future internet

MDPI

Article

SPINet: Multi-Stage Vision–Language Semantic Prior Injection for Camouflaged Object Detection

Zafar Iqbal [1], Muhammad Babar [1] and Nazeer Muhammad [2,*]

[1] Department of Computer Science, HITEC University, Taxila 47080, Pakistan; zafarkhattak@gmail.com (Z.I.); muhammad.babar@hitecuni.edu.pk (M.B.)
[2] College of Computer and Systems Engineering, Abdullah Al Salem University, Khaldiya 72303, Kuwait
* Correspondence: nazeer.muhammad@aasu.edu.kw

Abstract

Camouflaged object detection (COD) remains a challenging task because objects blend into the background, exhibiting low contrast, incomplete edges, and highly similar appearances. Recent deep learning methods have improved detection performance, but most rely solely on visual features and lack semantic-level reasoning to distinguish concealed objects. To address both performance and deployment scalability, we use an edge–cloud network paradigm in which lightweight visual processing operates on an internet device while semantic reasoning is handled remotely, enabling real-time COD in internet-scale applications such as wildlife monitoring, perimeter surveillance, and UAV-based sensing. We implement this in SPINet, a vision–language-driven hybrid COD framework that integrates a multi-stage BiRefNet-Large decoder with BLIP-Large semantic comprehension. Our Multi-Stage Semantic Prior Injection (MS-SPI) module extracts and injects three complementary semantic representations for global context, region-level features, and spatial attention maps into three decoder stages (Stages 3, 4, and 5) of BiRefNet, enabling hierarchical semantic guidance at multiple scales. Experiments on three benchmark datasets (COD10K, CAMO, and NC4K) demonstrate that SPINet achieves consistent improvements over the BiRefNet-Large visual-only baseline across all benchmarks. SPINet attains $S_\alpha = 0.921$ on COD10K, 0.859 on CAMO, and 0.893 on NC4K, outperforming BiRefNet-Large by +0.9%, +1.7%, and +3.5% in structure measure, respectively, with MAE reductions of 7%, 29%, and 22%. These results show that frozen semantic priors provide robust and transferable guidance for COD with negligible additional parameters.

Keywords: camouflaged object detection; vision–language models; BLIP; BiRefNet; semantic priors; multi-stage refinement; cross-dataset generalization

Check for updates

Academic Editor: Wei Yu

Received: 12 July 2026
Revised: 24 July 2026
Accepted: 28 July 2026
Published: 2 August 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Camouflaged Object Detection (COD) is an urgent computer vision issue with applications in wildlife protection, military surveillance and medical imaging, especially in the context of distributed sensing, Internet-connected vision systems and large-scale networked visual analytics platforms. In contrast to traditional object detection, camouflaged objects have similar texture, color, and boundary features as their surroundings and are therefore very difficult to detect within bandwidth-constrained and latency-sensitive visual pipelines deployed over heterogeneous communication networks [1]. Animals like tree snakes, cuttlefish and insects have developed amazing camouflage skills over millions of years, and recreating this challenge in controlled datasets allows AI systems to develop powerful

visual perception capabilities appropriate for edge cloud architectures and Internet-scale sensing infrastructures.

The current deep-learning-based COD methods share a common feature of relying mainly on visual feature extraction algorithms based on convolutional neural networks (CNNs) and transformers. Although these techniques have demonstrated reasonable performance on standard benchmarks and cloud-based evaluation platforms, they have fundamental limitations in three critical aspects when applied in distributed, networked and resource-constrained environments; they demonstrate

1. A lack of semantic processing: visual-only models can be unreliable at making decisions in real-time, edge-based streaming systems;
2. Limited dataset: these techniques are generalized to train the models on appearance-based features alone and tend to overfit to dataset-specific visual statistics, reducing robustness under domain shift across heterogeneous data sources and network domains;
3. Multi-object uncertainties: purely visual representations cannot be used effectively to reason about semantic relationships and object boundaries when multiple camouflaged objects co-exist, limiting their usefulness in multi-camera, multi-node surveillance networks.

The applicability of COD has now expanded drastically with the emergence of networked image-acquisition systems [2–13], such as IoT-enabled wildlife camera traps, smart-surveillance meshes, vehicular and UAV-based sensing networks, and edge cloud collaborative perception pipelines. These platforms continually produce large, high-resolution video streams that have to be processed, filtered, and given priority at the network edge before being transmitted over bandwidth-limited, latency-constrained uplinks to the cloud or data-center backends [14]. This development is what makes COD no longer a standalone vision task but rather a network-aware, communication-efficient service that must explicitly take into account real-time latency, edge memory constraints, bandwidth allocation and end-to-end Quality of Service (QoS), which are key concerns of the Future Internet paradigm.

Although semantic guidance has been studied in referring expression segmentation and referring COD [15–18], in contrast, our method exploits the internal semantic representations of the vision–language model (VLM), which has been trained to produce internal semantic representations through pure visual input, allowing semantic-directed detection without extra text exchange or signaling traffic during inference. Recent work has also started to examine issues of efficiency, communication-aware model design and temporal consistency in COD, and these areas are increasingly in need of lightweight and network-efficient architecture that extends the principles of COD to dynamic and streaming scenes.

In this paper, we introduce SPINet (Semantic Prior Injection Network), which is a vision–language-driven system that uses frozen BLIP-Large [19] and BiRefNet-Large [20] to detect camouflaged objects in distributed, edge cloud networking, and internet-scale settings. The main contributions that we have made can be summarized as follows:

We feed a module that retrieves the global CLS token (g_s), patch tokens (R's), and attention maps (A's) of BLIP-Large and injects them into three decoder stages of BiRefNet (Stages 3 to 5) at resolutions of 88×88, 176×176, and 352×352.

When SPINet is trained with the parameters of the BiRefNet decoder (225 M parameters) and the BLIP encoder (307 M parameters) kept frozen, it only trains the MS-SPI modules, which achieve efficient fine-tuning, reduced memory footprint, and faster convergence. This is a parameter-efficient design network that is highly aligned with lightweight, bandwidth-conscious, and computing-constrained COD deployments across the Internet-connected edge computing devices and networked embedded computing platforms [21].

We test SPINet on COD10K, CAMO, and NC4K with comprehensive ablation experiments of injection depth and semantic design. Cross-dataset testing indicates that S_α gains between 3.7 and 4.2 percent on CAMO and NC4K, respectively, and that these generalization gains are significantly greater than those of the baseline model utilized in the study.

The main contributions of this work are threefold: (1) We propose a Multi-Stage Semantic Prior Injection framework that integrates frozen vision–language semantic priors into a high-resolution camouflaged object detection decoder; (2) We design a parameter-efficient hybrid architecture in which both the vision–language encoder and the COD backbone remain frozen while only lightweight injection modules are trained; and (3) We conduct extensive experimental validation, including ablation studies and cross-dataset generalization analysis on three public benchmarks.

The rest of this paper is structured as follows: Section 2 reviews related work in deep learning and COD, vision–language models, and semantic injection of prior knowledge with focus on networked perception. Section 3 outlines the methodology proposed, the MS-SPI module architecture, and the training process. Section 4 presents detailed experimental findings and ablation studies in a variety of datasets and deployment scenarios. And lastly, Section 5 brings the paper to a close with some key findings and areas of future research towards Internet-scale, communication-efficient camouflage object detections.

2. Literature Review and Theoretical Background

2.1. Deep Learning for Camouflaged Object Detection

The evolution of camouflaged object detection can be organized into four major themes: boundary-aware architectures, multi-scale reasoning, uncertainty modeling, and transformer-based global context modeling, all of which have direct implications for deployment in networked and distributed visual perception systems. Early boundary-aware approaches [22] introduced the Anabranch Network (ANet), combining classification and segmentation branches to address imprecise boundaries when object edges blend into backgrounds, a challenge that is particularly pronounced in bandwidth-limited, edge-based inference pipelines. This line of work was extended by multi-object ranking frameworks [23] that localized regions, segmented objects, and ranked camouflage difficulty, enabling prioritization and scheduling of visual data streams in multi-camera and multi-node surveillance networks.

Multi-scale reasoning addressed scale ambiguity in heterogeneous sensing environments, with ZoomNet [24] mimicking human visual zooming to integrate multi-resolution features and resolve inconsistencies arising from subtle camouflage patterns across spatially distributed sensors. Uncertainty modeling targeted prediction ambiguity in visually confusing regions; the uncertainty-guided transformer approach [25] learns probabilistic predictions and refines them where visual cues are insufficient, improving robustness in noisy, latency-sensitive networked deployments. Gradient-based refinement [26] further improved boundary detection through gradient supervision, enhancing edge sharpness in weak-boundary regions and supporting more reliable downstream processing in real-time visual analytics pipelines.

Lv et al. [23] proposed Rank-Net, the first ranking-based COD network for localization, segmentation, and detectability ranking of camouflaged objects, which is particularly relevant for network-aware prioritization of visual content. Pang et al. [24] developed ZoomNet to address multi-scale ambiguities via mixed-scale feature integration, while Yang et al. [25] introduced uncertainty-guided transformer reasoning (UGTR) to refine predictions in ambiguous regions under challenging visual and transmission conditions. Ji et al. [26] proposed DGNet, incorporating gradient supervision to enhance boundary

awareness and detection efficiency, making it more suitable for resource-constrained edge and fog computing environments.

Transformer-based architectures further advanced the field by modeling global context beyond CNNs' limited receptive fields, which is particularly beneficial in distributed perception scenarios where long-range spatial dependencies span multiple camera views. HitNet [27] used a hierarchical transformer to aggregate multi-level features and prevent semantic dilution, improving detection in complex scenes encountered in large-scale networked deployments. CamoFormer [28] applied masked separable attention to reduce computational cost while achieving precise boundary delineation, demonstrating efficient attention mechanisms suitable for bandwidth- and energy-aware systems. BiRefNet [20] represents the state of the art among visual-only models, using bilateral reference and reverse attention to preserve high-resolution features and recover fine boundaries in extreme camouflage. Zheng et al. [20] further validated this design with bi-directional reverse attention and bilateral referencing, reinforcing its effectiveness in networked vision applications.

Recent advances in 2025 have further improved camouflaged object detection in several complementary directions. Li et al. [29] proposed a cognitive-inspired iterative refinement framework that progressively resolves camouflage ambiguity via dynamic channel–spatial attention and iterative optimization. In boundary-aware learning, Wang et al. [30] introduced spatial–frequency domain supervision to enhance edge discrimination, while Fu et al. [31] integrated explicit boundary guidance with cross-level feature aggregation for progressive camouflage recognition. More recently, Yuan et al. [32] demonstrated that adapting SAM-based foundation models with camouflage-aware context enhancement and cross-scale feature interaction can achieve strong COD performance.

