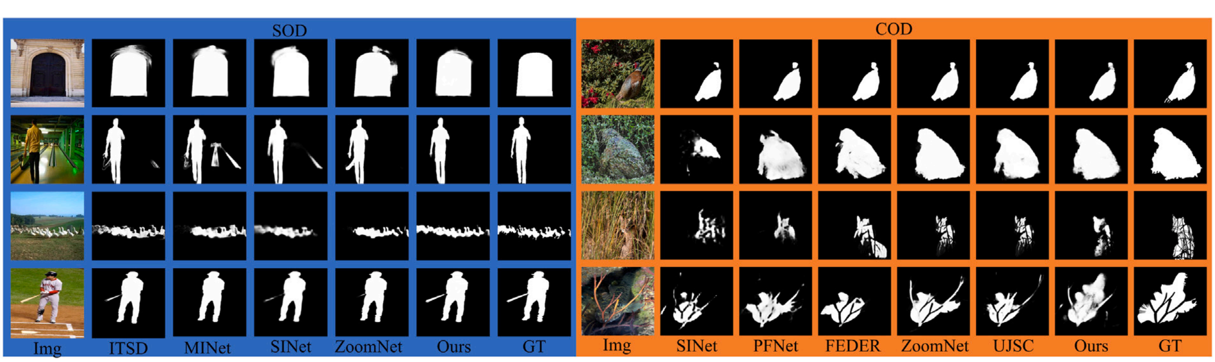

Fig. 5. Visual comparison for SOD and COD in the supervised setting.

Fig. 6. Visual comparison for SOD and COD in the unsupervised setting.

Table 4
Ablation study (%) for CDP in the supervised setting. "∗ −CDP" is achieved by replacing IGC with "∗" model in our framework.

			CPD (Wu et al., 2019)	CPD-CDP	ITSD (Zhou et al., 2020)	ITSD-CDP	MINet (Pang et al., 2020)	MINet-CDP
SOD	ECSSD	$MAE\downarrow$	**4.02**	4.60	4.01	**3.65**	**3.62**	3.95
		$F_\beta\uparrow$	**91.15**	89.36	91.01	**91.59**	**91.87**	91.32
		$S_m\uparrow$	**91.02**	89.62	91.42	**91.43**	**91.91**	91.07
		$E_m\uparrow$	**93.77**	91.76	93.75	**94.01**	**94.32**	93.45
	DUTS	$MAE\downarrow$	4.29	**3.19**	4.23	**3.24**	3.94	**3.4**
		$F_\beta\uparrow$	82.44	**86.15**	83.23	**86.27**	83.49	**85.89**
		$S_m\uparrow$	86.66	**89.26**	87.71	**89.30**	87.49	**88.82**
		$E_m\uparrow$	90.2	**93.21**	90.56	**93.29**	90.67	**92.54**
	DUT-O	$MAE\downarrow$	5.67	**4.91**	6.32	**4.93**	5.69	**5.43**
		$F_\beta\uparrow$	73.85	**76.75**	75.24	**77.96**	74.04	**75.48**
		$S_m\uparrow$	81.77	**83.79**	82.88	**84.47**	82.18	**82.82**
		$E_m\uparrow$	84.5	**86.88**	85.28	**88.11**	84.58	**86.16**
	PASCAL-S	$MAE\downarrow$	7.21	**6.36**	6.81	**5.97**	**6.39**	6.47
		$F_\beta\uparrow$	82.3	**83.38**	83.05	**84.09**	83.03	**83.19**
		$S_m\uparrow$	84.46	**85.3**	85.63	**85.78**	**85.54**	85.21
		$E_m\uparrow$	88.25	**88.53**	89.15	**89.56**	**89.36**	88.67
	HKU-IS	$MAE\uparrow$	**3.32**	3.57	3.46	**3.15**	**3.03**	3.19
		$F_\beta\uparrow$	**89.58**	88.88	89.4	**89.98**	**90.55**	90.14
		$S_m\uparrow$	**90.45**	89.75	**90.68**	90.52	**91.39**	90.6
		$E_m\uparrow$	**94.24**	92.88	**93.95**	93.90	**94.65**	93.91
COD	CAMO	$MAE\downarrow$	11.29	**7.92**	10.16	**8.33**	9.03	**8.8**
		$F_\beta\uparrow$	61.77	**75.37**	66.29	**73.90**	69.12	**71.97**
		$S_m\uparrow$	71.61	**79.43**	74.99	**77.83**	74.8	**76.56**
		$E_m\uparrow$	72.27	**84.94**	77.99	**83.87**	79.18	**82.35**
	CHAMELEON	$MAE\downarrow$	4.8	**3.26**	5.73	**3.41**	**3.58**	3.65
		$F_\beta\uparrow$	77.05	**81.65**	70.46	**81.52**	**80.24**	79.71
		$S_m\uparrow$	85.65	**86.81**	81.35	**86.14**	**85.48**	84.67
		$E_m\uparrow$	87.36	**93.06**	84.39	**92.69**	**91.42**	91.08
	COD10K	$MAE\downarrow$	5.29	**3.67**	5.11	**4.01**	4.17	**4.01**
		$F_\beta\uparrow$	59.53	**70.57**	61.51	**69.01**	65.69	**69.84**
		$S_m\uparrow$	75.01	**79.67**	76.68	**78.51**	76.97	**78.92**
		$E_m\uparrow$	77.63	**87.22**	80.83	**86.43**	83.24	**86.21**
	NC4K	$MAE\downarrow$	7.2	**4.82**	6.39	**5.10**	5.55	**5.12**
		$F_\beta\uparrow$	70.53	**79.71**	72.88	**78.78**	76.42	**78.94**
		$S_m\uparrow$	78.74	**83.65**	81.08	**82.94**	81.22	**82.86**
		$E_m\uparrow$	80.81	**88.96**	84.49	**88.51**	86.23	**88.17**

(UCOS Zhang & Wu, 2023), and one unsupervised universal image segmentation method (FOUND Siméoni et al., 2023).

