Depth-Aware Concealed Crop Detection in Dense Agricultural Scenes

Liqiong Wang[1,2†], Jinyu Yang[3,4†], Yanfu Zhang[5], Fangyi Wang[1,2*], Feng Zheng[3*]
[1]Hubei Key Laboratory of Intelligent Vision Based Monitoring for Hydroelectric Engineering, China Three Gorges University
[2]College of Computer and Information Technology, China Three Gorges University
[3]Southern University of Science and Technology [4]University of Birmingham
[5]Departmental of Computer Science, College of William and Mary
{liqiong.wang11,jinyu.yang96}@outlook.com yzhang105@wm.edu fy_wang@ctgu.edu.cn f.zheng@ieee.org

Abstract

Concealed Object Detection (COD) aims to identify objects visually embedded in their background. Existing COD datasets and methods predominantly focus on animals or humans, ignoring the agricultural domain, which often contains numerous, small, and concealed crops with severe occlusions. In this paper, we introduce Concealed Crop Detection (CCD), which extends classic COD to agricultural domains. Experimental study shows that unimodal data provides insufficient information for CCD. To address this gap, we first collect a large-scale RGB-D dataset, ACOD-12K, containing high-resolution crop images and depth maps. Then, we propose a foundational framework named Recurrent Iterative Segmentation Network (RISNet). To tackle the challenge of dense objects, we employ multi-scale receptive fields to capture objects of varying sizes, thus enhancing the detection performance for dense objects. By fusing depth features, our method can acquire spatial information about concealed objects to mitigate disturbances caused by intricate backgrounds and occlusions. Furthermore, our model adopts a multi-stage iterative approach, using predictions from each stage as gate attention to reinforce position information, thereby improving the detection accuracy for small objects. Extensive experimental results demonstrate that our RISNet achieves new state-of-the-art performance on both newly proposed CCD and classic COD tasks. All resources will be available at https://github.com/Kki2Eve/RISNet.

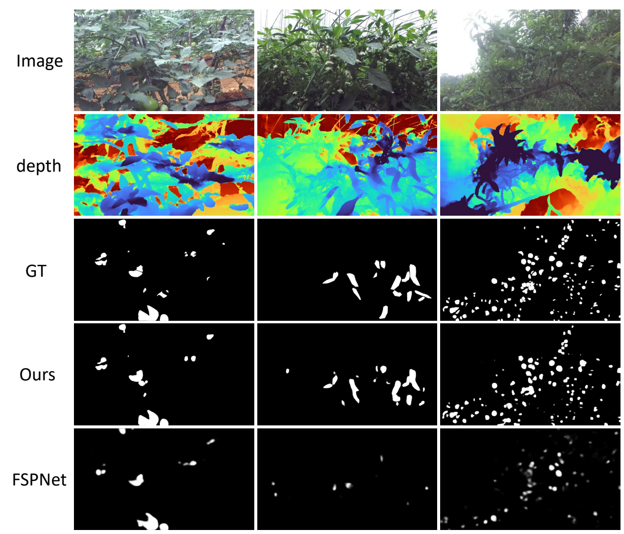

Figure 1. Results of FSPNet [28] and our RISNet for CCD. It is evident that our method is better equipped to tackle the challenges posed by severe occlusion and densely distributed small objects, resulting in superior performance.

1. Introduction

With the advancement of smart agriculture, there is a growing interest in integrating computer vision with agriculture [32]. Driven by economic demands, high-density planting of crops is becoming inevitable in agricultural production processes. Consequently, the analysis and understanding of dense scenes in agricultural vision problems [2, 4, 21, 33, 66] assume heightened significance. In these dense agricultural scenes, numerous small crops are concealed in the surrounding environment, causing significant interference in monitoring production [14].

Existing COD methodologies [13, 14, 22, 26, 28, 29, 31, 36, 42, 49, 51, 57, 65, 69, 75] primarily emphasize animal camouflage strategies, while CCD shifts its focus to densely packed small objects heavily occluded within complex scenes. As illustrated in Fig. 1, the state-of-the-art COD method can only generate inaccurate prediction maps

† Equal contribution. ∗ Corresponding author.
This work was done during Liqiong Wang visited SUSTech VIP lab.

(Col 1 and 3) and even fails to detect concealed objects (Col 2). This issue stems from the fact that objects in CCD do not employ the same camouflage strategy as animals. COD methods struggle to mitigate interference from occlusion and effectively capture visual features related to the densely packed small objects in complex environments.

In this paper, we introduce a new benchmark named Concealed Crop Detection (CCD), designed for identifying concealed objects in dense agricultural scenes. We observe that unimodal information lacks the capacity to discern subtle distinctions between objects and backgrounds. To overcome this limitation, we integrate depth maps to supplement spatial information absent in RGB data. The geometric priors from depth maps effectively mitigate interference caused by noise, thereby enhancing CCD performance.

To facilitate research on CCD, we have curated an extensive RGB-D dataset, *ACOD-12K*. Leveraging the ZED2i depth camera during fieldwork, we capture 6092 images of concealed objects within dense agricultural scenes, simultaneously recording corresponding depth images. As shown in Tab. 1, in comparison to the existing COD datasets, *ACOD-12K* exhibits several advantages:

- *ACOD-12K* is the sole existing multi-modal COD dataset.
- *ACOD-12K* is the largest-scale COD dataset with the highest image resolution among the existing datasets.
- *ACOD-12K* boasts a higher object density, with these objects situated in diverse scenes and distributed randomly across different positions within the images.
- In contrast to the current COD datasets, *ACOD-12K* focuses on the distinctive challenges presented by concealed objects in dense agricultural scenes.

CCD primarily faces four key challenges. Firstly, CCD scenes involve dense objects, where multiple objects of the same category are distributed across the image at varying distances, resulting in varying sizes for identical objects. Secondly, the challenge of intricate backgrounds arises, as objects closely resemble the background, creating a high level of background noise. Thirdly, severe occlusion compounds the complexity, as objects are concealed not only by intricate backgrounds but also by mutual occlusions among themselves. Lastly, small objects significantly increase the difficulty of precise detection.

To tackle these challenges, we introduce RISNet, a baseline method designed specifically for the CCD task. RISNet utilizes multi-scale, multi-modal, and multi-iteration approaches to discern subtle distinctions between objects and backgrounds, yielding robust detection outcomes. Specifically, we leverage multi-scale receptive fields to capture feature information of different-sized concealed objects, effectively addressing the challenges associated with dense objects. To handle complex backgrounds and occlusions, we incorporate depth data to enhance RGB information, providing crucial spatial context and emphasizing discriminative details. To address small objects, we employ a multi-iteration approach. We use detection results from the preceding iteration as gate attention to learn the position information of small objects, iteratively refining the detection results. Experiments show that RISNet outperforms all considered algorithms, demonstrating its effectiveness on CCD.

In summary, our contributions are listed as follows:

- We introduce a benchmark of Concealed Crop Detection (CCD), extending COD into agriculture and making COD more flexible and practical in real-world scenarios.
- To advance research on CCD, we introduce a new large-scale RGB-D dataset *ACOD-12K*, which is the first multi-modal dataset on COD tasks.
- We propose a new baseline framework, RISNet, which achieves new state-of-the-art performance on both classic COD and newly proposed CCD tasks.