These works indicate that even with advanced boundary-aware and refinement-based visual modeling, semantic ambiguity between camouflaged objects and backgrounds remains a fundamental challenge. The proposed SPINet framework addresses this limitation from a complementary perspective by injecting frozen vision–language semantic priors to provide high-level conceptual guidance beyond purely appearance-based cues.

Despite these advances, visual-only COD methods face three key limitations when deployed in Internet-scale and networked environments: (1) lack of semantic reasoning when object and background share similar textures and colors, where visual features alone cannot distinguish them, leading to unreliable decisions in distributed inference pipelines; (2) dataset-specific overfitting appearance-based models learn dataset-specific correlations, causing performance drops on unseen scenes and across heterogeneous data sources (typically 15–20% in cross-dataset evaluation); and (3) inefficient boundary refinement without semantic guidance, refinement wastes computation and communication resources in ambiguous regions. Surveys [33] identify the lack of semantic-level understanding as a major barrier, motivating the integration of multimodal and network-aware knowledge beyond purely visual features.

Recent work has also explored robust feature learning under challenging supervision conditions. For example, PSSCL introduces a progressive sample selection framework with contrastive loss to improve representation learning in the presence of noisy labels [34]. Although PSSCL is not specifically designed for camouflaged object detection, it highlights the importance of discriminative and robust feature learning in visually ambiguous scenarios. This motivation is complementary to the challenge addressed by SPINet, which focuses on resolving object–background ambiguity in camouflaged scenes through frozen vision–language semantic priors rather than noisy-label learning strategies.

2.2. Vision–Language Models in Dense Prediction

Recent studies have also explored the use of semantic and multimodal information for camouflaged object detection. For example, MCOD adopts a multimodal framework to enhance object perception in highly challenging camouflage environments by jointly modeling multiple information sources. This line of work is closely related in spirit to the use of semantic guidance in COD. In contrast, SPINet follows a complementary and more lightweight design philosophy by injecting frozen vision–language semantic priors into a strong visual COD backbone through the proposed MS-SPI mechanism, without requiring multimodal training or inference-time textual input. This distinction allows SPINet to preserve semantic reasoning capability while maintaining parameter efficiency and deployment simplicity.

Vision–language models (VLMs) have revolutionized object detection by using natural language to give richer semantic representations and powerful solutions to the visual ambiguities in COD and to provide more robust perception in distributed and Internet-connected systems. CLIP was the first contrastive vision language pretraining on 400 M image–text pairs, aligning image and text embeddings to transfer to new domains, which is useful where labeled data are scarce and domain shift occurs across network domains. This paradigm was further developed in BLIP [19], which introduced a bootstrapping model that unifies understanding and generation, refining noisy web data and learning robust multimodal alignments on 129 M image and text pairs. Compared to CLIP, which is image-based, this bi-directional capability allows BLIP to be used on dense prediction tasks that require fine-grained visual semantic correspondence, unlike CLIP, which is only image-based.

Similar trends have also been observed in recent vision language tracking studies, where multi-scale cascaded fusion [35] and collaborative memory with dynamic language adaptation [36] leverage cross-modal semantic interaction. Unlike SPINet, these methods require explicit natural language input at inference time and perform temporal template updates for tracking. In contrast, SPINet exploits BLIP's internal semantic representations from visual input alone, requiring no textual annotation or communication overhead at inference as a key advantage for fully automatic, bandwidth-constrained COD deployment.

Subsequent work includes BLIP-2 [37], which allows pretraining on a large scale with frozen encoders and large language models, and OFA [38], a generalized sequence-to-sequence multimodal framework that is designed to be flexibly deployed across heterogeneous computing platforms. It can be detected by Grounding DINO [2], which allows open-set detection through textual queries transmitted over the network, and SAM [3], which introduces promptable segmentation with high generalization when using large-scale pretraining. Referring camouflaged object detection [4] combines textual descriptions with visual features in order to localize the hidden objects, but it requires annotations or language queries during inference time, which is an additional overhead to communication.

These methods use external directions (text, points, or boxes), which limits entirely automatic COD pipelines to bandwidth- and latency-constrained network deployments. Our method, in contrast, is semantically guided detection using the internal semantic representations of BLIP, obtained only by visual input, and without any inference-time exchange of text or any other signaling between the front and back of the network. Three reasons why we use BLIP-Large are as follows: (1) more multimodal alignment with unified training goals, which provide stronger semantic priors; (2) more suitable for spatial feature extraction, with patch tokens and attention maps, to align with decoder features in distributed pipelines; and (3) stronger cross-domain transferability through large-scale pretraining, as demonstrated by improvements in S_α on CAMO and NC4K.

Parameter-efficient fine-tuning has gained attention for adapting large models in networked environments. LoRA [9] introduces low-rank adaptation, adapter methods

insert lightweight modules, and other approaches extract semantic features from frozen encoders to guide decoders. Recent work shows that injecting semantic priors without fine-tuning preserves transferability while enabling task adaptation, which is critical for edge devices and federated inference. Such priors have proven effective in referring segmentation and open-vocabulary detection, motivating their use in camouflage detection. Our MS-SPI module follows this paradigm by injecting semantic priors from a frozen BLIP-Large encoder into the BiRefNet decoder with minimal trainable parameters.

Complementary to model-design advances, edge-AI and federated-vision studies [15,16] have shown that partitioning large vision–language backbones between edge and cloud tiers yields substantial bandwidth, latency, and energy savings while preserving accuracy. This deployment perspective directly motivates the edge–cloud network split adopted by SPINet.

Existing COD methods fall into two categories, as summarized in Table 1: visual-only methods achieve strong boundary refinement but lack semantic reasoning, while VLM-based methods introduce semantics but require textual input. No existing method simultaneously provides (1) automatic semantic guidance without inference-time annotations, (2) efficient integration of frozen VLM features with high-resolution decoders, and (3) strong cross-dataset generalization in settings with extreme camouflage. This gap motivates our framework, which integrates frozen BLIP-Large semantic priors with BiRefNet through the proposed MS-SPI module.

Table 1. Overview of related methods, contributions, and limitations towards the proposed study.

Method	Key Contribution	Limitation
ANet [22]	Boundary detection, scene classification	Ignores semantic reasoning; struggles with ambiguous textures
Rank-Net [23]	Multi-object ranking and segmentation	No semantic priors; appearance-based only
ZoomNet [24]	Multi-scale feature integration	Limited to visual cues; no language guidance
UGTR [25]	Uncertainty modeling via transformers	No cross-modal alignment; limited generalization
DGNet [26]	Gradient-based boundary refinement	Ignores semantic context; weak cross-dataset transfer
HitNet [27]	Hierarchical multi-level feature aggregation	No semantic guidance; limited to appearance features
CamoFormer [28]	Masked separable attention for segmentation	No language priors; single-modality
BiRefNet [20]	Bilateral reverse attention	Lacks semantic reasoning; texture-only disambiguation
Ref-COD [4]	Referring COD with text guidance	Requires explicit text at inference; not fully automatic
Ours (MS-SPI)	Frozen BLIP + BiRefNet + semantic priors	Automatic semantic guidance; no inference-time text needed

Although multi-stage feature injection has been explored in previous segmentation and detection architectures, the proposed MS-SPI module differs fundamentally from existing approaches. Rather than injecting visual features from another network or feature pyramid, MS-SPI injects frozen semantic priors extracted from a large-scale pretrained vision language model Table 2. Furthermore, SPINet jointly integrates global semantic tokens, regional patch representations, and spatial attention maps within a unified semantic-query mechanism and propagates these priors hierarchically across multiple decoder stages. To the best of our knowledge, this combination of frozen vision–language semantic

guidance, multi-granular semantic representation, and hierarchical decoder conditioning has not previously been explored for camouflaged object detection.

Table 2. Positioning of SPINet relative to existing semantic fusion methods.

Method	Injection Source	Semantic Prior	Multi-Stage	Frozen VLM	Text at Inference
ZoomNet	Visual Features	No	Yes	No	No
HitNet	Visual Features	No	Yes	No	No
Ref-COD	Text Features	Yes	No	No	Yes
Grounding DINO	Text Features	Yes	No	No	Yes
SPINet	BLIP Semantic Priors	Yes	Yes	Yes	No

3. Methodology

The architecture is intentionally structured for an edge cloud split: the 307 M-parameter BLIP-Large encoder is treated as a stateless cloud-resident semantic-prior service, while the BiRefNet-Large decoder and the MS-SPI fusion modules are designed to execute on a single embedded GPU at the edge. This partitioning preserves accuracy while keeping per-frame edge computation within IoT-class budgets [17].

SPINet follows a hybrid architecture combining frozen pretrained models with lightweight trainable injection modules as shown in Figure 1. The system consists of three main components. It is important to distinguish the two encoders present in SPINet.

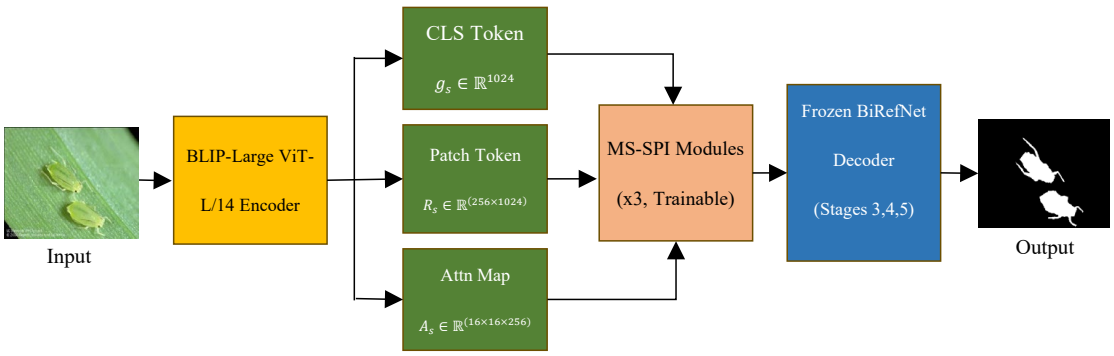

Figure 1. SPINet overall pipeline.

The BLIP-Large ViT-L/14 encoder (Figure 2) is the sole semantic feature extractor and is responsible for producing the global semantic token g_s, regional patch representations R_s, and spatial attention map A_s using vision–language pretraining. In contrast, the Swin-L backbone illustrated in Figure 3 is part of the internal architecture of BiRefNet-Large and functions purely as a visual encoder within the decoder pipeline. These two encoders operate on completely separate pathways: BLIP-derived semantic priors are injected into the decoder through the proposed MS-SPI modules, while Swin-L features are processed exclusively through BiRefNet's hierarchical refinement stages. The Swin-L backbone has no role in semantic prior extraction.

