# SWNet: A Cross-Spectral Network for Camouflaged Weed Detection

Henry O. Velesaca[1,3], Luigi Miranda[1], Angel D. Sappa[1,2]

[1]ESPOL Polytechnic University, Campus Gustavo Galindo, 090902, Guayaquil, Ecuador

[2]Computer Vision Center, Universitat Autònoma de Barcelona, 08193-Bellaterra, Barcelona, Spain

[3]Software Engineering Department, Research Center for Information and Communication Technologies (CITIC-UGR), University of Granada, 18071, Granada, Spain

{hvelesac,luidamir}@espol.edu.ec, sappa@ieee.org

## Abstract

*This paper presents SWNet, a bimodal end-to-end cross-spectral network specifically engineered for the detection of camouflaged weeds in dense agricultural environments. Plant camouflage, characterized by homochromatic blending where invasive species mimic the phenotypic traits of primary crops, poses a significant challenge for traditional computer vision systems. To overcome these limitations, SWNet utilizes a Pyramid Vision Transformer v2 backbone to capture long-range dependencies and a Bimodal Gated Fusion Module to dynamically integrate Visible and Near-Infrared information. By leveraging the physiological differences in chlorophyll reflectance captured in the NIR spectrum, the proposed architecture effectively discriminates targets that are otherwise indistinguishable in the visible range. Furthermore, an Edge-Aware Refinement module is employed to produce sharper object boundaries and reduce structural ambiguity. Experimental results on the Weeds-Banana dataset indicate that SWNet outperforms ten state-of-the-art methods. The study demonstrates that the integration of cross-spectral data and boundary-guided refinement is essential for high segmentation accuracy in complex crop canopies. The code is available on GitHub:* https://cod-espol.github.io/SWNet/.

## 1. Introduction

For over 150 years, plant camouflage has been extensively studied as an evolutionary strategy employed by flora to defend against herbivores and predators [16]. This visual blending relies heavily on environmental illumination and the spectral properties of chlorophyll, which allow certain plants to match their surroundings seamlessly [30]. In the context of modern precision agriculture, this natural phenomenon presents a significant challenge: invasive weeds often exhibit phenotypic characteristics—such as leaf shape, texture, and green coloration—that are nearly identical to those of the primary crops. This homochromatic blending effectively camouflages the weeds within the crop canopy, rendering traditional visual detection methods inadequate [6].

Consequently, there is a pressing need to develop and apply Camouflaged Object Detection (COD) methodologies specifically tailored for agricultural environments. To overcome the limitations of the visible light spectrum, researchers increasingly rely on multimodal approaches that combine standard RGB images with Near-Infrared (NIR) data. NIR imaging is particularly effective in agriculture because it captures distinct physiological differences in cellular structure and chlorophyll reflectance that are imperceptible in standard RGB imagery [18].

To address these challenges, this study evaluates the performance of state-of-the-art COD techniques applied to weed detection in dense crop environments. Specifically, we utilize the Weeds-Banana dataset to establish robust detection baselines. As presented in Table 1, metric evaluation results for each COD technique are reported for both the RGB and NIR baselines. Results are presented using the metric notation defined.

The manuscript is organized as follows. Section 2 introduces related work, recent SOTA COD techniques, and methods that address the problem of the COD approach. Section 3 presents the proposed architecture. Then, Section 4 shows the experimental results using SOTA COD techniques and the proposed approach. Finally, conclusions is given in Section 5.

## 2. Background

Automated weed detection has become a cornerstone of precision agriculture, significantly reducing the reliance on broad-spectrum herbicides. The evolution of visual recognition systems in this domain can be broadly categorized into classical computer vision methodologies and modern, deep learning-based camouflaged object detection frameworks.

### 2.1. Classical Weed Detection Techniques

Early approaches to weed detection relied heavily on handcrafted features and conventional image processing techniques. These methods primarily exploited the phenotypic differences between crops and weeds using morphological characteristics, texture descriptors, and color indices.

Also, traditional algorithms frequently utilize vegetation indices, such as the Excess Green (ExG) index, to separate plant matter from the soil background. Following segmentation, morphological operations and shape-based features (e.g., leaf area, perimeter, and aspect ratio) were extracted to classify the plants [29].

In order to handle more complex canopies, researchers incorporated texture analysis methods like the Gray-Level Co-occurrence Matrix (GLCM) or Histogram of Oriented Gradients (HOG), coupled with classical machine learning classifiers such as Support Vector Machines (SVM) or Random Forests [15].

While computationally efficient, these classical techniques exhibit significant degradation in performance under real-world field conditions. They are highly susceptible to variations in natural illumination, partial occlusions, and shadows. Most importantly, these methods fail when crops and weeds share highly similar spectral and morphological traits, a condition known as homochromatic blending, which renders traditional feature-extraction ineffective.

### 2.2. Camouflaged Weed Detection Techniques

To address the severe limitations of classical methods in dense agricultural environments, recent research has pivoted towards Camouflaged Object Detection (COD). Unlike standard object detection, COD is specifically designed to identify targets that are seamlessly embedded within their surroundings by mimicking the background's color, texture, and structural patterns [4].

Modern COD architectures leverage deep Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) to extract high-level semantic representations. Techniques such as boundary-guided attention, search-and-identification mechanisms, and multi-scale feature fusion have proven highly effective in distinguishing subtle boundary discrepancies between disguised weeds and the primary crop [23].

In agricultural COD, the visual similarity between weeds and crops (such as in the Weeds-Banana dataset) is often too high for RGB data alone. Consequently, the integration of Near-Infrared (NIR) imagery has emerged as a crucial strategy. Since different plant species exhibit unique physiological responses and chlorophyll reflectance in the NIR spectrum, fusing RGB spatial details with NIR spectral signatures allows neural networks to "break" the visual camouflage that occurs in the visible spectrum [18].