2. Related Work

Concealed Object Detection. Concealed object detection(COD) aims to identify objects that closely blend with the background, relying on subtle distinctions. To tackle this challenge, researchers have explored various methodologies. In the early stages of research, the prevalent approach involved manually crafted artificial features for COD [23, 46, 48]. However, these methods exhibit limited robustness, heavily relying on specific handcrafted feature information, making them susceptible to complex scenarios. With the advent of large-scale datasets in COD [13, 34], deep learning methods have surpassed traditional handcrafted feature-based approaches. These methods fall into three categories. The first category involves biomimetic networks. [13] drew inspiration from animal hunting processes, employing a progressive search and recognition approach to uncover concealed objects. [49] mimicked human behavior by zooming in and out to extract visual information. The second category focuses on intricately designed network architectures. [42] modeled concealment levels, contributing to a deeper understanding of visual information. [65] leveraged Bayesian distributions and attention mechanisms to handle uncertainty, enhancing concealed object detection. The third category introduces additional information to boost performance. [36] introduced joint training of SOD and COD, utilizing conflicting information to improve model performance. [22, 51, 75] incorporated edge information, elevating the precision of concealed object localization. *Unlike animals in classic COD employing various camouflage strategies for active concealment, dense agricultural scenes primarily involve the passive concealment of densely distributed small objects with complex backgrounds and severe occlusions. Existing COD models face limitations in addressing these challenges in new scenes, stemming from different kinds of objects [16, 30].*

Concealed Object Detection Datasets. There are currently

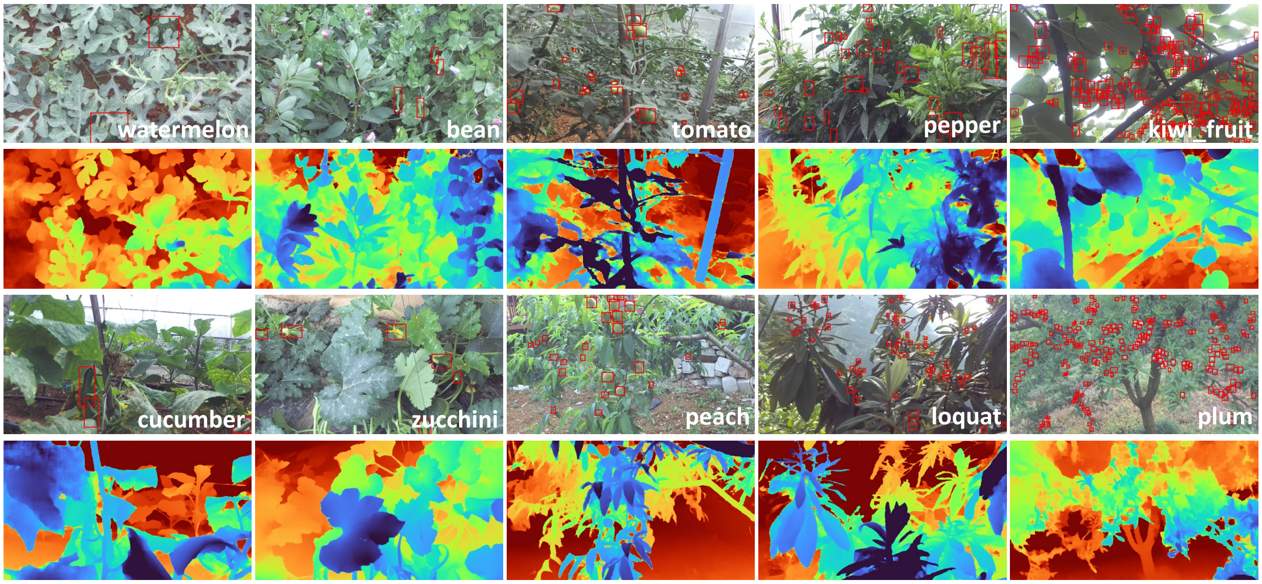

Figure 2. Example images from the proposed *ACOD-12K*. The concealed objects increase gradually from the *left* to the *right* column.

four existing COD datasets: *CHAMELEMON* [50], *CAMO* [34], *COD10K* [13] and *NC4K* [42]. *CHAMELEMON* [50] is an unreleased dataset comprising 76 concealed images downloaded from the internet. *CAMO* [34] consists of 1250 concealed images across eight categories in both natural and artificial scenes, with 1000 images allocated for the training set and 250 for the testing set. *COD10K* [13] is the largest and most challenging COD dataset, featuring 5066 concealed images spanning 69 categories. Among these, 3040 images are designated for training, while 2026 are reserved for testing. *NC4K* [42] serves as a comprehensive COD test set, comprising 4121 images designed to thoroughly assess the generalization capabilities of COD models. *As mentioned in [14], in the early stages of crop growth, many fruits share a visual similarity with green leaves, complicating production monitoring for farmers. The absence of relevant concealed object datasets in dense agricultural scenes hinders existing models from achieving optimal detection results. The proposal of ACOD-12K addresses this limitation, aiming to advance COD research.*

Dense Scenes. In computer vision, dense scenes often pose various challenging problems. One prominent task in dense scene visual analysis is counting, which includes extensively studied areas like crowd counting [5, 20, 27, 37–39, 43, 56, 64, 71], as well as specialized tasks such as vehicle counting [24, 47], penguin counting [3], plant counting [41], and cell counting [1]. Unlike counting, detection tasks in dense scenes are relatively uncommon. For instance, [63] introduced *DOTA*, a large-scale aerial image dataset tailored for object detection, with regions featuring a high concentration of instances, significantly amplifying the detection challenge. Similarly, [19] focused on precise object detection in artificially dense scenes, introducing *SKU-110K*, a novel dataset designed for retail-dense scenarios. *Due to high background noise and severe occlusions, dense agricultural scenes present greater challenges than typical dense scenes. According to [18], single-modal RGB data is susceptible to environmental interference. Thus, we capture RGB-D data to leverage multi-modal information, enhancing the model's comprehension of dense agricultural settings and improving CCD performance.*

Dataset	Year	Img	Avg.Res.	Free View	Mul.	Object Statistics Total	Min	Avg	Max	Link
CHAMELEON[50]	2018	76	742 × 981	✗	✗	79	1	1	3	N/A
CAMO[34]	2019	1250	509 × 653	✓	✗	1368	1	1	7	Link
COD10K[13]	2020	5066	737 × 964	✓	✗	5899	1	1	8	Link
NC4K[42]	2021	4121	530 × 709	✓	✗	4584	1	1	8	Link
ACOD-12K(Ours)	2023	6092	1080 × 1920	✓	✓	71417	1	11	412	Link

Table 1. Statistics of related datasets. "Avg.Res." indicates average resolution and "Mul." stands for multimodality.