4.3.2. Quantitative comparison

Table 3 lists the performance of different unsupervised methods on SOD and COD benchmarks. Similar to the supervised setting, we retrain the unsupervised SOD/COD models using the COD/SOD datasets to achieve COD/SOD results. As shown in Table 3, we conclude three findings: (i) The unsupervised SOD models trained using the COD datasets cannot well solve the COD task; (ii) The unsupervised COD method trained using the SOD datasets cannot well solve the SOD task either; (iii) Our task-agnostic model performs better than the previous task-specific SOD and COD methods on most metrics; (iv) Our model beats the previous universal FOUND (Siméoni et al., 2023) by a large margin.

4.3.3. Visual comparison

For better displaying the joint learning of SOD and COD, Fig. 6 shows visual detections for SOD and COD in the unsupervised setting, respectively. As displayed in Fig. 6 in the unsupervised setting, the previous methods mostly introduce lots of noise in the detected saliency map, and even cannot identify the camouflaged objects. Differently, our

Table 5
Ablation study (%) of CDP for SOD in the supervised setting.

Method	ECSSD (Shi et al., 2015)				DUTS (Wang et al., 2017)				DUT-O (Yang et al., 2013)				PASCAL-S (Li et al., 2014)				HKU-IS (Li & Yu, 2015)			
	$MAE\downarrow$	$F_\beta\uparrow$	$S_m\uparrow$	$E_m\uparrow$	$MAE\downarrow$	$F_\beta\uparrow$	$S_m\uparrow$	$E_m\uparrow$	$MAE\downarrow$	$F_\beta\uparrow$	$S_m\uparrow$	$E_m\uparrow$	$MAE\downarrow$	$F_\beta\uparrow$	$S_m\uparrow$	$E_m\uparrow$	$MAE\downarrow$	$F_\beta\uparrow$	$S_m\uparrow$	$E_m\uparrow$
CPD_D (Wu et al., 2019)	**4.02**	**91.15**	**91.02**	**93.77**	4.29	82.44	86.66	90.20	5.67	73.85	81.77	84.50	7.21	82.30	84.46	88.25	**3.32**	**89.58**	90.45	**94.24**
CPD_{D-C}	4.98	88.97	90.59	91.44	4.53	81.15	87.38	88.83	5.91	72.66	82.55	83.66	9.12	77.30	82.03	82.57	3.75	88.11	**90.80**	92.69
CPD-CDP_{D-C}	4.60	89.36	89.62	91.76	**3.19**	**86.15**	**89.26**	**93.21**	**4.91**	**76.75**	**83.79**	**86.88**	**6.36**	**83.38**	**85.30**	**88.53**	3.57	88.88	89.75	92.88
SINet_D (Fan et al., 2020)	**3.58**	**91.75**	**92.39**	**94.61**	4.10	82.83	87.87	91.78	5.63	75.16	**83.24**	85.81	**6.65**	**82.78**	**85.89**	**89.22**	**3.21**	**89.82**	**91.44**	**94.53**
SINet_{D-C}	4.46	90.23	91.00	92.69	3.98	83.44	**88.31**	90.77	5.54	74.35	82.99	85.03	7.45	81.72	84.74	86.83	3.29	89.61	91.38	94.00
SINet-CDP_{D-C}	3.91	91.20	91.26	93.73	**3.66**	**84.75**	88.06	**92.22**	**5.12**	**76.15**	**83.24**	**86.72**	6.87	82.68	84.71	88.36	3.33	89.64	90.31	93.86

Table 6
Ablation study (%) of CDP for COD in the supervised setting.

Method	CAMO (Le et al., 2019)				CHAMELEON (Skurowski et al., 2018)				COD10K (Fan et al., 2020)				NC4K (Lv et al., 2021)			
	$MAE\downarrow$	$F_\beta\uparrow$	$S_m\uparrow$	$E_m\uparrow$	$MAE\downarrow$	$F_\beta\uparrow$	$S_m\uparrow$	$E_m\uparrow$	$MAE\downarrow$	$F_\beta\uparrow$	$S_m\uparrow$	$E_m\uparrow$	$MAE\downarrow$	$F_\beta\uparrow$	$S_m\uparrow$	$E_m\uparrow$
CPD_C (Wu et al., 2019)	11.29	61.77	71.61	72.27	4.80	77.05	85.65	87.36	5.29	59.53	75.01	77.63	7.20	70.53	78.74	80.81
CPD_{D-C}	11.39	61.15	71.06	72.69	5.16	74.99	84.57	85.29	5.60	59.10	75.38	76.68	5.60	70.90	79.68	81.09
CPD-CDP_{D-C}	**7.92**	**75.37**	**79.43**	**84.94**	**3.26**	**81.65**	**86.81**	**93.06**	**3.67**	**70.57**	**79.67**	**87.22**	**4.82**	**79.70**	**83.65**	**88.96**
SINet_C (Fan et al., 2020)	**9.15**	**70.20**	**74.54**	**80.35**	**3.41**	**82.67**	87.20	**93.63**	**4.26**	**67.93**	**77.64**	**86.42**	5.76	76.86	80.80	87.13
SINet_{D-C}	10.10	66.58	74.51	76.54	3.75	80.50	**88.12**	89.63	4.79	64.07	77.41	80.50	6.13	74.75	81.45	84.18
SINet-CDP_{D-C}	9.62	68.23	74.50	79.72	3.49	80.41	85.96	92.69	4.40	66.72	77.35	84.72	**5.58**	**77.19**	**81.94**	**87.23**

model has the ability of identifying and segmenting the salient object and camouflage object well without human annotation.

