ESCNet: Edge-Semantic Collaborative Network for Camouflaged Object Detection

CONTENTS

01

Background & Motivation

What Is Camouflaged Object Detection?

Camouflaged Object Detection (COD) seeks to segment objects that adopt visual patterns nearly identical to their surroundings. It's a cornerstone problem for safety, health, and biodiversity, yet remains unsolved because appearance cues alone are unreliable.

Medical Imaging

Early tumour localisation.

Wildlife Monitoring

Tracking species in complex habitats.

Limitations of Existing Approaches

Fixed Geometric Priors

Standard convolutions embed fixed kernels that cannot flex to the fractal, stochastic boundaries of hidden animals.

Separate Feature Processing

Edge and texture streams are optimised separately, so the two informative cues never correct each other, leading to fragmented masks.

Lack of Calibration

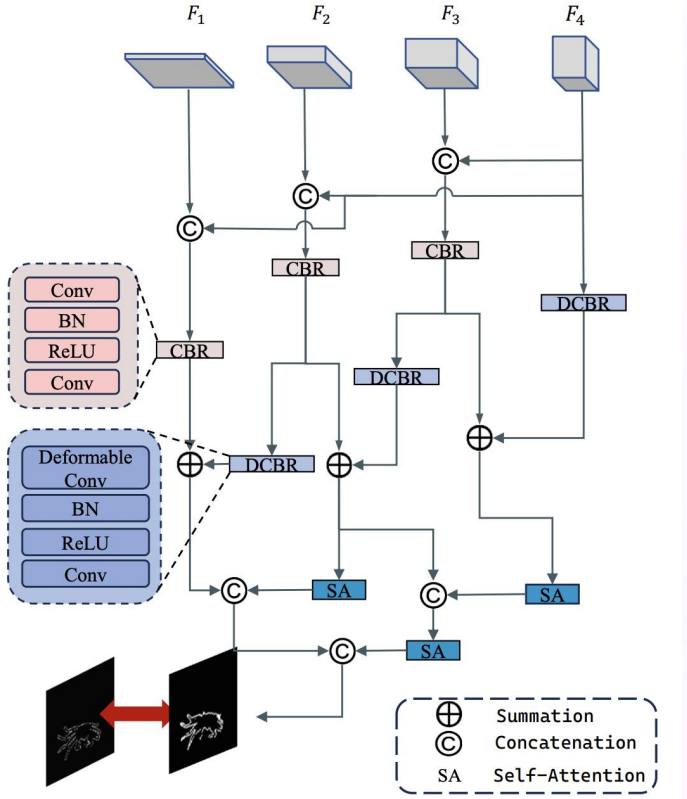
Multi-scale predictions are fused without dedicated calibration, so high-level semantics dilute fine edge evidence.

