

C3Net: Context-Contrast Network for Camouflaged Object Detection

Baber Jan, Aiman H. El-Maleh, Abdul Jabbar Siddiqui, Abdul Bais and Saeed Anwar

Abstract—Camouflaged object detection identifies objects that blend seamlessly with their surroundings through similar colors, textures, and patterns. This task challenges both traditional segmentation methods and modern foundation models, which fail dramatically on camouflaged objects. We identify six fundamental challenges in COD: Intrinsic Similarity, Edge Disruption, Extreme Scale Variation, Environmental Complexities, Contextual Dependencies, and Salient-Camouflaged Object Disambiguation. These challenges frequently co-occur and compound the difficulty of detection, requiring comprehensive architectural solutions. We propose C3Net, which addresses all challenges through a specialized dual-pathway decoder architecture. The Edge Refinement Pathway employs gradient-initialized Edge Enhancement Modules to recover precise boundaries from early features. The Contextual Localization Pathway utilizes our novel Image-based Context Guidance mechanism to achieve intrinsic saliency suppression without external models. An Attentive Fusion Module synergistically combines the two pathways via spatial gating. C3Net achieves state-of-the-art performance with S-measures of 0.898 on COD10K, 0.904 on CAMO, and 0.913 on NC4K, while maintaining efficient processing. C3Net demonstrates that complex, multifaceted detection challenges require architectural innovation, with specialized components working synergistically to achieve comprehensive coverage beyond isolated improvements. Code, model weights, and results are available at <https://github.com/Baber-Jan/C3Net>.

Impact Statement—C3Net advances camouflaged object detection with implications across multiple domains. Politically, standardized detection algorithms promote transparent and accountable automated systems. Economically, industrial defect detection applications can improve manufacturing quality control efficiency. Socially, enhanced polyp detection in colonoscopy screening supports early cancer diagnosis, addressing healthcare challenges where camouflaged lesions are frequently missed. Technologically, C3Net's dual-pathway architecture and intrinsic saliency suppression mechanism advance computer vision capabilities, demonstrating how architectural innovation solves complex detection challenges. Environmentally, automated wildlife monitoring enables efficient population tracking for conservation research. Legally, our research uses publicly available datasets without personal data collection, demonstrating privacy-conscious development practices. C3Net shows that specialized architectures can address multi-faceted vision challenges while maintaining ethical research standards. We encourage responsible deployment guided by domain-specific requirements and

regulatory frameworks.

Index Terms—Camouflaged Object Detection, Image segmentation, Deep learning, Edge detection, Pattern recognition

I. INTRODUCTION

CAMOUFLAGED objects blend into their surroundings through shared colors, textures, and patterns. Detecting such objects is termed Camouflaged Object Detection (COD) and differs fundamentally from conventional object detection [1]. While conventional detection targets objects with distinct visual features, camouflaged objects exhibit a high degree of intrinsic similarity to their backgrounds. This similarity renders standard segmentation methods ineffective, necessitating specialized approaches. State-of-the-art foundation models like SAM2 [2] achieve remarkable performance on general segmentation but experience dramatic degradation on camouflaged objects [3]. This performance gap motivates the development of specialized COD architectures. COD enables critical applications including medical polyp detection [4], wildlife monitoring [5], and industrial inspection [6].

Camouflaged object detection confronts six fundamental challenges that frequently co-occur and compound difficulty. Intrinsic Similarity (IS) forms the foundation where objects share identical colors, textures, and patterns with their backgrounds and thus become visually indistinguishable (Figure 1, row i). This similarity directly contributes to Edge Disruption (ED), where object boundaries fragment or vanish entirely and make precise segmentation impossible (Figure 1, row ii). These boundary ambiguities become critical with Extreme Scale Variation (ESV), where objects occupy minimal pixels or exhibit extreme aspect ratios and challenge detection networks (Figure 1, row iii). Environmental Complexities (EC) further degrade visibility through shadows, occlusions, and changes in illumination that obscure already-weak boundaries (Figure 1, row iv). Because local features fail under these conditions, Contextual Dependencies (CD) require models to integrate global scene information, yet local details remain essential for accuracy (Figure 1, row v). Compounding these difficulties, Salient-Camouflaged Object Disambiguation (SCOD) occurs where scenes contain both camouflaged and salient objects, and models must detect only the camouflaged target while classifying the prominent salient object as background (Figure 1, row vi). These interconnected challenges demand comprehensive architectures that address their complex interactions rather than isolated mechanisms.

Researchers have pursued various approaches to address these interconnected challenges, yet none achieve comprehensive coverage. Early CNN-based methods, such as SINet [9],

Manuscript submitted to IEEE Transactions on Artificial Intelligence, Month Day, Year.

B. Jan, A. H. El-Maleh, and A. J. Siddiqui are with the Computer Engineering Department, King Fahd University of Petroleum and Minerals, Dhahran 31261, Saudi Arabia (e-mail: baberjan008@gmail.com; aimane@kfupm.edu.sa; abduljabbar.siddiqui@kfupm.edu.sa).

A. Bais is with Electronic Systems Engineering, University of Regina, Canada (e-mail: Abdul.Bais@uregina.ca).

S. Anwar is with the Department of Computer Science and Software Engineering, The University of Western Australia, Perth, WA 6009, Australia (e-mail: saeed.anwar@uwa.edu.au).

Corresponding author: Saeed Anwar

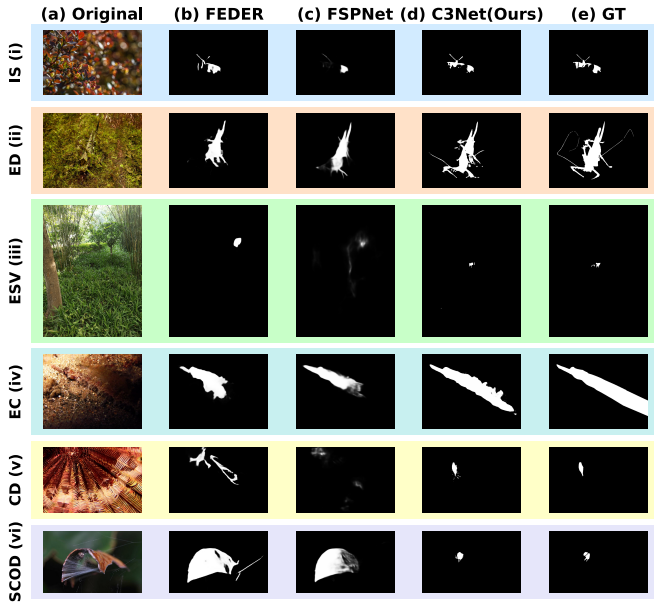


Fig. 1. Visual comparison demonstrating C3Net’s comprehensive handling of COD challenges. Each row illustrates a fundamental challenge: (i) Intrinsic Similarity (IS) - a camouflaged insect blends with leaves; (ii) Edge Disruption (ED) - an insect exhibits fragmented boundaries against ground; (iii) Extreme Scale Variation (ESV) - a bear barely visible in dense vegetation occupies minimal pixels; (iv) Environmental Complexities (EC) - shadows and terrain variations obscure half of the ground creature; (v) Contextual Dependencies (CD) - the insect requires global context for accurate segmentation; (vi) Salient-Camouflaged Object Disambiguation (SCOD) - a camouflaged insect must be distinguished from the prominent bark. C3Net consistently outperforms FEDER [7] (CNN-based SOTA) and FSPNet [8] (ViT-based SOTA) across all challenges.

introduced specialized modules for extracting subtle cues but struggled with integrating global context. Multi-scale fusion [10] and iterative refinement [11] improved local feature processing while boundary precision remained limited. Transformer architectures [8], [12] achieved superior global reasoning but generated coarse segmentation masks that missed fine details. Joint salient and camouflaged object detection [13] attempted to handle SCOD through multi-task learning, yet required external saliency models. These evolutionary improvements address specific challenges while leaving others unsolved. CNN methods excel at capturing local details but struggle with CD. Transformers capture global context but struggle with ED and ESV. Current architectures lack mechanisms to handle all six challenges systematically and particularly fail to suppress intrinsic saliency without external dependencies.

We propose C3Net (Context-Contrast Camouflaged Object Detection Network) to provide systematic coverage of all challenges. Our C3Net introduces specialized decoder pathways that process complementary visual cues rather than pursuing incremental improvements. Our approach separates edge refinement from contextual understanding, preventing signal dilution while ensuring synergistic operation. The architecture incorporates intrinsic mechanisms for each major challenge and integrates them through adaptive fusion. This design enables state-of-the-art performance across all benchmarks while maintaining processing efficiency. C3Net demonstrates

that effective detection emerges from architectural integration rather than isolated components.