3. Proposed Dataset

3.1. Image Collection

CCD encounters challenges in dense agricultural scenes, including dense objects (DO), complex backgrounds (CB), occlusions (OC) and small objects (SO). Due to the scarcity of suitable datasets, existing COD methods fail to deliver competitive results in such agricultural environments. To facilitate research on CCD, we conducted fieldwork and generated a comprehensive RGB-D dataset, *ACOD-12K*, the co-distribution of challenges is shown in Fig. 3.

In summary, we meticulously recorded 128 high-quality videos of varying durations across multiple orchards and farms using the ZED2i depth camera. Drawing from the significance of high-resolution priors in object edge and boundary detection, as highlighted in [55, 70], we maintained dataset effectiveness by capturing images at 1080×1920 resolution during the filming. After acquiring the videos, we preprocessed them to extract both RGB images and depth maps from the left camera's perspective. To curate a representative dataset, we selected one image for every 90 frames. Subsequently, we implemented a multi-stage filtering process to guarantee that each chosen image featured concealed objects in the foreground and had a clear and usable depth map. This filtering process consisted of three rounds: in the initial round, five researchers conducted the primary selection. Subsequently, two experts conducted a detailed review to eliminate any unsuitable images. Finally, a third round of selection, performed by a professional, concluded the entire dataset cleaning process. Our dataset now comprises 6092 RGB images, each paired with a corresponding depth map. See Fig. 2 for example images.

3.2. Image Annotation

Our data annotation process aims to provide mask annotations for all concealed objects in the images. Following [13], we adopt a multi-stage annotation method, *i.e.*, category $\rightarrow$ bounding box $\rightarrow$ mask. This ensures the precision and comprehensiveness of data labeling.

The annotation process consists of five steps. Initially, 700 images are selected, and three researchers annotate concealed objects in these images using bounding boxes to familiarize themselves with the process. Once these 700 images receive satisfactory annotations, the process advances to the next step. The entire dataset is then divided into three parts, with each researcher responsible for annotating one part. Following this, the researchers exchange their respective dataset portions, conduct a thorough review, and discuss any challenging annotations. Subsequently, a professional annotation company adds mask annotations to the dataset, building upon the existing bounding box annotations. In the final step, researchers perform a comprehensive review, rectifying any missed or inadequate annotations.

3.3. Dataset Information

ACOD-12K comprises 6092 images showcasing concealed agricultural objects spanning ten categories. We allocate 4600 images for training and reserve 1492 for testing. Notably, *ACOD-12K* is a groundbreaking RGB-D COD dataset, setting new standards for challenging datasets in the field. All images in our dataset are of high resolution, measuring 1080×1920 pixels, with over 82% containing small objects. Within our dataset, detection difficulty correlates with object-background similarity and density. For instance, watermelons, cucumbers, and zucchinis are straightforward, while peppers and plums pose challenges.

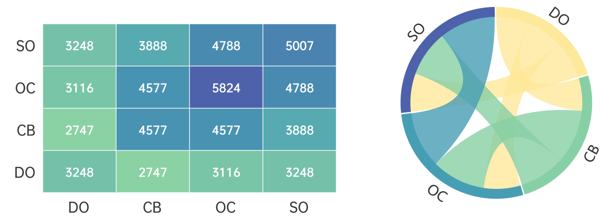

Figure 3. Left: Co-distribution of challenges in *ACOD-12K*, with numbers indicating total images per grid. Right: Multi-dependencies among challenges, with arc length indicates correlation probability.

4. Methodology

4.1. Overview

The holistic structure of our RISNet is depicted in Fig. 4. Given an input image and its corresponding depth map, we utilize the Concealed Feature Encoder (CFE) to extract multi-level feature information. To comprehensively capture information about dense objects, we integrate the Atrous Spatial Pyramid Pooling (ASPP) module [8], leveraging multi-scale receptive fields for detecting objects of varying sizes. The feature map is then fed into the Depth-Guided Feature Decoder (DFD). During this stage, RGB features are merged with depth features and passed through the cascaded residual decoder. This cascade decoder alleviates background and occlusion interference, enhancing the model's detection capacity for dense objects. For improved small object detection, we employ the Iterative Feature Refine (IFR) approach, using the results from the previous detection stage as gate attention to help the model accurately identify the features of small objects. In this paper, the number of model iterations is set to 3.

4.2. Concealed Feature Encoder

Following the success of the Transformer [53] in NLP, researchers are increasingly exploring their adaptation for computer vision tasks [10]. Similar to HitNet [26], we utilize the Pyramid Vision Transformer (PVT) [58] as the feature encoder. Initially, we convert the depth map into a three-channel image using a basic gray color mapping. For input images $f_r, f_d \in \mathbb{R}^{B\times 3\times H\times W}$, following [17], we concatenate the RGB map f_r and depth map f_d along the batch dimension. This ensures the model focuses on the shared regions of interest in both the RGB and depth modalities. Then we pass them through the base encoder, resulting in the feature set $\{f_k\}_{k=1}^{4} \in \mathbb{R}^{2B\times 3\times \frac{H}{2^{k+1}}\times \frac{W}{2^{k+1}}}$.

In dense agricultural scenes, multiple objects are dis-

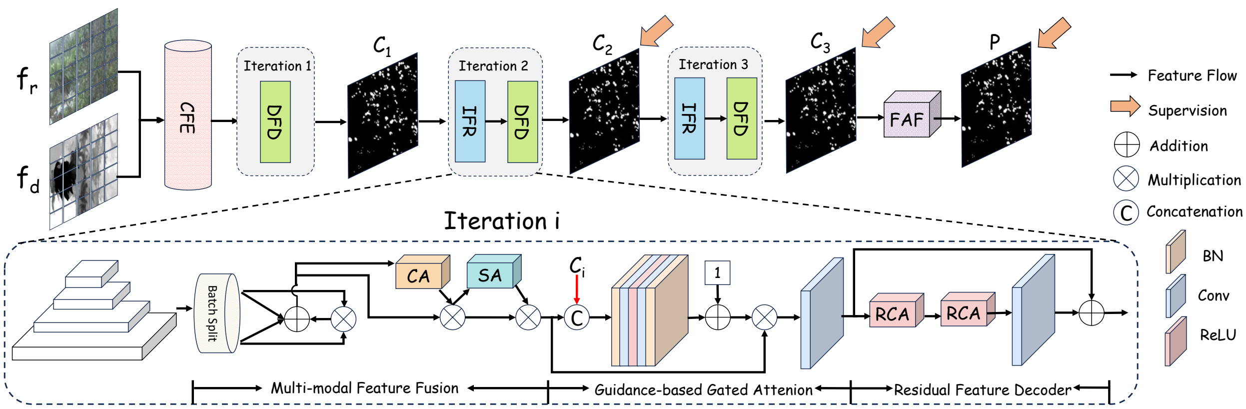

Figure 4. Overview of the proposed RISNet. Given input images f_r and depth images f_d, CFE is utilized to extract object features across multiple scales. DFD consists of MFF and RFD, during the feature decoding stage, MFF is employed to deeply integrate features from both modalities, followed by the progressive fusion of decoded features using RFD, from top to bottom, to yield a preliminary prediction C_i for the input image. IFR further enhances feature recognition iteratively by backpropagating the coarse prediction C_i. After multiple iterations, the final prediction image P is derived. Refer to §4 for more details.

tributed randomly across various locations within the image, resulting in these objects appearing in varying sizes. Due to occlusions between objects and between objects and the background, objects of the same category may exhibit different shapes, introducing substantial interference with our predictions. Different from HitNet [26], to comprehensively gather information from objects located at diverse positions within the image, we leverage the ASPP [8] architecture after obtaining multi-level features. With ASPP, we perform sampling on different levels of feature information using dilated convolutions with varying sampling rates. This effectively allows us to leverage multi-scale receptive fields to capture information from the input features at different scales and perceive contextual information at various proportions, ultimately yielding the feature set $\{f_k\}_{k=1}^{4} \in \mathbb{R}^{2B \times C \times \frac{H}{2^{k+1}} \times \frac{W}{2^{k+1}}}$.