4.4. Ablation study

4.4.1. CDP

Table 4 lists the performance of different plug-in models. Thanks to the CDP paradigm, the existing models can be improved further, *e.g.,* DUTS (Wang et al., 2017), DUT-O (Yang et al., 2013), CAMO (Le et al., 2019), CHAMELEON (Skurowski et al., 2018), COD10K (Fan et al., 2020), and NC4K (Lv et al., 2021). Especially our CDP paradigm helps the existing SOD methods get large-margin improvements for the COD task. This indicates that our CDP paradigm has great potentials for the agnostic tasks. Moreover, for the purpose of better describing the contribution of the proposed CDP to the joint learning of SOD and COD, we conduct the experiments of training the existing SOD or COD models using the joint SOD and COD datasets. As shown in Tables 5 and 6, using our CDP paradigm, the joint learning of SOD and COD will be enhanced, compared with the joint learning directly using the joint training datasets.

4.4.2. Different components

Table 7(a) lists the performance of different components. On top of BASE, our IGC gets obvious improvements for both SOD and COD tasks. The performance will be improved further by our CDP, which proves the power of our CDP paradigm for the unified framework of SOD and COD.

4.4.3. Different initial pseudo masks

Table 7(b) lists the performance of different initial pseudo masks generations, including TokenCut (Wang, Shen et al., 2022), SpectralSeg (Melas-Kyriazi et al., 2022), SelfMask (Shin et al., 2022), and our pseudo mask. Although SelfMask (Shin et al., 2022) performs well on two SOD benchmarks, our method achieves 27/36 best metrics on SOD/COD tasks. Besides, the update of pseudo labels during training is also crucial, which has been proved in Table 8(a) with the fact that the pseudo masks update improves the performance a lot for both SOD and COD tasks in the unsupervised setting.

4.4.4. Different-layer background semantics

Table 7(c) lists the performance of different layers, including $\mathbf{E}_0 \sim \mathbf{E}_4$, of the encoder for background semantics learning within the CDP paradigm. $\mathbf{E}_2$ gets all best metrics for SOD/COD, which indicates $\mathbf{E}_2$ can well help to extract the background semantics for both salient and camouflaged scenes. Based on this observation, we select $\mathbf{E}_2$ to learn the background semantics for SOD and COD.

4.4.5. Pseudo label

Following the pipeline of previous unsupervised salient object detection works, we select our initial pseudo labels to train CDP (Wu et al., 2019) (SOD method), ZoomNet (Pang et al., 2022) (COD method) in the unsupervised setting. As shown in Table 8(b), our approach beats CPD (Wu et al., 2019) and ZoomNet (Pang et al., 2022). This proves our advantage of the tolerance to coarse pseudo-labels, which comes from the update of pseudo labels during training.

4.4.6. Training separately

We list the results of the proposed method trained using SOD and COD datasets separately in Table 8(c) and (d) under both supervised and unsupervised settings, respectively. The results of joint training will be lower than that of separate manner, which is reasonable because joint SOD and COD datasets inevitably introduce domain bias. In Table 8, the phenomenon that joint learning performs worse than separate learning smoothly indeed reflects certain challenges faced by joint learning. Specifically, joint learning optimizes SOD and COD tasks simultaneously during training, which may lead to interference among tasks, especially when the task objectives or data characteristics differ. During joint learning, as the model attempts to optimize both tasks simultaneously, it may need to make trade-offs when learning shared representations. Such trade-offs could result in degraded performance, as the shared feature representations may not fully cater to the unique demands of each task.

4.4.7. Inference efficiency

Fig. 8 lists parameters, FLOPs, and speed of different methods. It is obvious that our model gets the competitively minimal parameters, least FLOPs, and most fast inference speed of 67 fps, compared with the previous methods, which indicates the potential of our framework for real-time and hardware-friendly applications.

Table 7
Ablation study (%) in the unsupervised setting.

