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

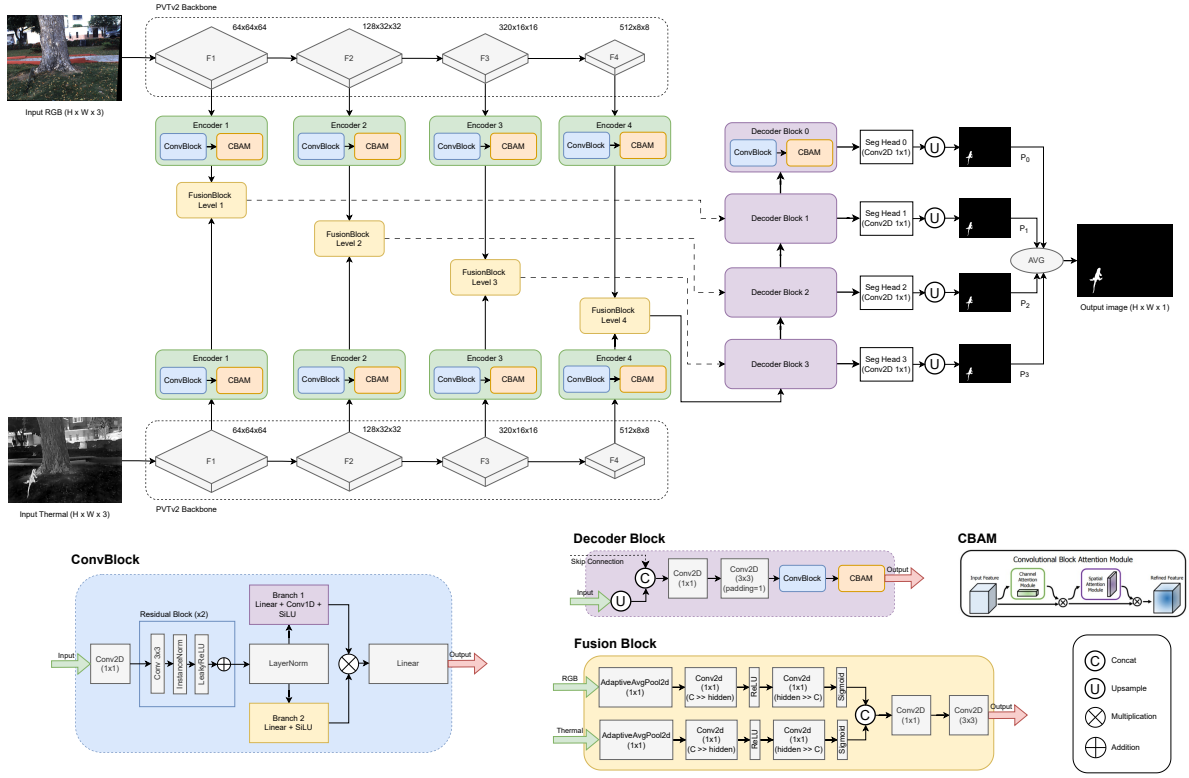

Figure 2: The overall architecture of the proposed AVNet.

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.

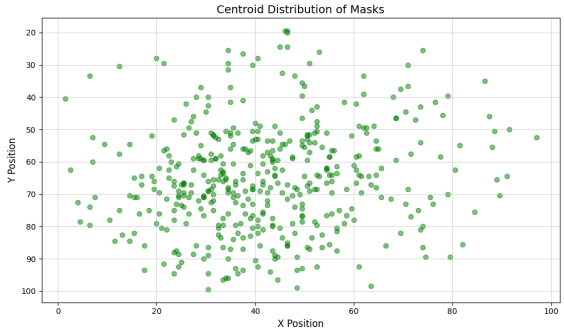

Figure 3: Centroid distribution of masks.

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.