The main contributions of this paper are summarized as follows:

- We design a dual-pathway decoder in which edge refinement and contextual localization operate at different feature levels. This separation prevents signal dilution and enables specialized learning for each visual cue type.
- We introduce Edge Enhancement Modules with multi-path convolutions initialized from gradient and Laplacian operators. These modules maintain classical edge detection principles while adapting to camouflage patterns through learning.
- We develop the Image-based Context Guidance mechanism that analyzes input appearance directly. This approach achieves intrinsic saliency suppression through contrast computation, eliminating the need for external models.
- We create an Attentive Fusion Module that spatially gates edge information via contextual pathways. This design emphasizes relevant boundaries while suppressing distractors at each spatial location.
- We formulate pathway-specific loss objectives with precision focus for saliency suppression and recall focus for complete capture. This strategy ensures each component learns its specialized role effectively.

II. RELATED WORK

Camouflaged object detection methods have evolved through distinct methodological approaches over the past decade. Each approach targets specific detection challenges while advancing the field toward more comprehensive solutions. This section examines these methodological directions and their contributions to addressing COD challenges.

Progressive Detection Strategies. Early COD research established fundamental detection paradigms through progressive refinement. SINet [9] pioneered COD as a distinct task and introduced search-identification modules to handle high intrinsic similarity between objects and backgrounds. PFNet [14] advanced this approach by incorporating distraction mining to explicitly address false positives and false negatives through positioning and focus modules. These methods demonstrated success in identifying potential camouflaged regions and progressively refining predictions. However, these methods primarily focused on local-region analysis and progressive refinement, without comprehensive global scene understanding. This limitation motivated the exploration of architectures with enhanced global modeling capabilities.

Multi-Scale and Iterative Approaches. Researchers developed multi-scale strategies to address the scale variation challenges in COD. ZoomNet [10] employed mixed-scale triplet networks to capture discriminative features at different zoom levels, while SegMaR [11] used iterative segment-magnify-reiterate strategies for progressive refinement. These methods improved object detection across scales through hierarchical processing. Yet, these methods primarily addressed scale-related aspects without comprehensive mechanisms for

other challenges, such as saliency suppression and contextual dependencies. The approaches revealed that scale handling alone cannot substitute for a comprehensive understanding of features.

Transformer-Based Global Modeling. The adoption of transformer architectures brought enhanced global context modeling to COD. FSPNet [8] addressed the limitations of standard transformers in locality modeling through non-local token enhancement and feature shrinkage pyramids. CamoFormer [15] employed masked separable attention to model foreground, background, and global context with distinct attention heads. These approaches significantly improved handling of contextual dependencies and global feature relationships. However, transformer architectures inherently prioritize global over local modeling, generating coarse feature representations. This architectural characteristic causes them to struggle with edge disruption (ED) and extreme scale variation (ESV), both of which require fine details. The trade-off reveals that global and local understanding require fundamentally different architectural treatments.

Edge and Boundary Enhancement. Specialized methods emerged to address the boundary precision challenges in COD. BGNet [16] combined boundary guidance with dual-branch global-local context integration to improve edge detection while maintaining contextual understanding. FEDER [7] explicitly tackled both intrinsic similarity and ambiguous boundaries through feature decomposition and ODE-inspired edge reconstruction. These approaches achieved significant improvements in boundary quality and demonstrated that edge enhancement benefits from integration with semantic understanding. However, their primary focus on boundary-related challenges left other aspects, such as extreme scale variation and comprehensive saliency handling, as secondary considerations. This specialization pattern shows that targeting specific challenges often comes at the expense of comprehensive coverage.

Joint Learning and Saliency Handling. The contradictory relationship between salient and camouflaged objects motivated joint learning approaches. Salient-Camouflaged Object Disambiguation (SCOD) occurs when scenes contain both prominent salient objects and subtle camouflaged objects, where models must detect only the camouflaged target while classifying the salient object as background. UJSC [13] pioneered this direction through uncertainty-aware training with data-wise correlation modeling, task-wise correlation modeling, and adversarial learning to distinguish between opposing visual characteristics. Zhao et al. [17] proposed saliency attribute transfer to spot camouflaged objects, while recent USCOD [18] introduced the CS12K dataset with four scene types and the Camouflage-Saliency Confusion Score metric for comprehensive evaluation. These approaches demonstrated that explicitly modeling the SOD-COD relationship improved detection accuracy and provided systematic evaluation frameworks. However, standard COD benchmarks lack salient object annotations and thus models requiring saliency supervision cannot be fairly evaluated on established datasets. Joint training frameworks also introduced computational overhead and

external dependencies through multi-task setups or saliency models. This reveals that while saliency handling is crucial for SCOD, intrinsic mechanisms without external dependencies remain an open challenge.

Integration and Hybrid Approaches. Researchers combined these methodological directions to create comprehensive solutions. UJSC [13] integrated joint saliency-camouflage learning with uncertainty modeling, while BGNet [16] merged boundary guidance with global-local processing. HDPNet [19] combined hourglass transformer structures with dual-path pyramids to balance global understanding and local detail preservation. These integration efforts demonstrated that combining complementary approaches could simultaneously address multiple challenges and achieve notable improvements across various benchmarks. However, the integration strategies primarily relied on established fusion techniques that may not fully exploit the complementary nature of different components. The field offers opportunities for architectures in which specialized components interact through learned, adaptive mechanisms that enhance their collective effectiveness. This suggests potential for designs that achieve integration through architectural innovation rather than additive complexity.

Existing Gaps and Opportunities. An analysis of existing methods reveals systematic gaps in comprehensively addressing COD challenges. While individual approaches excel at specific aspects, no current architecture provides integrated solutions for all six fundamental challenges. Edge-focused methods achieve precise boundaries but lack contextual understanding for SCOD. Transformer approaches capture global dependencies but sacrifice fine details essential for ED and ESV. Joint learning frameworks handle saliency disambiguation but introduce external dependencies that complicate deployment. Most critically, existing architectures lack intrinsic mechanisms for saliency suppression and rely on external models or multi-task setups. The field requires architectures that address these challenges through synergistic design rather than isolated improvements. Opportunities exist for methods that combine specialized processing pathways with adaptive fusion mechanisms, thereby maintaining both global coherence and local precision. Such architectures could achieve comprehensive coverage through architectural innovation rather than component accumulation.

III. METHODOLOGY

Our proposed C3Net model follows an encoder-decoder architecture with specialized components, as shown in Figure 2. Here, we provide an overview of our C3Net. Initially, a vision transformer encoder extracts features from the input image **I** containing camouflaged objects, which are then fed into our dual-pathway decoder architecture. Our Edge Enhancement Modules process early features to recover precise boundaries. Similarly, our Image-based Context Guidance mechanism processes deep features to identify objects while suppressing salient distractors. Both mentioned modules have different pathways that utilize content-adaptive upsampling to preserve fine details. Moreover, our Attentive Fusion Module combines