4.3. Depth-Guided Feature Decoder

4.3.1 Multi-modal Feature Fusion

In dense agricultural scenes, complex background noise and significant occlusion are unavoidable challenges. To attain precise detection results, it is crucial to alleviate these disturbances. The advancement of depth cameras has made it more cost-effective to access depth images, which provide essential geometric prior knowledge to help models effectively understand complex scenes. To combine features from both modalities, we devise a Multi-modal Feature Fusion (MFF) module. As illustrated in Fig. 4, we separate the extracted features $\{f_k\}_{k=1}^{4}$ along the batch dimension, reverting them to RGB features $\{f_k^r\}_{k=1}^{4}$ and depth features $\{f_k^d\}_{k=1}^{4}$. We observe that a mere concatenation along the channel dimension would lead to a model bias towards the RGB modality, which runs counter to our fusion objectives. Following [17], we use element-wise addition to explore complementarity of f_k^r and f_k^d, and element-wise multiplication to explore commonality of f_k^r and f_k^d:

$$f_k^{f'} = f_k^r \oplus f_k^d \oplus \left(f_k^r \otimes f_k^d\right), \tag{1}$$

where $\oplus$ denotes element-wise addition, $\otimes$ denotes element-wise multiplication. After obtaining the preliminary fused feature $\{f_k^{f'}\}_{k=1}^{4}$, we employ a dual attention mechanism [59] that encompasses both channel and spatial domains to further integrate noteworthy features. Consequently, the final fused feature $\{f_k^f\}_{k=1}^{4}$ is represented as:

$$f_k^f = (f_k^{f'} \otimes CA(f_k^{f'})) \otimes SA(f_k^{f'} \otimes CA(f_k^{f'})), \tag{2}$$

where CA($\cdot$) denotes channel attention module, SA($\cdot$) denotes spatial attention module.

4.3.2 Residual Feature Decoder

In CNN decoding, each feature channel is conventionally treated uniformly, but their importance varies across tasks. According to [25], explicitly modeling the interdependencies between feature channels enhances the representational capacity of the network. Inspired by [72], we integrate the Residual In Residual (RIR) structure into the decoding process. To capture subtle visual features for object-background discrimination, we directly propagate low-frequency information through long skip connections. Additionally, we employ residual channel attention to dynamically allocate channel weights, emphasizing the most relevant features. To tackle dense object challenges, we implement a multi-level cascaded decoder. Each decoder level focuses on different object scales, with higher-level decoder outputs serving as auxiliary features for lower-level decoders, enhancing the perception of dense objects. Specifically, our residual cascaded decoder is designed as follows:

$$\begin{aligned} f_3^o &= g_4 \oplus Conv3(RCA(g_4)), \\ f_2^o &= g_3 \oplus Conv3(RCA(Con(g_3, f_3^o)), \\ f_1^o &= g_2 \oplus Conv3(RCA(Con(g_2, f_2^o)), \\ C_i &= Up(Conv1(CBR(f_1^o)), \end{aligned} \tag{3}$$

where $\{g_k\}_{k=2}^4$ represents the outputs of the guidance-based gated attenion module, $\{f_k^o\}_{k=1}^3$ denotes the decoder outputs, RCA($\cdot$) refers to the residual channel attention module, Con($\cdot$) signifies channel concatenation, Conv3($\cdot$) is a 3×3 convolution, Conv1($\cdot$) is a 1×1 convolution, CBR($\cdot$) indicates stacked "Conv-BN-ReLU" layers, Up($\cdot$) denotes upsampling, and $\{C_i\}_{i=1}^3$ represents the coarse prediction maps generated by the model.

4.4. Iterative Feature Refinement

4.4.1 Guidance-based Gated Attenion

Deep network architectures tend to amalgamate various types of information, such as color, shape, and texture, during the prediction process. This blending of diverse information can potentially cause the model to overlook specific object details, thereby impairing its capacity to discern vital features [52], especially when dealing with small objects. For more effective small object detection, we implement a Guidance-based Gated Attention (GGA) module to learn location information specific to these objects:

$$g_k = Conv1(f_k^f \otimes (\sigma(GA(Con(f_k^f, C_i))) \oplus 1)), \tag{4}$$

where GA($\cdot$) refers to "BN-Conv-Relu-Conv-BN" layers, and σ denotes the sigmoid function.

4.4.2 Iterative Refinement Mechanism

When observing small objects in images, humans often start by roughly determining their position and then iteratively refine the details, resulting in a comprehensive observation. Drawing inspiration from this human observation strategy, we incorporate an iterative mechanism to enhance the model's detection of small objects. After obtaining the coarse prediction map C_i, we propagate it backward through the network, utilizing GGA to pinpoint the location information of small objects. This assists the model in focusing on the feature information within the object region. This process is iteratively performed to acquire a more accurate coarse detection map. Given that lower-level features contain finer details, we fuse the bottom-level feature f_1 with the final coarse prediction map C_3 to obtain the ultimate fusion result, denoted as P:

$$P = FAF(Con(f_1^f, C_3)), \tag{5}$$

Model	Publications	ACOD-12K		
		$S_\alpha \uparrow$	$F_\beta^\omega \uparrow$	$E_\theta \uparrow$
Concealed Object Detection				
SINet[13]	CVPR20	0.745	0.474	0.826
MGL[67]	CVPR21	0.808	0.685	0.872
PFNet[45]	CVPR21	0.805	0.685	0.942
UGTR[65]	ICCV21	0.798	0.632	0.858
SINet-V2[14]	TPAMI22	0.804	0.691	0.947
C2FNet[7]	TCSVT22	0.833	0.746	0.947
PreyNet[69]	MM22	0.832	0.760	0.937
SegMaR[31]	CVPR22	0.799	0.677	0.930
ZoomNet[49]	CVPR22	0.832	0.747	0.934
DaCOD[57]	MM23	0.803	0.705	0.910
PopNet[61]	ICCV23	**0.844**	**0.778**	**0.955**
HitNet[26]	AAAI23	**0.853**	**0.787**	**0.955**
FSPNet[28]	CVPR23	0.719	0.526	0.819
RGB-D Salient Object Detection				
CLNet[68]	ICCV21	0.826	0.747	0.936
SPNet[74]	ICCV21	0.818	0.731	0.949
DCMF[54]	TIP22	0.779	0.631	0.872
HINet[6]	PR22	0.776	0.651	0.853
SPSN[35]	ECCV22	0.834	0.739	0.930
CIRNet[9]	TIP22	0.794	0.675	0.865
HIDANet[60]	TIP23	0.822	0.734	0.950
XMSNet[62]	MM23	**0.844**	0.754	**0.961**
Ours		**0.866**	**0.803**	**0.967**

Table 2. Quantitative comparisons of different methods on CCD task. The best three results are highlighted in **red**, **blue** and **green**.

where FAF($\cdot$) denotes the Feature Adaptive Fusion module, comprising the ASPP module and convolution operations. FAF is designed for the multi-scale fusion of low-level semantic information to enhance detection results.