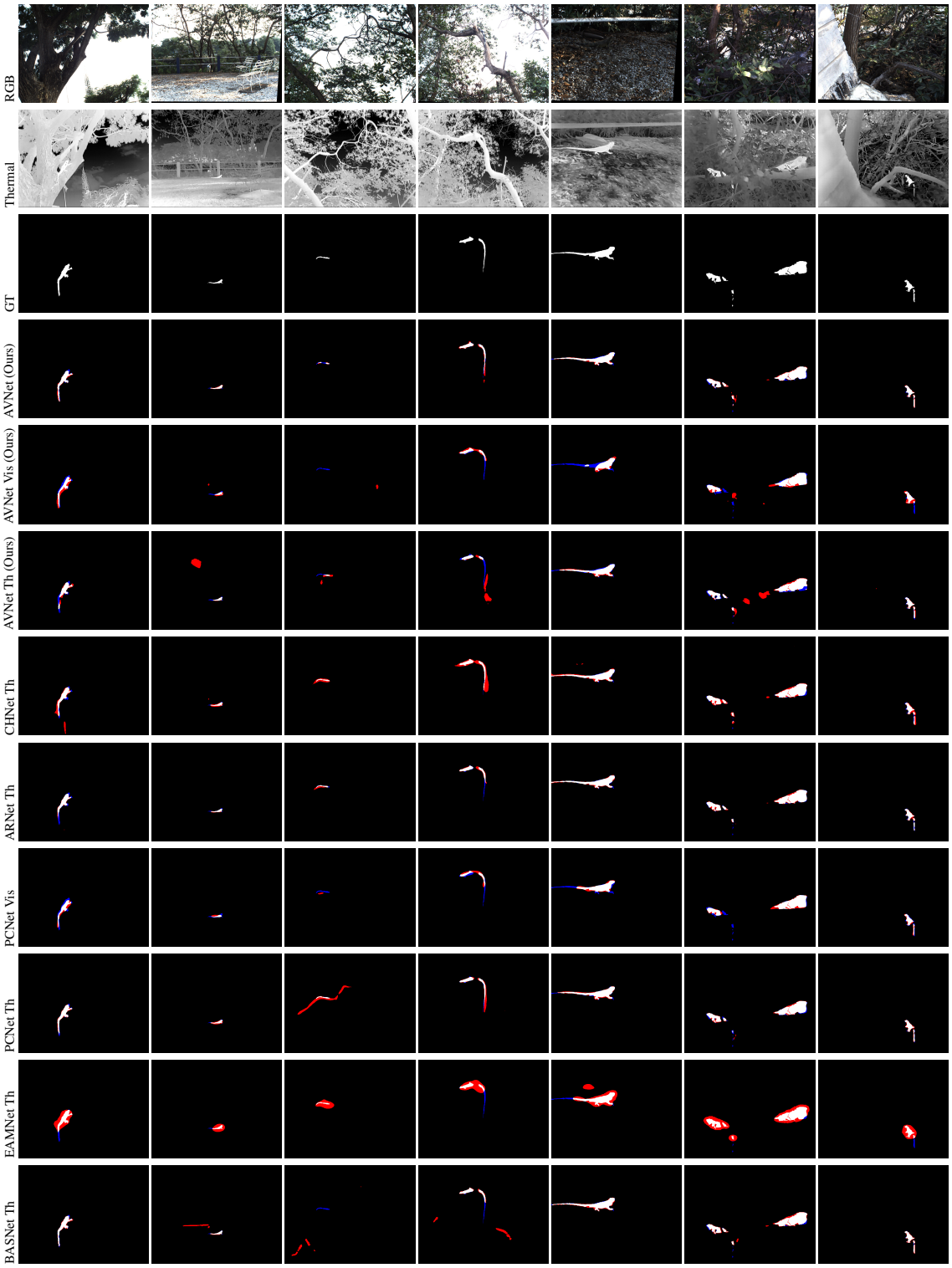

Figure 4: Results using SOTA COD techniques that have achieved first place in at least one of the metrics. Successful matches between GT and predicted masks (white areas); False positive regions (red areas, over-segmentation); and false negative regions (blue areas, miss-segmentation).

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.