Despite these advancements, many existing COD networks struggle with the computational overhead required for real-time field deployment or fail to optimally fuse multi-modal data in highly cluttered crop canopies, highlighting the need for more specialized architectures.

## 3. Proposed SWNet

Similar to most previous works (e.g., [4], [5], [11], [17], [31]), the current work adopts an encoder-decoder pipeline to build the proposed SWNet architecture. The proposed architecture, follows a bimodal encoder-decoder structure designed to extract and fuse complementary information from RGB and NIR spectra. SWNet is designed as an end-to-end trainable framework, as illustrated in Fig. 1.

### 3.1. Feature Extraction Backbone

The core of the encoder utilizes the Pyramid Vision Transformer v2 [27] (PVTv2-B2). Unlike traditional CNNs, the PVTv2 leverages a progressive shrinking pyramid and efficient self-attention mechanisms, allowing the model to capture long-range dependencies—a critical factor when identifying camouflaged objects that mimic their immediate surroundings. The backbone extracts features at four different stages, providing a multi-scale representation.

### 3.2. Processing Blocks: Residual and ConvBlocks

To refine the raw features extracted from the backbone, the architecture employs two fundamental units designed for feature enhancement and dimensional alignment. The ResidualBlock utilizes a $3 \times 3$ convolution followed by Instance Normalization and a LeakyReLU activation, incorporating a skip connection to ensure stable gradient flow and the preservation of fine spatial details. Building upon this, the ConvBlock serves as a bottleneck layer that first projects the backbone's channel dimensions into the internal feature space via a $1 \times 1$ convolution, subsequently applying a double residual refinement process to further strengthen local feature discrimination and representational power.

### 3.3. Convolutional Block Attention Module (CBAM)

To suppress background noise and highlight salient regions, the architecture integrates the Convolutional Block

Table 1. Distinctive characteristics of the evaluated SOTA COD techniques.

| Technique | Source | Source Type | Year | Image Size (px) | Backbone | #Param. (M) |
|---|---|---|---|---|---|---|
| SINet-v2 [5] | TPAMI | Journal | 2021 | $352 \times 352$ | Res2Net-50 [7] | 24.93 |
| BGNet [2] | IJCAI | Conference | 2022 | $416 \times 416$ | Res2Net-50 [7] | 77.80 |
| $C^2$F-Net [1] | TCSVT | Conference | 2022 | $352 \times 352$ | Res2Net-50 [7] | 26.36 |
| OCENet [13] | WACV | Conference | 2022 | $352 \times 352$ | ResNet-50 [8] | 58.17 |
| EAMNet [19] | ICME | Conference | 2023 | $384 \times 384$ | Res2Net-50 [7] | 30.51 |
| DGNet [10] | MIR | Journal | 2023 | $352 \times 352$ | EfficientNet [21] | 8.30 |
| HitNet [9] | AAAI | Conference | 2023 | $352 \times 352$ | PVTv2 [27] | 25.73 |
| ARNet [24] | ICMR | Conference | 2025 | $416 \times 416$ | SMT-Tiny [12] | 12.82 |
| CHNet [25] | ICMR | Conference | 2025 | $416 \times 416$ | SMT-Tiny [12] | 11.20 |
| ARNet-v2 [26] | arXiv | - | 2025 | $416 \times 416$ | Res2Net-50 [7] | 34.12 |
| SWNet (our) | CVPR | Conference | 2026 | $416 \times 416$ | PVTv2 [27] | 42.32 |

Table 2. Details of the training parameters used in evaluated SOTA COD techniques. Learning rate (LR); Batch size (BS).

| Technique | Optimizer | LR | BS | Epochs | Scheduler | Loss function |
|---|---|---|---|---|---|---|
| SINet-v2 [5] | Adam | 1e-4 | 16 | 150 | Custom (Adjust LR) | Structure loss (weighted BCE + weighted IOU) |
| BGNet [2] | Adam | 1e-4 | 12 | 100 | Custom (Poly LR) | Structure loss (weighted BCE + weighted IOU) + Dice loss (edge) |
| $C^2$F-Net [1] | AdaXW | 1e-4 | 32 | 50 | Custom (Poly LR) | Structure loss (weighted BCE + weighted IOU) |
| OCENet [13] | Adam | 1e-5 | 4 | 50 | StepLR | Uncertainty aware structure loss (weighted BCE + weighted IOU) |
| EAMNet [19] | AdamW | 5e-5 | 16 | 150 | Custom (Adjust LR) | Hybrid loss (weighted BCE + weighted IOU) + Edge loss (edge) |
| DGNet [10] | AdamW | 5e-5 | 16 | 150 | CosineAnnealingLR | Hybrid loss (weighted BCE + weighted IOU) + MSE loss (grad) |
| HitNet [9] | AdamW | 1e-4 | 8 | 150 | Custom (Adjust LR) | Structure loss (weighted BCE + weighted IOU) |
| ARNet [24] | Adam | 5e-5 | 8 | 100 | StepLR | Structure loss (wBCE + wIOU) + Edge loss |
| CHNet [25] | AdamW | 5e-5 | 8 | 100 | CosineAnnealingLR | Structure loss (wBCE + wIOU) |
| ARNet-v2 [26] | AdamW | 5e-5 | 4 | 100 | CosineAnnealingLR | Hybrid structure loss (wBCE + wIOU) |
| SWNet (our) | AdamW | 1e-4 | 10 | 200 | CosineAnnealingLR | Structure loss (weighted BCE + weighted IOU) + Edge loss (edge) |