			(a) Different components			(b) Different initial pseudo masks				(c) Different background semantics				
			BASE	BASE+ IGC	BASE+ IGC+ CDP	TokenCut (Wang, Shen et al., 2022)	SpectralSeg (Melas-Kyriazi et al., 2022)	SelfMask (Shin et al., 2022)	Ours	$\mathbf{E}_0$	$\mathbf{E}_1$	$\mathbf{E}_2$	$\mathbf{E}_3$	$\mathbf{E}_4$
SOD	ECSSD (Shi et al., 2015)	$MAE\downarrow$	5.09	4.63	**4.22**	8.55	14.21	5.89	**5.09**	4.35	4.38	**4.22**	4.36	4.38
		$F_\beta\uparrow$	89.23	90.71	**91.80**	84.84	69.80	89.19	**89.23**	91.27	91.26	**91.80**	91.27	91.18
		$S_m\uparrow$	87.36	88.73	**89.97**	82.77	75.98	86.24	**87.36**	89.75	89.76	**89.97**	89.78	89.63
		$E_m\uparrow$	93.01	92.87	**94.23**	86.97	80.37	91.90	**93.01**	94.03	93.99	**94.23**	94.01	93.85
	DUTS (Wang et al., 2017)	$MAE\downarrow$	6.22	5.53	**4.77**	8.87	16.42	**5.58**	6.22	5.48	5.39	**4.77**	5.46	5.40
		$F_\beta\uparrow$	73.87	77.85	**80.71**	74.17	52.53	**78.69**	73.87	77.83	78.23	**80.71**	77.91	77.95
		$S_m\uparrow$	79.85	82.41	**84.55**	77.18	66.20	**81.36**	79.85	83.25	83.32	**84.55**	83.19	83.30
		$E_m\uparrow$	85.90	87.54	**90.25**	82.67	69.01	**88.63**	85.90	88.27	88.55	**90.25**	88.33	88.48
	DUT-O (Yang et al., 2013)	$MAE\downarrow$	9.14	7.14	**6.65**	10.61	19.67	**6.55**	9.14	7.29	7.14	**6.65**	7.23	7.18
		$F_\beta\uparrow$	65.14	71.41	**73.16**	68.35	47.16	**73.12**	65.14	71.04	71.40	**72.01**	71.15	71.11
		$S_m\uparrow$	74.11	78.61	**80.33**	75.21	62.47	**72.01**	74.11	79.37	79.48	**79.76**	79.32	79.27
		$E_m\uparrow$	79.39	83.76	**84.83**	80.33	64.29	**85.78**	79.39	83.39	83.73	**84.83**	83.50	83.21
	PASCAL-S (Li et al., 2014)	$MAE\downarrow$	7.76	7.83	**6.90**	12.44	19.18	8.44	**7.76**	7.52	7.51	**6.90**	7.48	7.49
		$F_\beta\uparrow$	79.64	80.86	**82.41**	75.86	60.28	**81.10**	79.64	81.08	81.24	**82.41**	81.19	81.05
		$S_m\uparrow$	81.09	81.92	**83.50**	76.09	67.40	80.44	**81.09**	82.69	82.74	**83.50**	82.69	82.60
		$E_m\uparrow$	87.41	87.33	**89.12**	82.22	72.15	86.75	**87.41**	88.05	88.07	**89.12**	**88.12**	88.09
	HKU-IS (Li & Yu, 2015)	$MAE\downarrow$	4.27	3.54	**3.23**	7.02	11.35	5.06	**4.27**	3.34	3.35	**3.23**	3.35	3.36
		$F_\beta\uparrow$	87.69	89.53	**90.57**	81.67	67.44	86.80	**87.69**	89.69	89.78	**90.57**	89.80	89.50
		$S_m\uparrow$	86.57	88.27	**89.99**	78.66	76.02	84.92	**86.57**	89.63	89.68	**89.99**	89.65	89.62
		$E_m\uparrow$	93.44	93.67	**95.24**	84.02	80.80	92.35	**93.44**	94.84	94.90	**95.24**	94.85	94.84
COD	CAMO (Le et al., 2019)	$MAE\downarrow$	12.74	12.35	**11.35**	16.26	23.50	18.77	**12.74**	12.29	12.24	**11.35**	12.07	12.23
		$F_\beta\uparrow$	63.27	64.35	**67.83**	54.33	48.05	53.62	**63.27**	66.32	66.34	**67.83**	66.28	66.22
		$S_m\uparrow$	67.91	70.37	**72.29**	63.52	57.91	61.73	**67.91**	71.50	71.56	**72.29**	71.50	71.44
		$E_m\uparrow$	78.28	78.14	**79.77**	70.63	64.76	69.79	**78.28**	78.35	78.34	**79.77**	78.45	78.37
	CHAMELEON (Skurowski et al., 2018)	$MAE\downarrow$	9.38	8.55	**8.14**	13.18	21.95	17.59	**9.38**	8.43	8.38	**8.14**	8.36	8.46
		$F_\beta\uparrow$	59.02	64.26	**65.39**	53.56	43.96	48.09	**59.02**	64.90	64.99	**65.39**	65.21	64.46
		$S_m\uparrow$	68.17	72.19	**72.91**	65.35	57.47	61.89	**68.17**	73.10	72.98	**72.91**	73.08	72.83
		$E_m\uparrow$	80.62	82.08	**83.32**	73.96	62.79	67.51	**80.62**	82.19	83.04	**83.32**	82.35	81.99
	COD10K (Fan et al., 2020)	$MAE\downarrow$	8.75	7.98	**7.37**	10.34	19.32	13.09	**8.75**	7.98	7.85	**7.37**	7.84	7.90
		$F_\beta\uparrow$	51.93	53.94	**56.46**	50.24	38.81	46.92	**51.93**	54.56	54.78	**56.46**	54.76	54.65
		$S_m\uparrow$	66.87	69.77	**70.36**	65.78	57.51	63.71	**66.87**	69.48	69.56	**70.36**	69.57	69.58
		$E_m\uparrow$	76.14	75.90	**78.00**	73.50	59.52	67.87	**76.14**	75.90	76.13	**78.00**	76.17	76.09
	NC4K (Lv et al., 2021)	$MAE\downarrow$	8.59	8.13	**7.57**	10.12	15.88	11.41	**7.59**	7.98	7.94	**7.57**	7.90	7.93
		$F_\beta\uparrow$	67.35	69.92	**71.66**	64.88	56.18	63.38	**67.35**	70.34	70.44	**71.66**	70.56	70.38
		$S_m\uparrow$	74.21	75.47	**76.87**	72.45	66.91	71.57	**74.21**	76.50	75.79	**76.87**	76.32	76.39
		$E_m\uparrow$	83.04	82.76	**84.18**	80.21	71.90	77.65	**83.04**	83.18	83.05	**84.18**	83.27	83.22

Table 8
Improvement for the update of pseudo labels.

