

Arquitectura Avanzada de Aprendizaje Profundo para la Detección de Plagas Camufladas en Agricultura de Precisión

Henry Velesaca

Tutor(es):
Prof. Dr. Angel Sappa:

Una tesis submitida para el grado de:
Magister en Ciencias de la Computación

29-01-2026

Dedicatoria

A mi esposa, Alice Gómez, por su amor incondicional y paciencia infinita durante este camino. A mi hija, Amelia Idamia, por ser la luz que ilumina mis días y el motivo de mi esfuerzo.

Agradecimientos

Mi más sincero agradecimiento al Dr. Ángel Sappa, cuya guía y constante apoyo fueron decisivos para la culminación exitosa de esta tesis. Su mentoría ha sido una fuente de inspiración y aprendizaje continuo. De igual manera, agradezco a la ESPOL por el soporte institucional brindado, el cual ha facilitado significativamente el desarrollo de esta investigación.

Evaluadores



Prof. Dr. Angel Sappa
Tutor
Escuela Superior Politécnica del Litoral



Prof. Dr. Boris Vintimilla
Evaluador
Escuela Superior Politécnica del Litoral

Declaración expresa

Yo Henry Oswaldo Velesaca Lara acuerdo y reconozco que:

La titularidad de los derechos patrimoniales de autor (derechos de autor) del proyecto de graduación corresponderá al autor o autores, sin perjuicio de lo cual la ESPOL recibe en este acto una licencia gratuita de plazo indefinido para el uso no comercial y comercial de la obra con facultad de sublicenciar, incluyendo la autorización para su divulgación, así como para la creación y uso de obras derivadas. En el caso de usos comerciales se respetará el porcentaje de participación en beneficios que corresponda a favor del autor o autores.

La titularidad total y exclusiva sobre los derechos patrimoniales de patente de invención, modelo de utilidad, diseño industrial, secreto industrial, software o información no divulgada que corresponda o pueda corresponder respecto de cualquier investigación, desarrollo tecnológico o invención realizada por mí/nosotros durante el desarrollo del proyecto de graduación, pertenecerán de forma total, exclusiva e indivisible a la ESPOL, sin perjuicio del porcentaje que me/nos corresponda de los beneficios económicos que la ESPOL reciba por la explotación de mi/nuestra innovación, de ser el caso.

En los casos donde la Oficina de Transferencia de Resultados de Investigación (OTRI) de la ESPOL comunique al/los autor/es que existe una innovación potencialmente patentable sobre los resultados del proyecto de graduación, no se realizará publicación o divulgación alguna sin la autorización expresa y previa de la ESPOL.

Guayaquil, 29 de Enero 2026

Firmado electrónicamente por:
HENRY OSWALDO VELESACA LARA
Validar únicamente con FirmaEC

Henry Oswaldo Velesaca Lara

Notificación de Confidencialidad (solo si aplica)

Verificar si hay condiciones relacionadas a confidencialidad con el tutor(es).

Resumen

Esta tesis presenta AINet, una novedosa arquitectura de aprendizaje profundo diseñada para detectar objetos camuflados en entornos complejos y diversos. AINet aprovecha las ventajas de Mamba, un eficiente modelo de estados secuenciales para capturar dependencias de largo alcance, y Convolutional Block Attention Module (CBAM) para el refinamiento de características mediante mecanismos de atención. Detectar objetos camuflados supone un reto significativo en una amplia gama de aplicaciones del mundo real, como vigilancia, seguridad, imágenes médicas y sistemas autónomos, donde los objetos de interés pueden mimetizarse con el entorno y evadir los métodos de detección convencionales. Para demostrar su eficacia, AINet se evalúa en múltiples datasets, incluyendo benchmarks de detección de objetos camuflados como CAMO, COD10K y NC4K, así como un dataset para detección de plagas. Los resultados experimentales muestran que AINet supera a modelos recientes del estado del arte. El código está disponible públicamente en GitHub para su reproducibilidad: https://github.com/hvelesaca/AINet.

Palabras Clave: Detección de objetos camuflados, detección de plagas, agricultura de precisión, visión por computador.

Abstract

This thesis introduces AINet, a novel deep learning architecture designed for detecting camouflaged objects in complex and diverse environments. AINet leverages the strengths of Mamba, an efficient sequential state model for capturing long-range dependencies, and the Convolutional Block Attention Module (CBAM) for feature refinement through attention mechanisms. Detecting camouflaged objects is a significant challenge across a wide range of real-world applications, including surveillance, security, medical imaging, and autonomous systems, where objects of interest may blend into their backgrounds and evade conventional detection methods. To demonstrate its effectiveness, AINet is evaluated on multiple datasets, including standard camouflaged object detection benchmarks such as CAMO, COD10K, and NC4K, as well as pest detection datasets. Experimental results show that AINet outperforms existing state-of-the-art models. The code is publicly available on GitHub for reproducibility: https://github.com/hvelesaca/AINet.

Índice general

Resumen vi

Abstract vii

Abreviaturas x

Índice de figuras xi

Índice de cuadros xii

1 Introducción 1
1.1 Motivación . . . 1
1.2 Objetivos y contribuciones . . . 2
1.3 Estructura de la tesis . . . 2

2 Trabajos Relacionados 3
2.1 Detección de Objetos Genéricos (GOD) . . . 3
2.2 Detección de Objetos Salientes (SOD) . . . 4
2.3 Detección de Objetos Camuflados (COD) . . . 4
2.4 Aplicaciones en agricultura de precisión . . . 6

3 Conjunto de datos & Métricas 7
3.1 Conjunto de datos . . . 7
3.2 Generación de conjunto de datos . . . 8
3.3 Métricas . . . 9

4 Aumento de datos 13

5 YOLO 17
5.1 YOLOv11 . . . 17
5.2 SINet-v2 . . . 17
5.3 Mask R-CNN+PVT . . . 18
5.4 Métricas tradicionales . . . 18

6 Arquitectura Propuesta 21
6.1 Arquitectura general . . . 21
6.2 Bloques Mamba y CBAM . . . 22
6.3 Detalles de implementación . . . 23

7 Resultados **25**
7.1 Detalles de entrenamiento 25
7.2 Evaluación Cuantitativa 25
7.3 Evaluación Cualitativa 28
7.4 Estudio de Ablación 29

8 Conclusiones & Trabajos Futuros **34**

Referencias **35**

Archivos adjuntos **41**
Anexo A 42

Abreviaturas

Avg Promedio, por sus siglas en inglés (Average).

BCE Entropía Cruzada Binaria, por sus siglas en inglés (Binary Cross-Entropy).

CBAM Módulos de Atención de Bloques Convolucionales, por sus siglas en inglés (Convolutional Block Attention Module).

COD Detección de Objetos Camuflados, por sus siglas en inglés (Camouflaged Object Detection).

ED Disrupción de bordes, por sus siglas en inglés (Edge disruption).

GOD Detección de Objetos Genéricos, por sus siglas en inglés (Generic Object Detection).

Grad-CAM Mapeo de activación de clases ponderado por gradiente, por sus siglas en inglés (Gradient-weighted Class Activation Mapping).

GT Verdad Fundamental o Base, por sus siglas en inglés (Ground Truth).

IoU Intercepción sobre Unión, por sus siglas en inglés (Interception over Union).

IS Similitud intrínseca, por sus siglas en inglés (Intrinsic similarity).

MLP Multilayer Perceptron.

SOD Detección de Objetos Salientes, por sus siglas en inglés (Saliency Object Detection).

SOTA State-of-the-art o State of the Art sin guiones. Se refiere, en el ámbito científico, al Estado de Arte de una tarea.

Índice de figuras

3.1 Ejemplos de imágenes RGB y térmicas correctamente alineadas que contienen iguanas camufladas en un entorno natural del conjunto de datos BIOS. 8
3.2 Imágenes de ejemplo de los conjuntos de datos de referencia. 10

4.1 Ejemplo de imágenes generadas usando DiffuseMix [1]. Prompt: Watercolor Art (WA), Snowy (SN), Autumn (AU), Mosaic (MO) 14
4.2 Resultados de predicción de nueve técnicas COD entrenadas con aumento de datos, Albumentations [2] (*primera a cuarta columna*) y DiffuseMix [1] (*quinta a octava columna*). Las áreas blancas representan coincidencias exitosas entre GT y máscaras predichas; las áreas rojas indican regiones de falsos positivos (sobre-segmentación); y las áreas verdes indican regiones de falsos negativos (segmentación incorrecta). 16

5.1 Resultados de predicción utilizando Mask R-CNN+PVT, YOLOv11 y SINet-v2. 20

6.1 Arquitectura propuesta AINet. 22
6.2 Imágenes de ejemplo de los conjuntos de datos de referencia (CAMO [3, 4], COD10K [5, 6], NC4K [7]) y el dataset de plagas (Cotton Bollworm [8]) utilizados como caso de estudio. . . 24

7.1 Resultados cualitativos de seis técnicas SOTA COD y la arquitectura AINet propuesta, evaluados en imágenes de ejemplo de conjuntos de datos de referencia. Coincidencias exitosas entre GT y máscaras predichas (áreas blancas); regiones con falsos positivos (áreas rojas, sobresegmentación); y regiones con falsos negativos (áreas verdes, segmentación incorrecta). 27
7.2 Resultados cualitativos de siete técnicas SOTA COD y AINet, evaluados en imágenes de ejemplo del conjunto de datos Cotton Bollworm [8]. Coincidencias exitosas entre GT y máscaras predichas (áreas blancas); regiones con falsos positivos (áreas rojas, sobresegmentación); y regiones con falsos negativos (áreas verdes, segmentación incorrecta). 29
7.3 Resultados cualitativos de diferentes combinaciones de módulos utilizados, evaluados en imágenes de ejemplo del conjunto de datos Cotton Bollworm [8]. 31
7.4 Resultados cualitativos de las diferentes salidas de supervisión profunda, evaluados en imágenes de ejemplo del conjunto de datos Cotton Bollworm [8]. 32
7.5 Visualización de Grad-CAM utilizando las salidas de las cabezas de segmentación, evaluadas en imágenes de ejemplo del conjunto de datos Cotton Bollworm [8]. 33

Índice de cuadros

3.1 Conjunto de datos SOTA COD. 9

4.1 Evaluación de métricas para cada técnica COD. Los tres resultados con mejor rendimiento se resaltan mediante color: Primero, Segundo y Tercero respectivamente. 15
4.2 Diferencia porcentual entre los resultados obtenidos para cada técnica COD y cada método de aumento de datos se muestra en el Cuadro 4.1. Los valores positivos indican una mejora en el método basado en el modelo de difusión, mientras que los valores negativos indican un deterioro del rendimiento. La última fila muestra la mejora (%) de la DiffuseMix con respecto a Albumentation. 15

5.1 Métricas utilizadas para evaluar las técnicas. TP, FP, TN y FN representan Verdadero Positivo, Falso Positivo, Verdadero Negativo y Falso Negativo, respectivamente [8]. 19
5.2 Evaluación de métricas para cada técnica—notación como se presenta en el Cuadro 5.1. . 19

6.1 Distribución de los datasets utilizados en los resultados experimentales. 23

7.1 Detalles de los parámetros de entrenamiento utilizados en las técnicas SOTA COD evaluadas. Learning rate (LR); Batch size (BS). 25
7.2 Resultados experimentales de las técnicas SOTA COD y la arquitectura propuesta AINet en conjunto de datos de referencia. Los tres resultados con mejor rendimiento se resaltan en color: Primero, Segundo y Tercero, respectivamente. 26
7.3 Resultados cuantitativos de las técnicas SOTA COD y AINet en el conjunto de datos Cotton Bollworm [8]. Los tres resultados con mejor rendimiento se resaltan en color: Primero, Segundo y Tercero, respectivamente. 28
7.4 Evaluación de métricas con diferentes combinaciones de módulos (p. ej., Mamba y CBAM) en el conjunto de datos Cotton Bollworm [8]. Los mejores resultados de cada métrica se resaltan en negrita. 31
7.5 Evaluación de métricas con diferentes resultados en el conjunto de datos Cotton Bollworm [8]. Los tres resultados con mejor rendimiento se resaltan con color: Primero, Segundo y Tercero, respectivamente. 32

1

Introducción

1.1 Motivación

Con el avance de la visión por computadora basada en redes neuronales profundas, la investigación ha pasado de la detección y el reconocimiento de objetos genéricos (Generic Object Detection o GOD, por sus siglas en inglés) [9], cuyo objetivo es localizar y clasificar instancias de categorías predefinidas mediante cajas delimitadoras en escenas generales, a tareas más específicas como la detección de objetos salientes (Saliency Object Detection o SOD, por sus siglas en inglés) [10], que busca segmentar las regiones visualmente más destacadas de una imagen (típicamente a nivel de píxel) en función de contrastes y pistas perceptuales. En este contexto, y especialmente, surge la Detección de Objetos Camuflados (Camouflaged Object Detection o COD, por sus siglas en inglés), enfocada en identificar objetos que se mimetizan con el fondo y presentan diferencias sutiles respecto al entorno. Los métodos dominantes (p. ej., GOD y SOD) suelen funcionar cuando el objeto contrasta claramente con el fondo, pero fallan con frecuencia ante camuflaje natural o artificial, causando omisiones y falsas detecciones [11].

Detectar objetos camuflados en la actualidad es un problema desafiante y de gran relevancia en visión por computador [12]. Los objetos camuflados se caracterizan por su capacidad de integrarse imperceptiblemente en su entorno, lo que dificulta su detección mediante métodos convencionales [13]. Este reto es frecuente en una amplia variedad de escenarios del mundo real, incluidos videovigilancia, imagenología médica, vehículos autónomos y monitoreo ambiental, donde la identificación precisa de objetos ocultos o parcialmente ocluidos es crítica para la toma de decisiones y seguridad [14]. Además, COD no solo exige robustez frente a variaciones de iluminación, escala y desorden del fondo, sino también la capacidad de inferir diferencias semánticas y estructurales sutiles que pueden no ser evidentes a nivel de píxel, lo que lleva a los paradigmas actuales de detección a sus límites.

Los algoritmos tradicionales de detección de objetos suelen tener dificultades para detectar objetivos camuflados debido al contraste mínimo y a los límites ambiguos entre el objeto y el fondo [6]. Los avances recientes en aprendizaje profundo han logrado mejoras significativas en detección y segmentación

de objetos; sin embargo, la naturaleza particular del camuflaje continúa planteando obstáculos [15]. Abordar estos desafíos requiere modelos capaces de capturar tanto el contexto global como los detalles finos, así como mecanismos que enfoquen la atención en pistas sutiles que distingan a los objetos camuflados de su entorno [16].

1.2 Objetivos y contribuciones

Esta tesis presenta AINet, una arquitectura novedosa de aprendizaje profundo diseñada específicamente para mejorar la COD. AINet integra dos componentes potentes: Mamba [17], un modelo de estado secuencial eficiente capaz de modelar dependencias de largo alcance, y Módulos de Atención de Bloques Convolucionales (Convolutional Block Attention Module o CBAM, por sus siglas en inglés) [18], que refina las representaciones de características mediante mecanismos de atención espacial y por canal. Al combinar estos elementos, AINet puede resaltar y segmentar eficazmente objetos camuflados, incluso en escenas altamente complejas.

Las contribuciones clave de este trabajo incluyen:

- AINet introduce una novedosa arquitectura integrada con atención (Mamba + CBAM) para COD.
- La red propuesta alcanza un rendimiento superior en conjuntos de datos de referencia generales de COD (CAMO, COD10K, NC4K), superando de forma consistente a 26 métodos del Estado del Arte (State-of-the-art o SOTA, por sus siglas en inglés) recientes. Además, supera el desempeño en un conjunto de datos de detección de plagas agrícolas (Cotton Bollworm [8]).
- Estudios de ablación demuestran que la sinergia de los módulos Mamba y CBAM, así como la supervisión profunda multinivel, es crítica para lograr una precisión de segmentación óptima y una delimitación robusta de límites de los objetos.
- AINet obtiene resultados de detección altamente precisos y estables bajo condiciones diversas, incluyendo iluminación variable y fondos complejos, lo que la hace ideal para aplicaciones del mundo real.
- La arquitectura se valida en múltiples conjuntos de datos y métricas, mostrando mejoras significativas tanto en precisión como en exactitud de contornos en relación a técnicas SOTA.

1.3 Estructura de la tesis

La tesis está organizada de la siguiente manera. El Capítulo 2 presenta trabajos recientes relacionados a técnicas COD y métodos que abordan el problema del enfoque COD. El Capítulo 3 presenta conjuntos de datos SOTA y las métricas para evaluación de técnicas COD. El Capítulo 4 presenta un trabajo donde se comparó el uso de diferentes enfoques para la aumentación de datos. El Capítulo 5 presenta una primera aproximación a tareas COD utilizando YOLOv11 [19] como técnica de referencia. El Capítulo 6 describe la arquitectura propuesta. Luego, el Capítulo 7 muestra los resultados experimentales en diferentes conjuntos de datos, además de un estudio de ablación detallado. Finalmente, las conclusiones y trabajos futuros se presentan en el Capítulo 8.

2

Trabajos Relacionados

Esta sección revisa técnicas del estado del arte basadas en visión por computador para GOD, SOD, COD, analizando sus contribuciones y limitaciones en diversos dominios de aplicación.

2.1 Detección de Objetos Genéricos (GOD)

GOD busca localizar y clasificar instancias de múltiples categorías en una imagen, típicamente mediante bounding boxes y una etiqueta de clase por objeto. En esta línea, los detectores de dos etapas separan la generación de propuestas y la clasificación/regresión final: R-CNN popularizó el uso de propuestas de regiones combinadas con descriptores profundos, mejorando la precisión frente a enfoques previos [20]. Más adelante, Faster R-CNN integró en la red un Region Proposal Network (RPN), logrando un esquema más eficiente y estable para entrenamiento y despliegue [21].

En contraste, los detectores de una etapa (one-stage) formulan la detección como una predicción densa sobre una rejilla o un conjunto de anclas, buscando mayor velocidad con buena precisión. YOLO planteó una predicción unificada y en tiempo real de cajas y clases, destacando por su simplicidad y eficiencia [22]. SSD profundizó esta idea con predicciones multi-escala y anclas a diferentes resoluciones, lo que mejora la detección de objetos de distintos tamaños [23]. Posteriormente, RetinaNet abordó el desbalance extremo entre ejemplos positivos y negativos mediante pérdida focal, incrementando el rendimiento de detectores densos [24].

El progreso en GOD está fuertemente ligado a benchmarks y protocolos de evaluación: PASCAL VOC y MS COCO estandarizaron datasets, splits y métricas como mAP, facilitando comparaciones reproducibles entre métodos [25, 26]. Más recientemente, la detección se ha beneficiado de arquitecturas basadas en Transformers: DETR reformula la detección como un problema de asignación entre un conjunto fijo de predicciones y objetos reales, eliminando componentes heurísticos como anchor design y NMS en su formulación original [27]. En conjunto, estas líneas reflejan el espectro típico de GOD:

precisión vs. eficiencia y diseños con componentes explícitos vs. formulaciones “end-to-end”.

2.2 Detección de Objetos Salientes (SOD)

Por otra parte, SOD pretende identificar las regiones/objetos que destacan perceptualmente en una escena, normalmente produciendo un mapa de saliencia (frecuentemente a nivel de píxel) que resalta lo más llamativo para el observador. Una base influyente en visión computacional es el modelo de atención visual de Itti, Koch y Niebur, que formaliza la saliencia desde pistas de bajo nivel y mecanismos de competencia [28]. En enfoques clásicos de SOD, también se explotaron propiedades espectrales o contrastes globales: el método de spectral residual propone medir rarezas en el dominio de Fourier para ubicar regiones llamativas [29], mientras que la aproximación frequency-tuned busca regiones salientes mediante diferencias de color suavizadas y contrastes respecto a la imagen filtrada [30].

La evaluación en SOD suele reportar métricas de similitud entre mapas predichos y máscaras ground-truth (p.ej., F-measure, MAE), y se apoya en datasets con anotaciones de objetos salientes. Por ejemplo, DUT-OMRON se diseñó para escenas más complejas y desafiantes (múltiples objetos, fondos variados), siendo un referente en el análisis de robustez [31]. ECSSD se centra en imágenes con estructuras complejas y bordes ricos, comúnmente usada para medir la calidad de segmentación de regiones salientes [32]. Además, el dataset DUTS se volvió muy utilizado por su escala (entrenamiento/prueba) para aprendizaje profundo en SOD [33], y HKU-IS se emplea frecuentemente para escenarios con múltiples objetos y bajo contraste [34].

Con el auge del deep learning, SOD pasó de heurísticas de contraste a arquitecturas entrenables de extremo a extremo que combinan características multiescala y refinamiento progresivo. DSS introdujo supervisión profunda y conexiones cortas para mejorar la consistencia espacial y la recuperación de detalles, elevando el desempeño en varios benchmarks [35]. PiCANet incorporó mecanismos de atención contextual por píxel para capturar dependencias locales y globales relevantes a la saliencia [36]. BASNet enfatizó la preservación de bordes mediante módulos boundary-aware y pérdidas diseñadas para mejorar contornos y estructuras finas del objeto saliente [37]. En conjunto, estas contribuciones muestran que SOD moderno suele priorizar: (i) fusión multi-escala, (ii) contexto/atención y (iii) refinamiento de límites.

2.3 Detección de Objetos Camuflados (COD)

COD es un desafío persistente en visión por computador, ya que estos objetos se caracterizan por su capacidad de integrarse de manera imperceptible en fondos complejos, lo que dificulta distinguirlos mediante técnicas de detección convencionales. Los avances recientes en aprendizaje profundo han mejorado significativamente la segmentación de imágenes y la detección de objetos, permitiendo una identificación más precisa y eficiente de objetivos camuflados en una variedad de escenarios del mundo real, como videovigilancia, imagenología médica, monitoreo ambiental y sistemas autónomos.

Una de las primeras técnicas pioneras en COD es Search Identification Network (SINet) [5], que introduce un marco de dos etapas inspirado en el comportamiento de caza de los depredadores, compuesto por un Módulo de Búsqueda y un Módulo de Identificación. SINet presenta una arquitectura extremo a extremo simple pero efectiva basada en el conjunto de datos COD10K, ampliamente anotado. Sobre

esta base, SINet-V2 [6] mejora el diseño original incorporando capas densamente conectadas y un componente de campo receptivo para capturar características multinivel. Esta arquitectura mejorada muestra una aplicabilidad más amplia, incluyendo su posible uso en ámbitos militares, de seguridad y de conservación de la vida silvestre, donde la detección de objetos ocultos es esencial [38].

Otra contribución es la técnica SegMaR [39], que introduce un marco de refinamiento iterativo multietapa para la detección de objetos camuflados, simulando el proceso de detección de grueso a fino del sistema visual humano. La arquitectura consta de tres pasos centrales: *Segment*, *Magnify* y *Reiterate*. Inicialmente, una red de segmentación de camuflaje genera una máscara preliminar del objeto. Luego, el módulo *Magnify* amplía de forma adaptativa la región del objeto mediante muestreo basado en atención, haciéndola más distintiva dentro de la imagen. Finalmente, el módulo *Reiterate* refina la segmentación mediante retroalimentación iterativa, capturando progresivamente detalles más finos, especialmente en objetos pequeños o altamente camuflados. La arquitectura también incorpora un módulo de distracción para desacoplar características de primer plano y fondo, y emplea decodificadores paralelos para enfocarse en regiones clave del objeto y sus contornos. Sin embargo, SegMaR no puede entrenarse de extremo a extremo [40], lo que limita su adaptabilidad en ciertos escenarios.

Los enfoques COD enfrentan dos desafíos principales: (1) similitud intrínseca (IS), donde los objetos se asemejan visualmente al fondo, y (2) disrupción de bordes (ED), que produce límites poco claros. Métodos inspirados biológicamente como SINet [5] y SegMaR [39] intentan abordar estos problemas imitando el comportamiento de caza de depredadores o la cognición visual humana mediante procesamiento multietapa. Sin embargo, estos enfoques a menudo tienen dificultades con patrones de camuflaje complejos y pueden no capturar las señales sutiles necesarias para resolver IS y ED de forma efectiva. Para superar estas limitaciones, la técnica FEature Decomposition and Edge Reconstruction (FEDER) [41] adopta una estrategia dirigida de dos etapas. Utiliza wavelets aprendibles para descomponer características de la imagen e identificar las bandas de frecuencia más informativas mediante un módulo de atención de frecuencia y agregación de características. Para abordar ED, introduce una tarea auxiliar de reconstrucción de bordes inspirada en ecuaciones diferenciales, mejorando la precisión de los contornos y la exactitud global de detección. A pesar de resultados prometedores, la dependencia de FEDER en la reconstrucción de bordes puede ser menos efectiva cuando los objetos camuflados tienen texturas muy similares a su entorno, lo que puede provocar falsos negativos.

El campo de COD ha experimentado un progreso notable en los últimos años. El trabajo fundacional de Qin et al. [37] en detección de objetos salientes ha influido profundamente en la evolución de las metodologías de COD. En 2022, Chen et al. [42] propusieron un enfoque de fusión entre niveles con conciencia de contexto que mejoró la precisión en la identificación de objetos camuflados, mientras que Chen et al. [43] introdujeron una red guiada por bordes para mejorar la detección de contornos y características. Liu et al. [44] avanzaron aún más el campo al abordar el modelado de incertidumbre aleatoria en COD. Más recientemente, redes de retroalimentación iterativa de alta resolución [45], redes con conciencia de bordes [46] y técnicas de aprendizaje profundo de gradientes [47] han contribuido a mejoras significativas en eficiencia y precisión de detección. El alcance de la investigación en COD continúa expandiéndose, con trabajos recientes como PlantCamo [48] orientados a aplicaciones

especializadas en detección de camuflaje en plantas.

2.4 Aplicaciones en agricultura de precisión

Si bien gran parte de la investigación se ha centrado en COD de propósito general, estos avances también se han adaptado para abordar desafíos específicos de dominio. Por ejemplo, en entornos agrícolas, el camuflaje desempeña un papel crítico tanto en la detección de plagas como de frutos, lo que complica las tareas de monitoreo y manejo [49]. Estudios recientes han explorado el uso de arquitecturas avanzadas de aprendizaje profundo para estas aplicaciones. Meng et al. [8] abordaron la segmentación por instancias de plagas camufladas combinando Pyramid Vision Transformer (PVT) y Mask R-CNN, aprovechando la extracción jerárquica de características de PVT y las capacidades de segmentación a nivel de instancia de Mask R-CNN. De manera similar, Evangelista et al. [15] introdujeron FCNet, una arquitectura de segmentación basada en transformers y consciente del contexto para detectar frutos camuflados en entornos de huertos. Estos trabajos demuestran el potencial de adaptar metodologías generales de COD a la agricultura, aunque persisten desafíos relacionados con la generalización a condiciones diversas y la necesidad de grandes datasets etiquetados.

En conjunto, estos avances han establecido sistemas de detección de objetos camuflados más robustos y precisos, fijando nuevos referentes en el área y allanando el camino para futuras investigaciones e implementaciones prácticas. Se destaca cómo estos avances generales se han adaptado y extendido para abordar desafíos específicos en entornos agrícolas, como la detección de plagas.

3

Conjunto de datos & Métricas

Este capítulo presenta varios conjuntos de datos SOTA así como también las métricas de evaluación usadas para medir el rendimiento de las técnicas COD.

3.1 Conjunto de datos

El Cuadro 3.1 reúne diversos conjuntos de datos SOTA COD y trabajos representativos en el estudio de la detección de objetos camuflados, un problema particularmente desafiante porque el objeto de interés tiende a compartir texturas, colores y patrones con el fondo, reduciendo las pistas visuales que normalmente explotan los métodos de visión por computador. Por otra parte, la Figura 3.2 muestra imágenes de ejemplo de los diferentes conjuntos de datos de referencia.

En primer lugar, se presenta CAMO (1.250 imágenes) [3, 4], el cual se utiliza generalmente en la fase de entrenamiento como un conjunto de datos de validación para verificar si un modelo logra delimitar correctamente regiones camufladas, especialmente en bordes y zonas donde el objeto "se pierde" visualmente en la escena. Posteriormente apareció CAMO++ (5.500 imágenes) [50, 51]. Este conjunto de datos, a diferencia de CAMO, agrega el uso de instancias para cada clase de los objetos camuflados.

Un recurso de escala mayor es COD10K (10.000 imágenes) [52, 5], por su tamaño y calidad, este conjunto de datos de referencia es uno de los más usados en los trabajos que abordan el problema de COD, es usado para entrenar modelos con mayor diversidad de escenas y evaluar desempeño con menor sensibilidad a sesgos que un conjunto pequeño. Por otra parte, CPD1K (1.000 imágenes) [53] a diferencia de los conjuntos más "genéricos", este trabajo se centra en un caso de uso más concreto: personas con patrones de camuflaje, donde la dificultad no proviene solo del entorno natural sino del propio patrón de la vestimenta diseñado para romper la silueta.

Dentro de los conjuntos más reducidos, se encuentra CHAMELEON (76 imágenes). Aunque su tamaño es limitado, es útil para validaciones rápidas, especialmente cuando se desea analizar cualitati-

vamente fallos típicos del camuflaje (pérdida de contornos, confusión textura-fondo). Su uso suele ser complementario y solo enfocado en pruebas para observar comportamientos del modelo en ejemplos representativos.

El conjunto de datos que más se utiliza para evaluar es NC4K (4.121 imágenes) [7], el cual se plantea como el conjunto de datos de referencia de validación más grande, e incluye imagen RGB, GT y anotación a nivel de instancia (por ejemplo, para distinguir instancias cuando aplique).

Para aplicaciones más específicas, se presenta el conjunto de datos Cotton Bollworm (1.080 imágenes), aquí el camuflaje se estudia en un contexto agrícola: plagas que se confunden con la planta o el entorno, se presenta PlantCamo (1.250 imágenes) [48]. Este conjunto de datos enfoca el camuflaje en el dominio de plantas, donde la variabilidad de formas y patrones naturales, junto con fondos vegetales complejos, genera escenarios con alta confusión visual.

3.2 Generación de conjunto de datos

Finalmente y como parte del contexto de investigación del presente trabajo se puede mencionar la generación de un conjunto de datos de iguanas camufladas, dicho artículo ha sido aceptado en la conferencia VISAPP 2026 [54].



Figura 3.1: Ejemplos de imágenes RGB y térmicas correctamente alineadas que contienen iguanas camufladas en un entorno natural del conjunto de datos BIOS.

El conjunto de datos adquirido en la ciudad de Guayaquil y nombrado como BIOS contiene 148 pares de imágenes camufladas RGB-Térmicas (véase la Fig. 3.1). Este conjunto se centra en iguanas naturalmente camufladas en entornos exteriores complejos y está destinado a apoyar la investigación en monitoreo de fauna y detección de objetos camuflados.

El sistema de adquisición de imágenes del conjunto BIOS consta de una cámara Basler acA1300-60gc, utilizada para capturar imágenes del espectro visible con una resolución de 1280×1024 píxeles, un lente de 13 mm y una cámara térmica FLIR TAU2 con una resolución de 640×480 píxeles que utiliza un lente de 8 mm. Ambas cámaras están montadas rígidamente en una plataforma diseñada a medida para minimizar el paralaje y garantizar una línea base consistente entre los ejes ópticos. Esta configuración facilita un registro cruzado multimodal preciso, lo cual es crucial para alinear las imágenes térmicas y visibles.

Originalmente se capturó un total de 176 pares de imágenes en condiciones de luz diurna, en hábitats

Cuadro 3.1: Conjunto de datos SOTA COD.

Nombre	Artículo	Imágenes
CAMO	Anabranch network for camouflaged object segmentation (Computer Vision and Image Understanding) 2019 [3], [4]	1,250
CAMO++	Camoufinder: Finding camouflaged instances in images (The Thirty-Fifth AAAI Conference on Artificial Intelligence) 2021 [50], [51]	5,500
COD10K	Concealed object detection (PAMI) 2020 [52], [5]	10,000
CPD1K	Detection of people with camouflage pattern via dense deconvolution network (IEEE Signal Processing Letters) 2018 [53]	1,000
CHAMELEON	-	76
NC4K	Simultaneously Localize, Segment and Rank the Camouflaged Objects (CVPR) 2021 [7]	4,121
Cotton Bollworm	Camouflaged cotton bollworm instance segmentation based on PVT and Mask R-CNN (Computers and Electronics in Agriculture) 2024 [8]	1,080
PlantCamo	PlantCamo: Plant Camouflage Detection (arXiv) 2024 [48]	1,250

naturales de iguanas. Posteriormente, los pares de imágenes se alinean mediante un proceso de registro geométrico utilizando una técnica de alineamiento conocida como LightGlue [55]. Tras un exhaustivo procedimiento de validación y depuración que implicó eliminar muestras desalineadas o de baja calidad, el conjunto de datos final contiene 148 pares de imágenes de alta calidad, alineados espacialmente.

Este conjunto de datos resultante BIOS se presenta como un referente realista y desafiante para evaluar algoritmos de detección de objetos camuflados en escenarios del mundo real. Al integrar imágenes visibles y térmicas alineadas mediante la técnica LightGlue, el conjunto proporciona un recurso valioso para la comunidad de investigación en COD, especialmente para aplicaciones en monitoreo de biodiversidad, conservación ecológica y detección de fauna basada en visión térmica.

3.3 Métricas

A lo largo del presente trabajo se utilizaron cinco métricas ampliamente reconocidas para evaluar el rendimiento de técnicas COD. Estas métricas proporcionan un criterio de evaluación integral para analizar la precisión y la eficacia de la detección en diferentes modelos: Structure-measure (S_α) [56], weighted F-measure (F_β^w) [57], Mean Absolute Error (M) [58], E-measure (E_ϕ) [59] y F-measure (F_β) [30]. La métrica S_α cuantifica la similitud estructural entre la predicción y los mapas de GT. La F_β^w representa una métrica de evaluación mejorada que amplía la F_β tradicional al incorporar ponderaciones espaciales, lo que proporciona una mejor evaluación de la calidad de la segmentación, con énfasis en la precisión de los límites y la importancia de los píxeles detectados según la ubicación. La métrica M se centra en la evaluación del error a nivel de píxel entre la predicción normalizada y el GT. La métrica M se centra en la evaluación del error a nivel de píxel entre la predicción normalizada y el GT. La métrica E_ϕ evalúa simultáneamente la precisión global y local de COD basándose en los mecanismos de percepción visual humana. La métrica F_β proporciona una medida sintética que considera tanto los

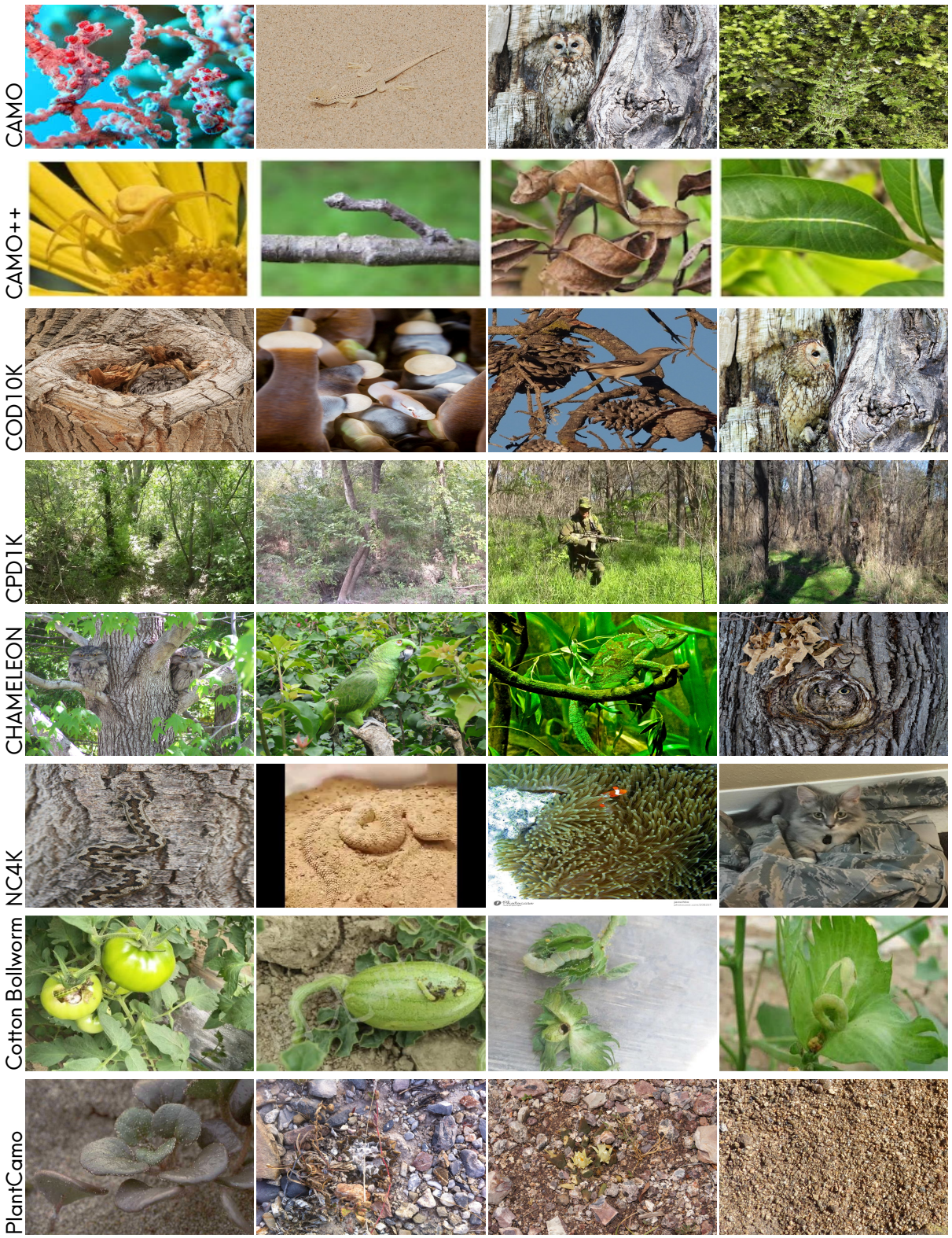

Figura 3.2: Imágenes de ejemplo de los conjuntos de datos de referencia.

componentes de precisión como de recuperación. Tanto para las métricas F-measure como E-measure, se pueden obtener diferentes puntuaciones según los distintos pares de precisión-recuperación. Esto conduce al cálculo de la media de F-measure (F_β^{mean}). De igual forma, E-measure utiliza variantes de la media, denotadas como E_ϕ^{mean}, que también se emplean como métricas de evaluación.

A continuación se define matemáticamente las ecuaciones que rigen cada métrica. Primero se define la curva de precisión-recuperación (PR) la cual se genera transformando la entrada en una máscara binaria M, segmentada en un rango de umbrales de 0 a 255. La precisión (P) y la recuperación (R) se calculan comparando la máscara binaria M con la máscara de GT en cada umbral, lo que genera la curva PR. P y R se pueden calcular como se muestra en la Ecuación 3.1.

$$P = \frac{|M(T) \cap GT|}{|M(T)|}, \qquad R = \frac{|M(T) \cup GT|}{|GT|} \tag{3.1}$$

donde $M(T)$ es la máscara obtenida al umbralizar el mapa de predicción no binario en el umbral T.

S-measure (S_α) cuantifica la similitud estructural espacial entre el mapa predicho (C) y el GT. Combina evaluaciones basadas en objetos (S_o) y regiones (S_r), como se muestra en la definición en la Ecuación 3.2.

$$S_\alpha = \alpha S_o + (1 - \alpha) S_r \tag{3.2}$$

donde $\alpha \in [0, 1]$ es un factor de ponderación, normalmente establecido en 0.5, que equilibra la contribución de S_o y S_r.

F-measure (F_β) se utiliza para calcular la relación entre la precisión (P) y la recuperación (R). Inicialmente, la entrada se transforma en una máscara binaria, M, segmentada en un rango de umbrales de 0 a 255. P y R se calculan comparando M con GT en estos umbrales. Las fórmulas para P y R se definen como se muestra en la Ecuación 3.3:

$$P = \frac{|M(T) \cap GT|}{|M(T)|}, \qquad R = \frac{|M(T) \cup GT|}{|GT|} \tag{3.3}$$

donde $M(T)$ representa la máscara binaria obtenida al umbralizar el mapa de predicción no binario en el umbral T, y $|\ |$ denota el área total de la máscara. F_β también muestra el valor medio armónico promedio entre ellos. La fórmula para F_β se define como se muestra en la Ecuación 3.3.

$$F_\beta = \frac{(\beta^2 + 1)\, PR}{\beta^2 P + R} \tag{3.4}$$

Con β^2 típicamente establecido en 0.3. A partir del rango de umbrales, se calculan 2 variantes de F_β: la media (F_β^{mean}) y la adaptativa (F_β^{adp}). Además, (F_β^w) es también una métrica ampliamente utilizada donde P y R son promedios ponderados.

E-measure (E_ϕ) evalúa tanto la similitud local como la global entre C y G. Se define según se muestra en la Ecuación 3.5:

$$E_\phi = \frac{1}{W \times H} \sum_{x=1}^{W} \sum_{y=1}^{H} \phi\big(C(x,y),\, G(x,y)\big) \tag{3.5}$$

donde ϕ representa una matriz de alineación mejorada, y W y H son el ancho y la altura de la imagen de entrada, respectivamente. E_ϕ también proporciona dos variantes: la media (E_ϕ^{mean}) y la adaptativa (E_ϕ^{adp}).

El Error Absoluto Medio (M) cuantifica la diferencia absoluta promedio por píxel entre el mapa predicho normalizado C y el mapa de GT, donde $C, GT \in [0, 1]$. El error absoluto medio M se formula según la Ecuación 3.6:

$$M = \frac{1}{H \times W} \sum_{x=1}^{W} \sum_{y=1}^{H} |C(x,y) - G(x,y)| \tag{3.6}$$

donde W y H representan el ancho y la altura de la imagen de entrada, y (x, y) representan las coordenadas de los píxeles. A diferencia de F_β, E_ϕ y S_α, un valor de M más bajo sugiere un mejor resultado.

4

Aumento de datos

Al igual que en toda solución basada en aprendizaje, el conjunto de datos usado para el entrenamiento es clave, tanto en calidad como en cantidad. En este capítulo se presenta una técnica que se implementó para aumentar los datos de entrenamiento.

De cara a evaluar técnicas de aumento de datos para ser usadas en COD, se comparan nueve técnicas COD: BASNet [37], SINet-v2 [6], BGNet [43], C^2F-Net [42], OCENet [44], EAMNet [46], DGNet [47], HitNet [45] y PCNet [48]. Técnicas como SINet-v2, C^2F-Net, OCENet, DGNet, HitNet y PCNet utilizan imágenes de entrada de 352×352 píxeles, mientras que BASNet, BGNet y EAMNet utilizan tamaños de 256×256, 416×416 y 384×384 píxeles, respectivamente. Esta consistencia dimensional entre la mayoría de los modelos indica un enfoque estandarizado en este campo.

Esta comparación se realizó sobre el conjunto de datos Cotton Bollworm [8]. Además, los resultados de esta evaluación fueron publicados en [60] en el que se compararon dos técnicas de aumento de datos: una con aumento clásico mediante Albumentations (AwA) [2] y otra con un enfoque basado en modelos de difusión (DiffuseMix, AwD) [1], obteniendo 4633 imágenes (AwA) y 4406 imágenes (AwD) respectivamente. Además, en la Figura 4.1 se presentan ejemplos de imágenes generadas con DiffuseMix usando distintos "prompts" (p. ej., Watercolor Art, Snowy, Autumn, Mosaic), con el objetivo de incrementar la diversidad, manteniendo coherencia con las etiquetas.

El siguiente paso es comparar los resultados entre el aumento de datos con Albumentations [2] y con DiffuseMix [1] para mejorar la tarea de COD. El Cuadro 4.1 muestra los resultados de esta evaluación. En general, DiffuseMix presenta un desempeño superior en la mayoría de las métricas. BGNet con DiffuseMix alcanza los mejores resultados en varias métricas clave, incluyendo S_α (0.8666), F_β^w (0.8237), F_β^{adp} (0.9356), F_β^{mean} (0.8431) y F_β^{max} (0.8584). HitNet con DiffuseMix también destaca, al lograr el mejor desempeño en la métrica M (0.0157, donde valores más bajos son mejores) y en las métricas E_ϕ^{adp} (0.9704) y E_ϕ^{mean} (0.9673). Entre los resultados obtenidos cuando el aumento se realiza con Albumentations, EAMNet y SINet-v2 se muestra un rendimiento notable, siendo SINet-v2 el que alcanza el mayor E_ϕ^{max} (0.9727) en todas las comparaciones. DGNet usando Albumentations obtiene el segundo mejor resultado en la métrica M (0.0163). Resulta interesante que la incorporación del modelo de

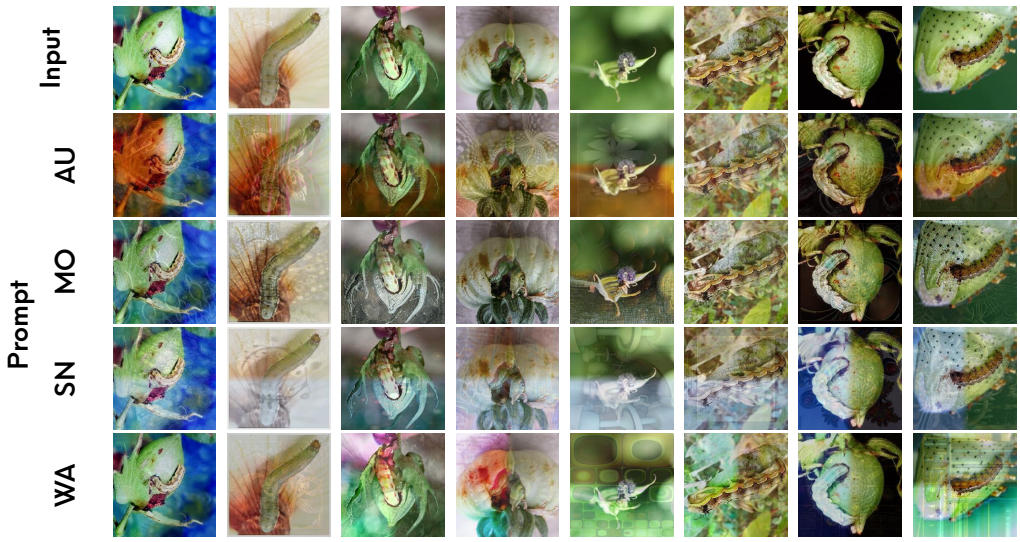

Figura 4.1: Ejemplo de imágenes generadas usando DiffuseMix [1]. Prompt: Watercolor Art (WA), Snowy (SN), Autumn (AU), Mosaic (MO)

difusión, en general, mejore el desempeño de las técnicas base, en particular en métricas relacionadas con la precisión y la detección de bordes.

Por otro lado, el Cuadro 4.2 presenta el análisis comparativo entre el aumento con Albumentations [2] y el aumento con DiffuseMix [1]; este revela patrones interesantes en las diferencias de rendimiento a través de diversas técnicas de COD. HitNet y BGNet muestran las mejoras más sustanciales al implementarse con DiffuseMix, evidenciando incrementos notables en todas las métricas de evaluación. De manera interesante, algunas técnicas como BASNet presentan un leve deterioro en el rendimiento con el enfoque DiffuseMix, con disminuciones que oscilan entre -0.64% y -1.22% según la métrica. La comparación de la diferencia mediana de rendimiento entre ambas estrategias indica que DiffuseMix, por lo general, supera a Albumentations, con diferencias positivas en las métricas S_α, F_β^w, E_ϕ^{adp}, E_ϕ^{max}, F_β^{adp}, F_β^{mean} y F_β^{max}, excepto en M, que permanece neutral (0.00%), y en E_ϕ^{mean}, que presenta una diferencia negativa de -0.27%. Las mejoras más sustanciales se observan en F_β^{mean} y E_ϕ^{max}, con incrementos de 1.50% y 0.91%, respectivamente. El enfoque basado en modelos de difusión mejora el rendimiento, aunque su efectividad varía según la técnica y la métrica.

La Figura 4.2 compara el GT y las máscaras predichas a través de distintas técnicas. En estas visualizaciones se resaltan tres regiones diferenciadas: las áreas blancas representan coincidencias exitosas entre la GT y las máscaras predichas, indicando una detección precisa; las áreas rojas denotan regiones de falsos positivos (sobre-segmentación), donde el modelo predice incorrectamente objetos camuflados en zonas no marcadas en la GT; y las áreas verdes indican regiones de falsos negativos (sub-segmentación u omisión), que corresponden a zonas donde el modelo no detecta objetos camuflados presentes en la GT. El análisis visual revela que las técnicas de COD entrenadas con el método Albumentations tienden a producir más falsos positivos (rojo) y falsos negativos (verde) en comparación con sus contrapartes entrenadas con conjuntos de datos aumentados mediante modelos de difusión. Esta observación sugiere que el enfoque basado en difusión conduce a una detección de objetos más precisa y exacta, con menos casos tanto de sobre-segmentación como de detecciones omitidas.

Cuadro 4.1: Evaluación de métricas para cada técnica COD. Los tres resultados con mejor rendimiento se resaltan mediante color: Primero, Segundo y Tercero respectivamente.

	Técnica	$S_\alpha \uparrow$	$F_\beta^w \uparrow$	$M \downarrow$	$E_\phi^{adp} \uparrow$	$E_\phi^{mean} \uparrow$	$E_\phi^{max} \uparrow$	$F_\beta^{adp} \uparrow$	$F_\beta^{mean} \uparrow$	$F_\beta^{max} \uparrow$
AwA [2]	BASNet [37]	0.7860	0.6839	0.0277	0.8938	0.8811	0.8890	0.7085	0.7073	0.7235
	SINet-v2 [6]	0.8588	0.7979	0.0184	0.9430	0.9571	0.9727	0.7837	0.8086	0.8392
	BGNet [43]	0.8573	0.7964	0.0185	0.9588	0.9532	0.9625	0.8264	0.8323	0.8527
	C^{2}F-Net [42]	0.8399	0.7184	0.0214	0.9368	0.9329	0.9441	0.7910	0.7989	0.8215
	OCENet [44]	0.8598	0.8042	0.0164	0.9421	0.9543	0.9647	0.7809	0.8115	0.8416
	EAMNet [46]	0.8627	0.8136	0.0174	0.9576	0.9611	0.9686	0.8164	0.8320	0.8570
	DGNet [47]	0.8565	0.8005	0.0163	0.9468	0.9585	0.9666	0.7827	0.8096	0.8377
	HitNet [45]	0.8422	0.7978	0.0178	0.9601	0.9663	0.9724	0.7991	0.8060	0.8244
	PCNet [48]	0.8463	0.7885	0.0184	0.9342	0.9471	0.9642	0.7691	0.7924	0.8311
AwD [1]	BASNet [37]	0.7761	0.6739	0.0269	0.9064	0.8726	0.9065	0.6992	0.7009	0.7113
	SINet-v2 [6]	0.8612	0.8123	0.0167	0.9493	0.9594	0.9689	0.7960	0.8199	0.8466
	BGNet [43]	0.8666	0.8237	0.0165	0.9698	0.9600	0.9708	0.9356	0.8431	0.8584
	C^{2}F-Net [42]	0.8487	0.7346	0.0212	0.9422	0.9509	0.9635	0.7931	0.8191	0.8450
	OCENet [44]	0.8548	0.7961	0.0179	0.9494	0.9528	0.9646	0.7831	0.8100	0.8352
	EAMNet [46]	0.8617	0.8180	0.0166	0.9524	0.9608	0.9674	0.8073	0.8303	0.8552
	DGNet [47]	0.8645	0.8068	0.0164	0.9454	0.9492	0.9597	0.7927	0.8196	0.8403
	HitNet [45]	0.8542	0.8206	0.0157	0.9704	0.9673	0.9713	0.8298	0.8318	0.8416
	PCNet [48]	0.8550	0.7949	0.0176	0.9425	0.9479	0.9535	0.7891	0.8057	0.8275

Cuadro 4.2: Diferencia porcentual entre los resultados obtenidos para cada técnica COD y cada método de aumento de datos se muestra en el Cuadro 4.1. Los valores positivos indican una mejora en el método basado en el modelo de difusión, mientras que los valores negativos indican un deterioro del rendimiento. La última fila muestra la mejora (%) de la DiffuseMix con respecto a Albumentation.

Técnica	ΔS_α	ΔF_β^w	ΔM	ΔE_ϕ^{adp}	ΔE_ϕ^{mean}	ΔE_ϕ^{max}	ΔF_β^{adp}	ΔF_β^{mean}	ΔF_β^{max}
BASNet [37]	-0.99	-1.00	-0.08	+1.26	-0.85	+1.75	-0.93	-0.64	-1.22
SINet-v2 [6]	+0.24	+1.44	-0.17	+0.63	+0.23	-0.38	+1.23	+1.13	+0.74
BGNet [43]	+0.93	+2.73	-0.20	+1.10	+0.68	+0.83	+10.92	+1.08	+0.57
C^{2}F-Net [42]	+1.62	-0.02	+0.54	+1.80	+1.94	+0.21	+0.27	+2.02	+2.35
OCENet [44]	-0.50	-0.81	+0.15	+0.73	-0.15	-0.01	+0.22	-0.15	-0.64
EAMNet [46]	-0.10	+0.44	-0.08	-0.08	-0.52	-0.03	-0.12	-0.17	-0.18
DGNet [47]	+0.80	+0.63	+0.01	+0.01	-0.14	-0.93	+0.69	+1.00	+0.26
HitNet [45]	+1.20	+2.28	-0.21	+1.03	+0.10	-0.11	+3.07	+2.58	+1.72
PCNet [48]	+0.87	+0.64	-0.08	+0.83	+0.08	+0.08	-1.07	+1.33	-0.36
Mejora(%)	+0.37	+0.54	0.00	+0.50	-0.27	+1.50	+0.31	+1.06	+0.24

Los resultados experimentales demuestran que la técnica de aumento de datos basada en modelos de difusión supera al método clásico. La comparación muestra una mejora de las métricas al utilizar DiffuseMix, lo que indica que la técnica basada en modelos de difusión, en general, supera al método clásico, con mejoras positivas en la mayoría de las métricas (S_α, F_β^w, E_ϕ^{adp}, E_ϕ^{max}, F_β^{adp}, F_β^{mean} y F_β^{max}), siendo las más sustanciales las de E_ϕ^{max} y F_β^{mean}, con incrementos de 1.50% y 1.06%, respectivamente.



Figura 4.2: Resultados de predicción de nueve técnicas COD entrenadas con aumento de datos, Albumentations [2] (*primera a cuarta columna*) y DiffuseMix [1] (*quinta a octava columna*). Las áreas blancas representan coincidencias exitosas entre GT y máscaras predichas; las áreas rojas indican regiones de falsos positivos (sobre-segmentación); y las áreas verdes indican regiones de falsos negativos (segmentación incorrecta).

5

YOLO

Como primera alternativa para detectar objetos camuflados en el contexto del trabajo de la presente tesis se evaluó el uso de la arquitectura YOLO, en particular YOLOv11. La evaluación se realizó sobre el dataset Cotton Bollworm [8] y los resultados fueron comparados con SINet-v2 [6] y Mask R-CNN+PVT [8].

El presente capítulo presenta en detalle los trabajos realizados, los cuales fueron publicados en [61].

5.1 YOLOv11

Al momento de realizar este trabajo, el modelo YOLOv11 era la última versión estable de la familia YOLO. Se utilizó YOLOv11 [19] con su variante para segmentación en [61], una de las características principales de esta arquitectura es la eficiencia computacional. Esta propiedad resulta especialmente relevante en escenarios agrícolas, donde se requiere un balance entre precisión y viabilidad operativa (p.ej., procesamiento cercano al tiempo real). Para el entrenamiento se empleó el modelo preentrenado *yolo11m-seg.pt* y se configuraron 500 épocas, tamaño de lote 8 y tamaño de imágenes de entrada de 640×640 píxeles. Adicionalmente, se aplicaron técnicas de aumento de datos mediante Albumentations [62] (rotaciones, recortes, ajustes de brillo/contraste, entre otros) con el fin de incrementar la robustez frente a variaciones de iluminación, perspectiva y fondos con alto mimetismo.

5.2 SINet-v2

Con el fin de analizar el desempeño de una arquitectura diseñada específicamente para objetos camuflados, se incluyó SINet-v2 [6]. Esta arquitectura aborda el problema COD mediante un esquema de dos etapas (búsqueda e identificación), apoyándose en un backbone eficiente y en módulos de refinamiento de características que potencian la localización de patrones sutiles. La implementación utilizada se entrenó en PyTorch empleando una función de pérdida compuesta por BCE ponderada e IoU, durante

100 épocas, con tamaño de lote 16 e imágenes de 352×352 píxeles. Se utilizó el optimizador Adam con tasa de aprendizaje inicial de 1×10^{-4}, ajustada cada 25 épocas.

5.3 Mask R-CNN+PVT

Finalmente, se consideró como línea base de comparación un enfoque de segmentación por instancias de dos etapas basado en Mask R-CNN+PVT [8]. Este método integra una arquitectura tipo Transformer (PVT) para capturar relaciones globales y características jerárquicas, mientras que Mask R-CNN aporta segmentación precisa a nivel de instancia. En este estudio, los valores asociados a Mask R-CNN+PVT se tomaron como referencia a partir de los resultados publicados en [8], utilizando el mismo conjunto de datos, lo cual permite contrastar el desempeño de la propuesta con un método reciente y competitivo.

En conjunto, la evaluación de YOLOv11, SINet-v2 y Mask R-CNN+PVT permite comparar estrategias con supuestos y fortalezas distintas: desde un modelo eficiente de una etapa orientado a despliegue práctico (YOLOv11), pasando por una red especializada en camuflaje (SINet-v2), hasta una aproximación de dos etapas con segmentación por instancias ampliamente reconocida en visión por computador (Mask R-CNN+PVT). Esta selección proporciona una base sólida para discutir tanto precisión como costo computacional en escenarios agrícolas reales.

5.4 Métricas tradicionales

Para evaluar las diferentes técnicas, se calculan las métricas como Average Precision (AP) (p. ej., para cuadros delimitadores y máscaras de segmentación), Recall, F1-score y Pixel Accuracy para determinar el rendimiento de las técnicas. El Cuadro 5.1 muestra los detalles de las métricas a calcular, así como su descripción.

El Cuadro 5.2 muestra que la arquitectura YOLOv11 alcanzó métricas destacables, con un AP^b de 71.3% para detección y un AP^s de 60.3% para segmentación, superando significativamente al modelo Mask R-CNN+PVT, que obtuvo 62.5% y 59.5%, respectivamente. Esta mejora en la precisión es particularmente relevante considerando la complejidad inherente de detectar plagas camufladas, y un aspecto notable de YOLOv11 es su eficiencia computacional. Estos resultados son particularmente impresionantes considerando la naturaleza compleja de las imágenes de prueba, que incluyen diversos patrones de camuflaje y condiciones ambientales variables.

Por otro lado, la Fig. 5.1 muestra la evaluación cualitativa de los resultados obtenidos por cada técnica, en la que se observa que con YOLOv11 se obtienen mejores *bounding boxes* y máscaras de segmentación; estos resultados son consistentes con las métricas cuantitativas obtenidas para las tareas de detección y segmentación. Estos resultados cualitativos evidencian la robustez de YOLOv11 en escenarios complejos, pudiendo igualar o mejorar los resultados de técnicas SOTA especializadas en tareas COD. Asimismo, SINet-v2 no obtiene buenos resultados y se limita a detectar y segmentar objetos salientes en este conjunto de datos en particular (véase el Cuadro 5.1, *primera fila, segunda columna*).

