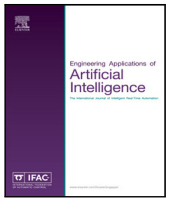
Engineering Applications of Artificial Intelligence 127 (2024) 107303

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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



# IPNet: Polarization-based Camouflaged Object Detection via dual-flow network

Xin Wang [a,b,*], Jiajia Ding [a], Zhao Zhang [a], Junfeng Xu [a], Jun Gao [a]

[a] School of Computer and Information, Hefei University of Technology, Hefei, Anhui, China
[b] Intelligent Interconnected Systems Laboratory of Anhui Province, Hefei, China

ARTICLE INFO

Dataset link: https://github.com/cvhfut/PCOD_1200

*Keywords:*
Camouflaged Object Detection
Polarization image
Feature fusion
Dataset

ABSTRACT

Camouflaged Object Detection (COD) is a critical task in a variety of domains, such as medicine and military applications. The main challenge in COD is accurately detecting and extracting the concealed object from the complex background. The similarity between the camouflaged objects and their background significantly reduces the accuracy of object extraction. Polarization information can provide valuable insights into the characteristics of objects with different material properties and surface roughness. It reflects the difference in polarization information between the object and the background, which increases the contrast between the two and improves the object detection accuracy even under complex scenes. In this paper, we propose IPNet, an efficient cross-modal fusion network that utilizes both RGB intensity and linear polarization cues to generate scene representation with high contrast. Our novel network architecture dynamically fuses RGB intensity and polarization cues using an efficient cross-modal fusion module, leveraging cross-level contextual information to achieve robust detection. For training and evaluating the proposed network, we construct a polarization-based PCOD_1200 dataset that contains 89 subclasses and 1200 samples. A comprehensive set of experiments demonstrates the effectiveness of IPNet to fuse polarization and RGB intensity information and shows that our approach outperforms state-of-the-art methods.

## 1. Introduction

The term “Camouflage” originally referred to the behavior of insects and other animals attempting to blend into their surrounding environment to evade natural predators. For instance, chameleons have the ability to modify their appearance to match the colors and patterns of their surroundings. Humans subsequently adopted this mechanism, applying it extensively on the battlefield by incorporating camouflage and concealment into the clothing and coloring of soldiers and war equipment. Camouflaged Object Detection (COD) is a technology that enables the identification of the entire scope of a camouflaged object. It has broad applications in various civil fields, including medicine (e.g. polyp segmentation Fan et al., 2020b), agriculture (e.g. pest identification Cheng et al., 2017), and ecological protection (e.g. wildlife preservation Roy et al., 2023).

Compared to object detection or segmentation techniques, COD presents a more difficult challenge, as the foreground objects often share similar textures with their surroundings. Conventional COD methods can be broadly categorized as traditional handcrafted feature-based or deep learning-based methods. Traditional methods primarily focus on integrating handcrafted features, such as color (Galun et al., 2003), texture (Sengottuvelan et al., 2008), shape (Song and Geng, 2010), gradient information (Kavitha et al., 2011), and optical properties (Liu et al., 2012), to accentuate the differences between camouflaged objects and their backgrounds. However, the detection results of these methods are not satisfactory in practical scenarios with clutter and noise interference. While emerging deep learning-based methods (Le et al., 2019; Fan et al., 2020a; Lv et al., 2021) have demonstrated promising performance, the low contrast between the object and background in RGB images hinders the network from learning the distinguishing features of camouflage. We posit that improving image contrast is necessary to obtain more critical camouflage cues.

Polarization is an innate characteristic of light that carries intrinsic information which can be exploited to discern objects with different material properties and surface roughness, through the changes in the polarization state. In fact, many animals such as crabs, dragonflies, and cuttlefish, possess a polarization vision system that enhances their ability to detect objects (Roberts et al., 2014). To detect and analyze polarization information, several imaging techniques have been developed, such as division of time, division of amplitude, division of

* Corresponding author at: School of Computer and Information, Hefei University of Technology, Hefei, Anhui, China.
*E-mail address:* wangxin@hfut.edu.cn (X. Wang).

https://doi.org/10.1016/j.engappai.2023.107303
Received 5 May 2023; Received in revised form 28 August 2023; Accepted 12 October 2023
Available online 20 October 2023

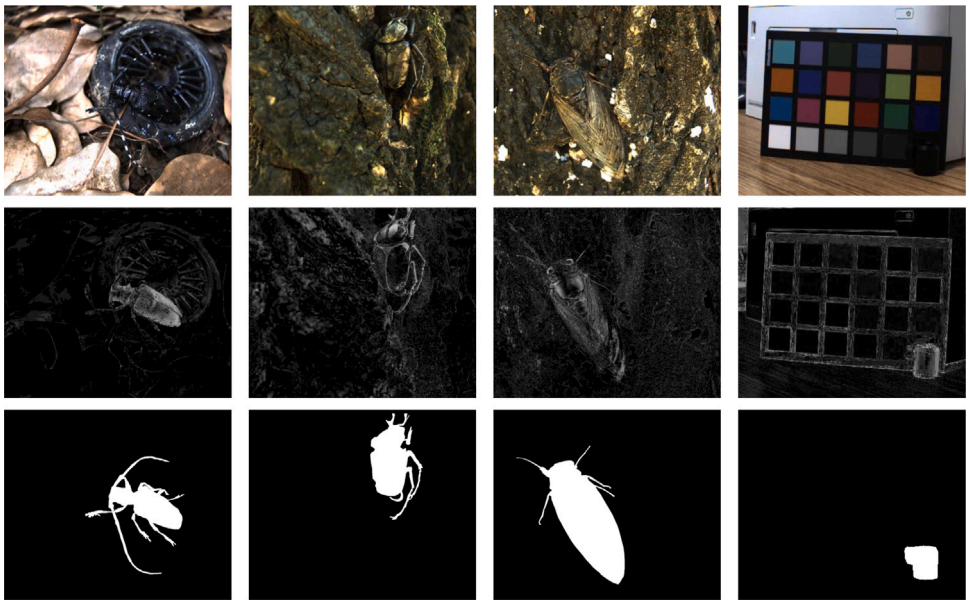


**Fig. 1.** The visual comparison of RGB image (1st row), DoLP image (2nd row), and ground truth (3rd row).

aperture, and division of focal plane (Tyo et al., 2006). These techniques provide both intensity and polarization information and have found applications in diverse fields, such as medical diagnoses (Buckley et al., 2020), defect detection (Giakos et al., 2004), remote sensing (Yan et al., 2020), and object detection (Tyo et al., 2006). Therefore, the incorporation of polarization information into the COD task has the potential to enhance performance.

Our current understanding of polarization is principally based on imaging techniques that employ Stokes parameter measurements to derive the angle and percentage of polarization. Typically, utilizing a polarization camera, one can capture optical intensity images, namely $I_0$, $I_{45}$, $I_{90}$ and $I_{135}$, in the polarization directions of 0°, 45°, 90° and 135°, respectively. Linear Stokes vectors ($S_0$, $S_1$ and $S_2$) are defined as follows:

$$\begin{cases} S_0 = & (I_0 + I_{45} + I_{90} + I_{135})/2 \\ S_1 = & I_0 - I_{90} \\ S_2 = & I_{45} - I_{135} \end{cases} . \tag{1}$$

$S_0$ represents the intensity image, while the degree of linear polarization (DoLP), which refers to the proportion of waves in a source of light that exhibit a particular polarization state, can be calculated as:

$$DoLP = \frac{\sqrt{S_1^2 + S_2^2}}{S_0}. \tag{2}$$

On the other hand, the angle of polarization (AoLP) describes the average orientation can be computed as:

$$AoLP = \frac{1}{2} tan^{-1} \frac{S_1}{S_2}. \tag{3}$$

The DoLP image accentuates the contrast between the object and the background, thereby highlighting differences in the polarization characteristics due to variations in materials within the same scene. While intensity images capture the reflected and transmitted light and typically describe the object's reflectivity and transmissivity, DoLP images document the polarization properties and provide detailed features, such as object surface shape, shading, roughness, and boundary information of camouflaged objects (see Fig. 1). Incorporating both intensity and DoLP images provides complementary information from distinct perspectives. The fusion of these modalities allows for the generation of a scene representation with enhanced contrast.

In this paper, we present a novel and efficient fusion network, IPNet, which is based on a coarse-to-fine structure for COD. Although the DoLP image can effectively reflect the geometric characteristics of the object and enhance its boundary information, it may lack crucial information under certain lighting conditions. Therefore, our approach integrates intensity and DoLP images to enhance image contrast and improve COD. Specifically, we first employ a Prediction Module (PM) to learn prediction maps from the intensity image, which are then used in a polarization Enrich Module (ERM) specially designed to enhance the DoLP image. The enriched DoLP image and the intensity image are then combined using a top-down approach to fuse their multimodal features for image contrast enhancement. Given that the contributions of intensity and DoLP modalities are distinct at every network level, we devise a Cross-modal Fusion Module (CMFM) that effectively integrates global and contextual information to accurately locate and highlight camouflaged objects. Additionally, we utilize a Cross-level Aggregation Module (CAM) that reintegrates the cross-modal fusion features at various levels to produce more precise results.

To train the IPNet, we construct a real-world, comprehensive and challenging polarization-based dataset, PCOD _1200, which includes 1200 manually annotated RGB intensity images and their corresponding DoLP images. The dataset contains scenes with diverse variations in size, texture, clutter, and illumination, captured using the Lucid Triton polarization camera, which captures four images with different polarization directions (0°, 45°, 90° and 135°) in one shoot by its division-of-focal-plane polarimeter.

Our experiments demonstrate the effectiveness of our approach and highlight the importance of intensity and DoLP image fusion for enhancing contrast in COD. We evaluate more than 10 object detection methods on PCOD_1200 and find the proposed IPNet outperforms all other methods. In brief, our contributions are as follows:

- We introduce polarization information to enlarge the discrepancies between camouflaged objects and their surroundings.
- We construct a polarization-based dataset, PCOD_1200, that comprises 1200 high-quality polarization scenes divided into 89 subclasses. For each scene, we provide four images with different polarization directions (0°, 45°, 90° and 135°), the intensity image, an object-level label, DoLP image, and AoLP image.
- We propose a novel polarization-based COD network that includes PM, ERM, CMFM, and AM to dynamically fuse RGB intensity and polarization cues (DoLP) to improve image contrast and achieves state-of-the-art COD performance.