		ECSSD (Shi et al., 2015)				HKU-IS (Li & Yu, 2015)				CAMO (Le et al., 2019)				COD10K (Fan et al., 2020)			
		$MAE\downarrow$	$F_\beta\uparrow$	$S_m\uparrow$	$E_m\uparrow$	$MAE\downarrow$	$F_\beta\uparrow$	$S_m\uparrow$	$E_m\uparrow$	$MAE\downarrow$	$F_\beta\uparrow$	$S_m\uparrow$	$E_m\uparrow$	$MAE\downarrow$	$F_\beta\uparrow$	$S_m\uparrow$	$E_m\uparrow$
(a)	Ours(w/o update)	4.94	90.82	88.52	92.21	4.22	88.47	88.73	93.48	12.28	63.92	70.92	77.53	8.04	52.37	68.93	76.32
	Ours	**4.22**	**91.80**	**89.87**	**94.23**	**3.23**	**90.57**	**89.99**	**95.24**	**11.35**	**67.83**	**72.29**	**79.77**	**7.37**	**56.46**	**70.36**	**78.00**
(b)	CPD (Wu et al., 2019)	5.04	90.12	89.96	92.77	3.92	88.89	89.82	93.58	12.97	63.13	70.92	76.23	8.28	53.52	69.93	75.32
	ZoomNet (Pang et al., 2022)	4.83	90.91	88.89	92.93	3.56	89.72	89.23	94.33	12.56	65.73	71.08	77.45	7.90	55.79	68.87	77.63
	Ours	**4.22**	**91.80**	**89.87**	**94.23**	**3.23**	**90.57**	**89.99**	**95.24**	**11.35**	**67.83**	**72.29**	**79.77**	**7.37**	**56.46**	**70.36**	**78.00**
(c)	Ours_sod	**3.16**	**93.04**	**92.82**	**95.20**	**2.52**	91.67	**92.25**	95.46	–	–	–	–	–	–	–	–
	Ours_cod	–	–	–	–	–	–	–	–	7.37	75.35	79.25	84.21	**3.60**	70.82	**79.81**	**86.78**
	Ours	3.39	92.59	92.24	94.81	2.59	**91.87**	92.16	**95.58**	**7.17**	**76.70**	**79.78**	**84.44**	3.62	**70.93**	79.12	86.29
(d)	Ours_sod	**4.02**	**92.35**	**90.29**	**94.35**	**3.14**	**91.54**	**90.18**	94.99	–	–	–	–	–	–	–	–
	Ours_cod	–	–	–	–	–	–	–	–	11.42	67.29	71.73	79.38	7.44	55.56	69.74	77.12
	Ours	4.22	91.80	89.87	94.23	3.23	90.57	89.99	**95.24**	**11.35**	**67.83**	**72.29**	**79.77**	**7.37**	**56.46**	**70.36**	**78.00**

5. Conclusion

In this paper, we have made a task-agnostic unified framework for SOD and COD via a contrastive distillation paradigm based on the agreeable binary segmentation nature. In the supervised setting, our framework performed competitively with the previous task-specific SOD and COD methods. In the unsupervised setting, our framework achieved superior performance on most SOD and COD benchmarks. As well, our work has a real-time inference speed.

Limitations and future work: Despite the superiority of the proposed method, it still has many limitations. For instance, the model may fail in the case of extreme contradiction between SOD and COD tasks since it may introduce significant interference during joint learning,

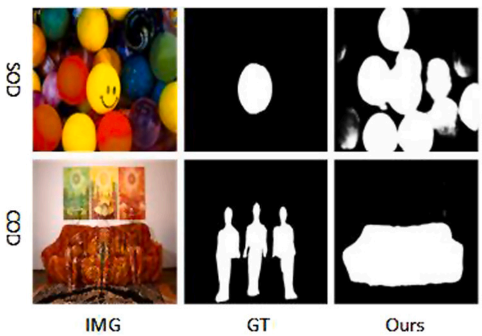

Fig. 7. Failure cases. From left to right: input images, ground truth, and detection results of the proposed method.

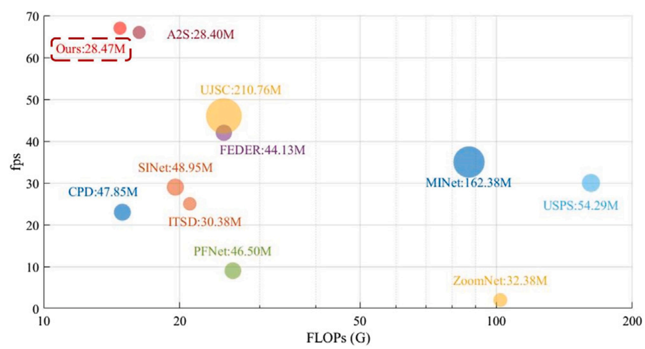

Fig. 8. Efficiency comparison.

making it challenging to balance their mutual influence effectively. This interference could result in the model struggling to fully accommodate the distinct demands of both tasks. As shown in Fig. 7, due to the extreme contradiction between salient and camouflaged scenes, the proposed method cannot obtain satisfactory detections. To address this issue, in the future, we will involve large language model (Hassanin, Keshk, Salim, Alsubaie, & Sharma, 2025; Pi, Yao, Gao, Zhang, & Zhang, 2024; Thirunavukarasu et al., 2023) to abstract the prompts of SOD and COD for contradiction learning and improve the performance.

CRediT authorship contribution statement

Yi Liu: Investigation, Writing – review & editing. **Chengxin Li:** Visualization, Software. **Xiaohui Dong:** Methodology, Software. **Lei Li:** Writing – review & editing. **Dingwen Zhang:** Supervision, Investigation, Writing – review & editing. **Shoukun Xu:** Investigation, Writing – review & editing. **Jungong Han:** Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: cczu.edu.cn, nwpu.edu.cn, tsinghua.edu.cn

Acknowledgements

This work is supported in part by the National Natural Science Foundation of Jiangsu Province under Grant No. BK20221379, in part by the Jiangsu Province Youth Science and Technology Talent Support Project, in part by the Changzhou Science and Technology Plan (Applied Basic Research Special Project) under Grant No. CJ20242060, in part by the State Key Laboratory of Reliability and Intelligence of Electrical Equipment under Grant EERI KF2022005, Hebei University of Technology, and in part by the National Natural Science Foundation of China under Grants No. 62441235, No. 62293543, No. 62322605 & No. U21B2048.

