CAMMEM: STRUCTURED CAMOUFLAGE-STATE MEMORY REPLAY WITH SAM2 FOR CAMOUFLAGED OBJECT DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Camouflaged object detection (COD) aims to segment objects concealed by their surroundings, where high foreground–background similarity and indistinct boundaries make accurate perception particularly challenging. Although Segment Anything Model 2 (SAM2) provides strong visual representations and memory mechanisms for video propagation, these mechanisms are not naturally tailored to single-image camouflage perception. In this work, we propose CamMem, a structured camouflage-state memory replay framework that adapts SAM2 to COD through an intra-image memory writing–memory replay–conflict correction process. Specifically, we first introduce Structured Camouflage-State Memory (SCSM), which organizes coarse predictions and visual features into task-specific camouflage states encoding foreground-background, boundary, reliability, and conflict cues. We then develop a Camouflage-State Replay Adapter (CSRA), which combines relation-calibrated state matching, evidence-directed replay, and spatial–channel gated injection to selectively propagate structured camouflage states across SAM2 encoder blocks. Finally, Conflict-Guided Anchor Correction (CGAC) retains the replay-conditioned prediction as the primary hypothesis and uses the initial prediction as a stable anchor, regulating anchor correction through prediction conflict and local-support preference with bounded residual mixing. Experimental results on four COD benchmarks, including CAMO, COD10K, NC4K, and CHAMELEON, show that CamMem achieves substantial performance gains over the SAM2-based baseline and delivers competitive results against state-of-the-art COD methods.

1 INTRODUCTION

Camouflaged object detection (COD) aims to segment objects that are visually concealed within their surroundings (Fan et al., 2020; 2022). Potential applications include wildlife monitoring and search-and-rescue scenarios (Le et al., 2019). However, the strong visual similarity between camouflaged objects and their surroundings, together with indistinct boundaries, makes accurate camouflage segmentation particularly challenging.

The emergence of vision foundation models has opened new possibilities for generic visual segmentation. Segment Anything Model (SAM) (Kirillov et al., 2023) provides transferable representations for promptable segmentation. Its successor, SAM2 (Ravi et al., 2025), further extends this capability by introducing a memory-driven mechanism that stores and reuses visual information during object propagation. More importantly, this design extends segmentation beyond isolated feed-forward inference by allowing previously acquired visual evidence to condition subsequent predictions.

Memory mechanisms have been widely explored in visual perception to retain and reuse previously observed information. Two representative forms are temporal memory, which stores observations from preceding frames for object propagation (Oh et al., 2019; Zhou et al., 2025b), and historical memory, which accumulates knowledge across samples for subsequent retrieval (Zhang et al., 2025). As illustrated in Fig. 1(a), both types of memory primarily depend on information beyond the current image. However, temporal memory is unavailable in single-image COD due to the absence of preceding frames, while historical memory focuses on knowledge accumulated across samples and

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

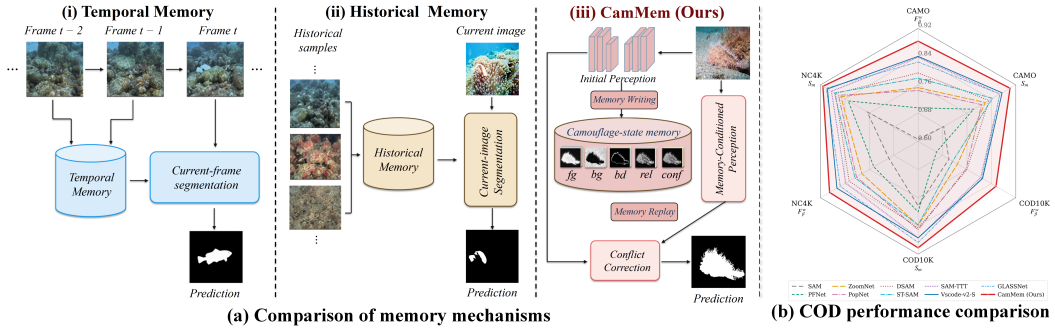


Figure 1: Motivation and performance comparison of CamMem. (a) Comparison of temporal, historical, and camouflage-state memory mechanisms. (b) Radar-chart comparison on CAMO, COD10K, and NC4K using F_β^w and S_m.

may not adequately capture the image-specific camouflage cues required for the current prediction. More importantly, the initial perception of a camouflaged image often contains heterogeneous and potentially conflicting responses, making it difficult to represent them with a single undifferentiated memory representation. Therefore, how to construct task-specific memory from the current image itself and effectively reuse it for subsequent camouflage perception remains insufficiently explored.

To address this limitation, we propose CamMem, a structured camouflage-state memory replay framework for single-image COD. CamMem establishes an intra-image memory writing–replay–correction process. First, Structured Camouflage-State Memory (SCSM) constructs foreground, background, boundary, reliable, and conflict prototypes from the initial prediction, determining *what to memorize*. Next, Camouflage-State Replay Adapter (CSRA) selectively reuses these memory states through relation-calibrated matching, evidence-directed replay, and spatial–channel gated injection, determining *how to replay*. Finally, Conflict-Guided Anchor Correction (CGAC) uses the initial prediction as a stable anchor to selectively correct inconsistent regions in the replay-conditioned prediction, determining *where to correct*. Together, these components enable image-specific memory construction, structured replay, and selective conflict correction. As shown in Fig. 1(b), CamMem also achieves consistently strong performance across multiple COD benchmarks.

Our main contributions are summarized as follows:

- We propose CamMem, a structured camouflage-state memory replay framework for single-image COD, which constructs and reuses memory from the model's own initial perception through an intra-image memory writing–memory replay–conflict correction process.
- We introduce Structured Camouflage-State Memory (SCSM), which organizes initial perception cues into foreground, background, boundary, reliable, and conflict prototypes to explicitly represent heterogeneous camouflage evidence.
- We develop a Camouflage-State Replay Adapter (CSRA) that combines relation-calibrated state matching, evidence-directed replay, and spatial–channel gated injection to selectively replay task-relevant camouflage states.
- We design Conflict-Guided Anchor Correction (CGAC), which uses the initial prediction as a stable anchor to selectively correct unreliable replay responses according to prediction conflict and local-support preference.

2 Related Work

2.1 Prediction-Guided Camouflaged Object Detection

Camouflaged object detection requires recovering object structure from weak foreground–background differences. Several methods use intermediate predictions to guide subsequent processing. SINet (Fan et al., 2020) uses a coarse camouflage prediction to guide feature enhancement

for identification, while PFNet (Mei et al., 2021) progressively refines coarse predictions through distraction mining. CamoFormer (Yin et al., 2024) uses intermediate predictions as foreground and background masks to guide separable attention during decoding. SegMaR (Jia et al., 2022) iteratively segments and magnifies object regions through discriminative masks and attention-based sampling. A related approach, UGTR (Yang et al., 2021), quantifies uncertainty in initial estimates and combines uncertainty-guided transformer reasoning with an auxiliary prototyping transformer. These methods motivate the use of intermediate predictions and their uncertainty to guide further processing. CamMem develops this direction by aggregating visual features according to cues derived from the initial prediction to construct five overlapping camouflage-state prototypes, encoding foreground, background, boundary, reliability, and conflict evidence. These states form an image-specific memory that is replayed into encoder representations, while conflict-guided anchor correction uses the initial prediction to selectively correct the replay-conditioned output.

2.2 Memory-Guided Camouflage Perception

Memory mechanisms have recently been explored in camouflage perception to reuse informative evidence beyond a single feed-forward prediction. For video camouflage perception, Zhou et al. (2025b) systematically investigate SAM2 for video camouflaged object segmentation, where its memory mechanism propagates information across frames. Feng et al. (2025) propose SRR, which selects informative memory reference frames and integrates long-term reference information with short-term motion cues. CamSAM2 (Zhou et al., 2025a) further exploits current- and previous-frame features and constructs object prototypes from high-quality historical observations to guide video camouflage segmentation.

Beyond temporal information, recent studies also explore historical or prototype-based knowledge across samples. RetroMem (Zhang et al., 2025) constructs a dynamic memory mechanism to retrieve relevant historical knowledge for current-sample inference. In referring COD, RefOnce (Wu et al., 2025c) distills reference images into a category-level prototype memory and generates query-conditioned prototype guidance at inference. RA-COD (Du et al., 2026) retrieves relevant visual prototypes from an external repository constructed using foundation models to support training-free camouflage detection. These studies demonstrate the value of reusing previously acquired camouflage knowledge, but their auxiliary evidence mainly originates from preceding frames, historical samples, or external prototype repositories. In contrast, CamMem constructs structured memory directly from the model's own initial perception of the current image, enabling memory writing and replay without relying on cross-frame or cross-sample information.

3 Method

The overall architecture of CamMem is illustrated in Fig. 2. Inspired by the memory-driven mechanism of SAM2, CamMem adapts it to single-image COD through three stages. In Stage I, SCSM derives camouflage cues from the initial prediction and aggregates visual features into foreground, background, boundary, reliable, and conflict prototypes, determining *what to memorize*. In Stage II, CSRA selectively replays these structured states into SAM2 encoder representations, determining *how to replay*. In Stage III, CGAC uses the initial prediction as a stable anchor to selectively correct inconsistent replay responses, determining *where to correct*.

3.1 Structured Camouflage-State Memory

SAM2 employs memory to store visual evidence from previous video frames and reuse it for temporal object propagation (Ravi et al., 2025). However, single-image COD provides no cross-frame observations, while its coarse prediction often contains heterogeneous perceptual evidence, including confident foreground and background responses, ambiguous boundary transitions, reliable foreground evidence, and conflict regions. Treating these mixed cues as an undifferentiated memory representation may obscure their distinct roles in camouflage perception.

To address this issue, we introduce Structured Camouflage-State Memory (SCSM), which adapts the idea of memory-guided propagation to single-image COD by constructing task-specific intra-image camouflage states. SCSM organizes the initial perception into foreground, background, boundary, reliable, and conflict states and further distills them into state-specific prototypes.

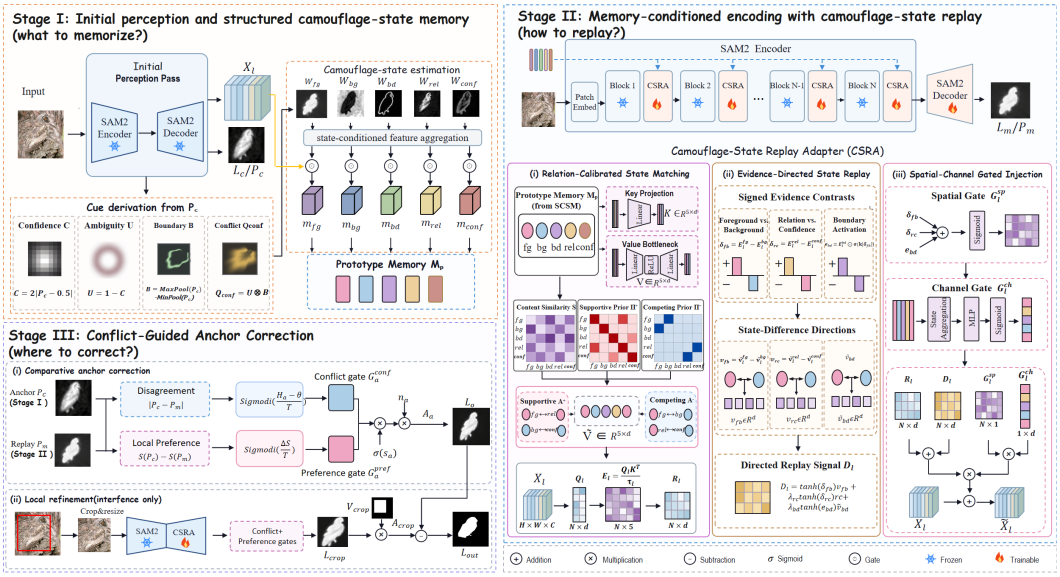


Figure 2: Overall framework of CamMem. Stage I constructs structured camouflage-state memory from the initial perception, Stage II replays the memory to refine SAM2 encoder representations, and Stage III performs conflict-guided anchor correction on the replay-conditioned prediction.

Camouflage-state estimation. During memory writing, CamMem first obtains a coarse prediction $\mathbf{L}_c$ and converts it into a probability map $\mathbf{P}_c = \sigma(\mathbf{L}_c)$. We then derive confidence, ambiguity, boundary, and conflict cues as

$$\mathbf{C} = 2|\mathbf{P}_c - 0.5|, \quad \mathbf{U} = 1 - \mathbf{C}, \quad \mathbf{B} = \mathrm{MaxPool}(\mathbf{P}_c) - \mathrm{MinPool}(\mathbf{P}_c), \quad \mathbf{Q}_{conf} = \mathbf{U} \odot \mathbf{B}. \tag{1}$$

Based on these cues, we construct five complementary camouflage-state maps $\mathcal{S} = \{fg, bg, bd, rel, conf\}$:

$$\begin{aligned} \mathbf{W}_{fg} &= \mathbf{P}_c \odot \mathbf{C}, \quad \mathbf{W}_{bg} = (1 - \mathbf{P}_c) \odot \mathbf{C}, \quad \mathbf{W}_{bd} = \mathbf{B} \odot (0.5 + \mathbf{U}), \\ \mathbf{W}_{rel} &= \mathbf{W}_{fg} \odot (1 - \mathbf{B}), \qquad \mathbf{W}_{conf} = \mathbf{P}_c \odot (\mathbf{U} + \mathbf{Q}_{conf}) + \mathbf{B} \odot \mathbf{Q}_{conf}. \end{aligned} \tag{2}$$

Here, $\mathbf{W}_{fg}$ and $\mathbf{W}_{bg}$ describe confident foreground and background evidence, $\mathbf{W}_{bd}$ emphasizes ambiguous structural transitions, $\mathbf{W}_{rel}$ retains reliable foreground responses away from boundaries, and $\mathbf{W}_{conf}$ aggregates uncertain object-like and conflicting structural cues. These state maps serve as overlapping soft aggregation weights rather than mutually exclusive region assignments.