(a) RGB images (b) DoLP image (c) AoLP image

**Fig. 2.** Display of diverse image modalities.

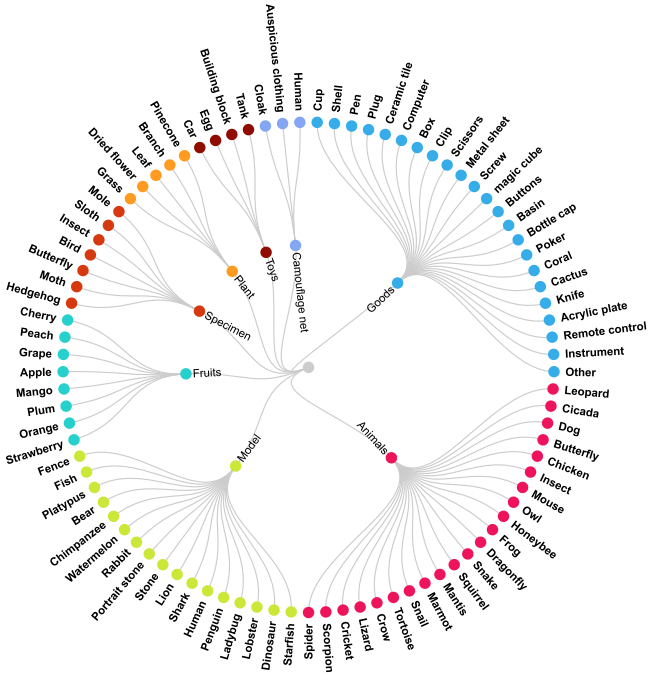



**Fig. 3.** Categories of camouflaged objects in the PCOD_1200 dataset.

## 2. Related work

**Camouflaged Object**. The phenomenon of camouflage, which can be attributed to the process of natural selection and adaptation, has a rich history in the biological sciences. Its impact, however, extends beyond the natural world and into human society, including fields such as art, popular culture, and design (Stevens and Merilaita, 2009). In computer vision, the study of camouflaged objects often pertains to salient object detection (SOD), which primarily concerns itself with identifying conspicuous objects in a scene (Borji et al., 2015). While saliency models are developed for general observation paradigms, where visually prominent objects are the focus, they are unsuitable for specific observation tasks, such as identifying concealed objects. To overcome this limitation, it is imperative to construct models that are tailored to the unique demands and specific data of the task, in order to effectively acquire the specialized expertise.

**Camouflaged Object Detection**. COD shares many similarities with the SOD task (Qin et al., 2019; Wei et al., 2020; Wu et al., 2019; Zhao et al., 2019). And COD can often be even more challenging as it involves distinguishing objects from their visually similar backgrounds. Early COD methods typically rely on handcrafted features, such as color (Galun et al., 2003), texture (Sengottuvelan et al., 2008), shape (Song and Geng, 2010), optical flow (Hou and Li, 2011), or gradient information (Kavitha et al., 2011), to highlight the camouflaged object. However, these methods are easily affected by variations in the surrounding environment (Bi et al., 2021). To overcome these limitations, several multi-modal mechanisms that additionally use hyperspectral analysis (Kim, 2015) and wavelet transform (Li et al., 2018) have been proposed to assist COD. However, traditional COD methods are highly dependent on the quality of handcrafted features, which limits their generalization ability in complex and changing scenes. In light of the advancements in deep learning, several COD benchmark datasets and deep learning-based COD techniques have been proposed (Le et al., 2019; Fan et al., 2020a; Lv et al., 2021; Sun et al., 2021; Wang et al., 2021a; Yan et al., 2021; Liu et al., 2022; Mei et al., 2021; Fan et al., 2021; Ren et al., 2021; Yang et al., 2021; Zhai et al., 2021). Most of these methods leverage a single image to extract camouflaged object features. Besides, several approaches have incorporated depth (Ren et al., 2021), boundary (Zhai et al., 2021), and flipped feature (Yan et al., 2021) to enhance prediction accuracy. Le et al. devised an end-to-end network consisting of a classification stream and a segmentation stream for the COD task. The classification stream predicted the likelihood of camouflaged objects being present in the input image. This probability map was then combined with the segmentation stream to improve the accuracy of camouflaged object segmentation (Le et al., 2019). In another work, Fan et al. developed a network called SINet that contains two main modules: the search module (SM) and the identification module (IM). The SM is responsible for searching for camouflaged objects while the IM is used to precisely detect them Fan et al. (2020a). Subsequently, Lyu et al. proposed the first joint network for SOD and COD within an adversarial learning framework that explicitly models the prediction uncertainty of each task (Lv et al., 2021). Mei et al. developed a novel distraction mining strategy for identifying and removing distractions, which was then used to construct a positioning and focus network (Mei et al., 2021). Fan et al. proposed an enhanced model based on SINet that achieved promising performance. This model included two well-elaborated sub-components: the neighbor connection decoder and group-reversal attention (Fan et al., 2021). Yang et al. proposed a framework that combines a probabilistic representational model with transformers to explicitly reason about uncertainties in camouflaged scenes (Yang et al., 2021). However, current deep learning-based COD methods overemphasize the extrinsic differences and overlook the intrinsic differences between camouflaged objects and their surroundings, leading to a heavy reliance on appearance features. Traditional RGB images provide limited information on the physical properties of camouflaged objects and exhibit low contrast, resulting in unstable detection outcomes in complex scenarios. In contrast, polarization information can robustly describe crucial physical properties such as surface geometry, material, and roughness, providing rich physical information, clear outlines, and high contrast. Therefore, within this paper, we are poised to incorporate polarization information

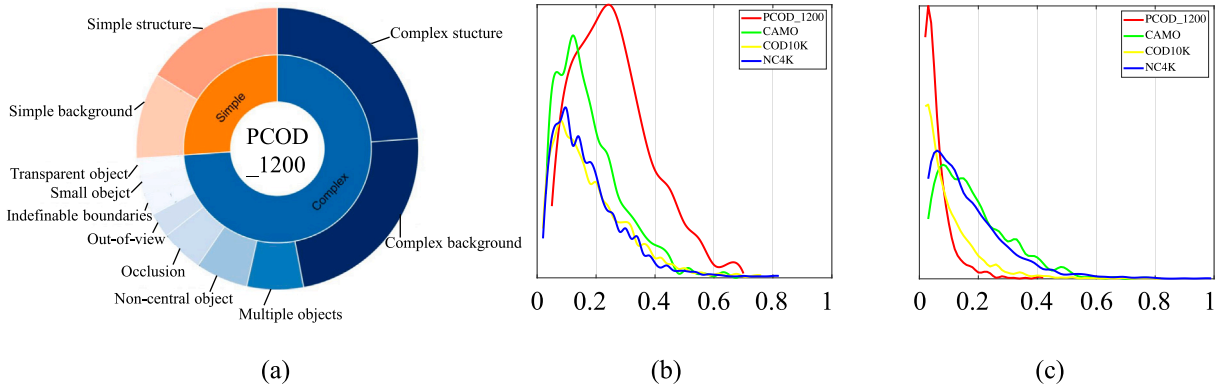



**Fig. 4.** Data distribution of PCOD_1200 dataset. (a) Statistics on scene complexity. (b) Object center to image center. (c) Normalized object size.

as a supplementary cue into the domain of COD. This endeavor exhibits promising prospects for elevating the accuracy and resilience of COD techniques.

**Polarization-based Object Detection**. Recently, the integration of polarization information has been widely explored in various multi-modal visual tasks. Kalra et al. demonstrated the application of polarization imaging techniques in the segmentation of transparent objects in cluttered scenes, and achieved good performance by inputting RGB images, polarization angle, and polarization degree into the same backbone network with a designed fusion module for feature-level fusion (Kalra et al., 2020). Fan et al. improved car detection by fusing polarization features and RGB image features (Fan et al., 2018a). Blin et al. leveraged polarization imaging and adaptive learning models to enhance object detection in road scenes under bad weather conditions (Blin et al., 2019). Zhang et al. improved road scene segmentation by using a scalable multi-level semantic segmentation fusion technique to fuse RGB images and polarized images (Zhang et al., 2019). Mei et al. presented a robust glass segmentation network that dynamically fused trichromatic intensity and polarization cues captured in-the-wild. The network paralleled input RGB, DoLP, and AoLP and enhanced local context cues with multi-scale pixel-wise correlation enhancement (Mei et al., 2022). Xiang et al. designed an efficient attention-bridges fusion network (EAFNet) to exploit complementary information from polarization sensors. EAFNet dynamically extracted attention weights of RGB and polarization branches, adjusted and fused multimodal features, and significantly improved segmentation performance, especially for classes with highly polarized characteristics such as glass and cars (Xiang et al., 2021). Within this paper, we will integrate polarization processing methodologies with the domain of deep learning techniques. This integration encompasses the curation of a specialized dataset customized for polarization-based COD. Subsequently, we will leverage this dataset to train specialized deep neural network models, specifically designed for COD, employing RGB intensity and polarization cues as key inputs.

## 3. Polarization COD dataset

Before delving into the details of the research presented in this paper, it is necessary to establish some fundamental concepts of polarimetry.

### 3.1. Polarization information

Polarized light is ubiquitous in nature, and unlike the majority of animal species, humans are unable to perceive polarization in our everyday lives. Imaging polarimetry enables the extraction of visual information from the polarization of light, in addition to the visual information conveyed by intensity and color. Due to the fact that the polarization of light can be altered by scattering and reflection processes, many animals utilize polarization sensitivity for various behavior-specific tasks, such as navigation (Wehner and Müller, 2006), communication (How et al., 2014), and contrast enhancement (How et al., 2015). In color-sensitive animals, polarization is used in conjunction with color to enhance object detection effectiveness (Kinoshita et al., 2011), while in color-blind species, polarization is the primary mechanism for achieving image segmentation (Cronin et al., 2003).