Data availability

Data will be made available on request.

References

Achanta, R., Hemami, S., Estrada, F., & Susstrunk, S. (2009). Frequency-tuned salient region detection. In *2009 IEEE conference on computer vision and pattern recognition* (pp. 1597–1604). IEEE.

Borji, A., Cheng, M.-M., Jiang, H., & Li, J. (2015). Salient object detection: A benchmark. *IEEE Transactions on Image Processing*, *24*(12), 5706–5722.

Bottou, L. (2012). Stochastic gradient descent tricks. In *Neural networks: Tricks of the trade: Second edition* (pp. 421–436). Springer.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., et al. (2021). Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 9650–9660).

Chen, S., Hong, Z., Xie, G.-S., Yang, W., Peng, Q., Wang, K., et al. (2022). Msdn: Mutually semantic distillation network for zero-shot learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 7612–7621).

Chen, S., Tan, X., Wang, B., & Hu, X. (2018). Reverse attention for salient object detection. In *Proceedings of the European conference on computer vision* (pp. 234–250).

Dehmollaian, M., & Sarabandi, K. (2006). Electromagnetic scattering from foliage camouflaged complex targets. *IEEE Transactions on Geoscience and Remote Sensing*, *44*(10), 2698–2709.

Diao, W., Sun, X., Zheng, X., Dou, F., Wang, H., & Fu, K. (2016). Efficient saliency-based object detection in remote sensing images using deep belief networks. *IEEE Geoscience and Remote Sensing Letters*, *13*(2), 137–141.

Fan, D.-P., Cheng, M.-M., Liu, Y., Li, T., & Borji, A. (2017). Structure-measure: A new way to evaluate foreground maps. In *Proceedings of the IEEE international conference on computer vision* (pp. 4548–4557).

Fan, D.-P., Gong, C., Cao, Y., Ren, B., Cheng, M.-M., & Borji, A. (2018). Enhanced-alignment measure for binary foreground map evaluation. arXiv preprint arXiv:1805.10421.

Fan, D.-P., Ji, G.-P., Sun, G., Cheng, M.-M., Shen, J., & Shao, L. (2020). Camouflaged object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 2777–2787).

Fang, H., Liu, A., Wan, J., Escalera, S., Zhao, C., Zhang, X., et al. (2023). Surveillance face anti-spoofing. *IEEE Transactions on Information Forensics and Security*.

Fang, C., Wang, Q., Cheng, L., Gao, Z., Pan, C., Cao, Z., et al. (2023). Reliable mutual distillation for medical image segmentation under imperfect annotations. *IEEE Transactions on Medical Imaging*, *42*(6), 1720–1734.

Guo, G., Zhang, D., Han, L., Liu, N., Cheng, M.-M., & Han, J. (2024). Pixel distillation: Cost-flexible distillation across image sizes and heterogeneous networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Han, J., Zhang, D., Hu, X., Guo, L., Ren, J., & Wu, F. (2014). Background prior-based salient object detection via deep reconstruction residual. *IEEE Transactions on Circuits and Systems for Video Technology*, *25*(8), 1309–1321.

Hassanin, M., Keshk, M., Salim, S., Alsubaie, M., & Sharma, D. (2025). Pllm-cs: Pre-trained large language model (llm) for cyber threat detection in satellite networks. *Ad Hoc Networks*, *166*, Article 103645.

He, C., Li, K., Zhang, Y., Tang, L., Zhang, Y., Guo, Z., et al. (2023). Camouflaged object detection with feature decomposition and edge reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 22046–22055).

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778).

Huang, Z., Dai, H., Xiang, T.-Z., Wang, S., Chen, H.-X., Qin, J., et al. (2023). Feature shrinkage pyramid for camouflaged object detection with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5557–5566).

Jiang, H., Wang, J., Yuan, Z., Wu, Y., Zheng, N., & Li, S. (2013). Salient object detection: A discriminative regional feature integration approach. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2083–2090).

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, *25*.

Lateef, F., Kas, M., & Ruichek, Y. (2021). Saliency heat-map as visual attention for autonomous driving using generative adversarial network (GAN). *IEEE Transactions on Intelligent Transportation Systems*, *23*(6), 5360–5373.

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

Li, T., Cao, P., Yuan, Y., Fan, L., Yang, Y., Feris, R. S., et al. (2022). Targeted supervised contrastive learning for long-tailed recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 6918–6928).

Li, Y., Hou, X., Koch, C., Rehg, J. M., & Yuille, A. L. (2014). The secrets of salient object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 280–287).

Li, Q., Hu, Q., Lin, C., Wu, D., & Shen, C. (2023). Revisiting gradient regularization: Inject robust saliency-aware weight bias for adversarial defense. *IEEE Transactions on Information Forensics and Security*.

Li, G., & Yu, Y. (2015). Visual saliency based on multiscale deep features. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 5455–5463).

Li, A., Zhang, J., Lv, Y., Liu, B., Zhang, T., & Dai, Y. (2021). Uncertainty-aware joint salient object and camouflaged object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10071–10081).

Lin, X., Wu, Z., Chen, G., Li, G., & Yu, Y. (2022). A causal debiasing framework for unsupervised salient object detection. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 1610–1619).

Liu, N., Han, J., & Yang, M.-H. (2018). Picanet: Learning pixel-wise contextual attention for saliency detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3089–3098).

Liu, Y., Han, J., Zhang, Q., & Wang, L. (2018). Salient object detection via two-stage graphs. *IEEE Transactions on Circuits and Systems for Video Technology*, *29*(4), 1023–1037.