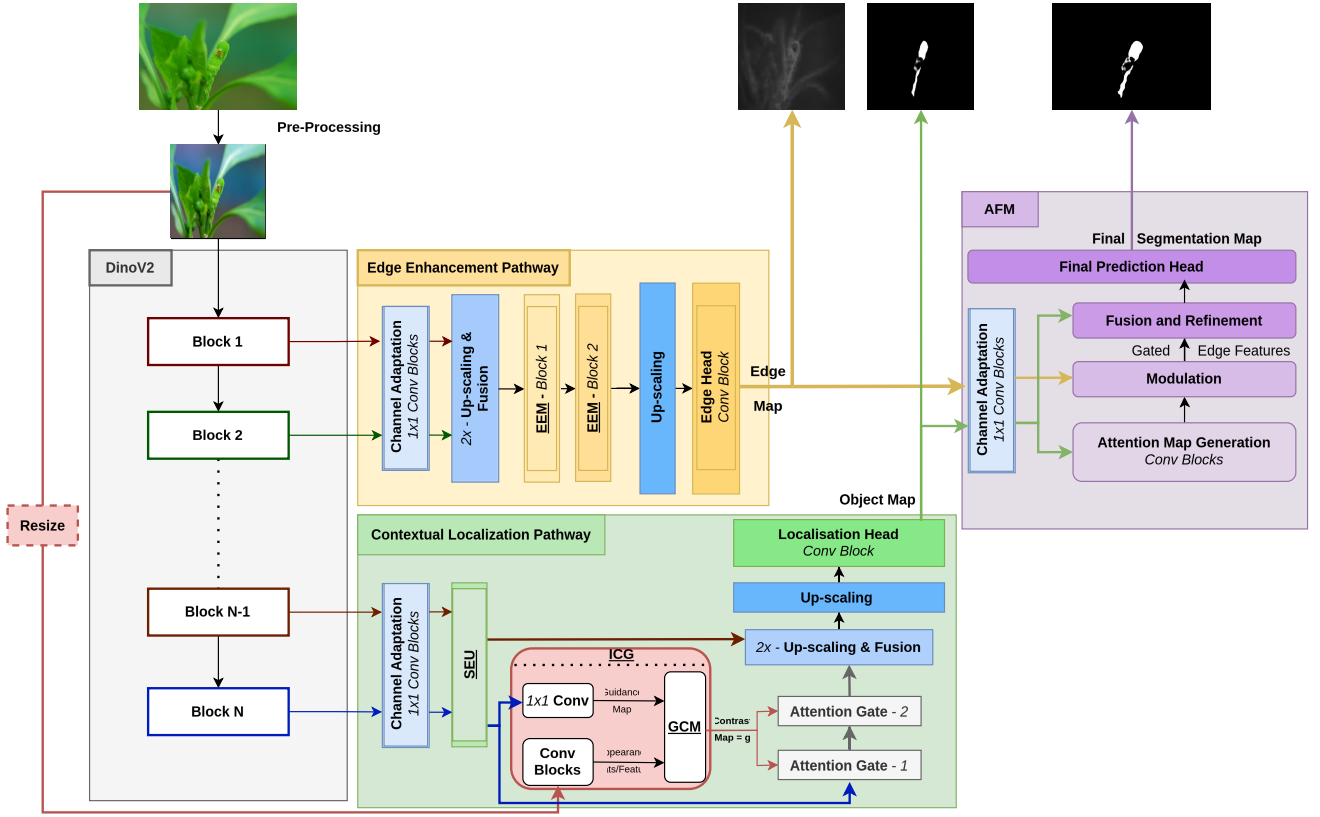


Fig. 2. Overview of the C3Net architecture. Input images are preprocessed and encoded to extract multi-scale features. Our dual-pathway decoder comprises two specialized branches. The Edge Refinement Pathway (top) processes early features through Edge Enhancement Modules (EEMs) to produce detailed edge maps. The Contextual Localization Pathway (bottom) processes deep features through Semantic Enhancement Units (SEUs) and our Image-based Context Guidance (ICG) mechanism to generate object maps with suppressed saliency. The ICG contains three components: appearance analysis from the input image, the Guided Contrast Module (GCM) for foreground-background differentiation, and iterative attention gates for saliency suppression. The Attentive Fusion Module (AFM) combines the outputs of both pathways through spatial gating to produce the final segmentation. Deep supervision is applied at three points: edge map, object map, and final prediction.

complementary outputs, and finally, multiple losses provide supervision at three prediction points.

C3Net learns a function $f : \mathbf{I} \rightarrow \mathbf{M}$ to predict a binary segmentation mask ($\mathbf{M} \in \{0, 1\}^{B \times H \times W}$) where 1 indicates camouflaged object pixels. The following subsections detail our dual-pathway architecture and its components.

A. Feature Encoding

We extract features from input images $\mathbf{I}$ using DINOv2 with register tokens [20] as the backbone encoder. DINOv2 is a vision transformer pre-trained via self-supervision without semantic labels. It learns unbiased features for detecting atypical objects, with register tokens that prevent artifacts and ensure clean, accurate upsampling for dense prediction. The encoder contains N transformer blocks that generate features at consistent spatial resolution. For the Large variant, $N = 24$ blocks produce features with dimensions $\mathbf{F}_i$.

We select four specific feature sets based on their complementary properties, which serve as inputs to our dual-pathway decoder architecture. Early features $\mathbf{F}_1$ and $\mathbf{F}_2$ from the first two blocks retain fine-grained spatial details and high-frequency information essential for edge detection. Deep features $\mathbf{F}_{n-1}$ and $\mathbf{F}_n$ from the final two blocks capture abstract semantic representations necessary for object understanding

and contextual reasoning. This selection strikes a balance between spatial precision and semantic richness.

B. Dual-Pathway Decoder

Our decoder separates edge and contextual processing into different pathways. This design arises from the observation that boundary detection and semantic understanding require fundamentally different feature-processing strategies. Unified decoders often dilute these distinct features through shared transformations. Our dual-pathway approach prevents this interference while enabling targeted learning. The Edge Refinement Pathway processes early encoder features $\mathbf{F}_1$ and $\mathbf{F}_2$, which are rich in spatial details. The Contextual Localization Pathway processes deep encoder features $\mathbf{F}_{n-1}$ and $\mathbf{F}_n$, which contain semantic information. Both pathways operate at progressively increasing resolutions through content-adaptive up-sampling. Their outputs merge through learned spatial gating rather than simple concatenation. This architectural separation enables each pathway to develop specialized representations optimized for specific tasks.

1) *Edge Refinement Pathway (ERP)*: The ERP extracts precise object boundaries from early encoder features $\mathbf{F}_1$ and $\mathbf{F}_2$. These features are first reduced in channel dimension via

1×1 convolutions, optimizing the representations for edge detection. Then, these features are progressively upsampled using DySample [21] module, which dynamically learns sampling locations based on feature content, thereby preserving fine boundary details during resolution recovery. After upsampling, these features are concatenated, followed by a convolution to produce $\mathbf{F}_{fused}^{edge}$, which captures complementary edge cues at different encoding depths. The fused representation is then processed through two cascaded Edge Enhancement Modules.

Edge Enhancement Module (EEM). is our core component designed to amplify edge-specific signals in camouflaged object detection. Each module employs a multi-path convolutional architecture that processes features across three parallel branches. The design leverages classical edge detection principles within a learnable framework. This combination provides strong inductive bias while maintaining adaptability to camouflage-specific patterns. The three paths within each EEM serve distinct purposes: i) Context Path: Captures general spatial patterns through depthwise separable convolutions, extracting contextual edge information without directional bias. ii) Gradient Path: Detects directional changes using learnable group convolutions initialized with Sobel weights. Sobel operators excel at detecting gradual intensity transitions—precisely the subtle boundaries found in camouflaged objects where sharp edges are deliberately avoided. iii) Discontinuity Path: Identifies edge points through learnable group convolutions initialized with Laplacian weights. The Laplacian’s zero-crossing detection captures texture boundaries where camouflaged patterns meet backgrounds, complementing gradient-based detection.

Each EEM processes the input features $\mathbf{F}_{edge}^{in}$ through the three paths simultaneously. Outputs from all three paths are concatenated along the channel dimension to combine their complementary edge information. This representation undergoes channel-wise recalibration using ECA [22]. The module produces output features $\mathbf{F}_{edge}^{out}$ that contain amplified edge signals. The ERP employs two EEMs in sequence to progressively enhance edge quality. The fused features $\mathbf{F}_{fused}^{edge}$ are fed into the first EEM for initial edge extraction. This module captures coarse edge structures and produces refined features. These refined features then enter the second EEM for further enhancement. The second EEM extracts fine boundary details and outputs the final edge features $\mathbf{F}_{edge}$. Additionally, these features are fed into a prediction head to generate an edge map $\mathbf{P}_{edge}$. Thus, the pathway generates two outputs: $\mathbf{F}_{edge}$ for subsequent fusion and $\mathbf{P}_{edge}$ for supervision.

2) *Contextual Localization Pathway (CLP)*: The CLP identifies semantic regions of camouflaged objects while suppressing salient distractors. This pathway receives encoder features $\mathbf{F}_{n-1}$ and $\mathbf{F}_n$ as inputs, which contain rich semantic information. Each feature is then processed through a Semantic Enhancement Unit (SEU).

Semantic Enhancement Unit (SEU). is our component that refines semantic features for improved object localization. Each unit processes its input adapted features through a series of transformations. First, depthwise separable convolutions extract initial spatial patterns $\mathbf{F}_{loc}^{conv}$ while maintaining efficiency. These then undergo enhancement through spatial and channel

attention mechanisms applied in parallel: i) Spatial Attention: Identifies discriminative regions by learning where to focus, adapting the concatenation strategy from CBAM [23]. The feature maps are aggregated using both average and max pooling along the channel dimension, capturing different spatial statistics, and are concatenated to form a comprehensive spatial descriptor. The descriptor passes through convolutional layers to generate spatial attention weights. ii) Channel Attention: Determines feature importance by recalibrating channel-wise responses [22]. It learns which channels contain the most discriminative information for camouflaged objects.

