Hierarchical Consistency Learning for Test-time Adaptation in Camouflage Perception

Mingfeng Zha, Tianyu Li, Guoqing Wang, *Member, IEEE,* Yunqiang Pei, Chaofan Qiao, Jiening Zhang Yang Yang, *Senior Member, IEEE,* and Heng Tao Shen, *Fellow, IEEE*

***Abstract*—Camouflaged object detection (COD) aims to localize targets that exhibit minimal perceptual differences from backgrounds through physical attributes. Existing methods, constrained by the static train-then-freeze paradigm, suffer from domain rigidity and annotation dependency, limiting their adaptability to scene variations and unseen camouflage patterns. To overcome these, we propose the hierarchical consistency learning (HCL) framework, which integrates test-time adaptation for dynamic representation recalibration. Specifically, we design the hierarchical representation reconstruction (HRR) to alleviate feature entanglement by synergizing spatial reconstruction with dual-stream frequency-domain decomposition, enhancing robustness against appearance homogenization. The pixel and spectrum inference provide structural and contextual priors. We further introduce task affinity guidance (TAG) to propagate knowledge across branches via channel-wise affinity, aligning local discriminative cues and mitigating semantic drift. To ensure semantic invariance, we formulate the prototype consistency calibration (PCC), which aggregates region features into compact prototypes and establishes prototype-feature similarity. This imposes implicit and hierarchical constraints that bridge task and representation gaps. Extensive experiments across four camouflaged and four underwater object benchmarks, under three degradation settings, demonstrate that our method consistently outperforms state-of-the-art approaches, highlighting its robustness and generalization under distribution shifts.**

***Index Terms*—Camouflage perception, Test-time adaptation, Consistency representation.**

I. Introduction

Existing segmentation and detection frameworks have made notable progress in perceiving salient entities, yet they struggle with concealed targets. When objects with highly similar appearances blend into surrounding contexts without clear boundaries, they may critically mislead downstream analysis, *e.g,* medical image diagnosis [1]. Camouflaged object detection (COD) focuses on identifying and localizing such hidden targets, providing critical visual cues for robust scene understanding and reliable decision-making.

COD presents fundamental challenges due to the strong visual entanglement between targets and backgrounds, *e.g,* texture mimicry and chromatic assimilation. In Figure 1 (a), traditional approaches mainly adhere to a *train-then-freeze* paradigm, where models are trained on offline datasets and deployed with fixed parameters. However, this paradigm exhibits significant limitations in open-world or degenerate scenarios (Figure 1 (b)): 1) **Domain Rigidity:** Pre-trained models lack the flexibility to adapt to diverse test-time scenes (*e.g.,* variations in illumination), resulting in feature misalignment and poor pixel-level discrimination. 2) **Annotation Dependency:** Heavy reliance on limited labeled training data restricts generalization, particularly in long-tail or evolving camouflage scenarios, where exhaustive annotation is infeasible. Test-time adaptation (TTA) [2], [3] offers a *train-then-adapt* paradigm that actively leverages intrinsic signals during inference. Rather than treating testing as a passive process, TTA enables models to dynamically recalibrate feature sensitivity, amplifying subtle yet crucial cues for camouflaged object perception. This shift allows models to self-optimize in the presence of distribution shifts. Motivated by this, we aim to formulate a task-specific TTA strategy for camouflage perception. However, seamlessly adapting existing TTA techniques to the COD task is non-trivial. Since camouflaged targets intentionally obscure boundaries and mimic background textures, directly applying standard self-supervised adaptation objectives often struggles to extract reliable discriminative representations. This raises three crucial questions regarding the conflicts between adaptation and camouflage characteristics:

How can we extract reliable self-supervised signals from heavily entangled multi-scale features? In COD, visual ambiguity arises from intrinsic multi-scale feature entanglement, where local texture similarities and global attribute alignment blur the distinction between targets and backgrounds. Existing TTA methods often rely on single-modality reconstruction in the spatial domain, attempting to differentiate targets through pixel-level autoencoding. However, this approach suffers from critical limitations: it is highly sensitive to local perturbations (*e.g.,* lighting variations), making it prone to overfitting on noise. Additionally, it decouples low-level details from high-level semantics, lacking robustness against occlusions and deformations that break spatial continuity and hinder global semantic reasoning. In camouflage scenarios, targets and backgrounds typically share predominant low-frequency features, with subtle differences hidden within high-frequency

This work was supported in part by the National Natural Science Foundation of China under grant U23B2011, 62102069, U20B2063 and 62220106008, the Key R&D Program of Zhejiang under grant 2024SSYS0091, the Sichuan Science and Technology Program under Grant 2024NSFTD0034.