A polarizer is an optical filter that selectively transmits light of a certain linear polarization angle known as the polarizer orientation. A polarization camera records four linear polarization states of light: $I_0$, $I_{45}$, $I_{90}$ and $I_{135}$, where $I_i$ represents the intensity image of light transmitted through a polarizer oriented at $i$ degrees to the horizontal.

The polarization state of light is described by a Stokes vector $S = [S_0, S_1, S_2]$, and the Stokes elements $S_0$, $S_1$ and $S_2$ can be computed from the measurements of $I_0$, $I_{45}$, $I_{90}$ and $I_{135}$. The values of DoLP and AoLP can be calculated using formulas (2) and (3), respectively.

The DoLP image is used to characterize polarization properties, and it reveals unique information about the surface states of an object that cannot be obtained through intensity or color. The polarization characteristics of the same object vary greatly with different surface states, highlighting the uniqueness of polarization detection in obtaining object texture features and surface state information. To demonstrate the benefits of using polarization information for scene perception, we generated a visualization of a set of DoLP and AoLP polarization images, as shown in Fig. 2. We observed that AoLP contains more noise and offers limited information, whereas DoLP enhances the contrast between the target and the background by showing differences in polarization information for objects of different categories or materials.

### 3.2. PCOD_1200 dataset

In order to establish a universal standard for Polarization-based COD, we have meticulously constructed a comprehensive, real-world, and challenging dataset, dubbed PCOD_1200. The PCOD_1200 dataset consists of a total of 1200 instances of camouflage object detection scenarios, with 970 scenes allocated for training and the remaining 230 for testing. Different from widely utilized COD datasets, which were collected from the Internet through the Google search engine using specific keywords, our dataset is meticulously curated from authentic real-world images captured using a polarization camera within carefully designed camouflage scenarios. We draw inspiration from a wide array of authentic camouflage scenarios found in nature, military contexts, industrial environments, and daily life. Utilizing elements such as plants, specimens, Ghillie suits, and simulation models, we meticulously design and construct camouflage object detection scenes across diverse backgrounds including grasslands, sandy terrains, snowy landscapes, barren woods, and camouflaged fabrics. Specifically, it encompasses a broad assortment of scenes categorized into 8 main classes, further subdivided into 89 subclasses, as depicted in Fig. 3.

Additionally, PCOD_1200 encompasses a considerable number of outdoor environments, including grasslands, dead leaves, dead trees, and other habitats, to further consider practical applications. Moreover, the camouflage scenes within PCOD_1200 encapsulate a variety of environmental lighting conditions, captured through passive detection mechanisms that forego any introduction of supplementary polarized light sources. This methodological choice is driven by the intent to authentically replicate natural conditions and real-world applications. To ensure the accurate representation of scene colors, each imaging scenario within the PCOD_1200 dataset is calibrated for white balance using Datacolor's 24-color standard color card.

Each polarization image in PCOD_1200 is captured using the Lucid Triton camera, capable of capturing four polarization images $I_0$, $I_{45}$, $I_{90}$ and $I_{135}$ in one single shot. It is worth nothing that the acquisition of polarization information comes at the expense of spatial resolution, whereby the spatial resolution of each image is 1224 × 1024. Each sample in PCOD_1200 comprises five elements, including the intensity image, an object-level label, DoLP image, AoLP image, and four images I0, I45, I90 and I135 with varying polarization directions. To ensure scene diversity, we captured the dataset from multiple locations, view angles, object sizes, object quantities, and object types.

For a more comprehensive analysis of PCOD_1200, we have conducted a series of statistical examinations focusing on object size and center bias. The object size distribution entails a comprehensive assessment of the area ratio occupied by the object within the entire image. By encompassing objects of various scales, the dataset's inclusivity is enhanced. For PCOD_1200, the object size ranges from 0.01% to 43.4% with an average of 10.2%. The center bias refers to the positional distance between an object and the central viewpoint within the scene. It is inherent in human nature to allocate greater attention to the central region of the field of view. However, real-world camouflage scenarios might entail objects concealed in diverse positions across the surroundings. We count the normalized Euclidean distance between the scene's center and the camouflaged object's barycenter as the center bias. The visual representation of the data distribution within the PCOD_1200 dataset is depicted in Fig. 4. The PCOD_1200 dataset will be promptly uploaded to a publicly accessible online repository, accessible via the link: https://github.com/cvhfut/PCOD_1200.

## 4. Polarization-based framework

The examples presented in Fig. 1 demonstrate the potential of DoLP as a strong physical cue for detecting camouflaged objects. However, incorporating DoLP directly into existing object detection networks may not necessarily lead to the expected performance improvement. This is due to the fact that DoLP can be weak or even absent under certain lighting conditions, resulting in a lack of meaningful cues for detection. In typical cases, both intensity and DoLP can provide informative cues for COD. Thus, it is crucial to effectively and dynamically fuse intensity and DoLP to achieve robust and multimodal COD.

To address this challenge, we propose a novel COD network called the Intensity-DoLP Network (IPNet). IPNet leverages both local and global contextual information to dynamically fuse multimodal intensity and DoLP for robust detection. IPNet follows an encoder–decoder architecture, as depicted in Fig. 5. During the encoding process, we introduce a Prediction Module (PM, Section 4.1) to generate a semi-refined feature that encodes the coarse information of camouflage objects. This semi-refined feature, together with the input DoLP, is then fed into a Enrich Module (ERM, Section 4.2) to generate enriched DoLP features, addressing the issue of weak or non-existent DoLP under certain lighting conditions. Finally, a Cross-modal Fusion Module (CMFM, Section 4.3) is employed to dynamically fuse the extracted local features from both intensity and DoLP inputs, guided by the global features.

During decoding, we rely on global contextual cues to integrate the fused features. To this end, we propose a Cross-level Aggregation Module (CAM, Section 4.4) to reintegrate the cross-modal fusion features at different levels and predict more accurate results.

### 4.1. PM

The primary objective of the PM is to convert the rough camouflage features into semi-refined ones. To achieve this, we first input the intensity image $I \in \mathbb{R}^{W\times H\times 3}$ into a Pyramid-Vision-Transformer-based (PVTv2) backbone (Wang et al., 2021b) and extract a set of features $f_k, k = 1, 2, 3, 4$ from the network. Each feature $f_k$ has a resolution of $H/2^{k+1} \times W/2^{k+1}$, $k = 1, 2, 3, 4$. Next, we use a Receptive Field Block (RFB) to enhance the receptive field in a specific layer and extract richer features. We use the same settings as in Wang et al. (2021b) and reduce the channel number of each feature level to 32. We denote the hierarchical features as $\{f_k^{rf}, k = 1, 2, 3, 4\}$. Then, these features are transferred to the PM for multi-level cross-fusion between features at all levels to obtain semi-refined and edge features, both with one feature channel. The edge features are utilized to promote the semi-refined feature to improve prediction accuracy. Mathematically, the PM is defined as follows:

$$\begin{cases} f_3^{pm} & = Bconvs(Cat(f_3^{rf}, Bconv(Up(f_4^{rf})))) \\ f_2^{pm} & = Bconvs(Cat(f_2^{rf}, Bconv(Up(f_3^{rf})))) \\ f_1^{pm} & = Bconvs(Cat(f_1^{rf}, Bconv(Up(f_2^{rf})))), \\ f_{sr} & = Bconv(f_1^{pm}) \\ f_{edge} & = Bconv(f_1^{pm}) \end{cases} \tag{4}$$

where $f_{sr}$ represents the semi-refined feature and $f_{edge}$ represents the edge feature. Up is the 2× upsample operation using bilinear interpolation, Cat represents the concatenation operation along the channel axis, Bconv is one 3 × 3 convolutional layer, and Bconvs represents one 3 × 3 and one 1 × 1 convolutional layers.

To ensure that the most relevant information is fused, we add intermediate supervisions on $f_{sr}$ and $f_{edge}$, explicitly identifying camouflage objects.

### 4.2. ERM

Given that DoLP may be weak or even non-existent under certain light conditions, we propose the ERM to effectively enrich DoLP, which eventually boosts the camouflage detection performance. Fig. 6 shows the structure of ERM, the DoLP and the semi refined features provided by PM are both fed to the ERM structure. Note that DoLP is expressed as $f_{dolp}$. Thus, ERM is defined by the following formula: In consideration of the potential weakness or absence of the DoLP under certain lighting conditions, we propose the ERM to effectively enhance the DoLP, thereby boosting the performance of camouflage detection. The structure of the ERM is presented in Fig. 6, where both the DoLP and semi refined features obtained from the PM are input into the ERM structure. Denoted as $f_{dolp}$, the DoLP is enriched and expressed as $f_{dolp}^{rich}$, according to the following formula:

$$f_{dolp}^{rich} = f_{dolp} \oplus (f_{dolp} \otimes f_{sr}), \tag{5}$$

where $\oplus$ and $\otimes$ represent element-wise addition and element-wise multiplication, respectively.

The enriched DoLP is then fed into the PVTv2 to extract hierarchical features, which are subjected to RFB for channel reduction, similar to the process applied to the intensity features. The hierarchical features extracted from the enriched DoLP are represented as $RD_k^{rf}, k = 1, 2, 3, 4$ which have the same dimensions as the intensity features $f_k^{rf}, k = 1, 2, 3, 4$ to enable cross-modal hierarchical fusion in subsequent stages. Fig. 6 provides an overview of the entire process. Additionally, intermediate supervisions are implemented to ensure the effectiveness of information fusion and the identification of camouflage objects.

