The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XLIX-B2-2026
XXV ISPRS Congress 2026 “From Imagery to Understanding”, Commission II, 4–11 July 2026, Toronto, Canada

Polarization-Aware Segmentation for Camouflaged Threat Detection from UAVs

Youssef Korny, Sunghwan Yoo, Gunho Sohn

Department of Earth and Space Science and Engineering, York University, Canada- (ykorny, jacobyoo, gsohn)@yorku.ca

Keywords: Semantic Segmentation, deep learning, camouflaged object detection, polarized imaging

Abstract

Surface-laid unexploded ordnance (UXO) and landmines constitute a critical humanitarian crisis. While unmanned aerial vehicles (UAVs) provide a scalable remote sensing solution, detecting modern, non-metallic explosive devices in cluttered environments remains a profound Camouflaged Object Detection (COD) challenge. Traditional optical sensors frequently suffer from foreground-background confusion when a target's texture mimics its surroundings. To overcome these physical bottlenecks, we introduce XPol-Net, a novel multimodal architecture synergizing the semantic reasoning of Vision Transformers with the deterministic physics of polarimetric imaging. Built on a hierarchical PVTv2 backbone, XPol-Net utilizes a progressive Dual Cross-Attention Strategy for effective modality fusion. In early stages, Channel Cross-Attention (CCA) filters material-specific Degree of Linear Polarization (DoLP) cues to suppress background clutter. In deeper stages, Spatial Cross-Attention (SCA) dynamically aligns high-level RGB semantics with strict structural boundaries. To enhance robustness and prevent modality collapse, we deploy a multi-task auxiliary learning framework that reconstructs the continuous Angle of Linear Polarization (AoLP) map. On the PCOD benchmark, XPol-Net achieves state-of-the-art results in global structural alignment (E_ϕ of 0.980 and 0.984 at 352×352 and 704×704, respectively). While minor trade-offs are observed in localized metrics such as S_α or F_β, XPol-Net remains highly competitive, consistently delivering superior results in E_ϕ and MAE. By prioritizing structural recall over localized strictness, XPol-Net ensures the complete discovery of concealed targets, establishing a reliable, physics-aware foundation for humanitarian demining operations.

1. Introduction

The proliferation of unexploded ordnance (UXO) and landmines across post-conflict regions constitutes a profound and enduring humanitarian crisis. Recent estimates suggest that over 100 million explosive remnants remain scattered globally, rendering vast tracts of land uninhabitable and severely stifling economic recovery (UN News, 2025, UC Merced Newsroom, 2025). Conventional detection modalities, such as handheld metal detectors and ground-penetrating radar, are frequently confounded by high-clutter environments and modern explosive devices that utilize non-metallic plastic casings (Baur, 2025, Kumar et al., 2025). To accelerate humanitarian demining, remote sensing via unmanned aerial vehicles (UAVs) equipped with optical payloads has emerged as a highly scalable and safer alternative (Javed et al., 2024, Baur, 2025).

However, identifying surface-laid or partially buried UXO in natural environments is fundamentally a Camouflaged Object Detection (COD) problem (see figure 1). Explosive devices are either manufactured with specific chromatic patterns designed to mimic natural backgrounds, or they become naturally occluded by vegetation and soil over time, resulting in visually perfect camouflage (Hwang and Ma, 2024, Cho et al., 2023). The deep learning era of COD has seen rapid algorithmic advancements aimed at separating foreground objects from visually homogenous backgrounds (Fan et al., 2020). Foundational RGB-based networks, however, face insurmountable physical bottlenecks; when an object's color and texture perfectly match its surroundings, convolutional filters extract statistically identical feature maps for both target and background, neutralizing the network's discriminative capacity. Attempts to circumvent this using RGB-Depth (RGB-D) fusion strategies suffer from severe practical vulnerabilities. RGB-D models lack robustness to noisy sensor data, particularly in outdoor scenarios (Abbas et al., 2014), and exhibit catastrophic failure when the background and camouflaged object are located at identical spatial ranges, such as a landmine resting flush on the soil surface, where depth variance approaches zero (Liu et al., 2024).

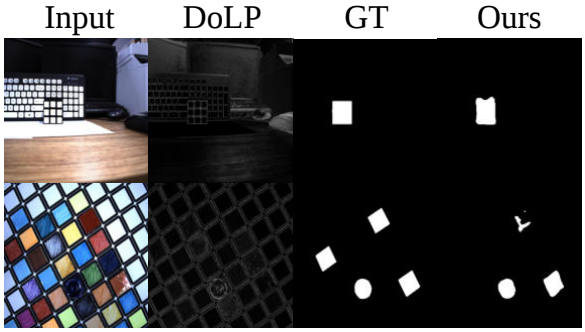

Figure 1. Example results of our proposed XPol-Net for Camouflaged Object Detection. By leveraging physical surface cues from the DoLP map, our method successfully localizes heavily concealed targets that are nearly indistinguishable in the optical RGB image.

To circumvent the inherent limitations of standard optical and depth sensors, polarimetric imaging has emerged as a transformative, physics-based sensing modality. Polarization is an innate characteristic of electromagnetic waves that carries intrinsic physical information completely invisible to the naked eye. When unpolarized ambient light interacts with a surface, the reflected light becomes partially polarized depending on the object's specific material properties, surface roughness, and micro-geometry. By capturing multi-angle linear polarization states, we can compute the Stokes parameters to derive two highly discriminative metrics: the Degree of Linear Polarization (DoLP) and the Angle of Linear Polarization (AoLP).

The **DoLP** measures the proportion of light that is linearly po-

larized. Because man-made objects, such as the smooth plastic casings of landmines, exhibit distinct specular reflection and dielectric properties compared to rough, depolarizing natural clutter, the DoLP effectively maximizes camouflage to background contrast regardless of visual color. Conversely, the **AoLP** describes the absolute geometric orientation of the polarization plane. It provides high-resolution structural changes, making it exceptionally sensitive to anomalies in surface structure, such as the disturbed earth surrounding a recently laid mine.