Semantic state prototypes. Prediction-weighted feature aggregation has been used to construct object-context representations for semantic segmentation (Yuan et al., 2020). Here, the aggregation weights encode five overlapping camouflage states. Given the encoder feature map $\mathbf{F}_c$, we aggregate visual semantics according to each state and project them into a shared memory space:

$$\bar{\mathbf{m}}_s = \frac{\sum_u \mathbf{W}_s(u)\mathbf{F}_c(u)}{\sum_u \mathbf{W}_s(u) + \epsilon}, \qquad \mathbf{m}_s = \phi_p(\bar{\mathbf{m}}_s) + \mathbf{e}_s, \quad s \in \mathcal{S}, \tag{3}$$

where u indexes spatial locations, ϵ ensures numerical stability, $\phi_p(\cdot)$ denotes prototype projection, and $\mathbf{e}_s$ is a learnable state-type embedding. The resulting state prototypes are collected as the structured prototype memory $\mathbf{M}_p = [\mathbf{m}_s]_{s \in \mathcal{S}}$.

3.2 Camouflage-State Replay Adapter

SCSM determines what camouflage evidence should be written into memory, but the resulting prototypes still need to be selectively interpreted and replayed to recalibrate visual representations. Due

to the strong foreground–background ambiguity in camouflaged scenes, state relevance alone is insufficient to resolve competing perceptual evidence. We therefore introduce the Camouflage-State Replay Adapter (CSRA), which models inter-state relations and converts competing state evidence into direction-sensitive corrections for SAM2 encoder representations.

Relation-calibrated state matching. Given the prototype memory $\mathbf{M}_p$, CSRA constructs normalized keys $\mathbf{K} = \mathrm{Norm}(\phi_k(\mathbf{M}_p))$ and lightweight replay values $\mathbf{V} = \phi_u(\phi_d(\mathbf{M}_p))$, where $\phi_k(\cdot)$ denotes key projection, while $\phi_d(\cdot)$ and $\phi_u(\cdot)$ form a lightweight bottleneck for replay values. The prototype memory is normalized as $\widehat{\mathbf{M}}_p = \mathrm{Norm}(\mathbf{M}_p)$, and the content-dependent inter-state similarity is computed as $\mathbf{S} = \widehat{\mathbf{M}}_p \widehat{\mathbf{M}}_p^{\mathsf{T}}$. Since the camouflage states are not independent, we introduce supportive and competing relation priors $\mathbf{\Pi}^+$ and $\mathbf{\Pi}^-$ to encode task-specific state relationships:

$$\mathbf{A}^+ = \mathrm{Softmax}\left(\mathbf{S} + \mathbf{\Pi}^+\right), \qquad \mathbf{A}^- = \mathrm{Softmax}\left(-\mathbf{S} + \mathbf{\Pi}^-\right). \tag{4}$$

Using these relations, the replay values are calibrated as

$$\widetilde{\mathbf{V}} = \mathbf{V} + \alpha_+ \left(\mathbf{A}^+\mathbf{V} - \mathbf{V}\right) + \alpha_- \left(\mathbf{V} - \mathbf{A}^-\mathbf{V}\right), \tag{5}$$

where α_+ and α_- control supportive attraction and competitive separation, respectively.

For the l-th encoder block, the current representation $\mathbf{X}_l$ is projected into normalized query tokens $\mathbf{Q}_l = \mathrm{Norm}(\phi_q(\mathbf{X}_l))$. Its matching evidence with the structured memory is computed as $\mathbf{E}_l = \mathbf{Q}_l\mathbf{K}^{\mathsf{T}}/\tau_l$, and the relevance-based replay feature is obtained by $\mathbf{R}_l = \mathrm{Softmax}(\mathbf{E}_l)\widetilde{\mathbf{V}}$, where τ_l is a learnable temperature parameter, the softmax is applied over the state dimension, and $\mathbf{E}_l^s$ denotes the matching evidence associated with state s.

Evidence-directed state replay. The relevance-based replay $\mathbf{R}_l$ aggregates memory states according to their matching scores, but does not explicitly indicate which competing state should dominate the adaptation. Let $\widetilde{\mathbf{v}}_s$ denote the calibrated replay value associated with state s in $\widetilde{\mathbf{V}}$. We therefore construct signed foreground–background and reliable–conflict contrasts, together with boundary-conditioned evidence and their corresponding state-difference directions:

$$\begin{gathered}\delta_{fb} = \mathbf{E}_l^{fg} - \mathbf{E}_l^{bg}, \quad \delta_{rc} = \mathbf{E}_l^{rel} - \mathbf{E}_l^{conf}, \quad e_{bd} = \mathbf{E}_l^{bd} \odot \sigma(\kappa|\delta_{fb}|), \\ \mathbf{v}_{fb} = \widetilde{\mathbf{v}}_{fg} - \widetilde{\mathbf{v}}_{bg}, \qquad \mathbf{v}_{rc} = \widetilde{\mathbf{v}}_{rel} - \widetilde{\mathbf{v}}_{conf}.\end{gathered} \tag{6}$$

The directed replay signal is then formulated as

$$\mathbf{D}_l = \tanh(\delta_{fb})\mathbf{v}_{fb} + \lambda_{rc}\tanh(\delta_{rc})\mathbf{v}_{rc} + \lambda_{bd}\sigma(e_{bd})\widetilde{\mathbf{v}}_{bd}, \tag{7}$$

where $\kappa = 4$, $\lambda_{rc} = 0.5$, and $\lambda_{bd} = 0.25$. This formulation converts competing state evidence into directional feature corrections rather than treating all retrieved states equally.

Spatial–channel gated injection. Since memory evidence is not equally informative across spatial locations and feature channels, CSRA further employs a spatial gate $\mathbf{G}_l^{sp}$ and a channel gate $\mathbf{G}_l^{ch}$ to regulate replay. The gated replay residual is formulated as

$$\Delta\mathbf{X}_l = \mathbf{G}_l^{sp} \odot \mathbf{G}_l^{ch} \odot (\mathbf{R}_l + \rho_l\mathbf{D}_l), \tag{8}$$

where ρ_l controls the contribution of directed replay. The encoder representation is then updated through residual injection as $\widetilde{\mathbf{X}}_l = \mathbf{X}_l + \boldsymbol{\gamma}_l \odot \Delta\mathbf{X}_l$, where $\boldsymbol{\gamma}_l$ is a learnable residual scaling factor.

3.3 Conflict-Guided Anchor Correction

To selectively refine regions with inconsistent camouflage responses, we introduce Conflict-Guided Anchor Correction (CGAC). CGAC retains the replay-conditioned prediction as the primary hypothesis and uses the initial prediction as a stable anchor to determine where corrective information should be introduced.

Let $\mathbf{L}_m$ and $\mathbf{P}_m = \sigma(\mathbf{L}_m)$ denote the replay-conditioned prediction, while $\mathbf{L}_c$ and $\mathbf{P}_c$ denote the initial anchor. The ambiguity and boundary cues $\mathbf{U}_m$ and $\mathbf{B}_m$ are computed from $\mathbf{P}_m$ analogously to SCSM. The initial–replay disagreement is defined as $\mathbf{D}_{cm} = |\mathbf{P}_c - \mathbf{P}_m|$, with adaptive cue weights $[\omega_u, \omega_b, \omega_d] = \mathrm{Softmax}(\mathbf{a})$. These cues are aggregated into a conflict evidence map $\mathbf{H}_a = \omega_u \mathbf{U}_m + \omega_b \mathbf{B}_m + \omega_d \mathbf{D}_{cm}$. The Anchor–Replay Conflict Gate is then defined as

$$\mathbf{G}_a^{conf} = \sigma\left(\frac{\mathbf{H}_a - \theta}{T}\right), \tag{9}$$

where θ and T are learnable threshold and temperature parameters. A larger $\mathbf{G}_a^{conf}$ indicates stronger evidence for potential inconsistency between the replay-conditioned prediction and the stable anchor.

Since conflict evidence alone cannot determine which prediction is more reliable, we further measure local support and construct the Local-Support Preference Gate:

$$\mathcal{S}(\mathbf{P}) = 2|\mathbf{P} - 0.5| \odot \left(1 - \left|\mathbf{P} - \mathrm{AvgPool}_{3\times3}(\mathbf{P})\right|\right), \qquad \mathbf{G}_a^{pref} = \sigma\left(\frac{\mathcal{S}(\mathbf{P}_c) - \mathcal{S}(\mathbf{P}_m)}{T}\right). \tag{10}$$

The two gates jointly regulate bounded anchor correction:

$$\mathbf{A}_a = \eta_a \sigma(s_a) \mathbf{G}_a^{conf} \odot \mathbf{G}_a^{pref}, \qquad \mathbf{L}_o = \mathbf{L}_m + \mathbf{A}_a \odot (\mathbf{L}_c - \mathbf{L}_m), \tag{11}$$

where $\eta_a = 0.25$ bounds the mixing coefficient and s_a is learnable. Thus, anchor correction is emphasized in regions with stronger conflict evidence and stronger local support from the initial prediction.

During inference, CGAC further performs target-centered local refinement. A candidate region is cropped, processed by the same memory-conditioned model, and restored to the original coordinates. Let $\mathbf{L}_{crop}$ and $\mathbf{V}_{crop}$ denote the restored crop logits and valid-region mask, respectively. The crop conflict and preference gates, $\mathbf{G}_{crop}^{conf}$ and $\mathbf{G}_{crop}^{pref}$, are computed analogously by comparing the restored crop prediction with the global prediction $\mathbf{L}_o$. The final prediction is

$$\mathbf{L}_{out} = \mathbf{L}_o + \eta_c \mathbf{G}_{crop}^{conf} \odot \mathbf{G}_{crop}^{pref} \odot \mathbf{V}_{crop} \odot (\mathbf{L}_{crop} - \mathbf{L}_o), \tag{12}$$

where $\eta_c = 0.10$ bounds the local refinement contribution.

4 Experiments

4.1 Experimental Settings

Datasets and evaluation metrics. Experiments are conducted on four COD benchmarks: CAMO (Le et al., 2019), CHAMELEON (Skurowski et al., 2017), COD10K (Fan et al., 2020), and NC4K (Lv et al., 2021). Following common settings (Yin et al., 2024; Wu et al., 2025b), 1,000 CAMO images and 3,040 COD10K images are used for training, while the corresponding CAMO and COD10K test splits, together with the full CHAMELEON and NC4K datasets, are used for evaluation. Performance is assessed using mean absolute error (M), weighted F-measure (F_β^w) (Margolin et al., 2014), S-measure (S_m) (Fan et al., 2017), and maximum E-measure (E_x) (Fan et al., 2018), where lower M and higher F_β^w, S_m, and E_x indicate better performance.

Implementation details. Our model is initialized from the pretrained SAM2-Hiera-L checkpoint (Ravi et al., 2025; Ryali et al., 2023). All images are resized to 512×512, and the experiments are performed on two NVIDIA GeForce RTX 4090 GPUs. The network is trained for 80 epochs using AdamW (Loshchilov and Hutter, 2019) with an initial learning rate of 5×10^{-4} and a weight decay of 1×10^{-5}. A five-epoch linear warm-up is adopted before normal optimization. For supervision, we employ a composite objective consisting of BCE loss, IoU loss, and MAE loss, with equal contributions from the three terms.

Table 1: Quantitative comparison of COD methods on CAMO, CHAMELEON, COD10K, and NC4K. The best and second-best results are highlighted in light red and light blue, and indicated by boldface and underlining, respectively.