4.5. Loss Function

The loss function of our RISNet primarily consists of the loss from the coarse prediction maps $\{C_i\}_{i=1}^3$ and the loss from the final prediction map P. According to [14], to better detect challenging pixels, we employ weighted binary cross-entropy loss $\mathcal{L}_{BCE}^\omega$ and weighted intersection-over-union loss $\mathcal{L}_{IoU}^\omega$ to supervise the prediction results, so our detection loss $\mathcal{L}_d = \mathcal{L}_{BCE}^\omega + \mathcal{L}_{IoU}^\omega$. Following [26], we apply different weights to supervise its coarse prediction maps at different iterative stages. Overall, given weight parameters γ, our total loss function is formulated as:

$$\mathcal{L}_{total} = \mathcal{L}_d(P, GT) + \sum_{i=2}^{3}(\gamma \times i)(\mathcal{L}_d(C_i, GT)), \tag{6}$$

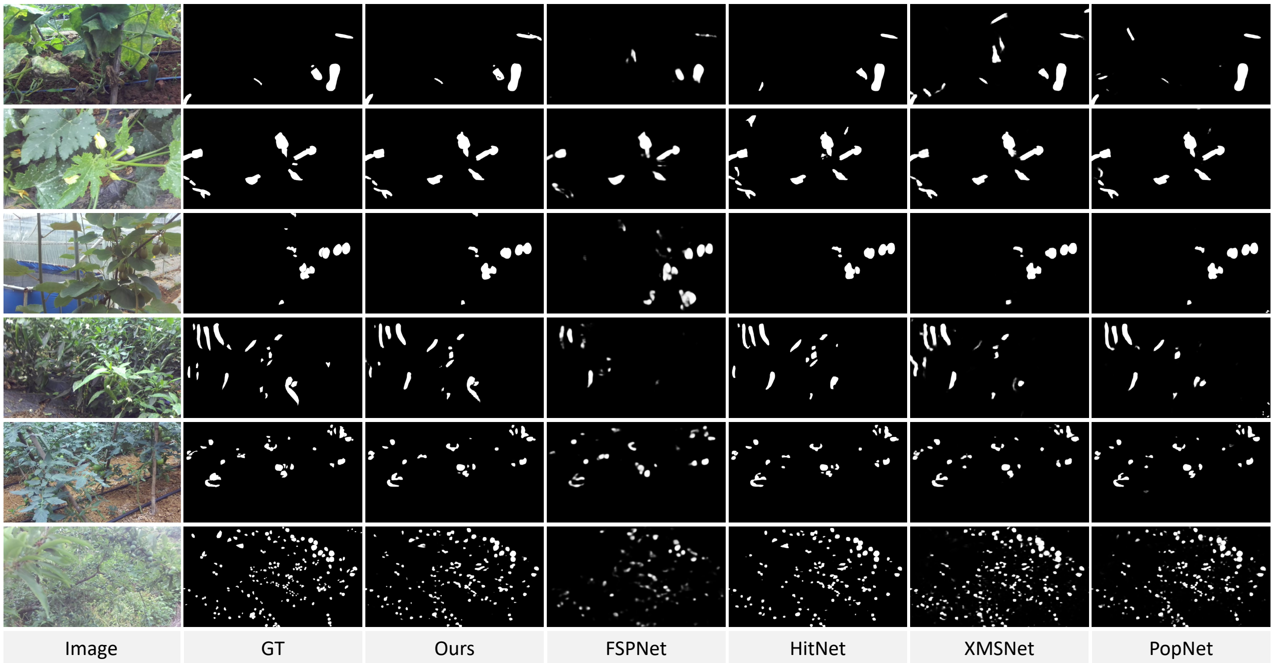

Figure 5. Visual comparisons with recent COD and RGB-D SOD methods on different types of samples. Please zoom in for more details.

Model	Publications	CAMO				COD10K				NC4K			
		$S_\alpha \uparrow$	$F_\beta^\omega \uparrow$	$E_\theta \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$F_\beta^\omega \uparrow$	$E_\theta \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$F_\beta^\omega \uparrow$	$E_\theta \uparrow$	$M \downarrow$
SINet[13]	CVPR20	0.745	0.644	0.804	0.092	0.776	0.631	0.864	0.043	0.808	0.723	0.871	0.058
LSR[42]	CVPR21	0.787	0.696	0.838	0.080	0.804	0.673	0.880	0.037	0.840	0.766	0.895	0.048
R-MGL[67]	CVPR21	0.775	0.673	0.812	0.088	0.814	0.666	0.852	0.035	0.833	0.740	0.867	0.052
JSCOD[36]	CVPR21	0.800	0.728	0.859	0.073	0.809	0.684	0.884	0.035	0.842	0.771	0.898	0.047
PFNet[45]	CVPR21	0.782	0.695	0.841	0.085	0.800	0.660	0.877	0.040	0.829	0.745	0.887	0.053
ZoomNet[49]	CVPR22	0.820	0.752	0.877	0.066	0.838	0.729	0.888	0.029	0.853	0.784	0.896	0.043
FDNet[73]	CVPR22	0.841	0.775	0.895	0.063	0.840	0.729	**0.919**	0.030	0.834	0.750	0.893	0.052
SegMaR[31]	CVPR22	0.815	0.753	0.874	0.071	0.833	0.724	0.899	0.034	0.841	0.781	0.896	0.046
DGNet[29]	MIR23	0.839	0.769	0.901	**0.057**	0.822	0.693	0.896	0.033	0.857	0.784	0.911	0.042
PopNet[61]	ICCV23	0.808	0.744	0.859	0.077	**0.851**	**0.757**	0.910	0.028	0.861	0.802	0.910	0.042
DaCOD[57]	MM23	**0.855**	0.796	**0.905**	**0.051**	0.840	0.729	0.907	0.028	**0.874**	0.814	**0.924**	**0.035**
HitNet[26]	AAAI23	0.844	**0.801**	**0.902**	**0.057**	**0.868**	**0.798**	**0.932**	**0.024**	0.870	**0.825**	**0.921**	**0.039**
FEDER[22]	CVPR23	0.822	0.738	0.886	0.067	**0.851**	0.716	0.917	0.028	0.863	0.789	0.917	0.042
FSPNet[28]	CVPR23	**0.856**	**0.799**	0.899	**0.050**	**0.851**	0.735	0.895	**0.026**	**0.879**	**0.816**	0.915	**0.035**
Ours		**0.870**	**0.827**	**0.922**	**0.050**	**0.873**	**0.799**	**0.931**	**0.025**	**0.882**	**0.834**	**0.925**	**0.037**

Table 3. Detailed comparison results of different methods on COD task. The best three results are highlighted in **red**, **blue** and **green**.