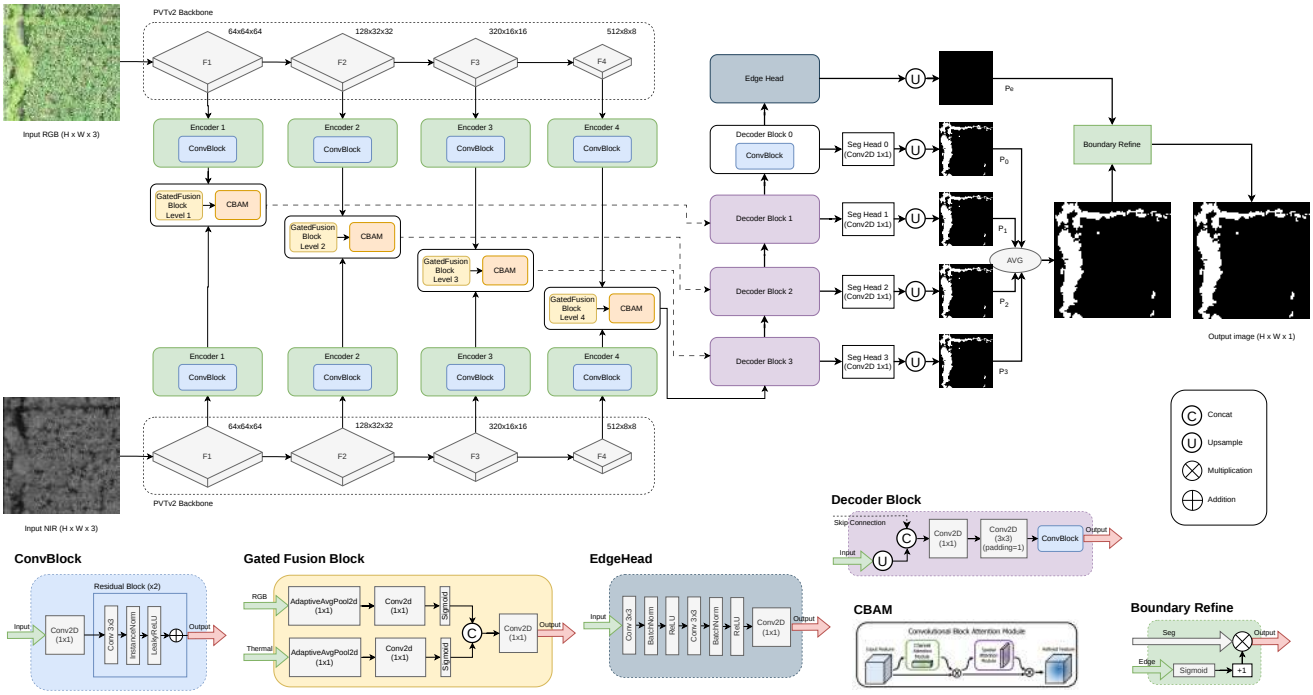


Figure 1. The overall architecture of the proposed SWNet.

Attention Module (CBAM), which processes information through two sequential sub-modules. The process begins with the Channel Gate, which aggregates spatial information by concurrently utilizing Average and Max pooling; these descriptors are then processed by a shared Multi-Layer Perceptron (MLP) to compute a 1D channel attention map, effectively identifying "what" specific features or modalities are most relevant to the task. Subsequently, the Spatial Gate receives these channel-refined features and applies a $7\times7$ convolution over the pooled channel axis to generate a 2D spatial mask, which serves to highlight "where" the camouflaged target is likely located within the scene.

### 3.4. Bimodal Gated Fusion Module

The fusion of RGB and NIR data is executed via a Gated Fusion mechanism that dynamically weights the contribution of each modality to ensure optimal information integration. This process begins with a Gating Mechanism, where independent global average pooling and $1 \times 1$ convolutions are employed to generate modality-specific gates ($\sigma$) that scale the input features based on their relative importance. During the Integration phase, these gated features are concatenated and projected into a unified feature space. Finally, a CBAM pass is applied to the integrated volume, ensuring that the fused representation prioritizes spatial and channel regions where both modalities consistently indicate the presence of a target, thereby significantly reducing the impact of sensor-specific noise or environmental artifacts.

### 3.5. Decoder and Feature Aggregation.

The decoder reconstructs the spatial resolution through a series of DecoderBlocks, which progressively restore the feature maps to the original input dimensions. Each block first performs Upsampling via bilinear interpolation to double the spatial resolution, providing a smoother alternative to learnable deconvolution. This is followed by a module, where the upsampled features are concatenated with the fused skip-connections from the encoder stages to recover lost spatial details. Finally, convolutional smoothing is applied to the aggregated volume to refine the feature representation and resolve the "checkerboard" artifacts typically associated with upsampling operations, ensuring a more spatially coherent output for the segmentation heads.

### 3.6. Edge-Aware Refinement

Detecting the boundaries of camouflaged objects is inherently challenging due to the seamless texture blending between the target and its environment. To address this, the architecture implements an Edge Head branch specifically designed to predict the object's contours from the final decoder stage. These predicted edges are then utilized by the Boundary Refinement module to modulate the primary segmentation mask. Mathematically, the final output is defined as

$$O_{final} = Mask \times (1 + \sigma(Edge))$$

, where the edge information acts as a regional enhancement factor. This operation enforces sharper transitions at the object boundaries, effectively "carving" the target out of the background and reducing the ambiguity typically found in the peripheral regions of camouflaged targets.

### 3.7. Deep Supervision

To prevent the vanishing gradient problem and encourage feature learning at multiple scales, Deep Supervision is applied. Four auxiliary segmentation heads produce intermediate masks from different decoder stages. During inference, these masks are averaged to produce a robust, multi-scale prediction, which is then refined by the edge map to yield the final output.