Method	Pub./Year	CAMO				CHAMELEON				COD10K				NC4K			
		$M\downarrow$	$F_\beta^w\uparrow$	$S_m\uparrow$	$E_x\uparrow$	$M\downarrow$	$F_\beta^w\uparrow$	$S_m\uparrow$	$E_x\uparrow$	$M\downarrow$	$F_\beta^w\uparrow$	$S_m\uparrow$	$E_x\uparrow$	$M\downarrow$	$F_\beta^w\uparrow$	$S_m\uparrow$	$E_x\uparrow$
Non-SAM COD Methods																	
SINet	$CVPR_{20}$	0.099	0.606	0.751	0.790	0.044	0.740	0.868	0.908	0.051	0.551	0.771	0.832	0.058	0.723	0.808	0.873
UGTR	$ICCV_{21}$	0.086	0.794	0.784	-	-	-	-	-	0.036	0.666	0.817	-	0.052	0.746	0.839	-
PFNet	$CVPR_{21}$	0.085	0.695	0.782	0.845	0.033	0.810	0.882	0.927	0.040	0.660	0.800	0.880	0.053	0.745	0.829	0.891
SegMaR	2022	0.071	0.742	0.815	-	-	-	-	-	0.033	0.724	0.833	-	0.046	0.781	0.841	-
SINetv2	$TPAMI_{22}$	0.070	0.743	0.820	0.882	0.030	0.816	0.888	-	0.037	0.680	0.815	0.887	0.048	0.770	0.847	0.903
FDNet	2022	0.062	0.778	0.844	-	-	-	-	-	0.030	0.731	0.837	-	0.052	0.750	0.834	-
ZoomNet	$CVPR_{22}$	0.066	0.752	0.820	-	0.023	0.845	0.902	-	0.029	0.729	0.838	-	0.043	0.784	0.853	-
TPRNet	$TVCJ_{23}$	0.076	0.781	0.814	-	-	-	-	-	0.034	0.725	0.829	-	0.047	0.790	0.854	-
PopNet	$ICCV_{23}$	0.077	0.744	0.808	-	0.020	0.875	0.917	-	0.028	0.757	0.851	-	0.042	0.802	0.861	-
EANet	$ICASSP_{24}$	0.051	0.793	0.841	-	-	-	-	-	0.029	0.709	0.825	-	0.039	0.798	0.860	-
CamoFormer-R	$TPAMI_{24}$	0.066	0.756	0.817	-	0.024	0.843	0.900	-	0.029	0.730	0.838	-	**0.024**	0.793	0.857	-
Vscode-S	$CVPR_{24}$	0.046	0.844	0.873	0.938	-	-	-	-	0.024	0.806	0.869	0.942	0.032	0.863	0.891	0.944
ZoomNeXt-R	$TPAMI_{24}$	0.069	0.760	0.822	-	0.020	0.864	0.912	-	0.026	0.758	0.855	-	0.038	0.808	0.869	-
UAT	TIP_{25}	0.052	0.810	0.858	-	0.021	0.875	0.914	-	0.025	0.772	0.859	-	0.035	0.832	0.879	-
CamoDiffusion	$TPAMI_{25}$	0.042	0.853	0.878	0.940	-	-	-	-	0.026	0.814	0.881	0.895	0.035	0.859	0.893	0.942
Vscode-v2-S	$TPAMI_{26}$	0.049	0.841	0.875	0.932	-	-	-	-	0.024	0.811	0.874	0.942	0.031	0.866	0.897	0.944
SAM-Based and SAM-Assisted COD Methods																	
SAM	$ICCV_{23}$	0.132	0.606	0.684	0.689	-	-	-	-	0.049	0.701	0.783	0.800	0.078	0.696	0.767	0.778
SAM-Adapter	$ICCVW_{23}$	0.070	0.765	0.847	0.873	0.033	0.824	0.896	0.919	0.025	0.801	0.883	0.918	-	-	-	-
SAM2-Adapter	$ArXiv_{24}$	0.051	0.810	0.855	0.909	**0.018**	0.889	0.915	0.955	**0.018**	0.850	0.899	0.950	-	-	-	-
WS-SAM	$NeurIPS_{23}$	0.092	0.742	0.759	-	0.056	0.767	0.805	-	0.039	0.6981	0.790	-	0.057	0.801	0.813	-
MedSAM	NC_{24}	0.065	0.780	0.820	0.913	-	-	-	-	0.033	0.751	0.841	0.926	0.041	0.821	0.866	0.937
DSAM	$ACMMM_{24}$	0.061	0.794	0.832	0.920	-	-	-	-	0.033	0.760	0.846	0.931	0.040	0.826	0.871	0.940
ST-SAM	$ACMMM_{25}$	0.058	0.779	0.844	0.890	0.032	0.804	0.876	0.926	0.030	0.713	0.837	0.874	0.037	0.807	0.874	0.911
HFS-SAM2	SPL_{25}	0.045	0.839	0.883	-	0.023	0.849	0.909	-	0.023	0.789	0.879	-	0.030	0.849	0.900	-
COD-SAM	PR_{25}	0.055	0.796	0.870	0.906	0.021	0.880	0.924	0.956	0.021	0.832	0.899	0.941	-	-	-	-
SAM-TTT	$ACMMM_{25}$	0.045	0.838	0.868	0.935	-	-	-	-	0.027	0.805	0.874	0.942	0.031	0.837	0.884	0.943
SAM2-UNet	$VisualIntell._{26}$	0.042	0.861	0.884	-	0.022	0.863	0.914	-	0.021	0.789	0.880	-	0.029	0.863	0.901	-
GLASSNet	$ArXiv_{26}$	0.044	0.825	0.868	-	-	-	-	-	0.020	0.809	0.887	-	0.029	0.855	0.901	-
DASWA-Net	$CVIU_{26}$	0.041	0.851	0.884	0.931	0.019	0.887	0.923	0.964	0.019	0.817	0.885	0.943	0.029	0.860	0.899	0.942
SAM-GDSOD	KBS_{26}	0.043	0.861	0.882	0.929	-	-	-	-	0.024	0.798	0.865	0.923	0.033	0.862	0.888	0.932
FCL-COD	$CVPRF_{26}$	0.050	0.824	0.862	0.915	0.038	0.842	0.882	0.932	0.022	0.808	0.878	0.928	0.033	0.846	0.885	0.928
Ours	–	**0.037**	**0.884**	**0.901**	**0.950**	**0.018**	**0.902**	**0.927**	**0.971**	**0.018**	**0.855**	**0.902**	**0.952**	0.027	**0.890**	**0.912**	**0.953**

4.2 Comparisons with State-of-the-Art Methods

Compared methods. We compare CamMem with 31 representative methods, covering both non-SAM and SAM-based approaches. The non-SAM methods include SINet (Fan et al., 2020), UGTR (Yang et al., 2021), PFNet (Mei et al., 2021), SegMaR (Jia et al., 2022), SINetv2 (Fan et al., 2022), FDNet (Zhong et al., 2022), ZoomNet (Pang et al., 2022), TPRNet (Zhang et al., 2023), PopNet (Wu et al., 2023), EANet (Liu et al., 2024), CamoFormer-R (Yin et al., 2024), Vscode-S (Luo et al., 2024), ZoomNeXt-R (Pang et al., 2024), UAT (Wu et al., 2025a), CamoDiffusion (Sun et al., 2025), and Vscode-v2-S (Luo et al., 2026). The SAM-based and SAM-assisted methods include SAM (Kirillov et al., 2023), SAM-Adapter (Chen et al., 2023), SAM2-Adapter (Chen et al., 2024), WS-SAM (He et al., 2023), MedSAM (Ma et al., 2024), DSAM (Yu et al., 2024), ST-SAM (Hu et al., 2025), HFS-SAM2 (Wu et al., 2025b), COD-SAM (Gao et al., 2025), SAM-TTT (Yu et al., 2025), SAM2-UNet (Xiong et al., 2026), GLASSNet (Moradi et al., 2026), DASWA-Net (Song et al., 2026), SAM-GDSOD (Zhang et al., 2026), and FCL-COD (Ni et al., 2026). These methods span different supervision regimes: WS-SAM, SAM-GDSOD, and FCL-COD use weak supervision, whereas ST-SAM uses semi-supervision. The comparison therefore covers diverse training settings as well as model architectures.

Quantitative comparison. As reported in Table 1, CamMem achieves the best or tied-best results on 15 of the 16 evaluation entries across CAMO, CHAMELEON, COD10K, and NC4K, while obtaining the second-best MAE on NC4K, demonstrating consistently strong performance across different camouflage scenarios. Compared with non-SAM COD methods, CamMem shows substantial gains over early representative approaches. On CAMO, compared with SINet, CamMem reduces MAE by 62.6% and improves F_β^w, S_m, and E_x by 45.9%, 20.0%, and 20.3%, respectively. Against recent methods, CamMem also surpasses CamoDiffusion on CAMO and improves over Vscode-v2-S on COD10K, reducing MAE from 0.024 to 0.018 while increasing F_β^w from 0.811 to 0.855. Com-

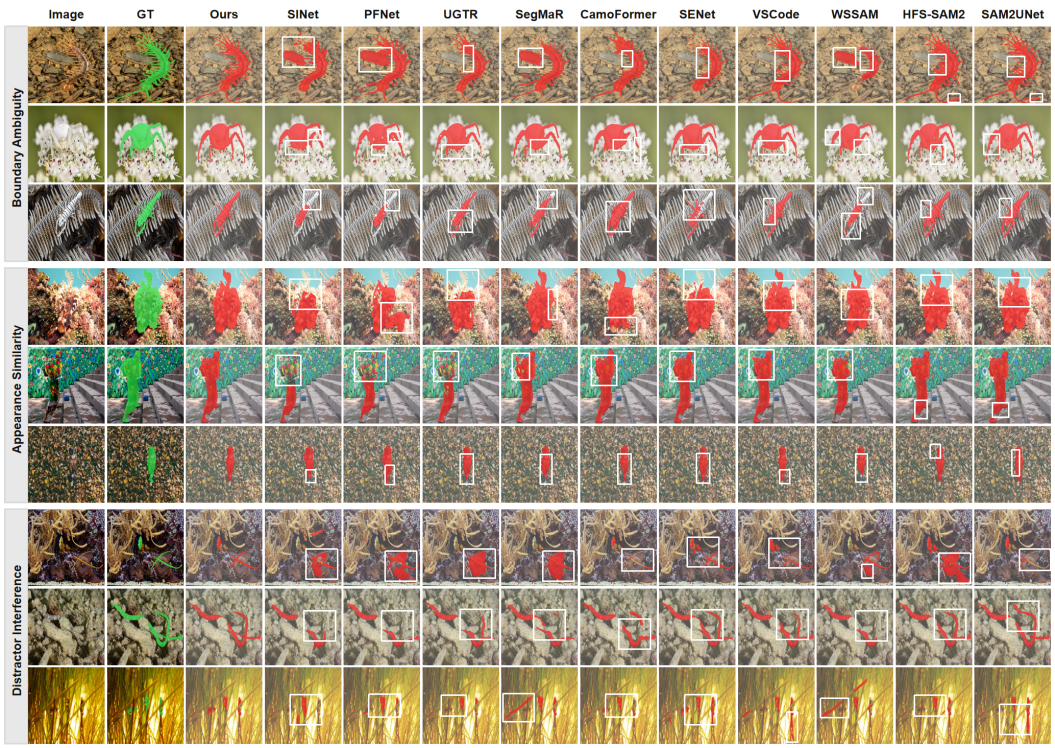


Figure 3: Qualitative comparison under three challenging scenarios: boundary ambiguity, appearance similarity, and distractor interference. CamMem achieves more complete object structures and more accurate localization than existing methods.

Table 2: Ablation study of the main components in CamMem.

CSRA	SCSM	CGAC	NC4K				CHAMELEON			
			$M\downarrow$	$F_\beta^w\uparrow$	$S_m\uparrow$	$E_x\uparrow$	$M\downarrow$	$F_\beta^w\uparrow$	$S_m\uparrow$	$E_x\uparrow$
–	–	–	0.037	0.848	0.886	0.929	0.030	0.848	0.892	0.940
✓	–	–	0.029	0.881	0.907	0.945	0.020	0.868	0.912	0.949
✓	✓	–	0.028	0.885	0.909	0.947	**0.018**	0.886	0.922	0.956
✓	✓	✓	**0.027**	**0.890**	**0.912**	**0.953**	**0.018**	**0.902**	**0.927**	**0.971**

Table 3: Ablation study of memory-state compositions.

Memory Composition	NC4K				CHAMELEON			
	$M\downarrow$	$F_\beta^w\uparrow$	$S_m\uparrow$	$E_x\uparrow$	$M\downarrow$	$F_\beta^w\uparrow$	$S_m\uparrow$	$E_x\uparrow$
Unified	0.029	0.881	0.907	0.945	0.020	0.868	0.912	0.949
FG + BG	0.028	0.883	0.907	0.944	0.019	0.873	0.916	0.953
FG + BG + BD	0.028	**0.886**	0.908	0.946	0.019	0.882	0.918	0.955
FG + BG + BD + REL	0.028	**0.886**	0.907	**0.948**	0.019	0.880	0.916	**0.956**
5-state SCSM	**0.027**	0.885	**0.909**	0.947	**0.018**	**0.886**	**0.922**	**0.956**

pared with SAM-based methods, CamMem also exhibits clear advantages. Compared with SAM-Adapter, F_β^w and E_x increase from 0.765 and 0.873 to 0.884 and 0.950, respectively. CamMem also consistently outperforms recent SAM-based COD methods such as COD-SAM and SAM-TTT on their reported benchmarks. Compared with existing SAM2-based COD methods, including SAM2-Adapter, HFS-SAM2, SAM2-UNet, GLASSNet, and DASWA-Net, CamMem achieves stronger performance on the available comparisons across the four benchmarks. This suggests that the gains are not solely attributable to the SAM2 backbone and further supports the effectiveness of structured camouflage-state memory construction, selective replay, and conflict-guided correction.

Qualitative comparison. Fig. 3 presents qualitative comparisons between CamMem and representative COD methods under three challenging scenarios, including boundary ambiguity, appearance similarity, and distractor interference. As highlighted by the white boxes, existing methods often exhibit missed object regions, inaccurate boundaries, or background false responses under difficult camouflage conditions. In contrast, CamMem better preserves fine structures, recovers more complete object regions, and suppresses distracting responses. These visual results further support the effectiveness of structured camouflage-state memory replay and are consistent with the quantitative results in Table 1.

Table 4: Ablation study of CSRA components.

Variant	NC4K				CHAMELEON			
	$M \downarrow$	$F_\beta^w \uparrow$	$S_m \uparrow$	$E_x \uparrow$	$M \downarrow$	$F_\beta^w \uparrow$	$S_m \uparrow$	$E_x \uparrow$
w/o Relation Calibration	0.028	**0.886**	0.908	0.946	0.019	0.882	0.921	0.955
w/o Evidence-Directed Replay	0.029	0.882	0.906	**0.947**	0.022	0.884	0.915	**0.957**
w/o Spatial–Channel Gating	0.028	0.884	0.907	0.945	0.022	0.869	0.908	0.941
Full CSRA	**0.027**	0.885	**0.909**	**0.947**	**0.018**	**0.886**	**0.922**	0.956

Table 5: Ablation study of CGAC components.