Aunque el conjunto de datos presentado constituye una excelente contribución, también presenta ciertas inconsistencias en el etiquetado. Por ejemplo, la Figura 5.1 (*quita fila, segunda columna*) muestra un error de etiquetado en el que el *GT* no contiene la máscara del gusano pequeño ubicado en el lado derecho, pero Mask R-CNN+PVT y YOLO sí logran detectarlo.

Cuadro 5.1: Métricas utilizadas para evaluar las técnicas. TP, FP, TN y FN representan Verdadero Positivo, Falso Positivo, Verdadero Negativo y Falso Negativo, respectivamente [8].

Métrica	Descripción
AP^b	Representa la precisión promedio para los cuadros delimitadores previstos, calculados a través de múltiples umbrales de IoU de 0,50 a 0,95 en pasos de 0,05
AP^b_{50}	Representa la precisión promedio para los cuadros delimitadores previstos, evaluados en un umbral de IoU de 0,50
AP^b_{75}	Representa la precisión promedio para los cuadros delimitadores previstos, evaluados en un umbral de IoU de 0,75
AP^s	Representa la precisión promedio para máscaras previstas, calculada a través de múltiples umbrales de IoU de 0,50 a 0,95 en pasos de 0,05
AP^s_{50}	Representa la precisión promedio para máscaras previstas, evaluada en un umbral de IoU de 0,5
AP^s_{75}	Representa la precisión promedio para máscaras previstas, evaluada en un umbral de IoU de 0,75
Pixel Accuracy	$PA = \frac{TP+TN}{TP+TN+FP+FN}$
Recall	$R = \frac{TP}{TP+FN}$
F1-Score	$F1 = \frac{2\times P\times R}{P+R}, P = \frac{TP}{TP+FP}$

Cuadro 5.2: Evaluación de métricas para cada técnica—notación como se presenta en el Cuadro 5.1.

Técnica	Detección			Segmentación			PA	R	$F1$
	AP^b	AP^b_{50}	AP^b_{75}	AP^s	AP^s_{50}	AP^s_{75}			
Mask R-CNN + PVT [8]	62.5	89.7	72.9	59.5	89.2	72.9	**98.61**	**99.12**	**98.95**
YOLOv11 [19]	**71.3**	**92.3**	**73.8**	**60.3**	**90.3**	**73.1**	98.34	96.47	97.79
SINet-v2 [6]	54.5	71.3	59.8	44.5	56.3	49.2	90.07	65.61	47.70

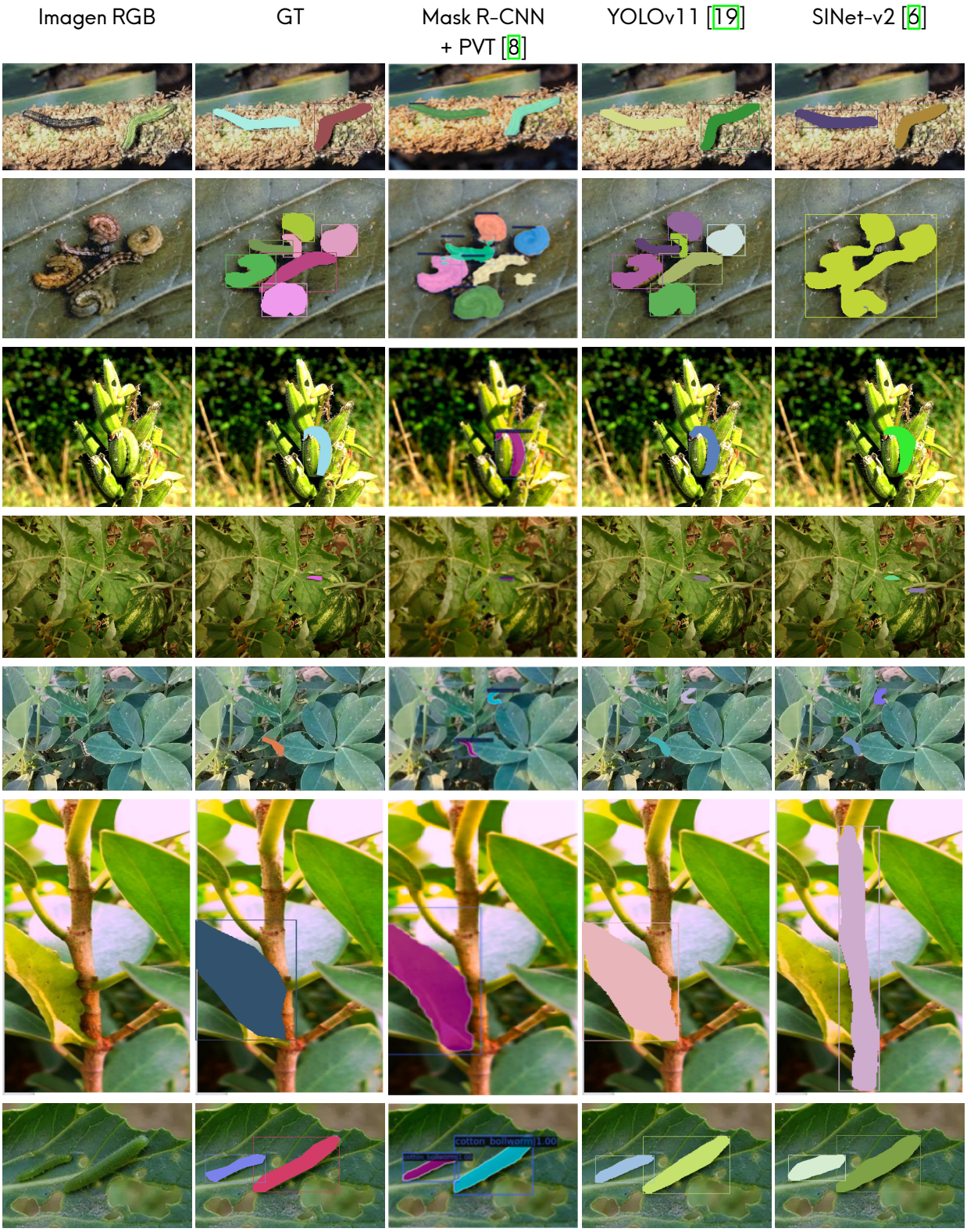

Figura 5.1: Resultados de predicción utilizando Mask R-CNN+PVT, YOLOv11 y SINet-v2.

6

Arquitectura Propuesta

En este capítulo se presentan los detalles de la arquitectura propuesta en el presente trabajo para abordar COD. Siguiendo la metodología de trabajos recientes en COD [5, 6, 63, 64, 65], se adopta un esquema codificador-decodificador para la arquitectura propuesta AINet. La arquitectura está diseñada para ser entrenable de extremo a extremo, como se ilustra en la Fig. 6.1.

6.1 Arquitectura general

AINet se construye integrando los módulos Mamba y CBAM en un diseño de codificador-decodificador adaptado para COD. Las capacidades de modelado selectivo del espacio de estados de Mamba se combinan con el refinamiento de características que ofrece el módulo CBAM, lo que permite la identificación eficaz de objetos que se mimetizan visualmente con su entorno.

Codificador. Se utiliza un backbone PVTv2-B2 [66] como codificador, que proporciona representaciones jerárquicas de características a múltiples escalas. A partir de la imagen RGB de entrada $I \in R^{H \times W \times 3}$, se extraen mapas de características a cuatro resoluciones diferentes, capturando tanto detalles de bajo nivel como información semántica de alto nivel necesaria para la detección de objetos camuflados.

Decodificador. Se utiliza una estrategia de refinamiento progresivo en el decodificador, donde los Bloques Decodificadores sobremuestrean las características de la etapa anterior y las agregan con conexiones de salto del codificador. La agregación de características se realiza mediante convoluciones de 1×1 y 3×3. Se incorporan múltiples cabezas de segmentación en diferentes niveles del decodificador para proporcionar una supervisión exhaustiva, mejorando el flujo de gradiente y el aprendizaje de características.

Función de Pérdida. Se adopta la función de pérdida propuesta por [67], siguiendo las prácticas establecidas [6, 68]. Las predicciones generadas por el decodificador se denotan como $\{P_i\}_{i=0}^{3}$. Durante el entrenamiento, cada predicción Pi se redimensiona al tamaño de entrada original y se supervisa mediante una combinación de pérdida de Entropía Cruzada Binaria (Binary Cross-Entropy o BCE, por sus

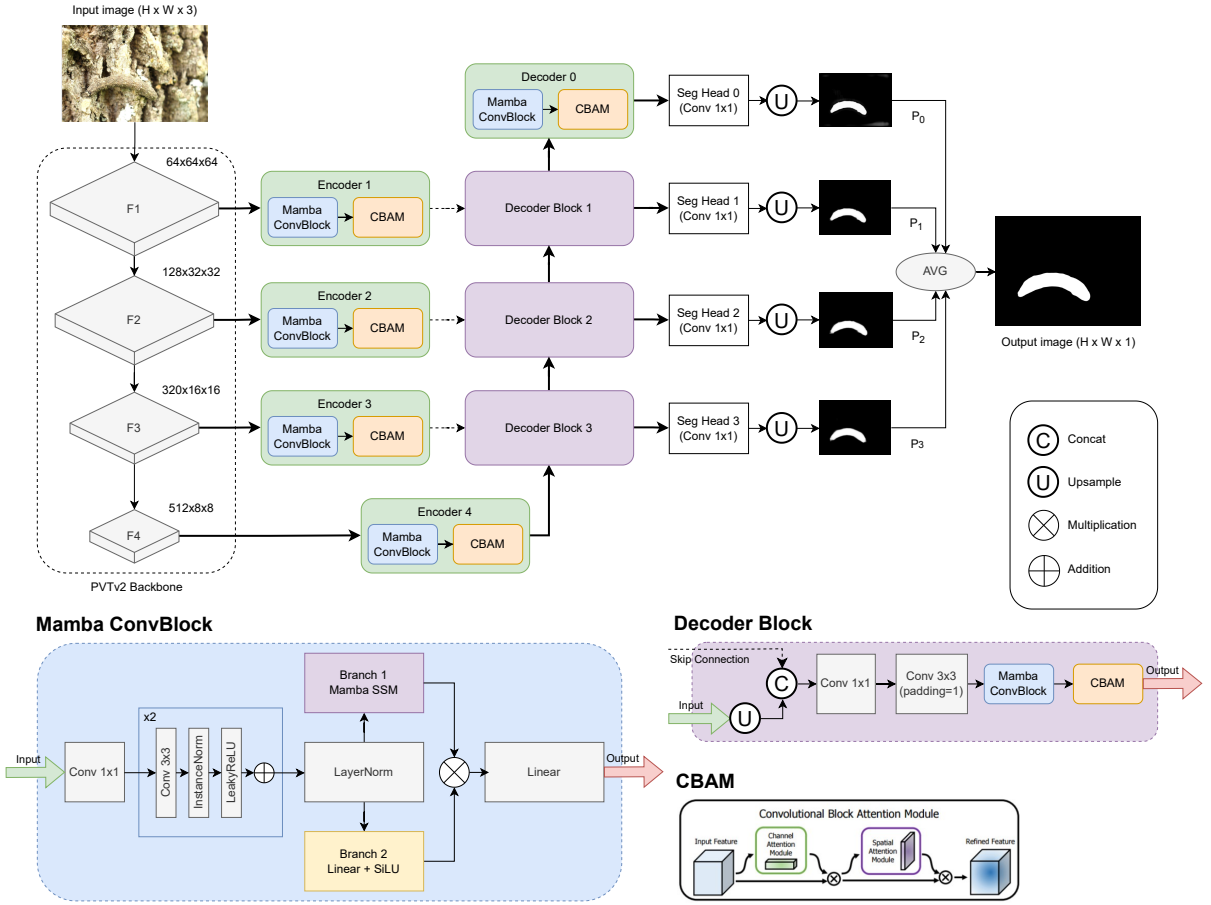

Figura 6.1: Arquitectura propuesta AINet.

siglas en inglés) [69] y pérdida de Intersección sobre Unión (Intersection over Union o IoU, por sus siglas en inglés) [70]. La pérdida total se calcula sumando las pérdidas de todas las etapas del decodificador, como se indica a continuación:

$$\mathcal{L}(P, GT) = \sum_{i=0}^{3} \mathcal{L}_{BCE}(P_i, GT) + \mathcal{L}_{IoU}(P_i, GT). \tag{6.1}$$

Para optimizar el aprendizaje de características a diferentes escalas, se emplea una supervisión profunda con múltiples cabezas de segmentación. Cada nivel del decodificador genera un mapa de segmentación supervisado por el GT, y la predicción final se obtiene promediando los resultados de todos los niveles. Esta estrategia de supervisión multinivel permite el aprendizaje de características robustas para detectar objetos camuflados en diferentes escalas y niveles de camuflaje.

6.2 Bloques Mamba y CBAM

Un componente central de AINet es el Bloque Mamba, que combina el aprendizaje residual con el modelado del espacio de estados [17], siguiendo el enfoque de [71]. El Bloque Mamba se compo-

ne de Bloques Residuales (dos bloques consecutivos con normalización de instancias y activación de LeakyReLU para mejorar la representación de características, preservando la información espacial); Procesamiento de doble rama, que incluye una rama Mamba (que procesa características mediante proyección lineal, convolución 1D y el modelo de espacio de estados Mamba para capturar dependencias de largo alcance) y una rama de activación (que aplica activación de SiLU para la preservación de la información local); y un Producto Hadamard, donde las salidas de ambas ramas se combinan mediante la multiplicación elemento a elemento para la fusión adaptativa de características. La Figura 6.1 ilustra el Bloque Mamba.

Por otro lado, el módulo CBAM [18] también se incorpora para refinar las características aplicando atención de canal y espacial. La atención de canal se calcula mediante agrupación de promedios y máximos, seguida de un MLP compartido para generar ponderaciones de atención por canal, mientras que la atención espacial se genera mediante agrupación de canales y convolución para generar un mapa de atención espacial.

6.3 Detalles de implementación

AINet se implementó usando la biblioteca PyTorch. El codificador se inicializa con un backbone PVTv2-B2 [66] preentrenada en ImageNet. Se utiliza el optimizador AdamW, con un weight decay de 1×10^{-4}. El learning rate inicial se establece en 1×10^{-4} y se usa el scheduler cosine annealing. Las imágenes de entrada se redimensionan a 352×352 tanto para el entrenamiento como para la inferencia. El modelo se entrena durante 100 épocas con un batch size de 16 en una GPU NVIDIA TESLA P100. Todos los experimentos se realizan en la plataforma Kaggle[1].

Tipo	Conjunto de datos	Número de imágenes		
		Train	Val	Test
Referencia	CAMO [3, 4]	797	203	250
	COD10K [5, 6]	2435	605	2026
	NC4K [7]	-	-	4121
Agricultura	Cotton Bollworm [8]	856	161	56

Cuadro 6.1: Distribución de los datasets utilizados en los resultados experimentales.

Conjunto de datos. Para evaluar exhaustivamente la eficacia y la generalización de AINet, se han realizado experimentos con conjuntos de datos de referencia y de plagas camufladas. Para la evaluación general de COD, se han utilizado tres conjuntos de datos de referencia ampliamente reconocidos: CAMO [3, 4], que contiene 2500 imágenes equilibradas entre objetos camuflados y no camuflados; COD10K [5, 6], que ofrece 5066 imágenes camufladas, 3000 de fondo y 1934 no camufladas; y NC4K [7], que aporta 4121 imágenes exclusivamente para pruebas. Siguiendo los protocolos establecidos [5, 64], se utilizan 1000 imágenes de CAMO y 3040 de COD10K para el entrenamiento, reservando el resto para la evaluación.

Para evaluar la aplicabilidad de AINet en entornos agrícolas, se usa el dataset Cotton Bollworm [8], que contiene 1073 imágenes de plagas camufladas. La Figura 6.2 presenta imágenes de muestra de los

[1] https://www.kaggle.com/

conjuntos de datos de referencia y de plagas agrícolas. La distribución de todos los datasets utilizados en este artículo se resume en el Cuadro 6.1.



Figura 6.2: Imágenes de ejemplo de los conjuntos de datos de referencia (CAMO [3, 4], COD10K [5, 6], NC4K [7]) y el dataset de plagas (Cotton Bollworm [8]) utilizados como caso de estudio.

7

Resultados

Este capítulo presenta una evaluación exhaustiva de AINet, centrándose primero en conjuntos de datos de referencia SOTA COD y, posteriormente, en un conjunto de datos de plagas agrícolas como caso de estudio. Se proporcionan análisis cuantitativos y cualitativos, junto con estudios de ablación para evaluar la contribución de los componentes clave de la arquitectura propuesta.

7.1 Detalles de entrenamiento

Esta sección resume la configuración de entrenamiento utilizada en todas las técnicas de COD para compararla con AINet. El Cuadro 7.1 indica el Optimizer, Learning Rate, Batch Size, Epochs, Scheduler y Loss Function empleados por cada modelo.

Cuadro 7.1: Detalles de los parámetros de entrenamiento utilizados en las técnicas SOTA COD evaluadas. Learning rate (LR); Batch size (BS).

Técnica	Optimizer	LR	BS	Epochs	Scheduler	Loss function
BASNet [37]	Adam	1e-3	8	1000	ReduceLROnPlateau	BCE + SSIM + IOU (multi-stage fusion)
SINet-v2 [8]	Adam	1e-4	16	150	Custom (Adjust LR)	Structure loss (weighted BCE + weighted IOU)
BGNet [43]	Adam	1e-4	12	100	Custom (Poly LR)	Structure loss (weighted BCE + weighted IOU) + Dice loss (edge)
C^2F-Net [42]	AdaXW	1e-4	32	50	Custom (Poly LR)	Structure loss (weighted BCE + weighted IOU)
OCENet [44]	Adam	1e-5	4	50	StepLR	Uncertainty aware structure loss (weighted BCE + weighted IOU)
DGNet [47]	AdamW	5e-5	16	150	CosineAnnealingLR	Hybrid loss (weighted BCE + weighted IOU) + MSE loss (grad)
PCNet [48]	AdamW	1e-4	8	150	Custom (Adjust LR)	Structure loss (weighted BCE + weighted IOU)
AINet	AdamW	1e-4	16	150	CosineAnnealingLR	Structure loss (weighted BCE + weighted IOU)

7.2 Evaluación Cuantitativa

Para validar el nivel de generalización, la arquitectura AINet propuesta se comparó con 26 modelos SOTA COD, que se enumeran en el Cuadro 7.2 (*primera columna*). Todos los resultados de las predicciones se obtuvieron directamente de los artículos originales o se generaron utilizando sus modelos

Cuadro 7.2: Resultados experimentales de las técnicas SOTA COD y la arquitectura propuesta AINet en conjunto de datos de referencia. Los tres resultados con mejor rendimiento se resaltan en color: Primero, Segundo y Tercero, respectivamente.

Técnica	Pub/Año	Backbone	CAMO-Test (250)				COD10K-Test (2,026)				NC4K (4,121)			
			$S_\alpha \uparrow$	$M \downarrow$	$F_\beta^m \uparrow$	$E_\phi^m \uparrow$	$S_\alpha \uparrow$	$M \downarrow$	$F_\beta^m \uparrow$	$E_\phi^m \uparrow$	$S_\alpha \uparrow$	$M \downarrow$	$F_\beta^m \uparrow$	$E_\phi^m \uparrow$
BASNet [37]	CVPR19	ResNet34	0.615	0.124	0.503	0.671	0.661	0.071	0.486	0.729	0.696	0.095	0.610	0.762
C^2F-Net [68]	IJCAI21	Res2Net50	0.796	0.080	0.762	0.854	0.813	0.036	0.723	0.890	0.838	0.049	0.795	0.897
LSR [72]	CVPR21	ResNet50	0.708	0.105	0.645	0.755	0.760	0.045	0.658	0.831	0.797	0.061	0.758	0.854
MGL [73]	CVPR21	ResNet50	0.775	0.088	0.726	0.812	0.814	0.035	0.711	0.852	0.833	0.052	0.782	0.867
PFNet [74]	CVPR21	ResNet50	0.782	0.085	0.746	0.841	0.800	0.040	0.701	0.877	0.829	0.053	0.784	0.887
TINet [75]	AAAI21	ResNet50	0.781	0.087	0.728	0.836	0.793	0.042	0.679	0.861	0.829	0.055	0.773	0.879
UGTR [76]	ICCV21	ResNet50	0.785	0.086	0.738	0.823	0.818	0.035	0.712	0.853	0.839	0.052	0.787	0.874
UR-SINet [77]	ACMMM21	ResNet50	0.741	0.091	0.649	0.804	0.775	0.041	0.643	0.869	0.806	0.057	0.731	0.873
BGNet [43]	IJCAI22	Res2Net50	0.812	0.073	0.789	0.870	0.831	0.033	0.753	0.901	0.851	0.044	0.820	0.907
C^2F-Net-V2 [42]	TCSVT22	Res2Net50	0.799	0.077	0.770	0.859	0.811	0.036	0.725	0.887	0.840	0.048	0.802	0.896
ERRNet [78]	PR22	ResNet50	0.779	0.085	0.729	0.842	0.786	0.043	0.675	0.867	0.827	0.054	0.778	0.887
FAP-Net [79]	TIP22	Res2Net50	0.815	0.076	0.776	0.865	0.822	0.036	0.731	0.888	0.851	0.047	0.810	0.899
FindNet [80]	TIP22	Res2Net	0.803	0.077	0.763	0.862	0.811	0.036	0.706	0.883	0.841	0.048	0.802	0.895
OCENet [44]	WACV22	ResNet50	0.802	0.080	0.766	0.852	0.827	0.033	0.741	0.894	0.853	0.045	0.818	0.902
PreyNet [81]	ACMMM22	ResNet50	0.790	0.077	0.757	0.842	0.813	0.034	0.736	0.881	0.834	0.050	0.803	0.887
SegMaR [39]	CVPR22	ResNet50	0.815	0.071	0.795	0.874	0.833	0.034	0.757	0.899	0.841	0.046	0.821	0.896
SINet-v2 [6]	TPAMI22	Res2Net50	0.820	0.070	0.782	0.882	0.815	0.037	0.718	0.887	0.847	0.048	0.805	0.903
DGNet [47]	MIR23	EfficientNet	0.839	0.057	0.806	0.901	0.822	0.033	0.728	0.896	0.857	0.042	0.814	0.911
FEDER [41]	CVPR23	ResNet50	0.807	0.069	0.785	0.873	0.823	0.032	0.740	0.900	0.846	0.045	0.817	0.905
LSR-V2 [13]	TCSVT23	ResNet50	0.789	0.079	0.751	0.840	0.805	0.037	0.711	0.880	0.840	0.048	0.801	0.896
MRR-Net [82]	TNNLS23	ResNet50	0.811	0.076	0.772	0.869	0.822	0.036	0.730	0.889	0.848	0.049	0.801	0.898
PUENet [83]	TIP23	ResNet50	0.794	0.080	0.762	0.857	0.813	0.035	0.727	0.887	0.836	0.050	0.798	0.892
UCOS-DA [84]	ICCVW23	DINO	0.701	0.127	0.646	0.784	0.689	0.086	0.546	0.740	0.755	0.085	0.689	0.819
UJSCOD-V2 [85]	arXiv23	ResNet50	0.803	0.071	0.768	0.858	0.817	0.033	0.733	0.895	0.856	0.040	0.824	0.913
WS-SAM [86]	NeurIPS23	ResNet50	0.759	0.092	0.742	0.818	0.803	0.038	0.719	0.878	0.829	0.052	0.802	0.886
GenSAM [87]	AAAI24	CLIP, BLIP2	0.719	0.113	0.659	0.775	0.775	0.067	0.681	0.838	—	—	—	—
AINet	-	PVTv2	0.840	0.057	0.815	0.909	0.835	0.028	0.757	0.913	0.868	0.036	0.838	0.926

preentrenados publicados oficialmente para una evaluación justa. Como se muestra en el Cuadro 7.2, la arquitectura AINet propuesta supera en las métricas de evaluación a las 26 técnicas SOTA en los tres conjuntos de datos de referencia. Considerando las tres mejores técnicas en diferentes métricas, AINet supera a técnicas como SINet-V2 [6], SegMaR [39], BGNet [43], OCENet [44], DGNet [47], FEDER [41] y UJSCOD-V2 [85]. Entre las técnicas SOTA, AINet supera a enfoques recientes como GenSAM [87], que emplea cadenas intermodales de estímulos mentales para generar estímulos visuales (p. ej., CLIP y BLIP2) y produce máscaras progresivamente, refinando iterativamente los resultados de detección. Otro método reciente que AINet supera es UCOS-DA [84], que implementa una adaptación de dominio no supervisada utilizando DINO a través de un enfoque autoantagónico contrastivo de primer plano-fondo.

Por otro lado, AINet demuestra un rendimiento excepcional para aplicaciones de detección de plagas, como lo demuestra el Cuadro 7.3, que muestra los resultados experimentales de la arquitectura AINet propuesta en comparación con las técnicas SOTA COD existentes probadas en el conjunto de datos Cotton Bollworm. AINet logra valores notables con un valor de S_α de 0,8787, F_β^w de 0,8428 y M más bajo, de 0,0132. En particular, AINet destaca en la detección mejorada de contornos, con las puntuaciones más altas de E_ϕ^{adp} de 0,9691 y E_ϕ^{mean} de 0,9727 y los valores de F_β^{adp} de 0,8330 y F_β^{mean} de 0,8453. Estos resultados superan a métodos consolidados como BASNet [37], SINet-v2 [6], BGNet [43] y PCNet [48], lo que posiciona a AINet como una solución eficaz para la detección de la plaga Cotton

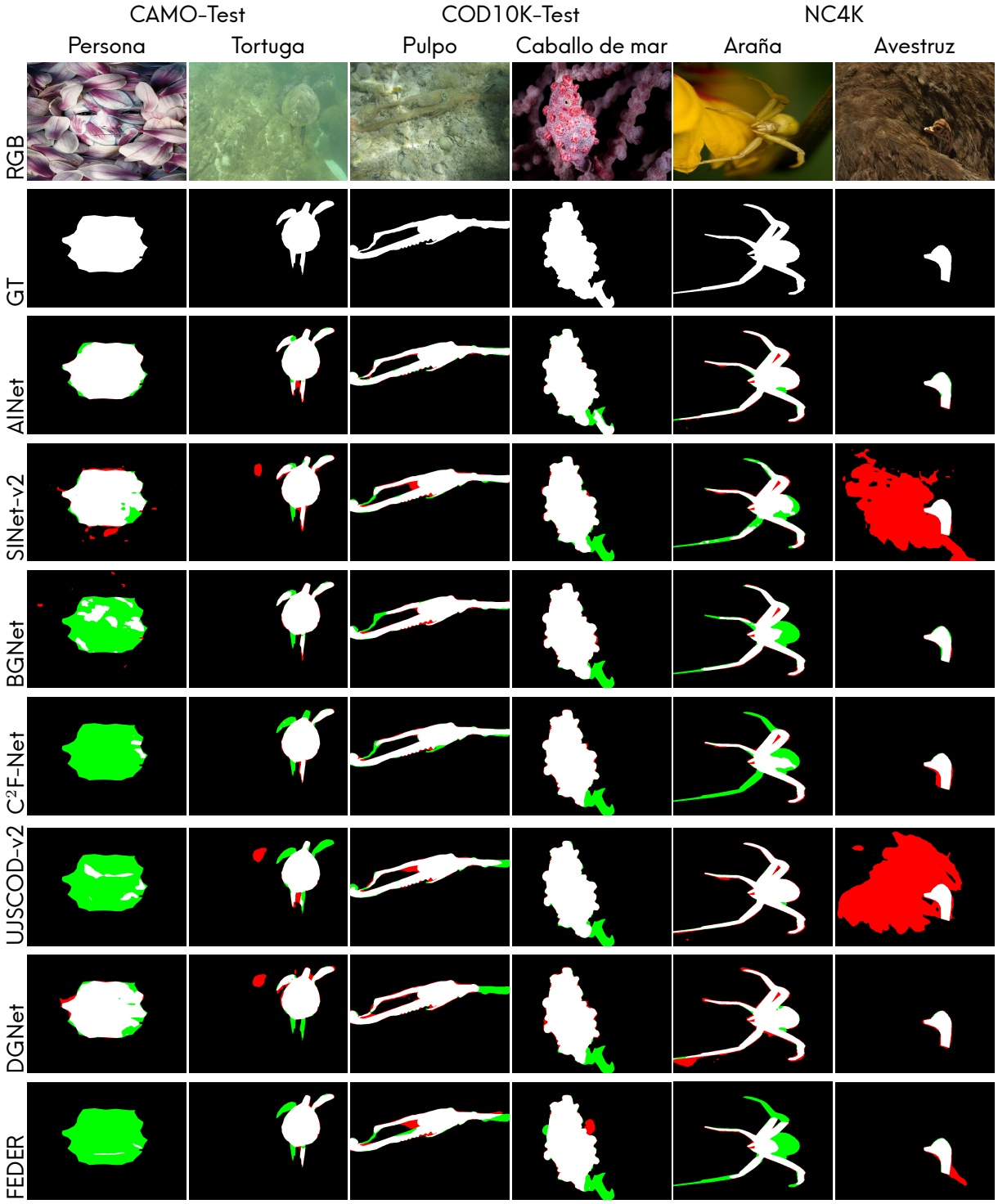

Figura 7.1: Resultados cualitativos de seis técnicas SOTA COD y la arquitectura AINet propuesta, evaluados en imágenes de ejemplo de conjuntos de datos de referencia. Coincidencias exitosas entre GT y máscaras predichas (áreas blancas); regiones con falsos positivos (áreas rojas, sobresegmentación); y regiones con falsos negativos (áreas verdes, segmentación incorrecta).

Bollworm, con mejoras significativas tanto en la precisión como en la exactitud de los contornos.

7.3 Evaluación Cualitativa

La Figura 7.1 presenta los resultados de la evaluación de la arquitectura propuesta AINet y seis técnicas SOTA COD evaluadas en conjunto de datos de referencia: CAMO-Test, COD10K-Test y NC4K. Para identificar visualmente qué técnica presenta el mejor rendimiento, se compara el GT con la máscara predicha para cada imagen. Las coincidencias exitosas entre el GT y las máscaras predichas se representan en blanco; las regiones con falsos positivos, en rojo (sobresegmentación); y las regiones con falsos negativos, en verde (falta de segmentación). Como se puede observar en las imágenes de ejemplo, AINet presenta los mejores resultados en cuanto a la delimitación del área del objeto camuflado, además de no presentar una sobresegmentación ni una falta de segmentación excesivas, superando claramente a las técnicas SOTA COD recientes.

La Figura 7.2 muestra los resultados comparativos del conjunto de datos Cotton Bollworm [8]. La tercera columna muestra los resultados de AINet, que demuestra una detección significativamente más precisa en comparación con otros métodos, identificando correctamente la forma y la ubicación de la plaga en diversos entornos. Se observa que AINet produce máscaras más nítidas y completas, con una mejor delimitación de los bordes y menos falsos positivos. Especialmente en casos difíciles donde el insecto se camufla con el fondo vegetal, AINet mantiene una alta precisión de segmentación, lo que destaca su eficacia para esta aplicación específica de detección de plagas agrícolas.

Finalmente, la robustez de AINet se evidencia en su capacidad para mantener detecciones consistentes y precisas en diversas condiciones de iluminación, desde luz natural intensa hasta zonas de sombra, a la vez que produce mapas de detección significativamente más nítidos, con menos artefactos y ruido que sus competidores. Los bordes detectados por AINet son notablemente más suaves y precisos. Mientras que otros métodos muestran una variabilidad considerable en sus resultados, AINet mantiene una estabilidad superior en diferentes escenarios, lo que lo consolida como la solución más fiable y eficaz para aplicaciones del mundo real.

Cuadro 7.3: Resultados cuantitativos de las técnicas SOTA COD y AINet en el conjunto de datos Cotton Bollworm [8]. Los tres resultados con mejor rendimiento se resaltan en color: Primero, Segundo y Tercero, respectivamente.

Técnica	$S_\alpha \uparrow$	$F_\beta^w \uparrow$	$M \downarrow$	$F_\beta^{adp} \uparrow$	$F_\beta^{mean} \uparrow$	$E_\phi^{adp} \uparrow$	$E_\phi^{mean} \uparrow$
BASNet [37]	0.8335	0.7500	0.0242	0.7538	0.7702	0.9270	0.9049
SINet-v2 [6]	0.9010	0.8677	0.0136	0.8483	0.8685	0.9668	0.9662
BGNet [43]	0.8953	0.8769	0.0133	0.8757	0.8864	0.9708	0.9635
C^2F-Net [42]	0.8861	0.7758	0.0178	0.8369	0.8571	0.9483	0.9341
OCENet [44]	0.9071	0.8692	0.0108	0.8431	0.8674	0.9586	0.9622
DGNet [47]	0.9001	0.8725	0.0104	0.8525	0.8747	0.9615	0.9590
PCNet [48]	0.9000	0.8652	0.0125	0.8532	0.8657	0.9651	0.9647
AINet	0.9208	0.9197	0.0076	0.9092	0.9156	0.9838	0.9820

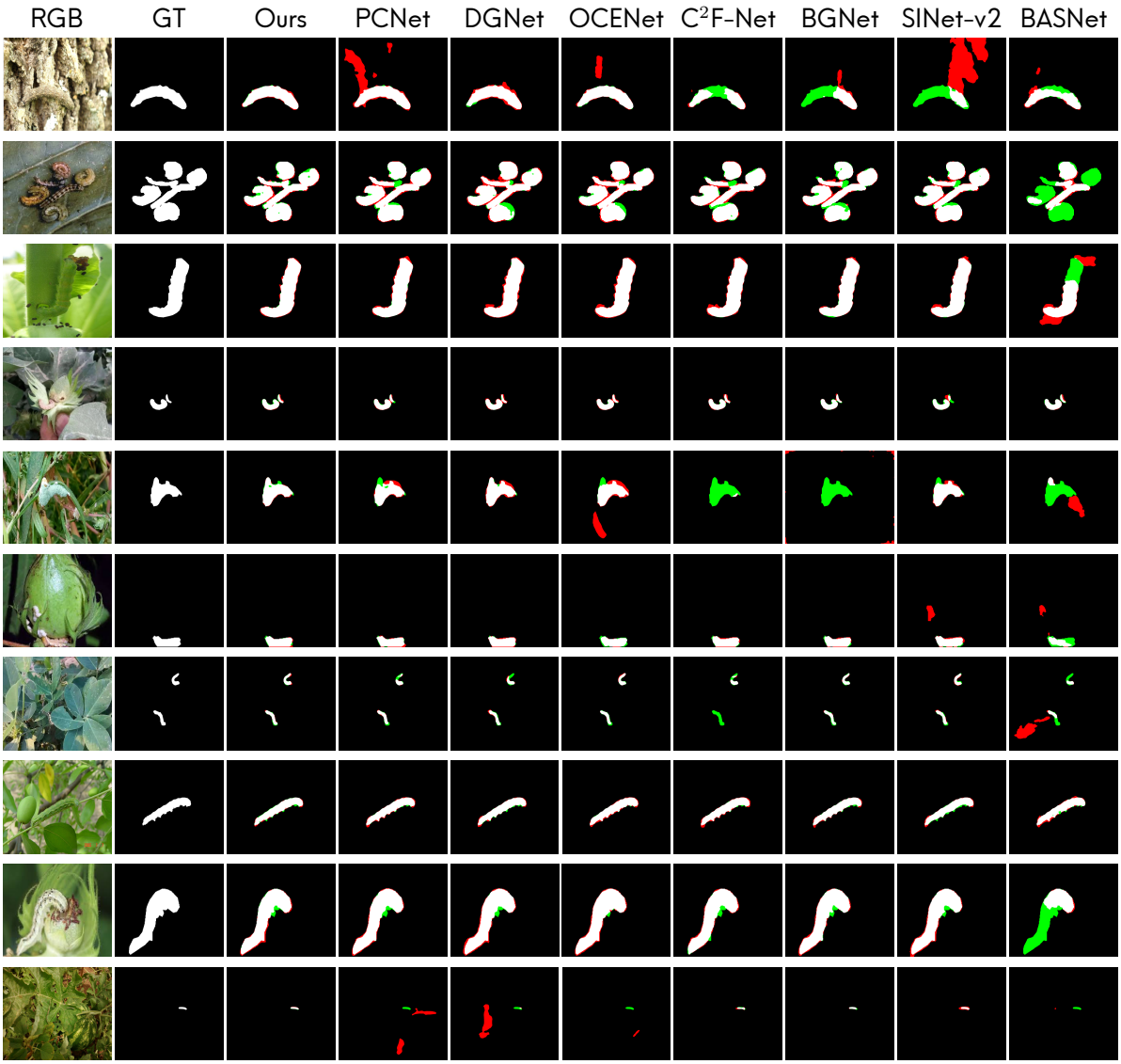

Figura 7.2: Resultados cualitativos de siete técnicas SOTA COD y AINet, evaluados en imágenes de ejemplo del conjunto de datos Cotton Bollworm [8]. Coincidencias exitosas entre GT y máscaras predichas (áreas blancas); regiones con falsos positivos (áreas rojas, sobresegmentación); y regiones con falsos negativos (áreas verdes, segmentación incorrecta).

7.4 Estudio de Ablación

Para evaluar sistemáticamente la contribución de cada componente de AINet, se realizó un análisis de ablación exhaustivo utilizando el conjunto de datos Cotton Bollworm. Los resultados, resumidos en el Cuadro 7.4, destacan el papel crucial de la integración modular. En concreto, la configuración que combina los módulos Mamba y CBAM alcanza el mayor rendimiento en el conjunto de datos Cotton Bollworm, con S_α = 0,9208, F_β^w = 0,9197 y M = 0,0076. En comparación, las configuraciones que utilizan solo Mamba o solo CBAM arrojan puntuaciones de S_α más bajas, de 0,9131 y 0,9118, respectivamente. Esta clara diferencia de rendimiento subraya el efecto de la integración de ambos módulos, ya que

cada uno aporta ventajas complementarias: Mamba mejora la capacidad del modelo para capturar dependencias de largo alcance y contexto secuencial, mientras que CBAM refina las representaciones de características al centrarse en las regiones salientes. Su combinación permite a AINet distinguir con mayor eficacia los objetos camuflados de fondos complejos.

Este rendimiento cuantitativo se corrobora con los resultados cualitativos de la Figura 7.3. Las máscaras de segmentación producidas por la configuración Mamba + CBAM presentan contornos más nítidos y precisos, y menos falsos positivos, mientras que las configuraciones de módulos individuales tienden a generar contornos menos precisos y son más propensas a errores, especialmente en escenarios de camuflaje complejos. El impacto de la supervisión multinivel se analiza en el Cuadro 7.5. La configuración M_1, que incorpora resultados de múltiples capas de predicción ($P_0 + P_1 + P_2 + P_3 + Avg$), logra los mejores resultados generales (S_α = 0.9208, F_β^w = 0.9197), superando a la configuración base Avg. Si bien la configuración Avg alcanza un S_α ligeramente superior (0,9261), presenta una falta de consistencia en otras métricas, lo que indica que la supervisión multinivel contribuye a una segmentación más robusta y fiable. La figura 7.4 corrobora visualmente estos hallazgos, mostrando que M produce máscaras más completas y coherentes, especialmente en regiones donde el camuflaje de los insectos es más pronunciado. Por el contrario, las configuraciones con supervisión reducida (M_3, M_4) suelen dar lugar a segmentaciones incompletas o fragmentadas.

Para analizar los resultados del enfoque propuesto, se utiliza la herramienta de visualización Mapeo de activación de clases ponderado por gradiente (Gradient-weighted Class Activation Mapping o Grad-CAM, por sus siglas en inglés). En el contexto de la detección de plagas camufladas, estas explicaciones visuales son cruciales para verificar que el modelo se basa en señales significativas relacionadas con las plagas y no en artefactos del fondo u otros elementos. La Figura 7.5 presenta visualizaciones de Grad-CAM para las cabezas de segmentación de AINet en el conjunto de datos Cotton Bollworm: la primera y la segunda fila muestran las imágenes RGB y las máscaras del GT correspondientes, seguidas de las máscaras de segmentación predichas, mientras que las filas posteriores representan los mapas de activación generados por cada cabeza de segmentación. Estos mapas de calor revelan que AINet se centra constantemente en las regiones camufladas de las plagas, incluso cuando los insectos están parcialmente ocultos o muy mimetizados con el fondo de la planta, con activaciones concentradas alrededor de las ubicaciones reales de los objetos, lo que demuestra la capacidad del modelo para capturar tanto el contexto global como los detalles precisos necesarios para una detección precisa de plagas.

En resumen, el estudio de ablación demuestra que tanto la integración de los módulos Mamba y CBAM como la aplicación de la supervisión multinivel son indispensables para lograr un rendimiento óptimo con AINet. Su efecto combinado permite una detección de objetos camuflados más precisa, fiable y generalizable, especialmente en el complejo contexto de las imágenes agrícolas.

Cuadro 7.4: Evaluación de métricas con diferentes combinaciones de módulos (p. ej., Mamba y CBAM) en el conjunto de datos Cotton Bollworm [8]. Los mejores resultados de cada métrica se resaltan en negrita.

Módulo	$S_\alpha \uparrow$	$F_\beta^w \uparrow$	$M \downarrow$	$F_\beta^{adp} \uparrow$	$F_\beta^{mean} \uparrow$	$E_\phi^{adp} \uparrow$	$E_\phi^{mean} \uparrow$
Mamba + CBAM	**0.9208**	**0.9197**	**0.0076**	**0.9092**	**0.9156**	**0.9838**	**0.9820**
solo Mamba	0.9131	0.8953	0.0084	0.8877	0.8932	0.9788	0.9784
solo CBAM	0.9118	0.8950	0.0087	0.8877	0.8923	0.9748	0.9744

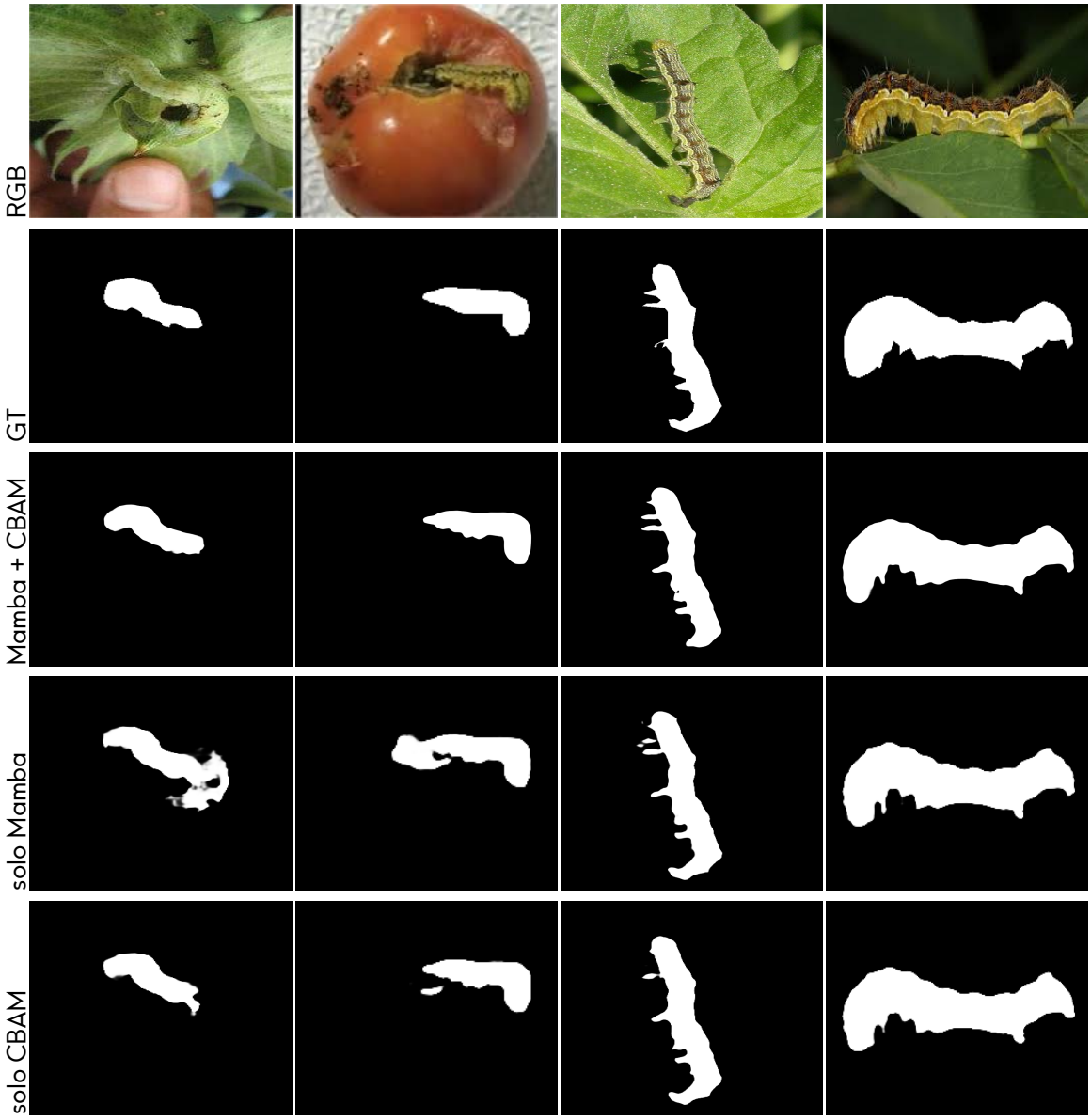

Figura 7.3: Resultados cualitativos de diferentes combinaciones de módulos utilizados, evaluados en imágenes de ejemplo del conjunto de datos Cotton Bollworm [8].

Cuadro 7.5: Evaluación de métricas con diferentes resultados en el conjunto de datos Cotton Bollworm [8]. Los tres resultados con mejor rendimiento se resaltan con color: Primero, Segundo y Tercero, respectivamente.

Nombre	Salida	$S_\alpha \uparrow$	$F_\beta^w \uparrow$	$M \downarrow$	$F_\beta^{adp} \uparrow$	$F_\beta^{mean} \uparrow$	$E_\phi^{adp} \uparrow$	$E_\phi^{mean} \uparrow$
Avg	Avg	0.9261	0.9137	0.0082	0.8885	0.9121	0.9795	0.9799
M_1	$P_0 + P_1 + P_2 + P_3 + Avg$	0.9208	0.9197	0.0076	0.9092	0.9156	0.9838	0.9820
M_2	$P_1 + P_2 + P_3 + Avg$	0.9206	0.9177	0.0077	0.9054	0.9132	0.9829	0.9817
M_3	$P_2 + P_3 + Avg$	0.9210	0.9161	0.0078	0.9017	0.9117	0.9821	0.9813
M_4	$P_3 + Avg$	0.9215	0.9131	0.0080	0.8946	0.9093	0.9804	0.9807

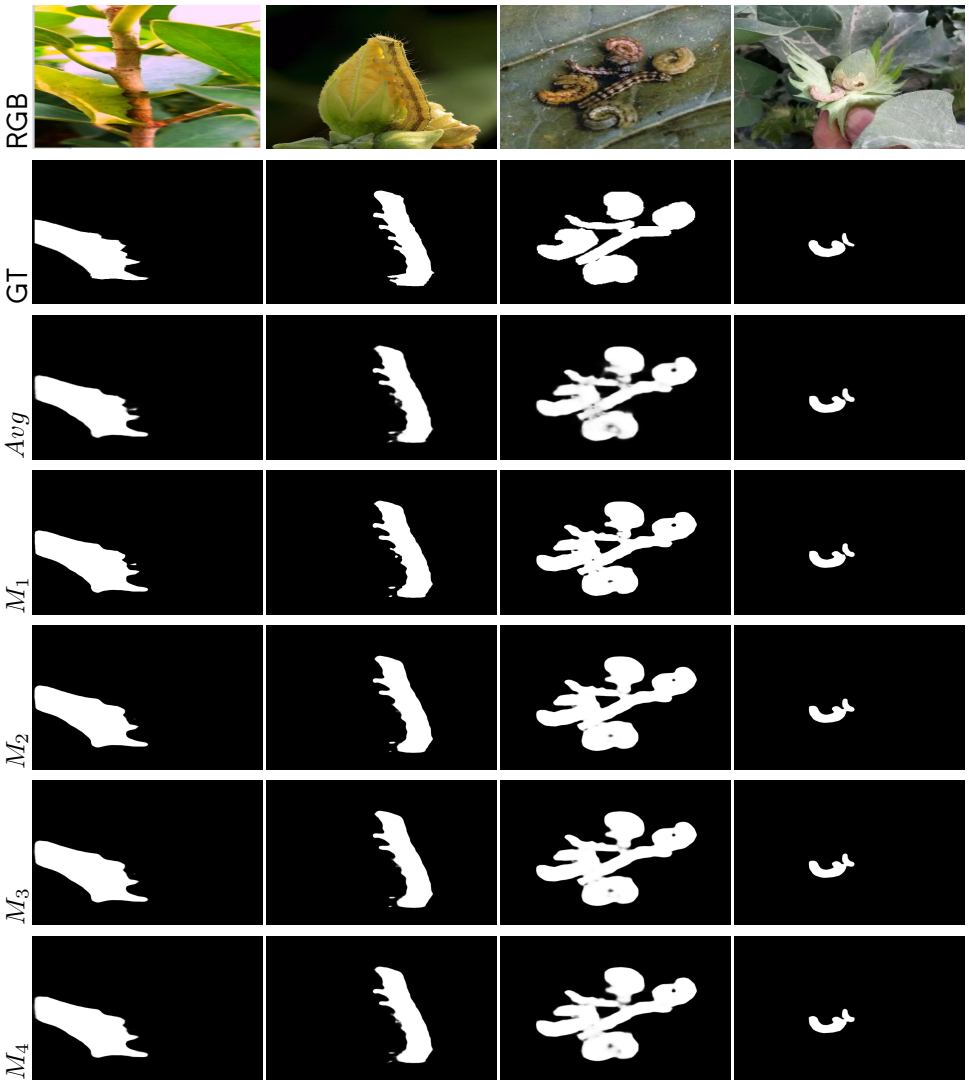

Figura 7.4: Resultados cualitativos de las diferentes salidas de supervisión profunda, evaluados en imágenes de ejemplo del conjunto de datos Cotton Bollworm [8].

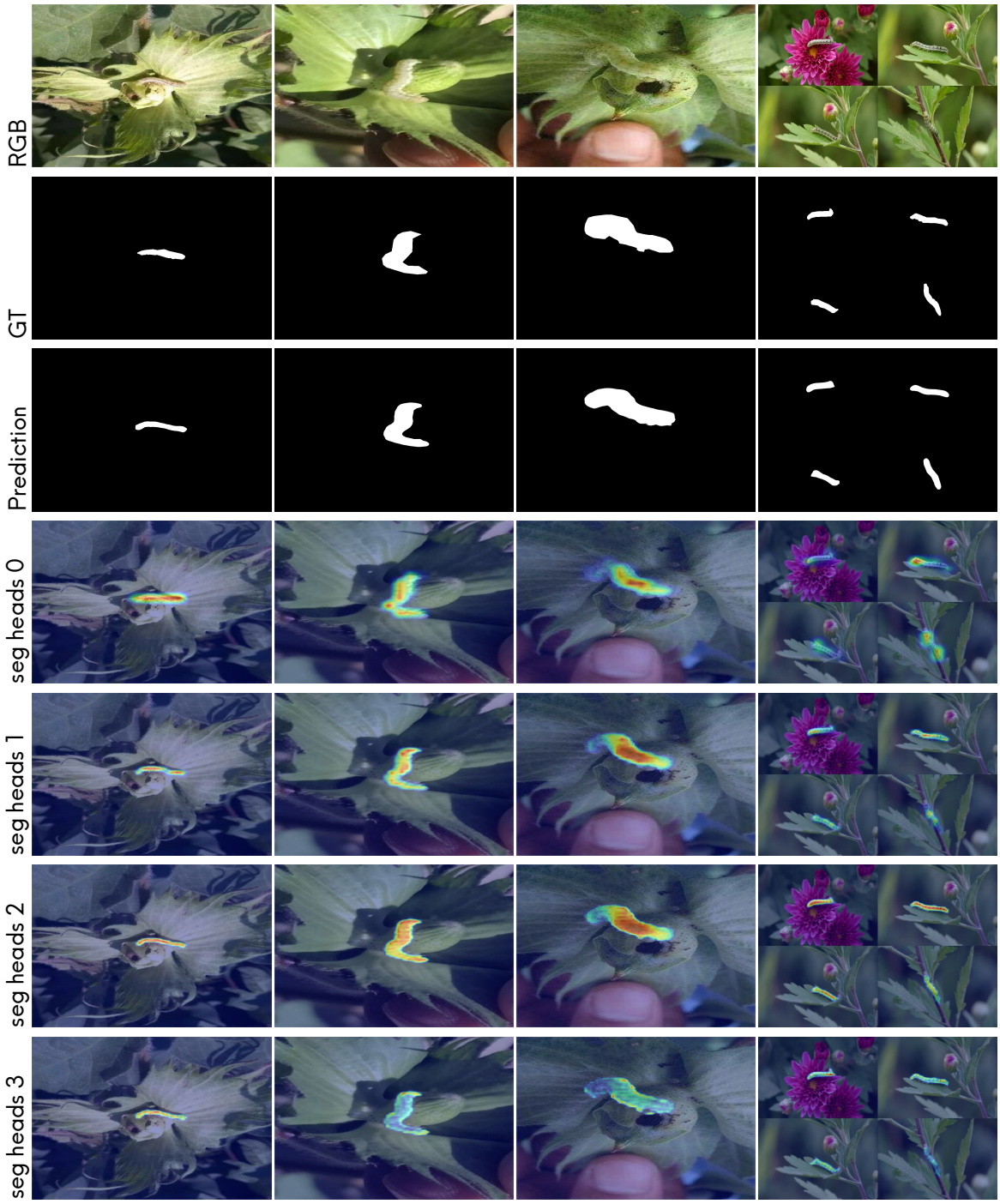

Figura 7.5: Visualización de Grad-CAM utilizando las salidas de las cabezas de segmentación, evaluadas en imágenes de ejemplo del conjunto de datos Cotton Bollworm [8].

8

Conclusiones & Trabajos Futuros

Este trabajo presenta AINet, una novedosa arquitectura para la detección de objetos camuflados que establece un nuevo estándar tanto en conjunto de datos de referencia como en un caso práctico de plaga agrícola. En conjunto de datos de referencia ampliamente reconocidos como (CAMO, COD10K, NC4K), AINet obtiene resultados consistentes, superando a los métodos existentes en todas las métricas evaluadas. El modelo demuestra una precisión de segmentación y delineación de bordes superior, detectando eficazmente una amplia gama de objetos camuflados y consolidándose como una solución líder para escenarios de COD complejos.

Más allá de las validaciones en entornos genéricos, la eficacia de AINet se valida mediante el uso de un conjunto de datos de plagas agrícolas. En el conjunto de datos Cotton Bollworm, AINet destaca en la detección y segmentación de la plaga, logrando una identificación altamente precisa y produciendo máscaras de segmentación nítidas y fiables incluso en entornos complejos donde la plaga es difícil de distinguir de la vegetación.

Los resultados demuestran que AINet no solo supone un avance en la detección general de objetos camuflados, sino que también ofrece importantes beneficios para la agricultura de precisión. Su sólido rendimiento, tanto en conjunto de datos de referencia como agrícolas, destaca su versatilidad y su potencial impacto, lo que favorece una gestión más eficaz de las plagas, reduce las pérdidas de cultivos y facilita el desarrollo de sistemas inteligentes y automatizados para la agricultura moderna.

En resumen, AINet surge como una solución generalizable para la detección de objetos camuflados, ofreciendo mejoras sustanciales en precisión y exactitud en diversas aplicaciones del mundo real.

Referencias

[1] K. Islam, M. Z. Zaheer, A. Mahmood, and K. Nandakumar, "Diffusemix: Label-preserving data augmentation with diffusion models," in *Conf. on Computer Vision and Pattern Recognition*, 2024, pp. 27 621–27 630.

[2] A. Buslaev, V. I. Iglovikov, E. Khvedchenya, A. Parinov, M. Druzhinin, and A. A. Kalinin, "Albumentations: fast and flexible image augmentations," *Information*, vol. 11, no. 2, p. 125, 2020.

[3] T.-N. Le, T. V. Nguyen, Z. Nie, M.-T. Tran, and A. Sugimoto, "Anabranch network for camouflaged object segmentation," *Journal of Computer Vision and Image Understanding*, vol. 184, pp. 45–56, 2019.

[4] J. Yan, T.-N. Le, K.-D. Nguyen, M.-T. Tran, T.-T. Do, and T. V. Nguyen, "Mirrornet: Bio-inspired camouflaged object segmentation," *IEEE Access*, vol. 9, pp. 43 290–43 300, 2021.

[5] D.-P. Fan, G.-P. Ji, G. Sun, M.-M. Cheng, J. Shen, and L. Shao, "Camouflaged object detection," in *CVPR*, 2020.

[6] D.-P. Fan, G.-P. Ji, M.-M. Cheng, and L. Shao, "Concealed object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[7] Y. Lyu, J. Zhang, Y. Dai, A. Li, B. Liu, N. Barnes, and D.-P. Fan, "Simultaneously localize, segment and rank the camouflaged objects," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[8] K. Meng, K. Xu, P. Cattani, and S. Mei, "Camouflaged cotton bollworm instance segmentation based on pvt and mask r-cnn," *Computers and Electronics in Agriculture*, vol. 226, p. 109450, 2024.

[9] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 6, pp. 1137–1149, 2016.

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

[11] J. Zhong, A. Wang, C. Ren, and J. Wu, "A survey on deep learning-based camouflaged object detection," *Multimedia Systems*, vol. 30, no. 5, p. 268, 2024.

[12] F. Xiao, S. Hu, Y. Shen, C. Fang, J. Huang, C. He, L. Tang, Z. Yang, and X. Li, "A survey of camouflaged object detection and beyond," *arXiv preprint arXiv:2408.14562*, 2024.

[13] Y. Lv, J. Zhang, Y. Dai, A. Li, N. Barnes, and D.-P. Fan, "Toward deeper understanding of camouflaged object detection," *IEEE transactions on circuits and systems for video technology*, vol. 33, no. 7, pp. 3462–3476, 2023.

[14] P. Jiang, Y. Chen, B. Liu, D. He, and C. Liang, "Real-time detection of apple leaf diseases using deep learning approach based on improved convolutional neural networks," *Ieee Access*, vol. 7, pp. 59 069–59 080, 2019.

[15] I. R. Evangelista, A. Bandala, and E. Dadios, "Fcnet: A transformer-based context-aware segmentation framework for detecting camouflaged fruits in orchard environments," *Technologies*, vol. 13, no. 8, p. 372, 2025.

[16] L. M. Dang, S. Danish, A. Khan, N. Alam, M. Fayaz, D. K. Nguyen, H.-K. Song, and H. Moon, "An efficient zero-labeling segmentation approach for pest monitoring on smartphone-based images," *European Journal of Agronomy*, vol. 160, p. 127331, 2024.

[17] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," *arXiv preprint arXiv:2312.00752*, 2023.

[18] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

[19] G. Jocher and J. Qiu, "Ultralytics yolo11," 2024. [Online]. Available: https://github.com/ultralytics/ultralytics

[20] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[21] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 28, 2015.

[22] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Conf. on Computer Vision and Pattern Recognition*, 2016.

[23] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *European Conference on Computer Vision*, 2016.