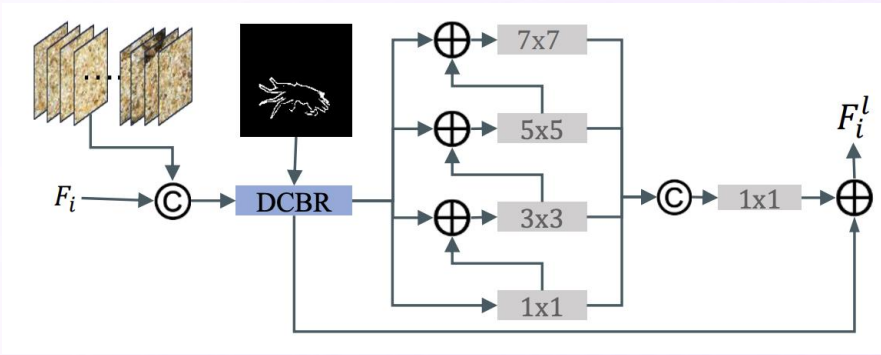
02

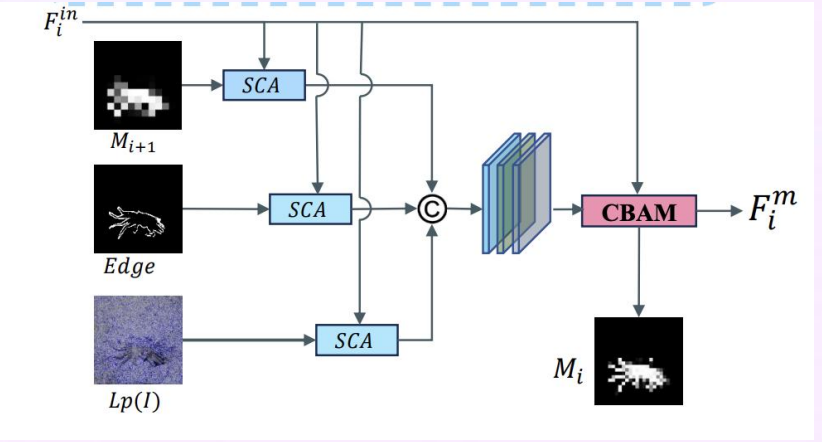
Contributions & Pipeline

Core Contributions of ESCNet

We introduce ESCNet, the first framework that couples edge and texture perception into a single feedback loop, establishing a new state of the art.







AETP

Builds an initial boundary hypothesis while receiving texture consistency signals.