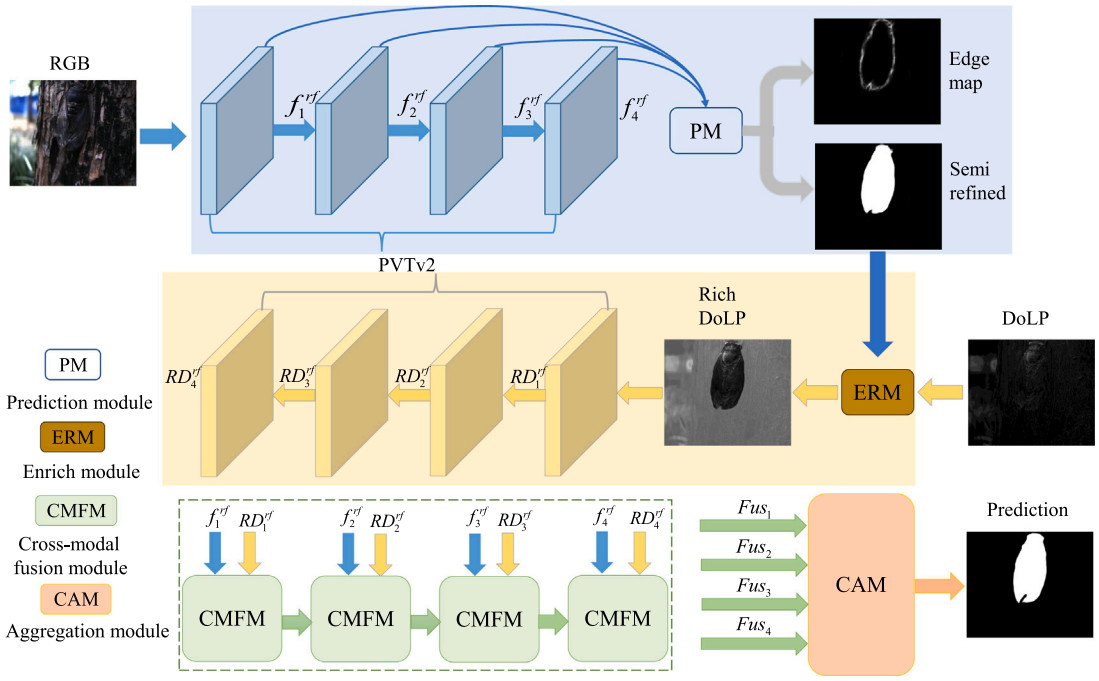



**Fig. 5.** The overall architecture of the proposed network. The PM is employed to generate a partially refined feature that encodes the coarse attributes of camouflage objects, the ERM serves to generate enhanced DoLP features, the CMFM dynamically fuses locally extracted features from both intensity and DoLP inputs, the CAM reinstates cross-modal fusion features across various levels and facilitates more precise predictions. Different colors are utilized to represent diverse categories of information. Blue signifies RGB data, yellow signifies DoLP data, gray signifies edge information, and green signifies the information resulting from the fusion of RGB and DoLP inputs.

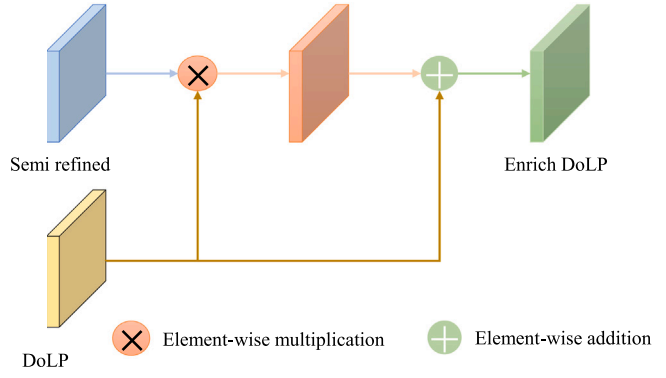



**Fig. 6.** Illustration of the ERM. Due to the potential weakness of DoLP under certain lighting conditions, which can result in a lack of meaningful cues for COD, we multiply the partially refined feature output from the PM with the DoLP information and subsequently enhancing it through a residual structure.

### 4.3. CMFM

The proposed CMFM aims to effectively extract and integrate complementary information from two modes in order to enhance image contrast. Due to the distinct contributions of intensity and DoLP at each level of the network, a mere concatenation or summation of the modal features could negatively impact the network's discriminative ability. To address this issue, an efficient CMFM has been proposed, comprising of four units: a gradient unit (GU), a weight acquisition unit (WCU), and two product units (PU), as shown in Fig. 7. Specifically, at each level $k = 1, 2, 3, 4$, the proposed CMFM takes as input the intensity feature $f_k^{rf}$ and the enriched DoLP feature $RD_k^{rf}$.

The GU is designed to extract gradient amplitude of features and achieve interaction of gradient information between intensity and DoLP. Firstly, the intensity feature $f_k^{rf}$ and enriched DoLP feature $RD_k^{rf}$ are passed through two convolution operations, resulting in $f_k^c$ and $RD_k^c$. These features are concatenated using a $1 \times 1$ convolution to adjust the number of channels, denoted as $frd_k$. The gradient operator (Sobel operator) and two convolution operations ($3 \times 3$ and $1 \times 1$) are then applied to $frd_k$ to produce a fine-grained detail feature map $frd_k^{sob}$. This process can be expressed as:

$$frd_k^{sob} = frd_k \oplus Bconvs(Sobel(frd_K)), \tag{6}$$

where $\oplus$ denotes element-wise addition.

The WCU computes two weights, $w_{k1}$ and $w_{k2}$, based on information from the two modalities. This is achieved through a series of operations, including two $3 \times 3$ convolution layers, one $1 \times 1$ convolution layer, one global average pooling layer, and one softmax function. A feature-wise attention vector $V_k^{weight} \in R^{1\times1\times2}$ is learned as:

$$V_k^{weight} = \delta(Avgpooling(conv(frd_k^{sob}))), \tag{7}$$

where $\delta$ denotes the softmax function. The weights for the two modalities are then obtained using the splitting operation (chunk) on $V_k^{weight}$, and are represented as $w_{k1}$ and $w_{k2}$:

$$w_{k1}, w_{k2} = chunk(V_k^{weight}). \tag{8}$$

In the PU, useful information is further extracted from the intensity and DoLP features based on the weights obtained by WCU. An outer product of $f_k^c$ and $w_{k1}$ is performed to extract more useful information from intensity features, and a similar operation is applied to $RD_k^c$ and $w_{k2}$. These extracted features are represented as $f_k^{cw}$ and $RD_k^{cw}$, respectively. Finally, the fusion feature of the $k$th level, $Fus_k$, is generated using concatenation and prediction operations.

$$Fus_k = PRE(Cat(PRE(f_k^{cw} \oplus RD_k^{cw}), frd_k)), \tag{9}$$

where $\oplus$ represents the element-wise addition, $Cat$ denotes the concatenation operation along the channel axis, and $PRE$ refers to the prediction operation carried out by a $1 \times 1$ convolution operation. However, experimental results in Section 5 show that simple concatenation alone leads to unsatisfactory performance compared to the CMFM, which exhibits a much stronger feature aggregation ability.

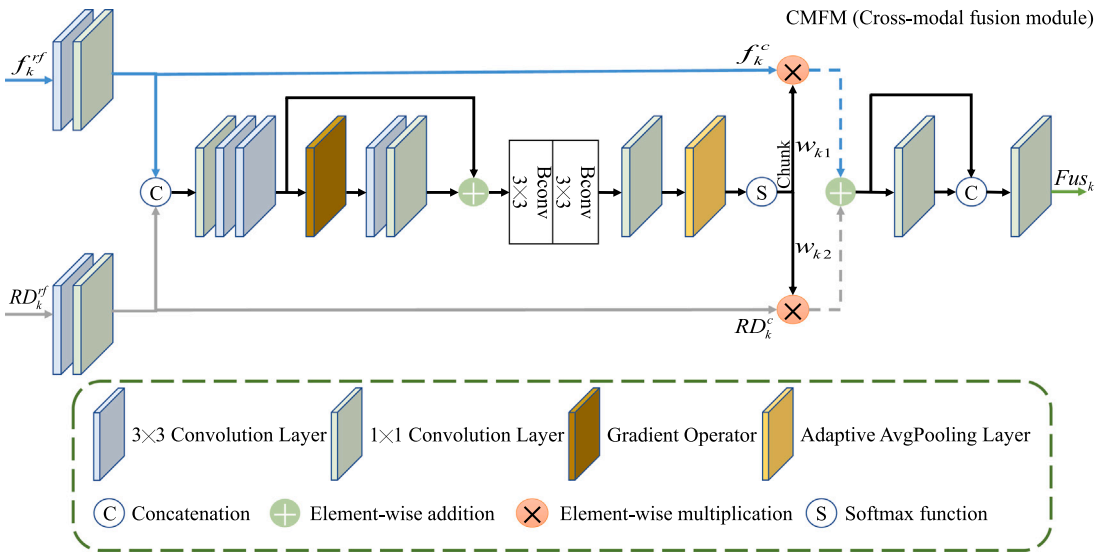



**Fig. 7.** Illustration of the CMFM. By extracting the gradient magnitudes of the concatenated RGB intensity image and DoLP image, we facilitate an interaction between the gradient information of these two modal images. Leveraging an attention mechanism, we proficiently extract and integrate complementary information, thus enhancing image contrast.

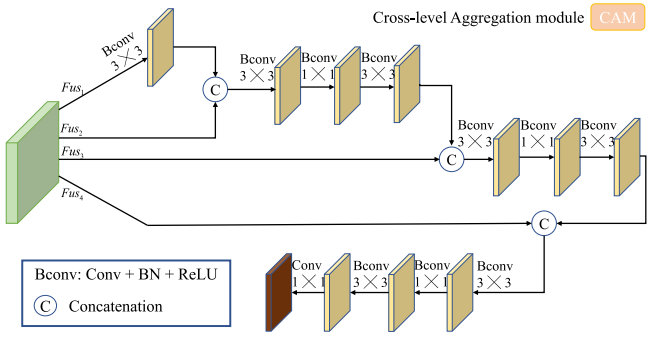



**Fig. 8.** Illustration of the CAM. By re-integrating the cross-modal fusion features at each level, more precise predictions can be achieved.

### 4.4. CAM

The proposed CAM is designed to aggregate the fused cross-modal feature of each level for accurate COD. Fig. 8 shows that CAM consists of a set of layer-wise fusion units.