Liu, T., Yuan, Z., Sun, J., Wang, J., Zheng, N., Tang, X., et al. (2010). Learning to detect a salient object. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *33*(2), 353–367.

Liu, Y., Zhang, D., Liu, N., Xu, S., & Han, J. (2022). Disentangled capsule routing for fast part-object relational saliency. *IEEE Transactions on Image Processing*, *31*, 6719–6732.

Liu, N., Zhang, N., Wan, K., Shao, L., & Han, J. (2021). Visual saliency transformer. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 4722–4732).

Liu, Y., Zhang, Q., Zhang, D., & Han, J. (2019). Employing deep part-object relationships for salient object detection. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 1232–1241).

Liu, Y., Zhang, D., Zhang, Q., & Han, J. (2021). Integrating part-object relationship and contrast for camouflaged object detection. *IEEE Transactions on Information Forensics and Security*, *16*, 5154–5166.

Liu, Y., Zhang, D., Zhang, Q., & Han, J. (2021c). Part-object relational visual saliency. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(7), 3688–3704.

Lo, Y.-C., Chang, C.-C., Chiu, H.-C., Huang, Y.-H., Chen, C.-P., Chang, Y.-L., et al. (2021). Clcc: Contrastive learning for color constancy. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 8053–8063).

Lv, Y., Zhang, J., Dai, Y., Li, A., Liu, B., Barnes, N., et al. (2021). Simultaneously localize, segment and rank the camouflaged objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 11591–11601).

Mei, H., Ji, G.-P., Wei, Z., Yang, X., Wei, X., & Fan, D.-P. (2021). Camouflaged object segmentation with distraction mining. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 8772–8781).

Melas-Kyriazi, L., Rupprecht, C., Laina, I., & Vedaldi, A. (2022). Deep spectral methods: A surprisingly strong baseline for unsupervised semantic segmentation and localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 8364–8375).

Meng, S., Shao, D., Guo, J., & Gao, S. (2023). Tracking without label: Unsupervised multiple object tracking via contrastive similarity learning. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 16264–16273).

Mondal, A. (2020). Camouflaged object detection and tracking: A survey. *International Journal of Image and Graphics*, *20*(04), Article 2050028.

Nguyen, T., Dax, M., Mummadi, C. K., Ngo, N., Nguyen, T. H. P., Lou, Z., et al. (2019). Deepusps: Deep robust unsupervised saliency prediction via self-supervision. *Advances in Neural Information Processing Systems*, *32*.

Pang, Y., Zhao, X., Xiang, T.-Z., Zhang, L., & Lu, H. (2022). Zoom in and out: A mixed-scale triplet network for camouflaged object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 2160–2170).

Pang, Y., Zhao, X., Zhang, L., & Lu, H. (2020). Multi-scale interactive network for salient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 9413–9422).

Pi, R., Yao, L., Gao, J., Zhang, J., & Zhang, T. (2024). Perceptiongpt: Effectively fusing visual perception into llm. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 27124–27133).

Qin, X., Zhang, Z., Huang, C., Gao, C., Dehghan, M., & Jagersand, M. (2019). Basnet: Boundary-aware salient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 7479–7489).

Shi, J., Yan, Q., Xu, L., & Jia, J. (2015). Hierarchical image saliency detection on extended CSSD. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *38*(4), 717–729.

Shin, G., Albanie, S., & Xie, W. (2022). Unsupervised salient object detection with spectral cluster voting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 3971–3980).

Siméoni, O., Sekkat, C., Puy, G., Vobeckỳ, A., Zablocki, É., & Pérez, P. (2023). Unsupervised object localization: Observing the background to discover objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 3176–3186).

Siris, A., Jiao, J., Tam, G. K., Xie, X., & Lau, R. W. (2021). Scene context-aware salient object detection. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 4156–4166).

Skurowski, P., Abdulameer, H., Błaszczyk, J., Depta, T., Kornacki, A., & Kozieł, P. (2018). Animal camouflage analysis: Chameleon database. *Unpublished Manuscript*, *2*(6), 7.

Song, R., Ai, Y., Tian, B., Chen, L., Zhu, F., & Fei, Y. (2022). Msfanet: A light weight object detector based on context aggregation and attention mechanism for autonomous mining truck. *IEEE Transactions on Intelligent Vehicles*.

Tang, L., Zhan, Y., Chen, Z., Yu, B., & Tao, D. (2022). Contrastive boundary learning for point cloud segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 8489–8499).

Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., Gutierrez, L., Tan, T. F., & Ting, D. S. W. (2023). Large language models in medicine. *Nature Medicine*, *29*(8), 1930–1940.

Tian, X., Zhang, J., Xiang, M., & Dai, Y. (2023). Modeling the distributional uncertainty for salient object detection models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 19660–19670).

Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, *13*(4), 600–612.

Wang, W., Lai, Q., Fu, H., Shen, J., Ling, H., & Yang, R. (2021). Salient object detection in the deep learning era: An in-depth survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(6), 3239–3259.

Wang, L., Lu, H., Wang, Y., Feng, M., Wang, D., Yin, B., et al. (2017). Learning to detect salient objects with image-level supervision. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 136–145).

Wang, W., Shen, J., Dong, X., & Borji, A. (2018). Salient object detection driven by fixation prediction. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1711–1720).

Wang, Y., Shen, X., Hu, S. X., Yuan, Y., Crowley, J. L., & Vaufreydaz, D. (2022). Self-supervised transformers for unsupervised object discovery using normalized cut. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 14543–14553).

Wang, Y., Zhang, W., Wang, L., Liu, T., & Lu, H. (2022). Multi-source uncertainty mining for deep unsupervised saliency detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 11727–11736).