[24] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Int. Conf. on Computer Vision*, 2017, pp. 2980–2988.

[25] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *Int. Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, 2010.

[26] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European Conference on Computer Vision*, 2014.

[27] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European Conference on Computer Vision*, 2020.

[28] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.

[29] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[30] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *Conf. on Computer Vision and Pattern Recognition*. IEEE, 2009, pp. 1597–1604.

[31] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang, "Saliency detection via graph-based manifold ranking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[32] Q. Yan, L. Xu, J. Shi, and J. Jia, "Hierarchical saliency detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[33] L. Wang, H. Lu, Y. Wang, M.-M. Feng, X. Ding, and X. Ruan, "Learning to detect salient objects with image-level supervision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[34] G. Li and Y. Yu, "Visual saliency based on multiscale deep features," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[35] Q. Hou, M.-M. Cheng, X. Hu, A. Borji, Z. Tu, and P. Torr, "Deeply supervised salient object detection with short connections," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[36] N. Liu, J. Han, and M.-H. Yang, "Picanet: Learning pixel-wise contextual attention for saliency detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[37] X. Qin, Z. Zhang, C. Huang, C. Gao, M. Dehghan, and M. Jagersand, "Basnet: Boundary-aware salient object detection," in *Conf. on Computer Vision and Pattern Recognition*, June 2019.

[38] C. A. Javalagi, K. M. Medha, N. T. Patil, S. Itagalli, U. Kulkarni, and S. Chikkamath, "Comparative study of cnns for camouflaged object detection," in *Proceedings of 4th International Conference on Recent Trends in Machine Learning, IoT, Smart Cities and Applications*, V. K. Gunjan and J. M. Zurada, Eds. Singapore: Springer Nature Singapore, 2024, pp. 207–220.

[39] Q. Jia, S. Yao, Y. Liu, X. Fan, R. Liu, and Z. Luo, "Segment, magnify and reiterate: Detecting camouflaged objects the hard way," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4713–4722.

[40] C. He, K. Li, Y. Zhang, Y. Zhang, Z. Guo, X. Li, M. Danelljan, and F. Yu, "Strategic preys make acute predators: Enhancing camouflaged object detectors by generating camouflaged objects," *arXiv preprint arXiv:2308.03166*, 2023.

[41] C. He, K. Li, Y. Zhang, L. Tang, Y. Zhang, Z. Guo, and X. Li, "Camouflaged object detection with feature decomposition and edge reconstruction," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 22 046–22 055.

[42] G. Chen, S.-J. Liu, Y.-J. Sun, G.-P. Ji, Y.-F. Wu, and T. Zhou, "Camouflaged object detection via context-aware cross-level fusion," *Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 10, pp. 6981–6993, 2022.

[43] T. Chen, J. Xiao, X. Hu, G. Zhang, and S. Wang, "Boundary-guided network for camouflaged object detection," *Knowledge-based systems*, vol. 248, p. 108901, 2022.

[44] J. Liu, J. Zhang, and N. Barnes, "Modeling aleatoric uncertainty for camouflaged object detection," in *Winter conference on applications of computer vision*, 2022, pp. 1445–1454.

[45] X. Hu, S. Wang, X. Qin, H. Dai, W. Ren, D. Luo, Y. Tai, and L. Shao, "High-resolution iterative feedback network for camouflaged object detection," in *Conf. on Artificial Intelligence*, vol. 37, no. 1, 2023, pp. 881–889.

[46] D. Sun, S. Jiang, and L. Qi, "Edge-aware mirror network for camouflaged object detection," in *Int. Conf. on Multimedia and Expo*. IEEE, 2023, pp. 2465–2470.

[47] G.-P. Ji, D.-P. Fan, Y.-C. Chou, D. Dai, A. Liniger, and L. Van Gool, "Deep gradient learning for efficient camouflaged object detection," *Machine Intelligence Research*, vol. 20, no. 1, pp. 92–108, 2023.

[48] J. Yang, Q. Wang, F. Zheng, P. Chen, A. Leonardis, and D.-P. Fan, "Plantcamo: Plant camouflage detection," *arXiv*, 2024.

[49] S. Lev-Yadun, A. Dafni, M. A. Flaishman, M. Inbar, I. Izhaki, G. Katzir, and G. Ne'eman, "Plant coloration undermines herbivorous insect camouflage," *BioEssays*, vol. 26, no. 10, pp. 1126–1130, 2004.

[50] T.-N. Le, V. Nguyen, C. Le, T.-C. Nguyen, M.-T. Tran, and T. V. Nguyen, "Camoufinder: Finding camouflaged instances in images," in *AAAI Conference on Artificial Intelligence*, 2021.

[51] T.-N. Le, Y. Cao, T.-C. Nguyen, M.-Q. Le, K.-D. Nguyen, T.-T. Do, M.-T. Tran, and T. V. Nguyen, "Camouflaged instance segmentation in-the-wild: Dataset, method, and benchmark suite," *IEEE Transactions on Image Processing*, vol. 31, pp. 287–300, 2022.

[52] D.-P. Fan, G.-P. Ji, M.-M. Cheng, and L. Shao, "Concealed object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[53] Y. Zheng, X. Zhang, F. Wang, T. Cao, M. Sun, and X. Wang, "Detection of people with camouflage pattern via dense deconvolution network," *IEEE Signal Processing Letters*, vol. 26, no. 1, pp. 29–33, 2018.

[54] H. Velesaca, H. Villegas, and A. Sappa, "Avnet: Cross-spectral attention-vision model for camouflaged object detection in ecological conservation," in *Proceedings of the 14th International Conference on Computer Vision Theory and Applications*, INSTICC. SciTePress, 2026, pp. 1–10.

[55] P. Lindenberger, P.-E. Sarlin, and M. Pollefeys, "Lightglue: Local feature matching at light speed," *arXiv preprint arXiv:2306.13643*, 2023.

[56] D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, and A. Borji, "Structure-measure: A new way to evaluate foreground maps," in *Int. Conference on Computer Vision*, 2017, pp. 4548–4557.

[57] R. Margolin, L. Zelnik-Manor, and A. Tal, "How to evaluate foreground maps?" in *Conf. on Computer Vision and Pattern Recognition*, 2014, pp. 248–255.

[58] F. Perazzi, P. Krähenbühl, Y. Pritch, and A. Hornung, "Saliency filters: Contrast based filtering for salient region detection," in *Conf. on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 733–740.

[59] D.-P. Fan, C. Gong, Y. Cao, B. Ren, M.-M. Cheng, and A. Borji, "Enhanced-alignment measure for binary foreground map evaluation," *arXiv*, 2018.

[60] H. O. Velesaca and A. D. Sappa, "Enhancing camouflaged object detection via diffusion model

augmentation," in *International Conference on Image Analysis and Processing*. Springer, 2025, pp. 352–364.

[61] ——, "Seeing the unseen: Ai-powered camouflaged pest detection," in *Multi Conference on Computer Science and Information Systems*, 2025, pp. 137–144.

[62] A. Buslaev, V. I. Iglovikov, E. Khvedchenya, A. Parinov, M. Druzhinin, and A. A. Kalinin, "Albumentations: Fast and flexible image augmentations," *Information*, vol. 11, no. 2, 2020. [Online]. Available: https://www.mdpi.com/2078-2489/11/2/125

[63] X. Jiang, W. Cai, Z. Zhang, B. Jiang, Z. Yang, and X. Wang, "Magnet: A camouflaged object detection network simulating the observation effect of a magnifier," *Entropy*, vol. 24, no. 12, p. 1804, 2022.

[64] Y. Pang, X. Zhao, T.-Z. Xiang, L. Zhang, and H. 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*, 2022, pp. 2160–2170.

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

[66] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pvt v2: Improved baselines with pyramid vision transformer," *Computational Visual Media*, vol. 8, no. 3, pp. 415–424, 2022.

[67] J. Wei, S. Wang, and Q. Huang, "F^3net: fusion, feedback and focus for salient object detection," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 07, 2020, pp. 12 321–12 328.

[68] Y. Sun, G. Chen, T. Zhou, Y. Zhang, and N. Liu, "Context-aware cross-level fusion network for camouflaged object detection," in *IJCAI*, 2021, pp. 1025–1031.

[69] P.-T. De Boer, D. P. Kroese, S. Mannor, and R. Y. Rubinstein, "A tutorial on the cross-entropy method," *Annals of operations research*, vol. 134, pp. 19–67, 2005.

[70] G. Máttyus, W. Luo, and R. Urtasun, "Deeproadmapper: Extracting road topology from aerial images," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 3438–3446.

[71] J. Ma, F. Li, and B. Wang, "U-mamba: Enhancing long-range dependency for biomedical image segmentation," *arXiv:2401.04722*, 2024.

[72] Y. Lv, J. Zhang, Y. Dai, A. Li, B. Liu, N. Barnes, and D.-P. Fan, "Simultaneously localize, segment and rank the camouflaged objects," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 11 591–11 601.

[73] Q. Zhai, X. Li, F. Yang, C. Chen, H. Cheng, and D.-P. Fan, "Mutual graph learning for camouflaged object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 12 997–13 007.

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

[75] J. Zhu, X. Zhang, S. Zhang, and J. Liu, "Inferring camouflaged objects by texture-aware interactive guidance network," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, no. 4, 2021, pp. 3599–3607.

[76] F. Yang, Q. Zhai, X. Li, R. Huang, A. Luo, H. Cheng, and D.-P. Fan, "Uncertainty-guided transformer reasoning for camouflaged object detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 4146–4155.

[77] N. Kajiura, H. Liu, and S. Satoh, "Improving camouflaged object detection with the uncertainty of pseudo-edge labels," in *Proceedings of the 3rd ACM International Conference on Multimedia in Asia*, 2021, pp. 1–7.

[78] G.-P. Ji, L. Zhu, M. Zhuge, and K. Fu, "Fast camouflaged object detection via edge-based reversible re-calibration network," *Pattern Recognition*, vol. 123, p. 108414, 2022.

[79] T. Zhou, Y. Zhou, C. Gong, J. Yang, and Y. Zhang, "Feature aggregation and propagation network for camouflaged object detection," *IEEE Transactions on Image Processing*, vol. 31, pp. 7036–7047, 2022.

[80] P. Li, X. Yan, H. Zhu, M. Wei, X.-P. Zhang, and J. Qin, "Findnet: Can you find me? boundary-and-texture enhancement network for camouflaged object detection," *IEEE Transactions on Image Processing*, vol. 31, pp. 6396–6411, 2022.

[81] M. Zhang, S. Xu, Y. Piao, D. Shi, S. Lin, and H. Lu, "Preynet: Preying on camouflaged objects," in *Proceedings of the 30th ACM international conference on multimedia*, 2022, pp. 5323–5332.

[82] X. Yan, M. Sun, Y. Han, and Z. Wang, "Camouflaged object segmentation based on matching–recognition–refinement network," *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[83] Y. Zhang, J. Zhang, W. Hamidouche, and O. Deforges, "Predictive uncertainty estimation for camouflaged object detection," *IEEE Transactions on Image Processing*, vol. 32, pp. 3580–3591, 2023.

[84] Y. Zhang and C. Wu, "Unsupervised camouflaged object segmentation as domain adaptation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4334–4344.

[85] A. Li, J. Zhang, Y. Lv, T. Zhang, Y. Zhong, M. He, and Y. Dai, "Joint salient object detection and camouflaged object detection via uncertainty-aware learning," *arXiv preprint arXiv:2307.04651*, 2023.

[86] C. He, K. Li, Y. Zhang, G. Xu, L. Tang, Y. Zhang, Z. Guo, and X. Li, "Weakly-supervised concealed object segmentation with sam-based pseudo labeling and multi-scale feature grouping," *Advances in Neural Information Processing Systems*, vol. 36, pp. 30 726–30 737, 2023.

[87] J. Hu, J. Lin, S. Gong, and W. Cai, "Relax image-specific prompt requirement in sam: A single generic prompt for segmenting camouflaged objects," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 11, 2024, pp. 12 511–12 518.

Archivos adjuntos

Anexo A

A continuación se presenta la lista de publicaciones (con sus respectivos PDFs) resultantes del trabajo de investigación de esta tesis.

Revistas

- Velesaca, Henry O and Mero, Andrea and Villegas, Hector and Sappa, Angel D. (2026). *Early Detection of Camouflaged Weeds in Agriculture using Near-Infrared*. Smart Agricultural Technology (Under Review)
- Velesaca, Henry O and Mero, Andrea and Reyes-Angulo, Abel and Angel Sappa. (2026). *AINet: Integrating Mamba and CBAM for Enhanced Camouflage Object Detection*. IEEE Access (Under Review).

Conferencias

- Velesaca, Henry O and Sappa, Angel D. (2025). *Seeing the Unseen: AI-Powered Camouflaged Pest Detection*. Multi Conference on Computer Science and Information Systems.
- Velesaca, Henry O and Villegas, Hector and Sappa, Angel D. (2025). *Exploring Camouflaged Object Detection Techniques for Invasive Vegetation Monitoring*. International Conference on Data Science, Technology and Applications.
- Velesaca, Henry O and Sappa, Angel D. (2025). *Enhancing Camouflaged Object Detection via Diffusion Model Augmentation*. International Conference on Image Analysis and Processing.
- Velesaca, Henry O and Mero, Andrea and Singh, Anjali and Sappa, Angel D. (2025). *The Hidden Cost of Poor Annotations: How Label Quality Affects Camouflaged Object Detection Performance*. International Symposium on Visual Computing.
- Velesaca, Henry O and Mero, Andrea and Rivadeneira, Rafael and Castillo, Guillermo A. and Sappa, Angel D. (2026). *AVNet: Cross-Spectral Attention-Vision Model for Camouflaged Object Detection in Ecological Conservation*. International Conference on Computer Vision Theory and Applications.

Unveiling the Hidden: Early Detection of Invasive Vegetation in Crops with UAV Multispectral Imaging

Henry O. Velesaca[a,c], Andrea Mero[a], Hector Villegas[a], Angel D. Sappa[a,b]

[a] *ESPOL Polytechnic University, FIEC, CIDIS, Campus Gustavo Galindo, 09-01-5863, Guayaquil, Ecuador*
[b] *Computer Vision Center, Universitat Autònoma de Barcelona, 08193-Bellaterra, Barcelona, Spain*
[c] *Software Engineering Department, Research Center for Information and Communication Technologies (CITIC-UGR), University of Granada, 18071, Granada, Spain*

Abstract

This work presents a comprehensive framework for the early detection of camouflaged weeds in agriculture using near-infrared (NIR) imagery. The proposed methodology combines RGB-NIR fusion techniques with camouflaged object detection (COD) approaches to address the challenge of identifying weeds that blend seamlessly into their agricultural environment. Fourteen state-of-the-art (SOTA) fusion techniques are evaluated, with the top three performing methods demonstrating superior integration of multispectral information. Subsequently, nine SOTA COD techniques are trained on both RGB and RGB-NIR fused datasets to assess detection performance across different spectral configurations. Experimental results on two UAV-based case studies—(i) our in-house Weeds-Banana dataset acquired in a commercial banana plantation and (ii) the SOTA WeedsGalore dataset collected in a maize field—demonstrate that RGB–NIR fusion consistently improves segmentation accuracy over RGB-only inputs. This multispectral approach offers a promising solution for precision agriculture, enabling early detection of camouflaged weeds and reducing dependence on chemical herbicides for sustainable crop management. The dataset is available at GitHub: https://github.com/hvelesaca/cod_rgb-nir_weed.

Email addresses: hvelesac@espol.edu.ec, hvelesaca@correo.ugr.es (Henry O. Velesaca), anmero@espol.edu.ec (Andrea Mero), hvillega@espol.edu.ec (Hector Villegas), asappa@espol.edu.ec, asappa@cvc.uab.es (Angel D. Sappa)

Keywords: `camouflaged object detection; computer vision approaches; unmanned aerial vehicle; weed detection; pest detection`

1. Introduction

The increasing global demand for agricultural production, driven by population growth, presents significant challenges for the implementation of intelligent systems in agriculture (e.g., [1],[2]). Moreover, the adoption of automated inspections in agriculture and the food industry has become an increasingly attractive solution for the final assessment of product quality [3]. Modern precision agriculture, without a doubt, relies on computer vision techniques to address critical tasks such as detecting plant diseases, predicting crop yields, recognizing species, identifying pests——including weeds——and managing irrigation [4].

One of the key tasks in computer vision is object detection, which is categorized into subfields: Generic Object Detection, Salient Object Detection, and Camouflaged Object Detection (COD) [5]. COD focuses on segmenting concealed objects, typically framed as a binary segmentation task [6]. In agriculture, pests often rely on camouflage as a survival mechanism [7], posing a real-world challenge due to variations in lighting, viewpoint, and environmental complexity. In general, pests blend into their environment, making them difficult to identify [8]. Similarly, weeds, which significantly hinder crop productivity, are often targeted by the widespread and unregulated application of chemical herbicides, leading to environmental degradation and contamination of agricultural land [4]. In such cases, the need for real-time COD becomes highly beneficial.

In recent years, several studies have explored the spectral signatures of camouflaged objects in greater depth by expanding spectral detection ranges and improving spectral resolution capabilities. According to [9], conventional approaches often rely on manually designed low-level features to identify essential image characteristics, while deep learning techniques autonomously learn hi-

erarchical feature representations directly from data, consistently outperforming traditional methods across various computer vision applications. Furthermore, [10] classifies COD techniques into several categories, including Human Vision, Infrared, Multispectral, Hyperspectral, and LiDAR. It is also important to highlight that exploring the other spectral bands, such as the thermal or near-infrared (NIR) ranges, offers significant advantages in low-visibility conditions. NIR imaging, in particular, can provide valuable insights into the material composition of objects [11]. Also, camouflage detection systems have become increasingly compatible with multispectral technology, intelligent adaptation, and integrated design. However, despite advances in multispectral composite detection technology, single-band camouflaged object detection remains inadequate to meet complex requirements [12]. Spectral imaging creates a multidimensional data cube that includes both spatial and spectral data, with each pixel represented as a spectral vector. This spectral vector, which contains radiance or reflectance values across multiple spectral bands, enhances the representation of targets compared to traditional RGB images, which only capture two-dimensional spatial information [13].

The main motivation of this study is to explore COD methodologies within the visible and NIR spectrum, addressing a gap in the current computer vision literature, as this area remains underexplored. The paper explores the fusion of RGB-NIR images for weed detection using camouflaged object detection techniques for precise segmentation.

The key contributions of this work include:

- New benchmark dataset using RGB and NIR images for weed detection using UAVs.
- RGB-NIR fusion improves early detection of camouflaged weeds in banana and maize fields.
- UAV-based dataset enables robust multispectral weed detection.
- RGB-NIR fusion outperforms RGB-only methods in all evaluated metrics.

- Combination PCNet-LatLRR and BASNet-FDPE achieves the highest accuracy for camouflaged weed segmentation in banana and maize crops respectively.

The manuscript is organized as follows. Section 2 provides a review of related work on RGB-NIR fusion methods and COD techniques. Section 3 details the proposed methodology for enhancing weed segmentation in crops through the integration of RGB-NIR fusion and COD approaches. Section 4 describes two case studies where the methodology is applied to UAV imagery from banana and maize plantations, along with the presentation of experimental results, comparative analyses, and discussion of various COD methods. Finally, Section 5 discusses the results and Section 6 summarizes the main conclusions of the study.

2. Background

This section reviews the most relevant existing techniques for detecting camouflaged agricultural pests, with a particular focus on fusion detection methods and camouflaged object detection (COD) strategies. These approaches form the foundation for the proposed methodology and support its development.

2.1. Fusion Approaches

Fusion detection methods combine information from multiple image modalities to generate a single, more informative representation, thereby enhancing application performance [14]. As described in [15], early fusion methods include the Cross-Bilateral Filter (CBF), which merges images by leveraging both intensity and geometric similarities. By using one image to guide the filtering of another, CBF preserves edges and reduces noise, making it especially effective for multi-focus and multisensor image fusion tasks. Building on the concept of edge-preserving fusion, [16] introduced the hybrid multiscale decomposition with a guided filter (HMSD-GF), specifically designed for fusing infrared and visible images to improve night vision. By integrating guided filtering within a multiscale decomposition framework, HMSD-GF effectively extracts complementary

information from both spectral domains, excelling at enhancing visibility in low-light conditions while maintaining the structural integrity of features from the visible spectrum.

In addition to these methods, Latent Low-Rank Representation (LatLRR) [17] provides a decomposition-based approach for IR-visible fusion. This technique separates each image into a low-rank component (capturing global structures) and a salient component (preserving local details). The low-rank layers are fused using weighted averaging to minimize redundancy, while salient layers are merged by summation to retain fine-grained features. This dual-branch fusion architecture achieves a balance between structural consistency and detail enhancement.

Expanding on these paradigms, the Guided Filter-based Context enhancement and Fusion model (GFCB) [16] advances night-vision image fusion by employing the guided filter for both adaptive visibility enhancement and multi-scale image decomposition. Initially, the GFCB model enhances the visible image through high dynamic range-inspired compression and contrast restoration, improving details in poorly illuminated regions while preserving edges. Subsequently, a hybrid multi-scale decomposition based solely on the guided filter extracts texture and edge features from both IR and visible images. These components are then fused using a perceptual saliency-driven weighting mechanism that dynamically adjusts the contribution of IR information, ensuring that important thermal features enhance the visible scene without compromising background details. This approach effectively balances noise suppression and detail preservation, thanks to the edge-aware smoothing properties of the guided filter, resulting in fused images with improved contextual clarity and suitability for human perception. Additionally, a perceptual-based parameter selection strategy, inspired by human visual system models, automatically adjusts fusion parameters, increasing robustness and reducing the need for manual tuning. Experimental results show that GFCB outperforms both traditional and other state-of-the-art methods in terms of visual quality and computational efficiency, making it a powerful tool for IR-visible image fusion in challenging

low-light scenarios [16].

All these fusion schemes highlight the transformative potential of integrating multimodal data, particularly in IR-visible fusion, to address challenges such as low-light conditions. Each spectral band contributes unique and complementary information essential for effective problem-solving in real-world applications. According to the literature, each reviewed method offers distinct advantages. Therefore, in this work, several state-of-the-art fusion techniques are evaluated to identify the most effective ones.

2.2. Camouflaged Object Detection Approaches

While fusion techniques leverage complementary information from different modalities to improve detection, COD presents unique challenges that require specialized solutions. Unlike conventional object detection, COD must address objects that intentionally blend into their surroundings through similarities in color, texture, and shape. This has led to the development of tailored methods that focus on detecting subtle visual cues and fine boundaries. Recent surveys, such as the comprehensive review by Xiao et al. [18], classify COD methodologies into traditional and deep learning-based approaches, illustrating the field's evolution. Yang et al. [19] introduced PlantCamo, the first dataset dedicated to camouflaged plant detection, containing 1,250 images across 58 plant categories and four camouflage types: background matching, disruptive coloration, masquerade, and decoration.

To address challenges such as irregular shapes and complex boundaries, PCNet introduces a Multi-scale Global Feature Enhancement module for bottom-up feature fusion and a Multi-scale Feature Refinement module with top-down iterative feedback for precise boundary delineation [19]. This Transformer-based architecture effectively captures both global context and fine details. Notably, PCNet accurately detects lithops—succulent plants that mimic stones—supporting biodiversity conservation and targeted weed control [19].

Another notable transformer-based approach is the High Resolution Iterative Feedback Network (HitNet) [20], which addresses COD by preserving and

refining high-resolution details often lost during down-sampling. HitNet's architecture includes three main modules: a Transformer-based Feature Extraction module utilizing a Pyramid Vision Transformer backbone [21] for efficient multi-scale feature extraction; a Multi-Resolution Iterative Refinement module that recursively enhances low-resolution features using high-resolution information via a global feedback mechanism; and an Iteration Feature Feedback (IFF) module that ensures progressive improvements in segmentation accuracy across iterations. Additionally, HitNet [20] incorporates a graph fusion module to integrate multi-scale features for precise boundary localization. An innovative application extends HitNet's capabilities by converting salient objects into camouflaged objects through cross-domain learning, thereby enriching the diversity of training data.

In contrast to transformer-based models like PCNet and HitNet, CNN-based methods such as Edge-Aware Mirror Network (EAMNet) and Boundary-Guided Network (BGNet) focus on explicit edge and boundary refinement for COD. EAMNet employs a dual-branch architecture featuring (1) a segmentation-induced edge aggregation module and (2) an edge-induced integrity aggregation module, which mutually refine each other through cross-branch guidance. The framework also includes a guided residual channel attention module that combines residual connections with gated convolution to enhance the extraction of structural details from low-level features. Similarly, BGNet [22] uses a Res2Net-50 backbone [23] to extract multi-level features from input images. It introduces three key modules: the Edge-Aware Module (EAM), Edge-Guidance Feature Module (EFM), and Context Aggregation Module (CAM). The EAM integrates low-level local edge details with high-level global location information to extract object-related edge semantics under explicit boundary supervision. EFM injects these edge cues into multi-level backbone features using local channel attention, enhancing feature representation with object structure information. CAM progressively aggregates multi-level fused features through cross-scale interactions using expanded convolutions to strengthen contextual semantics for accurate camouflaged object detection. The model employs multi-task super-

vision, combining mask and edge losses to improve boundary localization and object structure preservation. BGNet achieves strong performance on multiple benchmarks, delivering precise detection with fine and complete object boundaries, while maintaining balanced model complexity and inference speed. Its design mitigates the effects of edge disruption and background ambiguity in COD.

Transformer-based and CNN-based models thus offer complementary strategies for addressing COD through advanced feature fusion and boundary refinement. These methods improve detection accuracy by preserving high-resolution details and refining object boundaries in complex environments, making them especially valuable for agricultural applications. Most of the approaches reviewed in this section are evaluated in the present work to identify those best suited for the specific challenge of camouflaged weed detection.

3. Proposed Methodology

The proposed methodology introduces a comprehensive and systematic framework for the early detection of camouflaged weeds in agricultural environments, leveraging both visible (RGB) and near-infrared (NIR) imaging. The framework is composed of five main stages: (1) dataset acquisition, (2) preprocessing, (3) RGB-NIR image fusion, (4) training COD techniques, and (5) evaluation of multiple metrics. Figure 1 shows the overall pipeline of the proposed methodology. Each stage is described in detail below.

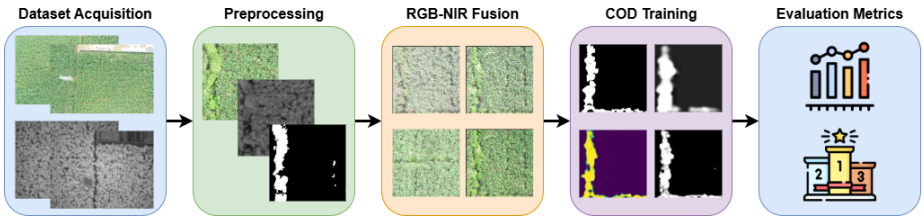

Figure 1: Overall pipeline of the proposed methodology.

3.1. Dataset

A high-quality and representative dataset is a key component for developing and validating automated weed detection and segmentation methods. To support this objective, data are collected in real agricultural field conditions to capture the variability typically encountered in operational scenarios, including changes in plant growth stages, weed density, soil appearance, and illumination.

Aerial surveys are conducted using an unmanned aerial vehicle (UAV) equipped with a multispectral sensor capable of acquiring both visible (RGB) and near-infrared (NIR) imagery. Flight plans are designed to ensure consistent ground sampling and sufficient image overlap for reliable mosaicking and georeferenced orthomosaic generation. Data acquisition is scheduled to promote stable lighting conditions and reduce shadow artifacts, improving the quality and consistency of the captured imagery across the surveyed area.

Table 1 summarizes the main real-world agricultural datasets with weed-segmentation annotations and positions our Weeds-Banana dataset within the current landscape. Carrot-Weed [24], one of the earliest works in this area (2015), focuses on carrot crops with 60 RGB images at a resolution of 1296×966. Later, CoFly-WeedDB [25] (2022) extends the domain to cotton crops, also in RGB modality, with 201 images of 1280×720. More recently, WeedsGalore [26] (2024) introduces multispectral (MSI) imagery for maize, comprising 156 samples at 600×600, while Potato-Weed [27] (2025) covers potato crops with 150 RGB images at 640×640. In this context, our Weeds-Banana dataset (2025) stands out by targeting banana crops—previously unexplored in this line of work—using multispectral modality and providing the largest number of images (272) with higher resolution (1024×1024), thus offering a richer and more detailed basis for developing and evaluating weed segmentation methods under real field conditions.

3.2. Preprocessing

The raw aerial imagery is first processed using standard photogrammetric and remote-sensing workflows to obtain analysis-ready products. This stage typ-

Dataset	Year	Crop	Modality	# images	Resolution
Carrot-Weed [24]	2015	Carrot	RGB	60	1296 × 966
CoFly-WeedDB [25]	2022	Cotton	RGB	201	1280 × 720
WeedsGalore [26]	2025	Maize	MSI	156	600 × 600
Potato-Weed [27]	2025	Potato	RGB	150	640 × 640
Weeds-Banana (Ours)	2025	Banana	MSI	272	1024 x 1024

Table 1: Comparison of available real-world agricultural datasets for weed segmentation.

ically includes radiometric normalization, geometric correction and alignment, and the generation of orthomosaics for each spectral band. The resulting orthomosaics are co-registered to ensure pixel-level correspondence between modalities, enabling consistent multi-band analysis. Weed presence is then delineated on the orthomosaic imagery through manual annotation using a dedicated labeling tool, producing segmentation masks aligned with the imagery.

To make the data suitable for deep learning training and evaluation, the large orthomosaics and their corresponding masks are converted into fixed-size samples. An automated preprocessing routine extracts image patches using a sliding-window strategy with configurable patch size and stride, producing a set of (multi-modal image, mask) sub-images with optional overlap. Patches are stored in an organized structure by modality, yielding a standardized and manageable dataset format. This patch-based preprocessing facilitates efficient batch loading during training, improves coverage of spatial variability, and increases the number of usable training samples without altering the underlying field content.

3.3. RGB-NIR Fusion Techniques

In agricultural scenes, RGB and NIR capture different physical drivers of vegetation appearance: visible bands are largely governed by pigment absorption, whereas NIR is dominated by scattering processes and canopy structure effects. This complementarity motivates multispectral fusion as a preprocessing

step to form a more informative representation before training segmentation/detection models [28, 29, 30].

To exploit the complementary information provided by RGB and NIR bands, a comprehensive evaluation of RGB-NIR image fusion techniques is performed. The fusion process is implemented using the VIFB framework proposed by [31], which supports a wide range of SOTA fusion algorithms. Fourteen fusion methods are selected for evaluation, representing diverse algorithmic families:

- Edge-preserving and filtering-based methods: Cross-Bilateral Filter (CBF) [32], Guided Filter-based Context Enhancement (GFCE) [33], Hybrid Multiscale Decomposition with Guided Filter (HMSD-GF) [33], Hybrid_MSD [34], Anisotropic Diffusion Fusion (ADF) [35].
- Decomposition-based methods: Latent Low-Rank Representation (LatLRR) [17], Multi-scale Guided Filter Fusion (MGFF) [36], Multi-resolution Singular Value Decomposition (MSVD) [37], Gradient Transfer Fusion (GTF) [38].
- Saliency and frequency-based methods: Two-scale Image Fusion (TIF) [39], Fourth Order Partial Differential Equation Fusion (FPDE) [40], Visual Saliency Map Weighted Least Squares (VSMWLS) [41].
- Deep learning-based methods: Convolutional Neural Network (CNN) fusion [42], Infrared Feature Extraction and Visual Information Preservation (IFEVIP) [43].

Each fusion method is applied to the registered RGB and NIR image pairs to generate fused images that combine spatial detail from the visible spectrum with spectral information from the NIR band.

3.4. COD Techniques

The fused images, along with the original RGB images, are used to train and evaluate nine state-of-the-art deep learning models for COD. These models are chosen to represent a variety of architectural paradigms, including convolutional

neural networks (CNNs), transformer-based networks, and hybrid approaches. Table 2 shows the techniques to be used and the main characteristics. The evaluated models are:

- BASNet [44]: A boundary-aware salient object detection network using a ResNet-34 backbone.
- SINet-v2 [45]: A Res2Net-50-based model designed for concealed object detection.
- BGNet [22]: A boundary-guided network with a Res2Net-50 backbone, optimized for precise edge detection.
- C^2F-Net [46]: A context-aware cross-level fusion network using Res2Net-50.
- OCENet [47]: A ResNet-50-based model that incorporates aleatoric uncertainty modeling.
- EAMNet [48]: An edge-aware mirror network with a Res2Net-50 backbone.
- DGNet [49]: A deep gradient learning network using EfficientNet.
- HitNet [20]: A high-resolution iterative feedback network based on PVTv2.
- PCNet [19]: A camouflage plant detection network using PVTv2, specifically designed for agricultural applications.

Each model is trained and validated on both the original RGB dataset and the three best-performing fusion datasets (CBF, HMSD-GF, LatLRR). Training is performed using standard data augmentation techniques to improve generalization, and hyperparameters are optimized for each model to ensure fair comparison. The evaluation is conducted on a held-out test set with manually annotated ground truth masks.

Table 2: Comparison between different characteristics of the camouflage techniques used.

Technique	**Source**	**Source Type**	**Year**	**Image Size (px)**	**Backbone**	**#Par. (M)**
BASNet [44]	CVPR	Conference	2019	256×256	ResNet-34 [50]	87.06
SINet-v2 [45]	TPAMI	Journal	2021	352×352	Res2Net-50 [23]	24.93
BGNet [22]	IJCAI	Conference	2022	416×416	Res2Net-50 [23]	77.80
C^2F-Net [46]	TCSVT	Conference	2022	352×352	Res2Net-50 [23]	26.36
OCENet [47]	WACV	Conference	2022	352×352	ResNet-50 [50]	58.17
EAMNet [48]	ICME	Conference	2023	384×384	Res2Net-50 [23]	30.51
DGNet [49]	MIR	Journal	2023	352×352	EfficientNet [51]	8.30
HitNet [20]	AAAI	Conference	2023	352×352	PVTv2 [52]	25.73
PCNet [19]	arXiv	-	2024	352×352	PVTv2 [52]	27.66

3.5. Evaluation Metrics

For RGB-NIR fusion quality evaluation of the fused images are quantitatively assessed using thirteen metrics, each capturing a different aspect of fusion performance. A point-based ranking system is used to identify the top-performing fusion methods, which are then selected for subsequent COD experiments. Table 3 shows each of the metrics and the category to which they belong.

On the other hand, for COD evaluation performance and to provide a rigorous assessment, five widely adopted metrics are used:

- S-measure (S_α) [64]: Measures the structural similarity between the predicted segmentation and the ground truth, capturing both region-aware and object-aware similarities.
- Weighted F-measure (F_β^w) [65]: An extension of the traditional F-measure that incorporates spatial weights, emphasizing boundary accuracy and the spatial relevance of detected pixels.
- Mean Absolute Error (M) [66]: Computes the average pixel-wise error between the normalized prediction and the ground truth mask.

Category	Metric	Description	+/-
Information theory-based	CE	Cross entropy [53]	-
	EN	Entropy [54]	+
	MI	Mutual information [55]	+
	$PSNR$	Peak signal-to-noise ration [56]	+
Structural similarity-based	$SSIM$	Structural similarity index measure [57]	+
	$RMSE$	Root mean squared error [56]	-
Image feature-based	AG	Average gradient [58]	+
	EI	Edge intensity [59]	+
	SD	Standard deviation [60]	+
	SF	Spatial frequency [61]	+
	$Q^{AB/F}$	Gradient-based fusion performance [61]	+
Human perception inspired	Q_{CB}	Chen-Blum metric [62]	+
	Q_{CV}	Chen-Varshney metric [63]	-

Table 3: Evaluation metrics used with SOTA fusion techniques. '+' means that a high value is a good performance, while '-' means that a small value denotes a good performance.

- E-measure (E_ϕ) [67]: Simultaneously evaluates global and local accuracy based on human visual perception.
- F-measure (F_β) [68]: Provides a balanced measure of precision and recall.

To capture different aspects of model performance, adaptive, mean, and maximum variants of both F-measure and E-measure (F_β^{adp}, F_β^{mean}, F_β^{max}, E_ϕ^{adp}, E_ϕ^{mean}, E_ϕ^{max}) are also computed. This comprehensive evaluation framework ensures that both the accuracy and the robustness of each model are thoroughly assessed across multiple datasets and fusion strategies.

3.6. Training details

This section summarizes the training configuration used across all COD techniques. Table 4 reports the optimizer, learning rate, batch size, number of epochs, scheduler, and loss functions employed per model. All models are



Figure 2: *(left)* DJI Mavic 2 Pro UAV. *(right)* MicaSense RedEdge-M multispectral camera.

trained under the same consistent splits and metrics described in Section 3.5 are used to enable a fair comparison.

4. Case Studies

This section presents the experimental results from the two case studies evaluated with the proposed methodology. The first case study presented uses its own dataset (Weeds-Banana) while the second case study uses a state-of-the-art dataset (WeedsGalore [26]). For the performance evaluation of this work, the metrics described in Sec. 3.5 are used. All experiments are conducted on a workstation configured with an Intel Core i9 processor and an NVIDIA Titan XP GPU. The framework proposed by [31] is used for the fusion techniques. This code has been implemented using MATLAB and is available at the following URL `https://github.com/xingchenzhang/VIFB`. For the nine COD techniques, all implementations are written in Python by the authors.

4.1. Datasets

Weed-Banana. For the first case study, a custom dataset created to validate the proposed methodology is used, named Weed-Banana. Dataset acquisition is systematically conducted in the Milagro canton, located in Guayas, Ecuador, a region characterized by a tropical climate with average temperatures ranging between 25°C and 27°C and annual precipitation between 1,100

Table 4: 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
BASNet [44]	Adam	1e-3	8	1000	ReduceLROnPlateau	BCE + SSIM + IOU (multi-stage fusion)
SINet-v2 [45]	Adam	1e-4	16	150	Custom (Adjust LR)	Structure loss (weighted BCE + weighted IOU)
BGNet [22]	Adam	1e-4	12	100	Custom (Poly LR)	Structure loss (weighted BCE + weighted IOU) + Dice loss (edge)
C^2F-Net [46]	AdaXW	1e-4	32	50	Custom (Poly LR)	Structure loss (weighted BCE + weighted IOU)
OCENet [47]	Adam	1e-5	4	50	StepLR	Uncertainty aware structure loss (weighted BCE + weighted IOU)
EAMNet [48]	AdamW	5e-5	16	150	Custom (Adjust LR)	Hybrid loss (weighted BCE + weighted IOU) + Edge loss (edge)
DGNet [49]	AdamW	5e-5	16	150	CosineAnnealingLR	Hybrid loss (weighted BCE + weighted IOU) + MSE loss (grad)
HitNet [20]	AdamW	1e-4	8	150	Custom (Adjust LR)	Structure loss (weighted BCE + weighted IOU)
PCNet [19]	AdamW	1e-4	8	150	Custom (Adjust LR)	Structure loss (weighted BCE + weighted IOU)

and 1,800 mm. The study focused on an experimental plot of approximately 4.3 hectares within a banana plantation, providing a representative agricultural environment for research on the detection of camouflaged weeds. Data collection utilized a DJI Mavic 2 Pro UAV equipped with MicaSense RedEdge-M multispectral camera, enabling simultaneous capture of visible and near-infrared spectral information (See Fig. 2). Flight operations are programmed with 85% lateral and frontal overlap at an altitude of 120 meters above the canopy, ensuring comprehensive coverage and sufficient image overlap for accurate mosaic generation. Image capture is conducted in April 2024 between 11:00 and 12:00 to minimize shadow effects and optimize lighting conditions for both RGB and

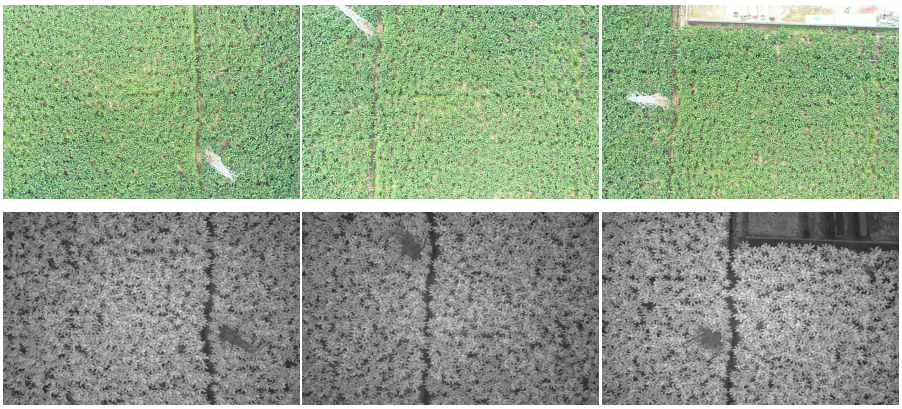

Figure 3: Example images used to build the RGB and NIR orthomosaics.

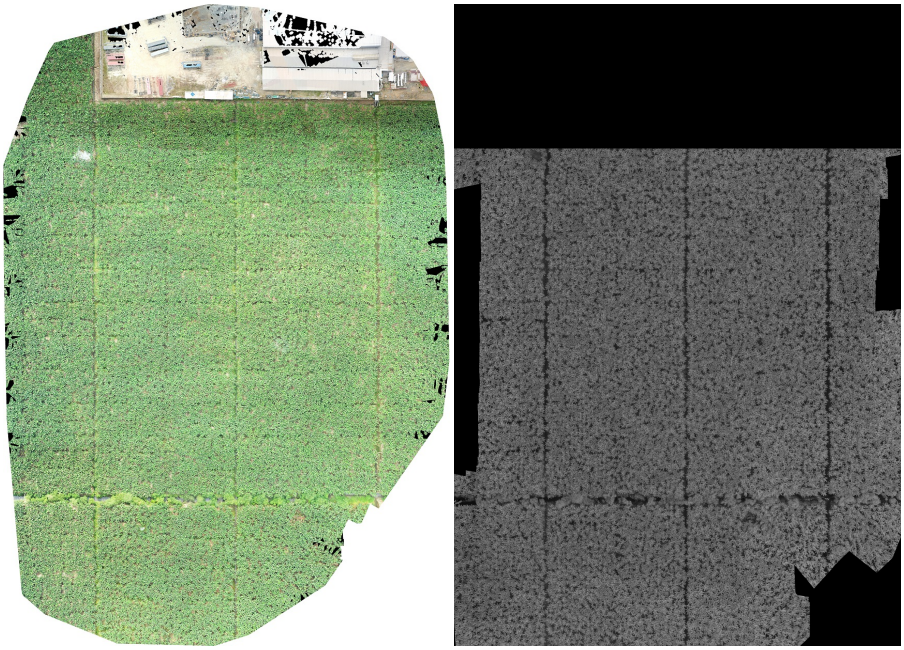

Figure 4: RGB and NIR orthomosaics registered using QGIS.

NIR imaging.

WeedsGalore [26]. For the second case study, a state-of-the-art dataset, WeedsGalore, is used. This dataset is systematically acquired in an experimental maize field near Marquardt, Potsdam, Germany (approx. 1,840 m^2),

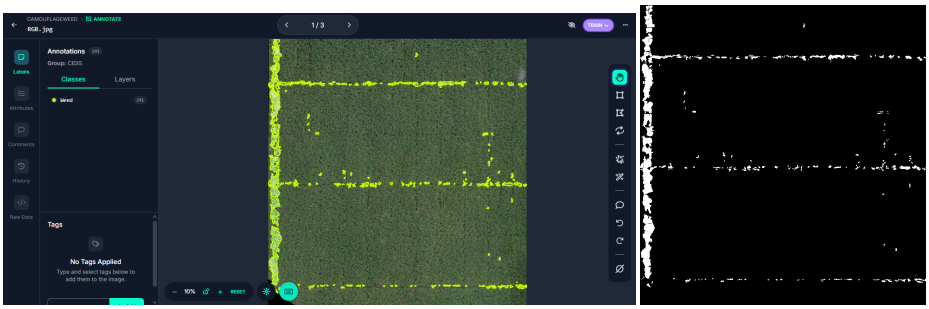

Figure 5: (left) View of the Roboflow annotation tool for the RGB orthomosaic image. (right) Labeled orthomosaic binary mask using Roboflow.

representative of conventional row-crop systems in Central Europe. Maize (Zea mays L.) is sown on 9 May 2023 under standard management practices, while naturally emerging weeds are allowed to develop, resulting in a dense and diverse weed community typical of regional maize production. Data collection is carried out using a DJI Phantom 4 Multispectral UAV, which provides RGB imagery together with five monochromatic multispectral bands (blue, green, red, red-edge, and near-infrared). Four flight campaigns are conducted on 25 May, 30 May, 6 June, and 15 June 2023, covering phenological stages from early maize emergence with sparse vegetation to almost complete ground cover by crops and weeds. Flights are performed at 5 m above ground level in Hover&Capture mode with 70% side and 60% front overlap, yielding a ground sampling distance of approximately 2.5 mm/pixel and enabling individual plant discrimination. From roughly 1,150 raw images per campaign, 48 geo-referenced locations are selected without spatial overlap, and for each location, a central 600×600-pixel window per band is extracted to form the annotated tiles used for semantic segmentation.

4.2. Preprocessing

Weeds-Banana. For the own Weeds-Banana dataset, the acquired images underwent systematic processing using PIX4Dmapper[1] software, involving ra-

[1] https://www.pix4d.com/

diometric calibration to ensure spectral accuracy, photographic alignment for geometric consistency, image geometry construction for spatial precision, and orthophotography generation to produce high-quality RGB and multispectral mosaics suitable for subsequent fusion and analysis processes. Figure 3 shows example images used to build the RGB and NIR orthomosaics. On the other hand, Fig. 4 shows the orthomosaic generated using RGB (*left*) and NIR (*right*) images.

In our Weeds-Banana dataset, the images have been registered using the QGIS[2] platform, a crucial step that will be useful in the subsequent stages. The resulting registered RGB and NIR image sizes are 5571×5855 pixels. After registering the RGB-NIR images, they are labeled using the RGB image as a reference to identify areas with weeds. To carry out the labeling task, the Roboflow[3] tool is used. Figure 5 (left) shows the labeling of the RGB orthomosaic image; the result of the annotation process is a binary mask of 5571×5855 pixels (see Fig. 5 (right)). The registered RGB-NIR images and the binary mask resulting from the annotation process will be available in the GitHub repository. Finally, after completing the annotation stage of the RGB orthomosaic image, a Python script called *"split_dataset.py"* (available in GitHub repository) is applied to generate 272 patches. The distribution of the images used to train and test the different models is training=218, validation=27, and testing=27. Figure 6 (left) shows examples of the patches created from the registered RGB-NIR orthomosaic.

WeedsGalore [26]. For the WeedsGalore [26] dataset, several preprocessing steps are applied to transform the raw UAV imagery into consistent tiles suitable for semantic and instance segmentation. For each of the four flight campaigns, the images captured by the DJI Phantom 4 Multispectral are processed at the level of individual frames rather than full-field orthomosaics. The five monochromatic multispectral bands (blue, green, red, red-edge, and near-

[2]https://qgis.org/
[3]https://app.roboflow.com/

infrared) are geometrically aligned and stacked, and, for every selected frame, a central 600×600-pixel window is cropped in all bands. From approximately 1,150 raw images per campaign, 48 geo-referenced locations are defined across the field, and the corresponding tiles are extracted such that there is no spatial overlap between annotated images, either within a date or across dates. This procedure results in 156 multispectral tiles at a ground sampling distance of about 2.5 mm/pixel, allowing individual crops and weeds to be clearly resolved. For each tile, paired semantic and instance annotations are generated using consistent class indices for maize and four weed categories, and the final dataset is organized into spatially disjoint training, validation, and test splits (104 / 26 / 26 tiles) using the official split files provided with the dataset. Figure 6 (right) shows patch examples used for the WeedGalore dataset.

4.3. Fusion Techniques Results

To establish the best Fusion techniques, a point system is defined based on ranking position (i.e., $1st = 3points$, $2nd = 2points$, $3rd = 1point$). The analysis is based on multiple complementary metrics that evaluate different aspects of fusion quality. Among them, some seek lower values, such as CE, $RMSE$, and Q_{CV}, while others, such as EN, MI, $PSNR$, AG, FI, $Q^{AB/F}$, SD, SF, $SSIM$, and Q_{CB}, are improved with higher values. This comparison is essential for selecting the most appropriate fusion method according to the particular needs for the specific application, and in particular in the context of the COD task. Also, Table 3 shows for each metric a description and category to which it belongs.

Weeds-Banana. Table 5 presents a comprehensive evaluation of SOTA fusion techniques using our Weeds-Banana dataset. CBF [32] stands out as the most effective method with 15 total points. HMSD-GF [33] follows it, and LatLRR [17], both with 12 points, are the top-3 of the best techniques, while the remaining methods obtain progressively lower scores, demonstrating a clear hierarchy in the performance of these techniques. CBF demonstrates superior performance across most metrics, justifying its dominant position in the overall

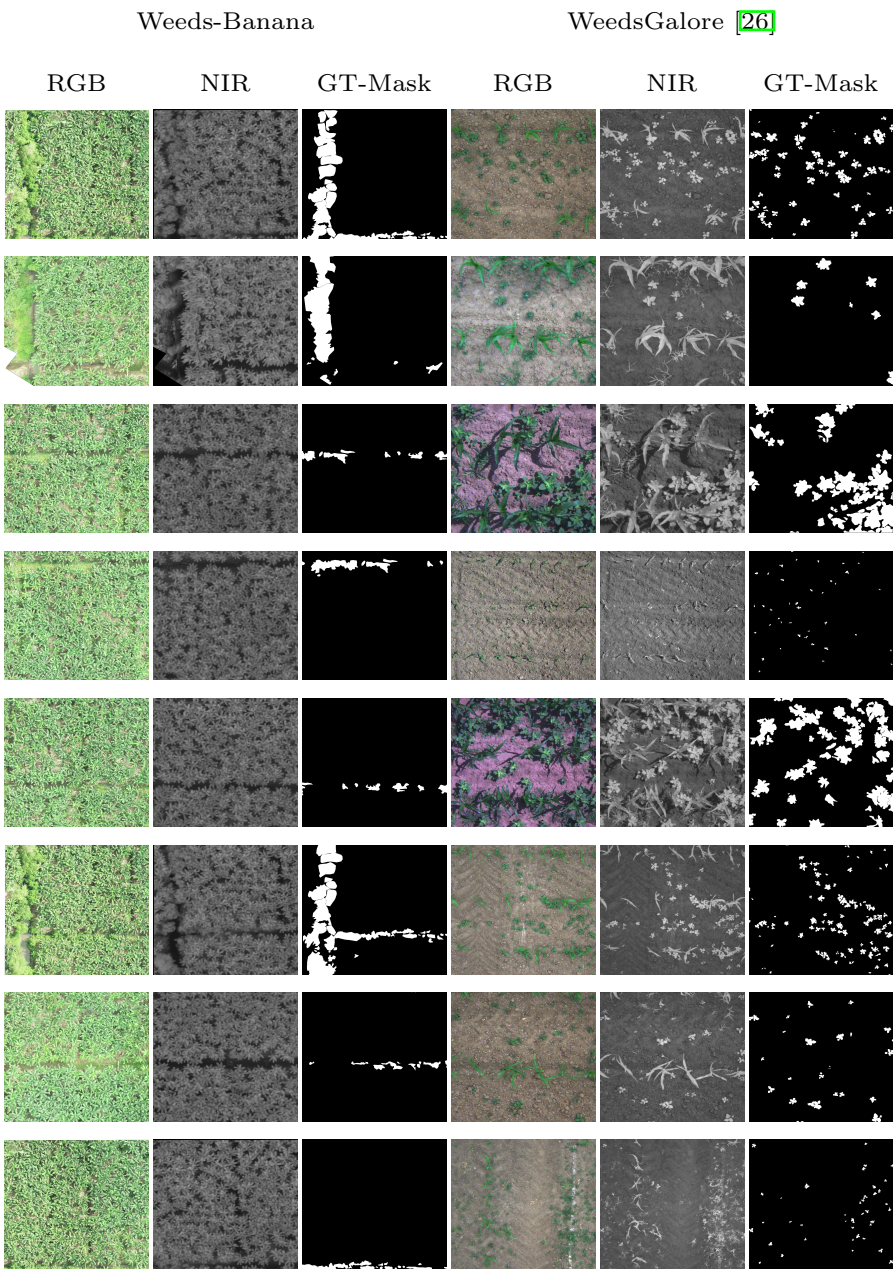

Figure 6: Example images of the patches created from the registered orthomosaic using our Weeds-Banana and the WeedsGalore [26] datasets.

ranking, although other methods, such as HMSD-GF, LatLRR, and GFCE, also exhibit significant strengths in specific metrics, offering viable alternatives depending on the priority requirements in each use case.

Technique	Ptos.	$CE \downarrow$	$EN \uparrow$	$MI \uparrow$	$PSNR \uparrow$	$AG \uparrow$	$EI \uparrow$	$Q^{AB/F} \uparrow$	$SD \uparrow$	$SF \uparrow$	$RMSE \downarrow$	$SSIM \uparrow$	$Q_{CB} \uparrow$	$Q_{CV} \downarrow$
CBF [32]	15	1.51	7.05	3.78	62.3390	14.07	141.00	0.8100	42.98	32.05	0.0391	1.0875	0.7248	152.00
HMSD-GF [33]	12	1.34	7.43	2.12	61.8620	15.41	154.45	0.7800	46.65	34.90	0.0432	1.0820	0.6320	148.50
LatLRR [17]	12	1.73	6.72	1.18	59.5820	24.58	246.15	0.4310	66.65	58.89	0.0717	0.8540	0.5706	223.47
GFCE [33]	11	1.55	7.62	1.06	60.2425	17.07	171.42	0.6191	54.80	38.83	0.0615	1.0051	0.5076	404.01
FPDE [40]	8	1.20	6.76	1.07	62.3835	9.74	94.39	0.5778	26.99	21.74	0.0387	1.0965	0.4654	339.28
Hybrid_MSD [34]	6	1.29	7.37	2.11	62.2400	14.35	143.94	0.7753	43.11	32.55	0.0398	1.0972	0.6419	152.11
MGFF [36]	3	1.13	7.46	1.13	61.9365	13.39	134.44	0.6724	44.00	30.56	0.0427	1.0940	0.5383	340.82
TIF [39]	3	1.05	7.33	1.35	62.0415	12.49	126.78	0.6694	40.32	28.44	0.0417	1.0730	0.5776	211.54
ADF [35]	2	1.24	6.89	1.07	62.3210	11.37	109.85	0.6507	29.45	26.00	0.0392	1.0862	0.4730	344.84
CNN [42]	2	1.32	7.40	1.68	61.9625	14.13	141.58	0.8043	43.76	32.21	0.0423	1.0939	0.5564	189.93
IFEVIP [43]	2	2.10	6.90	2.82	60.8045	10.26	103.25	0.5318	35.05	24.12	0.0542	1.0922	0.4802	378.28
VSMWLS [41]	2	1.09	7.20	1.21	62.1130	12.44	121.03	0.6252	36.55	29.20	0.0410	1.0795	0.4845	443.47
MSVD [37]	1	1.60	6.81	0.92	62.3170	9.59	89.80	0.4404	27.66	25.49	0.0392	1.0653	0.3997	426.00
GTF [38]	0	1.44	7.02	0.64	61.5470	11.90	115.70	0.6189	31.61	28.00	0.0468	0.9984	0.4687	836.72

Table 5: Metric evaluation results of SOTA Fusion techniques using our Weeds-Banana dataset. The best three performing results are highlighted using color: First, Second, and Third respectively.

WeedsGalore [26]. On the other hand, Table 6 presents a comprehensive evaluation of SOTA fusion techniques using the WeedsGalore [26] dataset. Contrary to the results presented in Table 5, the ranking position of each technique is different, indicating that the performance of each fusion technique depends on the characteristics of each dataset. LatLRR [17] had the best overall performance with a total of 18 points. GFCE [33] followed at a considerable distance with a total of 10 points, while FPDE [40] with 9 points completed the top three best fusion techniques.

4.4. SOTA COD Techniques Results

Weeds-Banana. As a result of the evaluation using our Weeds-Banana dataset, Table 7, Fig. 7 and Fig. 8 show the results obtained. Table 7 presents a comprehensive quantitative evaluation of nine SOTA COD techniques (i.e., BASNet [44], SINet-v2 [45], BGNet [22], C^2F-Net [46], OCENet [47], EAMNet [48], DGNet [49], HitNet [20], PCNet [19]), evaluated across 1 RGB dataset and 3 datasets generated with fusion techniques (i.e., CBF, HMSD-GF, and

Technique	Ptos.	$CE \downarrow$	$EN \uparrow$	$MI \uparrow$	$PSNR \uparrow$	$AG \uparrow$	$EI \uparrow$	$Q^{AB/F} \uparrow$	$SD \uparrow$	$SF \uparrow$	$RMSE \downarrow$	$SSIM \uparrow$	$Q_{CB} \uparrow$	$Q_{CV} \downarrow$
LatLRR [17]	18	1.52	7.44	1.86	58.2501	21.86	215.59	0.2501	52.30	48.32	0.0976	1.1310	0.6608	341.23
GFCE [33]	10	0.57	7.17	2.36	61.7139	13.68	133.48	0.5953	37.43	31.27	0.0459	1.5265	0.7006	144.28
FPDE [40]	9	0.32	6.43	2.52	64.7010	7.83	76.87	0.6511	22.31	17.52	0.0239	1.7425	0.7056	184.99
IFEVIP [43]	8	1.15	7.01	2.83	59.9904	11.64	115.48	0.6109	35.65	26.03	0.0662	1.5906	0.7194	209.53
ADF [35]	7	0.30	6.45	2.57	64.6922	8.04	78.61	0.6628	22.53	18.15	0.0239	1.7361	0.7061	189.01
CNN [42]	6	0.30	6.77	3.26	63.6096	10.16	97.71	0.6990	28.85	24.28	0.0307	1.6275	0.7072	118.90
TIF [39]	5	0.22	6.65	2.47	64.1743	9.03	88.90	0.6458	26.37	21.30	0.0266	1.6825	0.7185	135.33
MGFF [36]	4	0.28	7.01	2.20	63.3655	11.98	117.71	0.5874	33.06	27.32	0.0317	1.6027	0.6746	174.58
CBF [32]	3	0.32	6.60	2.63	64.2969	9.08	88.45	0.6788	25.09	21.18	0.0264	1.6808	0.6968	165.13
MSVD [37]	3	0.30	6.51	2.07	64.2872	8.89	83.54	0.5806	23.55	22.20	0.0258	1.6478	0.6772	199.47
VSMWLS [41]	3	0.21	6.72	2.52	64.0546	10.52	101.06	0.6456	27.40	24.91	0.0271	1.6661	0.6695	145.00
GTF [38]	2	0.29	6.63	2.46	63.6851	9.70	92.68	0.6859	26.65	23.42	0.0300	1.6232	0.6915	160.99
Hybrid_MSD [34]	2	0.26	6.81	2.64	63.7672	10.96	106.58	0.6558	29.14	25.86	0.0292	1.6580	0.7059	143.70
HMSD-GF [33]	1	0.38	6.90	2.56	62.8050	11.30	110.06	0.6517	31.00	26.12	0.0356	1.6396	0.7125	124.55

Table 6: Metric evaluation results of SOTA Fusion techniques using the WeedsGalore [26] dataset. The best three performing results are highlighted using color: First, Second, and Third respectively.