Variant	NC4K				CHAMELEON			
	$M \downarrow$	$F_\beta^w \uparrow$	$S_m \uparrow$	$E_x \uparrow$	$M \downarrow$	$F_\beta^w \uparrow$	$S_m \uparrow$	$E_x \uparrow$
w/o Conflict Gate	0.028	0.887	0.909	0.952	0.018	0.895	0.924	0.965
w/o Preference Gate	0.028	0.886	0.909	0.951	0.018	0.894	0.924	0.965
w/o Local Refinement	**0.027**	0.889	0.911	0.952	**0.017**	0.897	0.925	0.970
Full CGAC	**0.027**	**0.890**	**0.912**	**0.953**	0.018	**0.902**	**0.927**	**0.971**

4.3 Ablation

Main component ablation. Table 2 evaluates the contributions of CSRA, SCSM, and CGAC. Introducing CSRA with unified memory reduces MAE from 0.037 to 0.029 on NC4K and from 0.030 to 0.020 on CHAMELEON, while improving F_β^w by 0.033 and 0.020, respectively. Replacing unified memory with SCSM further improves CHAMELEON F_β^w from 0.868 to 0.886 and S_m from 0.912 to 0.922, confirming the benefit of structured camouflage-state memory. Finally, CGAC raises F_β^w from 0.885 to 0.890 and E_x from 0.947 to 0.953 on NC4K, while improving CHAMELEON F_β^w and E_x from 0.886 and 0.956 to 0.902 and 0.971. These progressive gains demonstrate the complementary effects of camouflage-state replay, structured memory construction, and conflict-guided anchor correction.

Ablation on memory-state composition. Compared with Unified memory, the complete five-state SCSM reduces MAE from 0.029 to 0.027 on NC4K and from 0.020 to 0.018 on CHAMELEON. More notably, on CHAMELEON, F_β^w and S_m increase from 0.868 and 0.912 to 0.886 and 0.922, respectively. Although some intermediate combinations achieve slightly higher individual scores on NC4K, the complete five-state design provides the best overall balance across both datasets, validating the complementary roles of foreground, background, boundary, reliable, and conflict states.

Ablation of CSRA components. Table 4 evaluates relation-calibrated state matching, evidence-directed state replay, and spatial–channel gated injection. Removing relation calibration decreases CHAMELEON F_β^w from 0.886 to 0.882 and increases MAE from 0.018 to 0.019, while removing directed replay increases MAE to 0.022 and reduces S_m from 0.922 to 0.915. The largest degradation is observed without spatial–channel gating: on CHAMELEON, F_β^w, S_m, and E_x drop from 0.886, 0.922, and 0.956 to 0.869, 0.908, and 0.941, respectively. These results confirm the complementary roles of relation calibration, evidence-directed replay, and spatial–channel gated injection in memory-conditioned feature refinement.

Ablation on CGAC components. Table 5 evaluates the internal components of CGAC. Removing the Anchor–Replay Conflict Gate decreases CHAMELEON F_β^w from 0.902 to 0.895 and E_x from 0.971 to 0.965, while removing the Local-Support Preference Gate further reduces F_β^w to 0.894. Without target-centered local refinement, F_β^w, S_m, and E_x on CHAMELEON decrease from 0.902, 0.927, and 0.971 to 0.897, 0.925, and 0.970, respectively, although MAE slightly improves from 0.018 to 0.017. On NC4K, the full CGAC achieves the highest F_β^w, S_m, and E_x of 0.890, 0.912, and 0.953. These results verify the complementary effects of conflict gating, local-support preference, and local refinement for selective anchor correction.

5 Conclusion

In this work, we propose CamMem, a structured camouflage-state memory replay framework for single-image COD. By constructing task-specific memory from the initial perception of the current image, CamMem establishes an intra-image memory writing–memory replay–conflict correction process. SCSM organizes heterogeneous camouflage evidence into structured state prototypes, CSRA selectively replays these states to refine SAM2 representations, and CGAC further improves prediction reliability through conflict-guided anchor correction. Extensive experiments on four COD benchmarks demonstrate the effectiveness of CamMem over conventional COD methods and recent SAM/SAM2-based approaches. Additional cross-task evaluations further indicate the generalization potential of the proposed image-specific memory mechanism. Future work will explore more adaptive memory construction and efficient memory utilization for broader visual perception tasks.

AI USE STATEMENT

In this work, we used generative AI tools solely for language polishing, including grammar correction, wording refinement, and improving the clarity and readability of the manuscript. Generative AI was not used to generate experimental results, perform data analysis, or make scientific conclusions. All AI-assisted revisions were carefully reviewed and verified by the authors to ensure technical accuracy and consistency with the original content. The authors take full responsibility for the final content of this work, including all text, claims, and results.

REPRODUCIBILITY STATEMENT

The source code for reproducing CamMem is available at the anonymous repository `https://anonymous.4open.science/r/Cammem-master-9AE3`. The repository provides the implementation and evaluation setup for the experiments reported in this paper. CamMem is evaluated on CAMO, CHAMELEON, COD10K, and NC4K using the metrics described in the experimental section.

REFERENCES

Bahareh Asheghi, Pedram Salehpour, Abdolhamid Moallemi Khiavi, Mahdi Hashemzadeh, and Amirhassan Monajemi. DASOD: Detail-aware salient object detection. *Image and Vision Computing*, 148:105154, 2024. URL https://doi.org/10.1016/j.imavis.2024.105154.

Shuyao Chai, Xixuan Zhao, Jiaming Zhang, and Jiangming Kan. Defocus blur detection based on transformer and complementary residual learning. *Multimedia Tools and Applications*, 83(17):53095–53118, 2024. URL https://doi.org/10.1007/s11042-023-17560-7.

Tianrun Chen, Lanyun Zhu, Chaotao Ding, Runlong Cao, Yan Wang, Shangzhan Zhang, Zejian Li, Lingyun Sun, Ying Zang, and Papa Mao. SAM-Adapter: Adapting segment anything in underperformed scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 3359–3367, 2023. URL https://doi.org/10.1109/iccvw60793.2023.00361.

Tianrun Chen, Ankang Lu, Lanyun Zhu, Chaotao Ding, Chunan Yu, Deyi Ji, Zejian Li, Lingyun Sun, Papa Mao, and Ying Zang. SAM2-Adapter: Evaluating & adapting Segment Anything 2 in downstream tasks: Camouflage, shadow, medical image segmentation, and more. *arXiv preprint arXiv:2408.04579*, 2024. URL https://doi.org/10.48550/arXiv.2408.04579.

Jing Dong, Wei Wang, and Tieniu Tan. CASIA image tampering detection evaluation database. In *Proceedings of the IEEE China Summit and International Conference on Signal and Information Processing*, pp. 422–426, 2013. URL https://doi.org/10.1109/chinasip.2013.6625374.

Ji Du, Jiesheng Wu, Desheng Kong, Fangwei Hao, Jing Xu, and Ping Li. RA-COD: Retrieval-augmented camouflaged object detection. *IEEE Transactions on Image Processing*, 35:5253–5265, 2026. URL https://doi.org/10.1109/tip.2026.3691679.

Rohit Venkata Sai Dulam and Chandra Kambhamettu. Salient object detection with dynamic convolutions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1683–1693, 2025. URL https://doi.org/10.1109/cvprw67362.2025.00157.

Deng-Ping Fan, Ming-Ming Cheng, Yun Liu, Tao Li, and Ali Borji. Structure-measure: A new way to evaluate foreground maps. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4558–4567, 2017. URL https://doi.org/10.1109/iccv.2017.487.

Deng-Ping Fan, Cheng Gong, Yang Cao, Bo Ren, Ming-Ming Cheng, and Ali Borji. Enhanced-alignment measure for binary foreground map evaluation. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pp. 698–704, 2018. URL https://doi.org/10.24963/ijcai.2018/97.

Deng-Ping Fan, Ge-Peng Ji, Guolei Sun, Ming-Ming Cheng, Jianbing Shen, and Ling Shao. Camouflaged object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2774–2784, 2020. URL https://doi.org/10.1109/cvpr42600.2020.00285.

Deng-Ping Fan, Ge-Peng Ji, Ming-Ming Cheng, and Ling Shao. Concealed object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6024–6042, 2022. URL https://doi.org/10.1109/tpami.2021.3085766.

Yu'ang Feng, Shuyong Gao, Fuzhen Yan, Yicheng Song, Lingyi Hong, Junjie Hu, and Wenqiang Zhang. Scoring, remember, and reference: Catching camouflaged objects in videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13043–13052, 2025. URL https://doi.org/10.1109/iccv51701.2025.01212.

Shixuan Gao, Pingping Zhang, Tianyu Yan, and Huchuan Lu. Multi-scale and detail-enhanced segment anything model for salient object detection. In *Proceedings of the ACM International Conference on Multimedia*, pp. 9894–9903, 2024. URL https://doi.org/10.1145/3664647.3680650.

Dongyang Gao, Yichao Zhou, Hui Yan, Chen Chen, and Xiyuan Hu. COD-SAM: Camouflage object detection using SAM. *Pattern Recognition*, 168:111826, 2025. URL https://doi.org/10.1016/j.patcog.2025.111826.

S. Alireza Golestaneh and Lina J. Karam. Spatially-varying blur detection based on multiscale fused and sorted transform coefficients of gradient magnitudes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 596–605, 2017. URL https://doi.org/10.1109/cvpr.2017.71.

Jing Hao, Zhixin Zhang, Shicai Yang, Di Xie, and Shiliang Pu. TransForensics: Image forgery localization with dense self-attention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15035–15044, 2021. URL https://doi.org/10.1109/iccv48922.2021.01478.

Chunming He, Kai Li, Yachao Zhang, Guoxia Xu, Longxiang Tang, Yulun Zhang, Zhenhua Guo, and Xiu Li. Weakly-supervised concealed object segmentation with SAM-based pseudo labeling and multi-scale feature grouping. In *Advances in Neural Information Processing Systems*, volume 36, pp. 30726–30737, 2023. URL https://doi.org/10.52202/075280-1339.

Xuefeng Hu, Zhihan Zhang, Zhenye Jiang, Syomantak Chaudhuri, Zhenheng Yang, and Ram Nevatia. SPAN: Spatial pyramid attention network for image manipulation localization. In *European Conference on Computer Vision*, pp. 312–328, 2020. URL https://doi.org/10.1007/978-3-030-58589-1_19.

Xihang Hu, Fuming Sun, Jiazhe Liu, Feilong Xu, and Xiaoli Zhang. ST-SAM: SAM-driven self-training framework for semi-supervised camouflaged object detection. In *Proceedings of the ACM International Conference on Multimedia*, pp. 8194–8203, 2025. URL https://doi.org/10.1145/3746027.3755355.

Kaixiang Ji, Feng Chen, Xin Guo, Yadong Xu, Jian Wang, and Jingdong Chen. Uncertainty-guided learning for improving image manipulation detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22399–22408, 2023. URL https://doi.org/10.1109/iccv51070.2023.02052.

Qi Jia, Shuilian Yao, Yu Liu, Xin Fan, Risheng Liu, and Zhongxuan Luo. Segment, magnify and reiterate: Detecting camouflaged objects the hard way. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4703–4712, 2022. URL https://doi.org/10.1109/cvpr52688.2022.00467.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment Anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3992–4003, 2023. URL https://doi.org/10.1109/iccv51070.2023.00371.

Trung-Nghia Le, Tam V. Nguyen, Zhongliang Nie, Minh-Triet Tran, and Akihiro Sugimoto. Anabranch network for camouflaged object segmentation. *Computer Vision and Image Understanding*, 184:45–56, 2019. URL https://doi.org/10.1016/j.cviu.2019.04.006.

Min Seok Lee, WooSeok Shin, and Sung Won Han. TRACER: Extreme attention guided salient object tracing network (Student Abstract). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, number 11, pp. 12993–12994, 2022. URL https://doi.org/10.1609/aaai.v36i11.21633.

Guanbin Li and Yizhou Yu. Visual saliency based on multiscale deep features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5455–5463, 2015. URL https://doi.org/10.1109/cvpr.2015.7299184.

Yin Li, Xiaodi Hou, Christof Koch, James M. Rehg, and Alan L. Yuille. The secrets of salient object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 280–287, 2014. URL https://doi.org/10.1109/cvpr.2014.43.

Feiran Li, Qianqian Xu, Shilong Bao, Zhiyong Yang, Runmin Cong, Xiaochun Cao, and Qingming Huang. Size-invariance matters: Rethinking metrics and losses for imbalanced multi-object salient object detection. In *Proceedings of the International Conference on Machine Learning*, volume 235, pp. 28989–29021, 2024. URL https://proceedings.mlr.press/v235/li24bx.html.

Jiang-Jiang Liu, Qibin Hou, and Ming-Ming Cheng. Dynamic feature integration for simultaneous detection of salient object, edge, and skeleton. *IEEE Transactions on Image Processing*, 29:8652–8667, 2020. URL https://doi.org/10.1109/tip.2020.3017352.

Nian Liu, Ni Zhang, Kaiyuan Wan, Ling Shao, and Junwei Han. Visual saliency transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4702–4712, 2021. URL https://doi.org/10.1109/iccv48922.2021.00468.

Xiaohong Liu, Yaojie Liu, Jun Chen, and Xiaoming Liu. PSCC-Net: Progressive spatio-channel correlation network for image manipulation detection and localization. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(11):7505–7517, 2022. URL https://doi.org/10.1109/tcsvt.2022.3189545.