Mingfeng Zha, Tianyu Li, Guoqing Wang, Yunqiang Pei, Chaofan Qiao, Jiening Zhang and Yang Yang are with the Center for Future Media and School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China. (Email: zhamf1116@gmail.com)

Heng Tao Shen is with the School of Computer Science and Technology, Tongji University, Shanghai 201804, China, with the Center for Future Media and School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China, and with the Peng Cheng Laboratory, Shenzhen 518066, China.

Corresponding author: Tianyu Li.

Project page: https://winter-flow.github.io/project/HCL

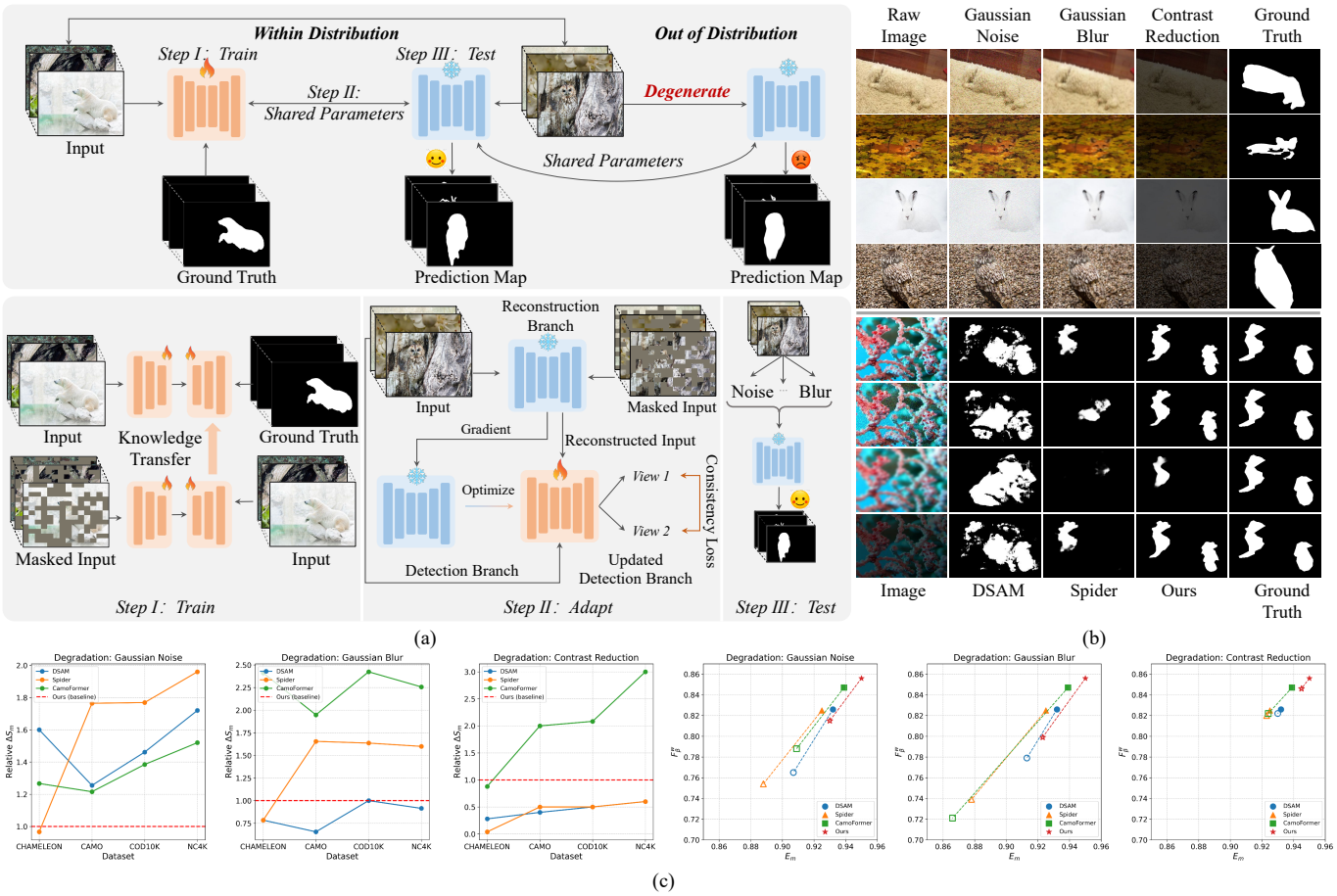

Fig. 1. (a) Traditional methods follow the *train-fix-test* paradigm (top), while our method adopts the *train-adapt-test* paradigm to dynamically perceive the scene (bottom). (b) Three testing sample degradation strategies (top) and qualitative comparisons (bottom). (c) Left: Performance drop $\triangle S_m$ of comparison methods, relative to our method. Right: Performance variation (F_β^w and E_m on the NC4K) before (solid markers) and after degradation (hollow markers). Best view by zooming in.

structural residuals. To address this, we propose a hierarchical feature disentanglement strategy through multi-spectral joint reconstruction. This method decomposes visual features in the spectrum space, suppressing high-frequency noise while modeling coarse target-background discrepancies. At the same time, it isolates and sharpens local details by enhancing edges and subtle textures, compensating for spatial ambiguity. Leveraging complementary constraints enables the model to disentangle complex visual patterns across different scales.

How can we bridge the optimization gap between self-supervised signals and main detection objectives to prevent learning drift? During dynamic adaptation, self-supervised objectives inherently focus on overall image restoration. When relying solely on reconstruction loss to adjust model parameters during the test phase, the auxiliary reconstruction branch may overfit to local restoration. Meanwhile, the main detection branch, lacking direct task-specific supervision, gradually drifts away from the discriminative representations. This leads to a severe decoupling between the extracted self-supervised signals and the ultimate detection objectives. To prevent this, we propose an explicit cross-branch knowledge-sharing mechanism. Rather than treating reconstruction and detection as independent processes, this approach transforms structural-semantic correlations learned during joint training into dynamic constraints. Concretely, structural