### 3.8. Loss Function

Following the methodology of previous studies [5, 20], this work adopts the loss function introduced by [28]. The SWNet decoder generates a set of predictions denoted as $\{P_i\}_{i=0}^{3}$. During training, each prediction $P_i$ is upsampled to match the original input dimensions and supervised using a combination of Binary Cross-Entropy ($\mathcal{L}_{BCE}$) [3] and Intersection over Union ($\mathcal{L}_{IoU}$) [14] losses. Consistent with [5], the total objective is calculated by aggregating losses across multiple stages. Furthermore, an Edge Loss is incorporated to refine the boundary detection branch, employing a binary cross-entropy objective against an edge ground truth ($GT_{edge}$) derived from the mask $GT$ via morphological operations (specifically, the difference between local max and min pooling). The total loss function for SWNet is defined as:

$$\mathcal{L}(P, GT) = \sum_{i=0}^{3} \mathcal{L}_{BCE}(P_i, GT) + \mathcal{L}_{IoU}(P_i, GT) \quad (1)$$

$$\mathcal{L}_{total} = \mathcal{L}(P, GT) + \mathcal{L}_{edge}(E, GT_{edge}) \quad (2)$$

### 3.9. Implementation Details

The proposed SWNet is implemented using the PyTorch framework. It employs the PVTv2-B2 backbone [27], pre-trained on ImageNet, as the primary encoder. Network optimization is performed using the AdamW optimizer with a weight decay of $1 \times 10^{-4}$. The learning rate is initialized at $1 \times 10^{-4}$ and adjusted following a cosine annealing schedule. All input images are resized to $416 \times 416$ for both the training and inference phases. The model is trained end-to-end for 200 epochs with a batch size of 10 on a GeForce RTX 4090 (24 GB) GPU. The complete training code is publicly available on GitHub.

### 3.10. Datasets

To validate the proposed architecture, a state-of-the-art multispectral dataset, Weeds-Banana [22] is used. The dataset consists of 272 high-resolution image pairs ($1024 \times 1024$ pixels) providing a specialized benchmark for camouflaged weed detection in dense agricultural canopies. Figure 2 shows examples of RGB, NIR, and mask images of the Weeds-Banana dataset[1].

### 3.11. Evaluation Metrics

To provide a multidimensional assessment of camouflaged object detection (COD) performance, five standard evaluation metrics are employed. The structural similarity between the predicted maps and the ground truth (GT) is quantified using the Structure-measure ($S_\alpha$), while the weighted F-measure ($F_\beta^w$) offers an enhanced evaluation by incorporating spatial weights that emphasize boundary accuracy. Pixel-level discrepancies are measured through the Mean Absolute Error ($M$), calculated between the normalized prediction and the GT. Furthermore, the E-measure ($E_\phi$) is utilized to capture both global and local accuracy based on human visual perception mechanisms, alongside the F-measure ($F_\beta$), which serves as the harmonic mean of precision and recall. To ensure a robust performance comparison across different models, the mean values for the latter two metrics ($E_\phi^{mean}$ and $F_\beta^{mean}$) are reported, as calculated across various thresholds.In order to guarantee a fair comparison with existing state-of-the-art methods, the training configurations for all evaluated models have been standardized. Comprehensive details regarding the specific optimizers, learning rates, schedulers, and loss functions utilized for each network within the benchmark are provided in Table 2.

## 4. Experimental Results

This section presents a comprehensive evaluation of the proposed SWNet against state-of-the-art (SOTA) Camouflaged Object Detection (COD) methods. We describe the quantitative performance across multiple metrics and provide a qualitative analysis of the detection results in challenging agricultural scenarios.

### 4.1. Quantitative Evaluation

The quantitative results, summarized in Table 3, demonstrate that SWNet significantly outperforms existing SOTA methods by effectively leveraging multimodal data. When comparing single-modality baselines, our model already shows competitive results; however, the integration of Visible (Vis) and Near-Infrared (NIR) spectra provides a definitive performance leap. Specifically, SWNet (Vis+NIR) achieves a weighted F-measure ($F_\beta^w$) of 0.8767, surpassing the previous best visible-light model, ARNet, which reached 0.8131. This improvement is also reflected in the Mean Absolute Error ($M$), where SWNet achieves a record low of 0.0070, indicating superior pixel-wise precision compared to the 0.0086 achieved by ARNet-v2. Furthermore, our model reaches an S-measure ($S_\alpha$) of 0.8966, outperforming the specialized HitNet (0.8773) and CHNet (0.8839), which suggests that the Bimodal Gated Fusion Module is more adept at capturing the subtle structural discrepancies between weeds and crops than traditional single-stream or attention-based backbones.

### 4.2. Qualitative Evaluation

The qualitative analysis illustrated in Fig. 4 reinforces the numerical findings, particularly in scenarios characterized by "homochromatic blending" where target and background textures are nearly indistinguishable. While competitive models like BGNet and HitNet often suffer from significant over-segmentation (false positives) or fail to capture the complete geometry of the weed (false negatives), SWNet produces segmentation masks that closely align with the GT. The effectiveness of the Edge-Aware Refinement module is evident in the sharp transitions at object boundaries, effectively "carving" the weed out of the crop canopy where other models produce blurred or fragmented results. By utilizing the NIR spectrum to "break" the visual camouflage, SWNet maintains structural integrity in the predicted masks, even in high-clutter areas where ARNet-v2 and CHNet show visible degradation.

### 4.3. Ablation study

To evaluate the effectiveness of the key components within the SWNet architecture, ablation experiments are conducted focusing on the two main enhancement modules: the Edge Refinement module and the CBAM. Table 4 compares three distinct configurations: a version utilizing only the Edge module, one utilizing only the CBAM module, and the complete SWNet integrating both. The results indicate that while the Edge module improves structural boundary definition ($S_\alpha$ of 0.8797) and the CBAM module enhances feature prioritization, their individual performances are surpassed by their combined integration. The full SWNet configuration achieved the highest scores across all metrics, notably reaching an $S_\alpha$ of 0.8966 and a substantial improvement in the weighted F-measure ($F_\beta^w = 0.8767$). These findings demonstrate that the synergy between spatial-channel attention and explicit edge refinement is essential for accurately discriminating camouflaged weeds from their surrounding crop environment, proving that both modules are critical for the network's success.