Weihuang Liu, Xi Shen, Chi-Man Pun, and Xiaodong Cun. Explicit visual prompting for low-level structure segmentations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19434–19445, 2023. URL https://doi.org/10.1109/cvpr52729.2023.01862.

Zijian Liu, Ping Jiang, Lixin Lin, and Xiaoheng Deng. Edge attention learning for efficient camouflaged object detection. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 5230–5234, 2024. URL https://doi.org/10.1109/icassp48485.2024.10448139.

Weihuang Liu, Xi Shen, Chi-Man Pun, and Xiaodong Cun. Explicit visual prompting for universal foreground segmentations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(2):1762–1777, 2026. URL https://doi.org/10.1109/tpami.2025.3619490.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1711.05101.

Ziyang Luo, Nian Liu, Wangbo Zhao, Xuguang Yang, Dingwen Zhang, Deng-Ping Fan, Fahad Shahbaz Khan, and Junwei Han. VSCode: General visual salient and camouflaged object detection with 2D prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17169–17180, 2024. URL https://doi.org/10.1109/cvpr52733.2024.01625.

Ziyang Luo, Nian Liu, Xuguang Yang, Dingwen Zhang, Deng-Ping Fan, Fahad Shahbaz Khan, and Junwei Han. VSCode-v2: Dynamic prompt learning for general visual salient and camouflaged object detection with two-stage optimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(3):3137–3153, 2026. URL https://doi.org/10.1109/tpami.2025.3635136.

Yunqiu Lv, Jing Zhang, Yuchao Dai, Aixuan Li, Bowen Liu, Nick Barnes, and Deng-Ping Fan. Simultaneously localize, segment and rank the camouflaged objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11586–11596, 2021. URL https://doi.org/10.1109/cvpr46437.2021.01142.

Mingcan Ma, Changqun Xia, Chenxi Xie, Xiaowu Chen, and Jia Li. Boosting broader receptive fields for salient object detection. *IEEE Transactions on Image Processing*, 32:1026–1038, 2023. URL https://doi.org/10.1109/tip.2022.3232209.

Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15(1):654, 2024. URL https://doi.org/10.1038/s41467-024-44824-z.

Ran Margolin, Lihi Zelnik-Manor, and Ayellet Tal. How to evaluate foreground maps? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2014. URL https://doi.org/10.1109/cvpr.2014.39.

Haiyang Mei, Ge-Peng Ji, Ziqi Wei, Xin Yang, Xiaopeng Wei, and Deng-Ping Fan. Camouflaged object segmentation with distraction mining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8768–8777, 2021. URL https://doi.org/10.1109/cvpr46437.2021.00866.

Morteza Moradi, Mohammad Moradi, Simone Palazzo, Ali Borji, and Concetto Spampinato. Global-local feature decoding with adapter-guided SAMv2 for salient object detection. *arXiv preprint arXiv:2605.02616*, 2026. URL https://doi.org/10.48550/arXiv.2605.02616.

Jingchen Ni, Quan Zhang, Dan Jiang, Keyu Lv, Ke Zhang, and Chun Yuan. FCL-COD: Weakly supervised camouflaged object detection with frequency-aware and contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pp. 7439–7449, 2026. URL https://openaccess.thecvf.com/content/CVPR2026F/html/Ni_FCL-COD_Weakly_Supervised_Camouflaged_Object_Detection_with_Frequency-aware_and_Contrastive_CVPRF_2026_paper.html.

Adam Novozamsky, Babak Mahdian, and Stanislav Saic. IMD2020: A large-scale annotated dataset tailored for detecting manipulated images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops*, pp. 71–80, 2020. URL https://doi.org/10.1109/wacvw50321.2020.9096940.

Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Video object segmentation using space-time memory networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9225–9234, 2019. URL https://doi.org/10.1109/iccv.2019.00932.

Youwei Pang, Xiaoqi Zhao, Lihe Zhang, and Huchuan Lu. Multi-scale interactive network for salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9410–9419, 2020. URL https://doi.org/10.1109/cvpr42600.2020.00943.

Youwei Pang, Xiaoqi Zhao, Tian-Zhu Xiang, Lihe Zhang, and Huchuan Lu. 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*, pp. 2150–2160, 2022. URL https://doi.org/10.1109/cvpr52688.2022.00220.

Youwei Pang, Xiaoqi Zhao, Tian-Zhu Xiang, Lihe Zhang, and Huchuan Lu. ZoomNeXt: A unified collaborative pyramid network for camouflaged object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):9205–9220, 2024. URL https://doi.org/10.1109/tpami.2024.3417329.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. SAM 2: Segment anything in images and videos. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2408.00714.

Chaitanya Ryali, Yuan-Ting Hu, Daniel Bolya, Chen Wei, Haoqi Fan, Po-Yao Huang, Vaibhav Aggarwal, Arkabandhu Chowdhury, Omid Poursaeed, Judy Hoffman, Jitendra Malik, Yanghao Li, and Christoph Feichtenhofer. Hiera: A hierarchical vision transformer without the bells-and-whistles. In *Proceedings of the International Conference on Machine Learning*, volume 202, pp. 29441–29454, 2023. URL https://proceedings.mlr.press/v202/ryali23a.html.

Jianping Shi, Li Xu, and Jiaya Jia. Discriminative blur detection features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2965–2972, 2014. URL https://doi.org/10.1109/cvpr.2014.379.

Jianping Shi, Qiong Yan, Li Xu, and Jiaya Jia. Hierarchical image saliency detection on extended CSSD. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(4):717–729, 2016. URL https://doi.org/10.1109/tpami.2015.2465960.

Przemysław Skurowski, Hassan Abdulameer, Jakub Błaszczyk, Tomasz Depta, Adam Kornacki, and Przemysław Kozieł. Animal camouflage analysis: CHAMELEON database. Silesian University of Technology, dataset webpage, 2017. URL https://www.polsl.pl/rau6/chameleon-database-animal-camouflage-analysis/.

Xiaogang Song, Kaixuan Yang, Peirui Li, Haoyu Yuan, Qin Zhao, and Xinhong Hei. Camouflaged object detection via deep assistance sparsely weighted attention. *Computer Vision and Image Understanding*, 271:104877, 2026. URL https://doi.org/10.1016/j.cviu.2026.104877.

Ke Sun, Zhongxi Chen, Xianming Lin, Xiaoshuai Sun, Hong Liu, and Rongrong Ji. Conditional diffusion models for camouflaged and salient object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(4):2833–2848, 2025. URL https://doi.org/10.1109/tpami.2025.3527469.

Chang Tang, Xinzhong Zhu, Xinwang Liu, Lizhe Wang, and Albert Y. Zomaya. DeFusionNET: Defocus blur detection via recurrently fusing and refining multi-scale deep features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2695–2704, 2019. URL https://doi.org/10.1109/cvpr.2019.00281.

Lijun Wang, Huchuan Lu, Yifan Wang, Mengyang Feng, Dong Wang, Baocai Yin, and Xiang Ruan. Learning to detect salient objects with image-level supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3796–3805, 2017. URL https://doi.org/10.1109/cvpr.2017.404.

Junke Wang, Zuxuan Wu, Jingjing Chen, Xintong Han, Abhinav Shrivastava, Ser-Nam Lim, and Yu-Gang Jiang. ObjectFormer for image manipulation detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2354–2363, 2022. URL https://doi.org/10.1109/cvpr52688.2022.00240.

Yi Wang, Ruili Wang, Xin Fan, Tianzhu Wang, and Xiangjian He. Pixels, regions, and objects: Multiple enhancement for salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10031–10040, 2023. URL https://doi.org/10.1109/cvpr52729.2023.00967.

Hui Wang, Yuqian Zhao, Fan Zhang, Gui Gui, Lingli Yu, Baifan Chen, Miao Liao, Chunhua Yang, and Weihua Gui. R-Net: Recursive decoder with edge refinement network for salient object detection. *Expert Systems with Applications*, 261:125562, 2025. URL https://doi.org/10.1016/j.eswa.2024.125562.

Jun Wei, Shuhui Wang, Zhe Wu, Chi Su, Qingming Huang, and Qi Tian. Label decoupling framework for salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13022–13031, 2020. URL https://doi.org/10.1109/cvpr42600.2020.01304.

Jun Wei, Shuhui Wang, and Qingming Huang. F3Net: Fusion, feedback and focus for salient object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 12321–12328, 2020. URL https://doi.org/10.1609/aaai.v34i07.6916.

Zhe Wu, Li Su, and Qingming Huang. Cascaded partial decoder for fast and accurate salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3902–3911, 2019. URL https://doi.org/10.1109/cvpr.2019.00403.

Yue Wu, Wael AbdAlmageed, and Premkumar Natarajan. ManTra-Net: Manipulation tracing network for detection and localization of image forgeries with anomalous features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9535–9544, 2019. URL https://doi.org/10.1109/cvpr.2019.00977.

Yu-Huan Wu, Yun Liu, Le Zhang, Ming-Ming Cheng, and Bo Ren. EDN: Salient object detection via extremely-downsampled network. *IEEE Transactions on Image Processing*, 31:3125–3136, 2022. URL https://doi.org/10.1109/tip.2022.3164550.

Zongwei Wu, Danda Pani Paudel, Deng-Ping Fan, Jingjing Wang, Shuo Wang, Cédric Demonceaux, Radu Timofte, and Luc Van Gool. Source-free depth for object pop-out. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1032–1042, 2023. URL https://doi.org/10.1109/iccv51070.2023.00101.

Ranwan Wu, Tian-Zhu Xiang, Guo-Sen Xie, Rongrong Gao, Xiangbo Shu, Fang Zhao, and Ling Shao. Uncertainty-aware transformer for referring camouflaged object detection. *IEEE Transactions on Image Processing*, 34:5341–5354, 2025. URL https://doi.org/10.1109/tip.2025.3587579.

Zihuang Wu, Xinyu Xiong, Guangwei Gao, Hongwei Li, and Hua Chen. HFS-SAM2: Segment anything model 2 with high-frequency feature supplementation for camouflaged object detection. *IEEE Signal Processing Letters*, 32:2923–2927, 2025. URL https://doi.org/10.1109/lsp.2025.3589944.

Yu-Huan Wu, Zi-Xuan Zhu, Yan Wang, Liangli Zhen, and Deng-Ping Fan. RefOnce: Distilling references into a prototype memory for referring camouflaged object detection. *arXiv preprint arXiv:2511.20989*, 2025. URL https://doi.org/10.48550/arXiv.2511.20989.

Xinyu Xiong, Zihuang Wu, Shuangyi Tan, Wenxue Li, Feilong Tang, Ying Chen, Siying Li, Jie Ma, and Guanbin Li. SAM2-UNet: Segment Anything 2 Makes Strong Encoder for Natural and Medical Image Segmentation. *Visual Intelligence*, 4:2, 2026. URL https://doi.org/10.1007/s44267-025-00106-w.

Chuan Yang, Lihe Zhang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. Saliency detection via graph-based manifold ranking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3166–3173, 2013. URL https://doi.org/10.1109/cvpr.2013.407.

Fan Yang, Qiang Zhai, Xin Li, Rui Huang, Ao Luo, Hong Cheng, and Deng-Ping Fan. Uncertainty-guided transformer reasoning for camouflaged object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4126–4135, 2021. URL https://doi.org/10.1109/iccv48922.2021.00411.

Xin Yi and Mark Eramian. LBP-based segmentation of defocus blur. *IEEE Transactions on Image Processing*, 25(4):1626–1638, 2016. URL https://doi.org/10.1109/tip.2016.2528042.

Yugen Yi, Ningyi Zhang, Wei Zhou, Yanjiao Shi, Gengsheng Xie, and Jianzhong Wang. GPONet: A two-stream gated progressive optimization network for salient object detection. *Pattern Recognition*, 150:110330, 2024. URL https://doi.org/10.1016/j.patcog.2024.110330.

Bowen Yin, Xuying Zhang, Deng-Ping Fan, Shaohui Jiao, Ming-Ming Cheng, Luc Van Gool, and Qibin Hou. CamoFormer: Masked separable attention for camouflaged object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):10362–10374, 2024. URL https://doi.org/10.1109/tpami.2024.3438565.

Zhenni Yu, Xiaoqin Zhang, Li Zhao, Yi Bin, and Guobao Xiao. Exploring deeper! Segment anything model with depth perception for camouflaged object detection. In *Proceedings of the ACM International Conference on Multimedia*, pp. 4322–4330, 2024. URL https://doi.org/10.1145/3664647.3681119.

Zhenni Yu, Li Zhao, Guobao Xiao, and Xiaoqin Zhang. SAM-TTT: Segment Anything Model via reverse parameter configuration and test-time training for camouflaged object detection. In *Proceedings of the ACM International Conference on Multimedia*, pp. 4030–4038, 2025. URL https://doi.org/10.1145/3746027.3755291.

Yuhui Yuan, Xilin Chen, and Jingdong Wang. Object-contextual representations for semantic segmentation. In *European Conference on Computer Vision*, pp. 173–190, 2020. URL https://doi.org/10.1007/978-3-030-58539-6_11.

Yi Ke Yun and Weisi Lin. Towards a complete and detail-preserved salient object detection. *IEEE Transactions on Multimedia*, 26:4667–4680, 2024. URL https://doi.org/10.1109/tmm.2023.3325731.

Qiao Zhang, Yanliang Ge, Cong Zhang, and Hongbo Bi. TPRNet: Camouflaged object detection via transformer-induced progressive refinement network. *The Visual Computer*, 39:4593–4607, 2023. URL https://doi.org/10.1007/s00371-022-02611-1.