Recognizing the value of these physical insights, pioneering models such as IPNet (Wang et al., 2024) and HIPFNet (Wang et al., 2025) have recently integrated polarization into COD architectures. However, these contemporary multimodal networks exhibit critical theoretical flaws stemming from their fusion paradigms. Models relying on sequential fusion rapidly dilute the physically deterministic, narrow-band properties of polarization within semantically noisy RGB color statistics during early convolutional processing. Conversely, late-fusion parallel streams fail to allow the modalities to interact during feature extraction, preventing the highly accurate structural contours of the polarization data from dynamically correcting the extraction of low-level visual features.

To bridge this semantic gap, we propose XPol-Net, a novel architecture that abandons flawed parallel fusion streams in favor of direct, attention-driven feature injection. Built upon a hierarchical Pyramid Vision Transformer v2 (PVTv2) (Wang et al., 2022) backbone, XPol-Net employs a highly specialized Dual Cross-Attention Strategy to optimally fuse RGB and polarimetric data. In early, high-resolution layers, Channel Cross Attention (CCA) explicitly maps channel correlations to mathematically filter material-specific features derived from the DoLP. In deeper, low-resolution layers, Spatial Cross Attention (SCA) utilizes AoLP spatial tokens to enforce strict geometric boundaries and correct foreground-background confusion. Furthermore, to mitigate overfitting and guarantee physical compliance, XPol-Net incorporates multi-task auxiliary learning. By forcing the network to simultaneously predict edge boundaries, coarse masks, and the physical AoLP map.

In summary, the primary contributions of this paper are:

- We introduce XPol-Net, a novel, highly robust deep learning framework for Camouflaged Object Detection specifically engineered for humanitarian demining and UAV geospatial intelligence.
- We propose a unique Dual Cross-Attention Strategy within a PVTv2 backbone that optimally injects polarimetric features into the visual stream. This strategy applies Channel Cross Attention (CCA) to filter material properties at early scales and Spatial Cross Attention (SCA) to align geometric structures at deeper scales, overcoming the limitations of sequential and late-fusion architectures.
- We implement a multi-task auxiliary learning framework, including a novel Angle of Linear Polarization (AoLP) prediction head, which functions as a profound network regularizer to enforce strict physical and structural compliance in the final segmentation mask.

2. Related Works

2.1 RGB-based Camouflaged Object Detection

The evolution of purely optical Camouflaged Object Detection (COD) has been defined by a transition toward global-reasoning Vision Transformers, boundary-aware optimization, and uncertainty-guided learning paradigms. Early deep learning approaches frequently suffered from localized perception, failing to segment holistic structures due to continuous background texture bleeding. To rectify this, recent architectures explicitly force representation learning to highlight object boundaries; for instance, BGNet utilizes extra object-related edge semantics to heavily penalize boundary misclassifications (Sun et al., 2022). Simultaneously, recognizing the inherent ambiguity of camouflaged scenes, models such as the Uncertainty-Aware Transformer (UAT) explicitly model aleatoric uncertainty, forcing the network to focus computational resources on highly ambiguous texture regions (Wu et al., 2025). Furthermore, high-resolution iterative feedback mechanisms, like those in HitNet, have been introduced to refine low-resolution contextual representations and prevent edge degradation (Hu et al., 2023). Despite these algorithmic interventions, purely optical COD faces a hard performance plateau. When a target perfectly mimics the spectral reflectance of its environment under flat illumination, deterministic RGB networks experience catastrophic foreground-background confusion (Bai et al., 2026). The introduction of recent multispectral (MCOD) and hyperspectral (HyperCOD) benchmarks mathematically proves that texture-matching cannot be fully resolved within the visible spectrum alone, establishing the absolute necessity for supplementary physical priors (Li et al., 2025a, Bai et al., 2026).

2.2 RGB-X Object Detection and Segmentation

To bypass the physical limitations of the visible spectrum, the field expanded into RGB-X multimodal learning, pairing standard RGB data with auxiliary sensing streams such as Depth (D), Thermal (T), or Event (E) cameras. However, the acquisition of perfectly aligned multimodal data imposes severe physical and hardware limitations. Active depth sensors (e.g., Time-of-Flight) suffer from catastrophic sunlight interference outdoors, while passive stereo vision fails in low-texture camouflage areas where disparity matching becomes ill-posed (Brenner et al., 2023). Both require highly sophisticated hardware-software triggers, such as the SmartDepthSync system, to achieve sub-millisecond temporal alignment (Faizullin et al., 2022). Thermal sensors (RGB-T) operate effectively in complete darkness but mandate the use of 50/50 beam-splitter camera rigs to eliminate parallax errors, which drastically reduces incident light and degrades RGB quality (Brenner et al., 2023). Furthermore, thermal detection fails entirely during ”thermal crossover,” when a target’s temperature equilibrates with its background (Brenner et al., 2023). Neuromorphic Event cameras (RGB-E) offer extreme dynamic range but are absolutely blind to static camouflaged objects, generating no events without motion (Xie et al., 2024). To fuse these heterogeneous modalities, recent architectures abandoned early and late fusion in favor of cross-attention; RDCRNet aligns statistical distributions via a Cross-Modal Feature Remapping Module (Li et al., 2025b). Despite these fusion advancements, the inherent environmental vulnerabilities of D, T, and E sensors underscore the necessity for a universally stable, geometry-aware modality.