anomalies captured by the reconstruction branch, e.g., edge discontinuity or illumination inconsistency, are leveraged through an affinity matrix to dynamically correct the local feature responses of the detection branch. By explicitly mapping these structural anomalies to refine semantic features, we ensure that the adaptation process remains firmly anchored to the primary goal of distinguishing the concealed target.

How can we enforce semantic invariance to prevent the adapted model from collapsing into local details? The discriminative core of camouflaged objects often lies in global semantic relationships rather than isolated local feature differences. Unconstrained pixel-level adaptation may introduce local noise or semantic deviations, especially when dealing with severe occlusions or heavy boundary blurring. Updating parameters based solely on local variations may lead the model to mistakenly assimilate target parts into the background. Guiding the model to ignore superficial perturbations and focus on high-level semantic invariance is therefore crucial. To achieve this, we construct a robust semantic representation space that shifts the optimization objective from pixel-level detail recovery to semantic consistency. By aggregating

features from regions into compact prototype vectors and establishing similarity alignments, this approach encourages the model to suppress local disturbances and focus on consistent semantic cues shared across samples. Dynamically balancing local reconstruction with global semantic coherence prevents the model from collapsing into noise, thereby significantly enhancing generalization to unseen camouflage patterns.

Technically, we introduce the HRR to enforce pixel-level and multi-frequency reconstruction while imposing domain consistency constraints. To further enhance knowledge transfer, we propose the TAG, which constructs channel-wise affinity maps to propagate local knowledge from the reconstruction branch to the main detection branch. To prevent the model from collapsing into trivial details, we formulate the PCC, which feeds the reconstructed image into the detection model to generate predictions. We employ entropy estimation and edge-guided mechanisms to produce confidence maps, which emphasize regions of high uncertainty and object boundaries. We then perform variational fusion of prototypes from the original and reconstructed images, followed by metric consistency computation to ensure robust and reliable predictions. In Figure 1 (c), our method yields promising results on the original benchmark and under diverse conditions.