Firstly, to ensure consistent resolution across all levels, we utilize an upsample operation with bilinear interpolation to reshape the fused features $\{Fus_k\}_{k=1}^{4}$ to the same size. We then apply a feature transformation operation to $Fus_4$, which includes a $3 \times 3$ convolution layer, a batch normalization (BN) (Ioffe and Szegedy, 2015) layer, and a Rectified Linear Unit (ReLU) activation function (Glorot et al., 2011). We concatenate $Fus_3$ and $Fus_4$ as $pred_{34}$, and use two feature transformation operations to further aggregate the two fusion features. The same operation is performed on $Fus_1$ and $Fus_2$. Mathematically, the output of CAM, $pred_{1234}$, is defined as:

$$\begin{cases} pred_{34} & = FT3(FT2(Cat(Fus_3, FT1(Fus_4)))) \\ pred_{234} & = FT3(FT2(Cat(Fus_2, FT1(pred_{34})))) \\ pred_{1234} & = FT3(FT2(Cat(Fus_1, FT1(pred_{234})))) \end{cases}, \tag{10}$$

where FT1, FT2 and FT3 are feature transformation operations, and $Cat$ represents the concatenation operation along the channel axis.

After CAM, we use a UNet-shaped Refine module (Qin et al., 2019) to predict the final camouflage map pred. Additionally, we include supervision to encourage CAM to learn the most discriminative information for COD.

## 5. Experiments and results

In this section, we begin by presenting the loss function, implementation details, and evaluation metrics used in our experiments. We then proceed to conduct ablation studies to demonstrate the effectiveness of our key modules. Finally, we present both quantitative and qualitative results, and compare them with state-of-the-art methods.

### 5.1. Loss function

The loss function utilized in the training of IPNet consists of two components: weighted binary cross-entropy (BCE) loss and structure loss, represented as the summation of the two, i.e. $L_{total} = L^e + \sum_v L_v^s, v \in \{f_{sr}, pred\}$. For the purpose of supervising the camouflaged mask, we employ a weighted BCE and a weighted intersection-over-union (IoU) loss $L^s = L_{BCE}^w + L_{IoU}^w$, with the former serving to counteract the negative impact of data imbalance, while the latter emphasizing the importance of global pixels. Our choice of loss function is consistent with that of Wei et al. (2020). As for the camouflaged edge supervision, we solely apply a weighted BCE loss $L^e = L_{BCE}^w$, as edge information lacks global context and the use of IoU loss may impede the learning process.

### 5.2. Implementation details

In our experiments with the PCOD_1200 dataset, we split it into 970 training images and 230 testing images. To augment the training dataset, we performed horizontal flipping, vertical flipping, and rotation (180°) operations, resulting in a total of 3880 training images. During inference, we resized each set of polarization images with four polarization directions (0°, 45°, 90° and 135°) to $352 \times 352$, and then used Eqs. (1) and (2) to obtain the corresponding RGB and DoLP images to feed into our model.

We implemented IPNet using Pytorch on a NVIDIA RTX 3090Ti GPU with 24 GB memory. The transformer-based feature extraction is initialized by PVTv2 (Wang et al., 2021b), and the remaining modules are initialized in a random manner. We optimized the overall parameters using the Adam optimization algorithm with an initial learning rate of 1e−4. Training the network with a batch size of 12 over 30 epochs took approximately 4 h to converge. Furthermore, the proposed model encompasses 120.16 million parameters (M), with an associated workload of 85.6 billion GFLOPs (Giga Floating-Point Operations), achieving an inference speed of 9.96 frames per second (fps).

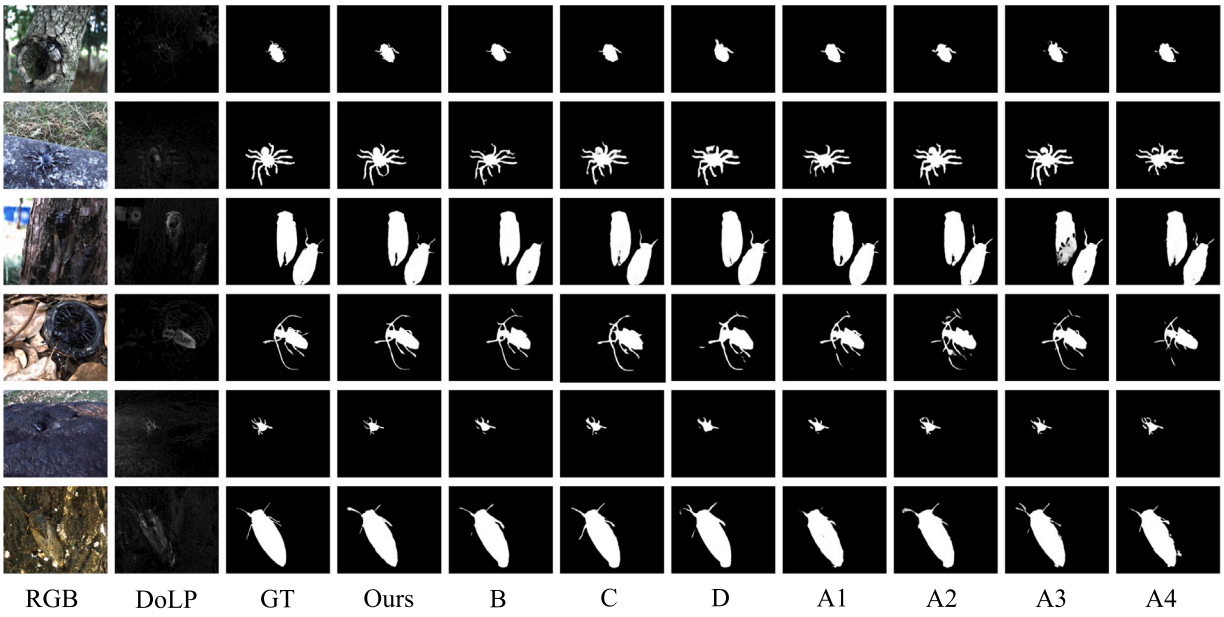



**Fig. 9.** Qualitative evaluation for ablation studies.

### 5.3. Evaluation metrics

We evaluate the efficacy of each approach using four metrics to compare the predicted camouflage map against the ground-truth (GT). The four metrics are described as follows:

(1) Structure-measure ($S_\alpha$) (Fan et al., 2017) is utilized to determine the structural similarity between the predicted map and the corresponding GT. This metric is defined as follows:

$$S_\alpha = (1-\alpha)S_r + \alpha S_0, \tag{11}$$

where $S_r$ and $S_0$ are the region-aware structural similarity and object-aware, respectively. The parameter $\alpha$ is set to 0.5 as Fan et al. (2017).

(2) Mean Enhanced-measure ($E_\phi$) (Fan et al., 2018b) devises an enhanced alignment matrix ($\phi_{FM}$) to capture both image-level statistics and local pixel matching information:

$$E_\phi = \frac{1}{W \times H}\sum_{i=1}^{W}\sum_{j=1}^{H}\phi_{FM}(i,j). \tag{12}$$

(3) Mean Absolute Error ($MAE$) (Perazzi et al., 2012) is a metric used to compute the average difference between the predicted camouflage map and the GT. It is calculated as follows:

$$MAE = \frac{1}{W \times H}\sum_{i=1}^{W}\sum_{j=1}^{H}|C(i,j) - G(i,j)|, \tag{13}$$

where $C(i,j)$ and $G(i,j)$ represent the predicted camouflage value and GT value, $W$ and $H$ denote the dimensions of the image.

(4) F-measure ($F_\beta$) (Achanta et al., 2009) as a harmonic average of precision and recall, defined as follows:

$$F_\beta = \frac{(1+\beta^2)\times P \times R}{\beta^2 \times P \times R}, \tag{14}$$

where precision and recall are denoted by $P$ and $R$, respectively. The parameter $\beta$ is a non-negative weight value, and we set it to $\beta^2 = 0.3$ according to Achanta et al. (2009).

### 5.4. Ablation study

We have conducted experiments to investigate the impact of polarization information on COD using IPNet, as well as the effectiveness of each component of the model. For each experiment, we fully retained the model.

**Table 1**
Quantitative comparisons for different inputs.

| | Methods | $S_\alpha \uparrow$ | $E_\phi \uparrow$ | $MAE \downarrow$ | $F_\beta \uparrow$ |
|---|---|---|---|---|---|
| (A) | IPNet(original) | **0.922** | **0.970** | **0.008** | **0.882** |
| (B) | Input intensity + AoLP | 0.914 | 0.966 | 0.010 | 0.875 |
| (C) | Input intensity + S | 0.907 | 0.965 | 0.011 | 0.860 |
| (D) | Input intensity only | 0.900 | 0.960 | 0.012 | 0.840 |

**Table 2**
Quantitative evaluation for ablation studies.

| Model | ERM | CMFM | CAM | $S_\alpha \uparrow$ | $E_\phi \uparrow$ | $MAE \downarrow$ | $F_\beta \uparrow$ |
|---|---|---|---|---|---|---|---|
| IPNet | √ | √ | √ | **0.922** | **0.970** | **0.008** | **0.882** |
| A1 | | √ | √ | 0.918 | 0.968 | 0.009 | 0.880 |
| A2 | √ | | √ | 0.912 | 0.966 | 0.010 | 0.863 |
| A3 | √ | √ | | 0.916 | 0.968 | 0.009 | 0.876 |
| A4 | | | | 0.910 | 0.964 | 0.009 | 0.866 |

**Impact of polarization information.** To demonstrate the effects of polarization information, we conducted a series of ablation experiments, as shown in Table 1, where (A) represents the IPNet baseline, (B) AoLP is used instead of DoLP, (C) $S$ is used instead of DoLP (where $S_0$, $S_1$ and $S_2$ are concatenated as $S$), and (D) only includes intensity. The comparison of (D) with (A), (B), or (C) reveals that adding any form of polarization information to the intensity information enhances the detection accuracy. Furthermore, we found that DoLP information has a greater impact than AoLP or S information. These quantitative observations are also supported by the visual results shown in Fig. 9.

**Effectiveness of ERM.** The results in Table 2 and Fig. 9 demonstrate that IPNet outperforms its ablated version without ERM (Table 2 A1), which shows the usefulness of the proposed ERM. Without ERM, the DoLP map may be weak or even non-existent under certain light conditions, which degrades the performance.