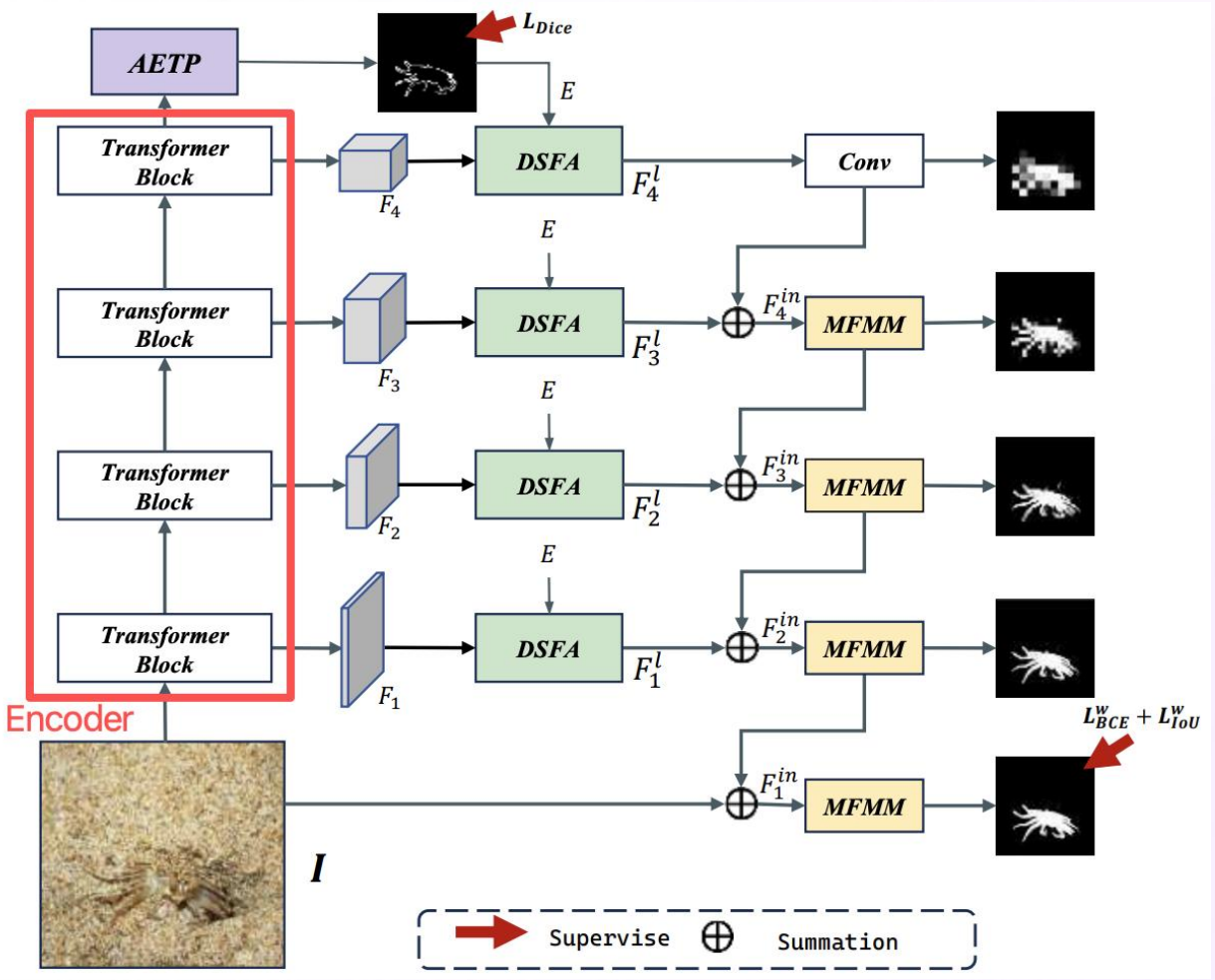
DSFA

Performs deformable sampling conditioned on edge orientation and texture complexity.

MFMM

Refines masks through cascaded edge-guided attention and hierarchical texture fusion.

Key Innovation: A self-reinforcing feedback loop where edge evidence rectifies texture discrimination and vice versa.



Architecture Overview