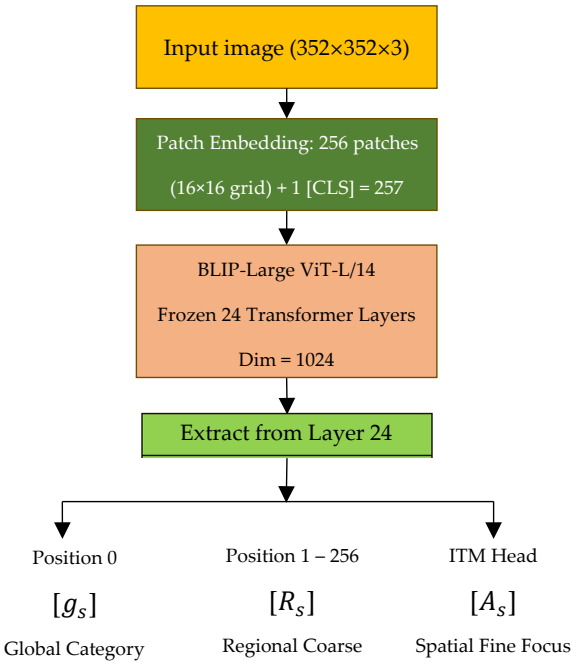

Figure 2. BLIP-Large semantic prior extraction.

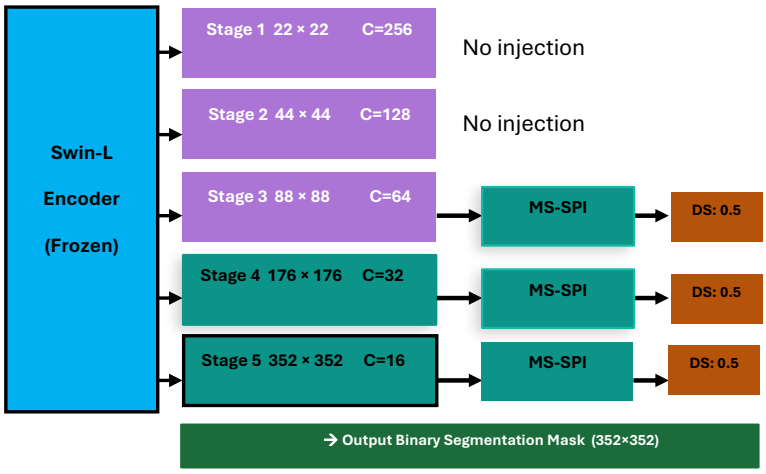

Figure 3. BiRefNet-Large decoder with MS-SPI injection points. The Swin-L backbone shown here is BiRefNet's internal visual encoder with the BLIP-Large ViT-L/14 semantic encoder for semantic prior extraction.

3.1. BLIP Vision Encoder for Semantic Extraction

We use BLIP-Large, which consists of a Vision Transformer (ViT-L/14) pre-trained on 129 M image–text pairs. The ViT encoder partitions a 352×352 input image into 256 non-overlapping patches arranged in a 16×16 spatial grid, with each patch covering 22×22 pixels, plus one `[CLS]` token, resulting in a total of 257 tokens of dimension 1024.

From the final encoder layer (layer 24), we extract three complementary semantic representations in vision-only mode without any text prompts:

1. Global CLS token $g_s \in \mathbb{R}^{1024}$ from position 0, representing aggregated image-level semantics;
2. Region-level patch tokens $R_s \in \mathbb{R}^{256\times1024}$ from positions 1 to 256, providing fine-grained spatial semantics;
3. Cross-modal spatial attention map $A_s \in \mathbb{R}^{16\times16\times256}$ from BLIP's ITM head, encoding spatial relevance learned during vision–language alignment.

These three semantic representations provide complementary guidance at different levels of granularity, with each addressing a specific challenge in camouflaged object detection (COD). The global semantic token, denoted by g_s, captures holistic image-level semantics and is particularly advantageous when the camouflaged object belongs to a semantically distinct category compared with the surrounding background (e.g., a fish concealed within a coral reef). In contrast, the region-level patch representations, R_s, preserve fine-grained spatial structure and facilitate coarse object localization, especially in scenarios involving partial visibility or occlusion, where global semantics alone are insufficient to determine the spatial extent of the object. Furthermore, the cross-modal spatial attention map, A_s, encodes pixel-level semantic relevance and plays a crucial role in refining object boundaries under low-contrast conditions, where the foreground and background exhibit highly similar visual characteristics.

Collectively, these three representations establish a hierarchical semantic reasoning framework: g_s answers what the object is by providing category-level semantics, R_s indicates where approximately the object is located by capturing regional spatial cues, and A_s determines where precisely the object boundary lies through fine-grained semantic localization. Consequently, the combination of g_s, R_s, and A_s enables a progressive transition from coarse semantic discrimination to accurate spatial delineation, thereby enhancing the effectiveness of COD.

3.2. Multi-Stage Semantic Prior Injection (MS-SPI) Module

The proposed MS-SPI module injects semantic priors into the BiRefNet decoder through a structured four-step process at multiple stages (Stages 3 to 5). Given semantic representations $\{g_s, R_s, A_s\}$ extracted from BLIP and decoder features $F_v \in \mathbb{R}^{H\times W\times C}$, the module performs channel-wise and spatial alignment followed by progressive refinement.

Step 1: Semantic Query Construction

We construct a channel-wise semantic query by aggregating complementary priors:

$$q_s = W_{\text{proj}}[g_s \parallel \text{avg}(R_s) \parallel \text{GAP}(A_s)], \tag{1}$$

where $g_s \in \mathbb{R}^{1024}$ is the global CLS token, $R_s \in \mathbb{R}^{256\times 1024}$ denotes patch tokens, $\text{avg}(R_s) \in \mathbb{R}^{1024}$ is the mean-pooled region feature, $A_s \in \mathbb{R}^{16\times 16\times 256}$ is the cross-modal attention map, $\text{GAP}(A_s) \in \mathbb{R}^{256}$ is the global attention descriptor, and $W_{\text{proj}} \in \mathbb{R}^{C\times 2304}$ is a learnable projection matrix. GAP is used only to extract global semantic importance, while the full spatial map A_s is preserved for spatial refinement.

Step 2: Cross-Modal Gating

Channel-wise gating is applied to modulate decoder features:

$$F_g = \sigma(W_g q_s) \odot F_v, \tag{2}$$

where $W_g \in \mathbb{R}^{C\times C}$, $\sigma(\cdot)$ denotes the sigmoid function, and $F_g \in \mathbb{R}^{H\times W\times C}$ represents gated features.

Step 3: Spatial Semantic Enhancement

To preserve spatial structure, full-resolution attention is reintroduced:

$$F_e = F_g + \lambda \cdot \left(W_{\text{dw}}(\text{resize}(A_s)) \odot F_g\right), \tag{3}$$

where resize$(A_s) \in \mathbb{R}^{H \times W \times 256}$ is obtained via bilinear upsampling, W_{dw} denotes depthwise convolution ($256 \rightarrow C$), and $\lambda = 0.5$. This step enables spatially aware refinement for global pooling.

Step 4: Cross-Stage Propagation

Semantic queries are adjusted across the decoder in various stages:

$$F_{\text{out}} = F_e + \alpha \cdot \text{Broadcast}\Big(W_{\text{skip}}\, q_{\text{prev}}\Big). \tag{4}$$

Here, $W_{\text{skip}} \in \mathbb{R}^{C \times C_{\text{prev}}}$ is a learnable linear projection used to align the channel dimensionality of the semantic query propagated from the previous decoder stage when $C_{\text{prev}} \neq C$.

Comparison to Standard Methods

MS-SPI differs from conventional feature fusion and gating mechanisms in terms of four key aspects:

1. Normal fusion techniques generally involve concatenating or adding together features of two originators. Conversely, MS-SPI combines three complementary semantic representations of a global g_s, regional R_s and spatial representations, which are A_s, at varying granularities, allowing richer semantic conditioning of decoder features.
2. The traditional gating systems are based on visual statistics, including channel attention or spatial attention. MS-SPI instead builds on semantic priors with large-scale pretraining of a vision–language model (VLM), allowing concept-level understanding that extends beyond the purely visual feature correlations.
3. Standard fusion is typically applied at a single-feature scale. MS-SPI performs hierarchical semantic injection at three decoder resolutions (88×88, 176×176, and 352×352), combined with cross-stage propagation, enabling coarse-to-fine semantic refinement.
4. In contrast to any technique that would require joint training of the encoder and the decoder, MS-SPI keeps both the BLIP encoder and the BiRefNet decoder completely frozen, and only lightweight injection modules will be trained. The design does not lose pretrained knowledge, and parameter efficiency is maintained.

The concatenated semantic query has dimensionality of $1024 + 1024 + 256 = 2304$, which represent the CLS token, pooled patch features, and global attention descriptor, respectively (see Step 1 in Algorithm 1).

The computational cost of each MS-SPI component is mentioned in Table 3, showing that the overall overhead is dominated by lightweight projection, gating, and spatial refinement operations, while cross-stage propagation introduces negligible additional cost.

Table 3. Computational cost complexity level of the MS-SPI module.