5. Experiment

5.1. Experimental Settings

Evaluation metrics. Traditional COD tasks typically use four evaluation metrics, which include mean absolute error M, weighted F-measure F_β^ω [44], mean E-measure E_θ [12] [15], and structure measure S_α [11]. However, the specific challenges in dense agricultural scenes, characterized by predominantly small objects, diminish the suitability of M. Even in the absence of many object pixels within high-resolution images, M has a limited impact, thereby reducing its effectiveness as an evaluation metric for CCD. Consequently, in experiments, we exclusive the M measure.

Implementation details. RISNet is implemented in PyTorch on an RTX 3090 GPU with the AdamW optimizer [40]. The training process spans 100 epochs with a batch size of 4, initiating with a learning rate of 1e-4 and dividing it by 10 every 50 epochs. The feedback loss weight parameter γ is set to 0.2, and the model undergoes three optimization iterations. To minimize information loss, high-resolution input at 704×704 pixels is employed.

5.2. Results on CCD

Quantitative Analysis. Tab. 2 presents quantitative results for our proposed RISNet compared to 11 state-of-the-art COD models and 8 state-of-the-art RGB-D SOD models on the CCD dataset. To ensure fairness, all compared models are trained using their default settings, and test results are evaluated using the same code. Clearly, our approach consistently outperforms other methodologies, resulting in significant improvements on both the S_α and F_β^ω metrics. On average, it surpasses the second-ranking method, HitNet, by 1.45%. It is noteworthy that, in contrast to single-modal approaches, RGB-D techniques exhibit a lower F_β^ω, Nevertheless, our F_β^ω still surpasses all single-modal methods, exceeding the second-best RGB-D method, XMSNet, by a substantial 4.9%. This can be attributed to the efficacy of our multi-scale deep-level modality integration, affirming the robustness of our model.

Qualitative Analysis. As shown in Fig. 5, visual comparisons of various methods on typical concealed objects are presented. These concealed objects are arranged by their density, ranging from sparse to dense. These objects are generally small, prone to severe occlusion, and possess colors similar to the background. Such challenges, prevalent in CCD images, can potentially confound existing COD and RGB-D SOD techniques, resulting in issues like incorrect detections and missing results. Visual results intuitively demonstrate that, in comparison to other approaches, our method delivers more comprehensive and accurate detection outcomes with clearer object outlines, showcasing the superior performance of our approach.

5.3. Ablation Study

Effect of RISNet. Following [14], we train our RISNet using 3040 images from *COD10K* and 1000 images from *CAMO*, excluding the multi-modal fusion module. Subsequently, tests are conducted on the remaining images. Remarkably, even without the multi-modal fusion module, RISNet maintains state-of-the-art performance. This can be attributed to the architecture of our model, which leverages multi-scale and multi-level feature information while iteratively optimizing detection results. Experimental results are presented in Tab. 3.

Effect of Each Module. In our proposed RISNet, we incorporate three crucial modules. We investigate their individual impacts on model performance systematically. Tab. 4 illustrates the effects of systematically disabling these modules. "w/o CFE" replaces PVT with Res2Net-50 and removes ASPP for multi-scale object information perception. "w/o DFD" involves simply concatenating information from the two modalities instead of using our carefully designed MFF for in-depth modality fusion, accompanied by the exclusion of our RFD module. "w/o IFR" removes the iterative optimization process, directly outputting prediction results. Results show an expected decline when each module is deactivated, emphasizing the significance of the collaborative efforts among these modules. Their synergy is vital for achieving optimal detection results, validating the rationale and effectiveness of our module design.

Metric	w/o CFE	w/o DFD	w/o IFR	RISNet
$S_\alpha \uparrow$	0.855	0.861	0.850	**0.866**
$E_\theta \uparrow$	0.964	0.965	0.949	**0.967**
$F_\beta^\omega \uparrow$	0.785	0.790	0.785	**0.803**

Table 4. Ablation study of Each Module.

Metric	in=1	in=2	in=3	in=4	in=5
$F_\beta^\omega \uparrow$	0.792	0.793	**0.803**	0.802	0.802

Table 5. Ablation study of Iteration Number.

Evaluation of Iteration Number. In Tab. 5, we illustrate the impact of iteration number on model performance in our iterative optimization mechanism. The results reveal a gradual improvement with an increased number of iterations. Considering both performance and efficiency, we find that 3 iterations represent the optimal choice.

6. Conclusion

We analyze and address the limitations inherent in classical COD tasks, particularly their inadequacy in dealing with concealed objects in agricultural environments. Building upon this foundation, we introduce a new benchmark called Concealed Crop Detection (CCD), aiming to identify concealed crops in dense agricultural settings. To facilitate CCD research, we compile a large-scale RGB-D agricultural concealed object dataset, *ACOD-12K*. We propose an effective baseline model, RISNet, which integrates depth information to unearth subtle visual cues for distinguishing concealed objects from the background. RISNet achieves state-of-the-art performance on both COD and CCD tasks, demonstrating the effectiveness of our framework. The CCD task we introduce extends classical COD tasks into the agricultural domain, opening up new applications such as crop growth monitoring, automated agricultural harvesting, weed control, and more.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant NO. 62122035).

References

[1] Mohammad Mahmudul Alam and Mohammad Tariqul Islam. Machine learning approach of automatic identification and counting of blood cells. *Healthcare technology letters*, 6 (4):103–108, 2019. 3

[2] Boaz Arad, Jos Balendonck, Ruud Barth, Ohad Ben-Shahar, Yael Edan, Thomas Hellström, Jochen Hemming, Polina Kurtser, Ola Ringdahl, Toon Tielen, et al. Development of a sweet pepper harvesting robot. *Journal of Field Robotics*, 37 (6):1027–1039, 2020. 1

[3] Carlos Arteta, Victor Lempitsky, and Andrew Zisserman. Counting in the wild. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII 14*, pages 483–498. Springer, 2016. 3

[4] C Wouter Bac, Eldert J Van Henten, Jochen Hemming, and Yael Edan. Harvesting robots for high-value crops: State-of-the-art review and challenges ahead. *Journal of Field Robotics*, 31(6):888–911, 2014. 1

[5] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017. 3

[6] Hongbo Bi, Ranwan Wu, Ziqi Liu, Huihui Zhu, Cong Zhang, and Tian-Zhu Xiang. Cross-modal hierarchical interaction network for rgb-d salient object detection. *Pattern Recognition*, 136:109194, 2023. 6

[7] Geng Chen, Si-Jie Liu, Yu-Jia Sun, Ge-Peng Ji, Ya-Feng Wu, and Tao Zhou. Camouflaged object detection via context-aware cross-level fusion. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(10):6981–6993, 2022. 6