2.3 Polarized Camouflaged Object Detection

Polarized Camouflaged Object Detection (PCOD) has emerged as the most robust frontier for resolving perfect texture mimicry. Because camouflage evolved primarily to deceive unpolarized visual systems, the specific polarization state of reflected light, dictated by microstructural surface roughness and underlying dielectric properties, remains undisguised. The Degree of Linear Polarization (DoLP) provides unparalleled contrast for distinguishing material differences regardless of color, while the Angle of Linear Polarization (AoLP) captures highly sensitive surface structure orientations (Wang et al., 2024). The formalization of this task was accelerated by the PCOD_1200 dataset and the IPNet baseline, which introduced a dual-flow network with a Polarization Enrich Module to dynamically fuse intensity and DoLP cues (Wang et al., 2024). Subsequent state-of-the-art models advanced this space: HIPFNet employs a high-resolution adaptive fusion network utilizing orthogonal linear polarization cues (Wang et al., 2025), while PRNet prioritized efficiency via a Cascaded Attention Perceptron (Hu et al., 2024). However, a critical architectural gap persists: existing models predominantly rely on static parallel-stream or sequential mid-level fusion (Wang et al., 2023). This structural reliance fails to account for the highly localized and dynamic nature of polarimetric reliability, leading to semantic discrepancies when models force fusion in noisy diffuse regions (Zhang et al., 2025). XPol-Net addresses this void by introducing a direct, attention-driven feature injection framework. By utilizing a dual cross-attention mechanism to intelligently query polarimetric representations based strictly on localized visual confidence, XPol-Net transcends static concatenation, establishing a new paradigm for adaptive, material-aware segmentation (Wang et al., 2024).

3. Methodology

3.1 Problem Formulation and Network Objective

Given an input optical image $I_{RGB} \in \mathbb{R}^{H\times W\times 3}$ and its corresponding Degree of Linear Polarization image $I_{DoLP} \in \mathbb{R}^{H\times W\times 1}$, the primary objective of XPol-Net is to predict a precise binary segmentation mask $\hat{M} \in \{0,1\}^{H\times W}$ that isolates the camouflaged object.

A prevalent vulnerability in multimodal fusion networks is modality collapse, where the network defaults to relying entirely on the dominant, feature-rich RGB stream and effectively ignores the auxiliary physical modalities. To prevent this degradation and force the active utilization of polarimetric features, we formulate XPol-Net's objective as a Multi-Task Learning (MTL) problem. We define the mapping function $\mathcal{F}$, parameterized by network weights Θ, to jointly predict three distinct outputs:

$$\mathcal{F}(I_{RGB}, I_{DoLP}; \Theta) \rightarrow (\hat{M}, \hat{E}, \hat{A})$$

where $\hat{E}$ represents the structural edge boundaries, and $\hat{A}$ represents the predicted Angle of Linear Polarization (AoLP) map. The prediction of $\hat{A}$ serves as a critical physical regularizer; by penalizing the network for failing to reconstruct the underlying geometry dictated by the polarization phase, we compel the features to deeply integrate the DoLP cues.

3.2 Preliminaries: Pyramid Vision Transformer v2 (PVTv2)

To extract robust semantic features while preserving spatial geometry, XPol-Net employs two parallel Vision Transformer encoders based on the Pyramid Vision Transformer v2 (PVTv2) architecture.

Unlike standard isotropic Vision Transformers (e.g., ViT) that maintain a constant token sequence length, PVTv2 employs a progressive shrinking pyramid structure, functionally mirroring the hierarchical feature extraction of traditional convolutional networks like ResNet. For a given input, PVTv2 extracts features across four sequential stages $i \in \{1,2,3,4\}$. Each stage consists of an overlapping patch embedding module followed by multiple Transformer encoder blocks.

To mitigate the quadratic computational complexity typically associated with self-attention at high resolutions, PVTv2 utilizes a Spatial Reduction Attention (SRA) mechanism. SRA mathematically reduces the spatial scale of the keys and values prior to the attention operation. Crucially, at the conclusion of each stage i, the 1D token sequence is systematically reshaped back into a 2D spatial feature map:

$$F_i \in \mathbb{R}^{C_i\times H_i\times W_i}$$

where C_i represents the expanding channel dimensions, and the spatial resolutions ($H_i \times W_i$) correspond to $\frac{1}{4}$, $\frac{1}{8}$, $\frac{1}{16}$, and $\frac{1}{32}$ of the original input dimensions, respectively.

This inherent ability to project Transformer tokens back into hierarchical 2D image topologies is the foundational mechanism of XPol-Net. While the DoLP encoder processes I_{DoLP} standardly to extract physical priors, the RGB encoder is heavily modified. We leverage these 2D reshaped outputs F_i after each transformer block to perform our targeted, scale-specific cross-attention injections.

3.3 Preliminaries: Polarimetric Imaging Principles

Unlike standard optical cameras that only capture the intensity and wavelength of light, polarization-aware sensors capture the oscillation orientation of light waves. In our dataset, the polarimetric data is acquired using a Division of Focal Plane (DoFP) microgrid polarization camera. This sensor simultaneously captures four spatially aligned intensity images at distinct polarization angles: I_{0°, I_{45°, I_{90°, and I_{135°.

As comprehensively illustrated in Figure 3, these raw captures are mathematically transformed into the final polarimetric features utilized by our network. We first compute the three primary Stokes parameters (S_0, S_1, S_2), which describe the linear polarization state of the reflected light:

$$S_0 = I_{0^\circ} + I_{90^\circ} \quad (1)$$

$$S_1 = I_{0^\circ} - I_{90^\circ} \quad (2)$$

$$S_2 = I_{45^\circ} - I_{135^\circ} \quad (3)$$

Here, S_0 represents the total unpolarized light intensity (functionally equivalent to a standard grayscale image). S_1 captures the difference between horizontal and vertical linear polarization, while S_2 captures the difference between the two diagonal linear polarizations.

Using these Stokes parameters, we derive the two critical physical properties injected into XPol-Net: the Degree of Linear Polarization (DoLP) and the Angle of Linear Polarization (AoLP).