[1] https://www.kaggle.com/datasets/hvelesaca/weedbananacod

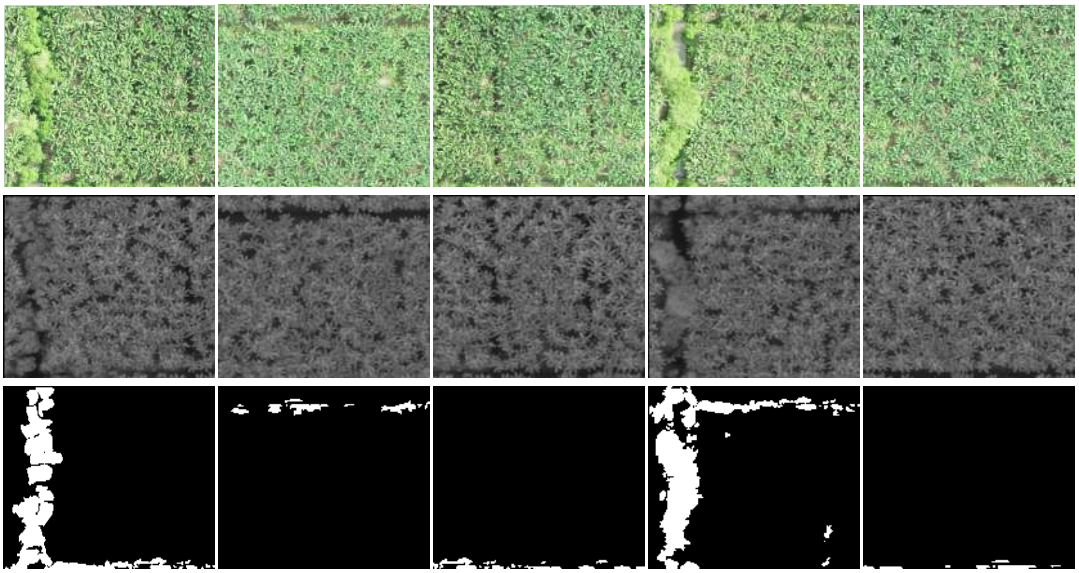

Figure 2. Example of RGB, NIR and mask images of the Weeds-Banana [22] dataset.

Table 3. Metric evaluation results for each COD technique on the Weeds-Banana [22] dataset, reported for the RGB and NIR baseline. Results are presented using the metric notation defined in Sec. 3.11, "↑ / ↓" indicates that larger or smaller is better. The best three performing results are highlighted using color: First, Second, and Third respectively.

| Technique | Input | $S_\alpha \uparrow$ | $F_\beta^w \uparrow$ | $M \downarrow$ | $E_\phi^{adp} \uparrow$ | $E_\phi^{mean} \uparrow$ | $E_\phi^{max} \uparrow$ | $F_\beta^{adp} \uparrow$ | $F_\beta^{mean} \uparrow$ | $F_\beta^{max} \uparrow$ |
|---|---|---|---|---|---|---|---|---|---|---|
| SINet-V2 [5] | Vis | 0.7917 | 0.5754 | 0.0179 | 0.8052 | 0.9175 | 0.9709 | 0.5382 | 0.6153 | 0.6753 |
| | NIR | 0.7522 | 0.5272 | 0.0215 | 0.7668 | 0.8773 | 0.9322 | 0.4945 | 0.5646 | 0.6347 |
| BGNet [2] | Vis | 0.7600 | 0.4362 | 0.0319 | 0.9543 | 0.9468 | 0.9946 | 0.8022 | 0.8469 | 0.9298 |
| | NIR | 0.7444 | 0.4217 | 0.0341 | 0.9017 | 0.9340 | 0.9910 | 0.7221 | 0.7962 | 0.8950 |
| $C^2F$-Net [1] | Vis | 0.6107 | 0.2293 | 0.0827 | 0.7488 | 0.8065 | 0.9729 | 0.5231 | 0.6072 | 0.7848 |
| | NIR | 0.6277 | 0.2522 | 0.0710 | 0.7731 | 0.8274 | 0.9662 | 0.5438 | 0.6188 | 0.7577 |
| OCENet [13] | Vis | 0.8207 | 0.7185 | 0.0133 | 0.9455 | 0.9416 | 0.9791 | 0.7151 | 0.7643 | 0.7835 |
| | NIR | 0.7651 | 0.5876 | 0.0175 | 0.9124 | 0.9341 | 0.9814 | 0.6070 | 0.6380 | 0.6575 |
| EAMNet [19] | Vis | 0.5796 | 0.1953 | 0.1031 | 0.7151 | 0.7685 | 0.9122 | 0.4630 | 0.4895 | 0.5943 |
| | NIR | 0.5467 | 0.1488 | 0.1232 | 0.7324 | 0.7701 | 0.9308 | 0.4393 | 0.4215 | 0.5255 |
| DGNet [10] | Vis | 0.8439 | 0.6908 | 0.0142 | 0.8299 | 0.9302 | 0.9768 | 0.6015 | 0.7120 | 0.7787 |
| | NIR | 0.8381 | 0.6829 | 0.0151 | 0.8530 | 0.9433 | 0.9855 | 0.6119 | 0.7117 | 0.7631 |
| HitNet [9] | Vis | 0.8773 | 0.8090 | 0.0088 | 0.9291 | 0.9652 | 0.9937 | 0.7393 | 0.7970 | 0.8582 |
| | NIR | 0.8705 | 0.7992 | 0.0089 | 0.9455 | 0.9463 | 0.9546 | 0.7813 | 0.8045 | 0.8264 |
| ARNet [24] | Vis | 0.8800 | 0.8131 | 0.0091 | 0.9827 | 0.9604 | 0.9904 | 0.8153 | 0.8492 | 0.8653 |
| | NIR | 0.8265 | 0.7229 | 0.0127 | 0.9429 | 0.9402 | 0.9860 | 0.7138 | 0.7670 | 0.7920 |
| CHNet [25] | Vis | 0.8839 | 0.8058 | 0.0090 | 0.9291 | 0.9713 | 0.9873 | 0.7358 | 0.8096 | 0.8619 |
| | NIR | 0.8027 | 0.6971 | 0.0124 | 0.9267 | 0.9112 | 0.9838 | 0.7200 | 0.7423 | 0.7631 |
| ARNet-v2 [26] | Vis | 0.9027 | 0.8229 | 0.0086 | 0.9493 | 0.9667 | 0.9901 | 0.7681 | 0.8425 | 0.8737 |
| | NIR | 0.8027 | 0.6971 | 0.0124 | 0.9267 | 0.9112 | 0.9838 | 0.7200 | 0.7423 | 0.7631 |
| SWNet (ours) | Vis | 0.7971 | 0.7227 | 0.0122 | 0.9305 | 0.9338 | 0.9391 | 0.7040 | 0.7175 | 0.7419 |
| | NIR | 0.8413 | 0.7624 | 0.0108 | 0.9300 | 0.9325 | 0.9385 | 0.7332 | 0.7460 | 0.7676 |
| | Vis+NIR | 0.8966 | 0.8767 | 0.0070 | 0.9857 | 0.9860 | 0.9906 | 0.8493 | 0.8590 | 0.8788 |