Both attention outputs are combined with the convolutional features $\mathbf{F}_{loc}^{conv}$ through element-wise multiplication. This combination enhances discriminative semantic patterns while suppressing irrelevant information. Finally, a residual connection adds the SEU’s input-adapted features to these attention-enhanced outputs. Thus, providing refined features that emphasize both spatially and channel-wise important information for contextual understanding.

As discussed above, each enhanced feature has distinct roles: i) Penultimate Layer Processing: The first SEU enhances the adapted $\mathbf{F}_{n-1}$ to produce enhanced features $\mathbf{F}_{n-1}^{enh}$, which are then upsampled using DySample [21] to produce $\mathbf{F}_{n-1}^{up}$, recovering spatial resolution while preserving localization details for fusion. ii) Final Layer Processing: The second SEU enhances the adapted $\mathbf{F}_n$ to produce $\mathbf{F}_n^{enh}$. These features are fed directly into our Image-based Context Guidance mechanism for saliency suppression and context-aware modulation. This design leverages deep features for semantic context and shallow features for spatial precision, optimizing each for effective camouflaged object detection.

Image-based Context Guidance (ICG). is for intrinsic saliency suppression and processes $\mathbf{F}_n^{enh}$ from the final encoder layer and the input image. This mechanism integrates three components to suppress salient distractors while enhancing camouflaged object cues: i) Appearance Analysis: Extracts low-level visual patterns directly from the input image. A lightweight CNN consisting of two convolutional blocks processes $\mathbf{I}$ to produce appearance features $\mathbf{A}_f$ that capture colour distributions and texture patterns independent of learned semantic representations and provide essential visual cues for contrast computation. ii) Guided Contrast Computation: Generates spatially-aware contrast maps using the appearance features $\mathbf{A}_f$. First, an initial object hypothesis $\mathbf{G}$ is generated from $\mathbf{F}_n^{enh}$. Then, the Guided Contrast Module (GCM) uses hypothesis $\mathbf{G}$ to weigh $\mathbf{A}_f$ spatially. This aggregation extracts local foreground representations using $\mathbf{G}$ as continuous weights. Meanwhile, global pooling on confident background regions produces robust background representations. Specifically, the GCM computes:

$$\mathbf{f}_{fg} = \frac{\sum_{(i,j)} \mathbf{A}_{f:,i,j} \cdot \mathbf{G}_{i,j}}{\sum_{(i,j)} \mathbf{G}_{i,j}}, \quad (1)$$

$$\mathbf{f}_{bg} = \frac{1}{|\Omega_{bg}|} \sum_{(i,j) \in \Omega_{bg}} \mathbf{A}_{f:,i,j}, \quad (2)$$

where $\Omega_{bg} = \{(i,j) | \mathbf{G}_{i,j} < 0.1\}$. The contrast map is generated by a contrast computation network that processes ap-

pearance features and extracted representations. This contrast highlights regions that subtly differ from their surroundings. ii) Iterative Attention Gating: Progressively refines features through dual attention gates. The first attention gate uses a contrast map as a gating signal to modulate $\mathbf{F}_n^{enh}$, producing intermediate features $\mathbf{F}_{loc}^{(1)}$. Subsequently, the second attention gate further refines intermediate features $\mathbf{F}_{loc}^{(1)}$ using the same contrast map, yielding final modulated features $\mathbf{F}_{loc}^{mod}$. This iterative process suppresses regions with high saliency but low contrast while enhancing regions with subtle differences. The dual-gate design ensures robust distractor suppression without losing fine camouflage cues. The ICG achieves intrinsic saliency suppression by combining appearance analysis and semantic features, enabling camouflaged object detection without external saliency models.

Furthermore, the ICG's output undergoes multi-scale fusion to produce the final contextual features. The modulated features $\mathbf{F}_{loc}^{mod}$ are first upsampled [21] to produce upsampled modulated features $\mathbf{F}_{loc}^{up}$ for accurate localization. These features are then fused with upsampled features $\mathbf{F}_{n-1}^{up}$. The fusion combines the context-aware deep features with the spatially detailed shallow features. This combination yields fused features $\mathbf{F}_{fused}^{loc}$.

The fused features undergo three processing stages to generate the final contextual features. First, a lightweight refinement module consisting of depthwise separable convolutions smooths the combined features and resolves potential conflicts between deep and shallow representations. Next, an initial upsampling recovers spatial detail for precise localization, and the subsequent upsampling produces the contextual localization features $\mathbf{F}_{loc}$. Additionally, a prediction head generates an intermediate object map by applying a 1×1 convolution followed by a sigmoid activation to produce a prediction map $\mathbf{P}_{loc}$. Thus, the CLP outputs both localization features $\mathbf{F}_{loc}$ for final fusion and prediction map $\mathbf{P}_{loc}$ for loss computation.

C. Attentive Fusion Module

The Attentive Fusion Module combines the complementary outputs from both pathways to generate the final segmentation mask. This module receives edge features $\mathbf{F}_{edge}$ from the ERP and localization features $\mathbf{F}_{loc}$ from the CLP. The fusion process leverages spatial attention to weigh each pathway's contribution according to its spatial relevance. First, both feature sets are transformed into $\mathbf{F}_{edge}^{adapt}$ and $\mathbf{F}_{loc}^{adapt}$ via channel reduction, integrating effective information while maintaining pathway-specific characteristics.

Subsequently, a spatial attention mechanism computes attention weights from the adapted localization features $\mathbf{F}_{loc}^{adapt}$ to guide selective fusion. The attention generation applies global average and max pooling operations to extract comprehensive spatial statistics. These pooled representations are concatenated and processed by convolutional layers to generate a spatial attention map, $\mathbf{A}_{spatial}$, that identifies regions requiring enhanced boundary precision.

The spatial attention map $\mathbf{A}_{spatial}$ then modulates the adapted edge features $\mathbf{F}_{edge}^{adapt}$ through element-wise multiplication to emphasize boundary information selectively. This

gating operation produces edge features $\mathbf{F}_{edge}^{att} = \mathbf{F}_{edge}^{adapt} \odot \mathbf{A}_{spatial}$ that focus on contextually relevant boundaries while suppressing spurious edge responses. Finally, the $\mathbf{F}_{edge}^{att}$ and $\mathbf{F}_{loc}^{adapt}$ are concatenated and refined through two convolutional blocks, producing a unified set of features, $\mathbf{F}_{fused}^{final}$. The final prediction head generates a segmentation mask, $\mathbf{P}_{final}$, for accurate camouflaged object detection.

D. Loss Function Strategy

The training objective combines three loss components corresponding to the edge prediction $\mathbf{P}_{edge}$, contextual prediction $\mathbf{P}_{loc}$, and final prediction $\mathbf{P}_{final}$. The total loss function is formulated as:

$$\mathcal{L}_{total} = w_{final}\mathcal{L}_{final} + w_{loc}\mathcal{L}_{loc} + w_{edge}\mathcal{L}_{edge}, \quad (3)$$

where w_{final} , w_{loc} , and w_{edge} are weighting coefficients that balance the contribution of each component during training.

The edge loss $\mathcal{L}_{edge}$ supervises the ERP to enhance boundary detection capabilities. This loss combines spatially-weighted Focal Loss [24] to address class imbalance around object boundaries and Total Variation (TV) loss to encourage smoothness in predicted background regions.

The contextual loss $\mathcal{L}_{loc}$ guides the CLP to achieve accurate object localization with saliency suppression. This loss pairs Focal Loss with Tversky Loss [25] configured to prioritize precision over recall. The precision emphasis reinforces the ICG mechanism's effectiveness in suppressing salient false positives, directly addressing the SCOD challenge through targeted optimization.

The final loss $\mathcal{L}_{final}$ optimizes the fused prediction for comprehensive object detection performance. This component combines Focal and Tversky losses configured to emphasize recall, thereby promoting complete object capture. Additionally, a weighted Dice Loss term, calculated specifically on ground-truth edge pixels, provides final refinement signals for boundary sharpness. This multi-metric approach ensures both accurate localization and precise boundary delineation.