Encoder (PVT-v2)

Extracts multi-scale features capturing global semantics and fine, deceptive patterns.

Decoder (ESCNet)

A cascade of AETP, DSFA, and MFMM blocks that progressively refine segmentation.

Skip Connections

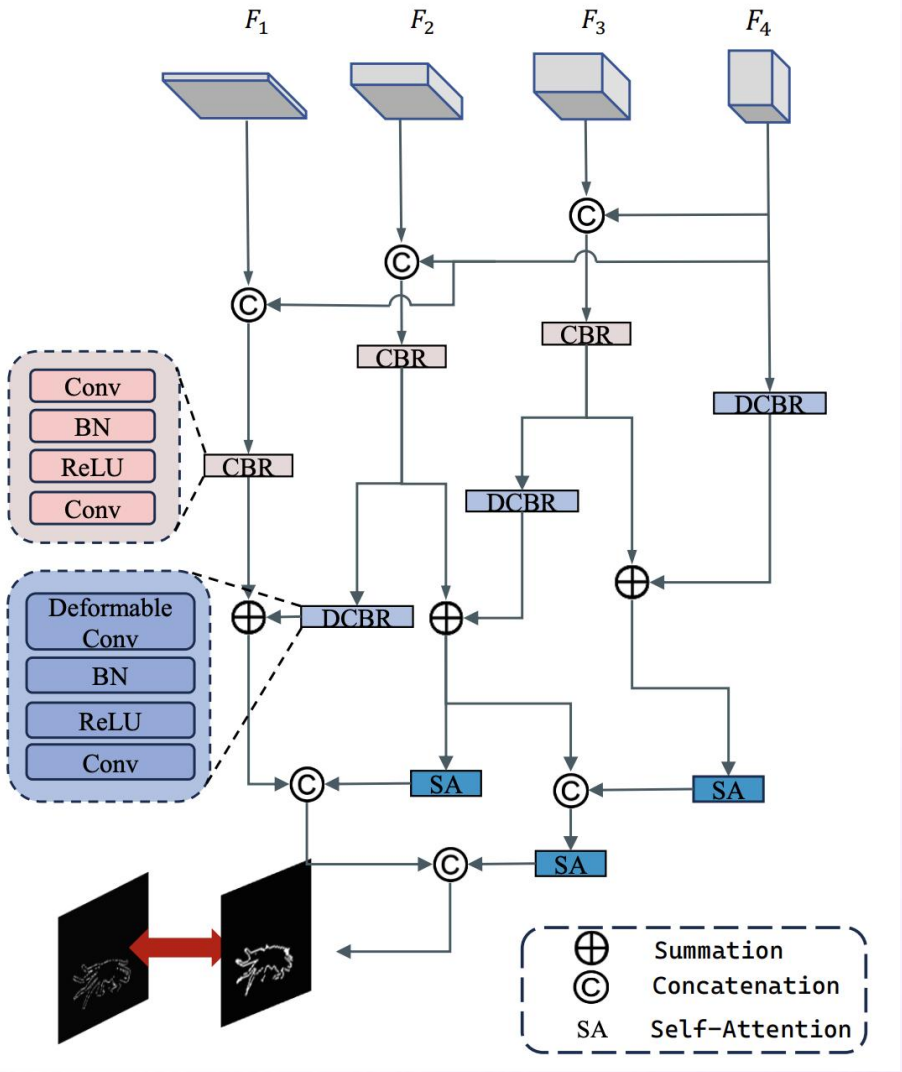
Deliver original-resolution texture patches to preserve local detail throughout the network.