LatLRR). The results show that PCNet [19] consistently stands out, obtaining the highest scores on F_β^w, M, E_ϕ^{adp}, E_ϕ^{mean}, E_ϕ^{max}, F_β^{adp}, and F_β^{mean} metrics. Furthermore, BASNet [44] shows the best result for the S_α metric, just as BGNet shows the best result for the F_β^{max} metric.

On the other hand, for our Weeds-Banana dataset, Fig. 7 shows qualitative results using two sample images from the test set evaluated for the top-3 techniques. It can be observed that RGB-NIR-based techniques outperform RGB-only techniques in small regions. Furthermore, corroborating the quantitative results, the PCNet [19] technique presents consistent results with well-defined segmented areas of weeds relative to the GT. In a more specific analysis focusing on the PCNet [19] COD technique Fig. 8 shows a comparison of the mask prediction results using RGB and fusion techniques where it can be observed that using RGB-NIR fusion techniques, masks with better contour definition and fewer false negatives are obtained, in this way the qualitative results presented in Table 7 can be confirmed.

WeedsGalore [26]. As a result of the evaluation using the WeedsGalore [26] dataset, Table 8, Fig. 9, and Fig. 10 summarize the quantitative and qualitative performance of the nine SOTA COD techniques across the RGB baseline and the three best-ranked fusion configurations for this dataset (i.e., FPDE,

GFCE, and LatLRR, as identified in Sec. 4.3). Overall, the results confirm that incorporating NIR information through RGB–NIR fusion benefits camouflaged weed segmentation also in this second case study, although the magnitude of the improvement is more method- and metric-dependent than in Weeds-Banana. In particular, the best-performing techniques on WeedsGalore are BASNet, PC-Net, and OCENet, which are therefore selected for qualitative inspection in Fig. 9.

From the quantitative results using the WeedsGalore dataset in Table 8, BASNet provides the most consistent trade-off across the evaluated metrics, achieving top-ranked scores when coupled with fusion inputs—especially under GFCE/FPDE, where improvements are observed in both region accuracy and boundary consistency. BASNet is highly competitive and yields very strong F-measure values on this dataset and remains robust and produces stable predictions, particularly in scenes where weeds are sparse and appear as small, low-contrast structures. PCNet and OCENet show good results but are far from the performance of BASNet. The qualitative examples in Fig. 9 show that fusion-based inputs (FPDE, GFCE, LatLRR) reduce miss-segmentation (false negatives) in difficult low-contrast weed regions compared to RGB-only inputs, while also helping mitigate over-segmentation in visually ambiguous areas. Finally, Fig. 10 focuses on BASNet, illustrating that fusion configurations generally lead to cleaner masks with fewer fragmented detections and improved alignment with the ground truth, corroborating the quantitative trends observed in Table 8.

In conclusion and in a general way, following the proposed methodology, it is shown that the approach using the fusion of the RGB and NIR images shows better results than only RGB images. Upon reviewing each camouflage technique, it can be observed that the datasets generated using the RGB-NIR fusion techniques exhibit better performance in all metrics (the first place for each COD technique is marked with an underline). In the case of our Weeds-Banana dataset, the combination of PCNet-LatLRR COD-Fusion technique presents the best overall result. While for WeedsGalore, the most competitive trade-offs are

obtained with BASNet-FPDE, where this combination provides the strongest overall performance.

Table 7: Metric evaluation results for each COD technique on our Weeds-Banana dataset, reported for the RGB baseline and the three best-ranked RGB–NIR fusion inputs (CBF, HMSD-GF, and LatLRR). Results are presented using the metric notation defined in Sec. 3.5. "↑ / ↓" indicates that larger or smaller is better. The best three performing results are highlighted using color: First, Second, and Third respectively.

T.	Method	$S_\alpha \uparrow$	$F_\beta^w \uparrow$	$M \downarrow$	$E_\phi^{adp} \uparrow$	$E_\phi^{mean} \uparrow$	$E_\phi^{max} \uparrow$	$F_\beta^{adp} \uparrow$	$F_\beta^{mean} \uparrow$	$F_\beta^{max} \uparrow$
BASNet [44]	RGB	0.9562	0.5125	0.0043	0.9718	0.9828	0.9960	0.4825	0.5192	0.5371
	CBF	0.9529	0.5075	0.0045	0.9601	0.9833	0.9961	0.4701	0.5113	0.5342
	HMSD-GF	0.9599	0.5151	0.0042	0.9674	0.9843	0.9968	0.4714	0.5182	0.5416
	LatLRR	0.9339	0.4747	0.0078	0.9491	0.9698	0.9825	0.4439	0.4840	0.5033
SINet-v2 [45]	RGB	0.8761	0.3410	0.0111	0.8774	0.9491	0.9826	0.3189	0.3646	0.4002
	CBF	0.8809	0.3553	0.0105	0.8762	0.9463	0.9756	0.3221	0.3738	0.4179
	HMSD-GF	0.8857	0.3637	0.0103	0.8832	0.9455	0.9837	0.3304	0.3818	0.4266
	LatLRR	0.8852	0.3672	0.0100	0.8856	0.9437	0.9797	0.3364	0.3831	0.4292
BGNet [22]	RGB	0.7232	0.2584	0.1534	0.9716	0.8333	0.9933	0.4754	0.5019	0.5510
	CBF	0.7294	0.2566	0.1490	0.9724	0.8363	0.9934	0.4742	0.4964	0.5465
	HMSD-GF	0.7123	0.2693	0.1711	0.9725	0.8153	0.9967	0.4735	0.4996	0.5498
	LatLRR	0.7275	0.2586	0.1476	0.9466	0.8361	0.9963	0.4460	0.4951	0.5493
C^2F-Net [46]	RGB	0.6931	0.1359	0.1252	0.8462	0.8079	0.9837	0.3100	0.3598	0.4651
	CBF	0.7111	0.1755	0.1219	0.8894	0.8331	0.9927	0.3546	0.4060	0.4862
	HMSD-GF	0.7045	0.1598	0.1248	0.8901	0.8278	0.9894	0.3545	0.3985	0.4799
	LatLRR	0.6783	0.1657	0.1517	0.8830	0.8041	0.9896	0.3503	0.4048	0.4746
OCENet [47]	RGB	0.8937	0.4258	0.0079	0.9674	0.9638	0.9875	0.4238	0.4529	0.4643
	CBF	0.8895	0.4292	0.0077	0.9368	0.9615	0.9878	0.4315	0.4569	0.4686
	HMSD-GF	0.8838	0.4116	0.0081	0.9698	0.9394	0.9800	0.4259	0.4469	0.4605
	LatLRR	0.8867	0.4217	0.0077	0.9429	0.9479	0.9915	0.4398	0.4577	0.4708
EAMNet [48]	RGB	0.7475	0.1157	0.0645	0.7928	0.8579	0.9478	0.2744	0.2901	0.3522
	CBF	0.6868	0.0803	0.1103	0.7202	0.7699	0.9137	0.1970	0.1882	0.2533
	HMSD-GF	0.7467	0.1257	0.0527	0.7887	0.8522	0.9410	0.2382	0.2584	0.2998
	LatLRR	0.7719	0.1839	0.0388	0.7862	0.8620	0.9529	0.2385	0.2850	0.3436
DGNet [49]	RGB	0.8772	0.4094	0.0387	0.8876	0.9278	0.9862	0.3565	0.4219	0.4614
	CBF	0.8930	0.4231	0.0309	0.9017	0.9443	0.9889	0.3748	0.4363	0.4651
	HMSD-GF	0.9079	0.4160	0.0127	0.8759	0.9616	0.9912	0.3660	0.4299	0.4579
	LatLRR	0.8708	0.4173	0.0526	0.8971	0.9182	0.9901	0.3639	0.4285	0.4607
HitNet [20]	RGB	0.9165	0.4794	0.0159	0.9196	0.9671	0.9963	0.4381	0.4723	0.5085
	CBF	0.9263	0.4840	0.0074	0.9359	0.9748	0.9962	0.4499	0.4746	0.5142
	HMSD-GF	0.9345	0.5004	0.0084	0.9437	0.9822	0.9957	0.4744	**0.4958**	0.5169
	LatLRR	0.9282	0.4926	0.0123	0.9326	0.9758	0.9957	0.4662	0.4875	0.5161
PCNet [19]	RGB	0.9558	0.5294	0.0034	0.9716	0.9901	0.9971	0.4954	0.5245	0.5448
	CBF	0.9559	0.5248	0.0035	0.9696	0.9907	0.9972	0.4981	0.5242	0.5384
	HMSD-GF	0.9555	0.5274	0.0036	0.9707	0.9908	0.9971	0.5021	0.5272	0.5409
	LatLRR	0.9590	0.5278	0.0034	0.9790	0.9903	0.9967	0.5011	0.5272	0.5411

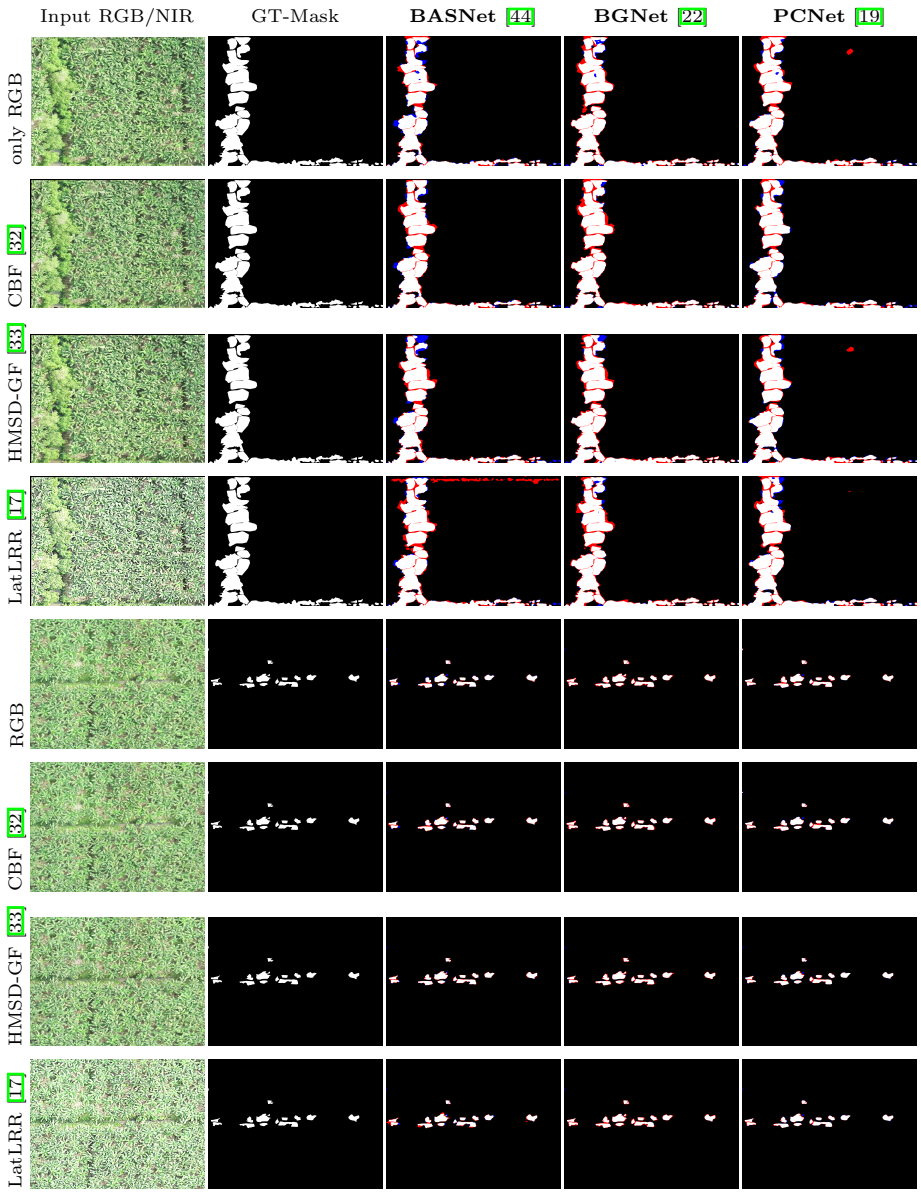

Figure 7: Prediction results for top-3 SOTA COD techniques using our Weeds-Banana dataset in two sample images. Successful matches between GT and predicted masks (white areas); False positive regions (red areas, over-segmentation); and false negative regions (blue areas, miss-segmentation).

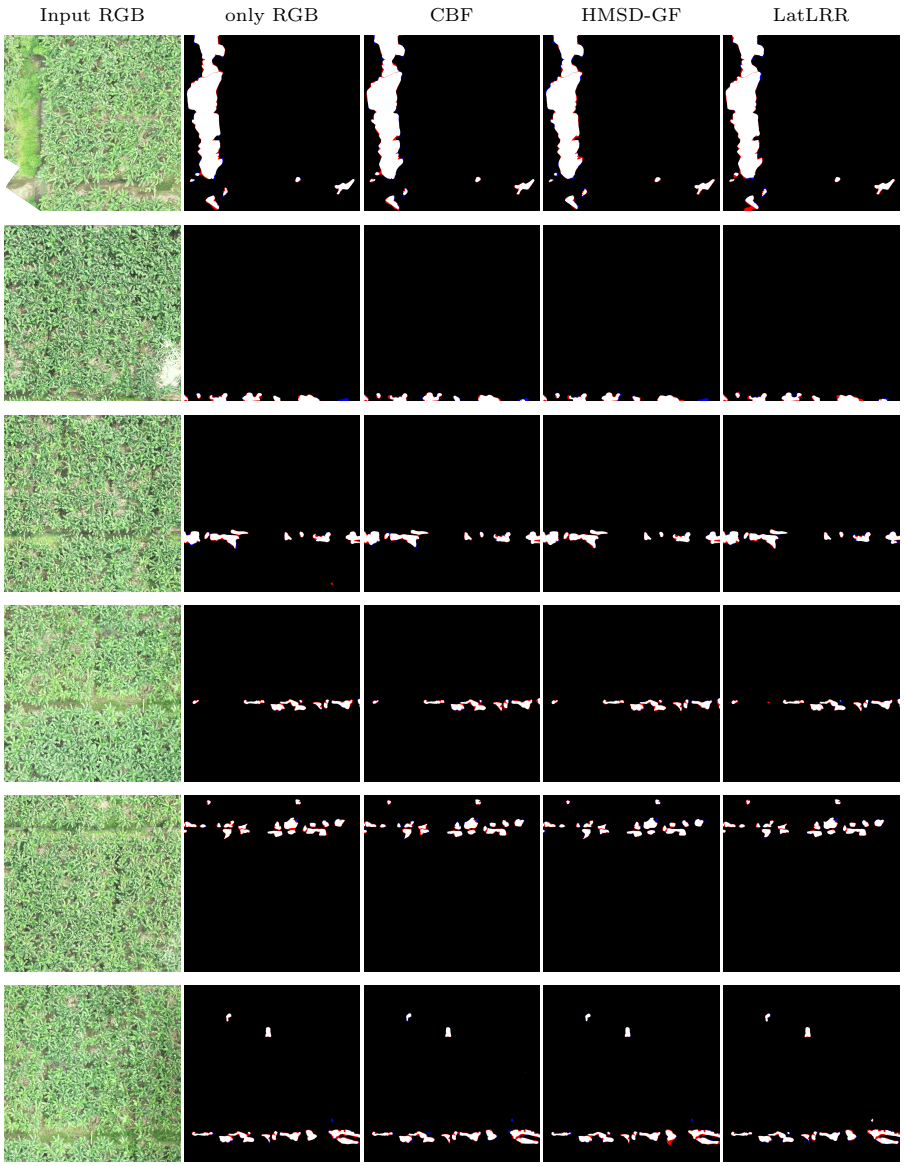

Figure 8: Prediction results for PCNet [19] the best overall technique using our Weeds-Banana dataset. Successful matches between GT and predicted masks (white areas); False positive regions (red areas, over-segmentation); and false negative regions (blue areas, miss-segmentation).

Table 8: Metric evaluation results for each COD technique on the WeedsGalore [26] dataset, reported for the RGB baseline and the three best-ranked RGB–NIR fusion inputs (FPDE, GFCE, and LatLRR). Results are presented using the metric notation defined in Sec. 3.5. "↑ / ↓" indicates that larger or smaller is better. The best three performing results are highlighted using color: First, Second, and Third respectively.

T.	Method	$S_\alpha \uparrow$	$F_\beta^w \uparrow$	$M \downarrow$	$E_\phi^{adp} \uparrow$	$E_\phi^{mean} \uparrow$	$E_\phi^{max} \uparrow$	$F_\beta^{adp} \uparrow$	$F_\beta^{mean} \uparrow$	$F_\beta^{max} \uparrow$
BASNet [44]	RGB	0.7533	0.4428	0.0150	0.7994	0.8385	0.9253	0.3991	0.4788	0.5150
	FPDE	0.7782	0.4548	0.0138	0.8804	0.8480	0.9070	0.4357	0.4880	0.5200
	GFCE	0.7696	0.4656	0.0145	0.8144	0.8533	0.9149	0.4207	0.5042	0.5474
	LatLRR	0.7541	0.4436	0.0138	0.8530	0.8304	0.8937	0.4003	0.4777	0.5122
SINet-v2 [45]	RGB	0.6350	0.2182	0.0316	0.5266	0.7242	0.8213	0.2015	0.2436	0.2654
	FPDE	0.6465	0.2472	0.0271	0.5476	0.7556	0.8414	0.2259	0.2699	0.3025
	GFCE	0.6433	0.2441	0.0280	0.5592	0.7359	0.7925	0.2307	0.2674	0.2908
	LatLRR	0.6435	0.2428	0.0274	0.5545	0.7229	0.7774	0.2230	0.2640	0.2890
BGNet [22]	RGB	0.5854	0.1478	0.1427	0.7178	0.7016	0.8954	0.3179	0.3383	0.4114
	FPDE	0.5931	0.1558	0.1312	0.7763	0.7115	0.8457	0.3516	0.3772	0.4518
	GFCE	0.5871	0.1442	0.1228	0.7471	0.7310	0.8986	0.3554	0.3939	0.4952
	LatLRR	0.5918	0.1479	0.1397	0.7476	0.7218	0.8788	0.3390	0.3703	0.4542
C^2F-Net [46]	RGB	0.5858	0.1361	0.1080	0.5482	0.6774	0.9141	0.1957	0.2722	0.3446
	FPDE	0.5736	0.1189	0.1323	0.5137	0.6255	0.8430	0.1842	0.2446	0.3217
	GFCE	0.5941	0.1332	0.0911	0.5293	0.6798	0.8927	0.2114	0.2699	0.3545
	LatLRR	0.5640	0.1171	0.1412	0.5162	0.6208	0.8602	0.1972	0.2294	0.2973
OCENet [47]	RGB	0.6543	0.2738	0.0286	0.6216	0.7718	0.8353	0.2277	0.3033	0.3336
	FPDE	0.6869	0.3114	0.0234	0.6097	0.8065	0.8761	0.2567	0.3395	0.3676
	GFCE	0.6929	0.3214	0.0225	0.6128	0.8362	0.9177	0.2603	0.3609	0.4007
	LatLRR	0.6755	0.2971	0.0216	0.6209	0.8002	0.8820	0.2499	0.3235	0.3553
EAMNet [48]	RGB	0.5547	0.0732	0.1025	0.5112	0.5907	0.7768	0.1540	0.1541	0.2101
	FPDE	0.5112	0.0871	0.1995	0.5067	0.5447	0.7012	0.1886	0.1854	0.2388
	GFCE	0.5548	0.0647	0.1088	0.5615	0.5911	0.7416	0.1959	0.1706	0.2381
	LatLRR	0.5454	0.0830	0.1249	0.4889	0.5726	0.8004	0.1566	0.1814	0.2573
DGNet [49]	RGB	0.6208	0.1876	0.0602	0.5132	0.7000	0.8426	0.1730	0.2624	0.3061
	FPDE	0.6177	0.1929	0.0802	0.5553	0.7026	0.8682	0.2007	0.2846	0.3404
	GFCE	0.6340	0.2122	0.0714	0.5642	0.7005	0.8083	0.2137	0.2845	0.3282
	LatLRR	0.6297	0.2097	0.0723	0.5518	0.7043	0.8615	0.2080	0.2736	0.3200
HitNet [20]	RGB	0.6514	0.2781	0.0433	0.6130	0.7567	0.8255	0.2771	0.3278	0.3707
	FPDE	0.6749	0.3258	0.0228	0.6443	0.8020	0.8925	0.2916	0.3506	0.3828
	GFCE	0.6694	0.3248	0.0242	0.6316	0.8113	0.8480	0.2850	0.3642	0.4002
	LatLRR	0.6433	0.2765	0.0330	0.5508	0.6917	0.8864	0.2480	0.3053	0.3732
PCNet [19]	RGB	0.7328	0.4087	0.0158	0.8583	0.8447	0.8719	0.4212	0.4358	0.4460
	FPDE	0.7260	0.4137	0.0164	0.8281	0.8357	0.9112	0.4123	0.4433	0.4568
	GFCE	0.7385	0.4210	0.0161	0.8247	0.8140	0.8481	0.4227	0.4496	0.4607
	LatLRR	0.7006	0.3443	0.0170	0.8076	0.7365	0.8382	0.3838	0.3708	0.3830

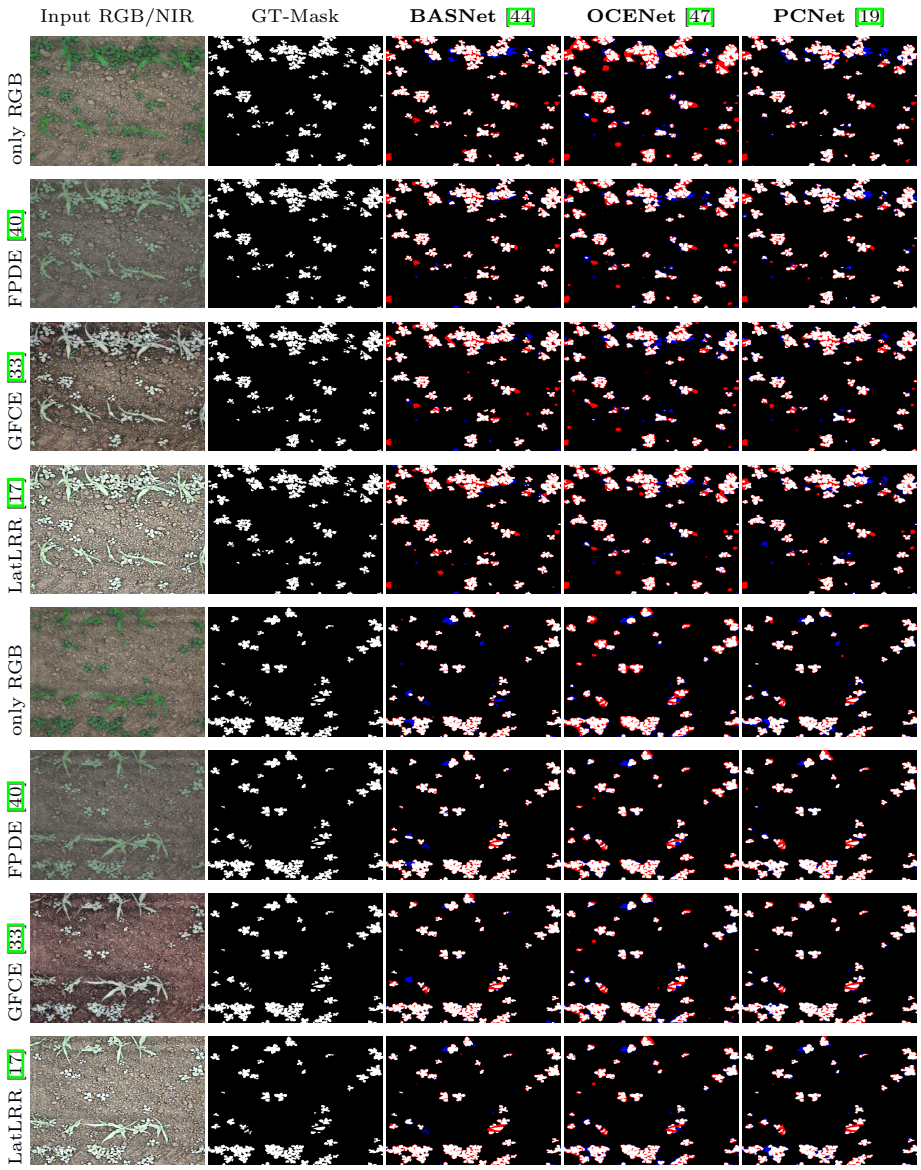

Figure 9: Prediction results for top-3 SOTA COD techniques using the WeedsGalore [26] dataset in two sample images. Successful matches between GT and predicted masks (white areas); False positive regions (red areas, over-segmentation); and false negative regions (blue areas, miss-segmentation).

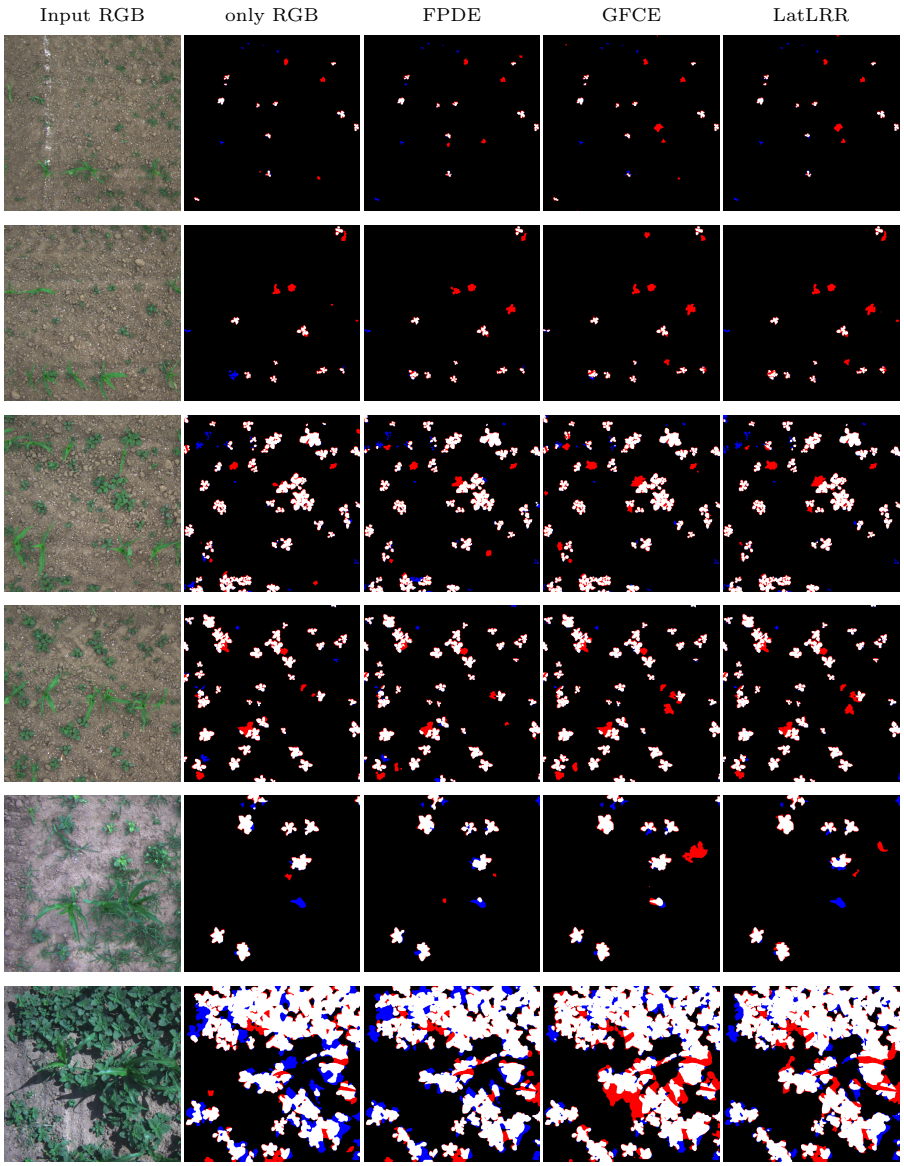

Figure 10: Prediction results for BASNet [44] the best overall technique using the WeedsGalore [26] dataset. Successful matches between GT and predicted masks (white areas); False positive regions (red areas, over-segmentation); and false negative regions (blue areas, miss-segmentation).

5. Discussion

The experimental results across the two case studies demonstrate that integrating NIR information with RGB imagery improves the segmentation of camouflaged weeds in real agricultural settings, while also showing that the optimal fusion strategy depends on the dataset characteristics.

In the banana plantation scenario, the fusion approaches CBF, HMSD-GF, and LatLRR consistently rank among the best, and the downstream COD results show clear gains when models are trained with fused inputs rather than RGB alone. This behavior is particularly evident in small or thin weed regions, where RGB-only segmentation tends to miss low-contrast areas. Among the evaluated COD methods, PCNet stands out as the most reliable approach, producing masks with sharper contours and fewer false negatives, which is consistent with its design for plant camouflage detection.

In the maize field scenario, the fusion ranking changes, and the best-performing fusion configurations become LatLRR, GFCE, and FPDE, confirming that fusion performance is sensitive to acquisition conditions, scene structure, and target appearance. While fusion still provides benefits over RGB-only inputs, the improvement is more heterogeneous across methods and metrics than in Weeds-Banana. This can be attributed to the different imaging conditions (very high spatial resolution tiles, different UAV/sensor, multi-temporal variability) and to the fact that weeds often appear as small, scattered objects, making the task more sensitive to small false positives/negatives. In this dataset, BASNet provides the strongest overall trade-off and stable qualitative behavior, while PCNet and OCENet remain highly competitive depending on the metric emphasis (e.g., region overlap vs. boundary consistency).

Across both datasets, a consistent conclusion is that RGB–NIR fusion improves the representation available to COD models, enabling earlier and more accurate weed localization, which is crucial for precision agriculture workflows that aim to reduce herbicide usage through targeted interventions. At the same time, the cross-dataset behavior highlights two practical implications: (1)

there is no universally optimal fusion method, and fusion should be selected (or learned) with awareness of the target domain; and (2) COD architectures exhibit dataset-dependent rankings, suggesting that model selection should consider the crop type, weed morphology, resolution, and density. Future work should therefore explore adaptive or learnable fusion strategies and broader validations across additional crops, sensors, and environmental conditions.

Despite the overall gains from RGB–NIR fusion, some failure modes are observed. First, very thin weed structures or extremely small weeds may be under-segmented (false negatives), particularly when their spatial footprint approaches the sensor resolution or when their NIR/RGB contrast is weak. Second, visually ambiguous transitions between soil residues and vegetation can produce localized over-segmentation (false positives), especially in high-texture backgrounds. Third, imperfect RGB–NIR registration and radiometric inconsistencies (e.g., brightness differences across the mosaic) can introduce edge artifacts that affect boundary precision. These observations suggest that further improvements could be obtained by (i) stronger domain-specific augmentation, (ii) more robust cross-spectral registration, and (iii) fusion strategies that explicitly model uncertainty or confidence at boundaries.

Finally, limitations of this study include (i) the dependency of fusion performance on the target domain (sensor characteristics, resolution, and scene structure), (ii) potential sensitivity to strong shadows, specular highlights, and radiometric drift across large mosaics, and (iii) the fact that COD model rankings can change with weed density, morphology, and crop row structure. Although we partially address domain shift by evaluating on WeedsGalore (different crop, country, platform, and multi-temporal acquisition), further multi-site and multi-season validation is required to fully characterize robustness.

6. Conclusions

The experimental results validate the proposed framework for detecting camouflaged weeds using RGB–NIR imagery and show its effectiveness across

two different agricultural scenarios: banana plantations (own Weeds-Banana dataset) and maize fields (WeedsGalore dataset). In both case studies, incorporating NIR information via RGB–NIR fusion improves weed segmentation performance compared to RGB-only inputs, confirming that additional spectral information provides more discriminative cues for camouflaged vegetation.

For Weeds-Banana, the best-performing fusion techniques are CBF, HMSD-GF, and LatLRR, and among the nine evaluated COD approaches, PCNet achieves the most consistent top performance across the majority of metrics, making it particularly suitable for plant camouflage segmentation under real field conditions. The combination with the best performance for this first case study is PCNet-LatLRR. For WeedsGalore, the fusion ranking differs—highlighting the dataset dependency of fusion quality—with LatLRR, GFCE, and FPDE emerging as the most competitive. In this second case study, BASNet provides the strongest overall trade-off and stable qualitative predictions, while PCNet and OCENet remain competitive depending on the metric emphasis. For this second case study, the winning combination is BASNet-FPDE.

Overall, these results show that RGB–NIR fusion is a practical and transferable strategy for improving early weed detection, but also that the best fusion and COD configurations should be selected with awareness of the target crop, sensor, and scene properties. Several research directions are proposed to extend and improve the presented work. First, expanding datasets to include different crop types, climatic conditions, and weed species would improve model generalization across diverse agricultural scenarios. Second, integrating additional spectral bands such as red-edge (already available in WeedsGalore) and thermal infrared could further enhance discrimination between crops and weeds. Third, developing computationally efficient variants would enable real-time implementation on UAV systems with limited computational resources. Fourth, incorporating temporal analysis (multi-date monitoring) would support weed growth tracking and proactive management. Finally, coupling segmentation outputs with selective spraying or mechanical weeding systems would move toward a fully automated weed management pipeline.

Acknowledgements

This work is supported in part by the Air Force Office of Scientific Research Under Award FA9550-24-1-0206; in part by the ESPOL project "Advancing Camouflaged Object Detection with a cost-effective Cross-Spectral vision system (ACODCS)" (CIDIS-003-2024).

References

[1] S. Ghazal, A. Munir, W. S. Qureshi, Computer vision in smart agriculture and precision farming: Techniques and applications, Artificial Intelligence in Agriculture 13 (2024) 64–83. doi:https://doi.org/10.1016/j.aiia.2024.06.004.

[2] A. Kamilaris, F. X. Prenafeta-Boldú, A review of the use of convolutional neural networks in agriculture, The Journal of Agricultural Science 156 (3) (2018) 312–322. doi:10.1017/S0021859618000436.

[3] J. F. S. Gomes, F. R. Leta, Applications of computer vision techniques in the agriculture and food industry: a review, European Food Research and Technology 235 (2012) 989–1000.

[4] K. R. S. R. Punithavathi, A. Delphin Carolina Rani, Computer vision and deep learning-enabled weed detection model for precision agriculture, Computer Systems Science and Engineering 44 (3) (2023) 2759–2774. doi:10.32604/csse.2023.027647.

[5] D.-P. Fan, G.-P. Ji, G. Sun, M.-M. Cheng, J. Shen, L. Shao, Camouflaged object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[6] Y. Lv, J. Zhang, Y. Dai, A. Li, N. Barnes, D.-P. Fan, Toward deeper understanding of camouflaged object detection, IEEE transactions on circuits and systems for video technology 33 (7) (2023) 3462–3476.

[7] S. Lev-Yadun, A. Dafni, M. A. Flaishman, M. Inbar, I. Izhaki, G. Katzir, G. Ne'eman, Plant coloration undermines herbivorous insect camouflage, BioEssays 26 (10) (2004) 1126–1130.

[8] S. M. Hussain, S. A. Zaidi, A. Hyder, S. M. A. Rizvi, M. Farhan, Enhanced camouflaged object detection for agricultural pest management: Insights from unified benchmark dataset analysis, in: 2024 21st International Bhurban Conference on Applied Sciences and Technology (IBCAST), IEEE, 2024, pp. 273–280.

[9] A. K. Gautam, P. Preet, T. S. Rawat, P. Roy Chowdhury, L. K. Sinha, Detection of camouflaged targets in hyperspectral images, in: P. S. Sastry, J. CV, D. Raghavamurthy, S. S. Rao (Eds.), Advances in Small Satellite Technologies, Springer Singapore, Singapore, 2020, pp. 155–161.

[10] C. L. Hi, K. L. Lew, C. Zheng, T. Kurniawan, S. A. Babale, C. S. Lee, A review of camouflage object detection techniques, International Journal on Robotics, Automation and Sciences 7 (1) (2025) 84–92. doi:10.33093/ijoras.2025.7.1.10.

[11] M. O. Gani, S. Kuiry, A. Das, M. Nasipuri, N. Das, Multispectral object detection with deep learning, in: P. Dutta, J. K. Mandal, S. Mukhopadhyay (Eds.), Computational Intelligence in Communications and Business Analytics, Springer International Publishing, Cham, 2021, pp. 105–117.

[12] Z. Zhang, Q. Wang, Z. Li, Z. Zhou, X. Xie, H. Sun, C. Chen, M. Wang, X. Dong, Y. Wu, Z. He, B. Liu, X. Qian, J. Li, L. Wang, F. Xiu, J. Liu, W. Huang, A skin-beyond multifrequency camouflage system with self-adaptive discoloration and radar-infrared stealth, Chemical Engineering Journal 494 (2024) 152867. doi:https://doi.org/10.1016/j.cej.2024.152867.

[13] S. Wang, D. Zeng, Y. Xu, G. Yang, F. Huang, L. Chen, Towards complex scenes: A deep learning-based camouflaged people detection method

for snapshot multispectral images, Defence Technology 34 (2024) 269–281. doi:https://doi.org/10.1016/j.dt.2023.12.011.

[14] H. Zhang, H. Xu, X. Tian, J. Jiang, J. Ma, Image fusion meets deep learning: A survey and perspective, Information Fusion 76 (2021) 323–336.

[15] B. K. Shreyamsha Kumar, Image fusion based on pixel significance using cross bilateral filter, Signal, Image and Video Processing 9 (5) (2015) 1193–1204. doi:10.1007/s11760-013-0556-9.

[16] Z. Zhou, M. Dong, X. Xie, Z. Gao, Fusion of infrared and visible images for night-vision context enhancement, Appl. Opt. 55 (23) (2016) 6480–6490. doi:10.1364/AO.55.006480.

[17] H. Li, X.-J. Wu, Infrared and visible image fusion using latent low-rank representation, arXiv preprint arXiv:1804.08992.

[18] F. Xiao, S. Hu, Y. Shen, C. Fang, J. Huang, C. He, L. Tang, Z. Yang, X. Li, A survey of camouflaged object detection and beyond (2024). arXiv:2408.14562.

[19] J. Yang, Q. Wang, F. Zheng, P. Chen, A. Leonardis, D.-P. Fan, Plantcamo: Plant camouflage detection, arXiv preprint arXiv:2410.17598.

[20] X. Hu, S. Wang, X. Qin, H. Dai, W. Ren, D. Luo, Y. Tai, L. Shao, High-resolution iterative feedback network for camouflaged object detection, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 37, 2023, pp. 881–889.

[21] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, L. Shao, Pyramid vision transformer: A versatile backbone for dense prediction without convolutions, in: Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 568–578.

[22] T. Chen, J. Xiao, X. Hu, G. Zhang, S. Wang, Boundary-guided network for camouflaged object detection, Knowledge-based systems 248 (2022) 108901.

[23] S.-H. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, P. Torr, Res2net: A new multi-scale backbone architecture, IEEE transactions on pattern analysis and machine intelligence 43 (2) (2019) 652–662.

[24] S. Haug, J. Ostermann, A crop/weed field image dataset for the evaluation of computer vision based precision agriculture tasks, in: European conference on computer vision, Springer, 2014, pp. 105–116.

[25] M. Krestenitis, E. K. Raptis, A. C. Kapoutsis, K. Ioannidis, E. B. Kosmatopoulos, S. Vrochidis, I. Kompatsiaris, Cofly-weeddb: A uav image dataset for weed detection and species identification, Data in Brief 45 (2022) 108575.

[26] E. Celikkan, T. Kunzmann, Y. Yeskaliyev, S. Itzerott, N. Klein, M. Herold, Weedsgalore: A multispectral and multitemporal uav-based dataset for crop and weed segmentation in agricultural maize fields (2025). arXiv:2502.13103.
URL https://arxiv.org/abs/2502.13103

[27] C. Lu, K. Gehring, S. Kopfinger, H. Bernhardt, M. Beck, S. Walther, T. Ebertseder, M. Minceva, Y. Hu, K. Yu, Weed instance segmentation from uav orthomosaic images based on deep learning, Smart Agricultural Technology (2025) 100966.

[28] D. M. Gates, H. J. Keegan, J. C. Schleter, V. R. Weidner, Spectral properties of plants, Applied Optics 4 (1) (1965) 11–20. doi:10.1364/AO.4.000011.
URL https://opg.optica.org/ao/abstract.cfm?uri=ao-4-1-11

[29] E. B. Knipling, Physical and physiological basis for the reflectance of visible and near-infrared radiation from vegetation, Remote Sensing of Environment 1 (3) (1970) 155–159. doi:10.1016/S0034-4257(70)80021-9.
URL https://www.sciencedirect.com/science/article/abs/pii/S0034425770800219

[30] S. V. Ollinger, Sources of variability in canopy reflectance and the convergent properties of plants, New Phytologist doi:10.1111/j.1469-8137.2010.03536.x.
URL https://nph.onlinelibrary.wiley.com/doi/10.1111/j.1469-8137.2010.03536.x

[31] X. Zhang, P. Ye, G. Xiao, Vifb: A visible and infrared image fusion benchmark, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020.

[32] B. Shreyamsha Kumar, Image fusion based on pixel significance using cross bilateral filter, Signal, image and video processing 9 (2015) 1193–1204.

[33] Z. Zhou, M. Dong, X. Xie, Z. Gao, Fusion of infrared and visible images for night-vision context enhancement, Applied optics 55 (23) (2016) 6480–6490.

[34] Z. Zhou, B. Wang, S. Li, M. Dong, Perceptual fusion of infrared and visible images through a hybrid multi-scale decomposition with gaussian and bilateral filters, Information fusion 30 (2016) 15–26.

[35] D. P. Bavirisetti, R. Dhuli, Fusion of infrared and visible sensor images based on anisotropic diffusion and karhunen-loeve transform, IEEE Sensors Journal 16 (1) (2015) 203–209.

[36] D. P. Bavirisetti, G. Xiao, J. Zhao, R. Dhuli, G. Liu, Multi-scale guided image and video fusion: A fast and efficient approach, Circuits, Systems, and Signal Processing 38 (12) (2019) 5576–5605.

[37] V. Naidu, Image fusion technique using multi-resolution singular value decomposition, Defence Science Journal 61 (5) (2011) 479.

[38] J. Ma, C. Chen, C. Li, J. Huang, Infrared and visible image fusion via gradient transfer and total variation minimization, Information Fusion 31 (2016) 100–109.

[39] D. P. Bavirisetti, R. Dhuli, Two-scale image fusion of visible and infrared images using saliency detection, Infrared Physics & Technology 76 (2016) 52–64.

[40] D. P. Bavirisetti, G. Xiao, G. Liu, Multi-sensor image fusion based on fourth order partial differential equations, in: 2017 20th International conference on information fusion (Fusion), IEEE, 2017, pp. 1–9.

[41] J. Ma, Z. Zhou, B. Wang, H. Zong, Infrared and visible image fusion based on visual saliency map and weighted least square optimization, Infrared Physics & Technology 82 (2017) 8–17.

[42] Y. Liu, X. Chen, J. Cheng, H. Peng, Z. Wang, Infrared and visible image fusion with convolutional neural networks, International Journal of Wavelets, Multiresolution and Information Processing 16 (03) (2018) 1850018.

[43] Y. Zhang, L. Zhang, X. Bai, L. Zhang, Infrared and visual image fusion through infrared feature extraction and visual information preservation, Infrared Physics & Technology 83 (2017) 227–237.

[44] X. Qin, Z. Zhang, C. Huang, C. Gao, M. Dehghan, M. Jagersand, Basnet: Boundary-aware salient object detection, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[45] D.-P. Fan, G.-P. Ji, M.-M. Cheng, L. Shao, Concealed object detection, IEEE transactions on pattern analysis and machine intelligence 44 (10) (2021) 6024–6042.

[46] G. Chen, S.-J. Liu, Y.-J. Sun, G.-P. Ji, Y.-F. Wu, T. Zhou, Camouflaged object detection via context-aware cross-level fusion, IEEE Transactions on Circuits and Systems for Video Technology 32 (10) (2022) 6981–6993.

[47] J. Liu, J. Zhang, N. Barnes, Modeling aleatoric uncertainty for camouflaged object detection, in: Proceedings of the IEEE/CVF winter conference on applications of computer vision, 2022, pp. 1445–1454.

[48] D. Sun, S. Jiang, L. Qi, Edge-aware mirror network for camouflaged object detection, in: 2023 IEEE International Conference on Multimedia and Expo (ICME), IEEE, 2023, pp. 2465–2470.

[49] G.-P. Ji, D.-P. Fan, Y.-C. Chou, D. Dai, A. Liniger, L. Van Gool, Deep gradient learning for efficient camouflaged object detection, Machine Intelligence Research 20 (1) (2023) 92–108.

[50] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Conf. on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[51] M. Tan, Q. Le, Efficientnet: Rethinking model scaling for convolutional neural networks, in: International conference on machine learning, PMLR, 2019, pp. 6105–6114.

[52] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, L. Shao, Pvt v2: Improved baselines with pyramid vision transformer, Computational Visual Media 8 (3) (2022) 415–424.

[53] D. Bulanon, T. Burks, V. Alchanatis, Image fusion of visible and thermal images for fruit detection, Biosystems engineering 103 (1) (2009) 12–22.

[54] J. W. Roberts, J. A. Van Aardt, F. B. Ahmed, Assessment of image fusion procedures using entropy, image quality, and multispectral classification, Journal of Applied Remote Sensing 2 (1) (2008) 023522.

[55] G. Qu, D. Zhang, P. Yan, Information measure for performance of image fusion, Electronics letters 38 (7) (2002) 313–315.

[56] P. Jagalingam, A. V. Hegde, A review of quality metrics for fused image, Aquatic Procedia 4 (2015) 133–142.

[57] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE transactions on image processing 13 (4) (2004) 600–612.

[58] G. Cui, H. Feng, Z. Xu, Q. Li, Y. Chen, Detail preserved fusion of visible and infrared images using regional saliency extraction and multi-scale image decomposition, Optics Communications 341 (2015) 199–209.

[59] B. Rajalingam, R. Priya, Hybrid multimodality medical image fusion technique for feature enhancement in medical diagnosis, International Journal of Engineering Science Invention 2 (Special issue) (2018) 52–60.

[60] Y.-J. Rao, In-fibre bragg grating sensors, Measurement science and technology 8 (4) (1997) 355.

[61] C. S. Xydeas, V. Petrovic, Objective image fusion performance measure, Electronics letters 36 (4) (2000) 308–309.

[62] Y. Chen, R. S. Blum, A new automated quality assessment algorithm for image fusion, Image and vision computing 27 (10) (2009) 1421–1432.

[63] H. Chen, P. K. Varshney, A human perception inspired quality metric for image fusion based on regional information, Information fusion 8 (2) (2007) 193–207.

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

[65] R. Margolin, L. Zelnik-Manor, A. Tal, How to evaluate foreground maps?, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 248–255.

[66] F. Perazzi, P. Krähenbühl, Y. Pritch, A. Hornung, Saliency filters: Contrast based filtering for salient region detection, in: 2012 IEEE conference on computer vision and pattern recognition, IEEE, 2012, pp. 733–740.

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

[68] R. Achanta, S. Hemami, F. Estrada, S. Susstrunk, Frequency-tuned salient region detection, in: 2009 IEEE conference on computer vision and pattern recognition, IEEE, 2009, pp. 1597–1604.

IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2024.0429000

AINet: Integrating Mamba and CBAM for Enhanced Camouflage Object Detection

HENRY O. VELESACA[1,4], ANDREA MERO P.[1,5], ABEL REYES-ANGULO[3], and ANGEL D. SAPPA[1,2]

[1]ESPOL Polytechnic University, FIEC, CIDIS, Campus Gustavo Galindo, 090902, Guayaquil, Ecuador (e-mail: hvelesac@espol.edu.ec, anmero@espol.edu.ec, asappa@espol.edu.ec)
[2]Computer Vision Center, Universitat Autònoma de Barcelona, 08193-Bellaterra, Barcelona, Spain (e-mail: asappa@cvc.uab.es)
[3]Michigan Technological University, Houghton, MI, EEUU (e-mail: areyesan@mtu.edu)
[4]Software Engineering Department, Research Center for Information and Communication Technologies (CITIC-UGR), University of Granada, 18071, Granada, Spain (e-mail: hvelesaca@correo.ugr.es)
[5]Università della Svizzera Italiana, Via Giuseppe Buffi 13, 6900, Lugano, Switzerland (e-mail: andrea.mero@usi.ch)

Corresponding author: Henry O. Velesaca (e-mail: hvelesac@espol.edu.ec).

This work is supported in part by the Air Force Office of Scientific Research Under Award FA9550-24-1-0206; in part by the ESPOL project "Advancing Camouflaged Object Detection with a cost-effective Cross-Spectral vision system (ACODCS)" (CIDIS-003-2024).

ABSTRACT This paper introduces AINet, a novel deep learning architecture designed for detecting camouflaged objects in complex and diverse environments. The objective of this work is to design an end-to-end camouflaged object detection architecture that simultaneously captures long-range dependencies and refines subtle camouflage cues, improving segmentation accuracy and boundary delineation across both standard COD benchmarks and real-world agricultural scenarios. AINet leverages the strengths of Mamba, an efficient sequential state model for capturing long-range dependencies, and the Convolutional Block Attention Module (CBAM) for feature refinement through attention mechanisms. Detecting camouflaged objects is a significant challenge across a wide range of real-world applications, including surveillance, security, medical imaging, and autonomous systems, where objects of interest may blend into their backgrounds and evade conventional detection methods. To demonstrate its effectiveness, AINet is evaluated on multiple datasets, including standard camouflaged object detection benchmarks such as CAMO, COD10K, and NC4K, as well as domain-specific datasets (such as pest and fruit detection). Experimental results show that AINet outperforms existing state-of-the-art models. The implementation is publicly available on GitHub for reproducibility: https://github.com/hvelesaca/AINet.

INDEX TERMS Camouflaged object detection, fruit harvest, pest detection, precision agriculture.

I. INTRODUCTION

Detecting camouflaged objects remains a persistent and challenging problem in computer vision [46]. Camouflaged objects are characterized by their ability to blend seamlessly into their surroundings, making them difficult to distinguish using conventional detection methods [30]. This challenge is prevalent in a wide array of real-world scenarios, including surveillance, security, medical imaging, autonomous vehicles, and environmental monitoring, where accurate identification of hidden or obscured objects is critical for decision-making and safety [21]. Moreover, camouflaged object detection not only demands robustness to variations in illumination, scale, and background clutter, but also the capacity to infer subtle semantic and structural differences that may not be immediately apparent at the pixel level, thereby pushing current detection paradigms to their limits.

Traditional object detection algorithms often struggle with camouflaged targets due to the minimal contrast and ambiguous boundaries between the object and its background [9]. Recent advances in deep learning have led to significant improvements in object detection and segmentation; however, the unique nature of camouflage still poses obstacles [6]. Addressing these challenges requires models capable of capturing both global context and fine-grained details, as well as mechanisms to focus attention on subtle cues that differentiate camouflaged objects from their environments [4].

This paper presents AINet, a novel deep learning architecture specifically designed to improve the detection of camouflaged objects. AINet integrates two powerful components: Mamba [11], an efficient sequential state model capable of modeling long-range dependencies, and the Convolutional Block Attention Module (CBAM) [45], which refines feature representations using spatial and per-channel attention mechanisms. By combining these elements, AINet can effectively

highlight and segment camouflaged objects, even in highly complex scenes.

The key contributions of this work include:

- AINet is proposed, a COD-oriented encoder–decoder architecture that integrates selective state-space modeling (Mamba) for efficient long-range dependency modeling with attention-based feature refinement (CBAM) to emphasize subtle camouflage cues.
- A comprehensive evaluation of COD standard benchmarks (CAMO, COD10K, NC4K) and agricultural case studies is offered, including quantitative, qualitative, and ablation analyses to validate the contribution of each component.
- Extensive ablation studies demonstrate that the synergy of Mamba and CBAM modules, as well as multi-level deep supervision, is critical for optimal segmentation accuracy and robust edge delineation.
- The architecture is validated across multiple datasets and evaluation metrics, demonstrating substantial improvements in accuracy, boundary precision, and overall segmentation mask quality.

The manuscript is organized as follows. Section II introduces related work, recent SOTA COD techniques, and methods that address the problem of the COD approach. Section III presents the proposed architecture. Then, Section IV shows the experimental results on different datasets, and an in-depth ablation study is presented in Section V. Finally, discussion and conclusions are given in Section VI and Section VII respectively.

II. RELATED WORKS

This section reviews key computer vision-based methods for camouflaged object detection (COD), analyzing their contributions and limitations across diverse application domains.

Detecting camouflaged objects is a persistent challenge in computer vision, as these objects are characterized by their ability to blend seamlessly into complex backgrounds, making them difficult to distinguish using conventional detection techniques. Recent advances in deep learning have significantly improved image segmentation and object detection, enabling more accurate and efficient identification of camouflaged targets in a variety of real-world scenarios, such as surveillance, security, medical imaging, environmental monitoring, and autonomous systems.

One of the pioneering state-of-the-art COD techniques is the Search Identification Network (SINet) [10], which introduces a two-stage framework inspired by predator hunting behavior, consisting of a Search Module (SM) and an Identification Module (IM). SINet presents a simple yet effective end-to-end architecture based on the richly annotated COD10K dataset, achieving visually appealing results compared to existing baselines. Building on this, SINet-V2 [9] enhances the original design by incorporating densely connected layers and a receptive field component to better capture multi-level features. This improved architecture demonstrates competitive performance and broader applicability, including potential use in military, security, and wildlife conservation, where the detection of concealed objects is essential [17].

Another notable contribution is the SegMaR technique [20], which introduces a multistage iterative refinement framework for camouflaged object detection, simulating the coarse-to-fine detection process of the human visual system. The framework consists of three core steps: *Segment*, *Magnify*, and *Reiterate*. Initially, a camouflaged segmentation network generates a preliminary mask for the object. The *Magnify* module then adaptively enlarges the object region using attention-based sampling, making it more distinct within the image. Finally, the *Reiterate* module refines the segmentation through iterative feedback, progressively capturing finer details, especially for small or highly camouflaged objects. The architecture also incorporates a distraction module to disentangle foreground and background features, and employs parallel decoders to focus on key object regions and contours. However, SegMaR cannot be trained end-to-end [14], which restricts its adaptability in certain scenarios.

COD approaches face two primary challenges: (1) intrinsic similarity (IS), where objects visually resemble the background, and (2) edge disruption (ED), which results in unclear boundaries. Biologically inspired methods like SINet [10] and SegMaR [20] attempt to address these by mimicking predator hunting behavior or human visual cognition through multistage processing. However, these approaches often struggle with complex camouflage patterns and may fail to capture the subtle cues needed to resolve IS and ED effectively. To address these limitations, the FEature Decomposition and Edge Reconstruction (FEDER) technique [12] adopts a targeted two-stage strategy. It uses learnable wavelets to decompose image features and identify the most informative frequency bands via a frequency attention and feature aggregation module. To address ED, it introduces an auxiliary edge reconstruction task inspired by differential equations, improving boundary precision and overall detection accuracy. Despite promising results, FEDER's reliance on edge reconstruction may be less effective when camouflaged objects have very similar textures to their surroundings, leading to potential false negatives.

The field of COD has seen remarkable progress in recent years. Foundational work by Qin et al. [40] on salient object detection has profoundly influenced the evolution of COD methodologies. In 2022, Chen et al. [2] proposed a context-aware cross-level fusion approach that enhanced camouflaged object identification accuracy, while Chen et al. [3] introduced a boundary-guided network to improve edge and feature detection. Liu et al. [29] further advanced the field by addressing aleatoric uncertainty modeling in COD. More recently, high-resolution iterative feedback networks [16], edge-aware networks [41], and deep gradient learning techniques [18] have contributed to significant improvements in detection efficiency and accuracy. The scope of COD research continues to expand, with recent work such as PlantCamo [50] targeting specialized applications in plant camouflage detection.

While much of the research has focused on general-purpose COD, these advances have also been adapted to address domain-specific challenges. For example, in agricultural environments, camouflage plays a critical role in both pest and fruit detection, complicating monitoring and management tasks [26]. Recent studies have explored the use of advanced deep learning architectures for these applications. Meng et al. [37] addressed camouflaged pest instance segmentation by combining Pyramid Vision Transformer (PVT) and Mask R-CNN, leveraging PVT's hierarchical feature extraction and Mask R-CNN's instance-level segmentation capabilities. Similarly, Evangelista et al. [6] introduced FCNet, a transformer-based, context-aware segmentation framework for detecting camouflaged fruits in complex orchard environments. These works demonstrate the potential of adapting general COD methodologies to agriculture, though challenges remain regarding generalization to diverse conditions and the need for large labeled datasets.

Collectively, these advances have established more robust and accurate camouflaged object detection systems, setting new benchmarks in the field and paving the way for future research and practical implementations. In the following, we highlight how these general advances have been adapted and extended to address specific challenges in agricultural environments, such as pest and fruit detection.

III. PROPOSED AINET

Consistent with recent advances in COD [9], [10], [22], [38], [55], an encoder-decoder pipeline is adopted for the proposed AINet architecture. The framework is designed to be end-to-end trainable, as illustrated in Fig. 1.

A. OVERALL ARCHITECTURE

AINet is constructed by integrating Mamba and CBAM modules within an encoder-decoder design tailored for COD. The selective state space modeling capabilities of Mamba are combined with the feature refinement provided by the Convolutional Block Attention Module (CBAM), enabling effective identification of objects that visually blend into their surroundings.

Encoder. A PVTv2-B2 backbone [43] is employed as the encoder, providing hierarchical feature representations at multiple scales. From the input RGB image $I \in R^{H\times W\times 3}$, feature maps at four different resolutions are extracted, capturing both low-level details and high-level semantic information necessary for camouflaged object detection.

PVTv2-B2 is adopted as the encoder because COD demands both fine-grained, multi-scale detail extraction and global context modeling to distinguish targets that blend into the background. In addition, several representative COD pipelines that rely on ResNet50-family backbones, as reported in Table 4 (e.g., LSR, MGL, PFNet, TINet, UGTR), achieve lower benchmark accuracy than the proposed AINet on CAMO, COD10K, and NC4K. Although overall performance depends on the full architecture rather than the backbone alone, these results support using a modern hierarchical encoder for this task.

Decoder. A progressive refinement strategy is utilized in the decoder, where Decoder Blocks upsample features from the previous stage and aggregate them with skip connections from the encoder. Feature Aggregation is performed using 1×1 and 3×3 convolutions. Multiple segmentation heads at different decoder levels are incorporated to provide deep supervision, enhancing gradient flow and feature learning.

Loss Function. The loss function proposed by [44] is adopted, following established practices [9], [42]. The predictions generated by the decoder are denoted as $\{P_i\}_{i=0}^{3}$. During training, each prediction Pi is resized to the original input size and supervised using a combination of binary cross-entropy (BCE) loss [5] and intersection-over-union (IoU) loss [35]. The total loss is computed by summing the losses from all decoder stages, as follows, where GT denotes the ground truth annotation:

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

To further enhance feature learning at different scales, deep supervision with multiple segmentation heads is employed. Each decoder level produces a segmentation map supervised by the ground truth, and the final prediction is obtained by averaging the outputs from all levels. This multi-level supervision strategy enables the learning of robust features for detecting camouflaged objects across varying scales and concealment levels.