Chenxi Zhang, Jiayun Wu, Qing Zhang, Yazhe Zhai, and Youwei Pang. Retrospective memory for camouflaged object detection. *arXiv preprint arXiv:2506.15244*, 2025. URL https://doi.org/10.48550/arXiv.2506.15244.

Dongdong Zhang, Chunping Wang, Qiang Fu, and Yao Song. SAM-guided depth-aware weakly supervised camouflaged object detection with spatial-frequency exploration. *Knowledge-Based Systems*, 336:115330, 2026. URL https://doi.org/10.1016/j.knosys.2026.115330.

Wenda Zhao, Fan Zhao, Dong Wang, and Huchuan Lu. Defocus blur detection via multi-stream bottom-top-bottom fully convolutional network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3080–3088, 2018. URL https://doi.org/10.1109/cvpr.2018.00325.

Wenda Zhao, Bowen Zheng, Qiuhua Lin, and Huchuan Lu. Enhancing diversity of defocus blur detectors via cross-ensemble network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8897–8905, 2019. URL https://doi.org/10.1109/cvpr.2019.00911.

Wenda Zhao, Fan Zhao, Dong Wang, and Huchuan Lu. Defocus blur detection via multi-stream bottom-top-bottom network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(8):1884–1897, 2020. URL https://doi.org/10.1109/tpami.2019.2906588.

Xiaoqi Zhao, Youwei Pang, Lihe Zhang, Huchuan Lu, and Lei Zhang. Suppress and balance: A simple gated network for salient object detection. In *European Conference on Computer Vision*, pp. 35–51, 2020. URL https://doi.org/10.1007/978-3-030-58536-5_3.

Wenda Zhao, Cai Shang, and Huchuan Lu. Self-generated defocus blur detection via dual adversarial discriminators. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6929–6938, 2021. URL https://doi.org/10.1109/cvpr46437.2021.00686.

Wenda Zhao, Xueqing Hou, You He, and Huchuan Lu. Defocus blur detection via boosting diversity of deep ensemble networks. *IEEE Transactions on Image Processing*, 30:5426–5438, 2021. URL https://doi.org/10.1109/tip.2021.3084101.

Wenda Zhao, Fei Wei, Haipeng Wang, You He, and Huchuan Lu. Full-scene defocus blur detection with DeFBD+ via multi-level distillation learning. *IEEE Transactions on Multimedia*, 25:9228–9240, 2023. URL https://doi.org/10.1109/tmm.2023.3248162.

Yijie Zhong, Bo Li, Lv Tang, Senyun Kuang, Shuang Wu, and Shouhong Ding. Detecting camouflaged object in frequency domain. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4494–4503, 2022. URL https://doi.org/10.1109/cvpr52688.2022.00446.

Jizhe Zhou, Xiaochen Ma, Xia Du, Ahmed Y. Alhammadi, and Wentao Feng. Pre-training-free image manipulation localization through non-mutually exclusive contrastive learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22289–22299, 2023. URL https://doi.org/10.1109/iccv51070.2023.02042.

Yuli Zhou, Yawei Li, Yuqian Fu, Luca Benini, Ender Konukoglu, and Guolei Sun. CamSAM2: Segment anything accurately in camouflaged videos. In *Advances in Neural Information Processing Systems*, volume 38, pp. 61752–61780, 2025. URL https://doi.org/10.52202/085713-1855.

Yuli Zhou, Guolei Sun, Yawei Li, Guo-Sen Xie, Luca Benini, and Ender Konukoglu. When SAM2 meets video camouflaged object segmentation: A comprehensive evaluation and adaptation. *Visual Intelligence*, 3:10, 2025. URL https://doi.org/10.1007/s44267-025-00082-1.

Jiayi Zhu, Xuebin Qin, and Abdulmotaleb Elsaddik. DC-Net: Divide-and-conquer for salient object detection. *Pattern Recognition*, 157:110903, 2025. URL https://doi.org/10.1016/j.patcog.2024.110903.

Mingchen Zhuge, Deng-Ping Fan, Nian Liu, Dingwen Zhang, Dong Xu, and Ling Shao. Salient object detection via integrity learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3738–3752, 2023. URL https://doi.org/10.1109/tpami.2022.3179526.

APPENDIX

A ADDITIONAL QUANTITATIVE ANALYSES

Curve-based Analysis. Fig. 4 presents the F_β-threshold and precision–recall (PR) curves on CAMO, CHAMELEON, COD10K, and NC4K. The comparison covers representative conventional COD methods as well as recent SAM- and SAM2-based approaches, providing a broad evaluation across different model families. CamMem remains at or near the upper envelope of the F_β curves over a broad range of thresholds, indicating robust performance under different decision thresholds. This advantage is especially evident on COD10K and NC4K, where CamMem maintains a clear margin over most competing methods.

The PR curves show a similar trend, with CamMem maintaining higher precision over a broad recall range across all four benchmarks. In the high-recall regime, its precision decreases more slowly than most competing methods, indicating a better balance between object completeness and false-positive suppression. These observations are consistent with the intended effects of structured camouflage-state memory replay and conflict-guided correction in preserving reliable object evidence and reducing ambiguous responses.

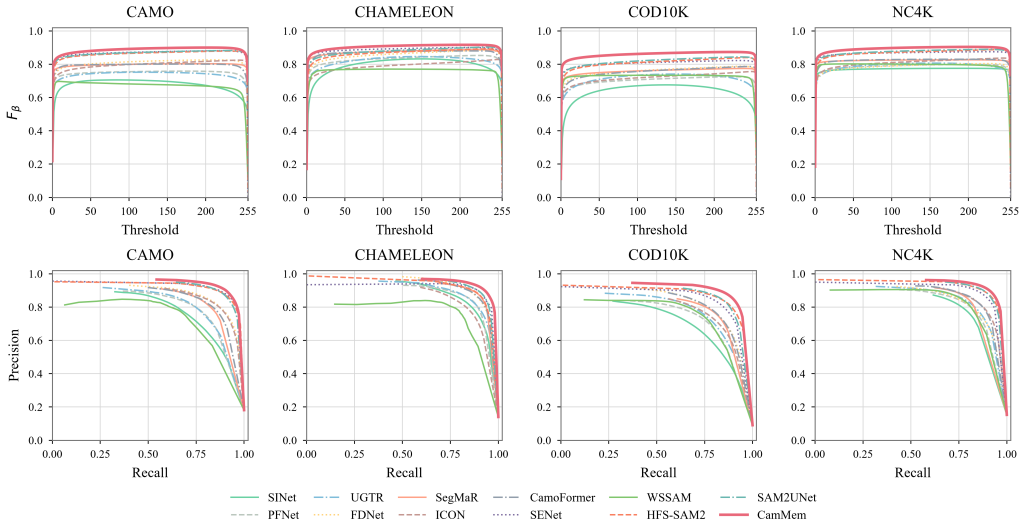


Figure 4: Comparison of F_β-threshold and precision–recall (PR) curves on CAMO, CHAMELEON, COD10K, and NC4K. The top and bottom rows show the F_β-threshold and PR curves, respectively.

Distribution-based Analysis. To examine whether the gains of CamMem are broadly distributed across test samples, Fig. 5 visualizes the per-image metric distributions of twelve representative conventional and SAM/SAM2-based COD methods on CAMO, CHAMELEON, COD10K, and NC4K. Each violin summarizes the sample-wise distribution of MAE, F_β^w, S_α, and E_m, where the white marker and inner vertical segment denote the median and interquartile range, respectively. Across the four benchmarks, CamMem generally concentrates its distributions toward the favorable regions, with lower MAE and higher F_β^w, S_α, and E_m. In particular, its F_β^w and S_α distributions remain concentrated near the upper performance range, while several competing methods exhibit heavier lower-score tails.

A similar trend is observed for MAE, where CamMem places more probability mass near zero, indicating fewer high-error samples. Although the E_m distributions of recent methods are relatively concentrated near the upper bound, CamMem still maintains consistently competitive distributions across all four datasets. These observations provide complementary evidence beyond dataset-level averages, suggesting that the improvements of CamMem are broadly shared across test samples rather than being driven by only a small subset of favorable cases.

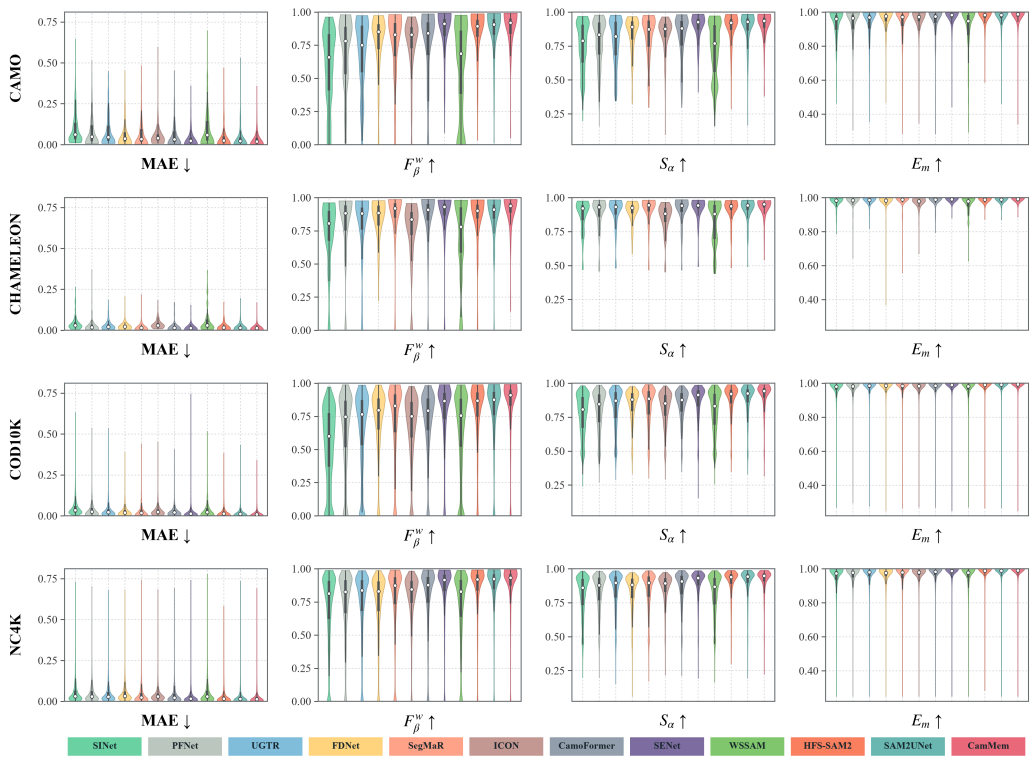


Figure 5: Per-image performance distributions of representative COD methods on CAMO, CHAMELEON, COD10K, and NC4K. The white marker and inner vertical segment denote the median and interquartile range, respectively. Lower MAE and higher F_β^w, S_α, and E_m indicate better performance.

Paired Per-image Analysis. To complement dataset-level evaluation, we further compare CamMem with SAM2-UNet, HFS-SAM2, and CamoFormer-R on exactly matched test images. Fig. 6 reports the paired differences in F_β^w and MAE, where positive values indicate improvements by CamMem. Across all dataset–method pairs, the distributions generally favor CamMem, with positive median differences for both metrics. The proportion of improved images ranges from 63.4% to 92.8% for F_β^w and from 60.4% to 92.8% for MAE reduction, showing that the aggregate gains are broadly distributed across individual test samples.

The advantage is particularly evident on COD10K, where CamMem improves F_β^w over SAM2-UNet, HFS-SAM2, and CamoFormer-R on 82.0%, 84.5%, and 92.8% of images, respectively. The corresponding proportions for MAE reduction are 81.4%, 84.5%, and 92.8%. On CAMO and NC4K, the paired distributions exhibit wider spreads and more pronounced negative tails, indicating greater sample-level variation; nevertheless, the medians remain positive and most images benefit from CamMem. These paired results provide complementary evidence that the improvements of CamMem are not driven by a small subset of favorable samples, while the remaining negative differences also reveal room for improvement on particularly difficult cases.