Furthermore, instance-level weighting based on object size is applied to $\mathcal{L}_{loc}$ and $\mathcal{L}_{final}$ to address scale variation. Objects with smaller foreground ratios receive higher weights to prevent bias toward larger objects, thus improving detection performance.

IV. EXPERIMENTS AND EVALUATION

A. Experimental Setup

This section details the experimental configuration used to evaluate C3Net. We first describe the benchmark datasets employed for training and testing. We then explain the evaluation metrics used to assess performance. Finally, we present the implementation details, including training procedures and hyperparameter settings.

a) *Datasets.*: We evaluate our C3Net on three established COD benchmarks widely adopted in the literature. All datasets include pixel-level ground truth annotations for precise evaluation. Additionally, we generate edge maps using the Canny detector on ground truth masks with a 5-pixel width. These

TABLE I

QUANTITATIVE COMPARISON WITH SOTA METHODS ON BENCHMARK DATASETS. NOTES: $\uparrow/\downarrow$ DENOTES THAT LARGER/SMALLER IS BETTER. THE BEST VALUES ARE IN **RED**, AND THE SECOND BEST ARE UNDERLINED IN BLUE. LIGHT-GRAY ROWS ARE CNN-BASED METHODS; WHITE ROWS ARE TRANSFORMER-BASED METHODS. ALL RESULTS ARE OBTAINED FROM AUTHOR-PROVIDED PREDICTIONS FOR FAIR COMPARISON.

Methods	CAMO (250)					COD10K (2,026)					NC4K (4,121)				
	$S_\alpha \uparrow$	$F_\beta^w \uparrow$	$F_\beta^m \uparrow$	$E_\phi \uparrow$	$\mathcal{M} \downarrow$	$S_\alpha \uparrow$	$F_\beta^w \uparrow$	$F_\beta^m \uparrow$	$E_\phi \uparrow$	$\mathcal{M} \downarrow$	$S_\alpha \uparrow$	$F_\beta^w \uparrow$	$F_\beta^m \uparrow$	$E_\phi \uparrow$	$\mathcal{M} \downarrow$
SINet ₂₀ [9]	.751	.606	.675	.831	.100	.771	.551	.634	.868	.051	.808	.723	.769	.883	.058
SLSR ₂₁ [26]	.787	.674	.744	.854	.080	.804	.673	.715	.892	.037	.840	.766	.804	.907	.048
PFNet ₂₁ [14]	.782	.695	.746	.855	.085	.800	.660	.701	.890	.040	.829	.745	.784	.898	.053
MGL ₂₁ [27]	.775	.673	.726	.842	.088	.814	.666	.711	.890	.035	.833	.740	.782	.893	.052
UJSC ₂₁ [13]	.800	.728	.772	.873	.073	.809	.684	.721	.891	.035	.842	.771	.806	.907	.047
C ² FNet ₂₁ [28]	.796	.719	.762	.864	.080	.813	.686	.723	.900	.036	.838	.762	.795	.904	.049
UGTR ₂₁ [12]	.784	.684	.735	.851	.086	.817	.666	.712	.890	.036	.839	.747	.787	.899	.052
PreyNet ₂₂ [29]	.790	.708	.757	.857	.077	.813	.697	.736	.891	.034	-	-	-	-	-
BSA-Net ₂₂ [30]	.794	.717	.763	.867	.079	.818	.699	.738	.901	.034	.841	.771	.808	.907	.048
OCE-Net ₂₂ [31]	.802	.723	.766	.865	.080	.827	.707	.741	.905	.033	.853	.785	.818	.913	.045
BGNet ₂₂ [16]	.812	.749	.789	.882	.073	.831	.722	.753	.911	.033	.851	.788	.820	.916	.044
SegMaR ₂₂ [11]	.815	.795	.794	.884	.071	.833	.724	.757	.906	.034	.841	.781	.820	.907	.046
ZoomNet ₂₂ [10]	.820	.752	.794	.892	.066	.830	.729	.766	.911	.029	.853	.784	.818	.912	.043
SINet-v2 ₂₂ [1]	.820	.743	.782	.895	.070	.815	.680	.718	.906	.037	.847	.770	.805	.914	.048
FDNet ₂₂ [32]	.828	.748	.781	.883	.068	.832	.706	.733	.907	.033	.834	.750	.784	.893	.051
DTINet ₂₂ [33]	.856	.796	-	.916	.050	.824	.695	-	.896	.034	.863	.792	-	.917	.041
OSFormer ₂₂ [34]	.799	-	-	.858	.073	.811	-	-	.881	.034	.832	-	-	.905	.049
FSPNet ₂₃ [8]	.856	.799	.830	.928	.050	.851	.735	.769	.930	.026	.879	.816	.843	.937	.035
TPRNet ₂₃ [35]	.814	.781	-	-	.076	.829	.725	-	-	.034	.854	.790	-	-	.047
FPNet ₂₃ [36]	.852	.806	-	.905	.056	.850	.748	-	.913	.029	-	-	-	-	-
OPNet ₂₃ [37]	.858	.817	-	.915	.050	.857	.767	-	.919	.026	.883	.838	-	.932	.034
HitNet ₂₃ [38]	.844	.801	-	.902	.057	.868	.798	-	.932	.024	.870	.825	-	.921	.039
SAM-Auto ₂₃ [39], [40]	.684	.606	.680	.687	.132	.783	.701	.756	.798	.050	.767	.696	.752	.776	.078
SAM-Prompt ₂₃ [39], [40]	.647	.520	-	-	.141	.696	.552	-	-	.094	.699	.591	-	-	.115
FEDER ₂₃ [7]	.807	.785	<u>.873</u>	<u>.947</u>	.069	.823	.740	.900	.911	.032	.846	.817	.905	.916	.045
SAM2-Auto ₂₄ [2], [3]	.444	.184	.207	.401	.236	.549	.271	.291	.521	.134	.512	.251	.268	.482	.186
SAM2-Prompt ₂₄ [2], [3]	.722	.633	-	-	.114	.754	.640	-	-	.078	.776	.700	-	-	.085
FocusDiffuser ₂₅ [41]	.881	<u>.851</u>	-	.939	.042	<u>.875</u>	<u>.809</u>	-	<u>.939</u>	<u>.020</u>	.891	<u>.854</u>	-	.940	<u>.029</u>
FSEL ₂₅ [42]	<u>.885</u>	<u>.851</u>	.864	.942	<u>.040</u>	.873	.800	.796	.928	.021	<u>.892</u>	.853	.864	<u>.941</u>	.030
C3Net (Ours)	.904	.889	.896	.951	.0311	.898	.851	<u>.859</u>	.961	.0162	.913	.895	<u>.903</u>	.958	.0220

edge maps serve as supervision signals for our ERP during training. The first benchmark is *COD10K* [9], which contains 10,000 images, making it the largest COD dataset. Among these images, 5,066 depict camouflaged instances spanning 10 super-classes and 78 sub-classes. These classes cover aquatic, terrestrial, flying, and amphibious categories. We follow the standard split using 3,040 images for training and 2,026 for testing. The second benchmark is *CAMO* [43], which provides 1,250 images featuring both natural and artificial camouflage. This dataset complements COD10K by including eight categories. We use the standard protocol of 1,000 images for training and 250 for testing. The third benchmark is *NC4K* [26], which serves as a test-only evaluation with 4,121 images. This dataset pushes detection limits by challenging natural camouflage in the presence of extremely similar backgrounds. It specifically tests cross-dataset generalization without fine-tuning. Following established practice, we train C3Net on the training sets of COD10K and CAMO to ensure a fair comparison. We then evaluate on all three benchmarks to assess both in-domain and cross-domain performance. NC4K serves as a held-out test set to measure cross-dataset generalization.

b) Evaluation Metrics.: We evaluate using five standard COD metrics. *Structure Measure* (S_α) [44] assesses structural similarity with $\alpha = 0.5$. *Enhanced Alignment Measure* (E_ϕ) [45] combines pixel-level and global statistics. *Weighted F-Measure* (F_β^w) [46] and *Mean F-Measure* (F_β^m) [47] evaluate precision-recall with $\beta^2 = 0.3$. *Mean Absolute Error* ($\mathcal{M}$) measures pixel-wise differences where lower is better. These metrics comprehensively evaluate structural accuracy

and boundary precision.