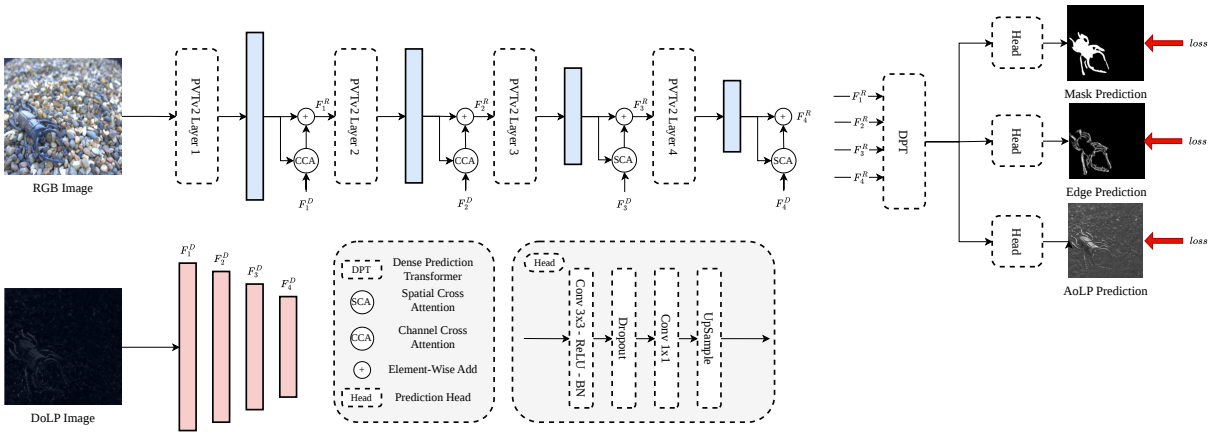

Figure 2. XPol-Net architecture with cross-attention fusion. The network processes complementary RGB and DoLP inputs using channel and spatial cross-attention (CCA and SCA) modules that dynamically inject polarization features into the visual backbone, enabling robust multi-scale refinement for segmentation and auxiliary predictions.

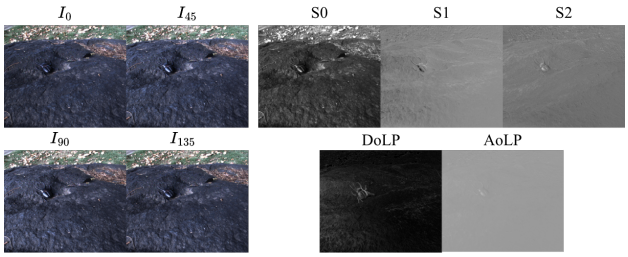

Figure 3. The polarimetric feature extraction pipeline. A DoFP sensor captures four raw directional intensity images ($I_{0^\circ}, I_{45^\circ}, I_{90^\circ}, I_{135^\circ}$). These are combined to calculate the Stokes parameters (S_0, S_1, S_2), which are then used to derive the physical Degree of Linear Polarization (DoLP) and Angle of Linear Polarization (AoLP) maps.

The DoLP measures the proportion of the light wave that is linearly polarized:

$$DoLP = \frac{\sqrt{S_1^2 + S_2^2}}{S_0} \tag{4}$$

The DoLP is highly sensitive to surface material properties. Smooth, manufactured surfaces (such as plastic landmine casings) reflect highly polarized light (high DoLP), whereas rough, natural environments (such as soil and foliage) scatter and depolarize the light (low DoLP). This provides the primary high-contrast cue for our Channel Cross-Attention module.

The AoLP represents the dominant orientation angle of the polarized light:

$$AoLP = \frac{1}{2} \arctan\left(\frac{S_2}{S_1}\right) \tag{5}$$

The AoLP is intrinsically linked to the surface structure of the object. Consequently, by forcing our network to reconstruct the continuous AoLP map, we mathematically compel the model to encode the true shape of the concealed threat, ensuring strict boundaries regardless of the optical camouflage.

3.4 Overall Architecture of XPol-Net

The end-to-end pipeline of XPol-Net, illustrated in Figure 2, is designed as a dual-stream architecture that processes aligned RGB and Degree of Linear Polarization (DoLP) images through parallel PVTv2 encoders. To overcome the common pitfalls of multi-modal fusion—specifically the semantic dilution found in early fusion and the spatial misalignment inherent in late fusion—we implement a progressive, layer-wise modulation strategy. Following each of the four hierarchical Transformer stages, the intermediate RGB features are dynamically recalibrated using physical priors from the DoLP stream via our novel Dual Cross-Attention Strategy. These modulated, multi-scale representations are subsequently aggregated by a Dense Prediction Transformer (DPT) (Ranftl et al., 2021) decoding head. This unified latent representation then feeds into three parallel prediction heads to execute our multi-task objective: primary mask segmentation, structural edge extraction, and auxiliary AoLP reconstruction.

3.4.1 Dual Cross-Attention Strategy The central innovation of XPol-Net is its scale-aware modality fusion, which adapts to the shifting feature characteristics across the network. While the DoLP encoder processes I_{DoLP} independently to maintain pristine physical priors, the RGB encoder undergoes active modulation. Within each PVTv2 block, after the Spatial Reduction Attention (SRA) mechanism, the 1D token sequences are reshaped into 2D spatial feature maps. Let F_i^R and F_i^D represent these reshaped features from the RGB and DoLP encoders at stage i. Our Dual Cross-Attention Strategy treats early and late stages differently to account for the evolution of spatial resolution and semantic density.

Channel Cross-Attention (CCA) for Stages 1 and 2 In the initial layers ($i \in \{1, 2\}$), feature maps exhibit high spatial resolutions ($\frac{H}{4} \times \frac{W}{4}$ and $\frac{H}{8} \times \frac{W}{8}$) but relatively low channel counts. At this shallow depth, features are dominated by low-level visual information, which often includes significant spatial noise and background clutter. Attempting spatial cross-attention here is not only computationally expensive but risks propagating diffuse noise into the backbone.

To address this, we employ Channel Cross-Attention (CCA) to filter material-specific properties. By computing attention across the channel dimension, the network learns to highlight RGB feature maps that correlate with high-contrast specular cues in the DoLP stream while suppressing channels associated with camouflage noise. We project Queries (Q_c) from the RGB stream, and Keys (K_c) and Values (V_c) from the DoLP stream, reshaping them to $\mathbb{R}^{C_i \times (H_i W_i)}$ to compute a $C_i \times C_i$

correlation matrix:

$$Q_c = \mathcal{W}_q F_i^R, \quad K_c = \mathcal{W}_k F_i^D, \quad V_c = \mathcal{W}_v F_i^D \tag{6}$$

$$\text{CCA}(F_i^R, F_i^D) = \alpha_c \cdot \left(\text{Softmax}\left(\frac{Q_c K_c^T}{\sqrt{H_i W_i}} \right) V_c \right) + F_i^R \tag{7}$$

Crucially, to preserve the pre-trained PVTv2 representation and avoid catastrophic interference from initialization noise, the learnable fusion weight (α_c) is strictly initialized to zero (Zhang et al., 2023, Hu et al., 2022). This ensures that at the start of training, the CCA module acts as an identity mapping, allowing the network to safely and progressively incorporate polarimetric priors.