[8] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 4, 5

[9] Runmin Cong, Qinwei Lin, Chen Zhang, Chongyi Li, Xiaochun Cao, Qingming Huang, and Yao Zhao. Cir-net: Cross-modality interaction and refinement for rgb-d salient object detection. *IEEE Transactions on Image Processing*, 31:6800–6815, 2022. 6

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 4

[11] Deng-Ping Fan, Ming-Ming Cheng, Yun Liu, Tao Li, and Ali Borji. Structure-measure: A new way to evaluate foreground maps. In *Proceedings of the IEEE international conference on computer vision*, pages 4548–4557, 2017. 7

[12] Deng-Ping Fan, Cheng Gong, Yang Cao, Bo Ren, Ming-Ming Cheng, and Ali Borji. Enhanced-alignment measure for binary foreground map evaluation. *arXiv preprint arXiv:1805.10421*, 2018. 7

[13] Deng-Ping Fan, Ge-Peng Ji, Guolei Sun, Ming-Ming Cheng, Jianbing Shen, and Ling Shao. Camouflaged object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2777–2787, 2020. 1, 2, 3, 4, 6, 7

[14] Deng-Ping Fan, Ge-Peng Ji, Ming-Ming Cheng, and Ling Shao. Concealed object detection. *IEEE transactions on pattern analysis and machine intelligence*, 44(10):6024–6042, 2021. 1, 3, 6, 8

[15] Deng-Ping Fan, Ge-Peng Ji, Xuebin Qin, and Ming-Ming Cheng. Cognitive vision inspired object segmentation metric and loss function. *Scientia Sinica Informationis*, 6(6), 2021. 7

[16] Deng-Ping Fan, Ge-Peng Ji, Peng Xu, Ming-Ming Cheng, Christos Sakaridis, and Luc Van Gool. Advances in deep concealed scene understanding. *Visual Intelligence*, 1(1):16, 2023. 2

[17] Keren Fu, Deng-Ping Fan, Ge-Peng Ji, Qijun Zhao, Jianbing Shen, and Ce Zhu. Siamese network for rgb-d salient object detection and beyond. *IEEE transactions on pattern analysis and machine intelligence*, 44(9):5541–5559, 2021. 4, 5

[18] Longsheng Fu, Fangfang Gao, Jingzhu Wu, Rui Li, Manoj Karkee, and Qin Zhang. Application of consumer rgb-d cameras for fruit detection and localization in field: A critical review. *Computers and Electronics in Agriculture*, 177: 105687, 2020. 3

[19] Eran Goldman, Roei Herzig, Aviv Eisenschtat, Jacob Goldberger, and Tal Hassner. Precise detection in densely packed scenes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5227–5236, 2019. 3

[20] Tao Han, Lei Bai, Junyu Gao, Qi Wang, and Wanli Ouyang. Dr. vic: Decomposition and reasoning for video individual counting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3083–3092, 2022. 3

[21] Leif O Harders, Vitali Czymmek, Florian J Knoll, and Stephan Hussmann. Area yield performance evaluation of a nonchemical weeding robot in organic farming. In *2021 IEEE International Instrumentation and Measurement Technology Conference (I2MTC)*, pages 1–6. IEEE, 2021. 1

[22] Chunming He, Kai Li, Yachao Zhang, Longxiang Tang, Yulun Zhang, Zhenhua Guo, and Xiu Li. Camouflaged object detection with feature decomposition and edge reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22046–22055, 2023. 1, 2, 7

[23] Jianqin Yin Yanbin Han Wendi Hou and Jinping Li. Detection of the mobile object with camouflage color under dynamic background based on optical flow. *Procedia Engineering*, 15:2201–2205, 2011. 2

[24] Meng-Ru Hsieh, Yen-Liang Lin, and Winston H Hsu. Drone-based object counting by spatially regularized regional proposal network. In *Proceedings of the IEEE international conference on computer vision*, pages 4145–4153, 2017. 3

[25] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 5

[26] Xiaobin Hu, Shuo Wang, Xuebin Qin, Hang Dai, Wenqi Ren, Donghao Luo, Ying Tai, and Ling Shao. High-resolution iterative feedback network for camouflaged object detection.

In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 881–889, 2023. 1, 4, 5, 6, 7

[27] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 3

[28] Zhou Huang, Hang Dai, Tian-Zhu Xiang, Shuo Wang, Huai-Xin Chen, Jie Qin, and Huan Xiong. Feature shrinkage pyramid for camouflaged object detection with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5557–5566, 2023. 1, 6, 7

[29] Ge-Peng Ji, Deng-Ping Fan, Yu-Cheng Chou, Dengxin Dai, Alexander Liniger, and Luc Van Gool. Deep gradient learning for efficient camouflaged object detection. *Machine Intelligence Research*, 20(1):92–108, 2023. 1, 7

[30] Ge-Peng Ji, Deng-Ping Fan, Peng Xu, Ming-Ming Cheng, Bowen Zhou, and Luc Van Gool. Sam struggles in concealed scenes–empirical study on" segment anything". *arXiv preprint arXiv:2304.06022*, 2023. 2

[31] Qi Jia, Shuilian Yao, Yu Liu, Xin Fan, Risheng Liu, and Zhongxuan Luo. Segment, magnify and reiterate: Detecting camouflaged objects the hard way. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4713–4722, 2022. 1, 6, 7

[32] Andreas Kamilaris and Francesc X Prenafeta-Boldú. Deep learning in agriculture: A survey. *Computers and electronics in agriculture*, 147:70–90, 2018. 1

[33] Polina Kurtser, Ola Ringdahl, Nati Rotstein, Ron Berenstein, and Yael Edan. In-field grape cluster size assessment for vine yield estimation using a mobile robot and a consumer level rgb-d camera. *IEEE Robotics and Automation Letters*, 5(2): 2031–2038, 2020. 1

[34] Trung-Nghia Le, Tam V Nguyen, Zhongliang Nie, Minh-Triet Tran, and Akihiro Sugimoto. Anabranch network for camouflaged object segmentation. *Computer vision and image understanding*, 184:45–56, 2019. 2, 3

[35] Minhyeok Lee, Chaewon Park, Suhwan Cho, and Sangyoun Lee. Spsn: Superpixel prototype sampling network for rgb-d salient object detection. In *European Conference on Computer Vision*, pages 630–647. Springer, 2022. 6

[36] Aixuan Li, Jing Zhang, Yunqiu Lv, Bowen Liu, Tong Zhang, and Yuchao Dai. Uncertainty-aware joint salient object and camouflaged object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10071–10081, 2021. 1, 2, 7

[37] Yuhong Li, Xiaofan Zhang, and Deming Chen. Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1091–1100, 2018. 3

[38] Jiang Liu, Chenqiang Gao, Deyu Meng, and Alexander G Hauptmann. Decidenet: Counting varying density crowds through attention guided detection and density estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5197–5206, 2018.