In summary, our main contributions are as follows:

- We revisit existing frameworks for the COD task and propose a sample-specific TTA strategy to handle data distribution shifts. This strategy requires no additional data and can be seamlessly integrated into other methods.
- We introduce three customized components: the HRR, which enables automatic adaptation to diverse scenes and imaging conditions; the TAG, which serves as a constraint to guide attention toward structurally inconsistent regions and refine features; and the PCC, which facilitates selective learning of decoupled and compact representations.
- Extensive experiments on eight benchmarks and three distribution shift settings validate the superiority of the proposed method and the effect of components.

II. RELATED WORK

A. Saliency Perception

In contrast to camouflage perception, saliency perception aims to identify visually distinctive content. Based on feature selection strategies, existing methods can be broadly categorized into handcrafted and learning-based approaches. The former leverages expert prior knowledge, such as gradient or geometric cues, while the latter adopts data-driven paradigms to generate high-dimensional latent representations with greater robustness. Learning-based methods evolved beyond conventional 2D natural scenes to diverse downstream scenarios, including underwater, mirror, and panoramic environments [4]–[8], and have progressed from single-modality frameworks to multi-modal collaboration (*e.g.,* depth and thermal data). Furthermore, some works focus on modeling relative saliency differences among targets, *i.e.,* saliency ranking [9]. Since fully supervised learning generally requires large-scale, multi-scene datasets to obtain generalizable representations, recent efforts explored data-efficient alternatives such as self-supervised learning [10]. In this work, we propose to discover discriminative regions during inference by leveraging the image itself as a supervisory signal.

B. Camouflage Perception

The development of COD has progressed rapidly. Based on supervised learning signals, we categorize it into four types: 1) Fully supervised, which utilizes edge [11], texture [12], frequency [13], depth [14], uncertainty [15], or text [16] guidance, or employs a coarse-to-fine progressive approach to mine potential clues [17]; 2) Weakly supervised [18], [19], which leverages manually annotated scribbles or points as ground truth and iteratively optimizes the labels; 3) Semi-supervised [20], which explores the relationships and consistencies between labeled and unlabeled data; 4) Zero/Few-shot learning [21], [22], which aims to transfer knowledge to unseen scenarios using only a limited number of samples. Additionally, some works [23]–[26] leverage features and semantic priors from pre-trained or large language/vision/multi-modal foundation models, introducing efficient fine-tuning mechanisms that achieve competitive performance with minimal learnable parameters. Furthermore, some studies investigate at the instance level [27], video level [28], [29], and across different scenarios [30]. Our work is most closely related to [31], but it overlooks scene adaptation during the testing phase by constructing a multi-task framework for pixel reconstruction and detection that is susceptible to noise interference. Pang *et al.* [32] proposed an open-world setting based on CLIP [33], utilizing text prompts. In contrast, our HCL does not rely on additional data or foundation models to facilitate the remapping of the model to address distribution shifts.

C. Test-time Adaptation

TTA adapts model parameters during the testing phase to dynamically align with new data distributions, facilitating the transition from *i.i.d.* modeling to out-of-distribution generalization. Sun *et al.* [2] proposed adapting models via online self-supervised learning, enabling real-time updates. Wang *et al.* [34] further improved efficiency by dynamically adjusting parameters via entropy minimization. To tackle error accumulation, Niu *et al.* [3] introduced anti-forgetting mechanisms with sample selection and regularization. Jang *et al.* [35] enhanced pseudo-label reliability by leveraging prototype matching. For lifelong adaptation, Brahma *et al.* [36] designed a probabilistic framework to balance new knowledge integration with prior preservation. Zhao *et al.* [37] tackled complex distribution shifts by combining statistical correction and sample reweighting. Ma *et al.* [38] boosted robustness through graph-structured label refinement and model averaging. Recently, some works introduced TTA into specific tasks, *e.g.,* video segmentation [39]. We introduce TTA into the COD and propose the HCL framework, which incorporates tailored components to mitigate the train-test gap, particularly under significant distribution shifts. Unlike [40], which relies on multiple foundation models and adapts only at the test stage through prompting, the HCL framework integrates seamlessly into the entire pipeline without requiring manual design (*e.g.,* prompts) or prior external knowledge.