Spatial Cross-Attention (SCA) for Stages 3 and 4 In the deeper stages ($i \in \{3, 4\}$), spatial dimensions are significantly reduced ($\frac{H}{16} \times \frac{W}{16}$ and $\frac{H}{32} \times \frac{W}{32}$), while the channel dimensions expand to capture complex semantics. While these layers excel at abstract object recognition, they often suffer from spatial degradation, leading to blurred geometric boundaries. Since the reduced spatial resolution makes spatial attention computationally efficient, we deploy Spatial Cross-Attention (SCA) to align semantic RGB tokens with the strict physical contours retained by the DoLP encoder.

Here, tensors are reshaped to $\mathbb{R}^{(H_i W_i) \times C_i}$ to compute an $(H_i W_i) \times (H_i W_i)$ spatial correlation matrix:

$$Q_s = \mathcal{W}_q F_i^R, \quad K_s = \mathcal{W}_k F_i^D, \quad V_s = \mathcal{W}_v F_i^D \tag{8}$$

$$\text{SCA}(F_i^R, F_i^D) = \alpha_s \cdot \left(\text{Softmax}\left(\frac{Q_s K_s^T}{\sqrt{C_i}} \right) V_s \right) + F_i^R \tag{9}$$

Following the same zero-initialization paradigm, the spatial fusion parameter (α_s) is set to zero initially. This ”progressive learning gate” allows the SCA module to eventually dictate exactly *where* the RGB stream should focus its semantic attention, forcing the feature maps to adhere to the physical geometric limits of the target object.

3.4.2 Multi-Task Dense Prediction Decoding Head After the hierarchical extraction and cross-modality injection, the modulated RGB features $\{F_1^R, F_2^R, F_3^R, F_4^R\}$ contain a fusion of visual semantics and polarimetric priors. XPol-Net uses a DPT (Ranftl et al., 2021) decoding head to progressively upsample and aggregate these maps into a high-resolution global representation. From this unified feature space, the model branches into three parallel convolutional heads to optimize our multi-task learning objectives: (1) a **Mask Prediction Head** that produces the primary binary segmentation mask $\hat{M}$ for the target; (2) an **Edge Prediction Head** that identifies high-frequency spatial gradients to predict the structural boundary map $\hat{E}$; and (3) an **AoLP Prediction Head** that serves as a physical regularizer by reconstructing the auxiliary Angle of Linear Polarization map $\hat{A}$. All prediction heads utilize the design architecture detailed in Figure 2.

3.4.3 Loss Function Formulation To optimize XPol-Net across these three distinct tasks, we formulate a joint loss function. For the segmentation of the camouflaged mask and the extraction of edge boundaries, we adopt the robust loss formulations utilized in recent polarization-aware networks (Wang et al., 2024).

Mask Supervision Loss ($\mathcal{L}_{mask}$) To supervise the camouflaged mask prediction, we employ a combination of losses including weighted Binary Cross-Entropy (BCE) loss and Intersection over Union (IoU) loss. The weighted BCE loss mitigates the negative impact of severe foreground-background data imbalance, while the weighted IoU loss emphasizes the structural alignment of global pixels. The mask loss is defined as:

$$\mathcal{L}_{mask} = \mathcal{L}_{BCE}^{w}(\hat{M}, M) + \mathcal{L}_{IoU}^{w}(\hat{M}, M) \tag{10}$$

where M denotes the ground-truth camouflage mask.

Edge Supervision Loss ($\mathcal{L}_{edge}$) For the supervision of the camouflaged object boundaries, we solely apply the weighted BCE loss. Because edge and boundary information fundamentally lacks global spatial context, the application of an IoU loss at this stage may impede the network's learning process. The edge loss is thus defined as:

$$\mathcal{L}_{edge} = \mathcal{L}_{BCE}^{w}(\hat{E}, E) \tag{11}$$

where E represents the ground-truth edge map derived from the mask annotations.

Auxiliary AoLP Reconstruction Loss ($\mathcal{L}_{AoLP}$) To force the network to actively utilize the polarimetric cues and encode surface structure, the AoLP prediction head must accurately reconstruct the physical polarization phase. Since AoLP prediction is a continuous regression task rather than a binary classification task, we employ the Mean Squared Error (MSE) loss:

$$\mathcal{L}_{AoLP} = \text{MSE}(\hat{A}, A) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} (\hat{A}_{i,j} - A_{i,j})^2 \tag{12}$$

where A represents the ground-truth Angle of Linear Polarization image.

Total Objective Function The total multi-task loss $\mathcal{L}_{total}$ used to optimize XPol-Net is the weighted summation of the three individual loss components:

$$\mathcal{L}_{total} = \mathcal{L}_{mask} + \lambda_1 \mathcal{L}_{edge} + \lambda_2 \mathcal{L}_{AoLP} \tag{13}$$

where λ_1 and λ_2 are empirically determined hyperparameters designed to balance the gradient contributions of the auxiliary tasks against the primary segmentation objective.

4. Experiments

4.1 Datasets and Evaluation Metrics

To rigorously evaluate the proposed XPol-Net architecture, we conduct experiments on the Polarized Camouflaged Object Detection (PCOD) benchmark (Wang et al., 2024). The PCOD dataset is currently the most comprehensive multimodal dataset for this domain, providing perfectly aligned RGB intensity images, Degree of Linear Polarization (DoLP) maps, Angle of Linear Polarization (AoLP) maps, and pixel-level ground-truth masks. The dataset consists of 1200 image pairs, which was divided by the author into a training set of 970 images and a testing set of 230 images. Vertical and horizontal flipping was applied for the training set by the author to increase the number of training samples.