B. MAMBA AND CBAM BLOCKS

A core component of AINet is the Mamba Block, in which residual learning is combined with state space modeling [11], following the approach in [33]. The Mamba Block is composed of Residual Blocks (two consecutive blocks with instance normalization and LeakyReLU activation to enhance feature representation while preserving spatial information); Dual-Branch Processing, including a Mamba Branch (processing features via linear projection, 1D convolution, and the Mamba state space model to capture long-range dependencies) and an Activation Branch (applying SiLU activation for local information preservation); and a Hadamard Product, where outputs from both branches are combined through element-wise multiplication for adaptive feature fusion. Figure 1 illustrates the Mamba Block.

On the other hand, the Convolutional Block Attention Module (CBAM) [45] is also incorporated to refine features by applying both channel and spatial attention. Channel attention is computed using average and max pooling, followed by a shared MLP to generate channel-wise attention weights, while spatial attention is generated through channel pooling and convolution to produce a spatial attention map.

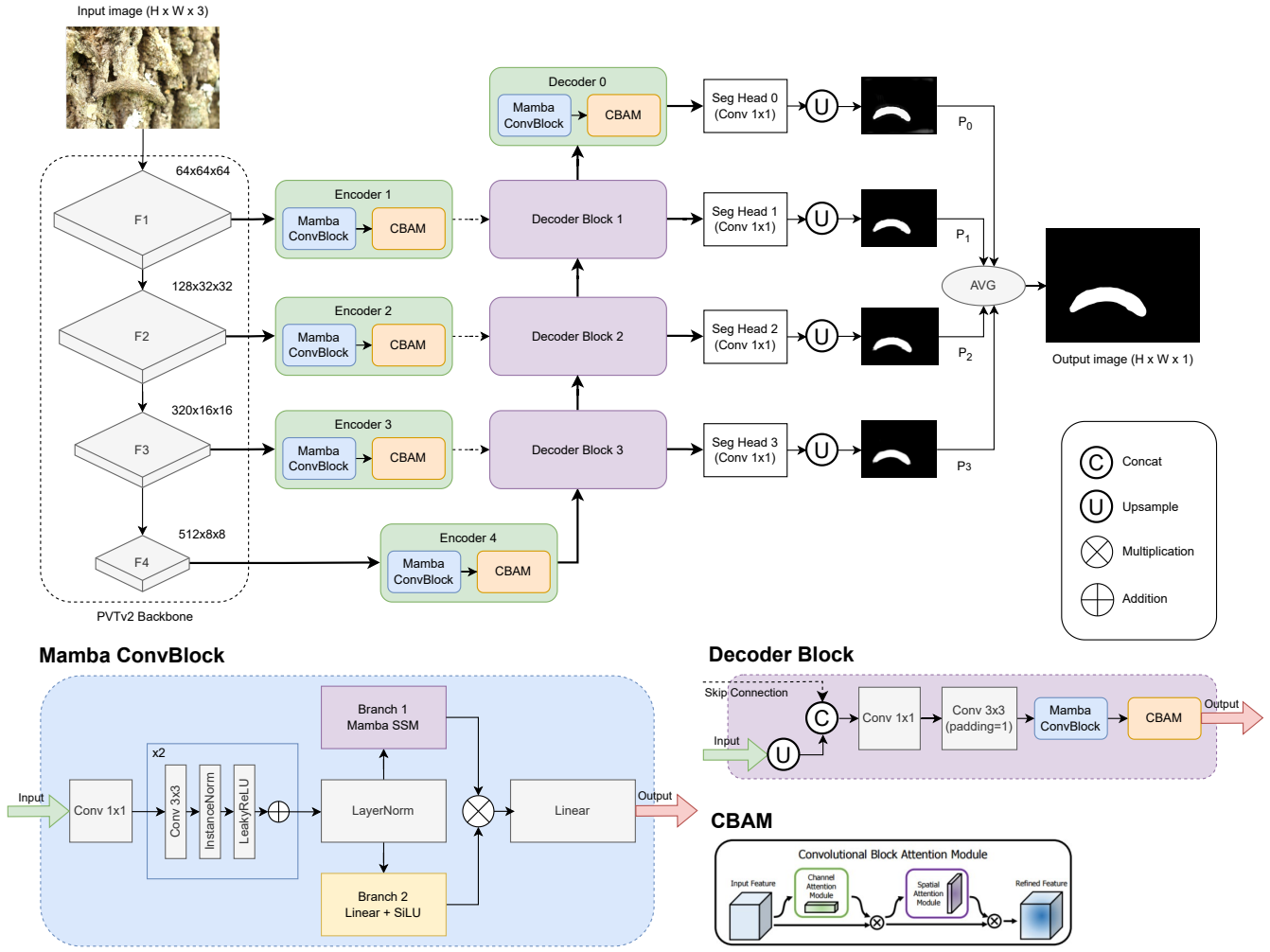

FIGURE 1. **The overall architecture of the proposed AINet.**

C. IMPLEMENTATION DETAILS

AINet has been implemented using the PyTorch library. The encoder is initialized with a PVTv2-B2 backbone [43] pre-trained on ImageNet. AdamW is used for optimization, with a weight decay of $1e^{-4}$. The initial learning rate is set to $1e^{-4}$ and is scheduled using cosine annealing. Input images are resized to 352×352 for both training and inference. The model is trained end-to-end for 100 epochs with a batch size of 16 on an NVIDIA TESLA P100 GPU. All experiments are conducted on the Kaggle platform[1].

Datasets. To comprehensively evaluate the effectiveness and generalization of AINet, experiments have been conducted on both standard benchmark datasets and domain-specific agricultural datasets. For general COD evaluation, three widely recognized benchmark datasets have been utilized: CAMO [25], [47], which contains 2,500 images balanced between camouflaged and non-camouflaged objects; COD10K [9], [10], which offers 5,066 camouflaged, 3,000 background, and 1,934 non-camouflaged images; and NC4K [32], contributing 4,121 images exclusively for testing.

Type	Dataset	Number of images		
		Train	Val	Test
Benchmark	CAMO [25], [47]	797	203	250
	COD10K [9], [10]	2435	605	2026
	NC4K [32]	-	-	4121
Agriculture	Cotton Bollworm [37]	856	161	56
	Mango [24]	1944	486	230

TABLE 1. **Distribution of the datasets used in experimental results**

Following established protocols by [10], [38], we use 1,000 images from CAMO and 3,040 images from COD10K for model development (train+val). Specifically, as summarized in Table 1, CAMO uses 797/203/250 images for train/val/test, and COD10K uses 2,435/605/2,026 images for train/val/test. NC4K is used exclusively for testing (4,121 images).

To further assess the applicability of AINet, two case stud-

[1] https://www.kaggle.com/

ies in agricultural scenarios have been selected. Hence, two domain-specific datasets are included: the Cotton Bollworm dataset [37], which contains 1,073 camouflaged pest images, and the Mango dataset [24], comprising 2,660 patch images for fruit detection. Figure 2 presents sample images from benchmark and agricultural datasets. The distribution of all datasets used in this study is summarized in Table 1.

Metrics. This study employs five widely recognized evaluation metrics to evaluate COD performance. These metrics provide a comprehensive assessment criterion for analyzing detection accuracy and effectiveness across different models. The Structure-measure (S_α) [7], weighted F-measure (F_β^w) [34], Mean Absolute Error (M) [39], E-measure (E_ϕ) [8], and F-measure (F_β) [1]. The S_α metric quantifies the structural similarity between prediction and GT maps. The F_β^w represents an enhanced evaluation metric that extends the traditional F_β by incorporating spatial weights, providing a better assessment of segmentation quality with emphasis on boundary accuracy and location-based importance of detected pixels. The M metric focuses on pixel-level error evaluation between the normalized prediction and GT. The E_ϕ metric simultaneously evaluates the global and local accuracy of COD based on human visual perception mechanisms. The F_β provides a synthetic measure that considers both precision and recall components. For both F-measure and E-measure metrics, different scores can be obtained according to different precision-recall pairs. This leads to the computation of mean F-measure (F_β^{mean}). Similarly, the E-measure utilizes mean variants, denoted as E_ϕ^{mean}, which are also employed as evaluation metrics.

Training details. This subsection summarizes the training configuration used across all COD techniques used to compare with AINet. Table 2 reports the optimizer, learning rate, batch size, number of epochs, scheduler, and loss functions employed per model.

IV. RESULTS

This section presents a comprehensive evaluation of AINet, focusing first on standard SOTA COD benchmarks and subsequently on agricultural datasets as a specialized case study. Both quantitative and qualitative analyses are provided, along with ablation studies to assess the contribution of key architectural components.

A. QUANTITATIVE EVALUATION

Table 3 introduces the deployment-oriented efficiency evaluation, reporting FLOPs, inference time, and parameters. Under a consistent input resolution, these results provide a practical view of AINet's computational footprint relative to representative COD baselines.

To validate the level of generalization, the proposed AINet has been compared with 26 SOTA COD models, which are listed in Table 4 (*1st column*). All prediction results for CAMO, COD10k, and NC4K datasets are either directly obtained from the original papers for fair evaluation. As shown in Table 4, the proposed AINet architecture outperforms all 26 SOTA techniques across all three benchmark datasets and evaluation metrics. Taking into account the top 3 best techniques in different metrics, AINet surpasses techniques such as SINet-V2 [9], SegMaR [20], BGNet [3], OCENet [29], DGNet [18], FEDER [12], and UJSCOD-V2 [27]. Among the SOTA techniques, AINet outperforms recent approaches such as GenSAM [15], which employs cross-modal chains of thought prompting to generate visual prompts (e.g., CLIP and BLIP2) and progressively produces masks, iteratively refining detection results. Another recent method that AINet surpasses is UCOS-DA [53], which implements source-free unsupervised domain adaptation using DINO through a foreground-background contrastive self-adversarial approach.

Unlike CAMO, COD10k or NC4K, the prediction results for the Cotton Bollworm and Mango datasets are generated using architectures with codes and weights available from the official author pages; for this case, the default configuration, training and inference code published by the authors are followed.

AINet demonstrates exceptional performance for pest detection applications, as evidenced by Table 5, which shows experimental results from the proposed AINet architecture compared to existing SOTA COD techniques on the Cotton Bollworm dataset. AINet achieves remarkable metrics with a structure measure (S_α) of 0.8787, F-measure (F_β^w) of 0.8428, and the lowest MAE (M) at 0.0132. Notably, our approach excels in enhanced boundary detection with the highest E-measure scores (E_ϕ^{adp} of 0.9691 and E_ϕ^{mean} of 0.9727) and F-measure values (F_β^{adp} of 0.8330 and F_β^{mean} of 0.8453). These results surpass well-established methods like BASNet [40], SINet-v2 [9], BGNet [3], and PCNet [50], positioning AINet as the most effective solution for cotton bollworm detection, with significant improvements in both accuracy and boundary precision.

Finally, Table 6 on the Mango dataset used for the harvesting operation. This advanced architecture achieves superior results, securing first place in both S_α (0.8662) and F_β^w (0.8280) metrics, while maintaining second place in F_β^{adp} (0.9545) and E_ϕ^{adp} (0.9558). AINet's consistent top-tier performance across multiple evaluation criteria makes it particularly well-suited for automated fruit detection and harvesting systems, where reliable object identification is crucial for agricultural robotics. The results confirm that AINet's architecture provides the robust and consistent performance required for real-world agricultural scenarios.

B. QUALITATIVE EVALUATION

Figure 3 presents the results of the evaluation of our AINet and six previous SOTA techniques on classical COD benchmark datasets-i.e., CAMO-Test, COD10K-Test, and NC4K. To visually identify which technique presents the best performance, the GT is compared with the predicted mask for each image. Successful matches between GT and predicted masks are painted with white color; False positive regions



FIGURE 2. Example images of the benchmark datasets (CAMO [25], [47], COD10K [9], [10], NC4K [32]) and agricultural datasets (Cotton Bollworm [37] and Mango [24]) used as case studies.

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
BASNet [40]	Adam	1e-3	8	1000	ReduceLROnPlateau	BCE + SSIM + IOU (multi-stage fusion)
SINet-v2 [9]	Adam	1e-4	16	150	Custom (Adjust LR)	Structure loss (weighted BCE + weighted IOU)
BGNet [3]	Adam	1e-4	12	100	Custom (Poly LR)	Structure loss (weighted BCE + weighted IOU) + Dice loss (edge)
C^2F-Net [2]	AdaXW	1e-4	32	50	Custom (Poly LR)	Structure loss (weighted BCE + weighted IOU)
OCENet [29]	Adam	1e-5	4	50	StepLR	Uncertainty aware structure loss (weighted BCE + weighted IOU)
DGNet [18]	AdamW	5e-5	16	150	CosineAnnealingLR	Hybrid loss (weighted BCE + weighted IOU) + MSE loss (grad)
PCNet [50]	AdamW	1e-4	8	150	Custom (Adjust LR)	Structure loss (weighted BCE + weighted IOU)
AINet (Ours)	AdamW	1e-4	16	150	CosineAnnealingLR	Structure loss (weighted BCE + weighted IOU)

TABLE 3. Comparison of architectural and efficiency characteristics of competing methods, including publication venue (source and type), input resolution, computational cost (FLOPs), inference time, and number of parameters.

Technique	Source	Source Type	Image Size (px)	FLOPs(G)	Inference time (ms)	# Params. (M)
BASNet [40]	CVPR	Conference	256×256	18.29	58.13	87.06
SINet-v2 [9]	TPAMI	Journal	352×352	4.96	41.20	24.93
BGNet [3]	IJCAI	Conference	416×416	16.93	51.28	77.80
C^2F-Net [2]	TCSVT	Conference	352×352	5.30	38.10	26.36
OCENet [29]	WACV	Conference	352×352	12.22	60.67	58.17
DGNet [18]	MIR	Journal	352×352	2.63	33.27	8.30
PCNet [50]	arXiv	-	352×352	5.65	65.94	27.66
AINet (Ours)	IEEE Access	Journal	352×352	7.78	55.92	36.05

TABLE 4. Experimental results for SOTA COD techniques and the proposed AINet architecture on benchmark datasets. The best three performing results are highlighted using color: First, Second, and Third respectively.

Technique	Pub/Year	Backbone	CAMO-Test (250)				COD10K-Test (2,026)				NC4K (4,121)			
			$S_\alpha \uparrow$	$M \downarrow$	$F_\beta^{mean} \uparrow$	$E_\phi^{mean} \uparrow$	$S_\alpha \uparrow$	$M \downarrow$	$F_\beta^{mean} \uparrow$	$E_\phi^{mean} \uparrow$	$S_\alpha \uparrow$	$M \downarrow$	$F_\beta^{mean} \uparrow$	$E_\phi^{mean} \uparrow$
BASNet [40]	CVPR-19	ResNet34	0.615	0.124	0.503	0.671	0.661	0.071	0.486	0.729	0.696	0.095	0.610	0.762
C^2F-Net [42]	IJCAI-21	Res2Net50	0.796	0.080	0.762	0.854	0.813	0.036	0.723	0.890	0.838	0.049	0.795	0.897
LSR [31]	CVPR-21	ResNet50	0.708	0.105	0.645	0.755	0.760	0.045	0.658	0.831	0.797	0.061	0.758	0.854
MGL [51]	CVPR-21	ResNet50	0.775	0.088	0.726	0.812	0.814	0.035	0.711	0.852	0.833	0.052	0.782	0.867
PFNet [36]	CVPR-21	ResNet50	0.782	0.085	0.746	0.841	0.800	0.040	0.701	0.877	0.829	0.053	0.784	0.887
TINet [57]	AAAI-21	ResNet50	0.781	0.087	0.728	0.836	0.793	0.042	0.679	0.861	0.829	0.055	0.773	0.879
UGTR [49]	ICCV-21	ResNet50	0.785	0.086	0.738	0.823	0.818	0.035	0.712	0.853	0.839	0.052	0.787	0.874
UR-SINet [23]	ACMMM-21	ResNet50	0.741	0.091	0.649	0.804	0.775	0.041	0.643	0.869	0.806	0.057	0.731	0.873
BGNet [3]	IJCAI22	Res2Net50	0.812	0.073	0.789	0.870	0.831	0.033	0.753	0.901	0.851	0.044	0.820	0.907
C^2F-Net-V2 [2]	TCSVT-22	Res2Net50	0.799	0.077	0.770	0.859	0.811	0.036	0.725	0.887	0.840	0.048	0.802	0.896
ERRNet [19]	PR-22	ResNet50	0.779	0.085	0.729	0.842	0.786	0.043	0.675	0.867	0.827	0.054	0.778	0.887
FAP-Net [56]	TIP-22	Res2Net50	0.815	0.076	0.776	0.865	0.822	0.036	0.731	0.888	0.851	0.047	0.810	0.899
FindNet [28]	TIP-22	Res2Net	0.803	0.077	0.763	0.862	0.811	0.036	0.706	0.883	0.841	0.048	0.802	0.895
OCENet [29]	WACV-22	ResNet50	0.802	0.080	0.766	0.852	0.827	0.033	0.741	0.894	0.853	0.045	0.818	0.902
PreyNet [52]	ACMMM-22	ResNet50	0.790	0.077	0.757	0.842	0.813	0.034	0.736	0.881	0.834	0.050	0.803	0.887
SegMaR [20]	CVPR-22	ResNet50	0.815	0.071	0.795	0.874	0.833	0.034	0.757	0.899	0.841	0.046	0.821	0.896
SINet-v2 [9]	TPAMI-22	Res2Net50	0.820	0.070	0.782	0.882	0.815	0.037	0.718	0.887	0.847	0.048	0.805	0.903
DGNet [18]	MIR-23	EfficientNet	0.839	0.057	0.806	0.901	0.822	0.033	0.728	0.896	0.857	0.042	0.814	0.911
FEDER [12]	CVPR-23	ResNet50	0.807	0.069	0.785	0.873	0.823	0.032	0.740	0.900	0.846	0.045	0.817	0.905
LSR-V2 [30]	TCSVT-23	ResNet50	0.789	0.079	0.751	0.840	0.805	0.037	0.711	0.880	0.840	0.048	0.801	0.896
MRR-Net [48]	TNNLS-23	ResNet50	0.811	0.076	0.772	0.869	0.822	0.036	0.730	0.889	0.848	0.049	0.801	0.898
PUENet [54]	TIP-23	ResNet50	0.794	0.080	0.762	0.857	0.813	0.035	0.727	0.887	0.836	0.050	0.798	0.892
UCOS-DA [53]	ICCVW-23	DINO	0.701	0.127	0.646	0.784	0.689	0.086	0.546	0.740	0.755	0.085	0.689	0.819
UJSCOD-V2 [27]	arXiv-23	ResNet50	0.803	0.071	0.768	0.858	0.817	0.033	0.733	0.895	0.856	0.040	0.824	0.913
WS-SAM [13]	NeurIPS-23	ResNet50	0.759	0.092	0.742	0.818	0.803	0.038	0.719	0.878	0.829	0.052	0.802	0.886
GenSAM [15]	AAAI-24	CLIP, BLIP2	0.719	0.113	0.659	0.775	0.775	0.067	0.681	0.838	—	—	—	—
AINet (Ours)	-	PVTv2	0.840	0.057	0.815	0.909	0.835	0.028	0.757	0.913	0.868	0.036	0.838	0.926

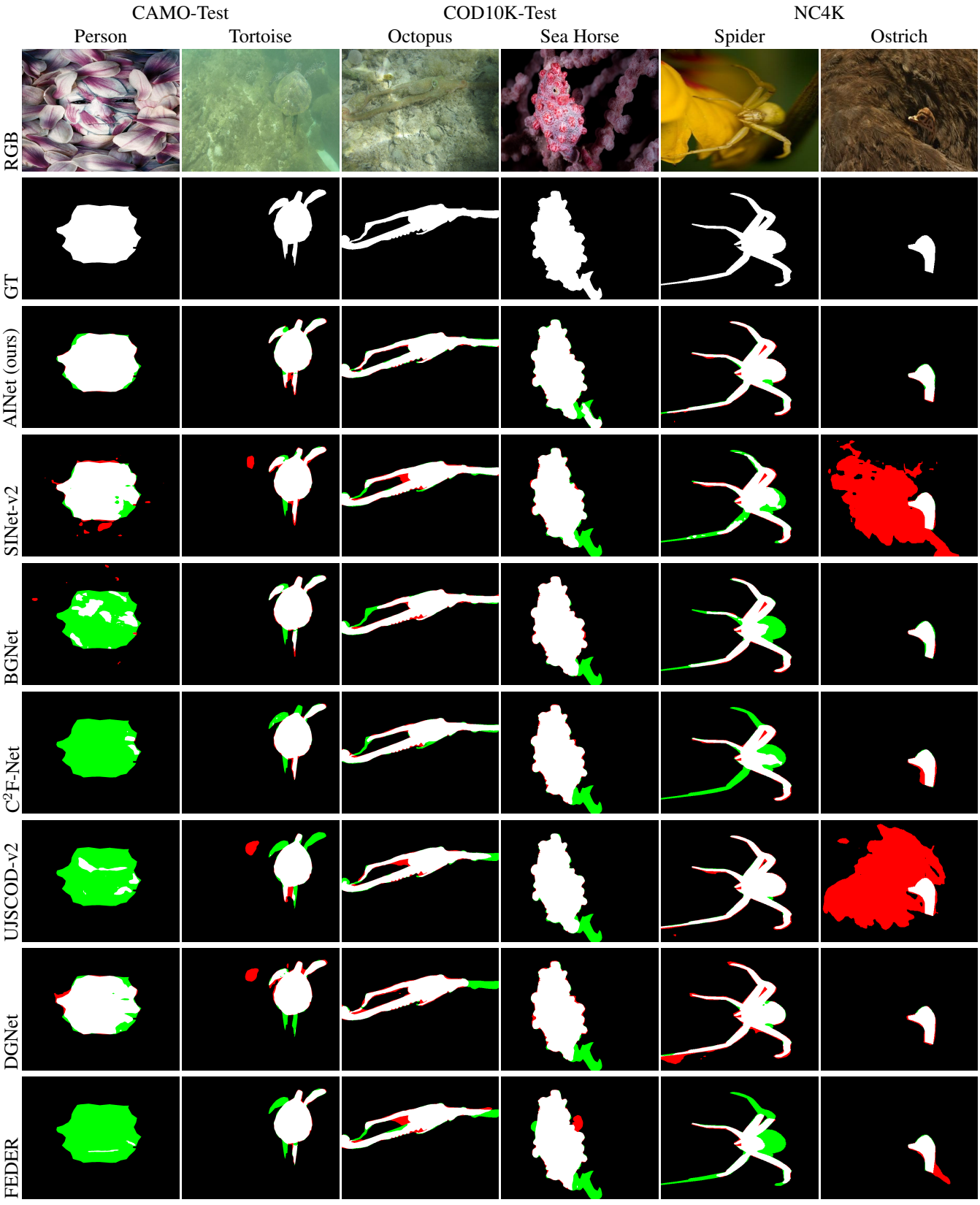

FIGURE 3. Quantitative results for six SOTA COD techniques and the proposed AINet architecture, evaluated on example images from benchmark datasets. Successful matches between GT and predicted masks (white areas); False positive regions (red areas, over-segmentation); and false negative regions (green areas, miss-segmentation).

with red color (over-segmentation); and false negative regions with green color (miss-segmentation). As can be seen in the example images, AINet presents exceptional results in terms of the delimitation of the camouflaged object area, as well as not presenting excessive over-segmentation or miss-segmentation, clearly surpassing recent SOTA COD techniques.

Figure 4 shows the comparative results on the Cotton Bollworm dataset [37]. The third column shows the results of AINet, which demonstrates significantly more accurate detection compared to other methods, correctly identifying the shape and location of cotton bollworms in various environments. AINet is observed to produce sharper and more complete masks, with better edge delineation and fewer false positives. Particularly in difficult cases where the insect is camouflaged with the plant background, AINet maintains high segmentation accuracy, highlighting its effectiveness for this specific application of agricultural pest detection.

On the other hand, the qualitative results on the Mango dataset strongly demonstrate the superiority of AINet as a leading architecture for automated fruit harvesting applications. AINet exhibits exceptional performance in accurately detecting multiple mangoes, achieving defined contours and minimal false positives compared to established methods such as PCNet, DGNet, and other SOTA approaches as shown in Fig. 5.

Finally, AINet's robustness is evident in its ability to maintain consistent and accurate detections under diverse lighting conditions, from bright natural light to shadowed areas, while producing significantly cleaner detection maps with fewer artifacts and noise than its competitors. The edges detected by AINet are noticeably smoother and more precise. While other methods show considerable variability in their results, AINet maintains superior stability across different scenarios, establishing itself as the most reliable and effective solution for real-world applications.

V. ABLATION STUDY

To systematically assess the contribution of each component within AINet, an extensive ablation analysis is performed on three standard COD benchmarks (CAMO, COD10K, and NC4K) as well as two domain-specific agricultural datasets (Cotton Bollworm and Mango). On the benchmark datasets, the study examines: (i) the effect of integrating Mamba and CBAM (module combinations), (ii) the impact of multi-level deep supervision through different output-fusion strategies, and (iii) model interpretability via Grad-CAM visualizations computed from the segmentation heads.

CAMO ablations. Table 7 reports the performance obtained with different module combinations on CAMO. The complete design (Mamba + CBAM) provides the best overall results, improving segmentation accuracy and reducing the pixel-wise error compared with single-module variants. This indicates that long-range dependency modeling (Mamba) and attention-based refinement (CBAM) play complementary roles when dealing with low-contrast camouflage. The qualitative examples in Fig. 6 support this observation, where Mamba + CBAM yields cleaner masks and sharper boundaries than using only Mamba or only CBAM.

Table 8 analyzes different output-fusion settings derived from the multi-level segmentation heads. Although the simple averaging baseline (Avg) attains a relatively high S_α, it is less consistent across complementary measures. In contrast, fusing multiple supervised outputs (notably M1/M2) leads to a more balanced behavior, improving the F-measure while reducing MAE (e.g., M2 achieves $F_\beta^w = 0.7898$ and $M = 0.0566$). Figure 7 visually reinforces these findings: incorporating multiple outputs produces more complete object regions and more coherent contours, which is particularly beneficial for small targets and ambiguous boundaries.

Beyond accuracy, we examine where the network focuses when segmenting camouflaged objects. Figure 8 shows Grad-CAM responses for the segmentation heads on CAMO, where the activation maps consistently concentrate on the true target regions rather than on background clutter. In addition, different heads exhibit complementary attention patterns, with some emphasizing coarse localization and others highlighting finer structures and boundaries. This behavior is consistent with the intended multi-level design and supports the effectiveness of the proposed feature refinement strategy.

COD10K ablations. Table 9 reports the same module-combination study on COD10K. The Mamba + CBAM configuration remains the strongest, outperforming both only-Mamba and only-CBAM across the evaluated metrics. Qualitative comparisons in Fig. 9 confirm that the combined design better preserves object completeness and reduces missed regions on highly camouflaged samples, demonstrating that the observed synergy extends to large-scale benchmark data.

Table 10 evaluates the role of deep supervision through different output-fusion strategies. While Avg achieves a high S_α, richer fusion with multi-level supervision (particularly M1) yields improved segmentation quality and lower error (M1: $F_\beta^w = 0.7371$, $M = 0.0283$ vs. Avg: $F_\beta^w = 0.7145$, $M = 0.0316$). Figure 10 illustrates that multi-level fusion provides more reliable masks, especially when the target shares texture and color with the background.

Figure 11 provides Grad-CAM visualizations for the segmentation heads on COD10K. The heatmaps remain well aligned with the true object locations even under strong camouflage. Moreover, the heads again show complementary attention: earlier heads tend to capture broader localization cues, whereas deeper heads emphasize boundary-sensitive details, which is consistent with the role of multi-level supervision in improving robustness.

NC4K ablations. Table 11 reports the module ablation on NC4K. As in the previous benchmarks, integrating Mamba + CBAM achieves the best overall performance compared to only-Mamba and only-CBAM. The qualitative examples in Fig. 12 further show that the full configuration improves boundary delineation while reducing both missing parts and spurious detections in complex natural scenes.

Table 12 analyzes different output-fusion variants on

TABLE 5. Quantitative results for SOTA COD techniques and AINet on Cotton Bollworm dataset [37]. The best three performing results are highlighted using color: First, Second, and Third respectively.

Technique	$S_\alpha \uparrow$	$F_\beta^w \uparrow$	$M \downarrow$	$F_\beta^{adp} \uparrow$	$F_\beta^{mean} \uparrow$	$E_\phi^{adp} \uparrow$	$E_\phi^{mean} \uparrow$
BASNet [40]	0.8335	0.7500	0.0242	0.7538	0.7702	0.9270	0.9049
SINet-v2 [9]	0.9010	0.8677	0.0136	0.8483	0.8685	0.9668	0.9662
BGNet [3]	0.8953	0.8769	0.0133	0.8757	0.8864	0.9708	0.9635
C^2F-Net [2]	0.8861	0.7758	0.0178	0.8369	0.8571	0.9483	0.9341
OCENet [29]	0.9071	0.8692	0.0108	0.8431	0.8674	0.9586	0.9622
DGNet [18]	0.9001	0.8725	0.0104	0.8525	0.8747	0.9615	0.9590
PCNet [50]	0.9000	0.8652	0.0125	0.8532	0.8657	0.9651	0.9647
AINet (Ours)	0.9208	0.9197	0.0076	0.9092	0.9156	0.9838	0.9820

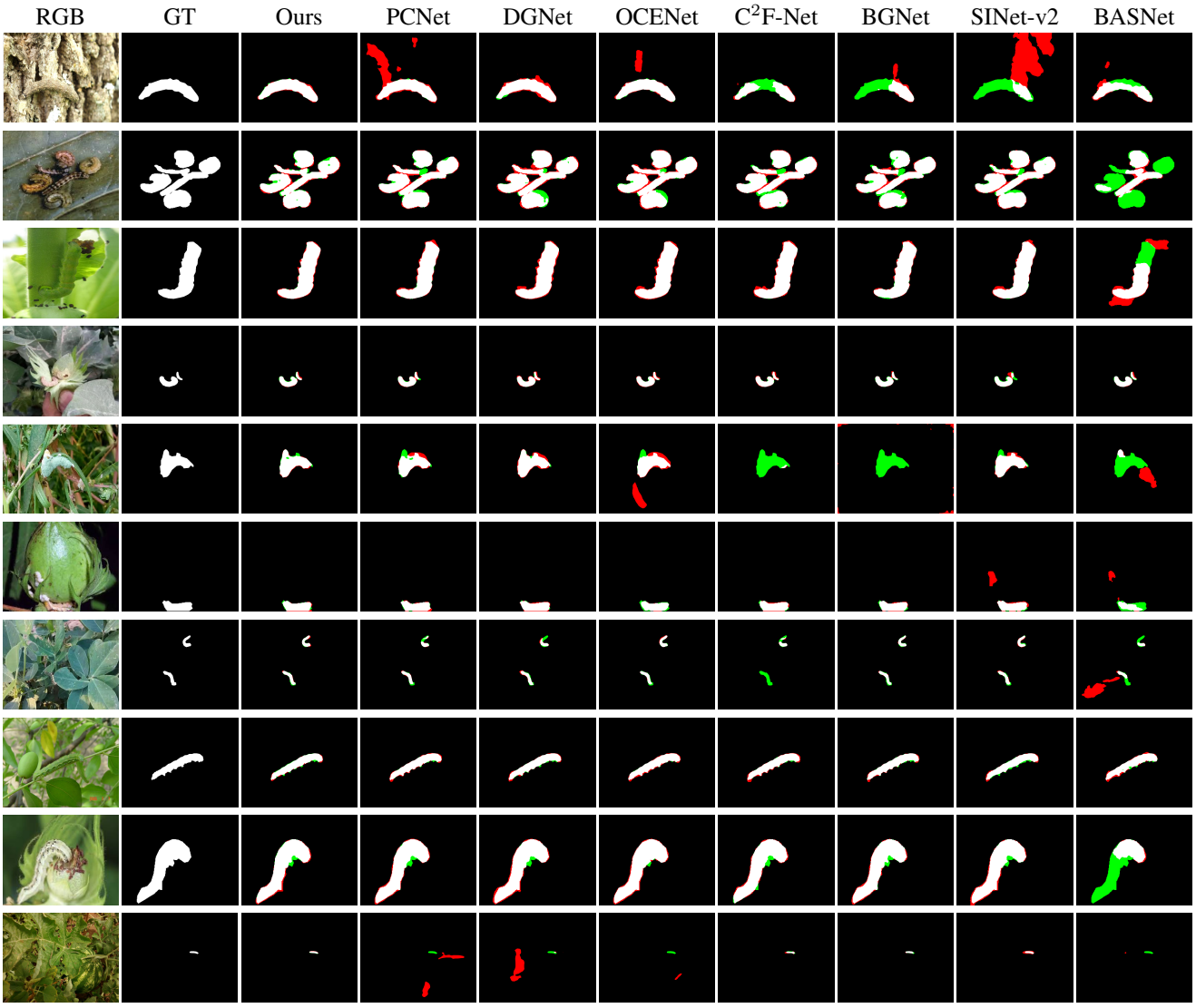

FIGURE 4. Qualitative results of seven SOTA COD techniques and AINet, evaluated on example images from the Cotton Bollworm dataset [37]. Successful matches between GT and predicted masks (white areas); False positive regions (red areas, over-segmentation); and false negative regions (green areas, miss-segmentation).

NC4K. Although Avg obtains a higher S_α, multi-level fusion (M1) offers a better trade-off across metrics by improving F-measure and reducing MAE (M1: $F_\beta^w = 0.8204$, $M = 0.0361$ vs. Avg: $F_\beta^w = 0.8007$, $M = 0.0404$). As shown in Fig. 13, multi-level fusion tends to produce more complete and stable segmentations, particularly for thin structures and low-contrast regions where reduced supervision may lead to fragmented masks.

TABLE 6. Quantitative results for SOTA COD techniques and AINet on Mango dataset [24]. The best three performing results are highlighted using color: First, Second, and Third respectively.

Technique	$S_\alpha \uparrow$	$F_\beta^w \uparrow$	$M \downarrow$	$F_\beta^{adp} \uparrow$	$F_\beta^{mean} \uparrow$	$E_\phi^{adp} \uparrow$	$E_\phi^{mean} \uparrow$
BASNet [40]	0.8508	0.8011	0.0119	0.7954	0.8224	0.9422	0.9336
SINet-v2 [9]	0.8543	0.8194	0.0106	0.7854	0.8290	0.9400	0.9562
BGNet [3]	0.8578	0.8278	0.0117	0.8200	0.8366	0.9513	0.9535
C^2F-Net [2]	0.8659	0.8234	0.0112	0.8167	0.8348	0.9545	0.9532
OCENet [29]	0.8505	0.8094	0.0111	0.8078	0.8301	0.9532	0.9452
DGNet [18]	0.8595	0.8000	0.0116	0.7621	0.8106	0.9284	0.9499
PCNet [50]	0.8582	0.8205	0.0103	0.7996	0.8310	0.9487	0.9558
AINet (Ours)	0.8662	0.8280	0.0100	0.8222	0.8285	0.9545	0.9558

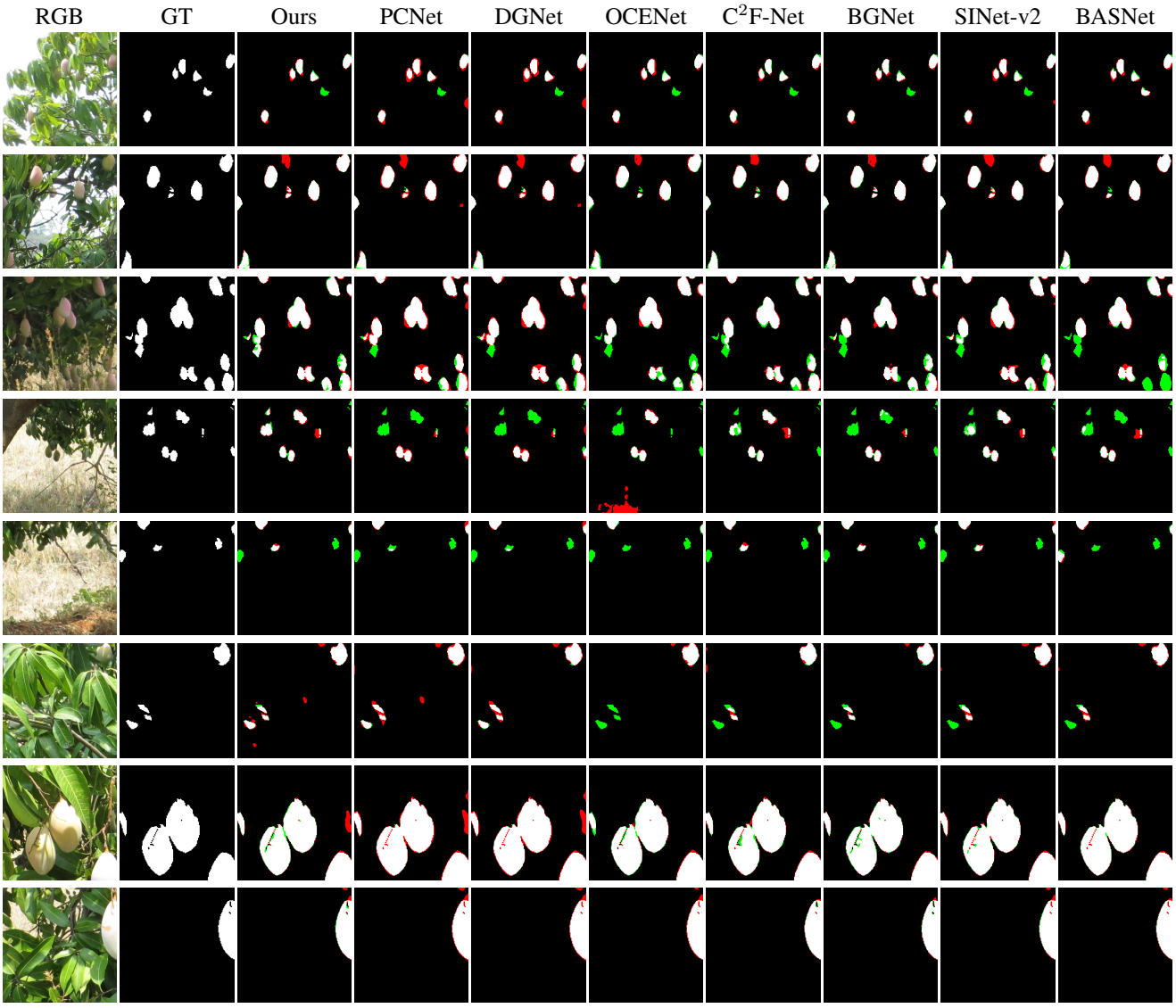

FIGURE 5. Qualitative results of seven SOTA COD techniques and AINet, evaluated on example images from the Mango dataset [24]. Successful matches between GT and predicted masks (white areas); False positive regions (red areas, over-segmentation); and false negative regions (green areas, miss-segmentation).

Figure 14 visualizes Grad-CAM responses for the segmentation heads on NC4K. The activations consistently align with the camouflaged targets across diverse scenes, indicating that predictions are driven by meaningful object cues rather than background artifacts. The complementary evidence across heads further supports the multi-level decoding strategy as a mechanism to aggregate distinct spatial cues for improved final segmentation.

Cotton Bollworm ablations. Table 13 reports the effect of different module combinations on the Cotton Bollworm dataset. The full configuration (Mamba + CBAM) achieves the strongest performance, whereas single-module variants are consistently weaker (e.g., only Mamba: S_α = 0.9131; only CBAM: S_α = 0.9118). This gap indicates that Mamba and CBAM contribute complementary benefits in agricultural scenes: Mamba strengthens long-range contextual modeling, while CBAM refines discriminative features to better separate camouflaged pests from visually similar plant textures. The qualitative examples in Fig. 15 reinforce this outcome, where Mamba + CBAM produces sharper boundaries and fewer false positives under challenging camouflage conditions.

Table 14 analyzes the role of multi-level deep supervision through different output-fusion settings. Although Avg attains a slightly higher S_α (0.9261), it is less consistent across the remaining metrics. In contrast, multi-output fusion (e.g., M_1: $P_0 + P_1 + P_2 + P_3 + Avg$) provides a more stable trade-off, achieving the best overall balance (S_α = 0.9208, F_β^w = 0.9197). As illustrated in Fig. 16, multi-level fusion tends to yield more complete and coherent pest regions, whereas reduced-supervision variants (e.g., M_3, M_4) more often lead to fragmented or partially missing segmentations.

Figure 17 presents Grad-CAM visualizations for the segmentation heads on the Cotton Bollworm dataset. The activation maps consistently align with the pest regions, even when insects are partially occluded or strongly blended with the plant background, suggesting that predictions are driven by meaningful object cues rather than background artifacts. In addition, the heads exhibit complementary attention patterns, where some heads emphasize coarse localization and others highlight fine structures and boundary-sensitive areas, supporting the intended multi-level decoding behavior.

Mango ablations. Table 15 reports the module-combination study on the Mango dataset. The combined design (Mamba + CBAM) again achieves the strongest performance (e.g., S_α = 0.8662, F_β^w = 0.8280), outperforming single-module alternatives. The qualitative results in Fig. 18 further show that integrating both modules improves mask completeness and preserves fruit contours under variable illumination and cluttered foliage.

Table 16 evaluates multi-level deep supervision through different output-fusion strategies on Mango. Multi-output fusion provides a clear advantage over simple averaging, improving robustness and yielding stronger overall segmentation quality (e.g., best $S_\alpha = 0.8662$ vs. Avg with $S_\alpha = 0.8478$). Figure 19 corroborates this trend, where multi-level fusion reduces both over-segmentation and under-segmentation in difficult cases with strong background similarity and partial occlusions.

Figure 20 provides Grad-CAM visualizations for the segmentation heads on the Mango dataset. The heatmaps remain well aligned with fruit regions across challenging scenarios, indicating that the model focuses on target-relevant cues. Similar to the other datasets, the heads show complementary attention behavior: earlier heads capture broader localization patterns, while deeper heads emphasize finer details and boundary regions, which is consistent with the purpose of multi-level supervision and staged decoding.

In summary, the ablation study demonstrates that (i) integrating Mamba and CBAM and (ii) employing multi-level deep supervision are both key to achieving strong and stable performance. Their combined effect yields more accurate, reliable, and generalizable camouflaged object segmentation on standard benchmarks and in challenging agricultural imagery.

VI. DISCUSSION

The experimental results demonstrate that AINet achieves state-of-the-art performance on both standard COD benchmarks and specialized agricultural datasets. The model's consistent superiority across diverse datasets underscores the effectiveness of integrating sequential state modeling (Mamba) with attention-based feature refinement (CBAM). This synergy enables AINet to capture long-range dependencies and focus on subtle visual cues, which are essential for detecting camouflaged objects in complex backgrounds.

The strong results on agricultural datasets highlight AINet's adaptability to real-world applications beyond generic COD tasks. Accurate detection of camouflaged pests and fruits is vital for precision agriculture, where reliable object identification supports automated monitoring and harvesting systems. The case study results suggest that AINet can be effectively deployed in such domains, offering both high accuracy and robustness under challenging conditions.

Despite these advances, some challenges remain. The reliance on large annotated datasets for training may limit the applicability of AINet in domains with scarce labeled data. Furthermore, while the model demonstrates strong generalization, future work should explore its performance in even more diverse and dynamic environments, as well as investigate strategies for reducing data requirements and improving real-time inference.

Despite the strong overall performance, AINet can still fail in extremely challenging camouflage scenarios. Figure 21 shows representative failure cases in which the target object exhibits near-identical texture/color to the background or appears very small and fragmented, making boundary cues ambiguous. In such cases, the model may produce false negatives (missed regions) due to insufficient discriminative contrast, or false positives when background structures mimic object-like patterns. These observations indicate that COD remains fundamentally difficult when both intrinsic similarity and edge disruption are simultaneously extreme. Addressing these cases may require incorporating additional cues (e.g., multi-modal data, stronger boundary priors, or specialized augmentation) and further improving robustness under scarce-label regimes.

In summary, AINet sets a new benchmark for camouflaged object detection, offering a versatile and high-performing solution for both academic research and practical deployment in fields such as agriculture, surveillance, and environmental monitoring.

TABLE 7. Metric evaluation with different module combinations (e.g., Mamba and CBAM) on the CAMO dataset [25], [47]. The best results by each metric are highlighted in bold.

Module	$S_\alpha \uparrow$	$F_\beta^w \uparrow$	$M \downarrow$	$F_\beta^{adp} \uparrow$	$F_\beta^{mean} \uparrow$	$E_\phi^{adp} \uparrow$	$E_\phi^{mean} \uparrow$
Mamba + CBAM	**0.8380**	**0.7896**	**0.0567**	**0.8149**	**0.8144**	**0.9104**	**0.9080**
only Mamba	0.8366	0.7861	0.0595	0.8120	0.8116	0.9079	0.9057
only CBAM	0.8285	0.7782	0.0622	0.8112	0.8113	0.9008	0.8983

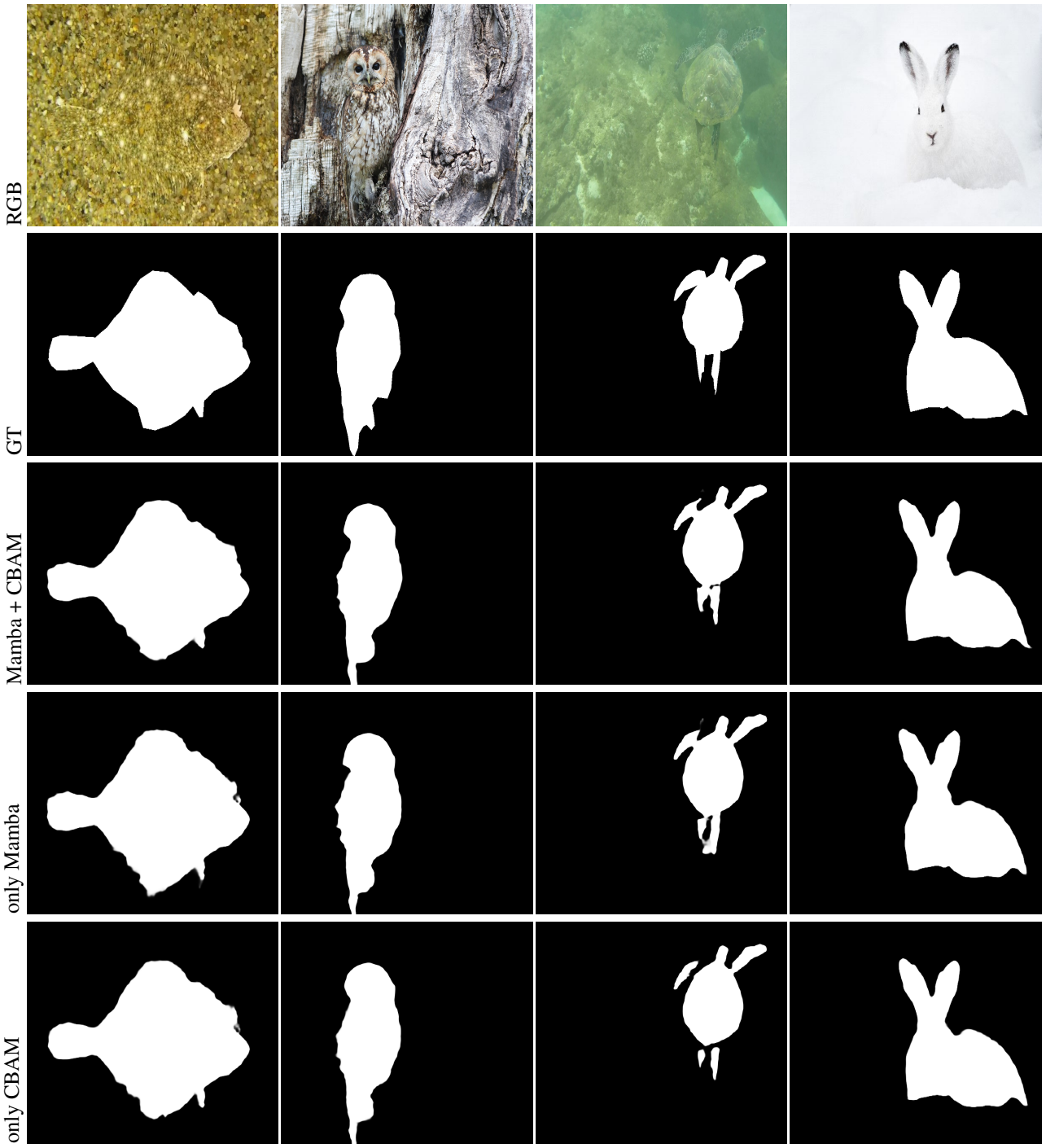

FIGURE 6. Qualitative results of different module combinations used, evaluated on example images from the CAMO dataset [25], [47].

TABLE 8. Metric evaluation with different outputs on the CAMO dataset [25], [47]. The best three performing results are highlighted using color: First, Second, and Third respectively.

Name	Output	$S_\alpha \uparrow$	$F_\beta^w \uparrow$	$M \downarrow$	$F_\beta^{adp} \uparrow$	$F_\beta^{mean} \uparrow$	$E_\phi^{adp} \uparrow$	$E_\phi^{mean} \uparrow$
Avg	*Avg*	0.8489	0.7669	0.0632	0.7962	0.7990	0.8964	0.8899
M_1	$P_0 + P_1 + P_2 + P_3 + Avg$	0.8380	0.7896	0.0567	0.8149	0.8144	0.9104	0.9080
M_2	$P_1 + P_2 + P_3 + Avg$	0.8404	0.7898	0.0566	0.8147	0.8150	0.9101	0.9079
M_3	$P_2 + P_3 + Avg$	0.8419	0.7879	0.0570	0.8125	0.8139	0.9086	0.9069
M_4	$P_3 + Avg$	0.8445	0.7835	0.0581	0.8080	0.8111	0.9055	0.9038

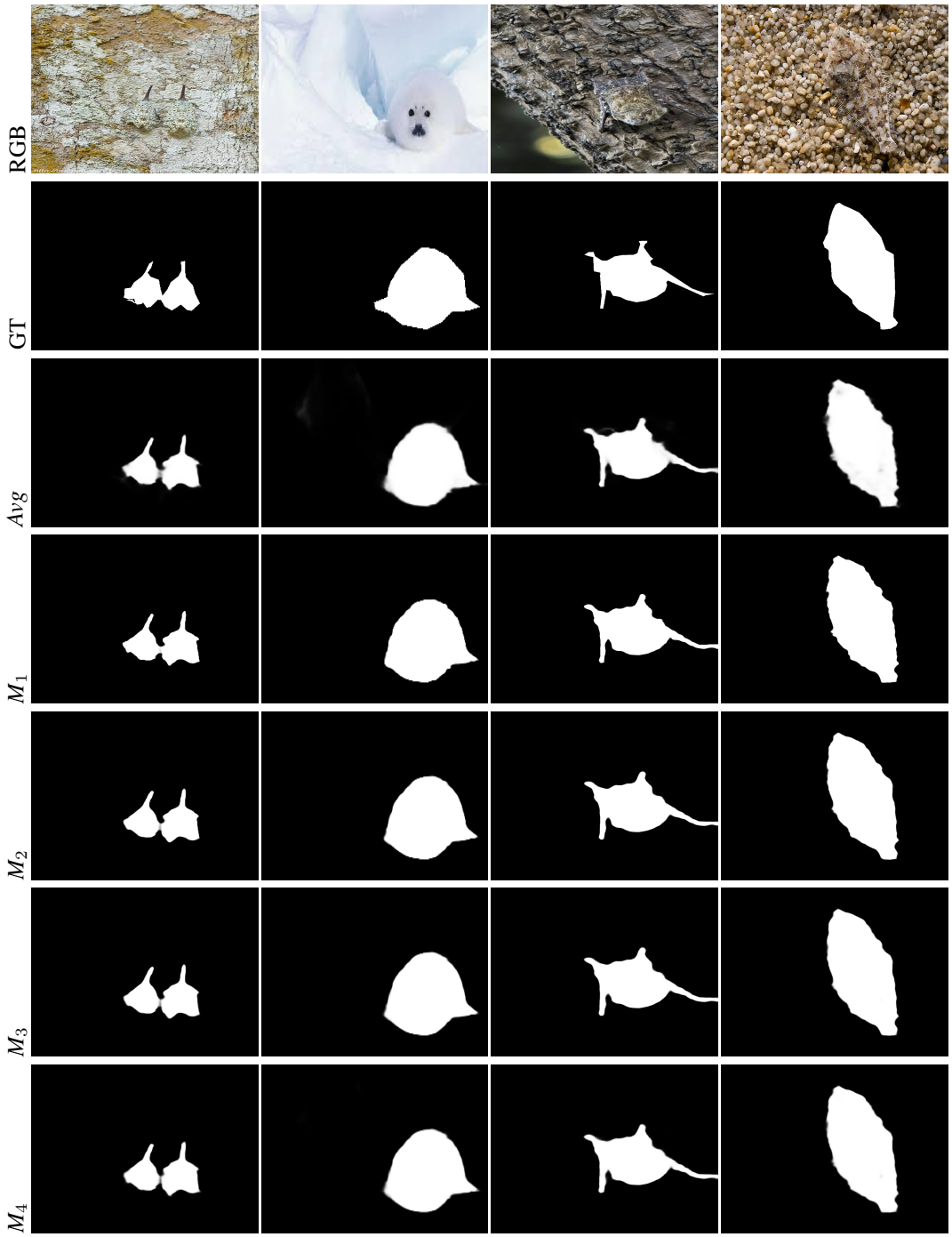

FIGURE 7. Qualitative results of different deep supervision outputs, evaluated on example images from the CAMO dataset [25], [47].

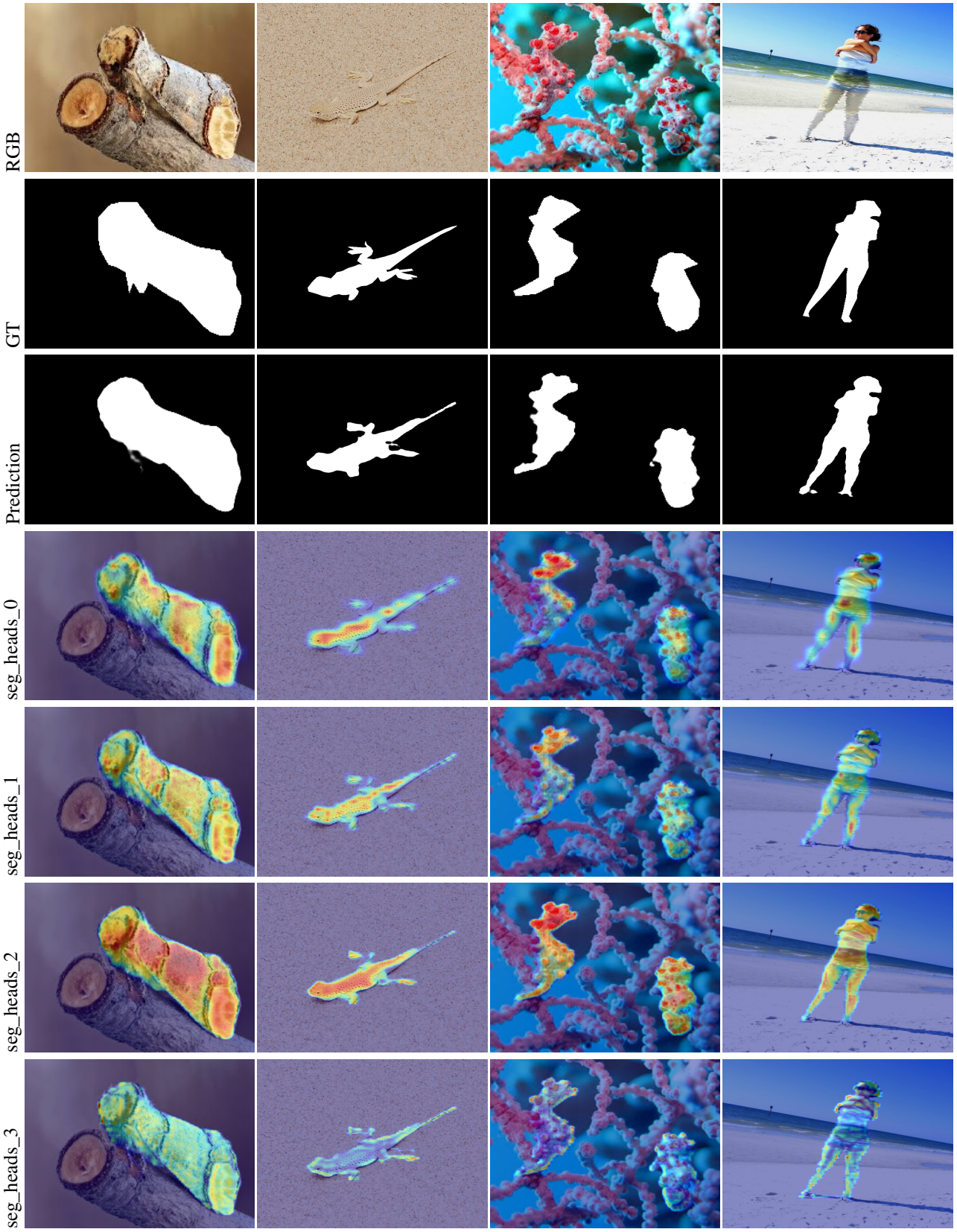

FIGURE 8. Grad-CAM visualization using segmentation heads outputs, evaluated on example images from the CAMO dataset [25], [47].

TABLE 9. Metric evaluation with different module combinations (e.g., Mamba and CBAM) on the COD10K dataset [9], [10]. The best results by each metric are highlighted in bold.

Module	$S_\alpha \uparrow$	$F_\beta^w \uparrow$	$M \downarrow$	$F_\beta^{adp} \uparrow$	$F_\beta^{mean} \uparrow$	$E_\phi^{adp} \uparrow$	$E_\phi^{mean} \uparrow$
Mamba + CBAM	**0.8352**	**0.7371**	**0.0283**	**0.7522**	**0.7568**	**0.9110**	**0.9131**
only Mamba	0.8267	0.7222	0.0316	0.7374	0.7420	0.8990	0.9010
only CBAM	0.8289	0.7242	0.0307	0.7430	0.7499	0.9030	0.9070



FIGURE 9. Qualitative results of different module combinations used, evaluated on example images from the COD10K dataset [9], [10].

TABLE 10. Metric evaluation with different outputs on COD10k dataset [9], [10]. The best three performing results are highlighted using color: First, Second, and Third respectively.