**Effectiveness of CMFM.** To verify the effectiveness of CMFM, we replace the simple concatenation operation with CMFM. A2 in Table 2 versus IPNet shows that our CMFM improves the performance. Intuitively, A2 in Fig. 9 versus IPNet shows that the predictions produced by adding the CMFM to improve contrast better locate the camouflage object. This advance confirms the superiority of our CMFM in effectively extracting and fusing paired complementary information of two modalities through gradient residuals.

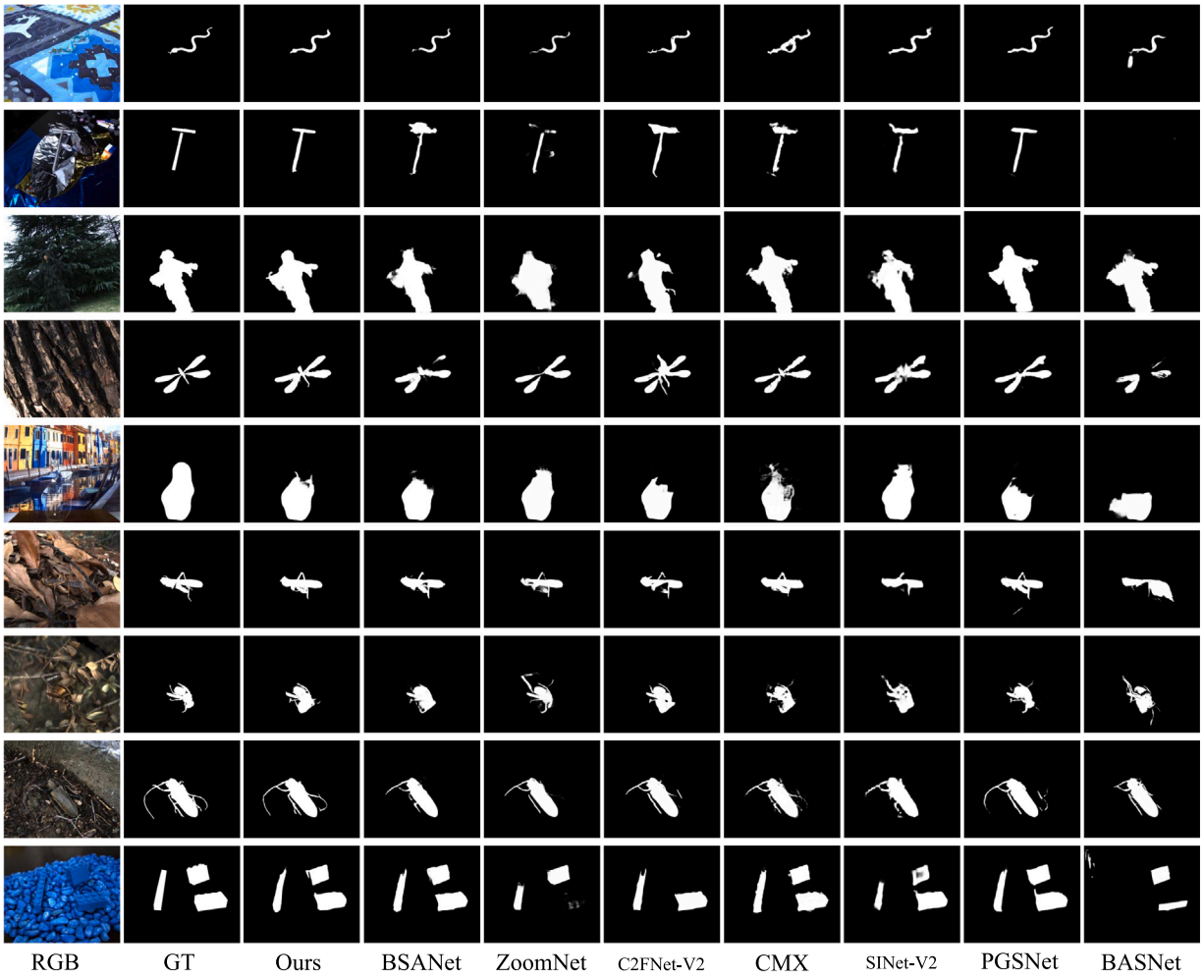



**Fig. 10.** Qualitative comparisons with SOTA methods.

**Effectiveness of CAM.** To provide evidence for the effectiveness of CAM, we simply fuse the four layers of features obtained from CMFM via concatenation operation. A3 in Table 2 versus IPNet shows that our proposed CAM achieves a significant improvement. Meanwhile, A3 in Fig. 9 versus IPNet shows that the predictions produced by CAM contain more complete information, clearly showing that the cross-level feature fusion module is necessary for improving performance.

**Effectiveness of IPNet components.** We demonstrate the influence of ERM, CMFM, and CAM that comprise IPNet by using their basic counterpart. A worst performance (A4 in Table 2) illustrates the importance of each component in IPNet.

### 5.5. Comparisons with state-of-the-art methods

We conduct a comprehensive comparison of our approach with 18 state-of-the-art (SOTA) methods across different related tasks, including 2 semantic segmentation methods (UNet++ Zhou et al., 2018, CMX Zhang et al., 2022), 5 SOD methods (EGNet Zhao et al., 2019, BASNet Qin et al., 2019, CPD Wu et al., 2019, PraNet Fan et al., 2020b, F3Net Wei et al., 2020), 10 COD methods (SINet-V1 Fan et al., 2020a, LSR Lv et al., 2021, PFNet Mei et al., 2021, C2FNet Sun et al., 2021, SINet-V2 Fan et al., 2021, OCENet Liu et al., 2022, ZoomNet Pang et al., 2022, BSANet Zhu et al., 2022, ERRNet Ji et al., 2022, C2FNet-V2 Chen et al., 2022) and 1 glass region segmentation method (Mei et al., 2022). For a fair comparison, all methods are retrained using open-source models on the PCOD_1200 dataset with the recommended hyperparameter settings by the authors.

**Quantitative Evaluation.** As shown in Table 3, our IPNet outperforms all 18 SOTA methods and achieves the best overall performance. Specifically, the fusion strategy based on DoLP and intensity contributes to improving the completeness of predictions, resulting in a 2.3% increase in on PCOD_1200 compared to the basic model, SINet-V2.

**Comparisons of Parameter count and GFLOPs.** Table 4 presents the parameter count and GFLOPs for the proposed model and representative models. Our model offers a performance-robust solution with a relatively modest parameter count for COD tasks. However, there might still exist some redundancy in the design of the inference structure. The application of the dual stream networks might result in additional inference costs.

**Qualitative Evaluation.** Fig. 10 shows the visual comparison of our IPNet with the top 7 competing methods. It can be observed that these methods fail to provide complete segmentation results for low-contrast camouflaged objects. In contrast, the proposed IPNet provides more accurate results closer to the GT due to the fusion strategy.

**Quantitative on CAMO and COD10K.** Comparisons between the proposed method and other different methods on CAMO (Le et al., 2019) and COD10K (Fan et al., 2020a) datasets are presented in Table 5. Since CAMO and COD10K datasets lack polarization information, the comparison experiments on these datasets involve replacing the input of our Dual-flow Network's original DoLP branch with RGB gradient

**Table 3**
Quantitative comparisons with SOTA methods on the PCOD_1200 dataset, with best scores highlighted in red, followed by green and blue. ↑ denotes the score the higher the better, ↓ indicates the lower the better.

| Methods | $S_\alpha$ ↑ | $E_\phi$ ↑ | $MAE$ ↓ | $F_\beta$ ↑ |
|---|---|---|---|---|
| EGNet (Zhao et al., 2019) | 0.861 | 0.902 | 0.015 | 0.787 |
| UNet++(Zhou et al., 2018) | 0.801 | 0.874 | 0.026 | 0.694 |
| BASNet (Qin et al., 2019) | 0.837 | 0.891 | 0.021 | 0.752 |
| CPD (Wu et al., 2019) | 0.855 | 0.899 | 0.016 | 0.784 |
| PraNet (Fan et al., 2020b) | 0.904 | 0.956 | 0.011 | 0.852 |
| F3Net (Wei et al., 2020) | 0.885 | 0.945 | 0.013 | 0.826 |
| SINet-V1 (Fan et al., 2020a) | 0.862 | 0.902 | 0.016 | 0.788 |
| LSR (Lv et al., 2021) | 0.888 | 0.938 | 0.011 | 0.845 |
| PFNet (Mei et al., 2021) | 0.873 | 0.935 | 0.014 | 0.817 |
| C2FNet (Sun et al., 2021) | 0.893 | 0.942 | 0.012 | 0.838 |
| SINet-V2 (Fan et al., 2021) | 0.882 | 0.941 | 0.013 | 0.819 |
| OCENet (Liu et al., 2022) | 0.883 | 0.945 | 0.013 | 0.827 |
| ZoomNet (Pang et al., 2022) | 0.897 | 0.922 | 0.010 | 0.842 |
| BSANet (Zhu et al., 2022) | 0.903 | 0.945 | 0.011 | 0.861 |
| ERRNet (Ji et al., 2022) | 0.833 | 0.901 | 0.023 | 0.704 |
| C2FNet-V2 (Chen et al., 2022) | 0.895 | 0.945 | 0.012 | 0.845 |
| PGSNet (Mei et al., 2022) | 0.916 | 0.965 | 0.010 | 0.868 |
| CMX (Zhang et al., 2022) | 0.922 | 0.965 | 0.009 | 0.876 |
| IPNet | 0.922 | 0.970 | 0.008 | 0.882 |

**Table 4**
Comparisons of Parameter count and GFLOPs across representative models. All evaluations are conducted based on the inference settings specified in the respective papers.

| Item | GFLOPs (G) | #Param (M) |
|---|---|---|
| EGNet (Zhao et al., 2019) | 276.2 | 103.0 |
| BASNet (Qin et al., 2019) | 448.6 | 87.1 |
| PraNet (Fan et al., 2020b) | 12.2 | 29.1 |
| SINet-V1(Fan et al., 2020a) | 38.8 | 48.9 |
| LSR (Lv et al., 2021) | 66.6 | 50.9 |
| PFNet (Mei et al., 2021) | 53.2 | 46.5 |
| SINet-V2(Fan et al., 2021) | 23.8 | 26.9 |
| ZoomNet (Pang et al., 2022) | 203.5 | 32.4 |
| BSANet (Zhu et al., 2022) | 23.2 | 27.5 |
| ERRNet (Ji et al., 2022) | 37.3 | 69.8 |
| C2FNet-V2(Chen et al., 2022) | 16.8 | 44.9 |
| PGSNet (Mei et al., 2022) | 193.8 | 288.7 |
| CMX (Chen et al., 2022) | 25.1 | 63.4 |
| IPNet | 85.6 | 120.2 |

image. It is worth noting that the absence of polarization information significantly impacts the performance of our network. However, even under the condition of not having polarization information as input, the proposed model still achieves the best performance.