We evaluate our model's performance using four standard evaluation metrics widely adopted in COD literature: Structure-measure (S_α) (Fan et al., 2017), Enhanced-alignment measure (E_ϕ) (Fan et al., 2018), F-measure (F_β), and Mean Absolute Error (MAE). For mission-critical geospatial applications such as unexploded ordnance (UXO) detection, minimizing false negatives is paramount. Therefore, we place a specific analytical emphasis on target discovery and overall structural recall, prioritizing the robust identification of the camouflaged threat over pixel-perfect boundary precision.

4.2 Implementation Details

XPol-Net is implemented using the PyTorch deep learning framework. The dual encoders are initialized with PVTv2 weights (Wang et al., 2022) pre-trained on ImageNet, while the DPT (Ranftl et al., 2021) decoding head and cross-attention modules are initialized randomly. Model training and inference are executed on a single NVIDIA L40s GPU.

During training, the input RGB, DoLP, and AoLP images are resized to two distinct spatial resolutions: 352×352 and 704×704, to evaluate the network's scalability and high-resolution feature extraction capabilities. We apply standard data augmentation techniques, including random horizontal flipping and multi-scale training. The network is optimized using the AdamW optimizer with an initial learning rate of 5×10^{-4}, a weight decay of 1×10^{-4}, and a batch size of 6. The training process spans 300 epochs, utilizing a step learning rate scheduler that decays the learning rate by a factor of 10 every 100 epochs. The empirical loss weights are set to $\lambda_1 = 0.5$ and $\lambda_2 = 0.05$.s

4.3 Comparison with State-of-the-Art Methods

4.3.1 Quantitative Analysis We quantitatively compare XPol-Net against state-of-the-art polarization-aware networks, including IPNet (Wang et al., 2024), HIPFNet (Wang et al., 2025), PRNet (Hu et al., 2024), and the recently introduced PMFNet (Zhang et al., 2025). Performance comparisons across multiple resolutions are detailed in Table 1.

At the 352×352 resolution, **XPol-Net** achieves state-of-the-art results in several key metrics. Most notably, it secures an impressive global structural alignment (E_ϕ) of 0.980 and the lowest MAE of 0.0078, outperforming both HIPFNet and PMFNet. While our localized spatial structure score (S_α) of 0.910 is narrowly surpassed by the top result (0.914), the performance remains highly competitive. This result continues to validate the efficacy of our Dual Cross-Attention Strategy in suppressing background noise and successfully coupling specular priors with the target's overarching shape.

At the 704×704 resolution, XPol-Net shares the top S_α score of 0.950 and the lowest MAE of 0.005 with PRNet, while establishing a new peak E_ϕ of 0.984. Although PRNet fractionally exceeds our F_β measure (0.940 vs. 0.930), our dominant E_ϕ highlights a deliberate architectural trade-off. By prioritizing AoLP for boundary reconstruction, XPol-Net trades fine-grained edge strictness for crucial gains in object discovery and global structural alignment. This ensures the complete, unfragmented localization of camouflaged threats, which is the primary operational priority for humanitarian demining.

4.3.2 Qualitative Analysis Visual comparisons further illustrate the distinct advantages of our attention-driven method over traditional fusion streams. As shown in Figure 4, we qualitatively compare XPol-Net against IPNet (Wang et al., 2024) and PRNet (Hu et al., 2024). Models such as HIPFNet (Wang et al., 2025) and PMFNet (Zhang et al., 2025) are excluded as their source codes and prediction masks remain publicly unavailable.

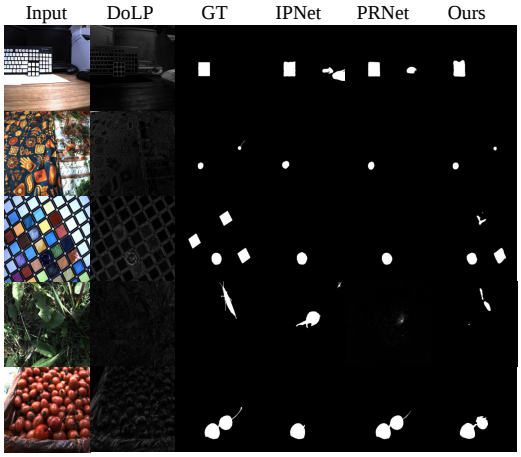

Figure 4. Qualitative comparison between IPNet, PRNet, and our XPol-Net. Our method successfully detects camouflaged objects missed by other models, although some fine structural details are not fully captured.

In scenarios where the target possesses visually perfect camouflage against highly cluttered backgrounds, baselines like IPNet and PRNet frequently experience foreground-background confusion. They often miss targets entirely (catastrophic false negatives) or fracture masks into disjointed segments. Conversely, XPol-Net successfully segments objects completely missed by other networks. By leveraging the AoLP auxiliary prediction to map continuous surface features, our architecture produces highly cohesive segmentation masks.

While our method successfully detects heavily obscured targets, some fine structural edges may not be perfectly captured. However, XPol-Net consistently achieves higher recall by locating the entire camouflaged threat. This focused gain in detection robustness is paramount for reliable target localization in high-risk geospatial imagery, proving its operational reliability for real-world demining.

5. Ablation Studies

To validate our architectural contributions, we conduct ablation studies on the PCOD dataset at a standard 352×352 resolution.

5.1 Effectiveness of the Dual Cross-Attention Strategy

We progressively integrate the Channel (CCA) and Spatial (SCA) Cross-Attention modules into a purely optical baseline (a single PVTv2 RGB encoder with a DPT head, omitting DoLP priors). The quantitative progression is detailed in Table 2.