03

Method Details

Adaptive Edge-Texture Perceptor (AETP)

AETP establishes a symbiotic relationship between boundary localization and texture discrimination, overcoming the fixed-kernel limitations of conventional detectors.



Texture-Aware Deformable Fusion

Uses deformable convolutions with offsets predicted by a texture complexity analysis branch.

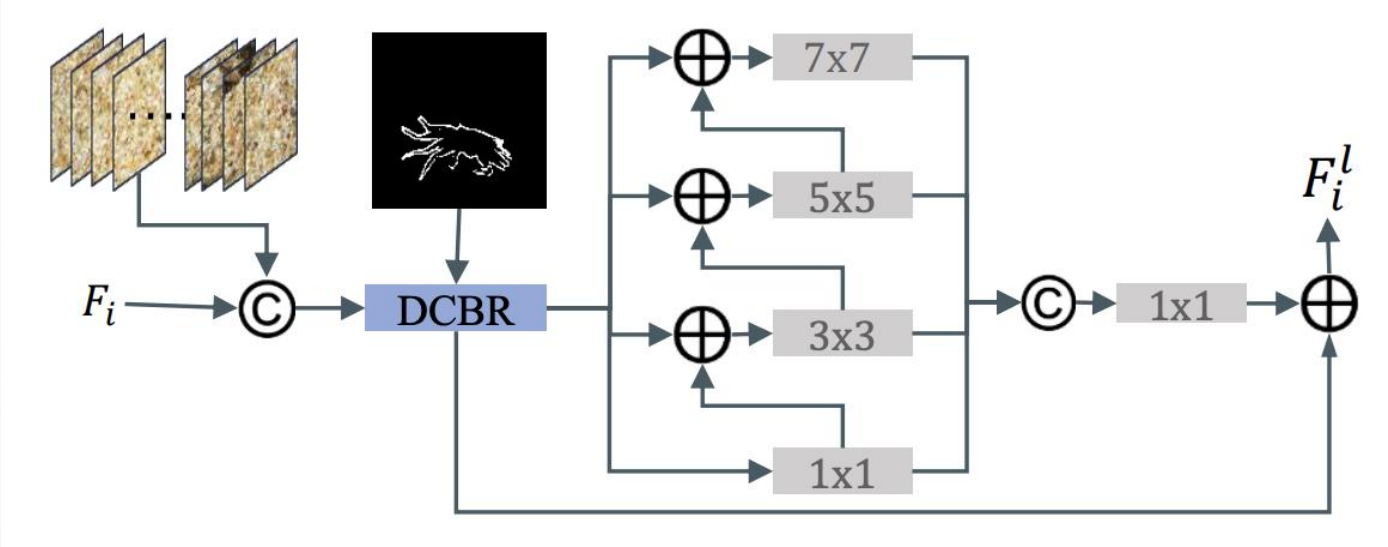
Cross-Scale Attention Guidance

A self-attention mechanism propagates long-range context, allowing ambiguous edge fragments to be completed by reliable texture regions.