**Fig. 11.** Illustration of three failure cases.

### 5.6. Limitations

During the experimental phase, we observed that certain camouflage scenarios within our constructed dataset exhibited poor polarization characteristics due to variations in lighting conditions. This phenomenon resulted in inaccurate predictions by our model, as illustrated in Fig. 11. In the case of PGSNet, its authors devised customized modules to enhance and rectify DoLP images, thereby mitigating similar issues. In contrast, our approach employs an ERM module to enhance DoLP images, but this module heavily relies on predictions from the PM module and lacks tailored treatment or control over lower-quality DoLP images. Moreover, while PCOD_1200 encompasses 1200 instances of camouflage object detection scenarios, and we have augmented the dataset using data augmentation techniques during model training, the current dataset size still imposes certain limitations on performance.

### 5.7. Potential directions

Given the current state of this research, we identify three potential work directions in the future. (1) Further expanding the scale and diversity of the PCOD_1200 dataset, coupled with an elevation in the intricacy and authenticity of camouflage scenarios, will inherently benefit the training of proposed models. (2) Incorporating an additional branch utilizing AoLP images as input could result in a three-input network model, potentially enhancing the effectiveness of COD. (3) Considering the potential of incorporating integrated multi-modal visual sensors, such as the development of an RGB-Polarization-Infrared dataset, offers an opportunity to address camouflage scenarios with insufficient visible light information by introducing additional modalities for data enhancement.

## 6. Conclusion

In conclusion, this paper introduces polarization information for COD task and proposes to fuse RGB intensity and polarization cues for enhancing the contrast between camouflaged objects and the background. To facilitate the evaluation of our method, we construct a challenging and comprehensive PCOD_1200 dataset consisting of 1200 high-quality polarization scenes divided into 89 subclasses, and propose a dual-flow network for efficient feature fusion. We conduct a thorough comparison of our approach with 18 existing COD methods, and the experimental results show that our approach outperforms the SOTAs both quantitatively and qualitatively.

**Table 5**
Quantitative comparisons with SOTA methods on the CAMO and COD10K, with best scores highlighted in red, followed by green and blue. ↑ denotes the score the higher the better, ↓ indicates the lower the better.

| Methods | CAMO | | | | COD10K | | | |
|---|---|---|---|---|---|---|---|---|
| | $S_\alpha$ ↑ | $E_\phi$ ↑ | $MAE$ ↓ | $F_\beta$ ↑ | $S_\alpha$ ↑ | $E_\phi$ ↑ | $MAE$ ↓ | $F_\beta$ ↑ |
| EGNet (Zhao et al., 2019) | 0.732 | 0.768 | 0.104 | 0.583 | 0.737 | 0.779 | 0.056 | 0.509 |
| UNet++ (Wu et al., 2019) | 0.599 | 0.653 | 0.149 | 0.392 | 0.623 | 0.672 | 0.086 | 0.350 |
| BASNet (Qin et al., 2019) | 0.618 | 0.661 | 0.159 | 0.413 | 0.634 | 0.678 | 0.105 | 0.365 |
| CPD (Wu et al., 2019) | 0.693 | 0.738 | 0.112 | 0.626 | 0.734 | 0.802 | 0.052 | 0.608 |
| PraNet (Fan et al., 2020b) | 0.769 | 0.837 | 0.094 | 0.710 | 0.789 | 0.879 | 0.045 | 0.671 |
| F3Net (Wei et al., 2020) | 0.711 | 0.780 | 0.109 | 0.616 | 0.739 | 0.819 | 0.051 | 0.593 |
| SINet-V1 (Fan et al., 2020a) | 0.751 | 0.771 | 0.100 | 0.606 | 0.771 | 0.806 | 0.051 | 0.551 |
| LSR (Lv et al., 2021) | 0.787 | 0.854 | 0.080 | 0.744 | 0.804 | 0.892 | 0.037 | 0.715 |
| PFNet (Mei et al., 2021) | 0.782 | 0.855 | 0.085 | 0.746 | 0.800 | 0.890 | 0.040 | 0.701 |
| SINet-V2 (Fan et al., 2021) | 0.820 | 0.882 | 0.070 | 0.782 | 0.815 | 0.887 | 0.037 | 0.718 |
| OCENet (Liu et al., 2022) | 0.802 | 0.852 | 0.080 | 0.766 | 0.827 | 0.894 | 0.033 | 0.741 |
| ZoomNet (Pang et al., 2022) | 0.820 | 0.877 | 0.066 | 0.794 | 0.838 | 0.888 | 0.029 | 0.766 |
| BSANet (Zhu et al., 2022) | 0.794 | 0.851 | 0.079 | 0.763 | 0.818 | 0.891 | 0.034 | 0.738 |
| ERRNet (Ji et al., 2022) | 0.761 | 0.817 | 0.088 | 0.660 | 0.780 | 0.867 | 0.044 | 0.629 |
| C2FNet-V2 (Chen et al., 2022) | 0.799 | 0.859 | 0.077 | 0.770 | 0.811 | 0.887 | 0.036 | 0.725 |
| IPNet | 0.864 | 0.924 | 0.047 | 0.846 | 0.850 | 0.922 | 0.026 | 0.785 |

## CRediT authorship contribution statement

**Xin Wang:** Writing – original draft, Review , Conceptualization, Methodology, Supervision. **Jiajia Ding:** Writing – original draft, Editing, Methodology, Programming, Data acquisition. **Zhao Zhang:** Writing – review & editing, Data analysis, Investigation. **Junfeng Xu:** Writing – review & editing , Data curation, Optimization. **Jun Gao:** Writing – review & editing, Conceptualization, Validation.

## Declaration of competing interest

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

## Data availability

Upon the acceptance of this manuscript, all the data and code will be promptly uploaded to a publicly accessible online repository, accessible via the link: https://github.com/cvhfut/PCOD_1200.

## Acknowledgments

This work was supported by the National Natural Science Foundation of China (62171178, 61801161, and 61971177), the Natural Science Foundation of Anhui Province, China (1908085QF282) and the Fundamental Research Funds for the Central Universities, China (JZ2020HGTB0048).

## References

Achanta, Radhakrishna, Hemami, Sheila, Estrada, Francisco, Susstrunk, Sabine, 2009. Frequency-tuned salient region detection. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 1597–1604.

Bi, Hongbo, Zhang, Cong, Wang, Kang, Tong, Jinghui, Zheng, Feng, 2021. Rethinking camouflaged object detection: Models and datasets. IEEE Trans. Circuits Syst. Video Technol. 32 (9), 5708–5724.

Blin, Rachel, Ainouz, Samia, Canu, Stéphane, Meriaudeau, Fabrice, 2019. Road scenes analysis in adverse weather conditions by polarization-encoded images and adapted deep learning. In: 2019 IEEE Intelligent Transportation Systems Conference (ITSC). IEEE, pp. 27–32.

Borji, Ali, Cheng, Ming-Ming, Jiang, Huaizu, Li, Jia, 2015. Salient object detection: A benchmark. IEEE Trans. Image Process. 24 (12), 5706–5722.

Buckley, Colman, Fabert, Marc, Kinet, Damien, Kucikas, Vytautas, Pagnoux, Dominique, 2020. Design of an endomicroscope including a resonant fiber-based microprobe dedicated to endoscopic polarimetric imaging for medical diagnosis. Biomed. Opt. Express 11 (12), 7032–7052.

Chen, Geng, Liu, Si-Jie, Sun, Yu-Jia, Ji, Ge-Peng, Wu, Ya-Feng, Zhou, Tao, 2022. Camouflaged object detection via context-aware cross-level fusion. IEEE Trans. Circuits Syst. Video Technol. 32 (10), 6981–6993.

Cheng, Xi, Zhang, Youhua, Chen, Yiqiong, Wu, Yunzhi, Yue, Yi, 2017. Pest identification via deep residual learning in complex background. Comput. Electron. Agric. 141, 351–356.

Cronin, Thomas W, Shashar, Nadav, Caldwell, Roy L, Marshall, Justin, Cheroske, Alexander G, Chiou, Tsyr-Huei, 2003. Polarization vision and its role in biological signaling. Integr. Comp. Biol. 43 (4), 549–558.

Fan, Wang, Ainouz, Samia, Meriaudeau, Fabrice, Bensrhair, Abdelaziz, 2018a. Polarization-based car detection. In: 2018 25th IEEE International Conference on Image Processing (ICIP). IEEE, pp. 3069–3073.

Fan, Deng-Ping, Cheng, Ming-Ming, Liu, Yun, Li, Tao, Borji, Ali, 2017. Structure-measure: A new way to evaluate foreground maps. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4548–4557.

Fan, Deng-Ping, Gong, Cheng, Cao, Yang, Ren, Bo, Cheng, Ming-Ming, Borji, Ali, 2018b. Enhanced-alignment measure for binary foreground map evaluation. arXiv preprint arXiv:1805.10421.

Fan, Deng-Ping, Ji, Ge-Peng, Cheng, Ming-Ming, Shao, Ling, 2021. Concealed object detection. IEEE Trans. Pattern Anal. Mach. Intell. 44 (10), 6024–6042.

Fan, Deng-Ping, Ji, Ge-Peng, Sun, Guolei, Cheng, Ming-Ming, Shen, Jianbing, Shao, Ling, 2020a. Camouflaged object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2777–2787.

Fan, Deng-Ping, Ji, Ge-Peng, Zhou, Tao, Chen, Geng, Fu, Huazhu, Shen, Jianbing, Shao, Ling, 2020b. Pranet: Parallel reverse attention network for polyp segmentation. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part VI 23. Springer, pp. 263–273.