III. PROPOSED METHOD

A. Overall Architecture

In Figure 2 and Algorithm 1, given an input $\mathbf{I} \in \mathbb{R}^{3\times H\times W}$, we generate a random mask $\mathbf{M}_{\text{spa}} \in \{0,1\}^{H\times W}$ and extract multi-scale features $\mathbf{X}^i \in \mathbb{R}^{C^i\times H^i\times W^i}$ using the encoder (C: Channel; H: Height; W: Width). We apply Fourier transform $\mathbb{F}$ to project the input into the frequency domain, where mask $\mathbf{M}_l$ (and $\mathbf{M}_h$) and inverse transformation $\mathbb{F}^{-1}$ are used to obtain the masked low-frequency component $\hat{\mathbf{I}}_l$ and the masked high-frequency component $\hat{\mathbf{I}}_h$. The subsequent operations mirror those in the spatial domain. We alternately use $\mathcal{L}_{\text{pix}}$ and $\mathcal{L}_{\text{freq}}$ as reconstruction and constraint losses. The masked feature $\mathbf{X}^i_{\text{rec}}$ transfers heterogeneous information to the unmasked feature $\mathbf{X}^i$ via the TAG as a bridge. The reconstructed image $\mathbf{I}_{\text{rec}}$ is then fed as an additional input to generate the prediction map $\mathbf{O}_{\text{rec}}$ and confidence map Φ, which are further used to derive the prototype $\mathbf{P}_{\text{rec}}$. We integrate $\mathbf{P}_{\text{rec}}$ with the prototype $\mathbf{P}$ from the original input, sequentially using the fused prototype $\mathbf{P}_{\text{fusion}}$ as an anchor to establish consistency across predictions and features.

B. Preliminary

According to Parseval's Theorem, a signal x retains the same total energy in both the spatial and frequency domains, though with different emphasis. Based on the 2D Discrete Fourier Transform (DFT) $\mathbb{F}$, we transform $x \in \mathbb{R}^{H\times W}$ into the Fourier spectrum,

$$\mathcal{F}(x)(u,v) = \sum_{h=0}^{H-1}\sum_{w=0}^{W-1} x(h,w)e^{-j2\pi\left(\frac{h}{H}u+\frac{w}{W}v\right)} \tag{1}$$

where u and v represent the horizontal and vertical indices in the spectrum, respectively. To transform back to the spatial domain, we apply the inverse Fourier transform $\mathbb{F}^{-1}$,

$$x(h,w) = \frac{1}{HW}\sum_{u=0}^{H-1}\sum_{v=0}^{W-1} \mathcal{F}(u,v)e^{j2\pi\left(\frac{uh}{H}+\frac{vw}{W}\right)} \tag{2}$$

In practice, we use the Fast Fourier Transform (FFT) version to improve computational efficiency.

C. Hierarchical Representation Reconstruction

Reconstruction is not a fundamental component of camouflage perception in terms of task definition. Instead, we employ it to obtain robust and effective self-supervised learning signals. We decompose the auxiliary reconstruction task into: spatial, high-frequency, and low-frequency, each accompanied by corresponding transformation constraints.

Spatial Reconstruction. To supervise the reconstruction process, we adopt the Mean Square Error (MSE) loss $\mathcal{L}_{\text{pix}}$ on the output $\mathbf{I}_{\text{rec}}$ from the pixel decoder, encouraging accurate recovery of the spatial regions masked by $\mathbf{M}_{\text{spa}}$,

$$\mathcal{L}_{\text{pix}} = \frac{1}{HWC}\sum_{i=0}^{H-1}\sum_{j=0}^{W-1}\sum_{c=0}^{C-1}\left(\mathbf{I}^{i,j,c}_{\text{rec}} - \mathbf{I}^{i,j,c}\right)^2 \tag{3}$$

Frequency Reconstruction. To avoid inherent reconstruction defects in the pixel domain, we employ the Fourier transform to convert the raw image into the frequency domain, partitioning the spectrum into high- and low- frequency part based on the predefined threshold (the distance of spectral points from the center). For low-frequency reconstruction, we use a central mask $\mathbf{M}_l \in \{0,1\}^{H_l\times H_w}$ to randomly suppress low-frequency energy while retaining high-frequency,

$$\mathbf{M}_l(u,v) = \begin{cases} 1, & \texttt{unmasked} \\ 0, & \texttt{masked} \end{cases} \tag{4}$$