B Extended Ablation Analysis

Performance and component analysis. Fig. 7 summarizes the overall performance and component-wise analyses of CamMem. As shown in Fig. 7(A), CamMem achieves the highest S_m scores of 0.901, 0.927, 0.902, and 0.912 on CAMO, CHAMELEON, COD10K, and NC4K, respectively, demonstrating consistent performance across different COD benchmarks. Fig. 7(B) shows that progressively introducing CSRA, SCSM, and CGAC improves S_m from 0.886/0.892 to 0.912/0.927 on NC4K/CHAMELEON. The memory-state analysis in Fig. 7(C) further shows that

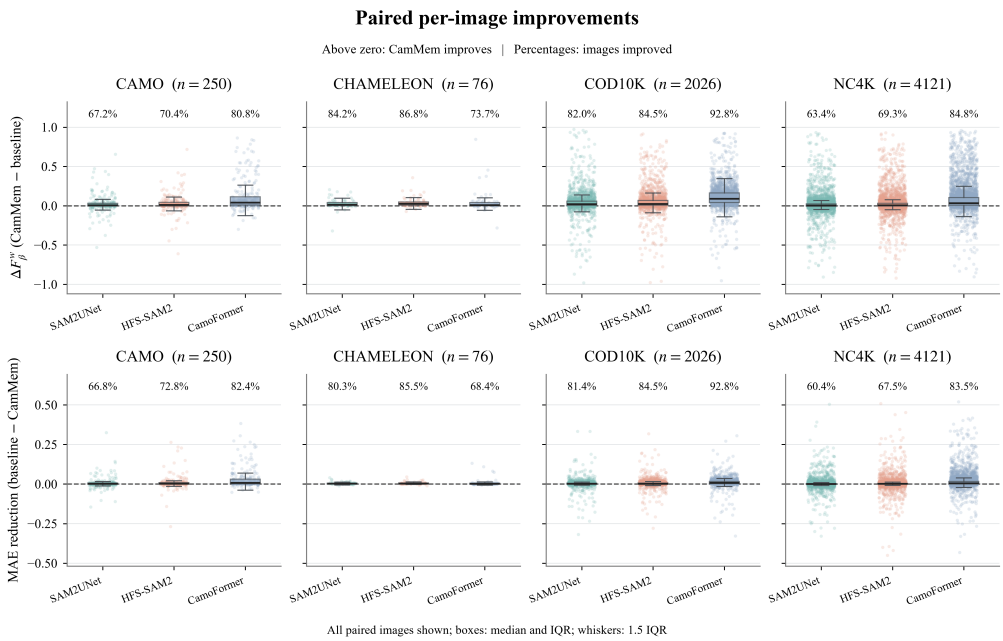


Figure 6: Paired per-image comparison of CamMem with three representative methods. The top row shows $F_\beta^w(\mathrm{CamMem}) - F_\beta^w(\mathrm{reference})$, while the bottom row shows $\mathrm{MAE}(\mathrm{reference}) - \mathrm{MAE}(\mathrm{CamMem})$. Positive values indicate improvements by CamMem. Points denote individual images, boxes show the median and interquartile range, and percentages indicate the proportion of images with positive differences.

the complete five-state SCSM achieves the highest S_m on both datasets. Fig. 7(D) confirms the complementary roles of relation calibration, evidence-directed replay, and spatial–channel gating in CSRA, while Fig. 7(E) verifies the contributions of the Anchor–Replay Conflict Gate, Local-Support Preference Gate, and local refinement in CGAC. Overall, these results further support the effectiveness of structured camouflage-state memory construction, selective replay, and conflict-guided anchor correction.

Qualitative Ablation Analysis. Fig. 8 visualizes the progressive effects of the proposed components, where green, red, and blue denote true-positive (TP), false-negative (FN), and false-positive (FP) regions, respectively. With CSRA alone, noticeable FN and FP errors remain in ambiguous structures and distracting regions. Incorporating SCSM substantially reduces these errors by providing structured foreground, background, boundary, reliable, and conflict evidence for camouflage-state replay. The full CamMem further corrects residual inconsistencies through CGAC, yielding more complete object regions and cleaner boundaries. The progressive reduction of FN and FP regions visually demonstrates the complementary effects of structured camouflage-state memory, state replay, and conflict-guided anchor correction.

C Cross-Task Generalization Evaluation

C.1 Salient Object Detection

Datasets and compared methods. We evaluate CamMem on five widely used SOD benchmarks, including DUTS-TE (Wang et al., 2017), DUT-OMRON (Yang et al., 2013), PASCAL-S (Li et al., 2014), HKU-IS (Li and Yu, 2015), and ECSSD (Shi et al., 2016). We compare CamMem with representative SOD methods, including CPD (Wu et al., 2019a), F3Net (Wei et al., 2020b), DFI (Liu et al., 2020), GateNet (Zhao et al., 2020b), MINet (Pang et al., 2020), LDF (Wei et al., 2020a), VST (Liu et al., 2021), TRACER (Lee et al., 2022), EDN (Wu et al., 2022), ICON (Zhuge et al., 2023), SelfReformer (Yun and Lin, 2024), BBRF (Ma et al., 2023), MENet (Wang et al., 2023), SI-EDN (Li et al., 2024), DASOD (Asheghi et al., 2024), Vscode-B (Luo et al., 2024), GPONet (Yi et al., 2024),

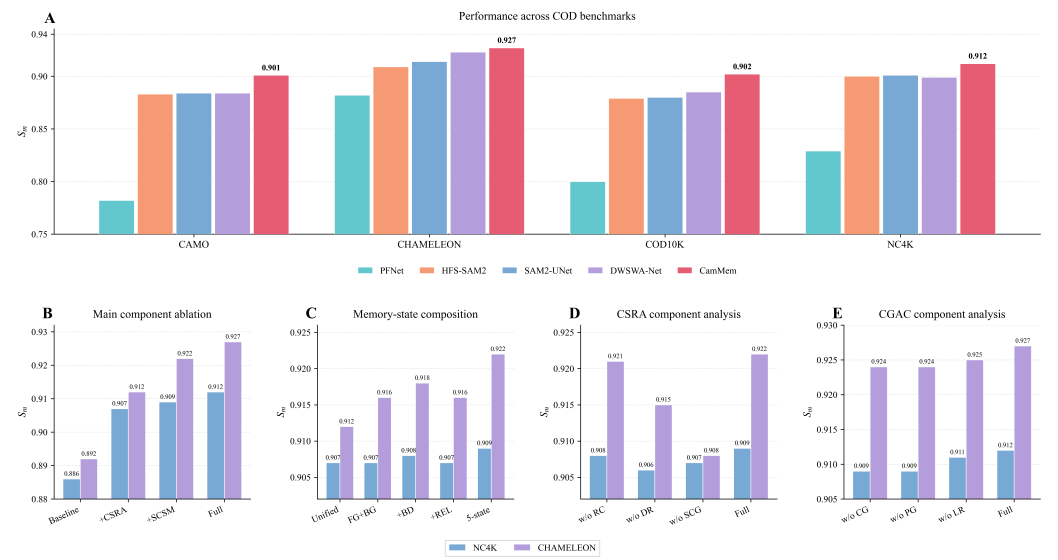


Figure 7: Extended performance and ablation analysis of CamMem. (A) S_m comparison with representative COD methods on CAMO, CHAMELEON, COD10K, and NC4K. (B) Ablation of the main CamMem components. (C) Ablation of the structured memory-state composition. (D) Ablation of CSRA components. (E) Ablation of CGAC components.

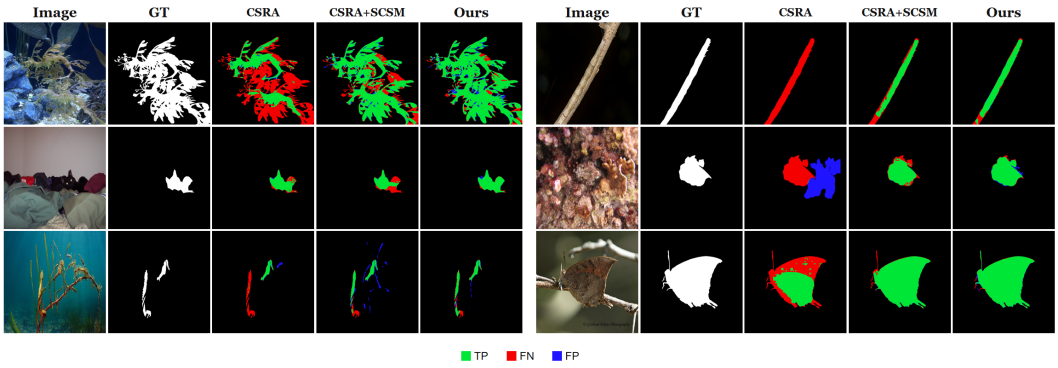


Figure 8: Qualitative ablation comparison of CamMem components. Green, red, and blue denote TP, FN, and FP regions, respectively. The progressive reduction of FN and FP errors illustrates the complementary effects of CSRA, SCSM, and CGAC.

R-Net (Wang et al., 2025), SODDCNet-L (Dulam and Kambhamettu, 2025), DC-Net (Zhu et al., 2025), and Vscode-v2-S (Luo et al., 2026). We further include SAM-based methods, including SAM (Kirillov et al., 2023), MDSAM (Gao et al., 2024), SAM2-UNet (Xiong et al., 2026), and GLASSNet (Moradi et al., 2026). Performance is evaluated using MAE (M), maximum F-measure ($F_\beta^{\max}$), S-measure (S_m), and mean E-measure (E_m).

Quantitative comparison. As shown in Table 6, CamMem demonstrates strong cross-task performance across five SOD benchmarks, achieving the best or tied-best results on 14 of the 20 evaluation entries. It performs particularly well on DUTS-TE, PASCAL-S, and HKU-IS, while remaining competitive on DUT-OMRON and ECSSD. Compared with earlier representative SOD methods, CamMem exhibits substantial improvements. Taking CPD as an example, CamMem reduces MAE by 58.1%, 28.6%, 40.8%, 50.0%, and 45.9% on DUTS-TE, DUT-OMRON, PASCAL-S, HKU-IS, and ECSSD, respectively. On DUTS-TE, it further improves $F_\beta^{\max}$, S_m, and E_m by 8.6%, 8.1%, and 7.1%, respectively, demonstrating a clear advantage over early SOD approaches.

Table 6: Quantitative comparison with state-of-the-art SOD methods. The best and second-best results are highlighted in light red and light blue, respectively.

Method	Pub./Year	DUTS-TE				DUT-OMRON				PASCAL-S				HKU-IS				ECSSD			
		$M\downarrow$	$F_\beta^{\max}\uparrow$	$S_m\uparrow$	$E_m\uparrow$	$M\downarrow$	$F_\beta^{\max}\uparrow$	$S_m\uparrow$	$E_m\uparrow$	$M\downarrow$	$F_\beta^{\max}\uparrow$	$S_m\uparrow$	$E_m\uparrow$	$M\downarrow$	$F_\beta^{\max}\uparrow$	$S_m\uparrow$	$E_m\uparrow$	$M\downarrow$	$F_\beta^{\max}\uparrow$	$S_m\uparrow$	$E_m\uparrow$
CPD	$CVPR_{19}$	0.043	0.872	0.869	0.898	0.056	0.818	0.825	0.847	0.071	0.876	0.848	0.882	0.034	0.828	0.905	0.938	0.037	0.946	0.918	0.942
F3Net	$AAAI_{20}$	0.035	0.905	0.888	0.920	0.053	0.841	0.838	0.864	0.061	0.892	0.861	0.898	0.028	0.943	0.917	0.952	0.033	0.957	0.924	0.948
DFI	TIP_{20}	0.039	0.896	0.887	0.912	0.055	0.818	0.839	0.865	0.065	0.885	0.857	0.861	0.031	0.934	0.920	0.951	0.035	0.949	0.927	0.924
GateNet	$ECCV_{20}$	0.035	0.908	0.897	0.916	0.051	0.847	0.849	0.865	0.064	0.892	0.865	0.895	0.029	0.946	0.925	0.947	0.035	0.957	0.929	0.944
MINet	$CVPR_{20}$	0.037	0.884	0.884	0.917	0.056	0.831	0.833	0.860	0.064	0.881	0.856	0.896	0.029	0.942	0.919	0.952	0.033	0.954	0.925	0.950
LDF	$CVPR_{20}$	0.034	0.905	0.892	0.925	0.052	0.835	0.839	0.865	0.060	0.887	0.863	0.903	0.028	0.943	0.919	0.953	0.034	0.956	0.924	0.948
VST	$ICCV_{21}$	0.037	0.895	0.896	0.919	0.058	0.836	0.850	0.871	0.061	0.882	0.872	0.902	0.029	0.946	0.928	0.952	0.033	0.954	0.932	0.951
TRACER	$AAAI_{22}$	0.023	0.932	0.920	0.954	0.045	0.849	0.856	0.883	0.047	0.906	0.883	0.928	0.021	0.953	0.934	0.967	0.026	0.962	0.936	0.959
EDN	TIP_{22}	0.035	0.898	0.892	0.925	0.049	0.839	0.849	0.877	0.062	0.887	0.865	0.902	-	-	-	-	-	-	-	-
ICON	$TPAMI_{23}$	0.025	0.924	0.917	0.954	0.043	0.862	0.869	0.900	0.048	0.903	0.885	0.924	0.022	0.954	0.935	0.968	0.023	0.962	0.914	0.968
SelfReformer	TMM_{23}	0.027	0.920	0.911	0.943	0.043	0.853	0.861	0.884	0.051	0.902	0.881	0.919	0.024	0.949	0.931	0.960	0.027	0.959	0.936	0.957
BBRF	TIP_{23}	0.025	0.911	0.909	0.949	0.044	0.839	0.861	0.896	0.049	0.887	0.878	0.923	0.020	0.949	0.932	0.969	0.022	0.961	0.939	0.969
MENet	$CVPR_{23}$	0.028	0.918	0.905	0.938	0.045	0.845	0.850	0.871	0.053	0.897	0.872	0.910	0.023	0.951	0.927	0.960	0.021	0.957	0.928	0.951
SAM	$ICCV_{23}$	0.030	0.921	0.909	0.937	0.044	0.865	0.869	0.896	0.061	0.897	0.866	0.902	0.022	0.956	0.935	0.965	0.025	0.968	0.944	0.964
SI-EDN	$ICML_{24}$	0.040	0.880	0.879	0.917	0.056	0.822	0.834	0.868	0.064	0.887	0.859	0.900	-	-	-	-	-	-	-	-
DASOD	IVC_{24}	0.033	–	–	–	0.052	–	–	–	0.059	–	–	–	0.024	–	–	–	0.027	–	–	–
Vscode-B	$CVPR_{24}$	0.022	0.945	0.931	0.959	0.043	0.884	0.880	0.905	0.047	0.913	0.891	0.927	0.020	0.962	0.941	0.967	0.020	0.970	0.950	0.968
MDSAM	$ACMMM_{24}$	0.024	0.937	0.920	0.949	0.039	**0.887**	0.878	0.910	0.052	0.907	0.882	0.917	0.019	0.963	0.941	0.969	0.021	**0.974**	0.948	0.967
GPONet	PR_{24}	0.027	–	0.912	0.937	0.045	–	0.865	0.889	0.055	–	0.870	0.908	0.023	–	0.936	0.962	0.025	–	0.942	0.964
R-Net	$ESWA_{25}$	0.033	0.893	0.892	0.908	0.049	0.813	0.832	0.859	0.059	0.876	0.858	0.857	0.025	0.941	0.924	0.955	0.030	0.950	0.930	0.928
SODDCNet-L	$CVPR_{25}$	0.025	0.927	0.922	0.928	0.043	0.852	0.872	0.902	0.052	0.915	0.882	0.885	0.022	0.955	0.938	0.964	0.025	0.966	0.944	0.932
DC-Net	PR_{25}	0.023	0.932	0.925	0.952	0.039	0.868	0.875	0.898	0.049	0.904	0.887	0.917	0.021	0.957	0.941	0.966	0.023	0.968	0.947	0.965
Vscode-v2-S	$TPAMI_{26}$	0.022	0.946	0.931	0.958	0.042	**0.887**	0.880	0.907	0.048	**0.916**	0.890	0.922	0.022	0.960	0.937	0.962	0.023	0.971	0.947	0.964
SAM2-UNet	$VisualIntell._{26}$	0.020	0.942	0.934	0.959	0.039	0.878	0.884	0.912	0.043	0.913	0.894	0.931	0.019	0.960	0.941	0.971	0.020	0.970	0.950	0.970
GLASSNet	$ArXiv_{26}$	0.019	0.935	0.936	0.960	**0.037**	0.860	**0.889**	**0.915**	0.045	0.904	0.893	0.931	0.019	0.955	0.943	0.971	**0.019**	0.966	**0.953**	**0.971**
Ours	–	**0.018**	**0.947**	**0.939**	**0.962**	0.040	0.875	0.883	0.908	**0.042**	**0.916**	**0.899**	**0.933**	**0.017**	**0.965**	**0.948**	**0.973**	0.020	**0.974**	**0.953**	0.969