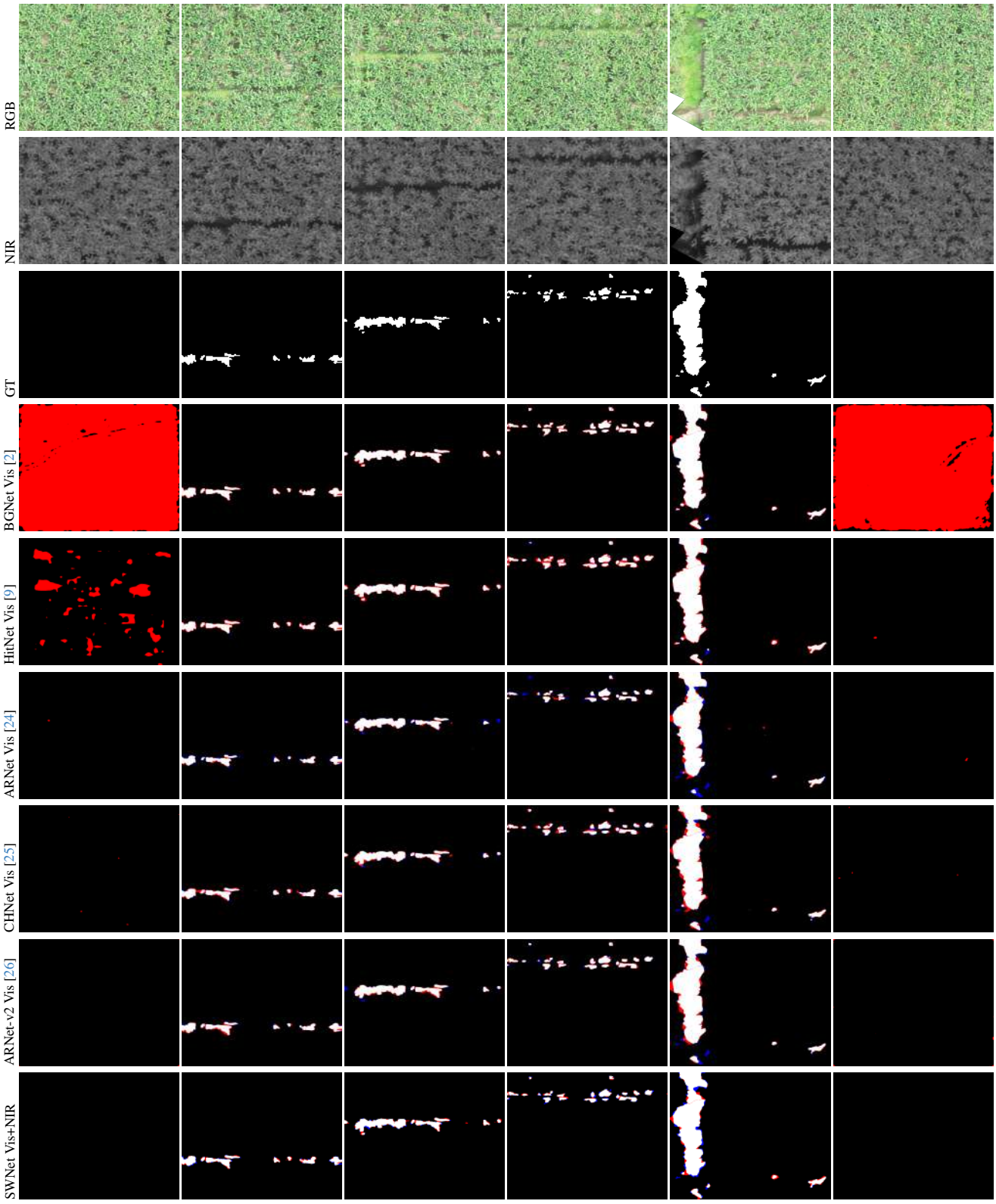


Figure 3. Results on COD techniques that have achieved first or second place in at least one of the metrics in Table 3. Successful matches between GT and predicted masks (white areas); False positive regions (red areas, over-segmentation); and false negative regions (blue areas, miss-segmentation).