As expected, the RGB baseline struggles to resolve severe camouflage, yielding an E_ϕ of 0.968 and an MAE of 0.0100. Integrating only the CCA module in the early, high-resolution stages improves global alignment (E_ϕ: 0.974) by actively filtering channel correlations to suppress background noise and isolate specular reflections. Conversely, applying only SCA in

Method	352×352				704×704			
	$S_\alpha \uparrow$	$F_\beta \uparrow$	$E_\phi \uparrow$	$MAE \downarrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	$E_\phi \uparrow$	$MAE \downarrow$
IPNet (Wang et al., 2024)	0.905	0.870	0.972	0.0085	-	-	-	-
PRNet (Hu et al., 2024)	-	-	-	-	**0.950**	**0.940**	0.981	**0.005**
HIPFNet (Wang et al., 2025)	**0.914**	0.882	0.970	0.008	0.944	0.928	0.980	**0.005**
PMFNet (Zhang et al., 2025)	**0.914**	**0.884**	0.971	0.008	0.941	0.926	0.977	**0.005**
XPol-Net (Ours)	0.910	**0.884**	**0.980**	**0.0078**	**0.950**	0.930	**0.984**	**0.005**

Table 1. Quantitative comparison of XPol-Net against state-of-the-art methods on the PCOD dataset across two spatial resolutions. The best results for each metric are highlighted in **bold**.

Model Variant	CCA	SCA	$S_\alpha \uparrow$	$F_\beta \uparrow$	$E_\phi \uparrow$	$MAE \downarrow$
RGB Baseline			0.900	0.868	0.968	0.0100
Baseline + CCA	✓		0.906	0.876	0.974	0.0087
Baseline + SCA		✓	0.907	0.870	0.970	0.0082
XPol-Net (Full)	✓	✓	**0.914**	**0.884**	**0.980**	**0.0078**

Table 2. Ablation of the Dual Cross-Attention Strategy at 352×352 resolution.

the deeper, semantic stages enhances structural precision (S_α: 0.907, MAE: 0.0082) by aligning high-level RGB tokens with DoLP-derived geometric contours.

The full XPol-Net integrates both mechanisms to achieve peak performance across all metrics. This confirms the critical synergy of our dual approach: early-stage CCA provides a noise-filtered, material-aware feature map, enabling the late-stage SCA to accurately enforce precise, physically compliant spatial boundaries.

5.2 Effectiveness of Multi-Task Auxiliary Learning

Beyond the attention mechanism, XPol-Net employs a multi-task learning objective, utilizing an auxiliary Angle of Linear Polarization (AoLP) prediction head. To validate its necessity, we compare the fully equipped XPol-Net against a variant that retains the dual cross-attention modules but removes the AoLP reconstruction loss ($\mathcal{L}_{AoLP}$), optimizing only for the primary segmentation mask and edges. The results are shown in Table 3.

Model Variant	AoLP Head	$S_\alpha \uparrow$	$F_\beta \uparrow$	$E_\phi \uparrow$	$MAE \downarrow$
Dual Attention w/o AoLP	×	0.906	0.876	0.974	0.0084
XPol-Net (Full)	✓	**0.914**	**0.884**	**0.980**	**0.0078**

Table 3. Ablation of the auxiliary AoLP prediction head at 352×352 resolution.

As demonstrated, removing the AoLP prediction head causes the network's performance to plummet, yielding results (E_ϕ: 0.974, MAE: 0.0084) that are nearly close to the purely optical RGB baseline (E_ϕ: 0.968, MAE: 0.0100). This exposes a critical vulnerability known as modality collapse: because the RGB stream is inherently feature-rich and the cross-attention fusion weights are initialized to zero, the network takes the path of least resistance during optimization. Even with the attention modules structurally present, the model effectively skips the DoLP features entirely, relying strictly on visual color and texture.

By explicitly forcing the network to simultaneously reconstruct the continuous physical AoLP map ($\mathcal{L}_{AoLP}$), we mathematically compel the shared features to utilize the polarimetric cues. This auxiliary task acts as a profound physical regularizer, preventing modality collapse and yielding the peak global structural alignment (E_ϕ: 0.980) observed in our final architecture.

6. Conclusion and Future Work

In this paper, we introduced XPol-Net, a novel multimodal architecture designed to overcome the critical limitations of standard optical sensors in Camouflaged Object Detection (COD). By synergizing a Pyramid Vision Transformer (PVTv2) backbone with our proposed Dual Cross-Attention Strategy, XPol-Net optimally fuses visual semantics with physically-grounded polarimetric priors. The early-stage Channel Cross-Attention (CCA) actively filters material-specific DoLP cues, while the late-stage Spatial Cross-Attention (SCA) enforces strict geometric boundaries. Furthermore, by introducing an explicit Angle of Linear Polarization (AoLP) reconstruction head as a multi-task regularizer, we successfully prevented modality collapse, yielding state-of-the-art global structural alignment (E_ϕ) on the generalized PCOD benchmark.

The primary operational motivation for this architecture is the acceleration of humanitarian demining. Modern unexploded ordnance (UXO) and landmines, which increasingly utilize non-metallic casings to evade traditional detection, exhibit distinct specular polarization signatures against rough, depolarizing natural terrains. XPol-Net's capacity to prioritize high recall and continuous surface mapping demonstrates immense potential for localizing these concealed threats. However, a significant barrier to immediate deployment is the profound lack of domain-specific data; currently, there is no publicly available, large-scale polarimetric dataset explicitly featuring surface-laid or partially buried explosive remnants in natural environments to rigorously test these capabilities.

Therefore, our immediate future work will address this critical data void. We intend to develop and publicly release a dedicated, UAV-acquired polarimetric dataset focused entirely on landmines and UXO across diverse environmental conditions and lighting scenarios. Following the establishment of this dataset, subsequent algorithmic research will concentrate on model distillation and efficiency optimization. By compressing XPol-Net into a lightweight architecture suitable for real-time, onboard UAV edge computing, we aim to deliver a highly scalable and reliable remote sensing tool for post-conflict regions.

Acknowledgment

The authors gratefully acknowledge the support of Dr. Connie Ko and Dr. Avishekh Pal from Teledyne Geospatial, and Dr.