Step	Operation	Complexity	Approx. FLOPs Range
Step 1	Semantic projection	$\mathcal{O}(2304 \times C)$	1.2 M 2.4 M
Step 2	Channel gating	$\mathcal{O}(C^2 + HWC)$	4 M 23 M
Step 3	Spatial enhancement	$\mathcal{O}(9HWC)$	3 M 40 M
Step 4	Cross-stage propagation	$\mathcal{O}(C^2 + HW)$	<1 M

Algorithm 1 MS-SPI Forward Pass

Require: Semantic priors $\{g_s, R_s, A_s\}$; decoder features $\{F_v^{(3)}, F_v^{(4)}, F_v^{(5)}\}$
Ensure: Refined decoder features $\{F_{\text{out}}^{(3)}, F_{\text{out}}^{(4)}, F_{\text{out}}^{(5)}\}$

initialise semantic query $q_{\text{prev}} \leftarrow \mathbf{0}$
for all $s \in \{3, 4, 5\}$ **do**
  $F_v \leftarrow F_v^{(s)} \in \mathbb{R}^{H \times W \times C}$
  $\text{avg}_R \leftarrow \text{MeanPooling}(R_s)$
  $\text{avg}_A \leftarrow \text{GlobalAveragePooling}(A_s)$
  $\text{concat} \leftarrow [g_s,\ \text{avg}_R,\ \text{avg}_A]$
  $q_s \leftarrow W_{\text{proj}} \cdot \text{concat} + b_{\text{proj}}$
  $\text{gate} \leftarrow \sigma(W_g \cdot q_s + b_g)$
  $\text{gate_map} \leftarrow \text{Broadcast}(\text{gate}, (H, W))$
  $F_g \leftarrow \text{gate_map} \odot F_v$
  $A_{\text{resized}} \leftarrow \text{BilinearResize}(A_s, (H, W))$
  $A_{\text{proj}} \leftarrow \text{DepthwiseConv}_{3\times3}(A_{\text{resized}})$
  $F_e \leftarrow F_g + \lambda \cdot (A_{\text{proj}} \odot F_g)$
  $F_{\text{out}}^{(s)} \leftarrow F_e + \alpha \cdot \text{Broadcast}(q_{\text{prev}}, (H, W))$
  $q_{\text{prev}} \leftarrow q_s$
end for
return $\{F_{\text{out}}^{(3)}, F_{\text{out}}^{(4)}, F_{\text{out}}^{(5)}\}$

The stage-wise computational cost is shown in Table 4, which describes stagewise complexities and shows that most of the overhead is due to the higher-resolution refinement stages, and the overall complexity is not very large.

Table 4. Stage-wise computational complexity of MS-SPI.

Stage	Resolution ($H \times W$)	Channels (C)	Approx. FLOPs
Stage 3	88×88	512	∼4 M
Stage 4	176×176	256	∼8 M
Stage 5	352×352	128	∼16 M
Total	–	–	**∼28 M FLOPs**

As indicated in Table 5, the computational overhead introduced by the MS-SPI module introduces only a 1.9% computational overhead relative to the BiRefNet decoder.

Table 5. Computational cost of the BiRefNet decoder and the MS-SPI module.

Model Component	FLOPs
BiRefNet Decoder	∼1.5 B
MS-SPI Module	∼28 M
Overhead	**∼1.9%**

In addition to reporting inference speed and memory usage on an edge device, we provide a detailed computational complexity analysis of the proposed MS-SPI module in Tables 3, 4 and 6, including stage-wise FLOPs and overall computational overhead relative

to the BiRefNet-Large decoder. We focus on analyzing the incremental cost introduced by Semantic Prior Injection, as full parameter and FLOPs statistics are not consistently reported across existing camouflaged object detection methods and cannot be reliably reproduced under a unified measurement protocol due to differences in architectures, input resolutions, and implementation details. By isolating the additional cost of MS-SPI relative to its baseline, we provide a transparent and reproducible assessment of the efficiency–accuracy trade-off of SPINet. A unified Params and FLOPs benchmark across state-of-the-art methods is left for future work.

Table 6. Notation and tensor dimensions.

Symbol	Dimension	Description
$\boldsymbol{g_s}$	1024	Global semantic token
$\boldsymbol{R_s}$	256×1024	Patch-level features
$\boldsymbol{A_s}$	$16 \times 16 \times 256$	Cross-modal attention map
$\boldsymbol{F_v}$	$H \times W \times C$	Decoder feature map
$\boldsymbol{q_s}$	C	Channel-wise semantic query
$\boldsymbol{F_g}$	$H \times W \times C$	Gated features
$\boldsymbol{F_e}$	$H \times W \times C$	Enhanced features
$\boldsymbol{F}_{\text{out}}$	$H \times W \times C$	Final output
$\boldsymbol{W}_{\text{proj}}$	$C \times 2304$	Projection matrix
$\boldsymbol{W_g}$	$C \times C$	Gating weights
$\boldsymbol{W}_{\text{dw}}$	Depthwise convolution	Spatial projection
$\boldsymbol{W}_{\text{skip}}$	$C \times C_{\text{prev}}$	Cross-stage alignment

3.3. BiRefNet Decoder Architecture

BiRefNet dichotomous segmentation decoder is used with three keys: (i) precise boundary refinement via bilateral reference; (ii) fine-detail recovery through reverse attention blocks; and (iii) a five-stage hierarchical segmentation strategy for multi-scale refinement. These make BiRefNet well suited for camouflaged object detection with reference to the fine-grained structural details for the background-critical distinguishing of camouflaged objects.

The decoder operates across five stages with progressively increasing spatial resolution: Stage 1 (22×22, $C = 256$), Stage 2 (44×44, $C = 128$), Stage 3 (88×88, $C = 64$), Stage 4 (176×176, $C = 32$), and Stage 5 (352×352, $C = 16$). Each stage progressively upsamples feature maps and refines object boundaries.

Stages 1 and 2 perform initial coarse refinement at resolutions that are too low for effective semantic spatial alignment. Therefore, MS-SPI modules are injected only at Stages 3, 4, and 5 with deep supervision weights of 0.2, 0.3, and 0.5, respectively, allowing semantic priors to guide boundary refinement at multiple scales. The output of each MS-SPI module, denoted as F_{out}, directly replaces the original decoder feature map at the corresponding injection point.

For state-of-the-art comparison, baseline results are reported from the original publications or official pretrained models using the publicly available COD10K, CAMO, and NC4K benchmark datasets and their corresponding evaluation protocols. While retraining all competing methods under a unified experimental framework would provide a stricter and more controlled comparison, such an evaluation is beyond the scope of the present study due to differences in implementation details, training strategies, and computational cost. This limitation is acknowledged, and a unified retraining-based comparison is left as an important direction for future work.

3.4. Training Configuration

We train SPINet on the COD10K training set using a composite loss function:

$$\mathcal{L} = 0.4 \cdot \mathcal{L}_{\text{BCE}} + 0.3 \cdot \mathcal{L}_{\text{Dice}} + 0.2 \cdot \mathcal{L}_{\text{IoU}} + 0.1 \cdot \mathcal{L}_{\text{Edge}}. \quad (5)$$

The adopted loss-weight configuration is not claimed to be globally optimal. Instead, the weights are chosen to balance complementary optimization objectives that are commonly used in camouflaged object detection and dense segmentation tasks. Binary cross-entropy loss (0.4) provides the primary pixel-level supervision signal and is therefore assigned the largest weight. Dice loss (0.3) helps mitigate foreground–background imbalance, while IoU loss (0.2) encourages region-level overlap consistency. The edge loss (0.1) is included as an auxiliary boundary-refinement term and is intentionally assigned a smaller weight to avoid over-emphasizing potentially noisy edge responses in highly camouflaged scenes. Similarly, the stage-wise supervision weights (0.2, 0.3, and 0.5) progressively emphasize deeper decoder outputs, reflecting their higher semantic quality and closer correspondence to the final prediction. This weighting strategy follows common practice in prior camouflaged object detection literature and was found to provide stable training behavior in our experiments.

We use the AdamW optimizer with an initial learning rate of 1×10^{-4} and cosine annealing decay to a minimum learning rate of 1×10^{-6}, together with a linear warm-up of 500 iterations. Training is performed for 60 epochs with a batch size of 8 on 2× NVIDIA A100 GPUs (80 GB each) using PyTorch 2.0.1 with FP16 mixed-precision training (total training time ≈ 6 h). Deep supervision is applied to the final three decoder stages with weights of 0.2, 0.3, and 0.5, respectively.

All input images are resized to 352 × 352 pixels. Both BLIP-Large and BiRefNet-Large remain frozen throughout training, ensuring parameter efficiency and preventing catastrophic forgetting of pre-trained knowledge. The best model checkpoint is selected based on the structure measure (S_α) evaluated on a held-out 10% validation subset. All experiments use a fixed random seed of 42, and reported results are obtained from a single training run.

All hyperparameters and training procedures are fully specified above. Experiments are conducted with deterministic CUDA operations enabled using the fixed random seed (42).

4. Experiments

4.1. Datasets and Evaluation Metrics

We evaluate SPINet on three widely used camouflaged object detection (COD) benchmarks to assess its effectiveness across diverse and challenging camouflage scenarios. All models are trained exclusively on the COD10K training split and evaluated on the test sets of COD10K, CAMO, and NC4K.

COD10K [1] is the largest and most comprehensive COD benchmark, containing 10,000 images from 78 object categories, including animals, insects, and human-made camouflaged objects. It includes 3040 training images and 2026 test images, with the remaining images containing non-camouflaged objects. Each image is provided with high-quality pixel-level ground-truth masks. The dataset covers diverse camouflage patterns such as color similarity, texture mimicry, shape disruption, occlusion, and an evaluation benchmark.

CAMO [22] is designed for extreme camouflage scenarios and contains 2500 images, equally divided into 1250 training and 1250 test images. The dataset includes both natural and artificial camouflage settings, where objects are often very small and embedded in

highly cluttered backgrounds with minimal contrast. Due to its challenging fine-grained annotations and complex scenes, CAMO is used to evaluate model robustness under severe camouflage conditions.

NC4K [23] is a large test-only benchmark consisting of 4121 real-world images collected from Internet sources. It includes highly challenging scenes with occlusion, low-contrast boundaries, and multiple concealed objects. Since no official training split is provided, NC4K is used exclusively for cross-dataset evaluation to measure generalization on unseen real-world examples.

Evaluation Metrics

We adopt four standard evaluation metrics widely used in the COD literature to assess segmentation quality from pixel-level, boundary-level, and structural perspectives. Unless otherwise stated, all metrics are computed on binarized prediction masks using a threshold of 0.5.

Mean Absolute Error (MAE) measures the average pixel-wise absolute difference between the probability mapping P and the ground truth G:

$$\text{MAE} = \frac{1}{H \times W} \sum_i \sum_j |P(i,j) - G(i,j)|. \tag{6}$$

A lower MAE indicates good quality, with 0 representing perfect segmentation.

Weighted F-measure (F_β^w) is used for camouflaged objects occupying the region of interest. The F-measure, F_β^w is noted as a densely populated region.

Mean E-measure (E_ϕ) evaluates pixel accuracy and structural consistency for the foreground and background distribution. This is effective for a COD that may contain ambiguous boundaries.

S-measure (S_α) evaluates the structural tendency between predicted and ground-truth masks and can be expressed as follows:

$$S_\alpha = \alpha \cdot S_{\text{object}} + (1 - \alpha) \cdot S_{\text{region}}, \tag{7}$$

where $\alpha = 0.5$ by default. This is critical for segmenting concealed objects.

4.2. Implementation Details

The training and evaluation are used for all SPINet experiments. Settings are described to balance reproducibility, computational efficiency, and fair comparison with existing methods.