c) Implementation Details.: C3Net uses the PyTorch framework on NVIDIA H100 GPUs. Images are resized to 392×392 pixels and normalized with $\mu = [0.485, 0.456, 0.406]$ and $\sigma = [0.229, 0.224, 0.225]$. We use the pretrained DINOv2-Large model, extracting features from blocks [1, 2, 23, 24]. Decoder channels are [512, 256, 128] with an output channel of 128. The Edge pathway utilizes two enhancement blocks, whereas the contextual pathway employs one. ICG uses 32 appearance channels and 16 contrast channels. DySample uses the 'lp' style with four groups. The fusion head uses 128 channels with two refinement blocks. Loss weights are $w_{edge} = 1.0$, $w_{loc} = 1.15$, and $w_{final} = 1.2$. Focal loss has the $\alpha = 0.25$ and $\gamma = 3.0$. Tversky for $\mathcal{L}_{loc}$ uses $\alpha = 0.6$ and $\beta = 0.4$ while $\mathcal{L}_{final}$ uses $\alpha = 0.4$ and $\beta = 0.6$. Edge loss employs a focal weight of 5.0 and a TV weight of 0.15, with a cutoff of 5. Final loss utilizes edge Dice weight 0.2. Instance weighting uses a factor of 3.0 with thresholds of 0.02 and 0.8. Training uses AdamW for 200 epochs with learning rates of 1×10^{-4} for the decoder and 2×10^{-5} for the encoder. Weight decay is 0.01, and gradient clipping is 1.0. We employ automatic mixed precision for efficiency. ReduceLROnPlateau uses a factor of 0.5, patience of 10, and a minimum of 1×10^{-6} . Batch size is 128. Inference uses identical preprocessing.

B. Comparison with State-of-the-Art

We benchmark C3Net against a comprehensive suite of 29 recent COD methods, including prominent CNN-based mod-

els, transformer-based approaches, and foundation models.

a) *Quantitative Results.*: Table I demonstrates C3Net’s state-of-the-art performance across benchmarks. On the primary COD10K benchmark, we achieve leading results with S_α of 0.898, F_β^w of 0.851, F_β^m of 0.859, and E_ϕ of 0.961. Our model outperforms the next-best method by 2.6% in structure measure (S_α) and 2.3% in enhanced alignment measure (E_ϕ), confirming that our dual-pathway design effectively captures both global coherence and local precision. The MAE of 0.0162 demonstrates precise pixel-level accuracy. This performance extends to CAMO, where we achieve S_α of 0.904, F_β^w of 0.889, and the leading $\mathcal{M}$ of 0.0311, improving over the next-best by 2.1% in structure measure. On NC4K, we maintain leadership with S_α of 0.913 and E_ϕ of 0.958. Across all three benchmarks, C3Net achieves the best performance on 13 of 15 metrics, validating our comprehensive challenge coverage rather than dataset-specific tuning. Foundation models confirm the necessity of specialized architecture, with SAM achieving only $S_\alpha = 0.783$ and SAM2 catastrophically failing at $S_\alpha = 0.549$. The performance gap between C3Net and SAM2 exceeds 63% for the S_α metric, quantifying why general segmentation fails on camouflaged objects. These results substantiate our core claim that architectural nitegration, achieved through specialized pathways and intrinsic mechanisms, surpasses the capabilities of isolated improvements.

b) *Qualitative Analysis.*: Figure 3 presents visual comparisons that validate our architectural design through consistent superior performance. In IS scenarios (Figure 3, row i) where objects share identical textures with backgrounds, C3Net accurately segments the camouflaged target. This success stems from our dual-pathway architecture, which extracts complementary features at multiple scales. ED cases (Figure 3, row ii) reveal fragmented boundaries in competing methods while C3Net produces complete masks. Our ERP with gradient-initialized EEMs specifically recovers these disrupted boundaries through specialized edge processing. CD examples (Figure 3, row iii) and multiple instance scenarios (Figure 3, row iv) demonstrate accurate segmentation where others fail or merge objects. The CLP enables this through deep semantic understanding combined with spatial precision. EC (Figure 3, row v) shows C3Net maintaining segmentation integrity under shadows and occlusions while competitors produce fragmented results. ESV is handled robustly with both small objects (Figure 3, row vi) and large objects (Figure 3, row vii) accurately segmented. Content-adaptive upsampling preserves fine details across these scales. Most significantly, the SCOD challenge (Figure 3, row viii) exposes fundamental limitations in existing methods. While OCENet, BGNet, ZoomNet, and FSPNet all erroneously segment the salient orange coral, C3Net correctly identifies only the camouflaged ray. This validates the ICG mechanism’s ability to suppress salient distractors through intrinsic context-contrast analysis, without relying on external saliency models. Across all challenges, C3Net consistently produces masks closest to ground truth while competing methods exhibit complete failures or severe artifacts.

TABLE II

ABLATION STUDY VALIDATING ARCHITECTURAL COMPONENTS AND IMPLEMENTATION CHOICES ACROSS ALL BENCHMARKS. NOTES: $\uparrow$ DENOTES THAT HIGHER VALUES ARE BETTER, WHILE $\downarrow$ DENOTES THAT LOWER VALUES ARE BETTER. WE EVALUATE ALTERNATIVE IMPLEMENTATIONS (ROWS 2-3) BY REPLACING KEY COMPONENTS WITH STANDARD ALTERNATIVES. WE THEN REMOVE ESSENTIAL COMPONENTS (ROWS 4-7) TO ANALYZE THEIR INDIVIDUAL CONTRIBUTIONS. BOLD VALUES INDICATE FULL C3NET BASELINE PERFORMANCE.

Configuration	CAMO (250)			COD10K (2,026)			NC4K (4,121)		
	$S_\alpha \uparrow$	$F_\beta^w \uparrow$	$\mathcal{M} \downarrow$	$S_\alpha \uparrow$	$F_\beta^w \uparrow$	$\mathcal{M} \downarrow$	$S_\alpha \uparrow$	$F_\beta^w \uparrow$	$\mathcal{M} \downarrow$
C3Net (Full)	0.904	0.889	0.0311	0.898	0.851	0.0162	0.913	0.895	0.0220
w/ ViT-L (no registers)	0.875	0.843	0.0412	0.871	0.806	0.0228	0.892	0.856	0.0278
w/ Bilinear upsampling	0.889	0.871	0.0337	0.880	0.832	0.0182	0.902	0.878	0.0239
w/o ERP	0.881	0.862	0.0341	0.872	0.825	0.0215	0.897	0.871	0.0245
w/o CLP	0.631	0.538	0.0847	0.608	0.520	0.0521	0.664	0.558	0.0598
w/o ICG mechanism	0.883	0.870	0.0362	0.875	0.829	0.0187	0.900	0.876	0.0257
w/ Random EEM init	0.893	0.876	0.0329	0.885	0.838	0.0186	0.902	0.880	0.0238

C. Ablation Studies

We validate C3Net’s architectural design through systematic ablation studies across all benchmarks. Table II investigates implementation alternatives and component contributions.

a) *Implementation Choices.*: We first examine encoder and upsampling alternatives. Replacing DINOv2 with standard ViT-Large decreases F_β^w by 5.3% on COD10K, 5.2% on CAMO, and 4.4% on NC4K. Similarly, replacing DySample with bilinear upsampling causes F_β^w reductions of 2.2% on COD10K, 2.0% on CAMO, and 1.9% on NC4K. These moderate degradations show that such choices optimize performance but do not drive core detection capability.

b) *Dual-Pathway Architecture.*: Next, we analyze the architectural pathways. Removing either pathway highlights their unique contributions and asymmetric significance. Without ERP, F_β^w decreases by 3.1% on COD10K, 3.0% on CAMO, and 2.7% on NC4K. These moderate losses confirm that ERP adds valuable boundary refinement. In sharp contrast, removing CLP causes a severe failure. On COD10K, S_α drops to 0.608 and F_β^w falls to 0.520, a 38.9% decrease in weighted F-measure. CAMO and NC4K experience similar drops, with F_β^w values of 0.538 and 0.558. The MAE rises by 221.6% on COD10K, 172.3% on CAMO, and 171.8% on NC4K. CLP has two parts that jointly enhance detection abilities. The SEU blocks provide semantic insight through deep feature processing, while ICG suppresses saliency to counter SCOD. Removing the entire CLP eliminates both functions simultaneously, resulting in a 39% collapse. This is a 13-fold difference compared to removing ERP. This shows that semantic understanding is far more critical than boundary refinement alone. These results support our dual-pathway design with its clear architectural hierarchy.