Dan Mashatan, Lyndon Chan, and Aidan Bowers from Teledyne FLIR, for their valuable technical guidance and collaboration throughout this research.

References

Abbas, M. A. et al., 2014. Outdoor rgb-d slam performance in slow mine detection. *Proceedings of the Australasian Conference on Robotics and Automation (ACRA)*.

Bai, S., Xu, T., Liu, P., Qiu, Y., Bai, H., Chen, H., Peng, Y., Li, J., 2026. Hypercod: The first challenging benchmark and baseline for hyperspectral camouflaged object detection.

Baur, J., 2025. Towards Operational UAS-Based Landmine Detection: Vegetation, Validation, and Geophysical Application. PhD thesis, Columbia University.

Brenner, M., Reyes, N. H., Susnjak, T., Barczak, A. L. C., 2023. RGB-D and Thermal Sensor Fusion: A Systematic Literature Review. *IEEE Access*, 11, 82410–82442. http://dx.doi.org/10.1109/ACCESS.2023.3301119.

Cho, S., Ma, J., Yakimenko, O. A., 2023. Aerial multi-spectral AI-based detection system for unexploded ordnance. *Defence Technology*, 27, 24–37.

Faizullin, M., Kornilova, A., Akhmetyanov, A., Pakulev, K., Sadkov, A., Ferrer, G., 2022. SmartDepthSync: Open Source Synchronized Video Recording System of Smartphone RGB and Depth Camera Range Image Frames With Sub-Millisecond Precision. *IEEE Sensors Journal*, 22(7), 7043-7052.

Fan, D.-P., Cheng, M.-M., Liu, Y., Li, T., Borji, A., 2017. Structure-measure: A new way to evaluate foreground maps.

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

Fan, D.-P., Ji, G.-P., Sun, G., Cheng, M.-M., Shen, J., Shao, L., 2020. Camouflaged object detection. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., 2022. Lora: Low-rank adaptation of large language models. *International Conference on Learning Representations (ICLR)*.

Hu, X., Wang, S., Qin, X., Dai, H., Ren, W., Luo, D., Tai, Y., Shao, L., 2023. High-Resolution Iterative Feedback Network for Camouflaged Object Detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(1), 881–889.

Hu, X., Zhang, X., Wang, F., Sun, J., Sun, F., 2024. Efficient Camouflaged Object Detection Network Based on Global Localization Perception and Local Guidance Refinement. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(7), 5452–5465.

Hwang, K.-S., Ma, J., 2024. Military camouflaged object detection with deep learning using dataset development and combination. *The Journal of Defense Modeling and Simulation*.

Javed, U., Butt, M. U. A., Naveed, Z., 2024. UAV-based detection of landmines using infrared thermography. *International Journal of Computational Vision and Robotics*, 1(1). http://dx.doi.org/10.1504/IJCVR.2024.10066473.

Kumar, H., Kumar, K., Mujumdar, A. A., Gaikwad, A., 2025. Sensor-Fusion Approach in a Smart Prodder for Enhanced Mine Detection Accuracy. *International Journal For Multidisciplinary Research (IJFMR)*, 7(5). https://www.ijfmr.com/papers/2025/5/55307.pdf.

Li, Y., Xu, T., Bai, S., Liu, P., Li, J., 2025a. MCOD: The First Challenging Benchmark for Multispectral Camouflaged Object Detection. *arXiv preprint arXiv:2509.15753*.

Li, Y., Zhan, W., Jiang, Y., Guo, J., 2025b. RDCRNet: RGB-T Object Detection Network Based on Cross-Modal Representation Model. *Entropy*, 27(4), 442.

Liu, X., Qi, L., Song, Y., Wen, Q., 2024. Depth awakens: A depth-perceptual attention fusion network for RGB-D camouflaged object detection. *Image and Vision Computing*, 143, 104924.

Ranftl, R., Bochkovskiy, A., Koltun, V., 2021. Vision transformers for dense prediction.

Sun, Y., Wang, S., Chen, C., Xiang, T.-Z., 2022. Boundary-guided camouflaged object detection. *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 186–194.

UC Merced Newsroom, 2025. UC Merced’s Berhe Joins Scientists in Warning of Global Land Mine Crisis. *UC Merced News*. https://news.ucmerced.edu/news/2025/uc-merceds-berhe-joins-scientists-warning-global-land-mine-crisis.

UN News, 2025. Ukraine: Mine contamination is lethal legacy of russia’s invasion. Accessed: 2026-03-24.

Wang, W., Xie, E., Li, X., Fan, D.-P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L., 2022. Pvtv2: Improved baselines with pyramid vision transformer. *Computational Visual Media*, 8(3), 1–10.

Wang, X., Ding, J., Zhang, Z., Xu, J., Gao, J., 2024. IPNet: Polarization-based Camouflaged Object Detection via dual-flow network. *Engineering Applications of Artificial Intelligence*, 127, 107303.

Wang, X., Xu, J., Ding, J., 2025. Polarization-based Camouflaged Object Detection with high-resolution adaptive fusion Network. *Engineering Applications of Artificial Intelligence*, 146, 110245.

Wang, X., Zhang, Z., Gao, J., 2023. Polarization-based Camouflaged Object Detection. *Pattern Recognition Letters*, 174, 106–111.

Wu, R., Xiang, T.-Z., Xie, G.-S., Gao, R., Shu, X., Zhao, F., Shao, L., 2025. Uncertainty-Aware Transformer for Referring Camouflaged Object Detection. *IEEE Transactions on Image Processing*.

Xie, B., Deng, Y., Shao, Z., Li, Y., 2024. EISNet: A Multi-Modal Fusion Network for Semantic Segmentation With Events and Images. *IEEE Transactions on Multimedia*, 26, 8639-8650.

Zhang, L., Rao, A., Agrawala, M., 2023. Adding conditional control to text-to-image diffusion models. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 3836–3847.

Zhang, X., Ru, J., Wang, Y., Wu, C., 2025. Polarization-driven camouflaged object detection: a multimodal fusion network with iterative polarimetric feature enhancement. *Applied Optics*, 64(27), 7899–7913.