In particular, CamMem achieves the best results on all four metrics on DUTS-TE, with $M = 0.018$, $F_\beta^{\max} = 0.947$, $S_m = 0.939$, and $E_m = 0.962$. On PASCAL-S, it obtains the best M, S_m, and E_m of 0.042, 0.899, and 0.933, respectively, together with a tied-best $F_\beta^{\max}$ of 0.916. Similar performance is observed on HKU-IS, where CamMem achieves the best results across all four metrics. On ECSSD, it obtains tied-best $F_\beta^{\max}$ and S_m scores of 0.974 and 0.953, respectively, together with a second-best M of 0.020. Although CamMem is less competitive on DUT-OMRON, its S_m of 0.883 remains only 0.001 below the second-best result. Overall, these results demonstrate strong cross-task transferability of CamMem across diverse SOD benchmarks.

Qualitative comparison. As shown in Fig. 9, CamMem generally produces more complete and accurate saliency predictions than representative SOD methods. In scenes containing thin structures or complex object shapes, several competing methods suffer from missing details or fragmented responses, whereas CamMem better preserves object integrity. For scenes with distracting backgrounds or ambiguous boundaries, CamMem also suppresses background responses more effectively and produces cleaner object contours. Moreover, for irregular objects and scenes containing multiple salient components, CamMem maintains more coherent foreground regions with fewer missing or spurious responses. These qualitative results further demonstrate the cross-task effectiveness of CamMem in preserving salient structures and distinguishing foreground objects from complex backgrounds.

C.2 Defocus Blur Detection

Datasets and compared methods. We evaluate CamMem on two widely used defocus blur detection benchmarks, CUHK (Shi et al., 2014) and DUT (Zhao et al., 2018). Following the protocol adopted by EVP (Liu et al., 2026), we use the 704-image defocus subset of CUHK, with 604 images for training and 100 for testing, while the 500-image DUT dataset is used for additional cross-dataset evaluation. We compare CamMem with representative defocus blur detection methods, including DBDF (Shi et al., 2014), LBP (Yi and Eramian, 2016), HiFST (Golestaneh and Karam, 2017), BTBNet (Zhao et al., 2018), BTBNet2 (Zhao et al., 2020a), DeFusionNet (Tang et al., 2019), CENet (Zhao et al., 2019), DAD (Zhao et al., 2021a), EFENet (Zhao et al., 2021b), MLDBD (Zhao et al., 2023), and TCRL (Chai et al., 2024). We further include EVPv1 (Liu et al., 2023) and the ResNet50, ViT-CLS, ViT-MAE, ViT-SAM, and SegFormer variants of EVPv2 (Liu et al., 2026). Performance is evaluated using F_β and mean absolute error (MAE), where higher F_β and lower MAE indicate better performance.

Quantitative comparison. As shown in Table 7, CamMem achieves the best results on all four evaluation entries. It obtains $F_\beta = 0.939$ and $M = 0.059$ on DUT, and $F_\beta = 0.960$ and $M = 0.039$

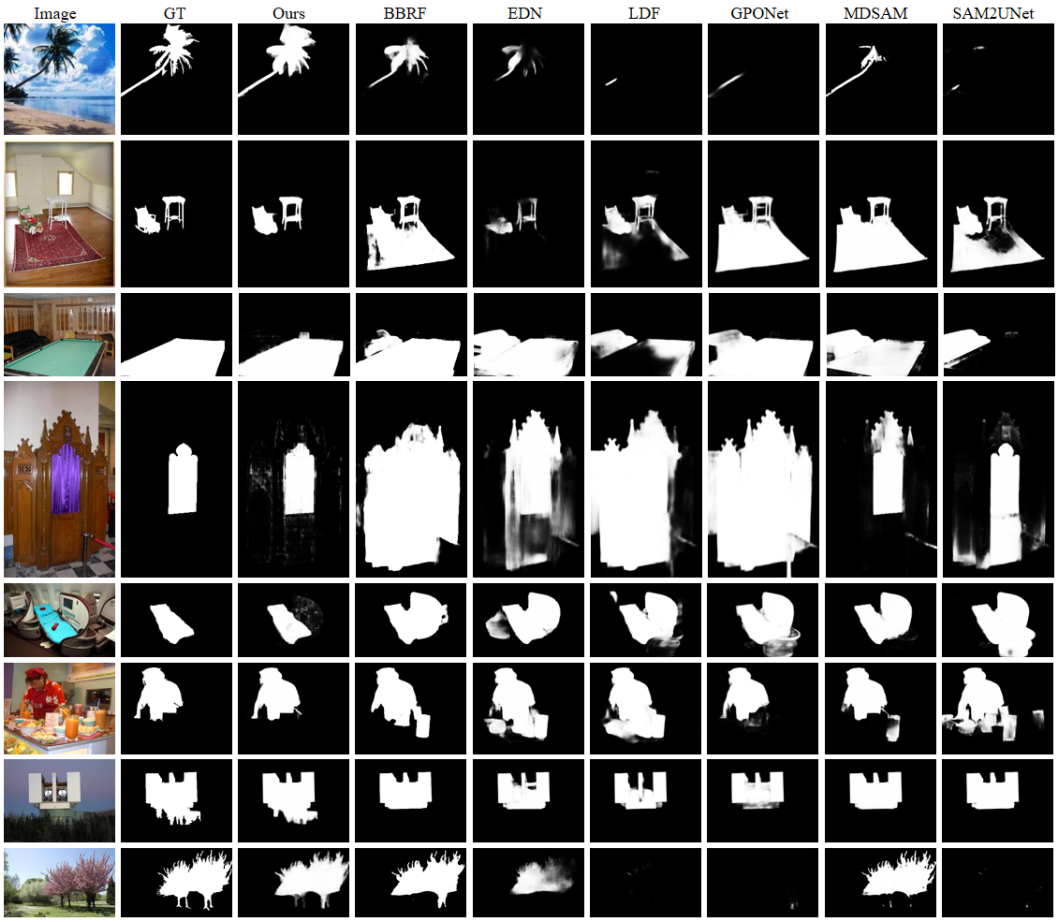


Figure 9: Qualitative comparison with representative SOD methods on diverse challenging scenes. CamMem shows improved target completeness, boundary preservation, and background suppression across different object structures and scene complexities.

on CUHK. Compared with the second-best results, CamMem improves F_β by 0.049 and 0.025 on DUT and CUHK, respectively, while reducing MAE by 0.009 and 0.003, corresponding to relative reductions of 13.2% and 7.1%.

CamMem also exhibits clear advantages over both task-specific DBD methods and unified foreground segmentation approaches. Compared with EVPv2 (ViT-SAM), for example, CamMem improves F_β from 0.866 to 0.939 and reduces MAE from 0.097 to 0.059 on DUT, while improving F_β from 0.932 to 0.960 and reducing MAE from 0.043 to 0.039 on CUHK. Notably, the DUT results are obtained by directly evaluating the model trained on CUHK, providing evidence of cross-dataset transferability. These results further indicate that the proposed structured image-specific memory mechanism can generalize beyond COD to defocus blur perception.

C.3 Forgery Detection

Datasets and compared methods. We evaluate CamMem on two widely used image forgery localization benchmarks, CASIA (Dong et al., 2013) and IMD20 (Novozamsky et al., 2020). For CamMem, we follow the evaluation setting adopted by EVP (Liu et al., 2023; 2026), using CASIA for model training and in-domain evaluation and IMD20 for additional cross-dataset evaluation. We compare CamMem with representative image manipulation localization methods, including ManTra-Net (Wu et al., 2019b), SPAN (Hu et al., 2020), PSCCNet (Liu et al., 2022), TransForensics (Hao et al., 2021), ObjectFormer (Wang et al., 2022), UEN (Ji et al., 2023), and NCL (Zhou et al., 2023), as well as the unified foreground segmentation methods EVPv1 (Liu et al., 2023) and EVPv2

Table 7: Quantitative comparison with state-of-the-art methods on defocus blur detection. $F_\beta \uparrow$ and $M \downarrow$ denote that higher and lower values are preferred, respectively.

Method	DUT		CUHK	
	$F_\beta \uparrow$	$M \downarrow$	$F_\beta \uparrow$	$M \downarrow$
DBDF	0.754	0.384	0.806	0.292
LBP	0.873	0.199	0.917	0.154
HiFST	0.839	0.313	0.803	0.223
BTBNet2	0.873	0.145	0.935	0.085
DeFusionNet	0.823	0.118	0.818	0.117
BTBNet	0.827	0.138	0.889	0.082
CENet	0.817	0.135	0.906	0.059
DAD	0.794	0.153	0.884	0.079
EFENet	0.854	0.094	0.914	0.053
MLDBD	0.875	0.111	0.921	0.049
TCRL	0.885	0.082	0.876	0.078
EVPv1 (SegFormer)	0.890	0.068	0.928	0.045
EVPv2 (ResNet50)	0.876	0.085	0.921	0.051
EVPv2 (ViT-CLS)	0.868	0.083	0.904	0.058
EVPv2 (ViT-MAE)	0.883	0.078	0.923	0.047
EVPv2 (ViT-SAM)	0.866	0.097	0.932	0.043
EVPv2 (SegFormer)	0.887	0.070	0.932	0.042
Ours	**0.939**	**0.059**	**0.960**	**0.039**

Table 8: Quantitative comparison on image forgery localization under a consistent CASIA/IMD20 evaluation protocol. Higher F1 and AUC indicate better performance.

Method	IMD20		CASIA	
	F1↑	AUC↑	F1↑	AUC↑
ManTra	–	0.748	–	0.817
SPAN	–	0.750	0.382	0.838
PSCCNet	–	0.806	0.554	0.875
TransForensics	–	0.848	0.627	0.837
ObjectFormer	–	0.821	0.579	0.882
UEN	–	–	0.629	0.877
NCL	–	–	0.598	0.864
EVPv1 (SegFormer)	0.443	0.807	0.636	0.862
EVPv2 (ResNet50)	0.382	0.756	0.557	0.798
EVPv2 (ViT-CLS)	0.371	0.755	0.471	0.784
EVPv2 (ViT-MAE)	0.404	0.772	0.619	0.858
EVPv2 (ViT-SAM)	0.418	0.792	0.614	0.842
EVPv2 (SegFormer)	**0.446**	0.811	0.654	0.876
Ours	0.429	**0.850**	**0.668**	**0.890**

(Liu et al., 2026). Performance is evaluated using pixel-level F1 score and area under the ROC curve (AUC), where higher values indicate better localization performance.

Quantitative comparison. As shown in Table 8, CamMem achieves the best results on three of the four evaluation entries across CASIA and IMD20. On CASIA, it obtains the highest F1 score of 0.668 and AUC of 0.890. On IMD20, CamMem achieves the highest AUC of 0.850, exceeding TransForensics (0.848), while obtaining an F1 score of 0.429.

Compared with EVPv2 (ViT-SAM), CamMem improves F1 and AUC from 0.614 and 0.842 to 0.668 and 0.890 on CASIA, and from 0.418 and 0.792 to 0.429 and 0.850 on IMD20, respectively. Compared with EVPv2 (SegFormer), CamMem further improves CASIA F1 from 0.654 to 0.668 and AUC from 0.876 to 0.890, while increasing IMD20 AUC from 0.811 to 0.850. These results demonstrate strong pixel-level localization performance under both in-domain and cross-dataset evaluation, further indicating that the proposed structured image-specific memory mechanism can transfer beyond COD to forgery detection.