c) *Saliency Suppression.*: Within CLP, we isolate ICG’s specific contribution to saliency handling. The ICG mechanism addresses SCOD by suppressing intrinsic saliency without the need for external models. Without ICG, F_β^w decreases by 2.6% on COD10K, 2.1% on CAMO, and 2.1% on NC4K. The MAE increases substantially by 15.4% on COD10K, 16.4% on CAMO, and 16.8% on NC4K. The MAE increases are particularly significant because SCOD failures manifest as false positives from salient distractors. These impacts demonstrate that ICG provides a crucial capability for handling saliency. This complements the semantic understanding from

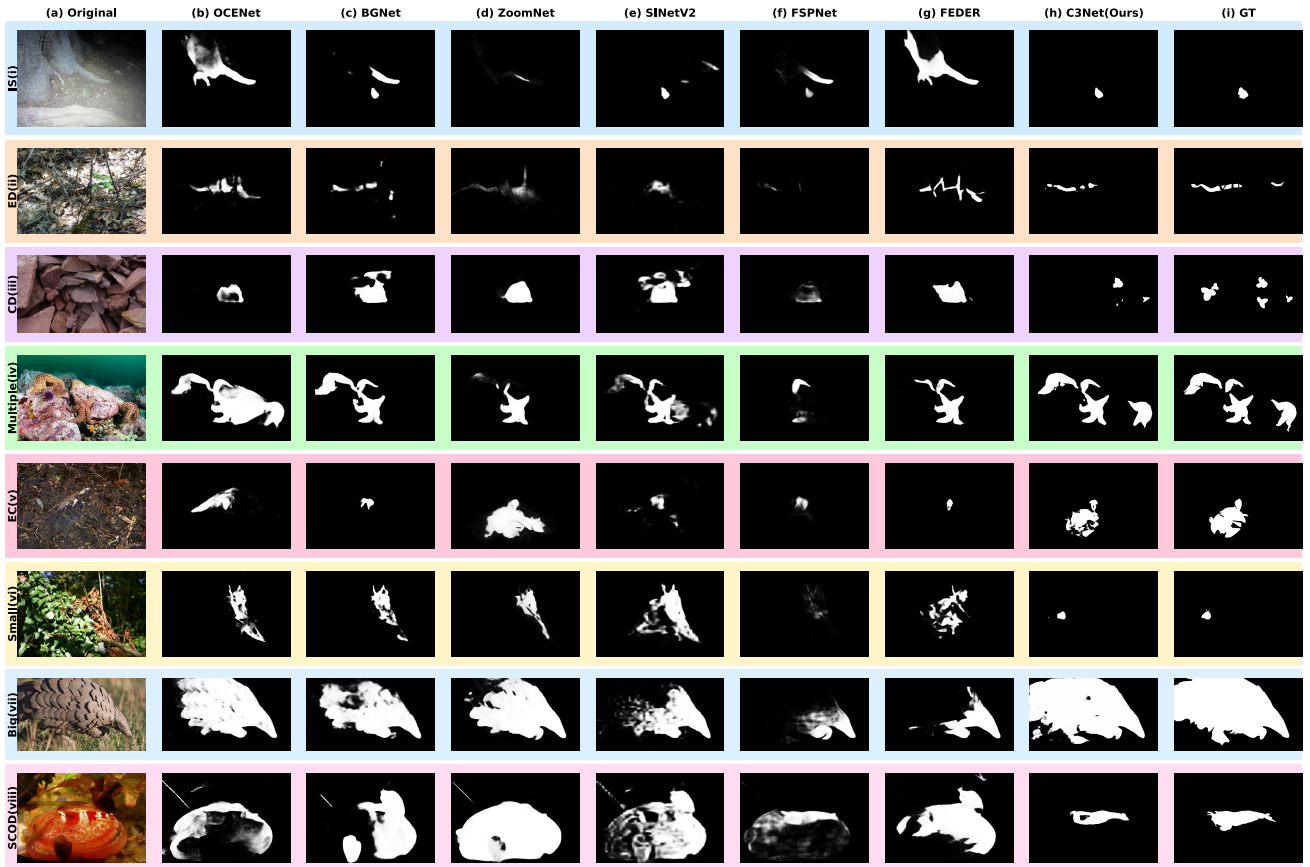


Fig. 3. Visual comparison of C3Net with state-of-the-art methods on challenging COD cases. Each row exemplifies a specific challenge: (i) Intrinsic Similarity (IS), (ii) Edge Disruption (ED), (iii) Contextual Dependencies (CD), (iv) Multiple Instances, (v) Environmental Complexities (EC), (vi) Small Objects (ESV aspect), (vii) Large Objects (ESV aspect), and (viii) Salient-Camouflaged Object Disambiguation (SCOD). For each row, columns are: (a) Input Image, (b) OCENet [31], (c) BGNet [16], (d) ZoomNet [10], (e) SINetV2 [1], (f) FSPNet [8], (g) FEDER [7], (h) C3Net (Ours), and (i) Ground Truth. Visual results for the compared methods are obtained from officially released predictions.

SEU blocks to achieve comprehensive detection performance.

d) Edge Initialization.: Eventually, we investigate the initialization strategy for the ERP components. Our gradient-based initialization strategy consistently improves performance. Random EEM initialization causes F_{β}^w drops of 1.5% on COD10K, 1.4% on CAMO, and 1.6% on NC4K. These results show that classical edge detection priors provide useful inductive biases for boundary detection.

These results establish clear architectural priorities. CLP provides foundational semantic understanding through its SEU blocks. ICG contributes crucial saliency suppression to address SCOD challenges. ERP provides valuable boundary refinement. Together, these architectural components enable core detection. In contrast, implementation choices such as encoder selection and upsampling method offer meaningful optimizations but are not essential for detection. Similarly, initialization strategy offers minor improvements through inductive biases. The 13-fold difference between CLP and ERP removal confirms that architectural design choices matter far more than implementation alternatives.

V. CONCLUSION

We presented C3Net for comprehensive camouflaged object detection, addressing six fundamental challenges. Our dual-pathway architecture achieves state-of-the-art performance

through specialized processing. The ERP with gradient-initialized EEMs captures precise boundaries, while the CLP with ICG suppresses salient distractors intrinsically without external models. Our architecture outperforms previous state-of-the-art methods while maintaining efficient processing. The key innovations—dual-pathway separation, intrinsic saliency suppression, gradient-initialized edge detection, and attentive fusion—enable critical applications in medical imaging and wildlife monitoring. Our C3Net demonstrates that complex vision challenges require specialized components working synergistically rather than isolated improvements.

REFERENCES

- [1] D.-P. Fan, G.-P. Ji, M.-M. Cheng, and L. Shao, “Concealed object detection,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 10, pp. 6024–6042, 2022.
- [2] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson *et al.*, “Sam 2: Segment anything in images and videos,” in *International Conference on Learning Representations*.
- [3] L. Tang and B. Li, “Evaluating sam2’s role in camouflaged object detection: From sam to sam2,” 2024.
- [4] D.-P. Fan, G.-P. Ji, T. Zhou, G. Chen, H. Fu, J. Shen, and L. Shao, “Pranet: Parallel reverse attention network for polyp segmentation,” in *MICCAI*, A. L. Martel, P. Abolmaesumi, D. Stoyanov, D. Mateus, M. A. Zuluaga, S. K. Zhou, D. Racoceanu, and L. Joskowicz, Eds., Springer. Cham: Springer International Publishing, October 2020, pp. 263–273.