Dual-Stream Feature Augmentor (DSFA)

DSFA enhances feature discriminability at critical boundary locations by concentrating network capacity on fractal boundaries and amorphous texture transitions.

- PatchRef: Crops high-res texture patches to compensate for info loss.
- Geometry-Texture Fusion: Uses edge-conditioned deformable convolutions.
- Feature-Texture Enhancement: Refines via cascaded residual blocks.



Multi-Feature Modulation Module (MFMM)

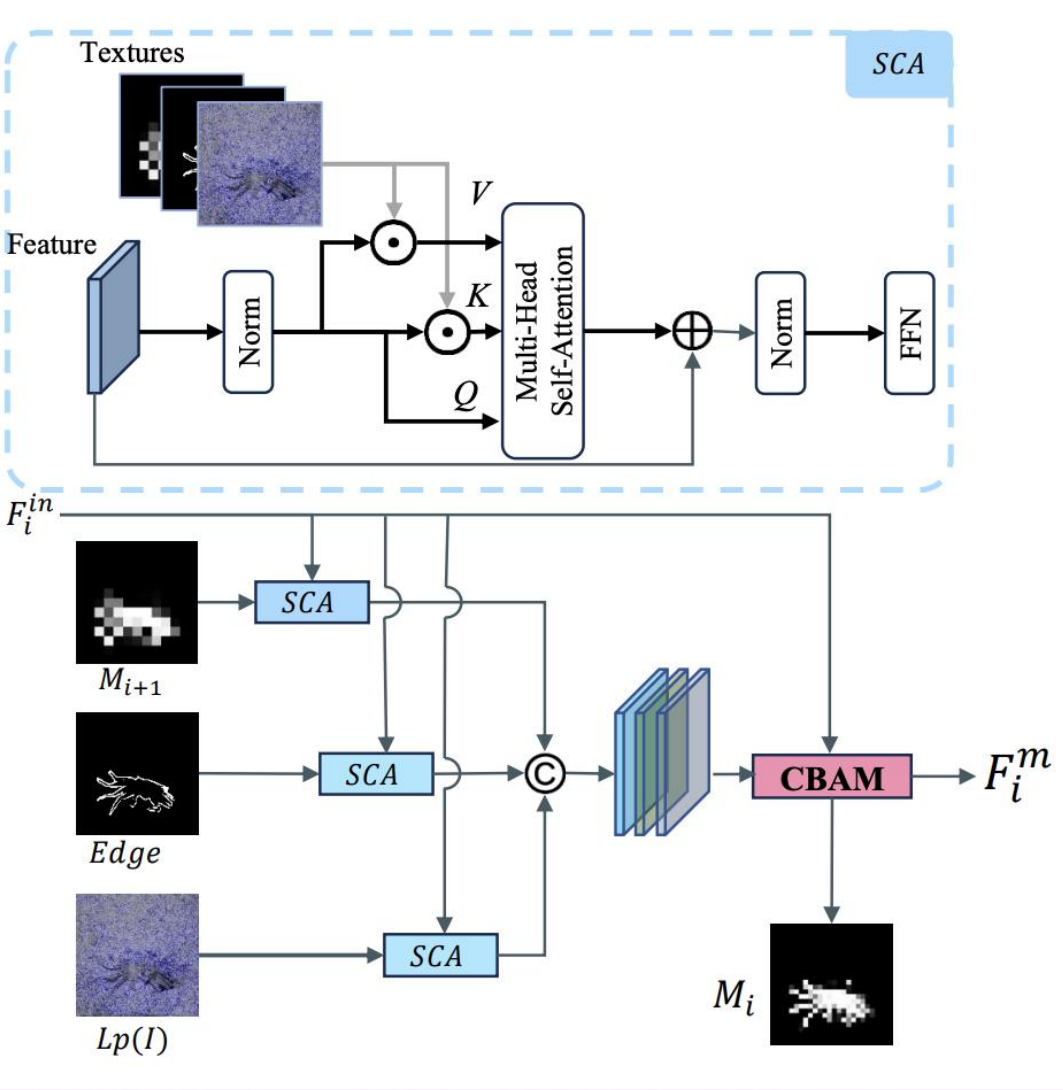
MFMM implements hierarchical edge-texture mutual verification to suppress fragmentation and progressively align object contours with internal texture patterns.