Galun, Meirav, Sharon, Eitan, Basri, Ronen, Brandt, Achi, 2003. Texture segmentation by multiscale aggregation of filter responses and shape elements. In: ICCV, Vol. 3. p. 716.

Giakos, George C, Fraiwan, Luay, Patnekar, N, Sumrain, S, Mertzios, George B, Periyathamby, S, 2004. A sensitive optical polarimetric imaging technique for surface defects detection of aircraft turbine engines. IEEE Trans. Instrum. Meas. 53 (1), 216–222.

Glorot, Xavier, Bordes, Antoine, Bengio, Yoshua, 2011. Deep sparse rectifier neural networks. In: Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics. JMLR Workshop and Conference Proceedings, pp. 315–323.

Hou, Jianqin Yin Yanbin Han Wendi, Li, Jinping, 2011. Detection of the mobile object with camouflage color under dynamic background based on optical flow. Procedia Eng. 15, 2201–2205.

How, Martin J, Christy, John H, Temple, Shelby E, Hemmi, Jan M, Marshall, N Justin, Roberts, Nicholas W, 2015. Target detection is enhanced by polarization vision in a fiddler crab. Curr. Biol. 25 (23), 3069–3073.

How, Martin J, Porter, Megan L, Radford, Andrew N, Feller, Kathryn D, Temple, Shelby E, Caldwell, Roy L, Marshall, N Justin, Cronin, Thomas W, Roberts, Nicholas W, 2014. Out of the blue: the evolution of horizontally polarized signals in haptosquilla (Crustacea, Stomatopoda, Protosquillidae). J. Exp. Biol. 217 (19), 3425–3431.

Ioffe, Sergey, Szegedy, Christian, 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: International Conference on Machine Learning. PMLR, pp. 448–456.

Ji, Ge-Peng, Zhu, Lei, Zhuge, Mingchen, Fu, Keren, 2022. Fast camouflaged object detection via edge-based reversible re-calibration network. Pattern Recognit. 123, 108414.

Kalra, Agastya, Taamazyan, Vage, Rao, Supreeth Krishna, Venkataraman, Kartik, Raskar, Ramesh, Kadambi, Achuta, 2020. Deep polarization cues for transparent object segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8602–8611.

Kavitha, Ch, Rao, B. Prabhakara, Govardhan, A., 2011. An efficient content based image retrieval using color and texture of image sub blocks. Int. J. Eng. Sci. Technol. (IJEST) 3 (2), 1060–1068.

Kim, Sungho, 2015. Unsupervised spectral-spatial feature selection-based camouflaged object detection using VNIR hyperspectral camera. Sci. World J. 2015.

Kinoshita, Michiyo, Yamazato, Kei, Arikawa, Kentaro, 2011. Polarization-based brightness discrimination in the foraging butterfly, papilio xuthus. Philos. Trans. R. Soc. B 366 (1565), 688–696.

Le, Trung-Nghia, Nguyen, Tam V, Nie, Zhongliang, Tran, Minh-Triet, Sugimoto, Akihiro, 2019. Anabranch network for camouflaged object segmentation. Comput. Vis. Image Underst. 184, 45–56.

Li, Shuai, Florencio, Dinei, Li, Wanqing, Zhao, Yaqin, Cook, Chris, 2018. A fusion framework for camouflaged moving foreground detection in the wavelet domain. IEEE Trans. Image Process. 27 (8), 3918–3930.

Liu, Zhou, Huang, Kaiqi, Tan, Tieniu, 2012. Foreground object detection using top-down information based on EM framework. IEEE Trans. Image Process. 21 (9), 4204–4217.

Liu, Jiawei, Zhang, Jing, Barnes, Nick, 2022. Modeling aleatoric uncertainty for camouflaged object detection. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1445–1454.

Lv, Yunqiu, Zhang, Jing, Dai, Yuchao, Li, Aixuan, Liu, Bowen, Barnes, Nick, Fan, Deng-Ping, 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, Haiyang, Dong, Bo, Dong, Wen, Yang, Jiaxi, Baek, Seung-Hwan, Heide, Felix, Peers, Pieter, Wei, Xiaopeng, Yang, Xin, 2022. Glass segmentation using intensity and spectral polarization cues. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12622–12631.

Mei, Haiyang, Ji, Ge-Peng, Wei, Ziqi, Yang, Xin, Wei, Xiaopeng, Fan, Deng-Ping, 2021. Camouflaged object segmentation with distraction mining. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8772–8781.

Pang, Youwei, Zhao, Xiaoqi, Xiang, Tian-Zhu, Zhang, Lihe, Lu, Huchuan, 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.

Perazzi, Federico, Krähenbühl, Philipp, Pritch, Yael, Hornung, Alexander, 2012. Saliency filters: Contrast based filtering for salient region detection. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 733–740.

Qin, Xuebin, Zhang, Zichen, Huang, Chenyang, Gao, Chao, Dehghan, Masood, Jagersand, Martin, 2019. Basnet: Boundary-aware salient object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7479–7489.

Ren, Jingjing, Hu, Xiaowei, Zhu, Lei, Xu, Xuemiao, Xu, Yangyang, Wang, Weiming, Deng, Zijun, Heng, Pheng-Ann, 2021. Deep texture-aware features for camouflaged object detection. IEEE Trans. Circuits Syst. Video Technol. 1157–1167.

Roberts, Nicholas W, How, Martin J, Porter, Megan L, Temple, Shelby E, Caldwell, Roy L, Powell, Samuel B, Gruev, Viktor, Marshall, N Justin, Cronin, Thomas W, 2014. Animal polarization imaging and implications for optical processing. Proc. IEEE 102 (10), 1427–1434.

Roy, Arunabha M, Bhaduri, Jayabrata, Kumar, Teerath, Raj, Kislay, 2023. WilDect-YOLO: An efficient and robust computer vision-based accurate object localization model for automated endangered wildlife detection. Ecol. Inform. 75, 101919.

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

Song, Liming, Geng, Weidong, 2010. A new camouflage texture evaluation method based on WSSIM and nature image features. In: 2010 International Conference on Multimedia Technology. IEEE, pp. 1–4.

Stevens, Martin, Merilaita, Sami, 2009. Animal camouflage: current issues and new perspectives. Philos. Trans. R. Soc. B 364 (1516), 423–427.

Sun, Yujia, Chen, Geng, Zhou, Tao, Zhang, Yi, Liu, Nian, 2021. Context-aware cross-level fusion network for camouflaged object detection. arXiv preprint arXiv:2105.12555.

Tyo, J Scott, Goldstein, Dennis L, Chenault, David B, Shaw, Joseph A, 2006. Review of passive imaging polarimetry for remote sensing applications. Appl. Opt. 45 (22), 5453–5469.

Wang, Kang, Bi, Hongbo, Zhang, Yi, Zhang, Cong, Liu, Ziqi, Zheng, Shuang, 2021a. C-net: A dual-branch, dual-guidance and cross-refine network for camouflaged object detection. IEEE Trans. Ind. Electron. 69 (5), 5364–5374.

Wang, Wenhai, Xie, Enze, Li, Xiang, Fan, Deng-Ping, Song, Kaitao, Liang, Ding, Lu, Tong, Luo, Ping, Shao, Ling, 2021b. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 568–578.

Wehner, Rüdiger, Müller, Martin, 2006. The significance of direct sunlight and polarized skylight in the ant's celestial system of navigation. Proc. Natl. Acad. Sci. 103 (33), 12575–12579.

Wei, Jun, Wang, Shuhui, Huang, Qingming, 2020. $F^3$net: fusion, feedback and focus for salient object detection. In: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34. pp. 12321–12328.

Wu, Zhe, Su, Li, Huang, Qingming, 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.

Xiang, Kaite, Yang, Kailun, Wang, Kaiwei, 2021. Polarization-driven semantic segmentation via efficient attention-bridged fusion. Opt. Express 29 (4), 4802–4820.

Yan, Jinnan, Le, Trung-Nghia, Nguyen, Khanh-Duy, Tran, Minh-Triet, Do, Thanh-Toan, Nguyen, Tam V, 2021. Mirrornet: Bio-inspired camouflaged object segmentation. IEEE Access 9, 43290–43300.

Yan, Lei, Li, Yanfei, Chandrasekar, V, Mortimer, Hugh, Peltoniemi, Jouni, Lin, Yi, 2020. General review of optical polarization remote sensing. Int. J. Remote Sens. 41 (13), 4853–4864.

Yang, Fan, Zhai, Qiang, Li, Xin, Huang, Rui, Luo, Ao, Cheng, Hong, Fan, Deng-Ping, 2021. Uncertainty-guided transformer reasoning for camouflaged object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4146–4155.

Zhai, Qiang, Li, Xin, Yang, Fan, Chen, Chenglizhao, Cheng, Hong, Fan, Deng-Ping, 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, Jiaming, Liu, Huayao, Yang, Kailun, Hu, Xinxin, Liu, Ruiping, Stiefelhagen, Rainer, 2022. CMX: Cross-modal fusion for RGB-X semantic segmentation with transformers. arXiv preprint arXiv:2203.04838.

Zhang, Yifei, Morel, Olivier, Blanchon, Marc, Seulin, Ralph, Rastgoo, Mojdeh, Sidibé, Désiré, 2019. Exploration of deep learning-based multimodal fusion for semantic road scene segmentation. pp. 336–343.

Zhao, Jia-Xing, Liu, Jiang-Jiang, Fan, Deng-Ping, Cao, Yang, Yang, Jufeng, Cheng, Ming-Ming, 2019. EGNet: Edge guidance network for salient object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8779–8788.

Zhou, Zongwei, Rahman Siddiquee, Md Mahfuzur, Tajbakhsh, Nima, Liang, Jianming, 2018. Unet++: A nested u-net architecture for medical image segmentation. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4. Springer, pp. 3–11.

Zhu, Hongwei, Li, Peng, Xie, Haoran, Yan, Xuefeng, Liang, Dong, Chen, Dapeng, Wei, Mingqiang, Qin, Jing, 2022. I can find you! boundary-guided separated attention network for camouflaged object detection. In: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 36. pp. 3608–3616.