Wang, X., Zhao, K., Zhang, R., Ding, S., Wang, Y., & Shen, W. (2022). Contrastmask: Contrastive learning to segment every thing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 11604–11613).

Wu, Z., Su, L., & Huang, Q. (2019). Cascaded partial decoder for fast and accurate salient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 3907–3916).

Wu, Z., Wang, L., Wang, W., Xia, Q., Chen, C., Hao, A., et al. (2023). Pixel is all you need: adversarial trajectory-ensemble active learning for salient object detection. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 2883–2891).

Wu, C., Wu, F., Lyu, L., Huang, Y., & Xie, X. (2022). Communication-efficient federated learning via knowledge distillation. *Nature Communications*, *13*(1), 2032.

Xing, H., Xiao, Z., Qu, R., Zhu, Z., & Zhao, B. (2022). An efficient federated distillation learning system for multitask time series classification. *IEEE Transactions on Instrumentation and Measurement*, *71*, 1–12.

Yang, C., Zhang, L., Lu, H., Ruan, X., & Yang, M.-H. (2013). Saliency detection via graph-based manifold ranking. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3166–3173).

Yu, B., Lu, J., Li, X., & Zhou, J. (2021). Salience-aware face presentation attack detection via deep reinforcement learning. *IEEE Transactions on Information Forensics and Security*, *17*, 413–427.

Zhai, Q., Li, X., Yang, F., Chen, C., Cheng, H., & Fan, D.-P. (2021). Mutual graph learning for camouflaged object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 12997–13007).

Zhang, X., Han, L., Xu, C., Zheng, Z., Ding, J., Fu, X., et al. (2024). MHKD: Multi-step hybrid knowledge distillation for low-resolution whole slide images glomerulus detection. *IEEE Journal of Biomedical and Health Informatics*.

Zhang, D., Han, J., & Zhang, Y. (2017). Supervision by fusion: Towards unsupervised learning of deep salient object detector. In *Proceedings of the IEEE international conference on computer vision* (pp. 4048–4056).

Zhang, P., Liu, W., Lu, H., & Shen, C. (2019). Salient object detection with lossless feature reflection and weighted structural loss. *IEEE Transactions on Image Processing*, *17*, 3048–3060.

Zhang, J., & Sclaroff, S. (2015). Exploiting surroundedness for saliency detection: a boolean map approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *38*(5), 889–902.

Zhang, J., Sclaroff, S., Lin, Z., Shen, X., Price, B., & Mech, R. (2015). Minimum barrier salient object detection at 80 fps. In *Proceedings of the IEEE international conference on computer vision* (pp. 1404–1412).

Zhang, X., Wang, W., Chen, Z., Xu, Y., Zhang, J., & Tao, D. (2023). Clamp: Prompt-based contrastive learning for connecting language and animal pose. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 23272–23281).

Zhang, P., Wang, D., Lu, H., Wang, H., & Ruan, X. (2017). Amulet: Aggregating multi-level convolutional features for salient object detection. In *Proceedings of the IEEE international conference on computer vision* (pp. 202–211).

Zhang, X., Wang, Q., Qin, M., Wang, Y., Ohtsuki, T., Adebisi, B., et al. (2024). Enhanced few-shot malware traffic classification via integrating knowledge transfer with neural architecture search. *IEEE Transactions on Information Forensics and Security*.

Zhang, Y., & Wu, C. (2023). Unsupervised camouflaged object segmentation as domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 4334–4344).

Zhang, J., Xie, J., Barnes, N., & Li, P. (2021). Learning generative vision transformer with energy-based latent space for saliency prediction. *Advances in Neural Information Processing Systems*, *34*, 15448–15463.

Zhang, J., Zhang, T., Dai, Y., Harandi, M., & Hartley, R. (2018). Deep unsupervised saliency detection: A multiple noisy labeling perspective. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 9029–9038).

Zhao, T., Han, J., Yang, L., Wang, B., & Zhang, D. (2021). SODA: Weakly supervised temporal action localization based on astute background response and self-distillation learning. *International Journal of Computer Vision*, *129*(8), 2474–2498.

Zhao, J.-X., Liu, J.-J., Fan, D.-P., Cao, Y., Yang, J., & Cheng, M.-M. (2019). Egnet: Edge guidance network for salient object detection. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 8779–8788).

Zhao, R., Ouyang, W., Li, H., & Wang, X. (2015). Saliency detection by multi-context deep learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1265–1274).

Zhao, X., Vemulapalli, R., Mansfield, P. A., Gong, B., Green, B., Shapira, L., et al. (2021). Contrastive learning for label efficient semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 10623–10633).

Zhong, Y., Li, B., Tang, L., Kuang, S., Wu, S., & Ding, S. (2022). Detecting camouflaged object in frequency domain. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 4504–4513).

Zhou, H., Chen, P., Yang, L., Xie, X., & Lai, J. (2022). Activation to saliency: Forming high-quality labels for unsupervised salient object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, *33*(2), 743–755.

Zhou, T., Fu, H., Chen, G., Zhou, Y., Fan, D.-P., & Shao, L. (2021). Specificity-preserving RGB-D saliency detection. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 4681–4691).

Zhou, H., Qiao, B., Yang, L., Lai, J., & Xie, X. (2023). Texture-guided saliency distilling for unsupervised salient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 7257–7267).

Zhou, H., Xie, X., Lai, J.-H., Chen, Z., & Yang, L. (2020). Interactive two-stream decoder for accurate and fast saliency detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 9141–9150).

Zhou, T., Zhou, Y., Gong, C., Yang, J., & Zhang, Y. (2022). Feature aggregation and propagation network for camouflaged object detection. *IEEE Transactions on Image Processing*, *31*, 7036–7047.