Name	Output	$S_\alpha \uparrow$	$F_\beta^w \uparrow$	$M \downarrow$	$F_\beta^{adp} \uparrow$	$F_\beta^{mean} \uparrow$	$E_\phi^{adp} \uparrow$	$E_\phi^{mean} \uparrow$
Avg	Avg	0.8430	0.7145	0.0316	0.7135	0.7453	0.8794	0.8992
M_1	$P_0 + P_1 + P_2 + P_3 + Avg$	0.8352	0.7371	0.0283	0.7522	0.7568	0.9110	0.9131
M_2	$P_1 + P_2 + P_3 + Avg$	0.8364	0.7356	0.0285	0.7465	0.7554	0.9078	0.9127
M_3	$P_2 + P_3 + Avg$	0.8368	0.7320	0.0289	0.7395	0.7527	0.9032	0.9111
M_4	$P_3 + Avg$	0.8389	0.7275	0.0295	0.7293	0.7505	0.8952	0.9084

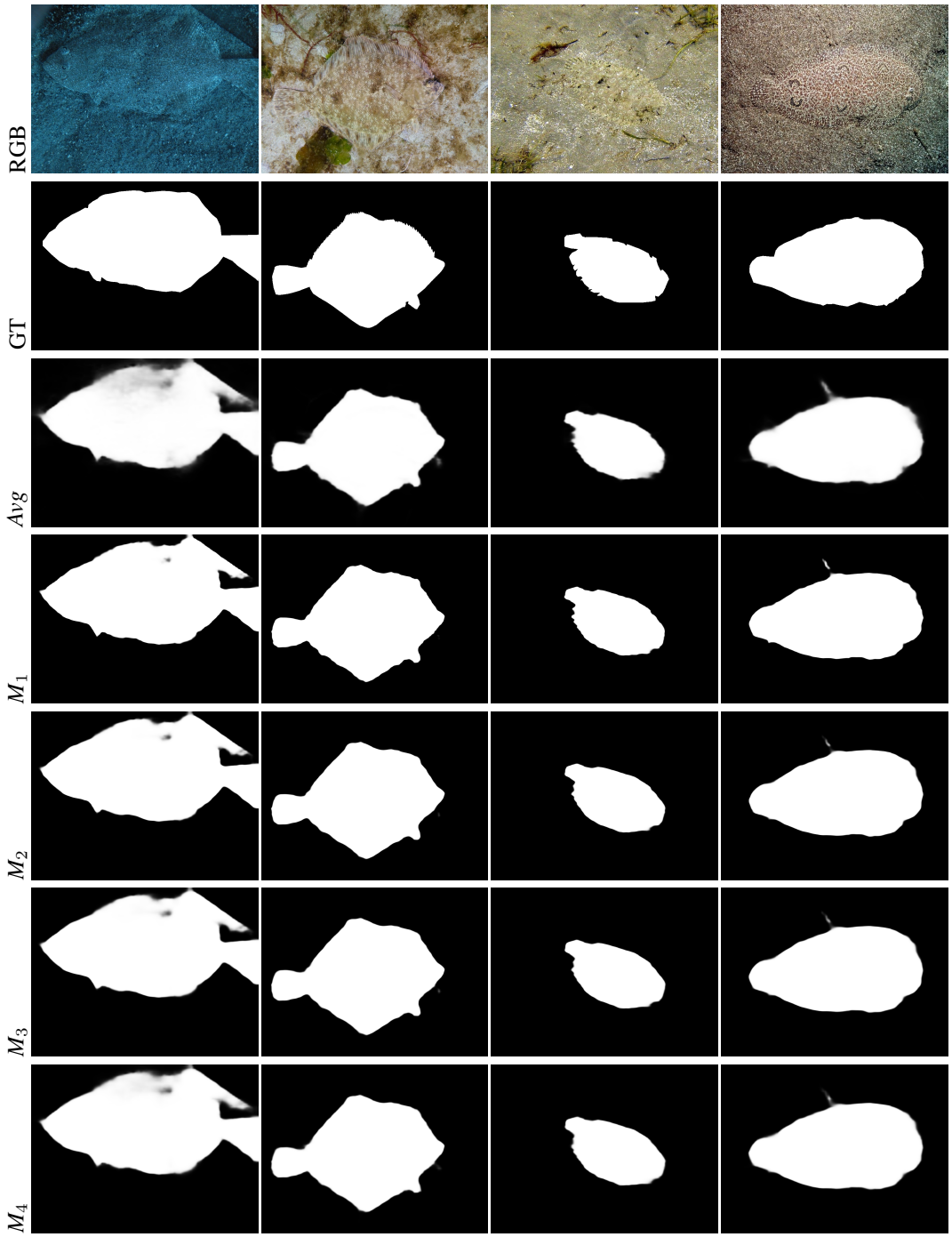

FIGURE 10. Qualitative results of different deep supervision outputs, evaluated on example images from the COD10K dataset [9], [10].

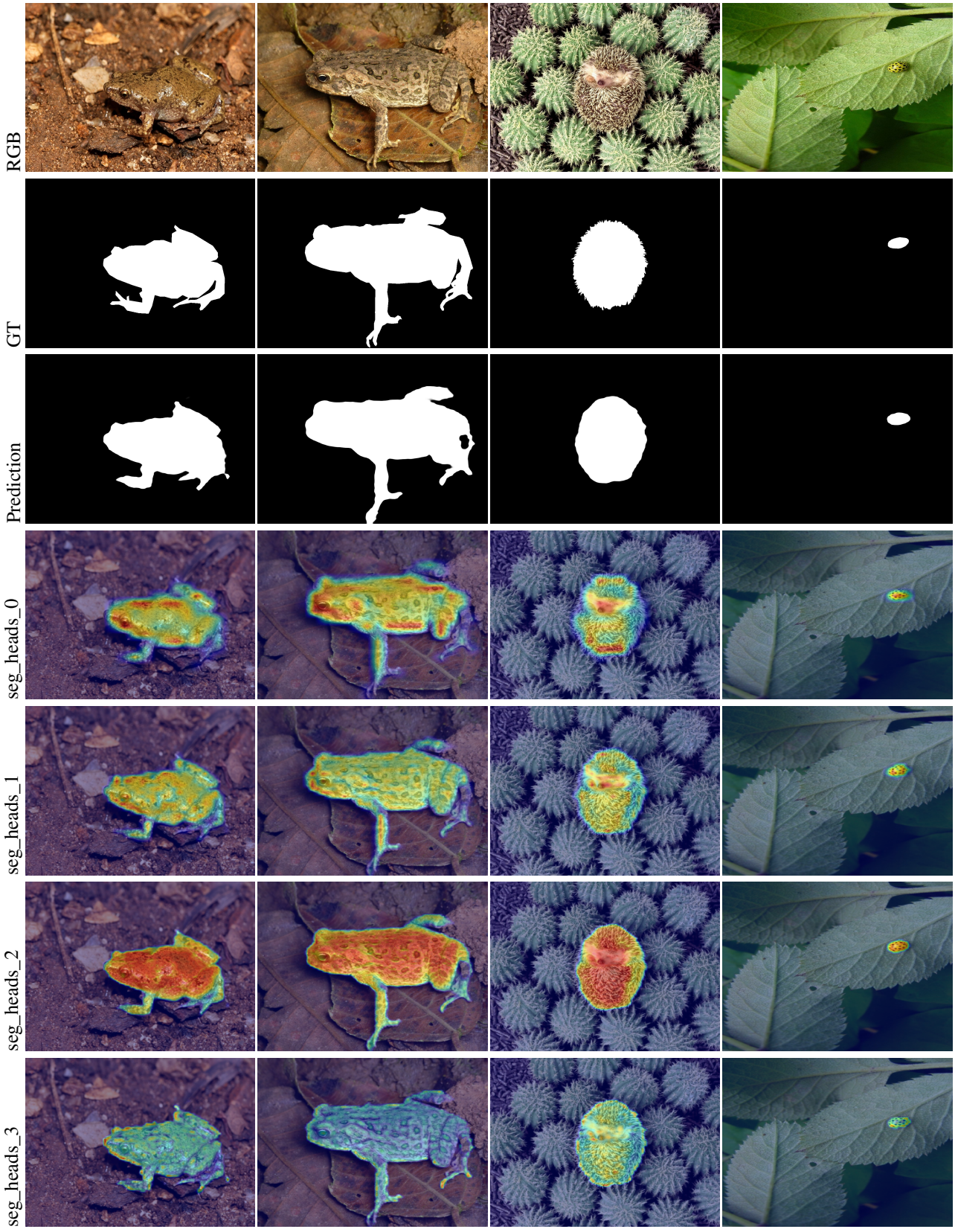

FIGURE 11. Grad-CAM visualization using segmentation heads outputs, evaluated on example images from the COD10K dataset [9], [10].

TABLE 11. Metric evaluation with different module combinations (e.g., Mamba and CBAM) on the NC4K dataset [32]. The best results by each metric are highlighted in bold.

Module	$S_\alpha \uparrow$	$F_\beta^w \uparrow$	$M \downarrow$	$F_\beta^{adp} \uparrow$	$F_\beta^{mean} \uparrow$	$E_\phi^{adp} \uparrow$	$E_\phi^{mean} \uparrow$
Mamba + CBAM	**0.8672**	**0.8204**	**0.0361**	**0.8387**	**0.8392**	**0.9287**	**0.9276**
only Mamba	0.8611	0.8093	0.0390	0.8270	0.8280	0.9207	0.9199
only CBAM	0.8605	0.8078	0.0394	0.8306	0.8323	0.9220	0.9210

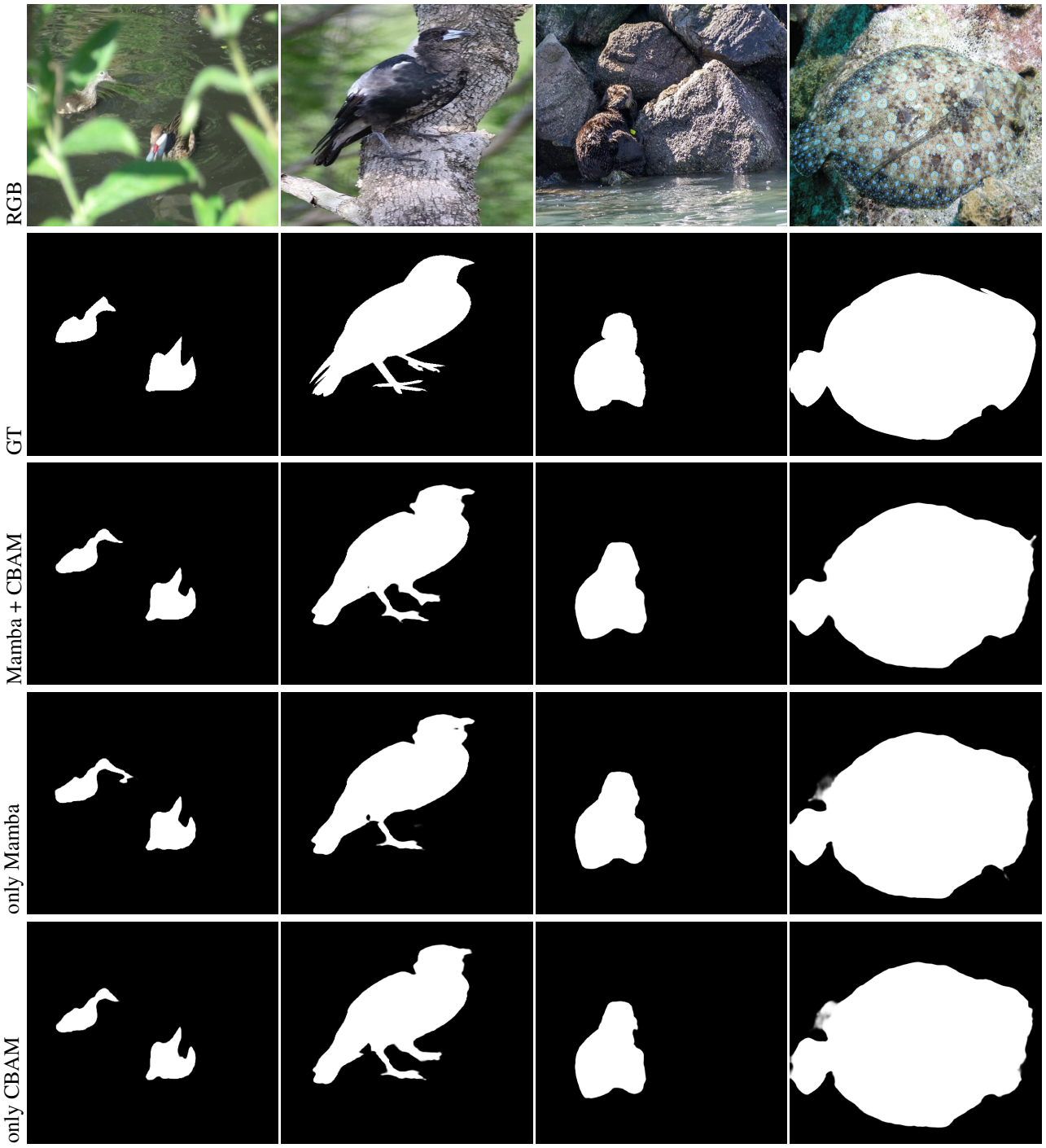

FIGURE 12. Qualitative results of different module combinations used, evaluated on example images from the NC4K dataset [32].

TABLE 12. Metric evaluation with different outputs on NC4K dataset [32]. The best three performing results are highlighted using color: First, Second, and Third respectively.

Name	Output	$S_\alpha \uparrow$	$F_\beta^w \uparrow$	$M \downarrow$	$F_\beta^{adp} \uparrow$	$F_\beta^{mean} \uparrow$	$E_\phi^{adp} \uparrow$	$E_\phi^{mean} \uparrow$
Avg	Avg	0.8756	0.8007	0.0404	0.8165	0.8273	0.9136	0.9151
M_1	$P_0 + P_1 + P_2 + P_3 + Avg$	0.8672	0.8204	0.0361	0.8387	0.8392	0.9289	0.9276
M_2	$P_1 + P_2 + P_3 + Avg$	0.8681	0.8188	0.0365	0.8358	0.8381	0.9275	0.9269
M_3	$P_2 + P_3 + Avg$	0.8689	0.8163	0.0369	0.8322	0.8363	0.9257	0.9259
M_4	$P_3 + Avg$	0.8711	0.8125	0.0377	0.8270	0.8341	0.9225	0.9238

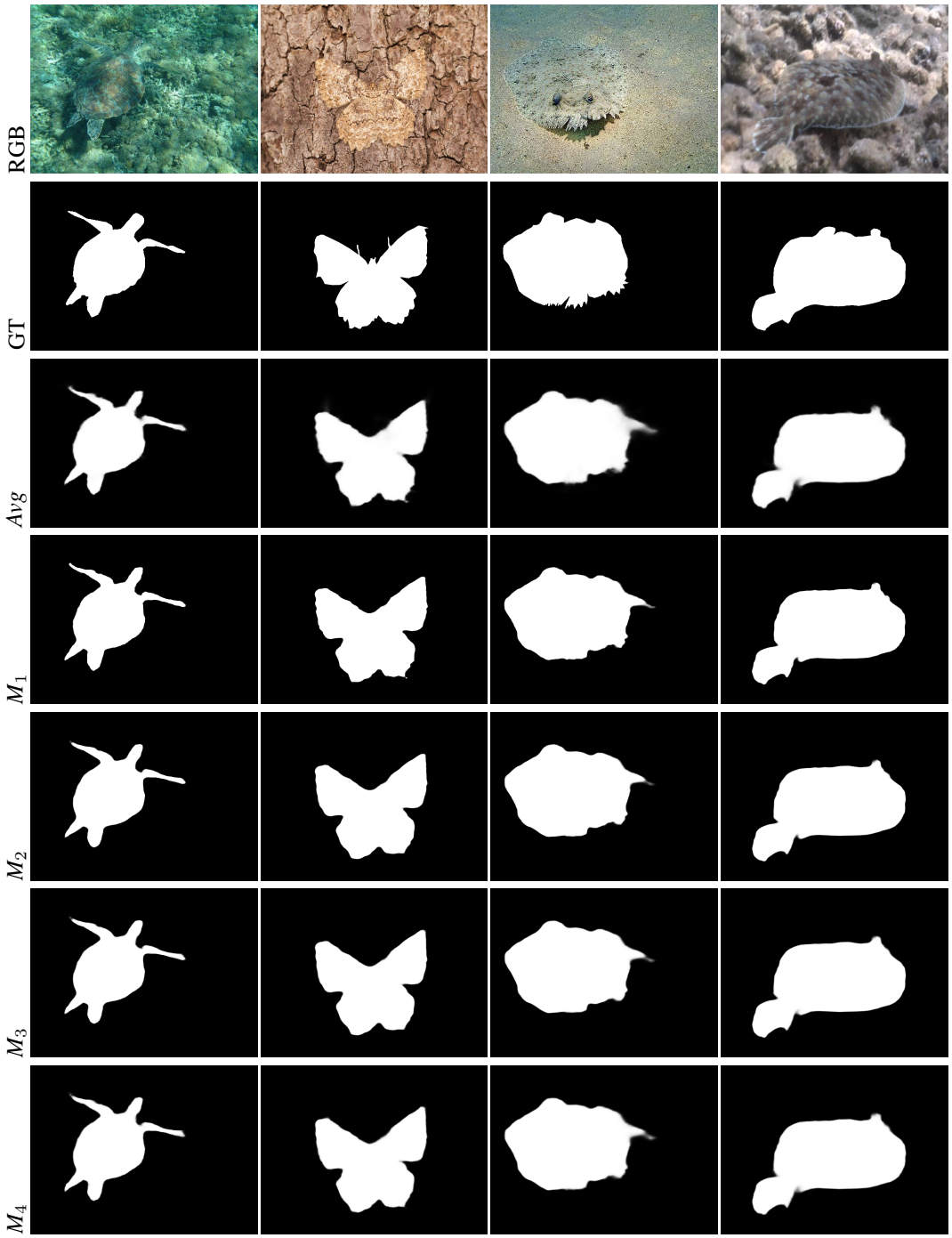

FIGURE 13. Qualitative results of different deep supervision outputs, evaluated on example images from the NC4K dataset [32].

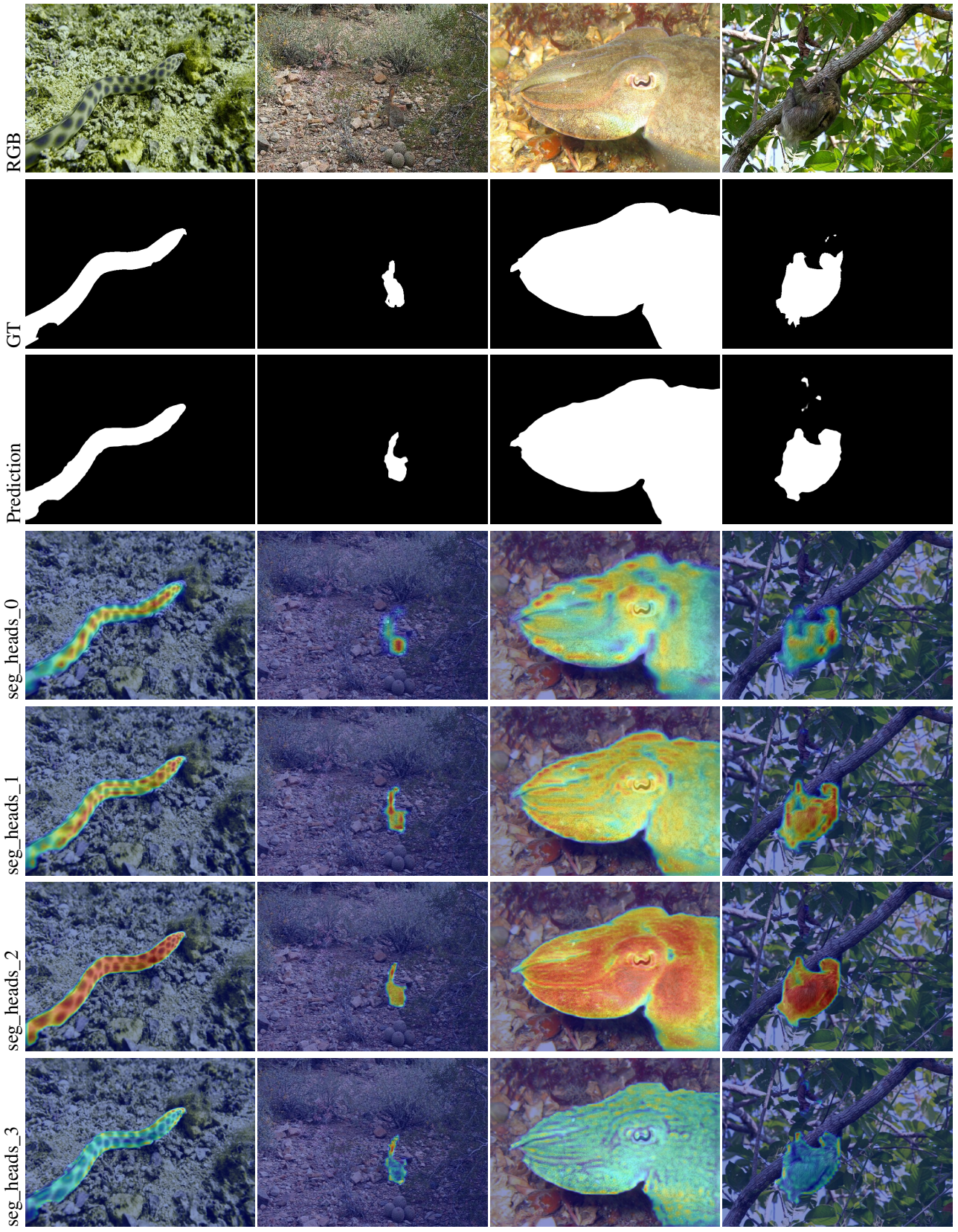

FIGURE 14. Grad-CAM visualization using segmentation heads outputs, evaluated on example images from the NC4K dataset [32].

TABLE 13. Metric evaluation with different module combinations (e.g., Mamba and CBAM) on the Cotton Bollworm dataset [37]. The best results by each metric are highlighted in bold.

Module	$S_\alpha \uparrow$	$F_\beta^w \uparrow$	$M \downarrow$	$F_\beta^{adp} \uparrow$	$F_\beta^{mean} \uparrow$	$E_\phi^{adp} \uparrow$	$E_\phi^{mean} \uparrow$
Mamba + CBAM	**0.9208**	**0.9197**	**0.0076**	**0.9092**	**0.9156**	**0.9838**	**0.9820**
only Mamba	0.9131	0.8953	0.0084	0.8877	0.8932	0.9788	0.9784
only CBAM	0.9118	0.8950	0.0087	0.8877	0.8923	0.9748	0.9744

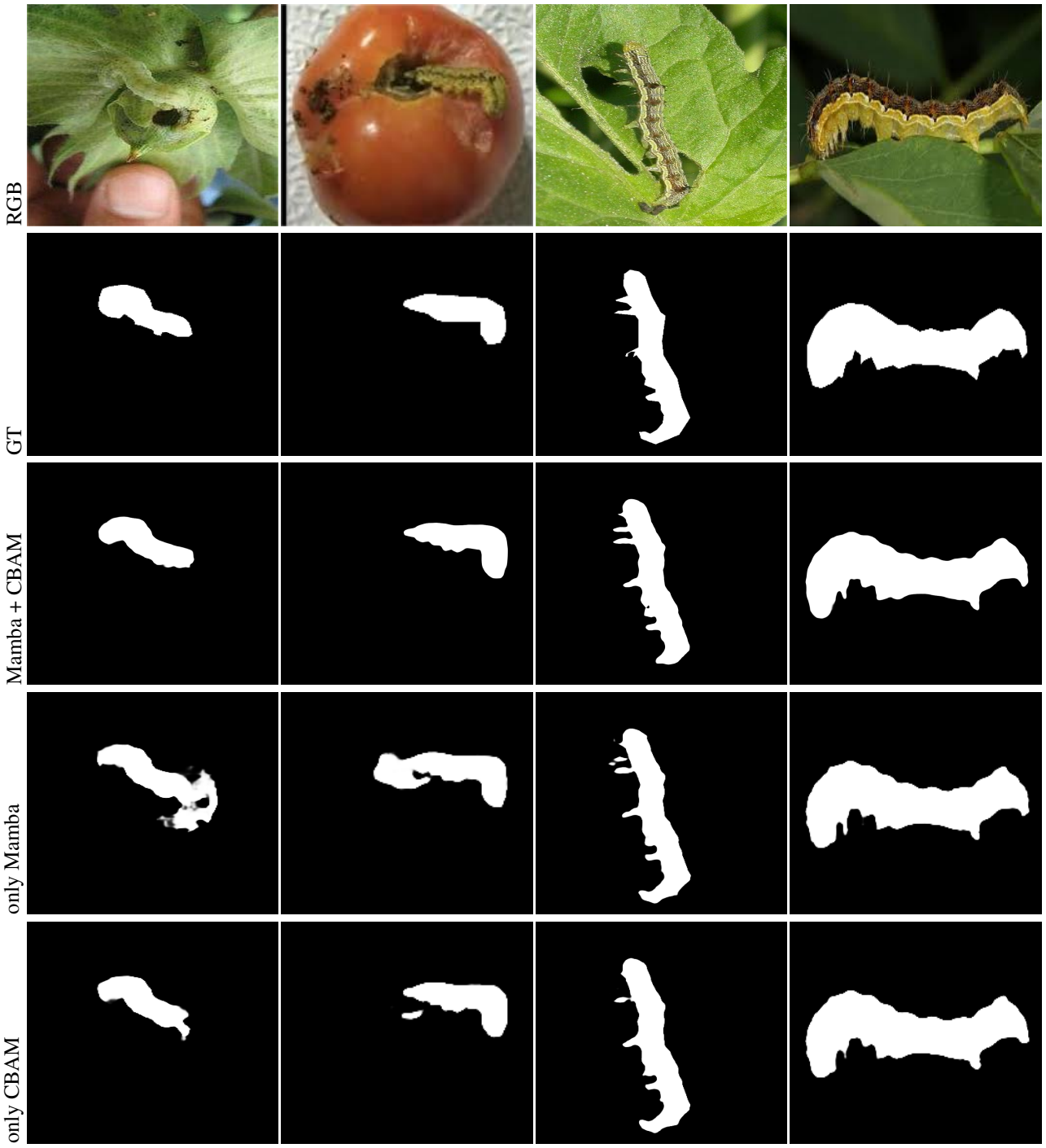

FIGURE 15. Qualitative results of different module combinations used, evaluated on example images from the Cotton Bollworm dataset [37].

TABLE 14. Metric evaluation with different outputs on Cotton Bollworm dataset [37]. The best three performing results are highlighted using color: First, Second, and Third respectively.

Name	Output	$S_\alpha \uparrow$	$F_\beta^w \uparrow$	$M \downarrow$	$F_\beta^{adp} \uparrow$	$F_\beta^{mean} \uparrow$	$E_\phi^{adp} \uparrow$	$E_\phi^{mean} \uparrow$
Avg	Avg	0.9261	0.9137	0.0082	0.8885	0.9121	0.9795	0.9799
M_1	$P_0 + P_1 + P_2 + P_3 + Avg$	0.9208	0.9197	0.0076	0.9092	0.9156	0.9838	0.9820
M_2	$P_1 + P_2 + P_3 + Avg$	0.9206	0.9177	0.0077	0.9054	0.9132	0.9829	0.9817
M_3	$P_2 + P_3 + Avg$	0.9210	0.9161	0.0078	0.9017	0.9117	0.9821	0.9813
M_4	$P_3 + Avg$	0.9215	0.9131	0.0080	0.8946	0.9093	0.9804	0.9807

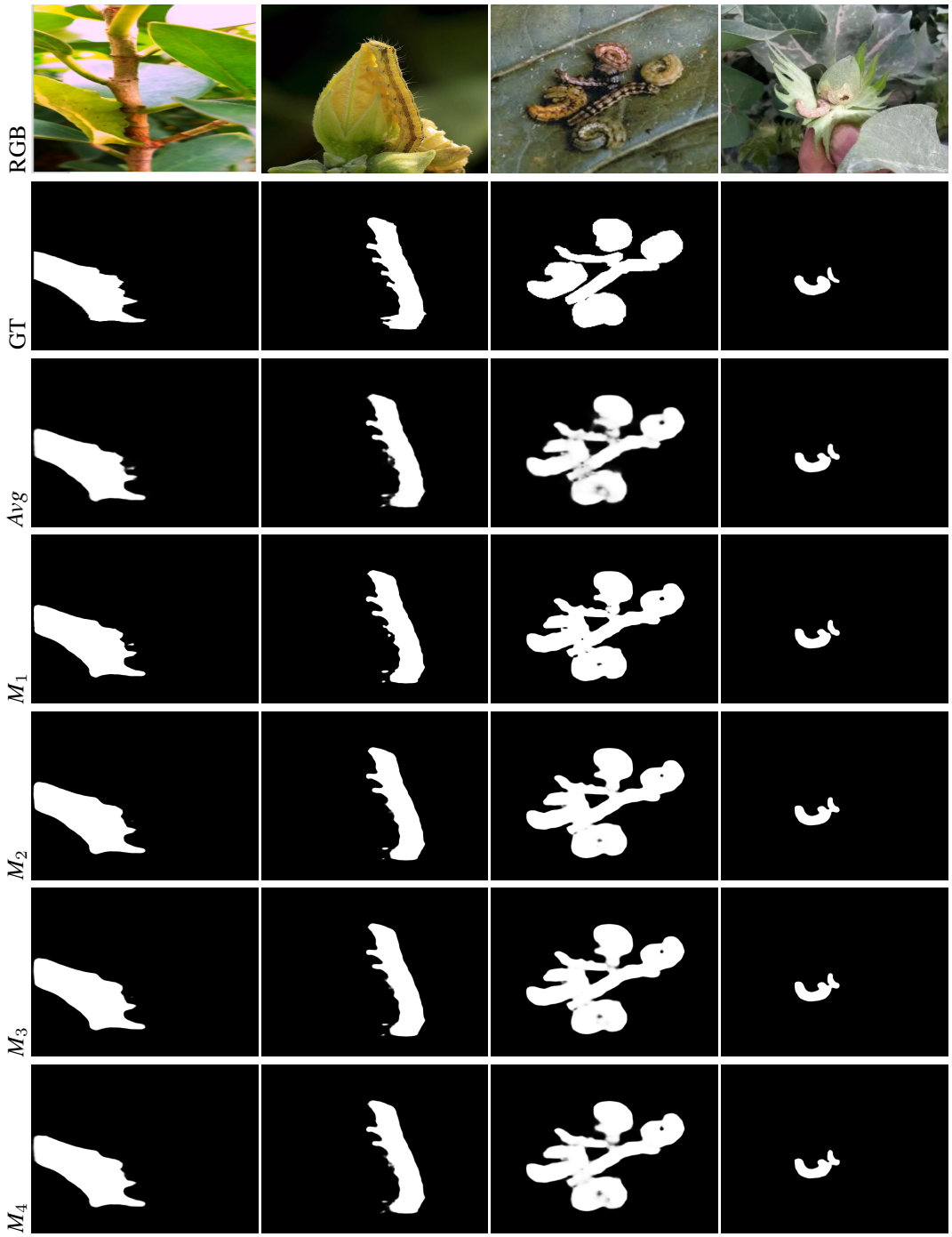

FIGURE 16. Qualitative results of different deep supervision outputs, evaluated on example images from the Cotton Bollworm dataset [37].

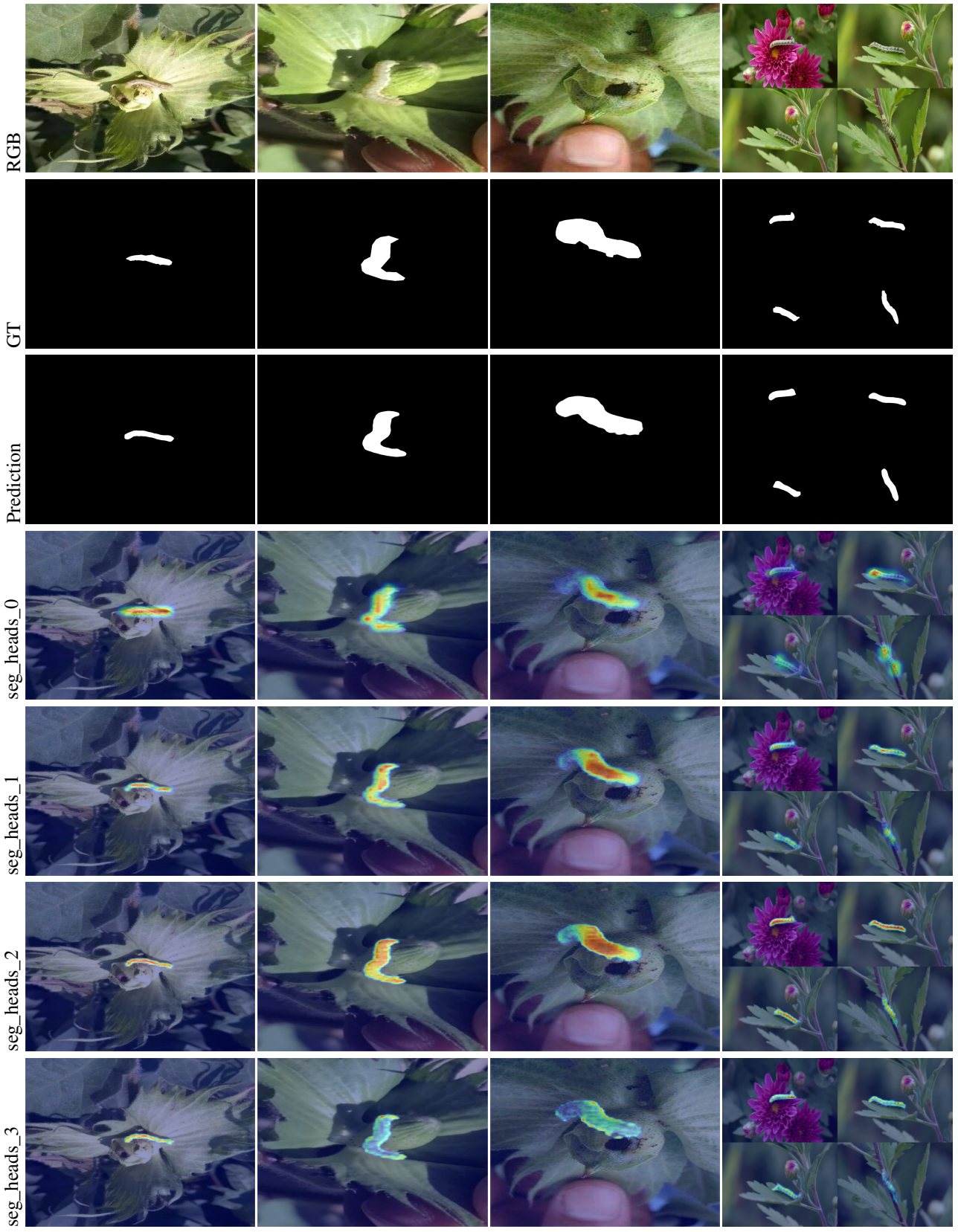

FIGURE 17. Grad-CAM visualization using segmentation heads outputs, evaluated on example images from the Cotton Bollworm dataset [37].

TABLE 15. Metric evaluation with different module combinations (e.g., Mamba and CBAM) on the Mango dataset [24]. The best results by each metric are highlighted in bold.

Module	$S_\alpha \uparrow$	$F_\beta^w \uparrow$	$M \downarrow$	$F_\beta^{adp} \uparrow$	$F_\beta^{mean} \uparrow$	$E_\phi^{adp} \uparrow$	$E_\phi^{mean} \uparrow$
Mamba + CBAM	**0.8662**	**0.8280**	**0.0100**	**0.8222**	**0.8285**	**0.9545**	**0.9558**
only Mamba	0.8646	0.8131	0.0108	0.8113	0.8174	0.9494	0.9507
only CBAM	0.8511	0.7936	0.0119	0.7921	0.8013	0.9457	0.9497

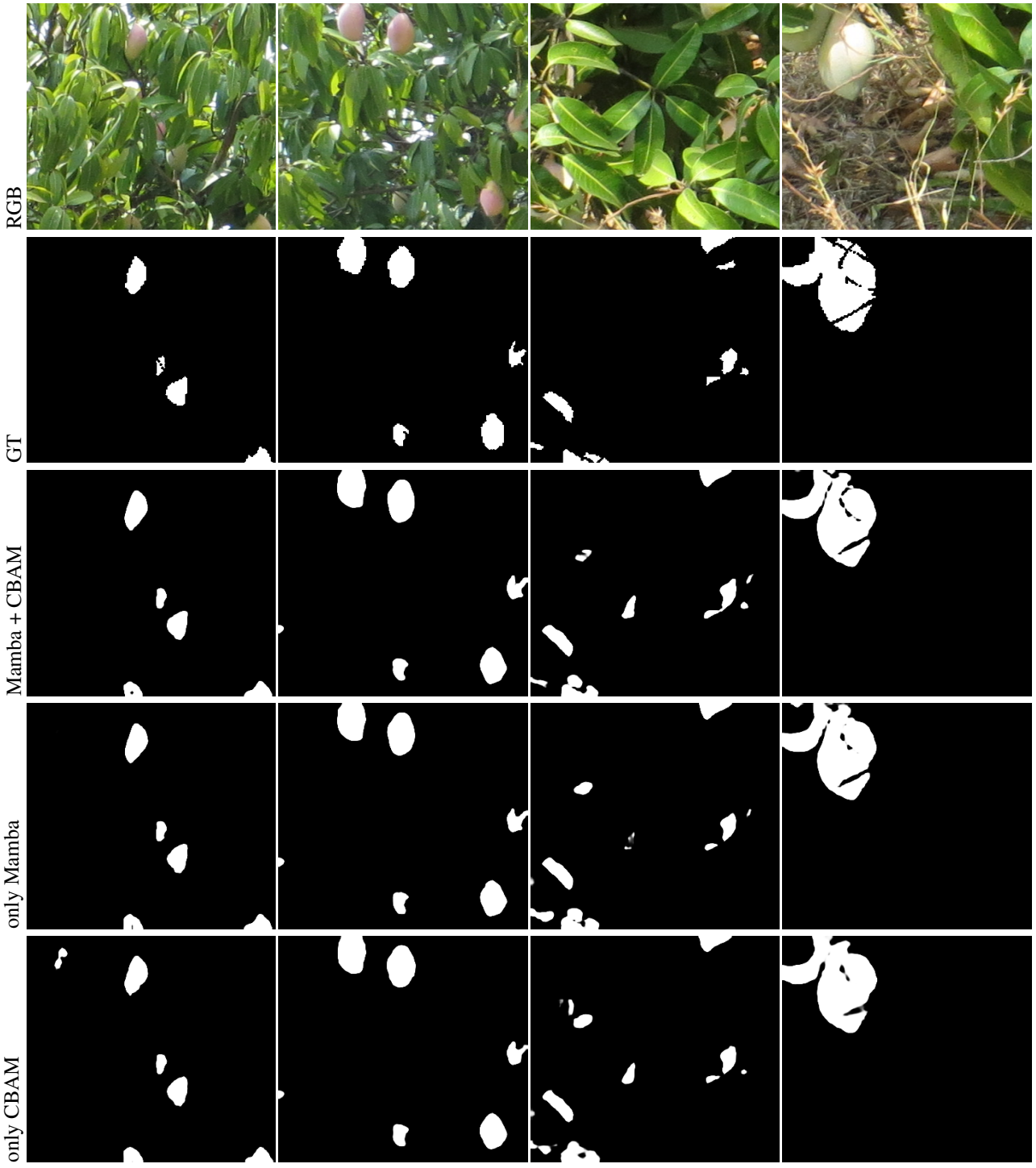

FIGURE 18. Qualitative results of different module combinations used, evaluated on example images from the Mango dataset [24].

TABLE 16. Metric evaluation with different outputs on Mango dataset [24]. The best three performing results are highlighted using color: First, Second, and Third respectively.

Name	Output	$S_\alpha \uparrow$	$F_\beta^w \uparrow$	$M \downarrow$	$F_\beta^{adp} \uparrow$	$F_\beta^{mean} \uparrow$	$E_\phi^{adp} \uparrow$	$E_\phi^{mean} \uparrow$
Avg	Avg	0.8478	0.8080	0.0113	0.7703	0.8199	0.9260	0.9491
M_1	$P_0 + P_1 + P_2 + P_3 + Avg$	0.8662	0.8280	0.0100	0.8222	0.8285	0.9544	0.9557
M_2	$P_1 + P_2 + P_3 + Avg$	0.8671	0.8263	0.0101	0.8146	0.8263	0.9510	0.9548
M_3	$P_2 + P_3 + Avg$	0.8663	0.8235	0.0103	0.8059	0.8240	0.9471	0.9539
M_4	$P_3 + Avg$	0.8634	0.8167	0.0107	0.7913	0.8208	0.9396	0.9541

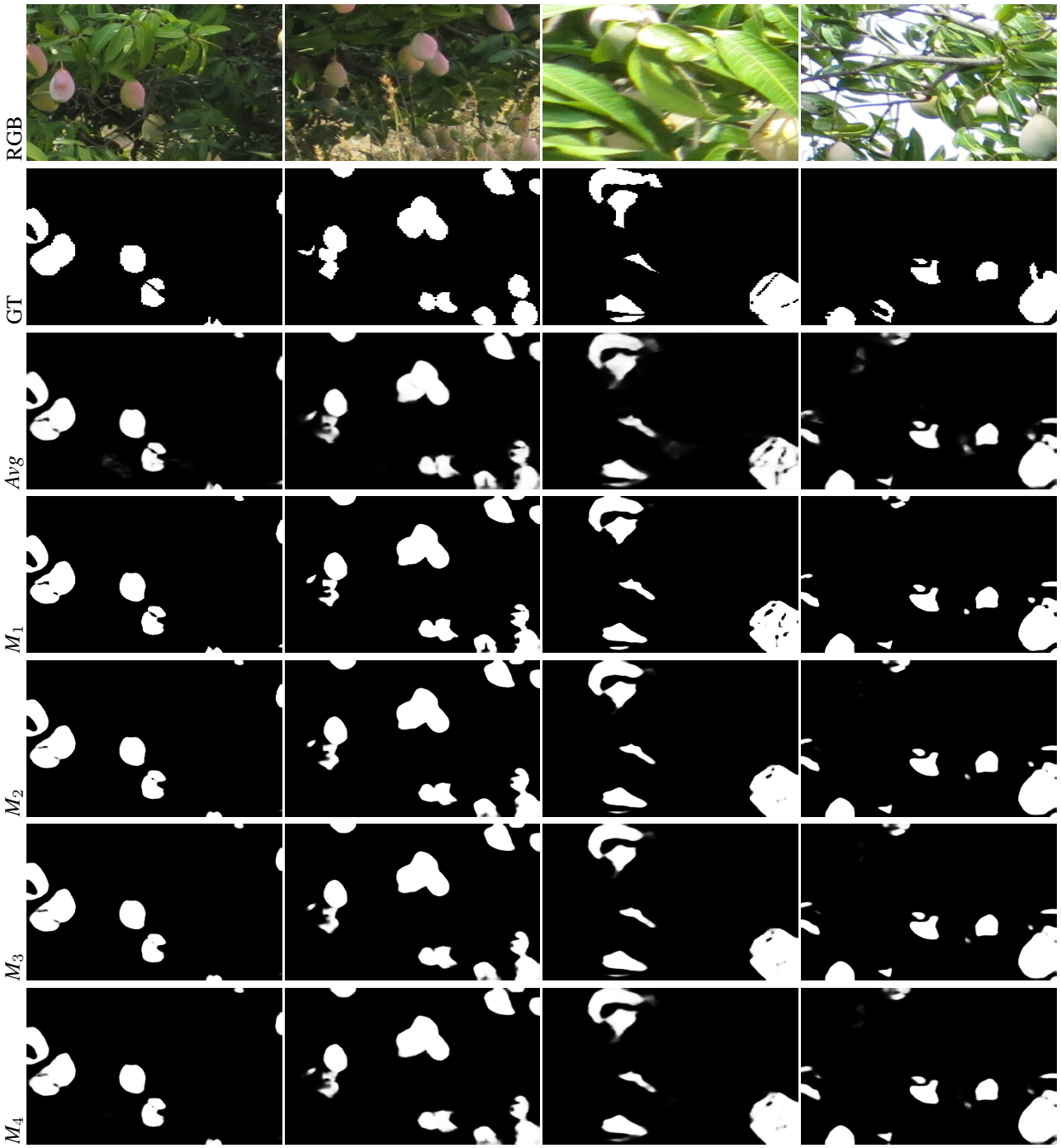

FIGURE 19. Qualitative results of different deep supervision outputs, evaluated on example images from the Mango dataset [24].

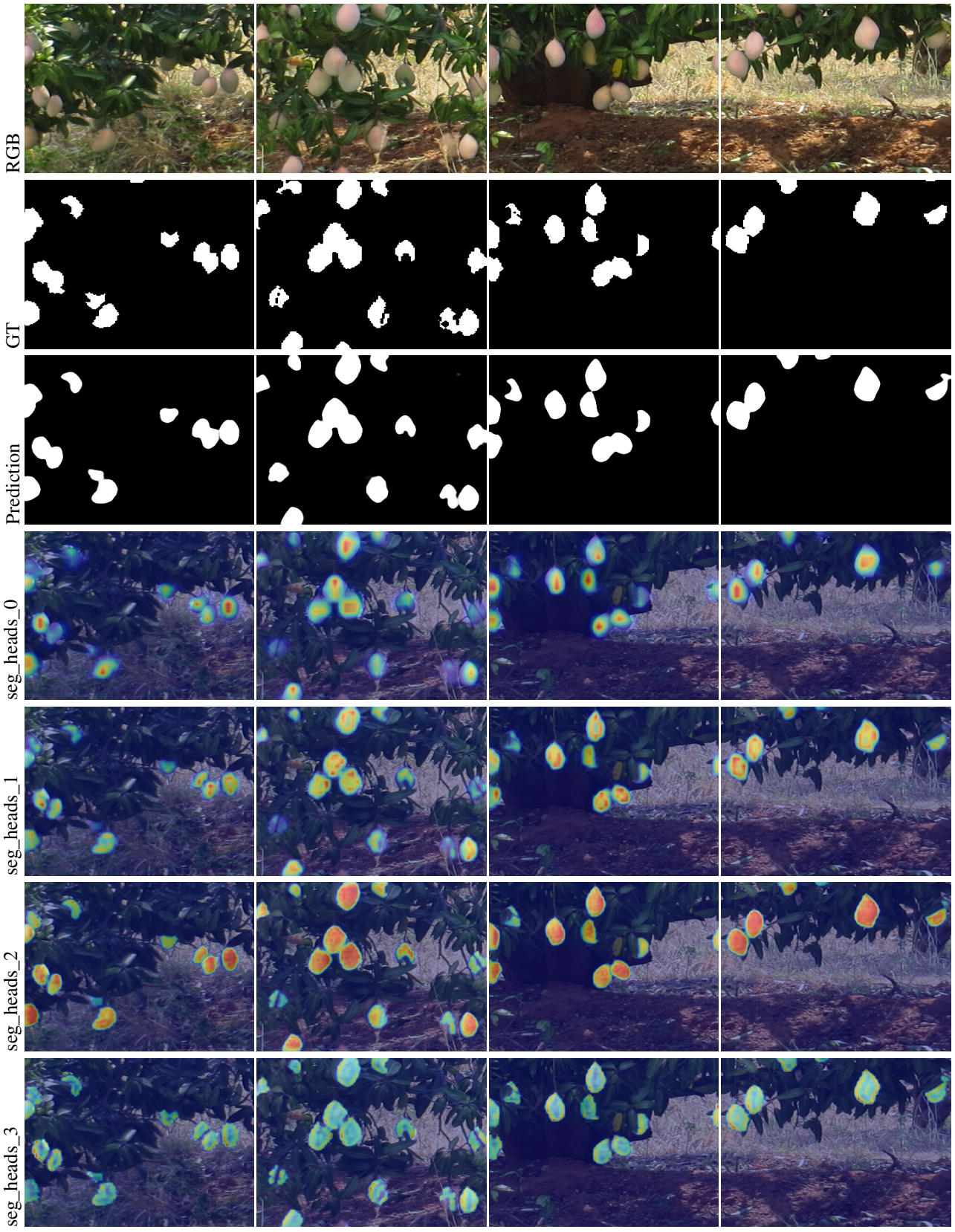

FIGURE 20. Grad-CAM visualization using segmentation heads outputs, evaluated on example images from the Mango dataset [24].

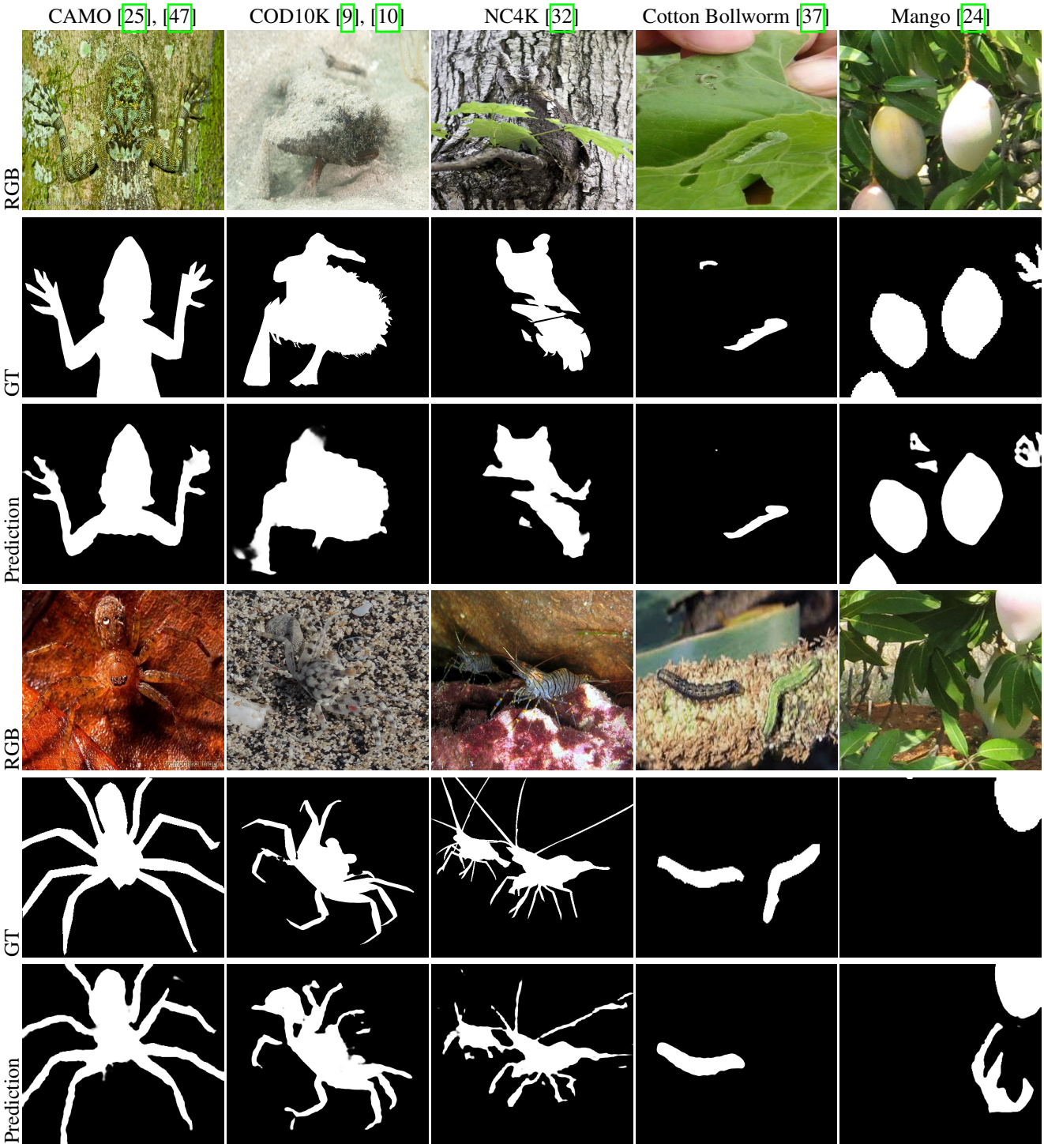

FIGURE 21. Analysis of AINet failure cases using two example images on standard COD benchmarks (CAMO, COD10K and NC4K datasets), as well as Cotton Bollworm and Mango datasets, summarizing common error patterns.

VII. CONCLUSIONS

This work introduces AINet, a novel architecture for camouflaged object detection that sets a new standard across both general benchmarks and agricultural case studies. On widely recognized benchmark datasets, AINet consistently achieves state-of-the-art results, surpassing existing methods in all evaluated metrics. The model demonstrates superior segmentation accuracy and edge delineation, effectively detecting a wide range of camouflaged objects and establishing itself as a leading solution for challenging COD scenarios.

Beyond general benchmarks, AINet's effectiveness is further validated through comprehensive case studies on agricul-

tural datasets. On the Cotton Bollworm dataset, AINet excels in pest detection, delivering highly accurate identification and producing sharp, reliable segmentation masks even in complex backgrounds where pests are difficult to distinguish from vegetation. Similarly, on the Mango dataset, AINet achieves outstanding fruit detection performance, maintaining precise contours and minimal false positives under diverse illumination conditions—key requirements for automated fruit harvesting and monitoring systems.

The results demonstrate that AINet not only advances the state of the art in generic camouflaged object detection but also offers significant benefits for precision agriculture. Its robust performance across both benchmark and agricultural datasets highlights its versatility and potential impact, supporting more effective pest management, reducing crop losses, and enabling the development of intelligent, automated systems for modern farming.

In summary, AINet emerges as a powerful and generalizable solution for camouflaged object detection, delivering substantial improvements in accuracy and boundary precision across diverse real-world applications.

ACKNOWLEDGEMENTS

This work is supported in part by the Air Force Office of Scientific Research Under Award FA9550-24-1-0206; in part by the ESPOL project "Advancing Camouflaged Object Detection with a cost-effective Cross-Spectral vision system (ACODCS)" (CIDIS-003-2024).

REFERENCES

[1] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk. Frequency-tuned salient region detection. In *Conf. on Computer Vision and Pattern Recognition*, pages 1597–1604. IEEE, 2009.

[2] G. Chen, S.-J. Liu, Y.-J. Sun, G.-P. Ji, Y.-F. Wu, and T. Zhou. Camouflaged object detection via context-aware cross-level fusion. *Transactions on Circuits and Systems for Video Technology*, 32(10):6981–6993, 2022.

[3] T. Chen, J. Xiao, X. Hu, G. Zhang, and S. Wang. Boundary-guided network for camouflaged object detection. *Knowledge-based systems*, 248:108901, 2022.

[4] L. M. Dang, S. Danish, A. Khan, N. Alam, M. Fayaz, D. K. Nguyen, H.-K. Song, and H. Moon. An efficient zero-labeling segmentation approach for pest monitoring on smartphone-based images. *European Journal of Agronomy*, 160:127331, 2024.

[5] P.-T. De Boer, D. P. Kroese, S. Mannor, and R. Y. Rubinstein. A tutorial on the cross-entropy method. *Annals of operations research*, 134:19–67, 2005.

[6] I. R. Evangelista, A. Bandala, and E. Dadios. Fcnet: A transformer-based context-aware segmentation framework for detecting camouflaged fruits in orchard environments. *Technologies*, 13(8):372, 2025.

[7] D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, and A. Borji. Structure-measure: A new way to evaluate foreground maps. In *Int. Conference on Computer Vision*, pages 4548–4557, 2017.

[8] D.-P. Fan, C. Gong, Y. Cao, B. Ren, M.-M. Cheng, and A. Borji. Enhanced-alignment measure for binary foreground map evaluation. *arXiv*, 2018.

[9] D.-P. Fan, G.-P. Ji, M.-M. Cheng, and L. Shao. Concealed object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

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

[11] A. Gu and T. Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

[12] C. He, K. Li, Y. Zhang, L. Tang, Y. Zhang, Z. Guo, and X. 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.

[13] C. He, K. Li, Y. Zhang, G. Xu, L. Tang, Y. Zhang, Z. Guo, and X. Li. Weakly-supervised concealed object segmentation with sam-based pseudo labeling and multi-scale feature grouping. *Advances in Neural Information Processing Systems*, 36:30726–30737, 2023.

[14] C. He, K. Li, Y. Zhang, Y. Zhang, Z. Guo, X. Li, M. Danelljan, and F. Yu. Strategic preys make acute predators: Enhancing camouflaged object detectors by generating camouflaged objects. *arXiv preprint arXiv:2308.03166*, 2023.

[15] J. Hu, J. Lin, S. Gong, and W. Cai. Relax image-specific prompt requirement in sam: A single generic prompt for segmenting camouflaged objects. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 12511–12518, 2024.

[16] X. Hu, S. Wang, X. Qin, H. Dai, W. Ren, D. Luo, Y. Tai, and L. Shao. High-resolution iterative feedback network for camouflaged object detection. In *Conf. on Artificial Intelligence*, volume 37, pages 881–889, 2023.

[17] C. A. Javalagi, K. M. Medha, N. T. Patil, S. Itagalli, U. Kulkarni, and S. Chikkamath. Comparative study of cnns for camouflaged object detection. In V. K. Gunjan and J. M. Zurada, editors, *Proceedings of 4th International Conference on Recent Trends in Machine Learning, IoT, Smart Cities and Applications*, pages 207–220, Singapore, 2024. Springer Nature Singapore.

[18] G.-P. Ji, D.-P. Fan, Y.-C. Chou, D. Dai, A. Liniger, and L. Van Gool. Deep gradient learning for efficient camouflaged object detection. *Machine Intelligence Research*, 20(1):92–108, 2023.

[19] G.-P. Ji, L. Zhu, M. Zhuge, and K. Fu. Fast camouflaged object detection via edge-based reversible re-calibration network. *Pattern Recognition*, 123:108414, 2022.

[20] Q. Jia, S. Yao, Y. Liu, X. Fan, R. Liu, and Z. 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.

[21] P. Jiang, Y. Chen, B. Liu, D. He, and C. Liang. Real-time detection of apple leaf diseases using deep learning approach based on improved convolutional neural networks. *Ieee Access*, 7:59069–59080, 2019.

[22] X. Jiang, W. Cai, Z. Zhang, B. Jiang, Z. Yang, and X. Wang. Magnet: A camouflaged object detection network simulating the observation effect of a magnifier. *Entropy*, 24(12):1804, 2022.

[23] N. Kajiura, H. Liu, and S. Satoh. Improving camouflaged object detection with the uncertainty of pseudo-edge labels. In *Proceedings of the 3rd ACM International Conference on Multimedia in Asia*, pages 1–7, 2021.

[24] R. Kestur, A. Meduri, and O. Narasipura. Mangonet: A deep semantic segmentation architecture for a method to detect and count mangoes in an open orchard. *Engineering Applications of Artificial Intelligence*, 77:59–69, 2019.

[25] T.-N. Le, T. V. Nguyen, Z. Nie, M.-T. Tran, and A. Sugimoto. Anabranch network for camouflaged object segmentation. *Journal of Computer Vision and Image Understanding*, 184:45–56, 2019.

[26] S. Lev-Yadun, A. Dafni, M. A. Flaishman, M. Inbar, I. Izhaki, G. Katzir, and G. Ne'eman. Plant coloration undermines herbivorous insect camouflage. *BioEssays*, 26(10):1126–1130, 2004.

[27] A. Li, J. Zhang, Y. Lv, T. Zhang, Y. Zhong, M. He, and Y. Dai. Joint salient object detection and camouflaged object detection via uncertainty-aware learning. *arXiv preprint arXiv:2307.04651*, 2023.

[28] P. Li, X. Yan, H. Zhu, M. Wei, X.-P. Zhang, and J. Qin. Findnet: Can you find me? boundary-and-texture enhancement network for camouflaged object detection. *IEEE Transactions on Image Processing*, 31:6396–6411, 2022.

[29] J. Liu, J. Zhang, and N. Barnes. Modeling aleatoric uncertainty for camouflaged object detection. In *Winter conference on applications of computer vision*, pages 1445–1454, 2022.

[30] Y. Lv, J. Zhang, Y. Dai, A. Li, N. Barnes, and D.-P. Fan. Toward deeper understanding of camouflaged object detection. *IEEE transactions on circuits and systems for video technology*, 33(7):3462–3476, 2023.

[31] Y. Lv, J. Zhang, Y. Dai, A. Li, B. Liu, N. Barnes, and D.-P. 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.

[32] Y. Lyu, J. Zhang, Y. Dai, A. Li, B. Liu, N. Barnes, and D.-P. Fan. Simultaneously localize, segment and rank the camouflaged objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[33] J. Ma, F. Li, and B. Wang. U-mamba: Enhancing long-range dependency for biomedical image segmentation. *arXiv:2401.04722*, 2024.