Mask Priors
Provides coarse-to-fine guidance. It gives a rough estimate of the object's location and shape.

Global Edge Map
Offers strong, explicit boundary constraints. This helps the model snap the predicted mask precisely to the object's real contours

Laplacian Texture
Captures high-frequency fine details. This allows the model to verify the internal texture of the potential object against the background

Attention Mechanism
Uses a Structure-Constrained Self-Attention (SCA) layer to focus on critical regions



These three branches are fused via Structure-Constrained Attention (SCA) and a CBAM block to produce the refined mask.

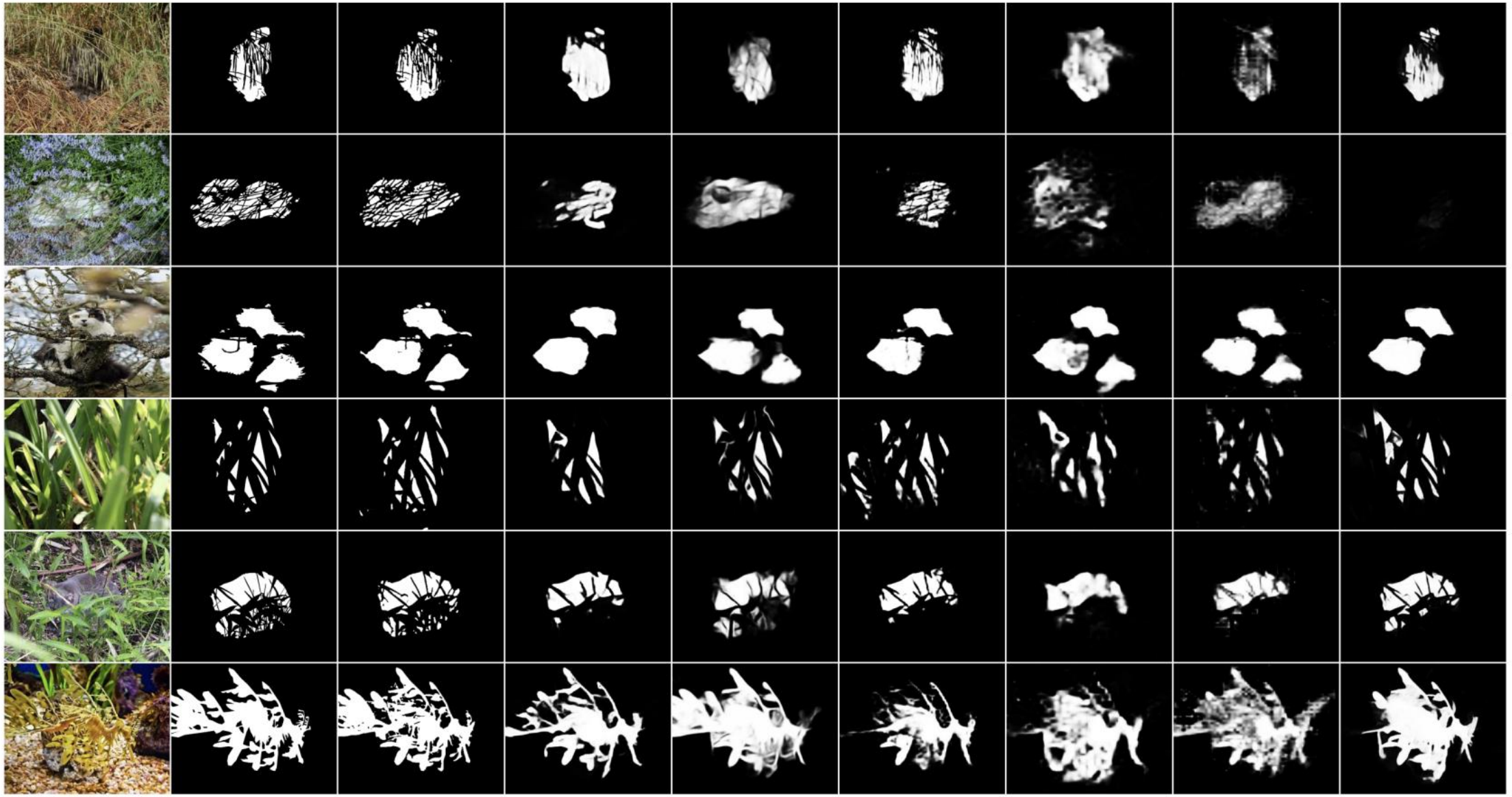
04

Experiments

Quantitative Comparison to SOTA

Methods	NC4K(4121)				COD10K(2026)				CAMO(250)			
	$F_\beta^w \uparrow$	$E_\phi \uparrow$	$S_m \uparrow$	$M \downarrow$	$F_\beta^w \uparrow$	$E_\phi \uparrow$	$S_m \uparrow$	$M \downarrow$	$F_\beta^w \uparrow$	$E_\phi \uparrow$	$S_m \uparrow$	$M \downarrow$
Convolution-based Backbone												
$SINet_{20}$ [12]	0.723	0.871	0.808	0.058	0.631	0.864	0.776	0.043	0.644	0.804	0.745	0.092
C^2FNet_{21} [46]	0.762	0.874	0.837	0.052	0.686	0.869	0.813	0.036	0.719	0.854	0.796	0.080
$PreyNet_{22}$ [56]	0.763	0.887	0.834	0.050	0.697	0.881	0.813	0.034	0.708	0.842	0.790	0.077
$SegMaR_{22}$ [26]	0.781	0.896	0.841	0.046	0.724	0.899	0.833	0.034	0.753	0.874	0.815	0.071
$BGNet_{22}$ [47]	0.788	0.907	0.851	0.044	0.722	0.901	0.831	0.033	0.751	0.871	0.816	0.069
$FindNet_{22}$ [30]	0.769	0.895	0.841	0.048	0.688	0.883	0.811	0.036	0.725	0.862	0.800	0.077
$ZoomNet_{22}$ [39]	0.784	0.896	0.853	0.043	0.729	0.888	0.838	0.029	0.752	0.877	0.820	0.066
$SINet\text{-}V2_{22}$ [15]	0.770	0.903	0.847	0.048	0.680	0.887	0.815	0.037	0.743	0.882	0.820	0.070
$DGNet_{23}$ [24]	0.784	0.911	0.857	0.042	0.693	0.896	0.822	0.033	0.901	0.769	0.839	0.057
$FEDER_{23}$ [19]	0.789	0.905	0.846	0.045	0.716	0.900	0.822	0.032	0.738	0.867	0.802	0.071
$Camouflageator_{24}$ [20]	0.835	0.922	0.869	0.041	0.763	0.920	0.843	0.028	0.805	0.891	0.829	0.066
Transformer-based Backbone												
$UGTR_{21}$ [51]	0.747	0.874	0.839	0.052	0.667	0.853	0.818	0.035	0.686	0.823	0.785	0.086
$FSPNet_{23}$ [23]	0.816	0.915	0.879	0.035	0.735	0.895	0.851	0.026	0.799	0.899	0.856	0.050
$HitNet_{23}$ [22]	0.834	0.926	0.875	0.037	**0.806**	0.935	0.871	0.023	0.809	0.906	0.849	0.055
$CamoFormer_{24}$ [53]	0.847	0.938	**0.892**	0.030	0.786	0.930	0.869	0.023	0.831	0.929	**0.872**	0.046
ESCNet(Ours)	**0.859**	**0.941**	**0.892**	**0.028**	0.804	**0.939**	**0.873**	**0.021**	**0.843**	**0.934**	0.871	**0.044**