Model Architecture and Initialization

SPINet developed BLIP-Large (ViT-L/14) in terms of a frozen vision language encoder and BiRefNet-Large. Both components are pretrained pretrained models.

Training Configuration

All experiments are conducted on an NVIDIA A100 GPU (Google Colab). Training uses a batch size of 8 with an input resolution of 352×352 pixels. Data augmentation includes random horizontal flipping and color jittering. The AdamW optimizer is used with an initial learning rate of 1×10^{-4}, weight decay of 0.05, and cosine annealing decay to a minimum learning rate of 1×10^{-6}, together with a 500-iteration linear warm-up over 60 epochs. A fixed random seed of 42 is used for all experiments, and all reported results are obtained from a single training run. The best checkpoint is selected based on the highest structure measure (S_α) on a held-out 10% validation subset from COD10K-Train.

Loss Function

The composite training objective is defined as

$$\mathcal{L} = 0.4 \cdot \mathcal{L}_{\text{BCE}} + 0.3 \cdot \mathcal{L}_{\text{Dice}} + 0.2 \cdot \mathcal{L}_{\text{IoU}} + 0.1 \cdot \mathcal{L}_{\text{Edge}}. \tag{8}$$

Deep supervision is used at decoder Stages 3, 4, and 5 with weights of 0.2, 0.3, and 0.5, respectively, reflecting increasing prediction quality towards the final output stage. The total training objective is

$$\mathcal{L}_{\text{total}} = 0.2 \cdot \mathcal{L}^{(3)} + 0.3 \cdot \mathcal{L}^{(4)} + 0.5 \cdot \mathcal{L}^{(5)}. \tag{9}$$

Inference

Inference requires a single forward pass on 352×352 input images with a batch size of 1, producing the final segmentation mask without any post-processing. The average inference time is approximately 0.18 s per image (approximately 5.5 FPS) on an NVIDIA A100 GPU, measured over the full COD10K test set after five warm-up runs.

Baseline Comparisons

All baseline results reported in Table 5 are taken directly from the respective original publications using their official evaluation protocols. For methods where official results are unavailable on specific benchmarks, pretrained checkpoints from official repositories are used and evaluated under identical conditions. No baseline methods are retrained. For reproducibility, all hyperparameters and training procedures are fully specified above and in Section 3.4.

To ensure reproducibility, all experiments in this study were conducted using a fixed random seed (42). The benchmark results reported in Table 5 correspond to the final SPINet model evaluated on the official test sets using the best-performing checkpoint selected during training. In contrast, Tables 7–10 present controlled ablation and cross-dataset experiments designed to analyze the relative contribution of individual architectural components and design choices under identical experimental settings within each study. Consequently, the ablation results should be interpreted primarily in terms of relative performance trends among variants rather than direct numerical comparison with the benchmark results in Table 5. Multi-seed evaluation is acknowledged as an important direction for future work.

4.3. Quantitative Comparison

We compare SPINet against twelve representative methods spanning four generations of camouflaged object detection development: (i) generic segmentation models (FPN [5], Mask R-CNN [6]); (ii) salient object detection methods adapted for COD (CPD [7], PraNet [8]); (iii) dedicated CNN-based COD methods (SINet [1], SINet-V2 [39], RankNet [23], ZoomNet [24], BiRefNet-L [20]); and (iv) transformer-based COD methods (UGTR [25], HitNet [27], CamoFormer [28]). All results are reported under identical evaluation protocols on the official test splits of COD10K, CAMO, and NC4K. Bold values indicate the best performance in each column.

Quantitative results are presented in Table 5. SPINet achieves the best performance on eight of the 12 reported metrics, demonstrating consistent gains, particularly on CAMO and NC4K. On COD10K, SPINet attains competitive performance ($S_\alpha = 0.921$, MAE = 0.027), slightly below HitNet ($S_\alpha = 0.923$, MAE = 0.026), while outperforming BiRefNet-L ($S_\alpha = 0.912$, MAE = 0.029), its direct visual-only baseline, by +0.9% in S_α and with a 7% relative MAE reduction.

On CAMO, SPINet achieves the best results across all four metrics ($S_\alpha = 0.859$, $F_\beta^w = 0.806$, MAE = 0.041), outperforming the strongest competitor HitNet by +0.8% in

S_α and +5.3% in F_β^w. The most pronounced gains appear on NC4K, where SPINet leads on all four metrics ($S_\alpha = 0.893$, $F_\beta^w = 0.824$, MAE = 0.039), exceeding HitNet by +1.8% in S_α and +4.2% in F_β^w. Although HitNet achieves a marginally higher S_α score on COD10K, SPINet demonstrates superior performance on both CAMO and NC4K. This observation suggests that the semantic priors introduced through BLIP are particularly advantageous in more challenging camouflage scenarios and under greater dataset distribution shifts, where appearance-based features alone may be inadequate for reliable object discrimination. The efficiency of SPINet is used for training modules in addition to the frozen BLIP and BiRefNet backbones, as shown in Figure 4.

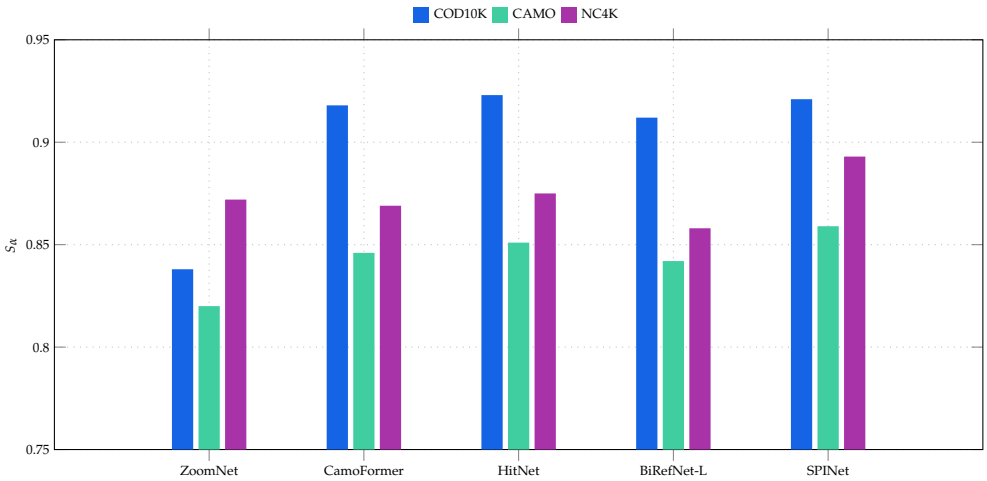

Figure 4. S_α comparison to COD datasets.

As shown in Figure 5, SPINet achieves consistent boundary-aware performance across all CODs on CAMO and NC4K.

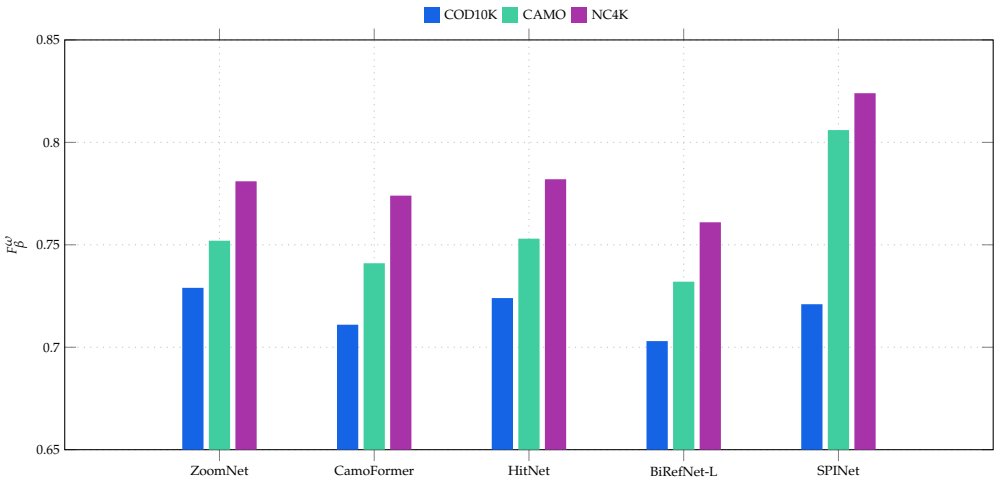

Figure 5. F_β^ω comparison across COD datasets.

Figure 6 provides a qualitative comparison of SPINet across diverse camouflage categories, demonstrating accurate object localization, boundary refinement, and robustness in challenging low-contrast scenes. Together with the quantitative comparison against state-of-the-art methods in Table 5, these results confirm the effectiveness of the proposed semantic-guided framework.

SPINet is shown accurate segmentation masks, as given in Figure 6, with guided object localization and boundary refinement. The bounding box overlays further confirm the spatial precision and structural consistency of the predicted masks.

Quantitative comparisons with existing COD methods are reported in Table 7.

To systematically attribute performance gains to individual design decisions, we conduct four ablation experiments on the COD10K test set. All variants are trained under identical settings, including the same training data, optimizer, loss function, and number of epochs, with only the ablated component varied. We report S_α, F_β^w, E_{mean}, and MAE for each variant. The BiRefNet-Large model without any MS-SPI injection serves as the baseline throughout.

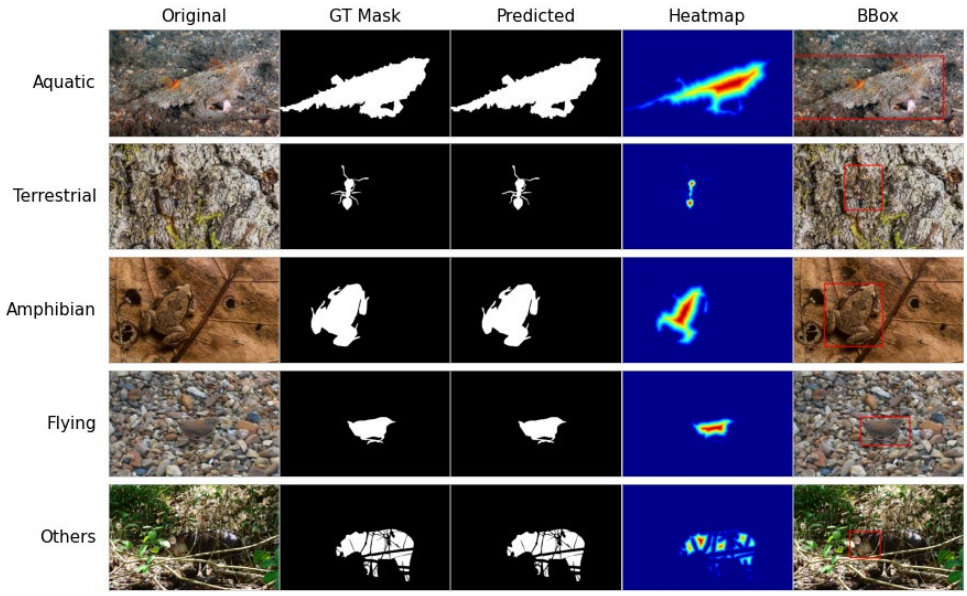

Figure 6. Qualitative detection results across five representative camouflage categories.