Table 4. Metric evaluation results for each COD technique on the Weeds-Banana [22] dataset, reported adding different module on proposed bimodal cross-spectral architecture. Results are presented using the metric notation defined in Sec. 3.11, "↑ / ↓" indicates that larger or smaller is better. The best three performing results are highlighted using color: First, Second, and Third respectively.

| Module | $S_\alpha \uparrow$ | $F_\beta^w \uparrow$ | $M \downarrow$ | $E_\phi^{adp} \uparrow$ | $E_\phi^{mean} \uparrow$ | $E_\phi^{max} \uparrow$ | $F_\beta^{adp} \uparrow$ | $F_\beta^{mean} \uparrow$ | $F_\beta^{max} \uparrow$ |
|---|---|---|---|---|---|---|---|---|---|
| only Edge | 0.8797 | 0.8332 | 0.0071 | 0.9467 | 0.9439 | 0.9480 | 0.8177 | 0.8179 | 0.8339 |
| only CBAM | 0.8714 | 0.8162 | 0.0077 | 0.9414 | 0.9415 | 0.9459 | 0.7918 | 0.8008 | 0.8204 |
| Edge + CBAM | **0.8966** | **0.8767** | **0.0070** | **0.9857** | **0.9860** | **0.9906** | **0.8493** | **0.8590** | **0.8788** |

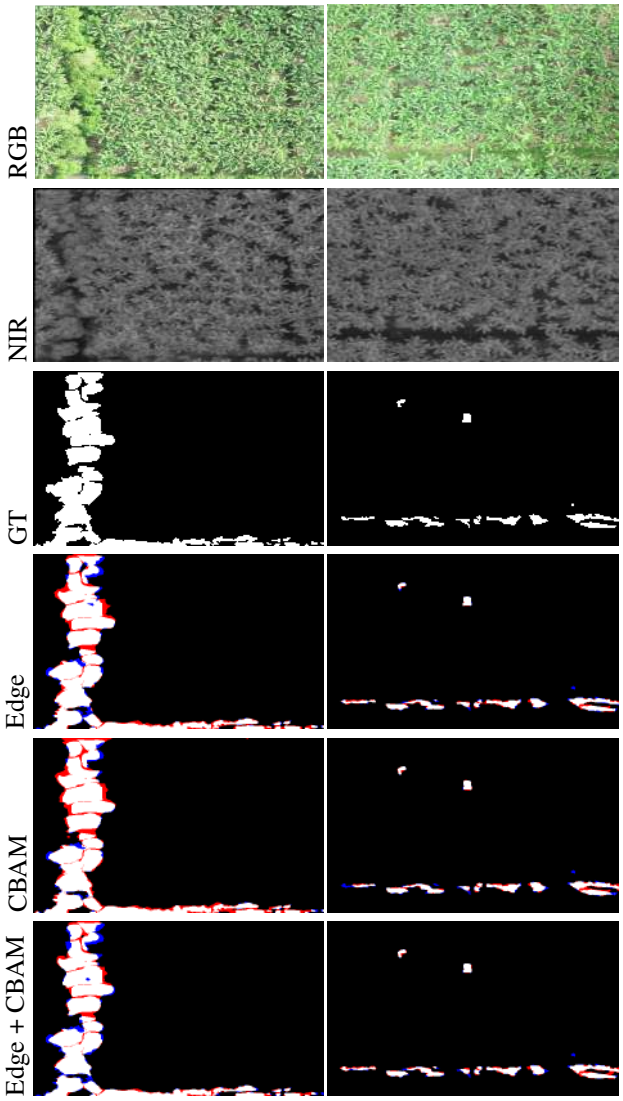


Figure 4. Comparison of three distinct configurations: a version using only the Edge module, one using only the CBAM module, and the complete SWNet integrating both. Successful matches between GT and predicted masks (white areas); False positive regions (red areas, over-segmentation); and false negative regions (blue areas, miss-segmentation).

## 5. Conclusions

This research introduced SWNet, a robust bimodal architecture designed to address the problem of camouflaged weed detection through cross-spectral analysis. Extensive evaluations on the Weeds-Banana dataset demonstrate that the fusion of RGB and NIR data is critical for breaking the natural camouflage of invasive plants. The implementation of the Bimodal Gated Fusion Module and the Convolutional Block Attention Module (CBAM) enables the network to prioritize relevant spatial and channel-wise features while suppressing environmental noise. Additionally, the Edge-Aware Refinement branch significantly improves the delineation of object boundaries, providing high-fidelity masks even in regions with severe visual overlap. SWNet establishes a new performance benchmark, exceeding the results of specialized models such as ARNet-v2 and HitNet across multiple structural and pixel-wise metrics. These findings suggest that multimodal frameworks offer a superior alternative for precision agriculture, facilitating more accurate weed management in challenging field conditions. Future developments could explore the optimization of this architecture for real-time inference on edge computing platforms used in autonomous agricultural robotics.

## Acknowledgements

This material is based upon work supported by the Air Force Office of Scientific Research under award number FA9550-24-1-0206; and partially supported by the Grant PID2021-128945NB-I00 funded by MICIU/AEI/ 10.13039/501100011033 and by ERDF/EU and Grant PID2024-162815NB-I00 funded by MICIU/AEI/ 10.13039/501100011033 and by ERDF/EU; and by the ESPOL project "Advancing Camouflaged Object Detection with a cost-effective Cross-Spectral vision system (ACODCS)" (CIDIS-003-2024). The authors acknowledge the support of the Generalitat de Catalunya CERCA Program to CVC's general activities.