Visual Comparison



Ablation Study

Modules ablation

Settings	NC4K (4121)				COD10K (2026)				CAMO (250)			
	$F_\beta^w \uparrow$	$E_\phi \uparrow$	$S_m \uparrow$	$M \downarrow$	$F_\beta^w \uparrow$	$E_\phi \uparrow$	$S_m \uparrow$	$M \downarrow$	$F_\beta^w \uparrow$	$E_\phi \uparrow$	$S_m \uparrow$	$M \downarrow$
Baseline	0.794	0.901	0.846	0.054	0.720	0.891	0.832	0.037	0.775	0.894	0.832	0.063
MFMM	0.836	0.927	0.870	0.035	0.793	0.925	0.863	0.025	0.827	0.919	0.859	0.048
AETP + MFMM	0.846	0.934	**0.893**	0.032	0.796	0.930	**0.877**	0.023	0.837	0.928	**0.876**	0.047
AETP + DSFA + MFMM	**0.859**	**0.941**	0.892	**0.028**	**0.804**	**0.939**	0.873	**0.021**	**0.843**	**0.934**	0.871	**0.044**

MFMM ablation

Exp.	MFMM			NC4K (4121)				COD10K (2026)				CAMO (250)			
	mask	edge	laplace	$F_\beta^w \uparrow$	$E_\phi \uparrow$	$S_m \uparrow$	$M \downarrow$	$F_\beta^w \uparrow$	$E_\phi \uparrow$	$S_m \uparrow$	$M \downarrow$	$F_\beta^w \uparrow$	$E_\phi \uparrow$	$S_m \uparrow$	$M \downarrow$
1	✓			0.845	0.929	0.868	0.033	0.788	0.927	0.855	0.025	0.833	0922	0.856	0.049
2		✓		0.842	0.920	0.863	0.035	0.785	0.925	0.850	0.026	0.829	0.917	0.850	0.051
3			✓	0.839	0.917	0.859	0.039	0.780	0.920	0.845	0.028	0.823	0.911	0.847	0.054
4		✓	✓	0.845	0.929	0.877	0.032	0.791	0.929	0.860	0.023	0.833	0.926	0.862	0.048
5	✓		✓	0.849	0.934	0.880	0.030	0.795	0.933	0.863	0.024	0.835	0.926	0.864	0.049
6	✓	✓		0.850	0.935	0.884	0.030	0.795	0.935	0.867	0.022	0.839	0.929	0.868	0.046
7	✓	✓	✓	**0.859**	**0.941**	**0.892**	**0.028**	**0.804**	**0.939**	**0.873**	**0.021**	**0.843**	**0.934**	**0.871**	**0.044**

05
Conclusion

Conclusion

ESCNet introduces a dynamic edge-texture collaborative paradigm that breaks the performance ceiling of single-cue COD.

- Mutual Reinforcement: AETP, DSFA, and MFMM convert camouflage into a jointly optimisable constraint.
- SOTA Performance: Achieves new state-of-the-art accuracy with sharper boundaries.
- Broader Impact: Offers a principled solution for any low-contrast segmentation task.