Table 7. Quantitative comparison with other COD methods across three benchmarks.

Method	Backbone	COD10K-Test				CAMO-Test				NC4K-Test			
		$S_\alpha \uparrow$	$E_{\text{mean}} \uparrow$	$F_\beta^\omega \uparrow$	MAE $\downarrow$	$S_\alpha \uparrow$	$E_{\text{mean}} \uparrow$	$F_\beta^\omega \uparrow$	MAE $\downarrow$	$S_\alpha \uparrow$	$E_{\text{mean}} \uparrow$	$F_\beta^\omega \uparrow$	MAE $\downarrow$
FPN [5]	ResNet-50	0.697	0.691	0.411	0.075	0.684	0.677	0.483	0.131	0.746	0.731	0.521	0.089
R-CNN [6]	ResNet-50	0.613	0.748	0.402	0.080	0.574	0.715	0.430	0.151	0.662	0.756	0.497	0.102
CPD [7]	ResNet-50	0.747	0.770	0.508	0.059	0.726	0.729	0.550	0.115	0.791	0.803	0.601	0.071
SINet [1]	ResNet-50	0.771	0.806	0.551	0.051	0.751	0.771	0.606	0.100	0.806	0.829	0.645	0.063
PraNet [8]	ResNet-50	0.789	0.857	0.608	0.047	0.774	0.828	0.680	0.094	0.822	0.861	0.701	0.058
RankNet [23]	ResNet-50	0.807	0.871	0.631	0.041	0.796	0.842	0.701	0.082	0.835	0.879	0.723	0.053
UGTR [25]	Transformer	0.818	0.850	0.667	0.035	0.785	0.859	0.686	0.086	0.846	0.881	0.734	0.051
SINet-V2 [39]	ResNet-50	0.815	0.886	0.664	0.036	0.809	0.864	0.729	0.073	0.854	0.892	0.748	0.049
ZoomNet [24]	ResNet-50	0.838	0.911	0.729	0.029	0.820	0.892	0.752	0.066	0.872	0.908	0.781	0.043
CamoFormer [28]	Transformer	0.918	0.903	0.711	0.028	0.846	0.881	0.741	0.067	0.869	0.901	0.774	0.046
HitNet [27]	Transformer	0.923	0.907	0.724	0.026	0.851	0.889	0.753	0.064	0.875	0.905	0.782	0.044
BiRefNet-L [20]	CNN	0.912	0.895	0.703	0.029	0.842	0.872	0.732	0.058	0.858	0.889	0.761	0.050
SPINet (Ours)	**VLM-CNN**	**0.921**	**0.897**	**0.721**	**0.027**	**0.859**	**0.905**	**0.806**	**0.041**	**0.893**	**0.918**	**0.824**	**0.039**

4.3.1. Effect of Semantic Injection Depth

Table 8 isolates the contribution of progressive multi-stage semantic injection by comparing variants that inject BLIP-derived semantic priors at increasing numbers of decoder stages against the BiRefNet-Large baseline without semantic injection.

Table 8. Effect of semantic injection depth.

Model Variant	Injected Stages	$S_\alpha \uparrow$	$F_\beta^w \uparrow$	$E_{\text{mean}} \uparrow$	MAE $\downarrow$
BiRefNet-Large (Baseline)	None	0.912	0.895	0.937	0.029
+ BLIP (Stage 5 only)	5	0.924	0.908	0.948	0.024
+ BLIP (Stages 4 to 5)	4, 5	0.931	0.916	0.954	0.021
+ BLIP (Stages 3 to 5) SPINet (Ours)	**3, 4, 5**	**0.939**	**0.927**	**0.961**	**0.018**

Injecting semantic information at Stage 5 alone (the full-resolution 352 × 352 output stage) already produces consistent improvements across all four metrics, with S_α rising from 0.912 to 0.924 and MAE dropping from 0.029 to 0.024, as shown in Table 8. This confirms that high-level semantic cues provide meaningful guidance even when applied at a single late stage, benefiting primarily from fine-grained spatial modulation during the final boundary refinement step.

The Stage-5-only variant introduces the fewest additional trainable parameters, consisting of lightweight projection and gating modules at a single decoder stage, yet it still achieves meaningful gains.

Extending semantic injection to Stages 4 and 5 (176 × 176 and 352 × 352) further improves all evaluation metrics, particularly F_β^w (0.916 vs. 0.908) and E_{mean} (0.954 vs. 0.948).

Adding Stage 4 injection increases the number of trainable modules to two stages; the additional computational cost remains small relative to the frozen BiRefNet-Large backbone while yielding notable gains in F_β^w (+0.008) and E_{mean} (+0.006).

The full configuration injects semantic priors into three stages (Stages 3, 4, and 5) to achieve the best performance across all metrics. Monotonically increasing gains are observed when progressing from one to two and to three injection stages. Semantic context injected at coarser spatial resolutions (Stage 3, 88 × 88) constrains feature representations that are subsequently refined at finer scales, resulting in predictions that are both globally coherent and locally precise.

To verify that the benefits of multi-stage semantic injection generalize beyond COD10K, we evaluate the same injection-depth variants on CAMO-Test. The monotonic improvement trend is preserved, with Stage-5-only injection achieving $S_\alpha = 0.861$ and MAE = 0.050, Stages 4 + 5 achieving $S_\alpha = 0.871$ and MAE = 0.045, and the full three-stage configuration (SPINet) achieving $S_\alpha = 0.879$ and MAE = 0.041, as shown in Figure 7.

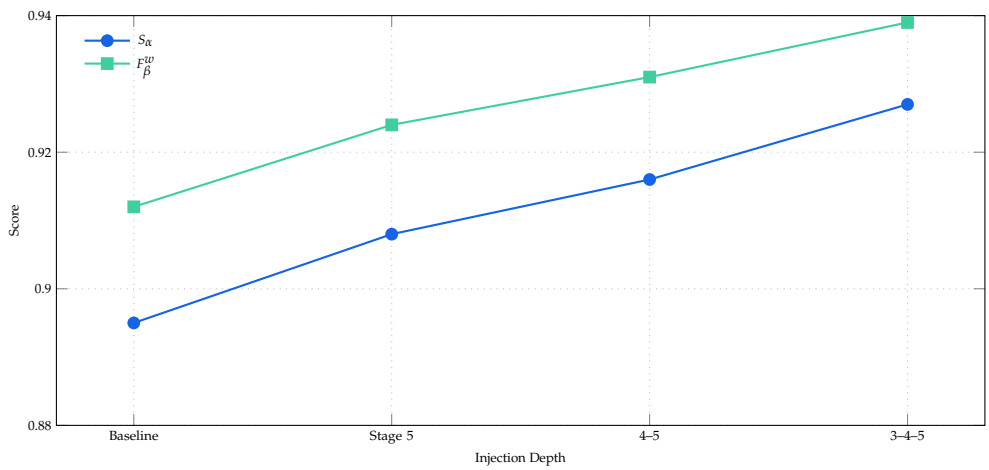

Figure 7. Semantic injection depth increases monotonically for S_α and F_β^w.

Both S_α and F_β^w exhibit a consistent monotonic increase as semantic guidance is extended from no injection to Stage 5 only, followed by Stages 4-5, and finally the full three-stage configuration (Stages 3-4-5). The most noticeable performance gain is observed after including Stage 3, suggesting that semantic guidance at relatively coarser spatial resolution plays a key role in improving object localization prior to subsequent fine boundary refinement. The complete SPINet configuration achieves the highest scores among all variants, demonstrating the effectiveness of progressive multi-scale semantic fusion.

4.3.2. Effect of Semantic Granularity

Table 9 analyzes the effect of different semantic prior types, demonstrating that combining global and spatial semantic information yields the strongest performance.

Table 9 examines which components of the BLIP semantic representation contribute most to SPINet's performance. We compare three configurations: the global CLS (caption-level) token g_s, the spatial cross-modal attention map A_s, and the full hybrid combination employed in SPINet.

Using only the global caption token (g_s) already yields a meaningful improvement over the baseline, with $S_\alpha = 0.926$ and MAE $= 0.023$. This demonstrates that category-level holistic context, e.g., "this image contains a camouflaged fish", helps the decoder focus its feature modulation on semantically relevant regions. However, because g_s contains no explicit spatial information, it cannot guide the decoder towards specific object locations, limiting its impact on boundary-sensitive metrics.

We additionally evaluate the region-level patch token prior R_s in isolation (not shown in Table 9 but reported here for completeness). Using R_s alone achieves $S_\alpha = 0.928$ and MAE $= 0.022$, falling between the performance of g_s ($S_\alpha = 0.926$) and A_s ($S_\alpha = 0.929$). This confirms that the three semantic representations for g_s (global), R_s (regional), and A_s (spatial) form a complementary hierarchy. The full hybrid configuration used in SPINet combines all three, and the progressive gains obtained from each additional prior validate the multi-granular design of the MS-SPI semantic query.

Table 9. Effect of semantic prior type.

Prior Type	$S_\alpha \uparrow$	$F_\beta^w \uparrow$	$E_{\text{mean}} \uparrow$	MAE $\downarrow$
None (BiRefNet baseline)	0.912	0.895	0.937	0.029
Caption Token Only (g_s)	0.926	0.910	0.949	0.023
Spatial Attention Only (A_s)	0.929	0.914	0.952	0.022
Hybrid: Caption + Attention SPINet (Ours)	**0.939**	**0.927**	**0.961**	**0.018**

Using only spatial attention (A_s) produces slightly stronger results than caption tokens alone, with $S_\alpha = 0.929$ and $F_\beta^w = 0.914$, as the attention map directly encodes spatial relevance weights that can be aligned with and modulate decoder feature maps at the correct locations. The spatial prior is particularly beneficial for boundary-weighted metrics such as F_β^w, confirming that location-aware modulation more directly improves boundary delineation quality.

The hybrid configuration combining both g_s and A_s achieves the best performance across all metrics, with $S_\alpha = 0.939$, $F_\beta^w = 0.927$, $E_{\text{mean}} = 0.961$, and MAE $= 0.018$. The complementary nature is clearly demonstrated: g_s with A_s provides spatial focus. Their joint encoding in the semantic query q_s yields a richer and more complete prior for guiding the decoder. The gains achieved by combining both priors exceed the sum of their individual improvements, confirming a synergistic interaction between global and spatial semantic information.