- [5] R. Pérez-de la Fuente, X. Delclòs, E. Peñalver, M. Speranza, J. Wierzbos, C. Ascaso, and M. S. Engel, "Early evolution and ecology of camouflage in insects," *National Academy of Sciences of the United States of America*, vol. 109, no. 52, pp. 21 414–21 419, 12 2012.
- [6] N. U. Bhajantri and P. Nagabhushan, "Camouflage defect identification: a novel approach," in *9th International Conference on Information Technology (ICIT'06)*. IEEE, 2006, pp. 145–148.
- [7] C. He, K. Li, Y. Zhang, L. Tang, Y. Zhang, Z. Guo, and X. Li, "Camouflaged object detection with feature decomposition and edge reconstruction," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2023, pp. 22 046–22 055.
- [8] Z. Huang, H. Dai, T.-Z. Xiang, S. Wang, H.-X. Chen, J. Qin, and H. Xiong, "Feature shrinkage pyramid for camouflaged object detection with transformers," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, June 2023, pp. 5557–5566.
- [9] D.-P. Fan, G.-P. Ji, G. Sun, M.-M. Cheng, J. Shen, and L. Shao, "Camouflaged object detection," in *CVPR*, June 2020, pp. 2777–2787.
- [10] Y. Pang, X. Zhao, T.-Z. Xiang, L. Zhang, and H. Lu, "Zoom in and out: A mixed-scale triplet network for camouflaged object detection," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2022, pp. 2160–2170.
- [11] Q. Jia, S. Yao, Y. Liu, X. Fan, R. Liu, and Z. Luo, "Segment, magnify and reiterate: Detecting camouflaged objects the hard way," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2022, pp. 4713–4722.
- [12] F. Yang, Q. Zhai, X. Li, R. Huang, A. Luo, H. Cheng, and D.-P. Fan, "Uncertainty-guided transformer reasoning for camouflaged object detection," in *IEEE/CVF International Conference on Computer Vision*, October 2021, pp. 4146–4155.
- [13] A. Li, J. Zhang, Y. Lv, B. Liu, T. Zhang, and Y. Dai, "Uncertainty-aware joint salient object and camouflaged object detection," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2021, pp. 10071–10081.
- [14] H. Mei, G.-P. Ji, Z. Wei, X. Yang, X. Wei, and D.-P. Fan, "Camouflaged object segmentation with distraction mining," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2021, pp. 8772–8781.
- [15] B. Yin, X. Zhang, D.-P. Fan, S. Jiao, M.-M. Cheng, L. Van Gool, and Q. Hou, "Camoformer: Masked separable attention for camouflaged object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.
- [16] Y. Sun, S. Wang, C. Chen, and T.-Z. Xiang, "Boundary-guided camouflaged object detection," in *Thirty-First International Joint Conference on Artificial Intelligence (IJCAI)*, July 2022, pp. 1335–1341.
- [17] W. Zhao, S. Xie, F. Zhao, Y. He, and H. Lu, "Nowhere to disguise: Spot camouflaged objects via saliency attribute transfer," *IEEE Transactions on Image Processing*, vol. 32, pp. 3108–3120, 2023.
- [18] Z. Zhou, Y. Li, C. Zhong, J. Huang, J. Pei, and H. Tang, "Unconstrained salient and camouflaged object detection," *arXiv preprint arXiv:2412.10943*, 2024.
- [19] J. He, B. Liu, and H. Chen, "Hdpnet: Hourglass vision transformer with dual-path feature pyramid for camouflaged object detection," in *2025 IEEE/CVF Winter Conference on Applications of Computer Vision*. IEEE, February 2025, pp. 8638–8647.
- [20] T. Darcet, M. Oquab, J. Mairal, and P. Bojanowski, "Vision transformers need registers," 2024.
- [21] W. Liu, H. Lu, H. Fu, and Z. Cao, "Learning to upsample by learning to sample," in *IEEE/CVF International Conference on Computer Vision*, October 2023, pp. 6027–6037.
- [22] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "Eca-net: Efficient channel attention for deep convolutional neural networks," in *IEEE/CVF conference on computer vision and pattern recognition*, June 2020, pp. 11 534–11 542.
- [23] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *European Conference on Computer Vision*, September 2018, pp. 3–19.
- [24] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *IEEE international conference on computer vision*, October 2017, pp. 2980–2988.
- [25] S. S. M. Salehi, D. Erdogmus, and A. Gholipour, "Tversky loss function for image segmentation using 3d fully convolutional deep networks," in *International workshop on machine learning in medical imaging*. Springer, September 2017, pp. 379–387.
- [26] Y. Lv, J. Zhang, Y. Dai, A. Li, B. Liu, N. Barnes, and D.-P. Fan, "Simultaneously localize, segment and rank the camouflaged objects," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2021, pp. 11 591–11 601.
- [27] Q. Zhai, X. Li, F. Yang, C. Chen, H. Cheng, and D.-P. Fan, "Mutual graph learning for camouflaged object detection," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2021, pp. 12 997–13 007.
- [28] Y. Sun, G. Chen, T. Zhou, Y. Zhang, and N. Liu, "Context-aware cross-level fusion network for camouflaged object detection," in *IJCAI*, July 2021, pp. 1025–1031.
- [29] M. Zhang, S. Xu, Y. Piao, D. Shi, S. Lin, and H. Lu, "Preynet: Preying on camouflaged objects," in *30th ACM International Conference on Multimedia*, ser. MM '22. New York, NY, USA: Association for Computing Machinery, October 2022, pp. 5323–5332.
- [30] H. Zhu, P. Li, H. Xie, X. Yan, D. Liang, D. Chen, M. Wei, and J. Qin, "I can find you! boundary-guided separated attention network for camouflaged object detection," in *AAAI Conference on Artificial Intelligence*, vol. 36, no. 3, June 2022, pp. 3608–3616.
- [31] J. Liu, J. Zhang, and N. Barnes, "Modeling aleatoric uncertainty for camouflaged object detection," in *2022 IEEE/CVF Winter Conference on Applications of Computer Vision*, January 2022, pp. 1445–1454.
- [32] Y. Zhong, B. Li, L. Tang, S. Kuang, S. Wu, and S. Ding, "Detecting camouflaged object in frequency domain," in *IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 4504–4513.
- [33] Z. Liu, Z. Zhang, Y. Tan, and W. Wu, "Boosting camouflaged object detection with dual-task interactive transformer," in *2022 26th International Conference on Pattern Recognition (ICPR)*. IEEE, August 2022, pp. 140–146.
- [34] J. Pei, T. Cheng, D.-P. Fan, H. Tang, C. Chen, and L. Van Gool, "Osformer: One-stage camouflaged instance segmentation with transformers," in *European conference on computer vision*. Springer, October 2022, pp. 19–37.
- [35] Q. Zhang, Y. Ge, C. Zhang, and H. Bi, "Tprnet: camouflaged object detection via transformer-induced progressive refinement network," *The Visual Computer*, vol. 39, no. 10, pp. 4593–4607, 2023.
- [36] R. Cong, M. Sun, S. Zhang, X. Zhou, W. Zhang, and Y. Zhao, "Frequency perception network for camouflaged object detection," in *31st ACM International Conference on Multimedia*, 2023, pp. 1179–1189.
- [37] H. Mei, K. Xu, Y. Zhou, Y. Wang, H. Piao, X. Wei, and X. Yang, "Camouflaged object segmentation with omni perception," *International Journal of Computer Vision*, vol. 131, no. 11, pp. 3019–3034, 2023.
- [38] X. Hu, S. Wang, X. Qin, H. Dai, W. Ren, D. Luo, Y. Tai, and L. Shao, "High-resolution iterative feedback network for camouflaged object detection," in *AAAI Conference on Artificial Intelligence*, vol. 37, no. 1, 2023, pp. 881–889.
- [39] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollar, and R. Girshick, "Segment anything," in *IEEE/CVF International Conference on Computer Vision*, October 2023, pp. 4015–4026.
- [40] L. Tang, H. Xiao, and B. Li, "Can sam segment anything? when sam meets camouflaged object detection," 2023.
- [41] J. Zhao, X. Li, F. Yang, Q. Zhai, A. Luo, Z. Jiao, and H. Cheng, "Focus-diffuser: Perceiving local disparities for camouflaged object detection," in *European Conference on Computer Vision*. Springer, 2025, pp. 181–198.
- [42] Y. Sun, C. Xu, J. Yang, H. Xuan, and L. Luo, "Frequency-spatial entanglement learning for camouflaged object detection," in *European Conference on Computer Vision*. Springer, 2025, pp. 343–360.
- [43] T.-N. Le, T. V. Nguyen, Z. Nie, M.-T. Tran, and A. Sugimoto, "Anabran network for camouflaged object segmentation," *Computer Vision and Image Understanding*, vol. 184, pp. 45–56, 2019.
- [44] M.-M. Cheng and D.-P. Fan, "Structure-measure: A new way to evaluate foreground maps," *International Journal of Computer Vision (IJCV)*, vol. 129, no. 9, pp. 2622–2638, 2021.
- [45] D.-P. Fan, C. Gong, Y. Cao, B. Ren, M.-M. Cheng, and A. Borji, "Enhanced-alignment measure for binary foreground map evaluation," in *Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*. International Joint Conferences on Artificial Intelligence Organization, July 2018, pp. 698–704.
- [46] R. Margolin, L. Zelnik-Manor, and A. Tal, "How to evaluate foreground maps?" in *IEEE Conference on Computer Vision and Pattern Recognition*, June 2014, pp. 248–255.
- [47] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, June 2009, pp. 1597–1604.