[39] Weizhe Liu, Mathieu Salzmann, and Pascal Fua. Context-aware crowd counting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5099–5108, 2019. 3

[40] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 7

[41] Hao Lu, Zhiguo Cao, Yang Xiao, Bohan Zhuang, and Chunhua Shen. Tasselnet: counting maize tassels in the wild via local counts regression network. *Plant methods*, 13(1):1–17, 2017. 3

[42] Yunqiu Lv, Jing Zhang, Yuchao Dai, Aixuan Li, Bowen Liu, Nick Barnes, and Deng-Ping Fan. Simultaneously localize, segment and rank the camouflaged objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11591–11601, 2021. 1, 2, 3, 7

[43] Zhiheng Ma, Xing Wei, Xiaopeng Hong, and Yihong Gong. Bayesian loss for crowd count estimation with point supervision. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6142–6151, 2019. 3

[44] Ran Margolin, Lihi Zelnik-Manor, and Ayellet Tal. How to evaluate foreground maps? In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 248–255, 2014. 7

[45] Haiyang Mei, Ge-Peng Ji, Ziqi Wei, Xin Yang, Xiaopeng Wei, and Deng-Ping Fan. Camouflaged object segmentation with distraction mining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8772–8781, 2021. 6, 7

[46] Ajoy Mondal, Susmita Ghosh, and Ashish Ghosh. Partially camouflaged object tracking using modified probabilistic neural network and fuzzy energy based active contour. *International Journal of Computer Vision*, 122:116–148, 2017. 2

[47] Daniel Onoro-Rubio and Roberto J López-Sastre. Towards perspective-free object counting with deep learning. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII 14*, pages 615–629. Springer, 2016. 3

[48] Yuxin Pan, Yiwang Chen, Qiang Fu, Ping Zhang, Xin Xu, et al. Study on the camouflaged target detection method based on 3d convexity. *Modern Applied Science*, 5(4):152, 2011. 2

[49] Youwei Pang, Xiaoqi Zhao, Tian-Zhu Xiang, Lihe Zhang, and Huchuan Lu. 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*, pages 2160–2170, 2022. 1, 2, 6, 7

[50] Przemysław Skurowski, Hassan Abdulameer, J Błaszczyk, Tomasz Depta, Adam Kornacki, and P Kozieł. Animal camouflage analysis: Chameleon database. *Unpublished manuscript*, 2(6):7, 2018. 3

[51] Yujia Sun, Shuo Wang, Chenglizhao Chen, and Tian-Zhu Xiang. Boundary-guided camouflaged object detection. *arXiv preprint arXiv:2207.00794*, 2022. 1, 2

[52] Towaki Takikawa, David Acuna, Varun Jampani, and Sanja Fidler. Gated-scnn: Gated shape cnns for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5229–5238, 2019. 6

[53] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 4

[54] Fengyun Wang, Jinshan Pan, Shoukun Xu, and Jinhui Tang. Learning discriminative cross-modality features for rgb-d saliency detection. *IEEE Transactions on Image Processing*, 31:1285–1297, 2022. 6

[55] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3349–3364, 2020. 4

[56] Qi Wang, Junyu Gao, Wei Lin, and Xuelong Li. Nwpu-crowd: A large-scale benchmark for crowd counting and localization. *IEEE transactions on pattern analysis and machine intelligence*, 43(6):2141–2149, 2020. 3

[57] Qingwei Wang, Jinyu Yang, Xiaosheng Yu, Fangyi Wang, Peng Chen, and Feng Zheng. Depth-aided camouflaged object detection. In *ACM Multimedia*, pages 3297–3306. ACM, 2023. 1, 6, 7

[58] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 568–578, 2021. 4

[59] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018. 5

[60] Zongwei Wu, Guillaume Allibert, Fabrice Meriaudeau, Chao Ma, and Cédric Demonceaux. Hidanet: Rgb-d salient object detection via hierarchical depth awareness. *IEEE Transactions on Image Processing*, 32:2160–2173, 2023. 6

[61] Zongwei Wu, Danda Pani Paudel, Deng-Ping Fan, Jingjing Wang, Shuo Wang, Cédric Demonceaux, Radu Timofte, and Luc Van Gool. Source-free depth for object pop-out. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1032–1042, 2023. 6, 7

[62] Zongwei Wu, Jingjing Wang, Zhuyun Zhou, Zhaochong An, Qiuping Jiang, Cédric Demonceaux, Guolei Sun, and Radu Timofte. Object segmentation by mining cross-modal semantics. In *ACMMM*, 2023. 6

[63] Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. Dota: A large-scale dataset for object detection in aerial images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3974–3983, 2018. 3

[64] Haipeng Xiong, Hao Lu, Chengxin Liu, Liang Liu, Zhiguo Cao, and Chunhua Shen. From open set to closed set: Counting objects by spatial divide-and-conquer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8362–8371, 2019. 3

[65] Fan Yang, Qiang Zhai, Xin Li, Rui Huang, Ao Luo, Hong Cheng, and Deng-Ping Fan. Uncertainty-guided transformer reasoning for camouflaged object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4146–4155, 2021. 1, 2, 6

[66] Azlan Zahid, Md Sultan Mahmud, Long He, Paul Heinemann, Daeun Choi, and James Schupp. Technological advancements towards developing a robotic pruner for apple trees: A review. *Computers and Electronics in Agriculture*, 189:106383, 2021. 1

[67] Qiang Zhai, Xin Li, Fan Yang, Chenglizhao Chen, Hong Cheng, and Deng-Ping Fan. Mutual graph learning for camouflaged object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12997–13007, 2021. 6, 7

[68] Jing Zhang, Deng-Ping Fan, Yuchao Dai, Xin Yu, Yiran Zhong, Nick Barnes, and Ling Shao. Rgb-d saliency detection via cascaded mutual information minimization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4338–4347, 2021. 6

[69] Miao Zhang, Shuang Xu, Yongri Piao, Dongxiang Shi, Shusen Lin, and Huchuan Lu. Preynet: Preying on camouflaged objects. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 5323–5332, 2022. 1, 6

[70] Pingping Zhang, Wei Liu, Yi Zeng, Yinjie Lei, and Huchuan Lu. Looking for the detail and context devils: High-resolution salient object detection. *IEEE Transactions on Image Processing*, 30:3204–3216, 2021. 4

[71] Yingying Zhang, Desen Zhou, Siqin Chen, Shenghua Gao, and Yi Ma. Single-image crowd counting via multi-column convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 589–597, 2016. 3

[72] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 286–301, 2018. 5

[73] Yijie Zhong, Bo Li, Lv Tang, Senyun Kuang, Shuang Wu, and Shouhong Ding. Detecting camouflaged object in frequency domain. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4504–4513, 2022. 7

[74] Tao Zhou, Huazhu Fu, Geng Chen, Yi Zhou, Deng-Ping Fan, and Ling Shao. Specificity-preserving rgb-d saliency detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4681–4691, 2021. 6

[75] Tao Zhou, Yi Zhou, Chen Gong, Jian Yang, and Yu Zhang. Feature aggregation and propagation network for camouflaged object detection. *IEEE Transactions on Image Processing*, 31:7036–7047, 2022. 1, 2