## References

[1] 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. 3, 6

[2] Tianyou Chen, Jin Xiao, Xiaoguang Hu, Guofeng Zhang, and Shaojie Wang. Boundary-guided network for camouflaged object detection. *Knowledge-based systems*, 248: 108901, 2022. 3, 6, 7

[3] Pieter-Tjerk De Boer, Dirk P Kroese, Shie Mannor, and Reuven Y Rubinstein. A tutorial on the cross-entropy method. *Annals of operations research*, 134:19–67, 2005. 4

[4] 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. 2

[5] 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. 2, 3, 4, 6

[6] Xijian Fan, Chunlei Ge, Xubing Yang, and Weice Wang. Cross-modal feature fusion for field weed mapping using rgb and near-infrared imagery. *Agriculture*, 14(12):2331, 2024. 1

[7] Shang-Hua Gao, Ming-Ming Cheng, Kai Zhao, Xin-Yu Zhang, Ming-Hsuan Yang, and Philip Torr. Res2net: A new multi-scale backbone architecture. *IEEE transactions on pattern analysis and machine intelligence*, 43(2):652–662, 2019. 3

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Conf. on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 3

[9] 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. 3, 6, 7

[10] 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. 3, 6

[11] Xinhao Jiang, Wei Cai, Zhili Zhang, Bo Jiang, Zhiyong Yang, and Xin Wang. Magnet: A camouflaged object detection network simulating the observation effect of a magnifier. *Entropy*, 24(12):1804, 2022. 2

[12] Weifeng Lin, Ziheng Wu, Jiayu Chen, Jun Huang, and Lianwen Jin. Scale-aware modulation meet transformer. In *Int. Conf. on Computer Vision*, pages 6015–6026, 2023. 3

[13] Jiawei Liu, Jing Zhang, and Nick Barnes. Modeling aleatoric uncertainty for camouflaged object detection. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1445–1454, 2022. 3, 6

[14] Gellért Máttyus, Wenjie Luo, and Raquel Urtasun. Deeproadmapper: Extracting road topology from aerial images. In *Int. Conf. on Computer Vision*, pages 3438–3446, 2017. 4

[15] Nafeesa Yousuf Murad, Tariq Mahmood, Abdur Rahim Mohammad Forkan, Ahsan Morshed, Prem Prakash Jayaraman, and Muhammad Shoaib Siddiqui. Weed detection using deep learning: A systematic literature review. *Sensors*, 23(7): 3670, 2023. 2

[16] Yang Niu, Hang Sun, and Martin Stevens. Plant camouflage: ecology, evolution, and implications. *Trends in ecology & evolution*, 33(8):608–618, 2018. 1

[17] 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 *Conf. on Computer Vision and Pattern Recognition*, pages 2160–2170, 2022. 2

[18] Inkyu Sa, Jong Yoon Lim, Ho Seok Ahn, and Bruce MacDonald. deepnir: Datasets for generating synthetic nir images and improved fruit detection system using deep learning techniques. *Sensors*, 22(13):4721, 2022. 1, 2

[19] Dongyue Sun, Shiyao Jiang, and Lin Qi. Edge-aware mirror network for camouflaged object detection. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, pages 2465–2470. IEEE, 2023. 3, 6

[20] Yujia Sun, Geng Chen, Tao Zhou, Yi Zhang, and Nian Liu. Context-aware cross-level fusion network for camouflaged object detection. In *IJCAI*, pages 1025–1031, 2021. 4

[21] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019. 3

[22] Henry O Velesaca, Andrea Mero, Hector Villegas, and Angel D Sappa. Unveiling the hidden: Early detection of invasive vegetation in crops with uav multispectral imaging. *Smart Agricultural Technology*, page 101875, 2026. 5, 6, 8

[23] Huiying Wang, Chunping Wang, Qiang Fu, Dongdong Zhang, Renke Kou, Ying Yu, and Jian Song. Cross-modal oriented object detection of uav aerial images based on image feature. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–21, 2024. 2

[24] Kuan Wang, Xiuhong Li, Yulong Bai, Songlin Li, Mengge Lu, and Zhenhong Jia. Assisted refinement network based on channel information interaction for camouflaged object detection. In *Int. Conf. on Multimedia Retrieval*, pages 2058–2062, 2025. 3, 6, 7

[25] Kuan Wang, Xiuhong Li, Songlin Li, Yulong Bai, Boyuan Li, Mengge Lu, and Zhenhong Jia. Efficient camouflaged object detection network based on channel reconstruction and hybrid attention. In *Int. Conf. on Multimedia Retrieval*, pages 2063–2067, 2025. 3, 6, 7

[26] Kuan Wang, Yanjun Qin, Mengge Lu, Liejun Wang, and Xiaoming Tao. Assisted refinement network based on channel information interaction for camouflaged and salient object detection. *arXiv preprint arXiv:2512.11369*, 2025. 3, 6, 7

[27] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pvt v2: Improved baselines with pyramid vision transformer. *Computational Visual Media*, 8(3):415–424, 2022. 2, 3, 4

[28] Jun Wei, Shuhui Wang, and Qingming Huang. $F^3$net: fusion, feedback and focus for salient object detection. In *Proceedings of the AAAI conference on artificial intelligence*, pages 12321–12328, 2020. 4

[29] David M Woebbecke, George E Meyer, Kenneth Von Bargen, and David A Mortensen. Color indices for weed identification under various soil, residue, and lighting conditions. *Transactions of the ASAE*, 38(1):259–269, 1995. 2

[30] Jinyu Yang, Qingwei Wang, Feng Zheng, Peng Chen, Aleš Leonardis, and Deng-Ping Fan. Plantcamo: Plant camouflage detection. *arXiv preprint arXiv:2410.17598*, 2024. 1

[31] Yijie Zhong, Bo Li, Lv Tang, Senyun Kuang, Shuang Wu, and Shouhong Ding. Detecting camouflaged object in frequency domain. In *Conf. on Computer Vision and Pattern Recognition*, pages 4504–4513, 2022. 2