[34] R. Margolin, L. Zelnik-Manor, and A. Tal. How to evaluate foreground maps? In *Conf. on Computer Vision and Pattern Recognition*, pages 248–255, 2014.

[35] G. Máttyus, W. Luo, and R. Urtasun. Deeproadmapper: Extracting road topology from aerial images. In *Proceedings of the IEEE international conference on computer vision*, pages 3438–3446, 2017.

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

[37] K. Meng, K. Xu, P. Cattani, and S. Mei. Camouflaged cotton bollworm instance segmentation based on pvt and mask r-cnn. *Computers and Electronics in Agriculture*, 226:109450, 2024.

[38] Y. Pang, X. Zhao, T.-Z. Xiang, L. Zhang, and H. 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.

[39] F. Perazzi, P. Krähenbühl, Y. Pritch, and A. Hornung. Saliency filters: Contrast based filtering for salient region detection. In *Conf. on Computer Vision and Pattern Recognition*, pages 733–740. IEEE, 2012.

[40] X. Qin, Z. Zhang, C. Huang, C. Gao, M. Dehghan, and M. Jagersand. Basnet: Boundary-aware salient object detection. In *Conf. on Computer Vision and Pattern Recognition*, June 2019.

[41] D. Sun, S. Jiang, and L. Qi. Edge-aware mirror network for camouflaged object detection. In *Int. Conf. on Multimedia and Expo*, pages 2465–2470. IEEE, 2023.

[42] Y. Sun, G. Chen, T. Zhou, Y. Zhang, and N. Liu. Context-aware cross-level fusion network for camouflaged object detection. In *IJCAI*, pages 1025–1031, 2021.

[43] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao. Pvt v2: Improved baselines with pyramid vision transformer. *Computational Visual Media*, 8(3):415–424, 2022.

[44] J. Wei, S. Wang, and Q. Huang. F^3net: fusion, feedback and focus for salient object detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 12321–12328, 2020.

[45] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

[46] F. Xiao, S. Hu, Y. Shen, C. Fang, J. Huang, C. He, L. Tang, Z. Yang, and X. Li. A survey of camouflaged object detection and beyond. *arXiv preprint arXiv:2408.14562*, 2024.

[47] J. Yan, T.-N. Le, K.-D. Nguyen, M.-T. Tran, T.-T. Do, and T. V. Nguyen. Mirrornet: Bio-inspired camouflaged object segmentation. *IEEE Access*, 9:43290–43300, 2021.

[48] X. Yan, M. Sun, Y. Han, and Z. Wang. Camouflaged object segmentation based on matching–recognition–refinement network. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[49] F. Yang, Q. Zhai, X. Li, R. Huang, A. Luo, H. Cheng, and D.-P. Fan. Uncertainty-guided transformer reasoning for camouflaged object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4146–4155, 2021.

[50] J. Yang, Q. Wang, F. Zheng, P. Chen, A. Leonardis, and D.-P. Fan. Plantcamo: Plant camouflage detection. *arXiv*, 2024.

[51] Q. Zhai, X. Li, F. Yang, C. Chen, H. Cheng, and D.-P. 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.

[52] M. Zhang, S. Xu, Y. Piao, D. Shi, S. Lin, and H. Lu. Preynet: Preying on camouflaged objects. In *Proceedings of the 30th ACM international conference on multimedia*, pages 5323–5332, 2022.

[53] Y. Zhang and C. Wu. Unsupervised camouflaged object segmentation as domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4334–4344, 2023.

[54] Y. Zhang, J. Zhang, W. Hamidouche, and O. Deforges. Predictive uncertainty estimation for camouflaged object detection. *IEEE Transactions on Image Processing*, 32:3580–3591, 2023.

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

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

[57] J. Zhu, X. Zhang, S. Zhang, and J. Liu. Inferring camouflaged objects by texture-aware interactive guidance network. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 3599–3607, 2021.



HENRY O. VELESACA received his degree in Computer Science Engineering from the Escuela Superior Politécnica del Litoral (ESPOL), Ecuador, in 2014. He obtained an Honors Diploma for Best Graduate of the Computer Science Engineering Degree in 2013. In 2017, he obtained a Master's in Software Development at the University of Granada. In the same year, he received Second Place in the First Agro-Food Innovation Contest of Grupo La Caña, Spain - Motril. He is currently a Ph.D. student in the Doctoral Program in Information and Communication Technologies at the University of Granada. Also works as a research technician at ESPOL. His research interests include deep learning, computer vision, and robust model design, with applications in agricultural, industrial automation, and sports analytics.



ANDREA MERO P. received the B.Sc. degree in computer science from the Escuela Superior Politécnica del Litoral (ESPOL), Guayaquil, Ecuador, in 2024. She is currently pursuing the M.Sc. degree in Artificial Intelligence with the Università della Svizzera italiana (USI), Lugano, Switzerland. She has held multiple research roles at ESPOL, including a position as a Computer Vision Research Assistant with the Centro de Investigación, Desarrollo e Innovación de Sistemas Computacionales (CIDIS). Her research contributions span computer vision, deep learning, and data analysis, with a focus on cross-spectral imagery, super-resolution, and camouflaged object detection.



ABEL REYES-ANGULO received the Ph.D. degree in Computational Science and Engineering from Michigan Technological University, Houghton, MI, USA, in 2025. He is currently a Senior Data Scientist with the Pricing and Assortment team at Walmart Global Tech and continues to collaborate as a researcher with academic partners in medical imaging and computer vision. His research interests include efficient deep learning architectures, medical image segmentation, representation learning, and robust model design, with applications in healthcare, remote sensing, and sports analytics. He has co-authored several peer-reviewed publications in these areas and is actively involved in the AI community, including service and mentoring roles in Latin American AI initiatives.



ANGEL D. SAPPA (Senior Member, IEEE) received the degree in electromechanical engineering from the National University of La Pampa, Argentina, in 1995, and the Ph.D. degree in industrial engineering from the Polytechnic University of Catalonia, Spain, in 1999. He is currently a Senior Scientist with the Computer Vision Center, Universitat Autónoma de Barcelona, Spain; and a full-time Professor with ESPOL Polytechnic University, Ecuador, where he leads the Computer Vision Team, CIDIS Research Center. He has over 25 years of experience in computer vision and artificial intelligence, with more than 300 publications in top-ranked journals and leading international conferences. From 2022 to 2024, he was included in Stanford University's top 2% most-cited researchers in the subfield of artificial intelligence (career-long ranking). He serves as an Editor in Chiev of ELCVIA Journal, an Associate Editor for Pattern Recognition and a Guest Editor for Robotics and Autonomous Systems and Journal of Computational Science.

SEEING THE UNSEEN: AI-POWERED CAMOUFLAGED PEST DETECTION

Henry O. Velesaca[1,3] and Angel D. Sappa[1,2]
[1]*Escuela Superior Politécnica del Litoral, ESPOL, Campus Gustavo Galindo Km. 30.5 Vía Perimetral, Guayaquil, 090902, Ecuador*
[2]*Computer Vision Center, Campus UAB, 08193 Bellaterra, Barcelona, Spain*
[3]*Software Engineering Department, Research Center for Information and Communication Technologies (CITIC-UGR), University of Granada, 18071, Granada, Spain*

ABSTRACT

This paper presents an innovative approach for detecting camouflaged agricultural pests using YOLOv11, a state-of-the-art deep learning architecture, and SINet-v2. The research focuses specifically on cotton bollworm detection, a significant threat to agricultural productivity. Our methodology leverages advanced neural networks to identify subtle patterns distinguishing concealed pests from their surrounding environment. Using a comprehensive dataset of 1,073 cotton bollworm images, the YOLOv11 and SINet-v2 architectures are evaluated against an existing approach. The results obtained by YOLOv11 demonstrate superior performance in the Average Precision metric for detection of 71.3% and segmentation of 60.3%. Compared to recently presented methods, like Mask R-CNN+PVT, our approach shows improved detection capabilities while requiring fewer computational resources. This work contributes to advancing automated pest management systems in agriculture, offering a more efficient and accurate solution for early pest detection.

KEYWORDS

Camouflaged Object Detection, Pest Detection, YOLO

1. INTRODUCTION

Agricultural pests pose a significant threat to global food security, causing substantial crop losses estimated at billions of dollars annually (https://www.fao.org/plant-production-protection/about/en) (Oerke, 2016). The challenge of detecting these pests is further complicated by their evolutionary adaptations, particularly camouflage mechanisms that make them nearly invisible to conventional detection methods (Skendžić et al., 2021).

In recent years, the emergence of deep learning technologies has revolutionized computer vision applications, offering unprecedented capabilities in automated pest detection and monitoring systems (Karar et al., 2021; Ullah at al., 2022). These advances have transformed traditional agricultural practices by introducing sophisticated neural network architectures like Convolutional Neural Networks (CNNs) (Venkatasaichandrakanthand et al., 2023), Region-based CNNs (R-CNN) (Rong et al., 2022), and You Only Look Once (YOLO) variants (Liu at al., 2020), which have demonstrated remarkable accuracy in identifying and classifying various pest species. The integration of these technologies has enabled real-time monitoring, early detection of pest infestations, and more precise application of pest control measures, leading to significant improvements in crop protection strategies and reduced pesticide usage.

The importance of this work lies in its potential to transform pest management practices, enabling early detection and targeted intervention strategies (Chen at al., 2020). This advancement could substantially reduce pesticide usage while improving crop protection efficiency, thereby contributing to more sustainable agricultural practices. Our approach not only addresses the immediate challenge of pest detection but also provides a scalable framework for future developments in agricultural monitoring systems.

This research presents an innovative approach for detecting camouflaged agricultural pests using state-of-the-art deep learning techniques. Our methodology leverages advanced convolutional neural networks (CNNs) and feature extraction algorithms to identify subtle patterns and characteristics that distinguish concealed pests from their surrounding environment. By combining high-resolution imaging with sophisticated neural network architectures (Peng et al, 2022).

To address this work in detail, the manuscript is organized as follows. Section 2 introduces the background on the use of pest detection techniques based on deep learning and addresses the problem as a camouflaged object detection approach. Section 3 presents the proposed approach to carry out the identification of pests using deep learning techniques. Then, section 4 shows the experimental results taking as a reference a case study for the camouflage detection of cotton bollworm. Finally, conclusions are presented in Section 5.

2. BACKGROUND

As described above, this document presents a deep learning-based approach to camouflaged agricultural pest detection. This section summarizes some of the most relevant techniques related to the present work. First, approaches based on Deep Learning are reviewed, followed by approaches based on Camouflaged Object Detection.

2.1 Deep Learning Approaches

Deep learning approaches, particularly CNN techniques, have demonstrated superior performance in pest detection. As is the case, the work proposed by Rivadeneira et al. (2023), which presented a novel approach that is especially valuable for camouflage scenarios, as it integrates super- resolution techniques with object detection methods like YOLOv8 (Jocher et al., 2023) to address the problem of very low-resolution thermal images. The most relevant feature of the camouflage area is that it allows the detection of objects that could remain undetected in visible spectrum images since it uses improved thermal images through a guided super-resolution strategy that is supported by a high-resolution image of the visible spectrum. This makes the system particularly useful for identifying camouflaged objects that will be difficult or impossible to detect using conventional images alone.

However, there are certain limitations and problems, such as the work presented by Fuentes et al. (2017) identified that these systems require massive amounts of labeled training data and significant computational resources. Their study on tomato disease detection showed that while achieving 96% accuracy under controlled conditions, performance dropped significantly when dealing with new or unseen pest varieties.

2.2 Camouflaged Object Detection Approaches

Recent research has specifically addressed the challenge of detecting pests as a Camouflaged Object Detection (COD) problem. Remarkable work by Wang et al. (2023) shows a camouflaged insect segmentation system using a Progressive Refinement Network that employs an innovative multi- stage approach to accurately detect and segment insects that blend into their environment. This architecture implements a hierarchical process that gradually refines visual features, starting with coarse features and progressing to fine details, achieving significant accuracy in detecting camouflaged insects. However, the system faces limitations such as dependence on large and well- labeled data sets, difficulties with variable lighting conditions, and extremely complex camouflage patterns, in addition to requiring considerable computational power for real-time processing.

Finally, the work recently presented by Meng at al. (2024) shows a segmentation system for camouflaged bollworm instances, which combines the Pyramid Vision Transformer (PVT) architecture with Mask R-CNN, creating a hybrid approach that improves detection and segmentation of this pest in agricultural environments. This integration leverages PVT's ability to capture global relationships and hierarchical features, while Mask R-CNN provides accurate segmentation at the instance level, achieving superior results in identifying cotton bollworms even when they are camouflaged in foliage. However, the system presents limitations in terms of required computational resources, sensitivity to varying environmental conditions, and the need for an extensive and diverse training dataset to maintain its effectiveness in different field scenarios.

3. PROPOSED APPROACH

This section details the different stages followed to carry out the proposed approach. First, the dataset is detailed, then the techniques used, and finally, the metrics needed for the quantitative evaluation.

3.1 Dataset Description

The cotton bollworm dataset represents a specialized collection of images focused on one of the most destructive agricultural pests affecting cotton crops worldwide. This dataset has been specifically created by Meng et al. (2024) to address the challenges in detecting these pests in various scenes and environmental conditions. The dataset is available at the following URL: https://www.kaggle.com/datasets/kexinmeng1/the-dataset-of-cotton-bollworms.

3.2 Deep Learning Techniques

For the segmentation task, YOLOv11 (Jocher et al., 2024) is selected as the first neural network architecture. This choice is driven by several key advantages: its exceptional processing efficiency and real-time detection capabilities, coupled with its ability to identify multiple objects simultaneously in a single forward pass. YOLOv11's sophisticated architecture offers remarkable flexibility, allowing for fine-tuned optimization to handle varying environmental conditions, including diverse lighting scenarios, and viewing perspectives. Furthermore, its open-source nature and robust community support ensure both straightforward implementation and sustainable long-term maintenance, making it an ideal choice for this application.

On the other hand, based on the state of the art, a comparison is made with the work proposed by Meng et al. (2024), which uses the Pyramid Vision Transformer (PVT) architecture with Mask R-CNN to detect camouflaged cotton bollworm instances. The dataset proposed by the same authors' team is used and compared with the metrics presented in the original paper.

Finally, it is also chosen to use SINet-v2 (Fan et al., 2021) which is a neural network used for COD tasks in two stages: search and identification. It uses EfficientNet as a backbone and incorporates two modules, neighbor connection decoder (NCD) and group-reversal attention (GRA) that allow it to excel in detecting camouflaged objects. This architecture is chosen because it has demonstrated state-of-the-art results on major reference datasets and is available as open source.

3.3 Metric Evaluation

For the evaluation of the proposed work, Average Precision (AP) (e.g., for bounding boxes and segmentation tasks), Recall, F1-score, and Pixel Accuracy metrics will be calculated to establish the performance of the proposed approach. Table 1 shows the details of the metrics to be calculated as well as a description of each of them.

Table 1. Metrics used to evaluate proposed work. TP, FP, TN, and FN represent True Positive, False Positive, True Negative, and False Negative, respectively \cite{meng2024camouflaged}

Metric	Description
AP^B	Average Precision for predicted bounding boxes, calculated across multiple IoU thresholds from 0.5 to 0.95 in steps of 0.05
AP_{B50}	Average Precision for predicted bounding boxes, evaluated at an IoU threshold of 0.5
AP_{B75}	Average Precision for predicted bounding boxes, evaluated at an IoU threshold of 0.75
AP^S	Average Precision for predicted masks, calculated across multiple IoU thresholds from 0.5 to 0.95 in steps of 0.05
AP_{S50}	Average Precision for predicted masks, evaluated at an IoU threshold of 0.5
AP_{S75}	Average Precision for predicted masks, evaluated at an IoU threshold of 0.75
Pixel Accuracy	PA = (TP + TN) / (TP + TN + FP + FN)
Recall	R = TP / (TP + FN)
F1-Score	F1 = 2 x P x R / (P + R); P = TP/ (TP + FP)

4. CASE STUDY

This section presents the case study experimental results obtained with the proposed framework. For the performance evaluation of the proposed approach, the metrics described in Table 1 are used.

4.1 System Implementation

All experiments are conducted on a high-performance workstation configured with an Intel Core i9 processor (base clock: 3.3 GHz, 16 cores/24 threads) and an NVIDIA Titan XP GPU (12GB GDDR5X, 3840 CUDA cores). This hardware configuration provided the computational resources required for efficient training, validation, and testing of the tested architectures in the context of COD tasks. The system's specifications are chosen to ensure optimal processing of complex neural network operations and handling of large-scale dataset operations.



Figure 1. Example of images from the cotton bollworm dataset presented by Meng et al. (2024)

4.2 Dataset

The dataset used for the different experiments carried out is obtained from (Meng et al., 2024). This dataset consists of 1080 images, but inconsistencies are found in 7 of these images, which are removed. Table 2 shows the distribution of the dataset for each of the subsets. Additionally, Figure 1 shows examples of the images that are part of the dataset.

Table 2. Distribution of dataset

Task	Cotton bollworm	
	# of Images	# of Instances
Training	856	932
Validation	161	172
Testing	56	69
Total	1073	1176

4.3 YOLOv11 Implementation

To enhance the model's training capabilities, the initial step involved expanding the training dataset through data augmentation techniques. This is accomplished using the Albumentations library (Buslaev et al., 2020), a powerful tool specifically designed for computer vision applications, including the development of object detection models like YOLOv11. The library provides a comprehensive suite of image transformation operations, encompassing techniques such as random cropping, rotational adjustments, brightness modifications, contrast alterations, and various other transformations. These modifications can be systematically applied during the data preprocessing phase, creating additional training samples that maintain realistic characteristics while introducing controlled variations to the original images. This approach effectively enriches the training dataset, promoting better model generalization and robustness. The model "*yolo11m-seg.pt*" with training epochs of 500, batch size of 8, and image size of 640 × 640 pixels are used as execution parameters.

4.4 SINet-v2

The SINet-V2 implementation is used for the detection of cotton bollworm camouflage. This architecture has been designed for the COD task and is considered for this work because it is a state-of-the-art architecture that has obtained very good results. The implementation uses PyTorch and features a custom loss function that combines weighted BCE (Binary Cross Entropy) and IoU (Intersection over Union). Training is carried out for 100 epochs with a batch size of 16 and an image size of 352×352 pixels, using the Adam optimizer with an initial learning rate of 0.0001 that is adjusted every 25 epochs.

4.5 Experimental Results

As a last stage to evaluate the performance of the proposed approach 56 test images are considered for the calculation of the metrics. Table 3 shows the result of the calculation of the metrics for each of the techniques. The values of the Mask R-CNN+PVT (Meng et al., 2024) technique are obtained from the original paper. In addition, Figure 2 (*3rd column*) uses the images of the results of the article proposed by Meng et al. (2024) for qualitative comparison in this work.

Table 3. Evaluation of metrics for each technique---notation as presented in Table 1

Technique	Detection			Segmentation			PA	R	F1
	AP^B	AP_{B50}	AP_{B75}	AP^S	AP_{S50}	AP_{S75}			
Mask R-CNN+PVT	62.5	89.7	72.9	59.5	89.2	72.9	**98.61**	**99.12**	**98.95**
YOLOv11	**71.3**	**92.3**	**73.8**	**60.3**	**90.3**	**73.1**	98.34	96.47	97.79
SINet-v2	54.5	71.3	59.8	44.5	56.3	49.2	90.07	65.61	47.70

Table 3 shows that YOLOv11 architecture achieved notable metrics with an AP^B of 71.3% for detection and an AP^S of 60.3% for segmentation, significantly outperforming the Mask R-CNN+PVT model which achieved 62.5% and 59.5% respectively. This improvement in accuracy is particularly relevant considering the inherent complexity of detecting camouflaged pests and a notable aspect of YOLOv11 is its computational efficiency. The proposed model requires substantially fewer resources, operating at 123.0 GFlops compared to the 154.78 GFlops of the Mask R-CNN+PVT while maintaining a smaller number of parameters (22.34M vs 41.78M) (see Table 4). This optimization in computational resources makes the system more viable for practical implementations in real agricultural environments. These results are particularly impressive considering the complex nature of the test images including various camouflage patterns and varying environmental conditions.

Table 4. Comparison between GFlops and the number of parameters

Technique	Input size (px)	GFlops	#Param. (M)
Mask R-CNN+PVT	800 x 640	154.78	41.78
YOLOv11	640 x 640	123.00	22.34
SINet-v2	352 x 352	144.48	26.98

On the other hand, Figure 2 shows the qualitative evaluation of the results obtained by each technique, in which better obtaining of bounding boxes and segmentation masks can be obtained using YOLOv11; these results are consistent with the quantitative metrics obtained for the detection and segmentation tasks. In addition, Figure 3 shows more challenging images when finding the cotton bollworm camouflaged in the different scenarios where they blend in with the environment, and it is more difficult to visualize them with the naked eye. These qualitative results demonstrate the robustness of YOLOv11 in complex scenarios, and this can equal or improve the results of state-of-the-art techniques specialized for COD tasks. Also, SINet-v2 does not obtain good results and is limited to detecting and segmenting salient objects in this particular dataset (see Table 2 *1st row, 2nd column*).

Although the presented dataset is an excellent contribution, it also presents certain inconsistencies in the labeling, for example, Table 2 (*5th row, 2nd column*) shows a labeling error where the GT does not contain the mask of the small worm located on the right side but Mask R-CNN+PVT and YOLO can detect it. Likewise, Table 3 (*1st row, 2nd column*) shows a labeling error where the GT does not show the mask of the

small worm located on the left side, but YOLO also detects it (*1st row and 4th column green mask*). This is one reason why the PA, R, and F1 metrics do not present higher values.



Figure 2. Prediction results using Mask R-CNN+PVT, YOLOv11, and SINet-v2

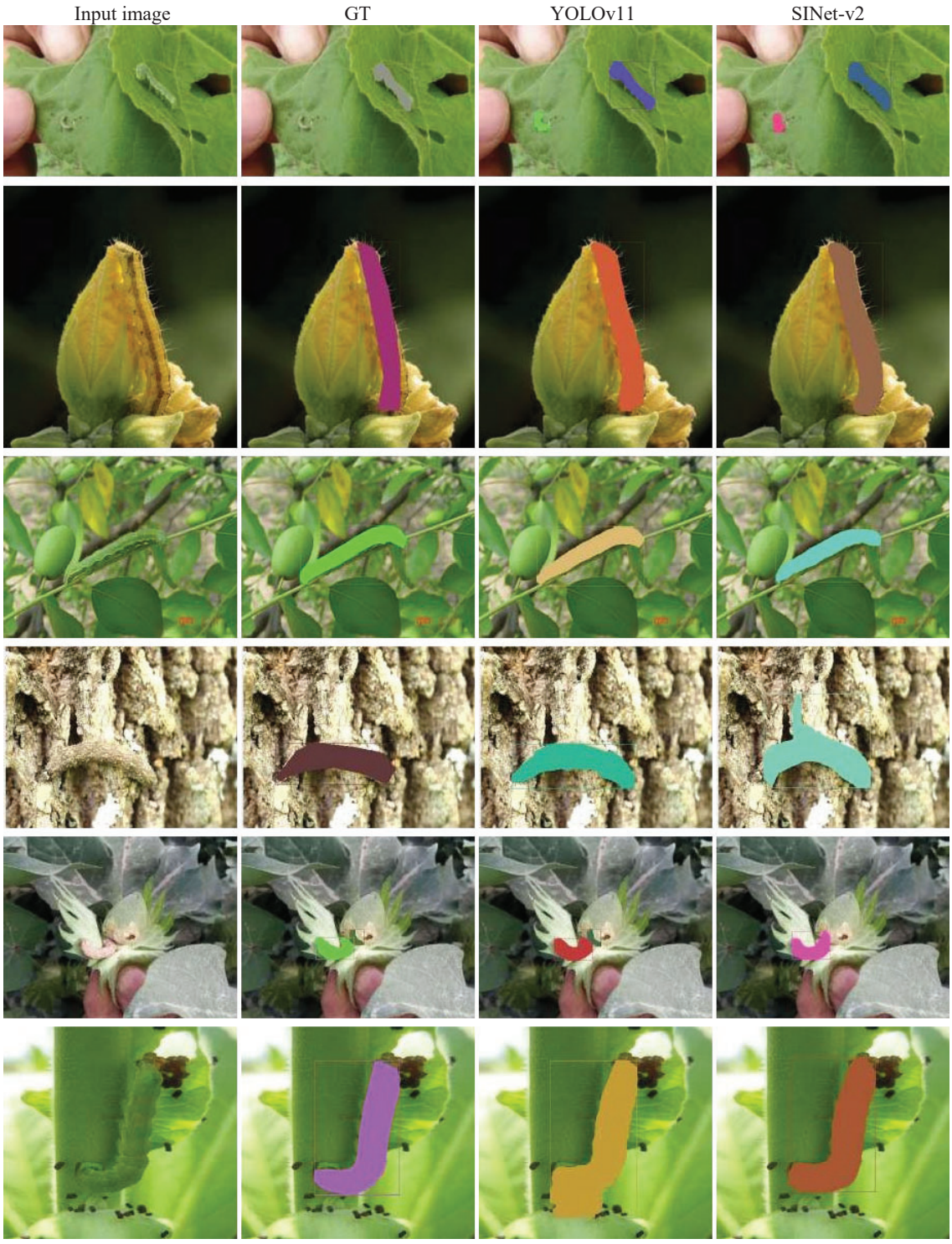

Figure 3. Prediction results using YOLOv11 and SINet-v2 on challenging camouflage images

5. CONCLUSION

This research demonstrates the effectiveness of the YOLOv11-based approach for camouflaged agricultural pest detection, significantly outperforming recent methods such as Mask R-CNN+PVT in both performance and computational efficiency. Another finding is that SINet-v2 exhibits lower performance than YOLOv11, despite SINet-v2 being a specialized architecture for COD tasks. The YOLOv11 evaluation architecture not only achieves better pest detection and segmentation but also requires fewer computational resources, making it more practical for deployments in real agricultural environments. Systems based on these types of technologies not only promise to reduce herbicide use but also offer more precise and environmentally friendly weed control solutions. The robustness of the system is evidenced by its ability to maintain high performance under various challenging scenarios with different camouflage patterns and varying environmental conditions, thus establishing a solid foundation for future developments.

ACKNOWLEDGEMENT

This material is based upon work supported by the Air Force Office of Scientific Research under award number FA9550-24-1-0206; in part by the ESPOL project CIDIS-003-2024-T; and in part by the University of Granada.

REFERENCES

Buslaev, A., Iglovikov, V.I., Khvedchenya, E., Parinov, A., Druzhinin, M. & Kalinin, A.A., (2020). Albumentations: fast and flexible image augmentations. Information, 11(2), p.125.

Chen, C.J., Huang, Y.Y., Li, Y.S., Chang, C.Y. & Huang, Y.M., (2020). An AIoT based smart agricultural system for pests detection. IEEE Access, 8, pp.180750-180761.

Fan, D.P., Ji, G.P., Cheng, M.M. & Shao, L., (2021). Concealed object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(10), pp.6024-6042.

Fuentes, A., Yoon, S., Kim, S.C. & Park, D.S., (2017). A robust deep-learning-based detector for real-time tomato plant diseases and pests recognition. Sensors, 17(9), p.2022.

Jocher, G., Chaurasia, A. & Qiu, J., (2023). Ultralytics yolov8. Available at: https://github.com/ultralytics/ultralytics

Jocher, G. & Qiu, J., (2024). Ultralytics yolo11. Available at: https://github.com/ultralytics/ultralytics

Karar, M.E., Alsunaydi, F., Albusaymi, S. & Alotaibi, S., (2021). A new mobile application of agricultural pests recognition using deep learning in cloud computing system. Alexandria Engineering Journal, 60(5), pp.4423-4432.

Liu, J. & Wang, X., (2020). Tomato diseases and pests detection based on improved Yolo V3 convolutional neural network. Frontiers in Plant Science, 11, p.898.

Meng, K., Xu, K., Cattani, P. & Mei, S., (2024). Camouflaged cotton bollworm instance segmentation based on PVT and Mask R-CNN. Computers and Electronics in Agriculture, 226, p.109450.

Oerke, E.C., (2006). Crop losses to pests. The Journal of Agricultural Science, 144(1), pp.31-43.

Peng, Y. & Wang, Y., (2022). CNN and transformer framework for insect pest classification. Ecological Informatics, 72

Rivadeneira, R.E., Velesaca, H.O. & Sappa, A., (2023), November. Object detection in very low-resolution thermal images through a guided-based super-resolution approach. In: 2023 17th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS), IEEE, pp.311-318.

Rong, M., Wang, Z., Ban, B. & Guo, X., (2022). Pest Identification and Counting of Yellow Plate in Field Based on Improved Mask R-CNN. Discrete Dynamics in Nature and Society, 2022(1), p.1913577.

Skendžić, S., Zovko, M., Živković, I.P., Lešić, V. & Lemić, D., (2021). The impact of climate change on agricultural insect pests. Insects, 12(5), p.440.

Ullah, N., Khan, J.A., Alharbi, L.A., Raza, A., Khan, W. & Ahmad, I., (2022). An efficient approach for crops pests recognition and classification based on novel deeppestnet deep learning model. IEEE Access, 10, pp.73019-73032.

Venkatasaichandrakanthand, P. & Iyapparaja, M., (2023). Pest detection and classification in peanut crops using CNN, MFO, and EViTA algorithms. IEEE Access, 11, pp.54045-54057.

Wang, J., Hong, M., Hu, X., Li, X., Huang, S., Wang, R. & Zhang, F., (2023). Camouflaged insect segmentation using a progressive refinement network. Electronics, 12(4), p.804.

Exploring Camouflaged Object Detection Techniques for Invasive Vegetation Monitoring

Henry O. Velesaca[1,2], Hector Villegas[1] and Angel D. Sappa[1,3]
[1]*Escuela Superior Politécnica del Litoral, ESPOL,*
Campus Gustavo Galindo Km. 30.5 Vía Perimetral, Guayaquil, 090902, Ecuador
[2]*Software Engineering Department, Research Center for Information and Communication Technologies (CITIC-UGR),*
University of Granada, 18071, Granada, Spain
[3]*Computer Vision Center, 08193-Bellaterra, Barcelona, Spain*

Keywords: Camouflaged Object Detection, Weed Detection, Pest Detection, Deep Learning, Computer Vision, UAV.

Abstract: This paper presents a novel approach to weed detection by leveraging state-of-the-art camouflaged object detection techniques. The work evaluates six camouflaged object detection architectures on agricultural datasets to identify weeds naturally blending with crops through similar physical characteristics. The proposed approach shows excellent results in detecting weeds using Unmanned Aerial Vehicle images. This work establishes a new framework for the challenging task of weed detection in agricultural settings using camouflaged object detection approaches, contributing to more efficient and sustainable farming practices.

1 INTRODUCTION

In recent years, the agricultural sector has faced increasing challenges in effective weed management, particularly when dealing with weeds that naturally blend with crops through similar physical characteristics (e.g., (Chauhan et al., 2017), (Westwood et al., 2018)). This phenomenon, known as biological mimicry or natural camouflage, presents a significant obstacle in traditional weed detection methods. The advancement of precision agriculture and computer vision technologies has opened new possibilities for addressing this complex challenge through innovative camouflage-based detection approaches (e.g., (Moldvai et al., 2024), (Wu et al., 2021)).

Traditional approaches to weed detection have primarily relied on conventional image processing techniques such as color analysis, texture features, and shape-based recognition. These methods typically employ techniques like RGB color space transformation, edge detection algorithms, and morphological operations to distinguish between crops and weeds (e.g., (Parra et al., 2020), (Agarwal et al., 2021)). Historically, researchers have utilized techniques such as the Normalized Difference Vegetation Index (NDVI), color indices, and spectral reflectance measurements to identify vegetation patterns. However, these conventional methods often fall short when confronted with sophisticated camouflage scenarios, where weeds closely mimic the visual characteristics of the desired crops (López-Granados et al., 2006).

Weed detection systems that can identify and differentiate camouflaged weeds from crops are becoming increasingly crucial for sustainable agriculture (Singh et al., 2024). These systems not only promise to reduce herbicide usage but also offer more precise and environmentally friendly weed control solutions. By leveraging advanced image processing techniques, machine learning algorithms, and sophisticated pattern recognition methods, camouflage-based weed detection represents a promising frontier in agricultural technology (Coleman et al., 2023).

This research explores various methodologies and techniques for detecting weeds that exhibit camouflage characteristics, focusing on overcoming the challenges posed by visual similarities between weeds and crops. The study examines both classical computer vision approaches and cutting-edge deep learning methods, aiming to develop more accurate and reliable detection systems for practical agricultural applications.

To address this work in detail, the manuscript is organized as follows. Section 2 introduces the background on using weed detection techniques based on classical and deep learning, and addresses the problem as a camouflaged object detection approach. Sec-

Velesaca, H. O., Villegas, H., Sappa and A. D.
Exploring Camouflaged Object Detection Techniques for Invasive Vegetation Monitoring.
DOI: 10.5220/0013630200003967
In *Proceedings of the 14th International Conference on Data Science, Technology and Applications (DATA 2025)*, pages 619-626
ISBN: 978-989-758-758-0; ISSN: 2184-285X
Copyright © 2025 by Paper published under CC license (CC BY-NC-ND 4.0)

tion 3 presents the proposed approach to identify weeds using camouflage detection techniques. Then, Section 4 shows the experimental results taking as reference a dataset with aerial images with the presence of weeds and a case study of banana crops evaluating COD techniques. Finally, conclusions are presented in Section 5.

2 BACKGROUND

The field of weed detection has evolved significantly over the past decades, transitioning from manual inspection methods to sophisticated automated systems. This evolution has been particularly marked by the integration of artificial intelligence and specialized detection techniques to address complex scenarios such as camouflaged weeds in agricultural settings.

Deep learning has revolutionized the field of computer vision and, by extension, weed detection systems. A recent work by Rehman et al. (Rehman et al., 2024) demonstrates the effectiveness of drone-based weed detection using feature-enriched deep learning approaches, achieving significant improvements in detection accuracy across various agricultural scenarios. This builds upon foundational work by Tang et al. (Tang et al., 2017), who pioneered the combination of K-means feature learning with convolutional neural networks for weed identification, establishing early benchmarks for automated detection systems. Further advances are made by Balabantaray et al. (Balabantaray et al., 2024), who have developed targeted weed management systems using robotics and YOLOv7 architecture, specifically focusing on Palmer amaranth detection and demonstrating the practical application of deep learning in real-world agricultural settings.

A particularly challenging aspect of weed detection involves scenarios where weeds exhibit camouflage characteristics within their environment. In (Singh et al., 2024), the authors address this complex issue through an innovative approach for identifying small and multiple weed patches using drone imagery. Their work specifically tackles the challenge of detecting weeds in scenarios where they naturally blend with crops, introducing novel techniques for distinguishing camouflaged weeds in complex agricultural environments. Their methodology demonstrated remarkable success in detecting small-scale infestations that are typically difficult to identify due to their visual similarity with surrounding vegetation. This breakthrough in handling camouflaged scenarios has opened new possibilities for addressing one of the most challenging aspects of automated weed detection systems.

3 PROPOSED STUDY

This section details the different stages followed to carry out the proposed study.

3.1 Dataset Description

Weed detection and classification datasets represent crucial resources in agricultural technology, playing a vital role in developing automated systems for sustainable farming practices. These specialized collections of images are specifically designed to address the challenges in identifying and managing unwanted vegetation in various agricultural settings. The datasets used for this work (e.g., (Future, 2024a), (Future, 2024b)) encompass different weed species captured under different lighting conditions, growth stages, and field scenarios, making them particularly valuable for developing robust detection systems. The images have been captured from aerial shots. These datasets serve as fundamental tools in developing more sophisticated and efficient weed management solutions, ultimately contributing to more sustainable and productive agricultural practices, while also highlighting the growing importance of data-driven approaches in modern agriculture.

3.2 COD Approaches

Camouflage Object Detection (COD) represents one of the most fascinating and complex challenges in computer vision. This problem, inspired by natural phenomena where certain organisms have evolved to blend with their surroundings, has motivated the development of various innovative techniques in recent years. The ability to identify objects that deliberately try to confuse themselves with their environment not only has applications in security and surveillance but also in biology, ecology, and agricultural applications, which is the focus of this investigation. Below are the most significant contributions of the state-of-the-art in the COD task that will be used to carry out the different experiments. It is worth mentioning that this approach to weed detection using COD techniques has not been widely addressed; there is only one precedent (e.g., (Singh et al., 2024)).

In the current work, off-the-shelf COD approaches are evaluated in the datasets mentioned above (e.g., (Future, 2024a), (Future, 2024b)). Table 1 shows the input size of the image, the backbone used as well as the number of parameters of each technique. These approaches are briefly described next.

The first chosen technique has been proposed by Fan et al. (Fan et al., 2021) (SINet-v2) it establishes a

Table 1: Comparison between different characteristics of the camouflage techniques used.

Technique	Year	Input size (px)	Backbone	#Param. (M)
SINet-v2 (Fan et al., 2021)	2021	352×352	Res2Net50 (Gao et al., 2019)	24.93
BGNet (Chen et al., 2022b)	2022	416×416	Res2Net50 (Gao et al., 2019)	77.80
C^2F-Net (Chen et al., 2022a)	2022	352×352	Res2Net50 (Gao et al., 2019)	26.36
DGNet (Ji et al., 2023)	2023	352×352	EfficientNet (Tan and Le, 2019)	8.30
HitNet (Hu et al., 2023)	2023	352×352	PVTv2 (Wang et al., 2022)	25.73
PCNet (Yang et al., 2024)	2024	352×352	PVTv2 (Wang et al., 2022)	27.66



Figure 1: Example of images from the datasets presented in (Future, 2024a) and (Future, 2024b).

fundamental benchmark in detecting camouflaged objects. Their pioneering research introduces a convolutional neural network-based approach that revolutionizes how this challenge is addressed. The technique is distinguished by its ability to extract distinctive features that allow differentiating camouflaged objects from their background, thus establishing a solid foundation for future research in the field. Building upon this foundational work and expanding its capabilities, Ji et al. (Ji et al., 2023) (DGNet) present an innovative approach based on deep gradient learning. Their method stands out for meticulously analyzing gradual changes in visual patterns, using gradient information to identify subtle differences between camouflaged objects and their surroundings. This technique excels not only in its computational efficiency but also in its ability to detect objects in highly complex camouflage situations, complementing and enhancing the features established by SINet-v2.

Taking this progress in a more specialized direction while incorporating these established principles, Yang et al. developed PlantCamo (Yang et al., 2024) (PCNet), a technique specifically designed for detecting camouflage in plants. This method represents a significant advance in understanding natural camouflage in the plant kingdom, incorporating specific biological knowledge and unique plant camouflage patterns to improve detection accuracy while building upon the gradient analysis concepts introduced by DGNet. Further advancing these developments and integrating previous insights, the contribution of Hu et al. (Hu et al., 2023) (HitNet) introduces a high-resolution iterative feedback network that marks an important milestone. Their innovative method uses an iterative process that progressively refines detection results, leveraging high-resolution information to capture the most subtle details in images. This feedback approach allows for continuous improvement in detection accuracy while incorporating elements from both SINet-v2's feature extraction and DGNet's gradient analysis.

Contrary to the previous approaches, Chen et al. (Chen et al., 2022b) (BGNet) propose a unique perspective with their boundary-guided network. This technique is distinguished by its focus on the precise identification of camouflaged object boundaries, using sophisticated edge and contour information to improve segmentation. The method proves particularly effective in cases where the boundaries between the camouflaged object and the background are especially diffuse, complementing the high-resolution analysis of HitNet. Building upon all these advances and synthesizing their strengths, Chen et al. (Chen et al., 2022a) (C^2F-Net) present an innovative method based on context-aware cross-level fusion. This technique represents a significant advance by integrating information from multiple feature levels, considering both spatial and semantic contexts. The intelligent fusion of different levels of information allows for a more holistic understanding of the scene, significantly improving detection robustness across various camouflage scenarios while incorporating the boundary

awareness of BGNet and the iterative refinement of HitNet.

Recent advances in camouflaged object detection have enhanced our detection capabilities, creating opportunities across biodiversity conservation, security, and scientific research. While various techniques offer complementary approaches that contribute to the field's development, and despite rapid progress in computer vision and machine learning, challenges persist in achieving accurate detection and segmentation of camouflaged objects in real-world scenarios.

3.3 Metric Evaluation

To evaluate results for COD approaches different evaluation metrics have been proposed in the literature in the current work, five widely used metrics for COD tasks are adopted to evaluate the detection results of each model, namely, the S-measure (S_α) (Fan et al., 2017), weighted F-measure (F_β^w) (Margolin et al., 2014), Mean Absolute Error (M) (Perazzi et al., 2012), E-measure (E_ϕ) (Fan et al., 2018), and F-measure (F_β) (Achanta et al., 2009). S_α computes the structural similarity between prediction and ground truth. F_β^w is an enhanced evaluation metric that extends the traditional F_β by incorporating spatial weights to better assess segmentation quality, particularly emphasizing boundary accuracy and location-based importance of detected pixels in object detection tasks. M focuses on evaluating the error at the pixel level between the normalized prediction and the ground truth. E_ϕ simultaneously assesses the overall and local accuracy of COD based on the human visual perception mechanism. F_β is an overall measure that synthetically considers both precision and recall.

Since different scores of F-measure can be obtained according to different precision-recall pairs, there are the mean F-measure (F_β^{mean}) and the maximum F-measure (F_β^{max}). Similar to the F-measure, maximum, and mean denoted as E_ϕ^{mean} and E_ϕ^{max} are also used as evaluation metrics.

4 RESULTS AND DISCUSSION

This section presents the results obtained with the proposed study. For the performance evaluation, the metrics described in Sec. 3.3 are used.

4.1 Dataset

The dataset used for the comparisons is obtained from (Future, 2024a) and (Future, 2024b), each dataset consists of 672 and 318 images, respectively, and both datasets are merged to get a unique dataset of 990 images. Table 2 shows the dataset distribution for each subset. Additionally, Figure 1 shows examples of the images that are part of the dataset.

Table 2: Distribution of dataset.

Task	**# of images with class weed**		
	Dataset 1	**Dataset 2**	**Dataset**
Training	543	300	843
Validation	91	10	101
Testing	38	8	46
Total	672	318	990

4.2 COD Techniques

To carry out the study, six different COD techniques are used (i.e., SINet-v2 (Fan et al., 2021), BGNet (Chen et al., 2022b), C^2F-Net(Chen et al., 2022a), DGNet (Ji et al., 2023), HitNet (Hu et al., 2023), PCNet (Yang et al., 2024)). Table 1 shows the comparison between different characteristics of each of the camouflage techniques. With each architecture, a fine-tuning of 100 epochs is carried out. Most techniques operate with a standard input size of 352×352 pixels, except BGNet, which uses a slightly larger input size of 416×416 pixels. This consistency in input size among most models suggests a standardized approach to image processing in this field.

Regarding the backbone architectures, there is a notable variety in the choices made by different researchers. Three of the techniques (SINet-v2, BGNet, and C^2F-Net) utilize Res2Net50 as their backbone network. PCNet and HitNet opt for PVT-V2 (Pyramid Vision Transformer V2), while DGNet employs EfficientNet, showing the diversity in architectural approaches to the problem.

The number of parameters varies significantly across these models, ranging from 8.30 million in DGNet to 77.80 million in BGNet. DGNet stands out as the most parameter-efficient model with just 8.30M parameters, while BGNet represents the other extreme with 77.80M parameters. The remaining models (SINet-v2, PCNet, HitNet, and C^2F-Net) fall within a relatively similar range, between 24-28 million parameters, suggesting a common sweet spot for model complexity in this domain.

4.3 Comparisons

The comprehensive evaluation of six different COD architectures reveals significant insights into their performance in weed detection scenarios.

The quantitative analysis for the dataset described

Table 3: Metric evaluation results for each COD techniques using the state-of-the-art datasets (e.g., (Future, 2024a), (Future, 2024b))—notation as presented in Sec. 3.3. Top 3 results are shown in red, blue, and green.

Technique	$S_\alpha \uparrow$	$F_\beta^w \uparrow$	$M \downarrow$	$E_\phi^{mean} \uparrow$	$E_\phi^{max} \uparrow$	$F_\beta^{mean} \uparrow$	$F_\beta^{max} \uparrow$
SINet-v2 (Fan et al., 2021)	0.7251	0.4738	0.0611	0.7943	0.8302	0.5110	0.5358
BGNet (Chen et al., 2022b)	0.7075	0.4468	0.0945	0.8203	0.8922	0.5799	0.6201
C^2F-Net (Chen et al., 2022a)	0.6297	0.2547	0.1357	0.6557	0.7386	0.3627	0.3962
DGNet (Ji et al., 2023)	0.6931	0.4019	0.0779	0.7481	0.8013	0.4379	0.4621
HitNet (Hu et al., 2023)	0.7631	0.5594	0.0458	0.8617	0.8828	0.5816	0.5972
PCNet (Yang et al., 2024)	0.7612	0.5516	0.0450	0.8517	0.8877	0.5780	0.5938

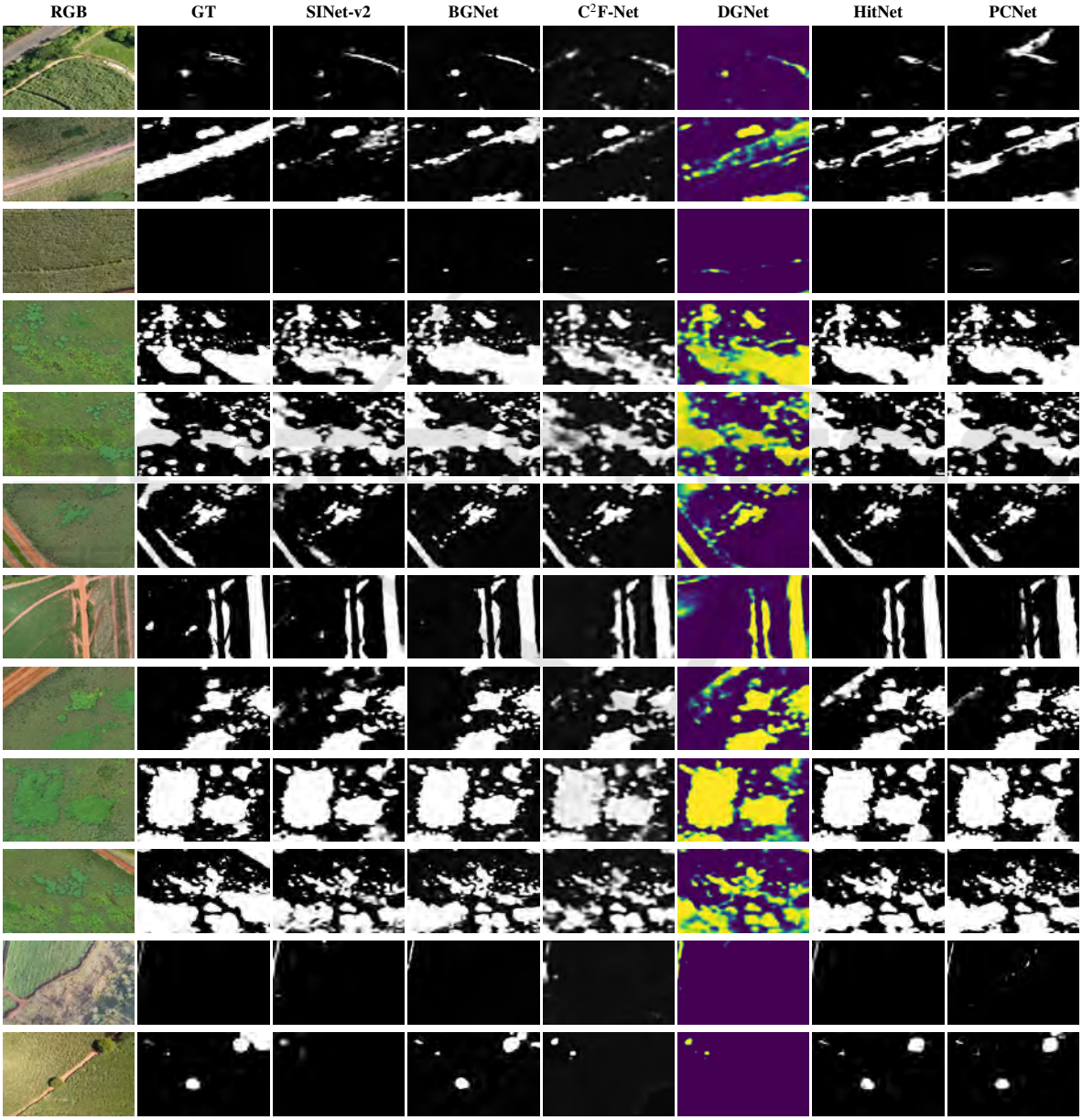

Figure 2: Prediction results using different state-of-the-art camouflage techniques. These example UAV images are part of the testing set (e.g., (Future, 2024a), (Future, 2024b)).

Table 4: Metric evaluation results on banana testing images with the presence of weeds. This is our dataset and consists of three images. Top 3 results are shown in red, blue, and green.

Technique	$S_\alpha \uparrow$	$F_\beta^w \uparrow$	$M \downarrow$	$E_\phi^{mean} \uparrow$	$E_\phi^{max} \uparrow$	$F_\beta^{mean} \uparrow$	$F_\beta^{max} \uparrow$
SINet-v2 (Fan et al., 2021)	0.7254	0.5228	0.0348	0.8383	0.9631	0.5833	0.6353
BGNet (Chen et al., 2022b)	0.3301	0.0262	0.5185	0.2579	0.8176	0.0311	0.0582
C^2F-Net (Chen et al., 2022a)	0.4521	0.0528	0.1867	0.5711	0.8206	0.0710	0.0857
DGNet (Ji et al., 2023)	0.7048	0.4341	0.0498	0.8034	0.9489	0.5283	0.6262
HitNet (Hu et al., 2023)	0.8013	0.6280	0.0267	0.9394	0.9692	0.6849	0.6976
PCNet (Yang et al., 2024)	0.6655	0.3199	0.1033	0.7060	0.7848	0.3846	0.4697

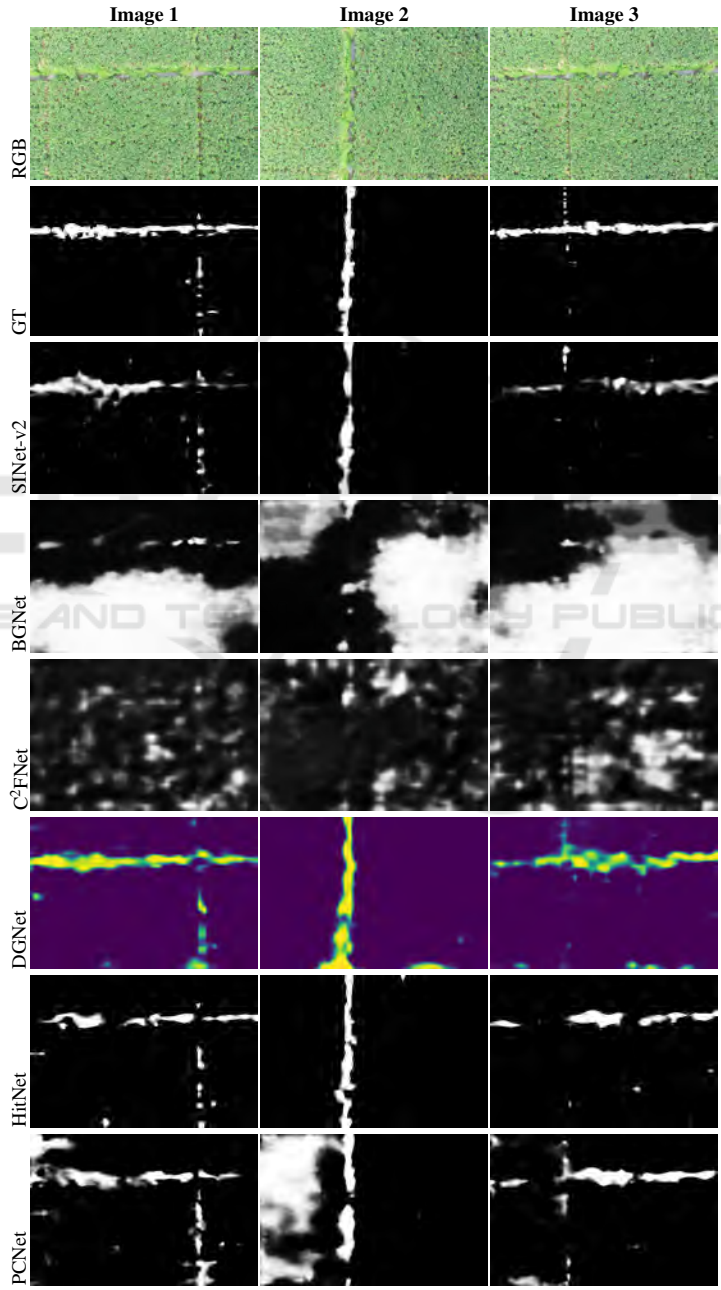

Figure 3: Prediction results of test UAV images belonging to banana crops with the presence of weeds. This is our dataset and consists of three images.

in Section 4.1 demonstrates that HitNet and PCNet consistently emerged as the leading performers across multiple evaluation metrics. HitNet achieved particularly impressive results with a S_α of 0.7631 and F_β^w of 0.5594, while PCNet closely follows with comparable performance metrics. In contrast, C^{2}F-Net shows the lowest performance across most metrics, indicating potential limitations in its application to weed detection tasks. The error analysis provides further validation of these findings, with PCNet and HitNet achieving the lowest M scores of 0.0450 and 0.0458, respectively. This represents a significant improvement over C^{2}F-Net, which shows the highest error rate with an M of 0.1357. The substantial difference in M between the best and worst-performing models (0.0907) shows the importance of architectural choices in achieving reliable weed detection results. In terms of boundary detection, BGNet demonstrates strong performance with an E_ϕ^{max} of 0.8922, while both HitNet and PCNet maintain consistent performance across both mean and maximum metrics. Table 3 shows metric evaluation results for each COD technique.

The qualitative analysis through visual results for the dataset described in Section 4.1 reveals important practical implications. HitNet and PCNet produce notably clearer boundaries between weed and non-weed regions, while BGNet shows good boundary detection but exhibits some over-segmentation tendencies. SINet-v2 and DGNet demonstrate moderate levels of false positives, whereas HitNet achieves a better balance between detection accuracy and false positives. C^{2}F-Net's tendency to under-detect camouflaged regions suggests limitations in its ability to handle subtle vegetation differences. On the other hand, it is also important to mention that the first four techniques in Fig. 2 (i.e., SINet-v2 (Fan et al., 2021), DGNet (Ji et al., 2023), PCNet (Yang et al., 2024), and HitNet (Hu et al., 2023)) can identify areas with very low weed density (see Fig. 2 *row 9 to 14*) identifying images where the presence of weeds is below 2% of the entire image, which indicates a good generalization capacity of these models for this type of datasets. Figure 2 shows some illustrations of the results obtained with the approaches evaluated in the current work.

To test the generalization of the different COD techniques, a dataset is captured and labeled that includes three images of banana crops in the presence of weeds. This dataset is not part of the training stage, only the pre-trained weights are used to test the generalization capacity of the architectures. Among the results obtained in banana crop scenarios, it can be highlighted the that HitNet significantly outperforms other models, obtaining the best result in all metrics. On the other hand, the SINet-v2 and DGNet metrics, in second and third place respectively, show regular performance, although not comparable with HitNet. Table 4 shows metric evaluation results on banana testing images in the presence of weeds. This specialized performance indicates HitNet's robust capability in handling complex agricultural scenarios. Although BGNet and C^{2}F-Net struggle significantly with banana crop images, SINet-v2 and DGNet maintain consistent performance, although PCNet shows somewhat reduced effectiveness compared to general testing scenarios. Figure 3 shows the prediction results of test images belonging to banana crops in the presence of weeds.

5 CONCLUSIONS

The comprehensive evaluation proposed in this work of six state-of-the-art COD techniques for weed detection has yielded several significant findings. HitNet and PCNet demonstrate superior performance in quantitative evaluation metrics for the dataset described in Section 4.1, with HitNet achieving the best overall results. These results can be contrasted with the qualitative analysis, showing that SINet-v2, PCNet, and HitNet present visual results where the weeds are correctly segmented. On the other hand, for the analysis carried out with images of banana crops with the presence of weeds, HitNet demonstrated exceptional performance in this type of scenario, indicating its robust capability of the model in specialized agricultural contexts despite not having been trained with images of banana crops and weeds. The significant variation in performance across different architectures shows the importance of model selection for specific agricultural applications. In summary, the success of the application of these COD techniques in the weed detection context demonstrates the viability of treating weed identification as a camouflage detection problem, opening new avenues for the field of precision agriculture.

ACKNOWLEDGEMENTS

This work was supported in part by the Air Force Office of Scientific Research Under Award FA9550-24-1-0206; in part by the ESPOL project "Advancing Camouflaged Object Detection with a cost-effective Cross-Spectral vision system (ACODCS)" (CIDIS-003-2024); and in part by the University of Granada.

REFERENCES

Achanta, R., Hemami, S., Estrada, F., and Susstrunk, S. (2009). Frequency-tuned salient region detection. In *Conf. on Computer Vision and Pattern Recognition*, pages 1597–1604. IEEE.

Agarwal, R., Hariharan, S., Rao, M. N., and Agarwal, A. (2021). Weed identification using k-means clustering with color spaces features in multi-spectral images taken by uav. In *Int. Geoscience and Remote Sensing Symposium*, pages 7047–7050. IEEE.

Balabantaray, A., Behera, S., Liew, C., Chamara, N., Singh, M., Jhala, A. J., and Pitla, S. (2024). Targeted weed management of palmer amaranth using robotics and deep learning (yolov7). *Frontiers in Robotics and AI*, 11:1441371.

Chauhan, B. S., Matloob, A., Mahajan, G., Aslam, F., Florentine, S. K., and Jha, P. (2017). Emerging challenges and opportunities for education and research in weed science. *Frontiers in plant science*, 8:1537.

Chen, G., Liu, S.-J., Sun, Y.-J., Ji, G.-P., Wu, Y.-F., and Zhou, T. (2022a). Camouflaged object detection via context-aware cross-level fusion. *Transactions on Circuits and Systems for Video Technology*, 32(10):6981–6993.

Chen, T., Xiao, J., Hu, X., Zhang, G., and Wang, S. (2022b). Boundary-guided network for camouflaged object detection. *Knowledge-based systems*, 248:108901.

Coleman, G. R., Bender, A., Walsh, M. J., and Neve, P. (2023). Image-based weed recognition and control: Can it select for crop mimicry? *Weed Research*, 63(2):77–82.

Fan, D.-P., Cheng, M.-M., Liu, Y., Li, T., and Borji, A. (2017). Structure-measure: A new way to evaluate foreground maps. In *Int. Conf. on Computer Vision*, pages 4548–4557.

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

Fan, D.-P., Ji, G.-P., Cheng, M.-M., and Shao, L. (2021). Concealed object detection. *Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6024–6042.

Future (2024a). Machine segmentation dataset. https://universe.roboflow.com/future-cn6m4/machine-segmentation. visited on 2024-12-27.

Future (2024b). Weed segmentation dataset. https://universe.roboflow.com/future-cn6m4/weed-segmentation-mjudm. visited on 2024-12-27.