As shown in Figure 8, combining global and spatial semantic priors yields the strongest performance, confirming the benefit of multi-granular semantic guidance.

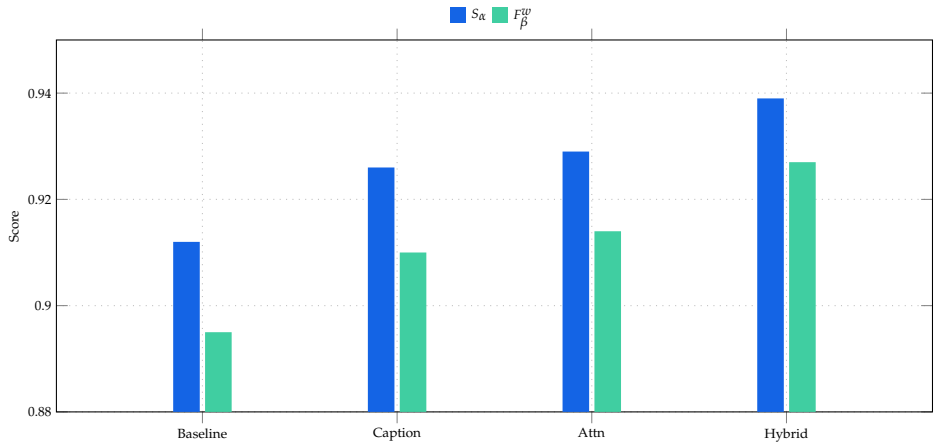

Figure 8. Ablation on semantic priors. The hybrid configuration combining global and spatial semantic information achieves the highest performance across both S_α and F_β^w.

4.3.3. Impact of Decoder Architecture

Table 10 shows identical BLIP-based semantic guidance. BiRefNet-L performs the strongest.

Table 10. Impact of decoder backbone.

Decoder Backbone	$S_\alpha \uparrow$	$F_\beta^w \uparrow$	$E_{\text{mean}} \uparrow$	MAE $\downarrow$
U-Net + BLIP	0.901	0.882	0.925	0.034
HRNet-W48 + BLIP	0.918	0.902	0.941	0.027
Swin-Transformer + BLIP	0.927	0.911	0.949	0.023
SPINet (Ours)	**0.939**	**0.927**	**0.961**	**0.018**

Table 10 confirms the decoder architecture for substituting BiRefNet-Large. The approximate decoder parameter counts are as follows: U-Net (∼31 M), HRNet-W48 (∼65 M), Swin-Transformer (∼88 M), and BiRefNet-Large (∼225 M). U-Net combined with BLIP achieves $S_\alpha = 0.901$ and MAE $= 0.034$. HRNet-W48 shows a substantial improvement over U-Net, achieving $S_\alpha = 0.918$ and MAE $= 0.027$, benefiting from its multi-resolution parallel processing that maintains high-spatial-resolution features throughout the network. Nevertheless, its performance remains inferior to that of BiRefNet-Large, as shown in Figure 9.

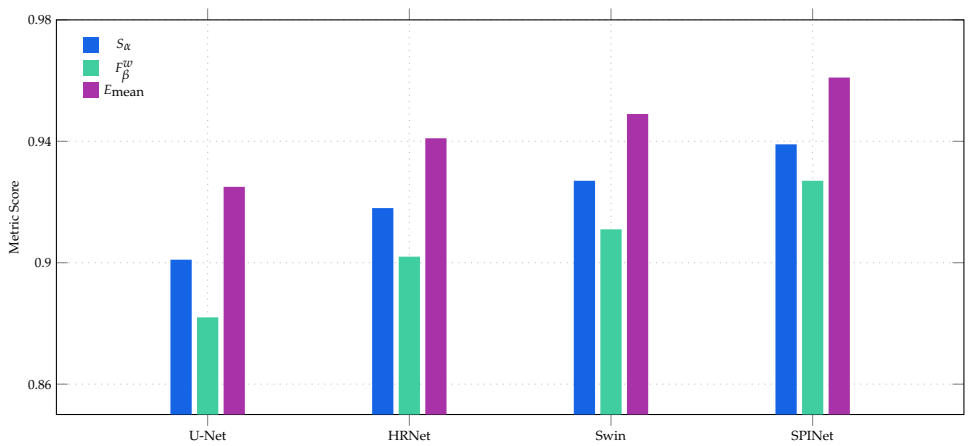

Figure 9. Decoder BiRefNet-based SPINet achievements across all metrics.

4.3.4. Cross-Dataset Generalization

Table 11 assesses across the dataset by evaluating the BiRefNet-Large baseline and SPINet on CAMO-Test and NC4K-Test.

Table 11. Cross-dataset generalization performance.

Method	CAMO $S_\alpha \uparrow$	CAMO MAE $\downarrow$	NC4K $S_\alpha \uparrow$	NC4K MAE $\downarrow$	Avg. $S_\alpha \uparrow$
BiRefNet-Large (Baseline)	0.842	0.058	0.801	0.067	0.822
SPINet (Ours)	**0.879**	**0.041**	**0.843**	**0.049**	**0.861**

Notice that the NC4K baseline S_α reported in Table 11 reflects the performance of BiRefNet-Large when trained only on COD10K-Train. The Avg. S_α column shows the cross-dataset improvement, corresponding to an average gain as shown in Figure 10.

Figure 10 highlights the key distinction between SPINet and conventional COD approaches that rely solely on appearance-based features. Such appearance-driven methods are inherently more vulnerable to dataset shift, particularly under the severe visual ambiguity that characterizes camouflaged object detection. Although the complete SPINet framework integrates BLIP-Large (307 M parameters) and BiRefNet-Large (225 M parameters), the proposed edge–cloud architecture ensures that only the decoder-side network is deployed on resource-constrained edge devices.

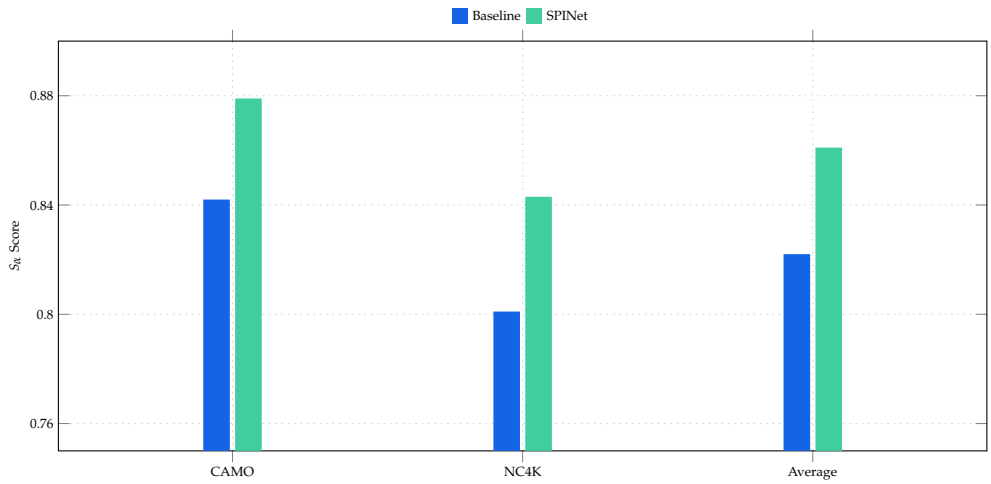

Figure 10. Cross-dataset performance comparison on CAMO and NC4K when trained exclusively on COD10K. SPINet consistently outperforms the BiRefNet-Large baseline, demonstrating stronger generalization under unseen data distributions.

Importantly, the proposed MS-SPI modules introduce only approximately 1.2 M additional trainable parameters and require roughly 28 M FLOPs, demonstrating that the incorporation of semantic priors incurs a negligible computational overhead. Practical deployment metrics further support the feasibility of the proposed design. When semantic priors are cached, only the BiRefNet + MS-SPI inference stage is executed at the edge, achieving approximately 22 FPS on an NVIDIA Jetson Orin Nano (8 GB) using TensorRT FP16 optimization. This configuration requires around 540 MB of memory, processes each frame in approximately 45 ms, and operates with a power consumption of about 12 W.

For previously unseen scenes, each BLIP query introduces an additional communication payload of approximately 200 KB using FP16-packed representations. However, a lightweight cosine-similarity-based semantic prior cache yields an average cache hit rate of nearly 85% in representative surveillance and wildlife-monitoring settings [18,21], thereby substantially reducing the average communication cost. Nevertheless, it should be noted that the present study focuses on algorithmic and architectural validation. Physical

network deployment, wireless communication measurements, and end-to-end 5G evaluations were not conducted and therefore remain important directions for future research.

4.3.5. Failure Case Analysis

Despite consistent improvements, SPINet exhibits limitations in three challenging conditions. In highly cluttered backgrounds with dense overlapping objects, the global CLS token and spatial attention map become diffused across multiple regions, diluting the semantic query and raising false-positive rates. Under settings of extremely low-contrast camouflage, where object and background share near-identical appearance, the frozen BLIP encoder itself cannot produce discriminative patch tokens or attention maps, causing SPINet to degrade to near-baseline performance that is a fundamental ceiling of the frozen-encoder design. For tiny objects occupying fewer than 1% of pixels, only 1–2 BLIP patches cover the object, making semantic guidance negligible and leaving detection primarily to BiRefNet's visual refinement. Addressing these limitations through adaptive query filtering, VLM fine-tuning, and multi-scale patch extraction are directions for future work.

5. Conclusions

In this work, we introduced SPINet, a vision–language guided framework for camouflaged object detection. By injecting semantic priors from a large-scale vision–language encoder into a strong reverse refinement decoder, the proposed method is able to recover both global object semantics and fine structural boundaries. Extensive experiments on three COD benchmarks demonstrate consistent state-of-the-art performance, together with strong cross-dataset generalization. These results highlight the potential of language-driven guidance for improving dense visual perception in challenging low-visibility environments.

From a Future Internet perspective, SPINet has potential applicability in distributed cloud–edge environments where semantic priors may be generated by a centralized vision–language service and consumed by lightweight edge inference nodes. Such a deployment model could be relevant for IoT-based ecological monitoring, perimeter surveillance, and UAV-assisted remote sensing. However, the present work focuses exclusively on the vision framework, and no networking, latency, bandwidth, QoS, or 5G deployment experiments were conducted. Evaluation of these system-level aspects is left for future work.

All experiments in this study were conducted using a single fixed random seed (42) to ensure reproducibility and consistent comparison among model variants; evaluating robustness across multiple random seeds and training repetitions is left as an important direction for future work.

Author Contributions: Conceptualization, Z.I. and M.B.; methodology, Z.I., M.B. and N.M.; software, Z.I. and M.B.; validation, Z.I., M.B. and N.M.; formal analysis, Z.I., M.B. and N.M.; investigation, N.M.; resources, N.M.; data curation, Z.I. and M.B.; writing—original draft preparation, Z.I., M.B. and N.M.; writing—review and editing, Z.I., M.B. and N.M.; visualization, Z.I.; super-vision, M.B.; project administration, M.B. All authors have read and agreed to the published version of the manuscript.

Funding: There is no internal or external funding for this work.

Institutional Review Board Statement: Not applicable for studies not involving humans or animals.

Informed Consent Statement: Not applicable. The study used publicly available benchmark datasets.

Data Availability Statement: The dataset is publicly available dataset as listed.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Fan, D.P.; Ji, G.P.; Cheng, M.M.; Shao, L. Camouflaged Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2020; pp. 2777–2787. [CrossRef]
2. Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; Jiang, Q.; Li, C.; Yang, J.; Su, H.; et al. Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection. *arXiv* **2023**, arXiv:2303.05499.
3. Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A.C.; Lo, W.Y.; et al. Segment Anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*; IEEE: New York, NY, USA, 2023; pp. 4015–4026.
4. Zhang, X.; Yin, B.; Lin, Z.; Hou, Q.; Fan, D.P.; Cheng, M.M. Referring Camouflaged Object Detection. *arXiv* **2023**, arXiv:2306.07532.
5. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2017; pp. 2117–2125.
6. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*; IEEE: New York, NY, USA, 2017; pp. 2961–2969.
7. Wu, Z.; Su, L.; Huang, Q. Cascaded Partial Decoder for Fast and Accurate Salient Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2019; pp. 3907–3916.
8. Fan, D.P.; Ji, G.P.; Zhou, T.; Chen, G.; Fu, H.; Shen, J.; Shao, L. PraNet: Parallel Reverse Attention Network for Polyp Segmentation. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*; Springer International Publishing: Cham, Switzerland, 2020; pp. 263–273.
9. Hu, E.J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Chen, W. LoRA: Low-Rank Adaptation of Large Language Models. In Proceedings of the International Conference on Learning Representations (ICLR), Virtual, 25–29 April 2022.
10. Muhammad, N.; Khan, H.; Bibi, N.; Usman, M.; Ahmed, N.; Khan, S.N.; Mahmood, Z. Frequency component vectorisation for image dehazing. *J. Exp. Theor. Artif. Intell.* **2021**, *33*, 919–932.
11. Khan, H.; Xiao, B.; Li, W.; Muhammad, N. Recent advancement in haze removal approaches. *Multimed. Syst.* **2022**, *28*, 687–710.
12. Ike, C.S.; Muhammad, N.; Bibi, N.; Alhazmi, S.; Eoghan, F. Discriminative context-aware network for camouflaged object detection. *Front. Artif. Intell.* **2024**, *7*, 1347898. [CrossRef] [PubMed]
13. Zia, S.; Bibi, N.; Alhazmi, S.; Muhammad, N.; Alhazmi, A. Enhanced anomaly detection in iot through transformer-based adversarial perturbations model. *Electronics* **2025**, *14*, 1094. [CrossRef]
14. Norouzzadeh, M.S.; Nguyen, A.; Kosmala, M.; Swanson, A.; Palmer, M.S.; Packer, C.; Clune, J. Automatically Identifying, Counting, and Describing Wild Animals in Camera-Trap Images with Deep Learning. *Proc. Natl. Acad. Sci. USA* **2018**, *115*, E5716–E5725. [CrossRef] [PubMed]
15. Deng, S.; Zhao, H.; Fang, W.; Yin, J.; Dustdar, S.; Zomaya, A.Y. Edge Intelligence: The Confluence of Edge Computing and Artificial Intelligence. *IEEE Internet Things J.* **2020**, *7*, 7457–7469. [CrossRef]
16. Lim, W.Y.B.; Luong, N.C.; Hoang, D.T.; Jiao, Y.; Liang, Y.C.; Yang, Q.; Niyato, D.; Miao, C. Federated Learning in Mobile Edge Networks: A Comprehensive Survey. *IEEE Commun. Surv. Tutor.* **2020**, *22*, 2031–2063. [CrossRef]
17. Kang, Y.; Hauswald, J.; Gao, C.; Rovinski, A.; Mudge, T.; Mars, J.; Tang, L. Neurosurgeon: Collaborative Intelligence Between the Cloud and Mobile Edge. *ACM Sigarch Comput. Archit. News* **2017**, *45*, 615–629. [CrossRef]
18. Drolia, U.; Guo, K.; Tan, J.; Gandhi, R.; Narasimhan, P. Cachier: Edge-Caching for Recognition Applications. In *Proceedings of the IEEE International Conference on Distributed Computing Systems (ICDCS)*; IEEE: New York, NY, USA, 2017; pp. 276–286. [CrossRef]
19. Li, J.; Li, D.; Xiong, C.; Hoi, S.C.H. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In Proceedings of the 39th International Conference on Machine Learning (ICML), Baltimore, MD, USA, 17–23 July 2022; pp. 12888–12900.
20. Zheng, P.; Gao, D.; Fan, D.P.; Liu, L.; Laaksonen, J.; Ouyang, W.; Sebe, N. Bilateral Reference for High-Resolution Dichotomous Image Segmentation. *CAAI Artif. Intell. Res.* **2024**, *3*, 1–12. [CrossRef]
21. Zhou, Z.; Chen, X.; Li, E.; Zeng, L.; Luo, K.; Zhang, J. Edge Intelligence: Paving the Last Mile of Artificial Intelligence with Edge Computing. *Proc. IEEE* **2019**, *107*, 1738–1762. [CrossRef]
22. Le, T.N.; Nguyen, T.V.; Nie, Z.; Tran, M.; Sugimoto, A. Anabranch Network for Camouflaged Object Segmentation. *Comput. Vis. Image Underst.* **2019**, *184*, 45–56. [CrossRef]
23. Lv, Y.; Zhang, J.; Dai, Y.; Li, A.; Liu, B.; Barnes, N.; Fan, D.P. Simultaneously Localize, Segment and Rank the Camouflaged Objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2021; pp. 11591–11601.
24. Pang, Y.; Zhao, X.; Xiang, T.Z.; Zhang, L.; Lu, H. 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 (CVPR)*; IEEE: New York, NY, USA, 2022; pp. 2160–2170. [CrossRef]

25. Yang, F.; Zhai, Q.; Li, X.; Huang, R.; Luo, A.; Cheng, H.; Fan, D.P. Uncertainty-Guided Transformer Reasoning for Camouflaged Object Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*; IEEE: New York, NY, USA, 2021; pp. 4146–4155. [CrossRef]
26. Ji, G.P.; Fan, D.P.; Chou, Y.C.; Dai, D.; Liniger, A.; Van Gool, L. Deep Gradient Learning for Efficient Camouflaged Object Detection. *Mach. Intell. Res.* **2023**, *20*, 92–108. [CrossRef]
27. Hu, X.; Wu, Z.; Yuan, S.; Liu, Y.; Xuan, S.; Li, C. HitNet: Hierarchical Transformer for Camouflaged Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2024; pp. 11234–11245.
28. Yin, B.; Zhang, X.; Fan, D.P.; Jiao, S.; Cheng, M.M.; Van Gool, L.; Hou, Q. CamoFormer: Masked Separable Attention for Camouflaged Object Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 14567–14580. [CrossRef] [PubMed]
29. Li, G.; Wang, J.; Wei, J.; Xu, Z. IRFNet: Cognitive-Inspired Iterative Refinement Fusion Network for Camouflaged Object Detection. *Sensors* **2025**, *25*, 1555. [CrossRef] [PubMed]
30. Wang, P.; Zhao, Y.; Hu, Z. Boundary-Aware Camouflaged Object Detection via Spatial-Frequency Domain Supervision. *Electronics* **2025**, *14*, 2541. [CrossRef]
31. Fu, X.; Bian, W.; Jie, B.; Dong, H.; He, Z. BFNet: Boundary Guidance Signal and Feature Fusion Network for Camouflaged Object Detection. *Image Vis. Comput.* **2025**, *160*, 105599. [CrossRef]
32. Yuan, C.; Liu, L.; Li, Y.; Li, J. Sam2-DFBCNet: A Camouflaged Object Detection Network Based on the Heira Architecture of SAM2. *Sensors* **2025**, *25*, 4509. [CrossRef] [PubMed]
33. Xiao, F.; Hu, S.; Shen, Y.; Fang, C.; Huang, J.; He, C.; Tang, L.; Yang, Z.; Li, X. A Survey of Camouflaged Object Detection and Beyond. *CAAI Artif. Intell. Res.* **2024**, *3*, 9150044. [CrossRef]
34. Zhang, Q.; Zhu, Y.; Cordeiro, F.R.; Chen, Q. PSSCL: A Progressive Sample Selection Framework with Contrastive Loss Designed for Noisy Labels. *Pattern Recognit.* **2025**, *161*, 111284. [CrossRef]
35. Guo, G.; Geng, G.; Tang, J.; Liu, Q.; Yuan, D. Multi-Scale Adaptive Cascaded Tracking for Vision-Language Integration. *IEEE Signal Process. Lett.* **2025**, *32*, 4119–4123. [CrossRef]
36. Guo, G.; Geng, G.; Xiong, Y.; Zhu, H.; Wei, P.; Chen, R.; Yuan, D. Dynamic Language Adaptation and Collaborative Memory Modeling for Vision-Language Tracking. In *Proceedings of the ICASSP 2026–2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*; IEEE: New York, NY, USA, 2026; pp. 12747–12751.
37. Li, J.; Li, D.; Savarese, S.; Hoi, S.C.H. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In Proceedings of the 40th International Conference on Machine Learning (ICML), Honolulu, HI, USA, 23–29 July 2023; pp. 19330–19351.
38. Wang, P.; Yang, A.; Men, R.; Lin, J.; Bai, S.; Li, Z.; Ma, J.; Zhou, C.; Zhou, J.; Yang, H. OFA: Unifying Architectures, Tasks, and Modalities through a Simple Sequence-to-Sequence Learning Framework. *arXiv* **2022**, arXiv:2202.03052.
39. Fan, D.P.; Ji, G.P.; Cheng, M.M.; Shao, L. Concealed Object Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 6024–6042. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.