Gao, S.-H., Cheng, M.-M., Zhao, K., Zhang, X.-Y., Yang, M.-H., and Torr, P. (2019). Res2net: A new multi-scale backbone architecture. *Transactions on Pattern Analysis and Machine Intelligence*, 43(2):652–662.

Hu, X., Wang, S., Qin, X., Dai, H., Ren, W., Luo, D., Tai, Y., and Shao, L. (2023). High-resolution iterative feedback network for camouflaged object detection. In *AAAI Conf. on Artificial Intelligence*, volume 37, pages 881–889.

Ji, G.-P., Fan, D.-P., Chou, Y.-C., Dai, D., Liniger, A., and Van Gool, L. (2023). Deep gradient learning for efficient camouflaged object detection. *Machine Intelligence Research*, 20(1):92–108.

López-Granados, F., Jurado-Expósito, M., Peña-Barragán, J. M., and García-Torres, L. (2006). Using remote sensing for identification of late-season grass weed patches in wheat. *Weed Science*, 54(2):346–353.

Margolin, R., Zelnik-Manor, L., and Tal, A. (2014). How to evaluate foreground maps? In *Conference on Computer Vision and Pattern Recognition*, pages 248–255.

Moldvai, L., Ambrus, B., Teschner, G., and Nyéki, A. (2024). Weed detection in agricultural fields using machine vision. In *BIO Web of Conferences*, volume 125. EDP Sciences.

Parra, L., Marin, J., Yousfi, S., Rincón, G., Mauri, P. V., and Lloret, J. (2020). Edge detection for weed recognition in lawns. *Computers and Electronics in Agriculture*, 176:105684.

Perazzi, F., Krähenbühl, P., Pritch, Y., and Hornung, A. (2012). Saliency filters: Contrast based filtering for salient region detection. In *Conf. on Computer Vision and Pattern Recognition*, pages 733–740. IEEE.

Rehman, M. U., Eesaar, H., Abbas, Z., Seneviratne, L., Hussain, I., and Chong, K. T. (2024). Advanced drone-based weed detection using feature-enriched deep learning approach. *Knowledge-Based Systems*, 305:112655.

Singh, V., Singh, D., and Kumar, H. (2024). Efficient application of deep neural networks for identifying small and multiple weed patches using drone images. *Access*.

Tan, M. and Le, Q. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114. PMLR.

Tang, J., Wang, D., Zhang, Z., He, L., Xin, J., and Xu, Y. (2017). Weed identification based on k-means feature learning combined with convolutional neural network. *Computers and electronics in agriculture*, 135:63–70.

Wang, W., Xie, E., Li, X., Fan, D.-P., Song, K., Liang, D., Lu, T., Luo, P., and Shao, L. (2022). Pvt v2: Improved baselines with pyramid vision transformer. *Computational Visual Media*, 8(3):415–424.

Westwood, J. H., Charudattan, R., Duke, S. O., Fennimore, S. A., Marrone, P., Slaughter, D. C., Swanton, C., and Zollinger, R. (2018). Weed management in 2050: Perspectives on the future of weed science. *Weed science*, 66(3):275–285.

Wu, Z., Chen, Y., Zhao, B., Kang, X., and Ding, Y. (2021). Review of weed detection methods based on computer vision. *Sensors*, 21(11):3647.

Yang, J., Wang, Q., Zheng, F., Chen, P., Leonardis, A., and Fan, D.-P. (2024). Plantcamo: Plant camouflage detection. *arXiv*.



Enhancing Camouflaged Object Detection via Diffusion Model Augmentation

Henry O. Velesaca[1,2] and Angel D. Sappa[1,3(✉)]

[1] ESPOL Polytechnic University, Campus Gustavo Galindo, Guayaquil, Ecuador
hvelesac@espol.edu.ec
[2] Software Engineering Department, Research Center for Information and Communication Technologies (CITIC-UGR), University of Granada, 18071 Granada, Spain
[3] Computer Vision Center, 08193 Bellaterra, Spain
sappa@ieee.org

Abstract. This paper proposes a novel approach to improve camouflaged object detection (COD) techniques using a data augmentation strategy based on a diffusion model. COD represents a significant challenge in computer vision, being crucial in applications such as military surveillance, medical image analysis, and ecological studies. Although recent advances in deep learning have improved COD performance, they heavily rely on well-annotated datasets, which are particularly difficult to obtain in the context of camouflaged objects. The proposed method addresses the problem of data sparsity while maintaining crucial features of camouflaged objects and their relationships with the environment. Incorporating a diffusion model into the data augmentation pipeline shows improvement in the performance of the COD models, which varies between the different techniques and the proposed evaluation metrics.

Keywords: Camouflaged object detection · Computer vision · Diffusion models · Pest detection

1 Introduction

Camouflaged Object Detection (COD) represents a significant challenge in computer vision, where the goal is to identify objects that are deliberately or naturally blended with their surroundings [5]. This task is particularly crucial in various applications, including military surveillance, medical image analysis, and ecological studies [17]. Unlike traditional object detection, COD faces unique challenges due to the inherent similarity between target objects and their backgrounds, making it difficult for conventional detection methods to achieve satisfactory results [28]. Recent advances in deep learning have significantly improved COD performance [31]. However, these improvements are heavily dependent on large-scale and well-annotated datasets, which are particularly challenging to obtain in the context of camouflaged objects due to the time-consuming nature

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
E. Rodolà et al. (Eds.): ICIAP 2025 Workshops, LNCS 16169, pp. 352–364, 2026.
https://doi.org/10.1007/978-3-032-11317-7_30

of data collection and annotation [14]. This limitation often leads to models that underperform in real-world scenarios or fail to generalize across different camouflage patterns and environments.

To address these challenges, data augmentation has emerged as a promising solution [13]. Traditional augmentation techniques, such as geometric transformations and color adjustments, while useful, often fail to capture the complex nature of camouflage patterns and their interaction with the environment [30]. This limitation has motivated the exploration of more sophisticated approaches to data augmentation.

In recent years, diffusion models have demonstrated remarkable capabilities in generating high-quality, diverse images [19]. These models, which learn to gradually denoise images through an iterative process, have shown particular promise in maintaining structural coherence and semantic relationships in generated images [20]. However, their application to the specific challenge of COD data augmentation remains largely unexplored. This paper proposes a methodology to improve COD techniques using data augmentation with diffusion models. Our method addresses the data scarcity problem while maintaining the crucial characteristics of camouflaged objects and their relationships with their surroundings. By incorporating diffusion models into the data augmentation pipeline, we can generate synthetic yet realistic training samples that enhance the robustness and generalization capabilities of COD models.

The manuscript is organized as follows. Section 2 presents work related to camouflaged object detection and data augmentation using classical and diffusion model approaches. Section 3 presents the proposed methodology to improve camouflaged object detection using specialized techniques for this task. Experimental results, comparisons, and discussions with different COD approaches to validate the proposed methodology are presented in Sect. 4. Finally, conclusions are presented in Sect. 5.

2 Background

This section summarizes some of the most relevant techniques related to the camouflaged object detection techniques and data augmentation approaches.

2.1 Camouflaged Object Detection Approaches

Camouflaged Object Detection has emerged as a crucial research field in computer vision, with significant advances in recent years. The early work of Qin et al. [18] on salient object detection significantly influences the development of COD techniques. In 2021, Fan et al. [5] establish the fundamental foundations for concealed object detection, providing a robust theoretical framework for this challenging problem. In 2022, Chen et al. [3] propose a context-aware cross-level fusion-based approach, improving the accuracy of camouflaged object identification, while Chen et al. [4] introduce a boundary-guided network that enhances the precision in detecting edges and features of camouflaged objects. In the same

year, Liu et al. [16] address the modeling of aleatoric uncertainty in camouflaged object detection. The field continues to evolve in 2023 with significant contributions such as the high-resolution iterative feedback network presented by Hu et al. [9], and the edge-aware network for COD presented by Sun et al. [24]. In the same year, Ji et al. [11] introduce an innovative deep gradient learning technique that significantly improves the efficiency of camouflaged object detection. More recently, in 2024, Yang et al. [29] expand the field towards specific applications with PlantCamo, focusing on camouflage detection in plants.

These advances contribute to creating more robust and accurate systems for camouflaged object detection, setting new standards in the field and paving the way for future research and practical applications.

2.2 Data Augmentation Approaches

Data augmentation has become a fundamental technique in deep learning, especially in computer vision tasks, where the availability of labeled data may be limited. Classic augmentation techniques include geometric transformations such as rotation, flipping, cropping, and rescaling, as well as color space modifications such as changes in brightness, contrast, and saturation [22]. These basic transformations help artificially increase the size of the training dataset and improve the generalization capabilities of models. A popular tool that has revolutionized the implementation of these techniques is Albumentations [1], a Python library that provides an efficient and optimized interface for applying data augmentation transformations. This library has become a standard in the computer vision community due to its superior performance and wide range of available transformations.

Recent research uses synthetic data and domain adaptation to combat data scarcity. Virtual environments effectively train deep learning models for camera pose estimation [2] and image dehazing [21]. Synthetic datasets aid instance segmentation [26] and agricultural applications like corn kernel classification [23], reducing annotation costs. This approach, combined with traditional augmentation, improves model performance while decreasing reliance on manual labeling.

In a significant advancement in the field of deep learning and data augmentation, Islam et al. [10] present DiffuseMix, a groundbreaking data augmentation technique that preserves labels using diffusion models. This method addresses one of the fundamental challenges in machine learning: generating synthetic data that maintains the integrity of the original labels. DiffuseMix leverages the power of diffusion models to create realistic and meaningful variations of the training data while ensuring that semantic labels remain consistent and accurate. This approach not only improves the diversity of the training dataset but also helps prevent overfitting and improves the robustness of deep learning models. The technique demonstrates promising results on several benchmark datasets, establishing itself as a valuable contribution to the computer vision and machine learning community.

The Hidden Cost of Poor Annotations: How Label Quality Affects Camouflaged Object Detection Performance

Heny O. Velesaca[1,2(✉)], Andrea Mero[1,5], Anjali Singh[3], and Angel Sappa[1,4]

[1] ESPOL Polytechnic University, ESPOL, Campus Gustavo Galindo, Km. 30.5 Vía Perimetral, 090902 Guayaquil, Ecuador
{hvelesac,anmero,asappa}@espol.edu.ec

[2] Software Engineering Department, Research Center for Information and Communication Technologies (CITIC), University of Granada, 18071 Granada, Spain
hvelesaca@correo.ugr.es

[3] Birsa Institute of Technology, Sindri, Jharkhand University of Technology, JUT, Ranchi 834010, Jharkhand, India
director@bitsindri.ac.in

[4] Computer Vision Center, Universitat Autónoma de Barcelona, 08193 Bellaterra, Barcelona, Spain
asappa@cvc.uab.es

[5] Università della Svizzera italiana (USI), Via Giuseppe Buffi 13, Lugano, Switzerland
andrea.mero@usi.ch

Abstract. This paper presents an in-depth study on the impact of high-quality, comprehensive annotations on camouflaged object detection (COD) performance. We evaluate 13 state-of-the-art COD models trained on original annotations versus a re-annotated version created under stricter, more consistent guidelines using the Cotton Bollworm dataset. Experimental results demonstrate that enhanced annotation quality significantly improves both Intersection over Union (IoU) scores and instance recall, reducing undetected camouflaged objects by an average of 4.6% in Structure-measure and 7.0% in weighted F-measure. The re-annotation process identified 1.4% additional instances, with an average area refinement of 6.3%, primarily through improvements in boundary precision and the detection of previously missed instances. These findings underscore the crucial role of precise annotations in advancing COD performance and validate the data-centric AI paradigm, suggesting that systematic refinement of annotations should be prioritized in computer vision pipelines. The re-annotated dataset is available on GitHub https://cod-espol.github.io/ReannotatedCottonBollworm/.

Keywords: Camouflaged object detection · Data-centric · Annotation quality · Pest detection · Computer vision · Dataset re-annotation

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
G. Bebis et al. (Eds.): ISVC 2025, LNCS 16396, pp. 285–296, 2026.
https://doi.org/10.1007/978-3-032-14492-8_22

1 Introduction

Artificial intelligence (AI) systems rely on both models and data. Most of the recent advances have concentrated on improving model architectures, where data is often treated as *ground truth* without thorough examination of annotation quality. Recently, a new paradigm—known as data-centric AI—has emerged, first introduced by [5]. This approach emphasizes improving annotation quality rather than focusing solely on model refinement. The principle applies to all supervised learning-based methods that require annotated data for training. For instance, in the case of object detection, annotation quality is particularly crucial, as it involves not only assigning the correct class to each object but also accurately localizing all instances in an image with precise and consistent bounding boxes (e.g., [6,19]). Similarly, in edge detection, studies have shown that precise and consistent image annotations are critical to model performance, with re-annotation efforts significantly pushing the boundaries of achievable accuracy [23]. The previous examples are just a couple of cases that benefit from the improvement in the annotations.

Data-centric AI has recently been explored in the object detection problem. In this case, studies have proven that incomplete annotations, such as when some objects of a category are left without bounding boxes, not only reduce the accuracy measured by Intersection over Union (IoU) but also introduce false negatives that can bias the learning process and degrade the model's ability to detect all instances in real-world scenarios. Therefore, the impact of high-quality annotation extends beyond a quantitative improvement in IoU; it directly affects the completeness and exhaustiveness of object detection.

In the current work, we investigate how improving annotation quality affects both the IoU-based accuracy and the model's capacity to detect every instance of interest in the camouflage object detection (COD) task. Specifically, we compare the performance of 13 SOTA COD models trained and evaluated on the original dataset annotations against that obtained using a re-annotated version created under stricter, more meticulous guidelines. Our results show that high-quality, comprehensive annotations not only boost standard performance metrics (e.g., IoU) but also improve instance recall, reducing the number of missed camouflaged objects.

The remainder of the paper is organized as follows. Section 2 summarizes works related to data-centric approaches as well as state-of-the-art camouflaged detection approaches. Section 3 describes the dataset and re-annotation process together with a short introduction of the COD approaches evaluated in the current work; additionally, the metrics used for the evaluation are introduced. Finally, Sect. 4 presents experimental results and discussions. Conclusions are provided in Sect. 5.

2 Background

Data-centric AI is based on the key understanding that data is the most important resource for creating effective and high-performing AI systems. Unlike

model-centric AI development, which typically emphasizes model enhancements, data-centric AI directs attention toward methodically improving the data pipeline [33]. While data-centric AI and dataset improvement are related, they are not equivalent. The main difference between them is that data-centric AI modifies datasets primarily to increase model accuracy, whereas dataset improvement may involve broader motivations, such as redefining a problem [19]. In the current section, some of the most relevant data-centric approaches for the object detection task are presented.

One of the seminal data-centric approaches has been presented in [19], where the authors follow specific guidelines to improve annotation quality in ImageNet and MS COCO datasets. Similarly, [2] proposes a high-quality annotation process that includes data annotations and multistage hashing to avoid duplicate instances and noisy labels. In [15], the authors highlight the importance of high-quality annotations and emphasize how inconsistent data labeling can degrade model accuracy. Maintaining annotation consistency is challenging yet essential to mitigate bias and ensure fairness in models. Additionally, the high cost of human annotation further complicates this issue. These challenges may soon drive the adoption of systematic processes to enhance data quality, particularly in annotation practices.

Object detection represents a fundamental challenge in computer vision (CV), with three primary subfields: Generic Object Detection (GOD), Salient Object Detection (SOD), and Camouflaged Object Detection (COD) [10]. A study by [6] indicates that effective data annotation practices are fundamental to object detection performance, particularly for accurate feature extraction. The current work is focused on the COD category, which specifically addresses the segmentation of visually concealed objects, typically formulated as a binary segmentation problem [18]. Recent advances in COD, as demonstrated by [35], have been facilitated by improved availability of detailed segmentation labels. However, the annotation process for camouflaged objects remains exceptionally labor-intensive, creating a critical bottleneck for research progress. This challenge aligns with [31]'s observation that supervised learning approaches are fundamentally constrained by their dependence on large-scale, high-quality datasets, a requirement particularly difficult to satisfy in real-world applications where annotation complexity and data scarcity pose substantial limitations. Building upon these foundational challenges, Sect. 3 systematically examines SOTA COD techniques, analyzing their performance characteristics while considering these data constraints. To our knowledge, there is no in-depth study on the relation between the results of COD approaches and the quality of camouflaged object annotations.

3 Materials and Methods

This section presents the methodology employed to evaluate the impact of annotation quality on COD performance. It describes the Cotton Bollworm dataset and re-annotation process, introduces the 13 SOTA COD techniques evaluated, and details the comprehensive evaluation metrics used for performance assessment.

AVNet: Cross-Spectral Attention-Vision Model for Camouflaged Object Detection in Ecological Conservation

Henry O. Velesaca[1,2][a], Andrea Mero[1][b], Rafael E. Rivadeneira[1][c], Guillermo A. Castillo[1][d] and Angel D. Sappa[1,3][e]

[1]*ESPOL Polytechnic University, ESPOL, Campus Gustavo Galindo Km. 30.5 Via Perimetral, Guayaquil, Ecuador*

[2]*Software Engineering Department, University of Granada, C. Periodista Daniel Saucedo Aranda, Granada, Spain*

[3]*Computer Vision Center, Edificio O Campus UAB Bellaterra, Barcelona, Spain*

{*hvelesac, amero, rrivaden, guancast, asappa*}*@espol.edu.ec*

Keywords: Camouflaged Object Detection, Ecological Conservation, Bimodal Dataset, Computer Vision, Deep Learning.

Abstract: This work introduced AVNet, a novel attention-vision architecture for Camouflaged Object Detection (COD), optimized for ecological conservation. The proposed approach integrates an RGB-Thermal fusion approach with the Convolutional Block Attention Model (CBAM) within an encoder-decoder framework, enabling accurate detection of low-contrast and highly camouflaged targets. As an additional contribution, this study introduces the Bimodal Iguana Observational Set (BIOS), comprising 148 camouflaged RGB-Thermal registered image pairs, specifically collected to support COD research in wildlife conservation. Experimental results validate the model's robustness under challenging real-world conditions. The original code and dataset presented in the study are openly available in GitHub at https://cod-espol.github.io/AVNet.

1 INTRODUCTION

Camouflaged Object Detection (COD) aims to localize targets that intentionally blend into their surroundings, often exhibiting low contrast, complex textures, and misleading boundaries (Fan et al., 2020). This problem is particularly critical in ecological conservation, where the robust detection of wildlife under natural occlusion, foliage clutter, and varying illumination is essential for population monitoring, behavioral studies, and effective protection strategies (Velesaca and Sappa, 2025). Conventional RGB-based detectors tend to fail under these conditions due to weak photometric cues and the high similarity between targets and background. Thermal imaging offers complementary information by capturing emissivity patterns that are less sensitive to color and texture, but thermal cues alone can be noisy, low-resolution, or ambiguous in heterogeneous outdoor environments. These challenges motivate cross-spectral fusion strategies that can jointly exploit the complementary strengths of visible and thermal modalities ((Fan et al., 2022), (Patel and Chaudhary, 2019), (Davis and Sharma, 2007)).

Recent progress in COD has been driven by encoder–decoder architectures augmented with attention and multi-scale feature aggregation, as well as transformer-based backbones that enhance long-range context modeling ((Pang et al., 2022), (Sun et al., 2021)). While these approaches significantly improve structural fidelity and boundary precision, their performance in real-world conservation settings remains limited by low-contrast camouflage, scale and pose variation, and domain-specific artifacts such as foliage shadows or thermal reflections (Fan et al., 2022). Moreover, the lack of domain-focused, spatially-registered RGB–Thermal datasets with camouflaged wildlife further impedes progress and fair benchmarking for cross-spectral COD ((Yan et al., 2021a), (Le et al., 2019)).

To address these limitations, AVNet is introduced as a cross-spectral attention–vision model designed for COD in ecological conservation scenarios. AVNet integrates a PVTv2-B2 encoder to obtain hierarchical, multi-scale features and the Convolutional Block Attention Module (CBAM) to refine channel- and spatial-level representations, sup-

[a] https://orcid.org/0000-0003-0266-2465
[b] https://orcid.org/0000-0002-0319-2090
[c] https://orcid.org/0000-0002-5327-2048
[d] https://orcid.org/0009-0002-6105-4590
[e] https://orcid.org/0000-0003-2468-0031

pressing distractors while emphasizing subtle target cues ((Wang et al., 2022), (Woo et al., 2018)). A dedicated fusion block adaptively combines visible and thermal features through modality-specific attention gates and aggregation, and a multi-level supervision scheme stabilizes learning across scales with deep supervision (Xu et al., 2020). As a second contribution, we present the Bimodal Iguana Observational Set (BIOS), a domain-specific dataset of spatially registered RGB–Thermal image pairs featuring naturally camouflaged iguanas in outdoor habitats, constructed via robust feature matching (Lindenberger et al., 2023). BIOS provides a challenging and realistic benchmark for cross-spectral COD under variable illumination and complex vegetation backgrounds (see Fig. 1).

The manuscript is organized as follows. Section 2 introduces related work in 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 on the BIOS dataset. Finally, discussion and conclusions are given in Section 5 and Section 6 respectively.

2 RELATED WORKS

COD has attracted increasing attention in recent years, driven by encoder–decoder architectures with multi-scale feature aggregation and attention mechanisms ((Fan et al., 2020), (Pang et al., 2022), (Zhong et al., 2022)). Most existing approaches operate on the visible spectrum only and are typically designed and evaluated on generic COD datasets with limited focus on ecological conservation scenarios.

Within RGB-based COD, several recent methods are particularly relevant to the proposed approach. EAMNet (Sun et al., 2023) introduces an edge-aware mirror architecture that explicitly enhances boundary information while modeling contextual cues. Its design focuses on fine-grained contour refinement through specialized edge branches and feature mirroring, which improves the localization of thin or irregular structures that are common in camouflaged targets. However, EAMNet is restricted to RGB inputs and thus remains sensitive to low-contrast conditions and illumination changes, which frequently occur in outdoor ecological environments.

PCNet (Yang et al., 2024) is a PVTv2-based architecture tailored for plant camouflage detection. By leveraging pyramid vision transformers, PCNet captures long-range dependencies and mixed-scale patterns, making it effective at disentangling targets from highly textured and cluttered vegetation. Although PCNet demonstrates strong performance on plant-focused COD benchmarks, it still operates purely in the visible domain, limiting its robustness in situations where photometric cues are extremely weak or misleading, such as in heavily shadowed or color-similar backgrounds typical of wildlife habitats.

More recently, ARNet (Wang et al., 2025a) proposes an assisted refinement network built on a lightweight SMT-Tiny backbone. ARNet employs channel-wise information interaction and hierarchical refinement modules to enhance target representations while suppressing background noise. Its design offers a good trade-off between accuracy and efficiency, and achieves competitive results on several COD datasets. Nonetheless, similar to EAMNet and PCNet, ARNet does not exploit thermal information and thus cannot fully leverage modality complementarity in challenging outdoor conditions.

Beyond purely visible architectures, there has been growing interest in cross-spectral and thermal-based methods for detection and segmentation tasks, motivated by the complementary nature of RGB and infrared signals. Early works on infrared–visible fusion for surveillance and general object detection ((Davis and Sharma, 2007), (Patel and Chaudhary, 2019)) illustrated that thermal imagery can provide stable cues under poor lighting or low contrast, while RGB images contribute rich texture and color details. In ecological conservation, thermal cameras have been used to locate animals hidden by vegetation or operating at night, although many systems rely on heuristic fusion or operate on a single modality, which can lead to false alarms in thermally ambiguous backgrounds (e.g., sun-heated rocks or branches).

Despite these advances, few COD methods have been explicitly designed for cross-spectral RGB–Thermal fusion, and even fewer that target wildlife monitoring under realistic ecological conditions. Most existing RGB-Thermal architectures for object detection or segmentation are not tailored to camouflaged targets and typically lack attention mechanisms specialized for subtle, low-contrast boundaries. In addition, the limited availability of spatially registered RGB–Thermal datasets containing camouflaged animals remains a significant barrier, hindering both methodological progress and fair benchmarking in this domain ((Xu et al., 2020), (Yan et al., 2021b)).

In contrast to the above approaches, AVNet is a cross-spectral attention-vision model that jointly exploits visible and thermal modalities through an encoder–decoder architecture with CBAM attention

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

Technique	Source	Source Type	Year	Image Size (px)	Backbone	#Param. (M)
BASNet (Qin et al., 2019)	CVPR	Conference	2019	256×256	ResNet-34 (He et al., 2016)	87.06
SINet-v2 (Fan et al., 2022)	TPAMI	Journal	2021	352×352	Res2Net-50 (Gao et al., 2019)	24.93
BGNet (Chen et al., 2022b)	IJCAI	Conference	2022	416×416	Res2Net-50 (Gao et al., 2019)	77.80
C^2F-Net (Chen et al., 2022a)	TCSVT	Conference	2022	352×352	Res2Net-50 (Gao et al., 2019)	26.36
OCENet (Liu et al., 2022)	WACV	Conference	2022	352×352	ResNet-50 (He et al., 2016)	58.17
EAMNet (Sun et al., 2023)	ICME	Conference	2023	384×384	Res2Net-50 (Gao et al., 2019)	30.51
DGNet (Ji et al., 2023)	MIR	Journal	2023	352×352	EfficientNet (Tan and Le, 2019)	8.30
HitNet (Hu et al., 2023)	AAAI	Conference	2023	352×352	PVTv2 (Wang et al., 2022)	25.73
PCNet (Yang et al., 2024)	arXiv	-	2024	352×352	PVTv2 (Wang et al., 2022)	27.66
ARNet (Wang et al., 2025a)	ICMR	Conference	2025	416×416	SMT-Tiny (Lin et al., 2023)	12.82
CHNet (Wang et al., 2025b)	ICMR	Conference	2025	416×416	SMT-Tiny (Lin et al., 2023)	11.20
CTF-Net (Zhang et al., 2025)	CVIU	Journal	2025	384×384	PVTv2 (Wang et al., 2022)	64.48
AVNet (Ours)	VISAPP	Conference	2026	352×352	PVTv2 (Wang et al., 2022)	48.04



Figure 1: Examples of captured images showing corresponding registered RGB and Thermal pairs containing camouflaged iguanas in a natural environment from the BIOS dataset.

and an adaptive fusion block. While EAMNet, PCNet, and ARNet demonstrate the effectiveness of edge-aware refinement, transformer-based context modeling, and efficient channel interaction in the visible domain, AVNet extends these ideas to a bimodal RGB–Thermal setting and targets ecological conservation explicitly. Moreover, by introducing the BIOS dataset of spatially registered RGB-Thermal iguana images, this work provides both an architecture and a domain-specific benchmark for cross-spectral COD in realistic wildlife monitoring scenarios.

3 PROPOSED AVNet

Similar to most previous works (e.g., (Fan et al., 2020), (Fan et al., 2022), (Jiang et al., 2022), (Pang et al., 2022), (Zhong et al., 2022)), the current work adopts an encoder-decoder pipeline to build the proposed AVNet architecture. AVNet is designed as an end-to-end trainable framework, as illustrated in Fig. 2.

Overall Architecture. The proposed AVNet architecture, based on RGB-Thermal fusion and CBAM component, follows an encoder-decoder design for COD. The architecture integrates the attention refinement of the CBAM to identify objects that visually blend with their environment effectively. Additionally, utilize RGB-Thermal information to enhance the detection of objects present in the scene.

Encoder. The PVTv2-B2 (Wang et al., 2022) backbone encoder is adopted to provide hierarchical feature representations at different scales. Given an input RGB image $I \in R^{H \times W \times 3}$, the encoder extracts multi-scale feature maps across four different resolutions. These representations capture both low-level details and high-level semantics information essential for camouflaged objects.

Decoder. The architecture incorporates a progressive refinement strategy along the decoder path. The decoder is composed of a series of Decoder Blocks that perform upsampling of features from the previous level and integrate them with skip connections from the encoder. A dedicated Feature Aggregation module fuses features from multiple sources through 1×1 and 3×3 convolutions. Multiple segmentation heads at different decoder levels provide deep supervision signals, enhancing gradient flow and feature learning.

CBAM. Another important element within the

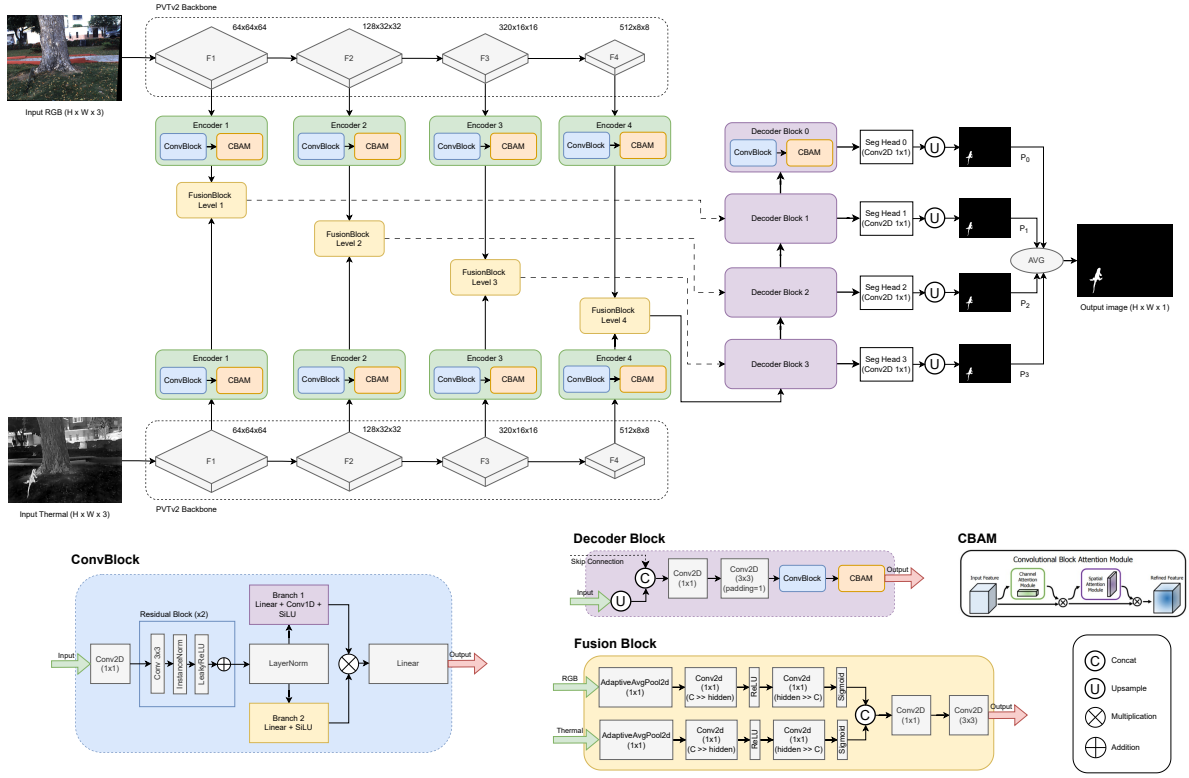

Figure 2: The overall architecture of the proposed AVNet.

proposed architecture is the Convolutional Block Attention Module (CBAM) (Woo et al., 2018), which refines features by applying both channel and spatial attention. The channel attention uses both average and max pooling operations, followed by a shared MLP to generate channel-wise attention weights. The spatial attention uses a channel pooling mechanism followed by a convolution to generate a spatial attention map.

Fusion Block. This is a neural network block designed to adaptively fuse two sets of features, from different modalities such as RGB and Thermal images. This block uses independent attention gates for each input, which generate channel-by-channel attention masks through a combination of pooling, convolutions, and activation functions. These masks allow the network to learn to dynamically highlight or attenuate the most relevant information from each source. Subsequently, the modulated features are concatenated and refined through an aggregation block, producing a fused representation that efficiently integrates the complementary information from both inputs. This mechanism improves the model's ability to leverage the strengths of each modality and facilitates smarter and more effective fusion for the segmentation task.

Multi-Level Supervision. To enhance the learning of discriminative features at different scales, it employs deep supervision with multiple segmentation heads. Each decoder level produces a segmentation map that is supervised by the GT. The final prediction is obtained by averaging the outputs from all levels. This multi-level supervision strategy helps the network learn more robust features for detecting camouflaged pests at different scales and with varying degrees of concealment.

Loss Function. In line with previous studies ((Fan et al., 2022), (Sun et al., 2021)), the loss function introduced by (Wei et al., 2020) is used in this work. The predictions produced by the AVNet decoder are denoted as $\{P_i\}_{i=0}^{3}$. During training, each prediction Pi is resized to the original size to match the input image dimensions and supervised using a combination of Binary Cross-entropy loss ($\mathcal{L}_{BCE}$) (De Boer et al., 2005) and the Intersection over Union loss ($\mathcal{L}_{IoU}$) (Máttyus et al., 2017). As described in (Fan et al., 2022), the overall loss is obtained by summing the losses from multiple prediction stages. Therefore, the total loss function for the AVNet model is thus defined as follows, where GT is the ground truth annotation:

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

4 EXPERIMENTAL RESULTS

Datasets. To evaluate the effectiveness of the proposed architecture in ecological conservation scenarios, a domain-specific dataset is employed: the BIOS dataset, which contains 148 camouflaged RGB-Thermal image pairs (see Fig. 1). This dataset focuses on iguanas naturally camouflaged in complex outdoor environments and is intended to support research in wildlife monitoring and camouflage object detection.

The image acquisition system for the BIOS dataset consists of a Basler acA1300-60gc camera, used to capture visible spectrum images at a resolution of 1280×1024 pixels with a 13mm lens, and a FLIR TAU2 thermal camera with a resolution of 640×480 pixels using an 8mm lens. Both cameras are rigidly mounted on a custom-designed platform to minimize parallax and ensure a consistent baseline between optical axes. This setup facilitates accurate cross-modal registration, which is crucial for aligning thermal and visible images ((Velesaca et al., 2024), (Rivadeneira et al., 2024)).

A total of 176 image pairs were originally captured under daylight conditions in natural iguana habitats. The pairs of images are subsequently aligned through a geometric registration process using LightGlue (Lindenberger et al., 2023) matching framework. After a thorough validation and cleaning procedure that involved removing misaligned or low-quality samples, the final dataset is curated to include 148 high-quality, spatially aligned image pairs.

This resulting BIOS dataset offers a realistic and challenging benchmark for evaluating camouflaged object detection algorithms in real-world scenarios. By integrating visible and thermal imagery registered using the LightGlue technique, the dataset provides a valuable resource for the COD research community, particularly for applications in biodiversity monitoring, ecological conservation, and thermal-vision-based wildlife detection.

Figure 3 shows a scatter plot of mask centroids. Each green dot marks one mask centroid on the image plane. Centroids are spread across the field, with a slightly higher concentration in the central region (around $X \approx 20$–60 and $Y \approx 50$–90). Fewer points appear near the edges, and this pattern suggests a broadly uniform distribution without distinct clusters or directional trends. For the purpose of displaying this graphic, the binary mask images are normalized to a size of 100×100 pixels for a standard reference.

Implementation Details. AVNet is implemented using the PyTorch library. The encoder utilizes a PVTv2-B2 (Wang et al., 2022) model pretrained on the ImageNet dataset. Optimization is performed using the AdamW algorithm with a weight decay $1e^{-4}$. The initial learning rate is set to $1e^{-4}$ and follows a cosine annealing schedule throughout training. All input images are resized to 416×416 for both training and inference. The model is trained end-to-end for 100 epochs with a batch size of 16 on an NVIDIA TESLA P100 GPU. All experiments are conducted on the Kaggle platform[1]. Also, Table 1 shows distinctive characteristics of the evaluated SOTA COD techniques in this work.

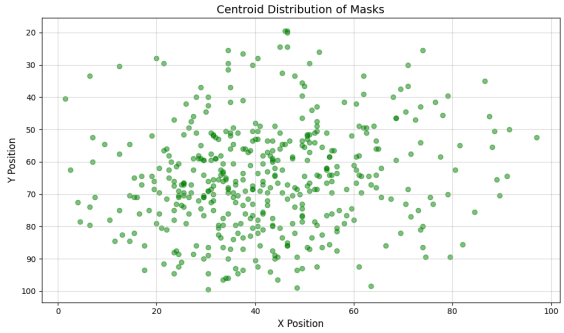

Figure 3: Centroid distribution of masks.

Metrics. This study employs five widely recognized evaluation metrics to evaluate COD performance. These metrics provide a comprehensive assessment criterion for analyzing detection accuracy and effectiveness across different models. The Structure-measure (S_α) (Fan et al., 2017), weighted F-measure (F_β^w) (Margolin et al., 2014), Mean Absolute Error (M) (Perazzi et al., 2012), E-measure (E_ϕ) (Fan et al., 2018), and F-measure (F_β) (Achanta et al., 2009). The S_α metric quantifies the structural similarity between prediction and GT maps. The F_β^w represents an enhanced evaluation metric that extends the traditional F_β by incorporating spatial weights, providing a better assessment of segmentation quality with emphasis on boundary accuracy and location-based importance of detected pixels. The M metric focuses on pixel-level error evaluation between the normalized prediction and GT. The E_ϕ metric simultaneously evaluates the global and local accuracy of COD based on human visual perception mechanisms. The F_β provides a synthetic measure that considers both precision and recall components. For both F-measure and E-measure metrics, different scores can be obtained according to different precision-recall pairs. This leads to the computation of mean F-measure (F_β^{mean}) and max F-measure (F_β^{mean}). Similarly, the E-measure utilizes mean variants, denoted as E_ϕ^{mean} and max E-measure as E_ϕ^{max}, which are also employed as evaluation metrics.

[1] https://www.kaggle.com/

4.1 Quantitative Evaluation

AVNet on the BIOS dataset is evaluated using standard COD metrics, including S_α, F_β^w, M, E_ϕ, and F_β, which together assess structural fidelity, boundary-aware precision–recall, pixel-wise error, and perceptual alignment. Table 2 shows quantitative results for SOTA COD techniques and AVNet. In the RGB-thermal modality, AVNet attains the best overall performance among all compared methods, achieving first place in $S_\alpha = 0.8951$, $F_\beta^w = 0.8404$, along with leading E_ϕ scores ($E_\phi^{adp} = 0.9835$, $E_\phi^{mean} = 0.9785$) and strong precision–recall behavior ($F_\beta^{adp} = 0.8190$, $F_\beta^{mean} = 0.8255$, $F_\beta^{max} = 0.8585$); and second place in $M = 0.0028$ and $E_\phi^{max} = 0.9837$. These results surpass recent strong baselines, such as ARNet and PCNet, particularly in terms of structural integrity and boundary precision, while reducing pixel-wise error. In the single visible modality, AVNet gets a second place in $E_\phi^{adp} = 0.9837$ metric. On the other hand, in single thermal mode, AVNet achieves acceptable but not exceptional results, ranking among the top performers in any metric. However, the cross-spectral design of AVNet yields more consistent gains, highlighting the benefits of RGB–Thermal fusion and CBAM-guided attention for robust COD.

4.2 Qualitative Evaluation

Qualitative comparisons (Fig. 4) demonstrate that AVNet generates cleaner, more complete segmentation masks with fewer missed regions (blue) and fewer over-segmented areas (red) relative to state-of-the-art baselines. In scenes characterized by foliage patterns, textured clutter, and low contrast between targets and background, the CBAM component enhances spatial focus. It suppresses distractors, while cross-spectral fusion leverages thermal cues to recover visually indistinct regions in RGB. The encoder–decoder design with multi-level supervision further stabilizes detection across scales and poses, yielding tighter boundary adherence and higher internal mask consistency with fewer holes and spurious fragments. These visual observations align with the quantitative gains on S_α, F_β, and E_ϕ, reinforcing AVNet's effectiveness in challenging ecological conditions.

5 DISCUSSION

This study demonstrates that cross-spectral fusion, when paired with targeted attention mechanisms, can substantially improve camouflaged object detection in realistic conservation settings. On the BIOS dataset, AVNet consistently outperforms strong RGB-only and Thermal-only baselines, particularly in structure-aware metrics (S_α) and boundary-sensitive measures (F_β, E_ϕ), and exhibits competitive or best-in-class pixel-wise error (M). The gains are most pronounced in scenes with low RGB contrast and heavy texture clutter, where thermal cues help recover target regions that are visually indistinguishable. Conversely, in thermally ambiguous conditions (e.g., sun-warmed rocks or branches), RGB cues mitigate false activations, illustrating the complementary nature of the two modalities. CBAM further enhances this synergy by suppressing background distractors and sharpening spatial focus around faint target boundaries, while multi-level supervision stabilizes learning across scales and improves mask completeness.

Beyond raw performance, the qualitative analysis indicates that AVNet reduces typical COD failure modes: missed fine structures (false negatives), boundary bleeding into background (false positives), and internal mask holes. These improvements are valuable for downstream ecological tasks such as automated counting, occupancy estimation, and behavioral analytics, where both detection sensitivity and spatial accuracy are critical. Importantly, AVNet's encoder–decoder design with lightweight attention remains compatible with real-time or near-real-time deployment on field hardware, especially if distilled or pruned variants are explored.

However, the dataset has some limitations. First, BIOS, while carefully curated and spatially registered, this process is very time-consuming since some of the images are registered manually. Second, thermal imaging quality and calibration can vary across devices and weather conditions, introducing cross-device variability that may affect fusion robustness.

These observations suggest several avenues for future work. Incorporating domain adaptation and calibration-aware fusion could improve robustness across environments. Finally, exploring semi- or weakly-supervised learning, as well as active learning strategies, could reduce annotation costs while maintaining or improving detection quality.

In summary, the evidence indicates that attention-guided cross-spectral fusion is a compelling direction for camouflaged object detection in ecological conservation. By coupling modality-complementary cues with principled attention and multi-scale supervision, AVNet delivers reliable, high-fidelity detections under realistic field conditions and establishes a strong foundation for next-generation, conservation-oriented COD systems.

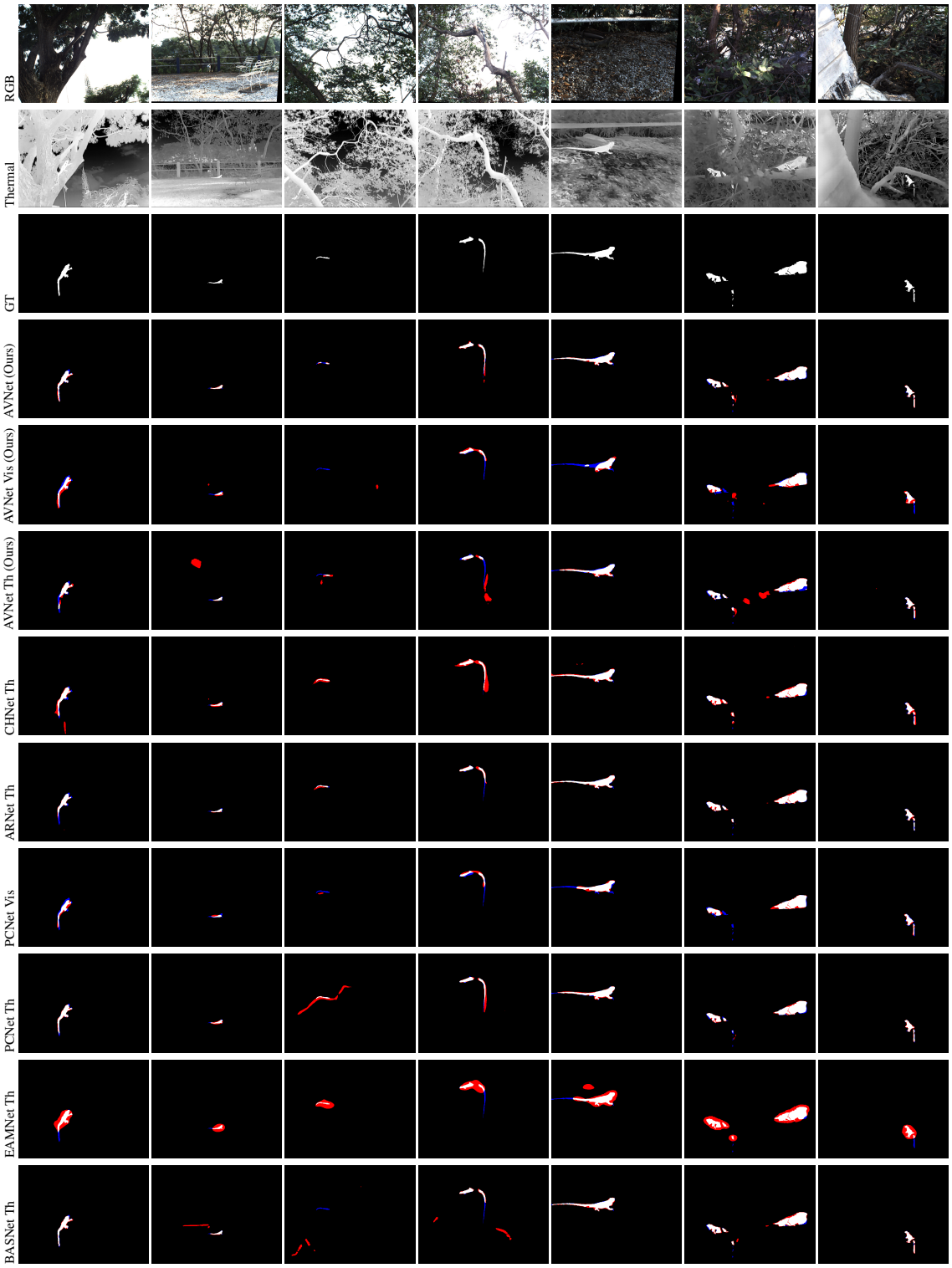

Figure 4: Results using SOTA COD techniques that have achieved first place in at least one of the metrics. Successful matches between GT and predicted masks (white areas); False positive regions (red areas, over-segmentation); and false negative regions (blue areas, miss-segmentation).

Table 2: Experimental results for SOTA COD techniques and AVNet on the BIOS dataset. The best three performing results are highlighted using color: First, Second, and Third respectively.

	Technique	$S_\alpha \uparrow$	$F_\beta^w \uparrow$	$M \downarrow$	$E_\phi^{adp} \uparrow$	$E_\phi^{mean} \uparrow$	$E_\phi^{max} \uparrow$	$F_\beta^{adp} \uparrow$	$F_\beta^{mean} \uparrow$	$F_\beta^{max} \uparrow$
Visible	BASNet (Qin et al., 2019)	0.6356	0.3248	0.0132	0.7028	0.7466	0.8156	0.3366	0.3439	0.3508
	SINet-V2 (Fan et al., 2022)	0.7093	0.4429	0.0075	0.6865	0.7930	0.9093	0.4052	0.4858	0.5141
	BGNet (Chen et al., 2022b)	0.5439	0.0866	0.1467	0.9186	0.7316	0.9469	0.4650	0.4499	0.5324
	C^2F-Net (Chen et al., 2022a)	0.5529	0.0892	0.0954	0.5829	0.6658	0.8760	0.3221	0.3772	0.4908
	OCENet (Liu et al., 2022)	0.7003	0.4371	0.0079	0.8096	0.8960	0.9159	0.4444	0.4838	0.4981
	EAMNet (Sun et al., 2023)	0.5432	0.0982	0.0923	0.6807	0.7451	0.8998	0.3589	0.3628	0.4818
	DGNet (Ji et al., 2023)	0.6832	0.3800	0.0325	0.6753	0.7831	0.9292	0.3922	0.4535	0.4777
	Hitnet (Hu et al., 2023)	0.7122	0.4612	0.0089	0.7189	0.8696	0.9246	0.4462	0.5025	0.5140
	PCNet (Yang et al., 2024)	0.7615	0.5761	0.0059	0.8890	0.9021	0.9869	0.6155	0.6222	0.6315
	ARNet (Wang et al., 2025a)	0.7285	0.5133	0.0056	0.9061	0.8302	0.9585	0.5907	0.5766	0.6080
	CHNet (Wang et al., 2025b)	0.7687	0.5809	0.0061	0.8549	0.9235	0.9602	0.5502	0.6199	0.6501
	CTF-Net (Zhang et al., 2025)	0.5180	0.0730	0.1500	0.5475	0.6365	0.9181	0.2600	0.3310	0.4881
	AVNet (Ours)	0.7648	0.5899	0.0063	0.9682	0.9538	0.9677	0.5978	0.6047	0.6179
Thermal	BASNet (Qin et al., 2019)	0.8154	0.6707	0.0036	0.8352	0.9260	0.9814	0.6023	0.6834	0.7276
	SINet-V2 (Fan et al., 2022)	0.7865	0.6207	0.0060	0.7756	0.8766	0.9523	0.5373	0.6232	0.6781
	BGNet (Chen et al., 2022b)	0.5521	0.0900	0.1050	0.7716	0.7809	0.9666	0.5353	0.5624	0.7641
	C^2F-Net (Chen et al., 2022a)	0.5647	0.1012	0.0849	0.6811	0.7363	0.9004	0.4384	0.5011	0.6218
	OCENet (Liu et al., 2022)	0.7909	0.6555	0.0056	0.7971	0.8746	0.9534	0.5754	0.6555	0.7517
	EAMNet (Sun et al., 2023)	0.5541	0.1196	0.0647	0.6725	0.7576	0.9832	0.4005	0.4544	0.6268
	DGNet (Ji et al., 2023)	0.7504	0.4439	0.0101	0.6635	0.8735	0.9568	0.4350	0.6275	0.6761
	HitNet (Hu et al., 2023)	0.7473	0.5402	0.0068	0.7020	0.9028	0.9420	0.4371	0.5921	0.6372
	PCNet (Yang et al., 2024)	0.8446	0.7331	0.0043	0.8605	0.8989	0.9139	0.6736	0.7302	0.7561
	ARNet (Wang et al., 2025a)	0.8761	0.8133	0.0027	0.9482	0.9688	0.9756	0.7639	0.8150	0.8331
	CHNet (Wang et al., 2025b)	0.8001	0.6969	0.0046	0.8284	0.8995	0.9474	0.6086	0.6874	0.7827
	CTF-Net (Zhang et al., 2025)	0.5919	0.1554	0.0830	0.5080	0.6748	0.8501	0.2518	0.4223	0.5690
	AVNet (Ours)	0.8301	0.6089	0.0059	0.8810	0.9072	0.9369	0.6016	0.6183	0.6597
	AVNet (Ours)	0.8951	0.8404	0.0028	0.9835	0.9785	0.9837	0.8190	0.8255	0.8585

6 CONCLUSIONS

This study presents AVNet, a cross-spectral attention–vision architecture designed for camouflaged object detection (COD) in an ecological conservation context. The model integrates PVTv2-based hierarchical encoding, CBAM attention mechanisms, an adaptive RGB–Thermal fusion block, and multi-level supervision to effectively detect camouflaged wildlife. AVNet demonstrates state-of-the-art performance on the BIOS dataset, particularly in the thermal modality, where camouflaged targets exhibit higher discriminability. The model shows strong robustness to challenges such as low target–background contrast, complex natural environments, and variations in object scale and pose. These capabilities contribute to reducing false positives and negatives, as well as improving boundary localization. These properties make AVNet a strong candidate for practical conservation monitoring in real-world field conditions. Future work will expand the BIOS dataset to additional species and habitats, explore lightweight variants for edge deployment, and investigate semi- and weakly-supervised learning strategies to reduce annotation costs while maintaining high detection accuracy.

ACKNOWLEDGEMENTS

This work was supported in part by the Air Force Office of Scientific Research Under Award FA9550-24-1-0206; in part by the ESPOL project “Advancing Camouflaged Object Detection with a cost-effective Cross-Spectral vision system (ACODCS)” (CIDIS-003-2024); in part by Grant PID2024-162815NB-I00 funded by MICIU/AEI/ 10.13039/501100011033 and by ERDF/EU, and Grant PID2021-128945NB-I00 funded by MICIU/AEI/ 10.13039/501100011033 and by ERDF/EU. The authors acknowledge the support of the Generalitat de Catalunya CERCA Program to CVC’s general activities, and the Departament de Recerca i Universitats from Generalitat de Catalunya to the SGR Research Group 2021 MACO (reference 2021 SGR 01499).

REFERENCES

Achanta, R., Hemami, S., Estrada, F., and Susstrunk, S. (2009). Frequency-tuned salient region detection. In *Conf. on Computer Vision and Pattern Recognition*, pages 1597–1604. IEEE.

Chen, G., Liu, S.-J., Sun, Y.-J., Ji, G.-P., Wu, Y.-F., and Zhou, T. (2022a). Camouflaged object detection via context-aware cross-level fusion. *Transactions on Circuits and Systems for Video Technology*, 32(10):6981–6993.

Chen, T., Xiao, J., Hu, X., Zhang, G., and Wang, S. (2022b). Boundary-guided network for camouflaged object detection. *Knowledge-based systems*, 248:108901.

Davis, J. W. and Sharma, V. (2007). Background-subtraction using contour-based fusion of thermal and visible imagery. *Computer vision and image understanding*, 106(2-3):162–182.

De Boer, P.-T., Kroese, D. P., Mannor, S., and Rubinstein, R. Y. (2005). A tutorial on the cross-entropy method. *Annals of operations research*, 134:19–67.

Fan, D.-P., Cheng, M.-M., Liu, Y., Li, T., and Borji, A. (2017). Structure-measure: A new way to evaluate foreground maps. In *Int. Conference on Computer Vision*, pages 4548–4557.

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

Fan, D.-P., Ji, G.-P., Cheng, M.-M., and Shao, L. (2022). Concealed object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

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

Gao, S.-H., Cheng, M.-M., Zhao, K., Zhang, X.-Y., Yang, M.-H., and Torr, P. (2019). Res2net: A new multi-scale backbone architecture. *Transactions on Pattern Analysis and Machine Intelligence*, 43(2):652–662.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Conf. on Computer Vision and Pattern Recognition*, pages 770–778.

Hu, X., Wang, S., Qin, X., Dai, H., Ren, W., Luo, D., Tai, Y., and Shao, L. (2023). High-resolution iterative feedback network for camouflaged object detection. In *Conf. on Artificial Intelligence*, volume 37, pages 881–889.

Ji, G.-P., Fan, D.-P., Chou, Y.-C., Dai, D., Liniger, A., and Van Gool, L. (2023). Deep gradient learning for efficient camouflaged object detection. *Machine Intelligence Research*, 20(1):92–108.

Jiang, X., Cai, W., Zhang, Z., Jiang, B., Yang, Z., and Wang, X. (2022). Magnet: A camouflaged object detection network simulating the observation effect of a magnifier. *Entropy*, 24(12):1804.

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

Lin, W., Wu, Z., Chen, J., Huang, J., and Jin, L. (2023). Scale-aware modulation meet transformer. In *Int. Conf. on Computer Vision*, pages 6015–6026.

Lindenberger, P., Sarlin, P.-E., and Pollefeys, M. (2023). Lightglue: Local feature matching at light speed. *arXiv preprint arXiv:2306.13643*.

Liu, J., Zhang, J., and Barnes, N. (2022). Modeling aleatoric uncertainty for camouflaged object detection. In *Winter Conference on Applications of Computer Vision*, pages 1445–1454.

Margolin, R., Zelnik-Manor, L., and Tal, A. (2014). How to evaluate foreground maps? In *Conf. on Computer Vision and Pattern Recognition*, pages 248–255.

Máttyus, G., Luo, W., and Urtasun, R. (2017). Deeproadmapper: Extracting road topology from aerial images. In *Int. Conf. on Computer Vision*, pages 3438–3446.

Pang, Y., Zhao, X., Xiang, T.-Z., Zhang, L., and Lu, H. (2022). Zoom in and out: A mixed-scale triplet network for camouflaged object detection. In *Conf. on Computer Vision and Pattern Recognition*, pages 2160–2170.

Patel, A. and Chaudhary, J. (2019). A review on infrared and visible image fusion techniques. *Intelligent Communication Technologies and Virtual Mobile Networks*, pages 127–144.

Perazzi, F., Krähenbühl, P., Pritch, Y., and Hornung, A. (2012). Saliency filters: Contrast based filtering for salient region detection. In *Conf. on Computer Vision and Pattern Recognition*, pages 733–740. IEEE.

Qin, X., Zhang, Z., Huang, C., Gao, C., Dehghan, M., and Jagersand, M. (2019). Basnet: Boundary-aware salient object detection. In *Conf. on Computer Vision and Pattern Recognition*.

Rivadeneira, R. E., Velesaca, H. O., and Sappa, A. (2024). Cross-spectral image registration: a comparative study and a new benchmark dataset. In *Int. Conf. on Innovations in Computational Intelligence and Computer Vision*, pages 1–12. Springer.

Sun, D., Jiang, S., and Qi, L. (2023). Edge-aware mirror network for camouflaged object detection. In *Int. Conf. on Multimedia and Expo*, pages 2465–2470. IEEE.

Sun, Y., Chen, G., Zhou, T., Zhang, Y., and Liu, N. (2021). Context-aware cross-level fusion network for camouflaged object detection. In *IJCAI*, pages 1025–1031.

Tan, M. and Le, Q. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. In *Int. Conf. on Machine Learning*, pages 6105–6114. PMLR.

Velesaca, H. O., Bastidas, G., Rouhani, M., and Sappa, A. D. (2024). Multimodal image registration techniques: a comprehensive survey. *Multimedia Tools and Applications*, 83(23):63919–63947.

Velesaca, H. O. and Sappa, A. D. (2025). Seeing the unseen: Ai-powered camouflaged pest detection. In *Multi Conference on Computer Science and Information Systems*, pages 137–144.

Wang, K., Li, X., Bai, Y., Li, S., Lu, M., and Jia, Z. (2025a). Assisted refinement network based on channel infor-

mation interaction for camouflaged object detection. In *Int. Conf. on Multimedia Retrieval*, pages 2058–2062.

Wang, K., Li, X., Li, S., Bai, Y., Li, B., Lu, M., and Jia, Z. (2025b). Efficient camouflaged object detection network based on channel reconstruction and hybrid attention. In *Int. Conf. on Multimedia Retrieval*, pages 2063–2067.

Wang, W., Xie, E., Li, X., Fan, D.-P., Song, K., Liang, D., Lu, T., Luo, P., and Shao, L. (2022). Pvt v2: Improved baselines with pyramid vision transformer. *Computational Visual Media*, 8(3):415–424.

Wei, J., Wang, S., and Huang, Q. (2020). F^3net: fusion, feedback and focus for salient object detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 12321–12328.

Woo, S., Park, J., Lee, J.-Y., and Kweon, I. S. (2018). Cbam: Convolutional block attention module. In *European Conference on Computer Vision*, pages 3–19.

Xu, H., Ma, J., Le, Z., Jiang, J., and Guo, X. (2020). Fusiondn: A unified densely connected network for image fusion. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 12484–12491.

Yan, J., Le, T.-N., Nguyen, K.-D., Tran, M.-T., Do, T.-T., and Nguyen, T. V. (2021a). Mirrornet: Bio-inspired camouflaged object segmentation. *IEEE Access*, 9:43290–43300.

Yan, J., Le, T.-N., Nguyen, K.-D., Tran, M.-T., Do, T.-T., and Nguyen, T. V. (2021b). Mirrornet: Bio-inspired camouflaged object segmentation. *IEEE access*, 9:43290–43300.

Yang, J., Wang, Q., Zheng, F., Chen, P., Leonardis, A., and Fan, D.-P. (2024). Plantcamo: Plant camouflage detection. *arXiv*.

Zhang, D., Wang, C., Wang, H., Fu, Q., and Li, Z. (2025). An effective cnn and transformer fusion network for camouflaged object detection. *Computer Vision and Image Understanding*, page 104431.

Zhong, Y., Li, B., Tang, L., Kuang, S., Wu, S., and Ding, S. (2022). Detecting camouflaged object in frequency domain. In *Conf. on Computer Vision and Pattern Recognition*, pages 4504–4513.