Therefore, we can obtain the degraded spectrum $\hat{\mathcal{F}}$,

$$\hat{\mathcal{F}} = \sum_{c=0}^{C-1} \mathcal{F}^c_{\text{low}} \odot \mathbf{M}^c_l + \mathcal{F}^c_{\text{high}} \tag{5}$$

where $\odot$ denotes the Hadamard product. We further utilize the inverse Fourier transform to generate the degraded image $\hat{\mathbf{I}}_l$. For high-frequency reconstruction, we perform similar operations and define the random mask $\mathbf{M}_h$, obtaining $\hat{\mathbf{I}}_h$. After generating features using the encoder, we reconstruct the high-frequency and low-frequency images from tokens, *i.e.,* $\mathbf{I}_{\text{rec(h)}}$ and $\mathbf{I}_{\text{rec(l)}}$.

In spatial visual recognition, different regions contribute unequally to decision-making. Likewise, in frequency decoding, hard samples demand greater learning attention. To address this, we introduce focal frequency loss [41] to capture high-value clues and emphasize challenging frequency components,

$$\mathcal{L}_{\text{freq}} = \frac{1}{HW}\sum_{u=0}^{H-1}\sum_{v=0}^{W-1} \omega(u,v) \odot \gamma(\mathcal{F}(u,v), \mathcal{F}_{\text{GT}}(u,v))^2 \tag{6}$$

where $\mathcal{F}(u,v)$ and $\mathcal{F}_{\text{GT}}(u,v)$ represent the components of the reconstructed spectrum $\mathbf{L}_{\text{rec}}$ (or $\mathbf{H}_{\text{rec}}$) and the raw spectrum $\mathbf{L}$ (or $\mathbf{H}$) at position (i,j), respectively. w and γ denote the weight parameter and distance metric, which can be formulated as,

$$\begin{aligned} \omega(u,v) &= \gamma(\mathcal{F}(u,v), \mathcal{F}_{\text{GT}}(u,v))^\beta \\ \gamma(\mathcal{F}, \mathcal{F}_{\text{GT}}) &= \sqrt{(\mathcal{R}-\mathcal{R}_{\text{GT}})^2 + (\mathcal{I}-\mathcal{I}_{\text{GT}})^2} \end{aligned} \tag{7}$$

where $\mathcal{R}$ and $\mathcal{I}$ denote the real and imaginary parts of the spectrum, respectively. β is a scaling factor, which is set to 1 by default.

Transformation Consistency. In theory, when $\mathcal{L}_{\text{pix}} \to 0$ or $\mathcal{L}_{\text{freq}} \to 0$, the model can accurately predict the masked regions, resulting in the reconstructed image that matches the raw image. However, $\mathcal{L}_{\text{pix}}$ directly enforces pixel-level matching, ensuring the recovery of local details but is insensitive to global structures, *e.g.,* edge continuity. Conversely, $\mathcal{L}_{\text{freq}}$ enhances the perception of global structures but is less sensitive to local pixel shifts. In other words, a sole focus on pixel matching may lead to blurriness, while an emphasis on spectral matching may introduce artifacts. Therefore, for each reconstruction branch, we apply the corresponding domain loss as a regularization term to achieve transformation consistency. We can formulate the total loss $\mathcal{L}_{\text{HRR}}$ for the HRR,

$$\begin{aligned} \mathcal{L}_{\text{HRR}} = \mathcal{L}_{\text{pix(rec)}}(\mathbf{I}_{\text{rec}}, \mathbf{I}) + \mathcal{L}_{\text{freq(con)}}(\mathcal{F}, \mathcal{F}_{\text{GT}}) + \\ \sum_{k\in\{\mathbf{L},\mathbf{H}\}} \lambda_k \left(\mathcal{L}_{\text{freq(rec)}}(k_{\text{rec}}, k) + \mathcal{L}_{\text{pix(con)}}(\mathbb{F}^{-1}(k), \mathbf{I})\right) \end{aligned} \tag{8}$$

where $\mathcal{L}_{\sim\text{(rec)}}$ and $\mathcal{L}_{\sim\text{(con)}}$ represent the reconstruction and consistency (constraint) losses, respectively. λ is the balancing hyperparameter. We empirically set $\lambda_{\mathbf{L}}$ and $\lambda_{\mathbf{H}}$ to 0.4 and 0.6.

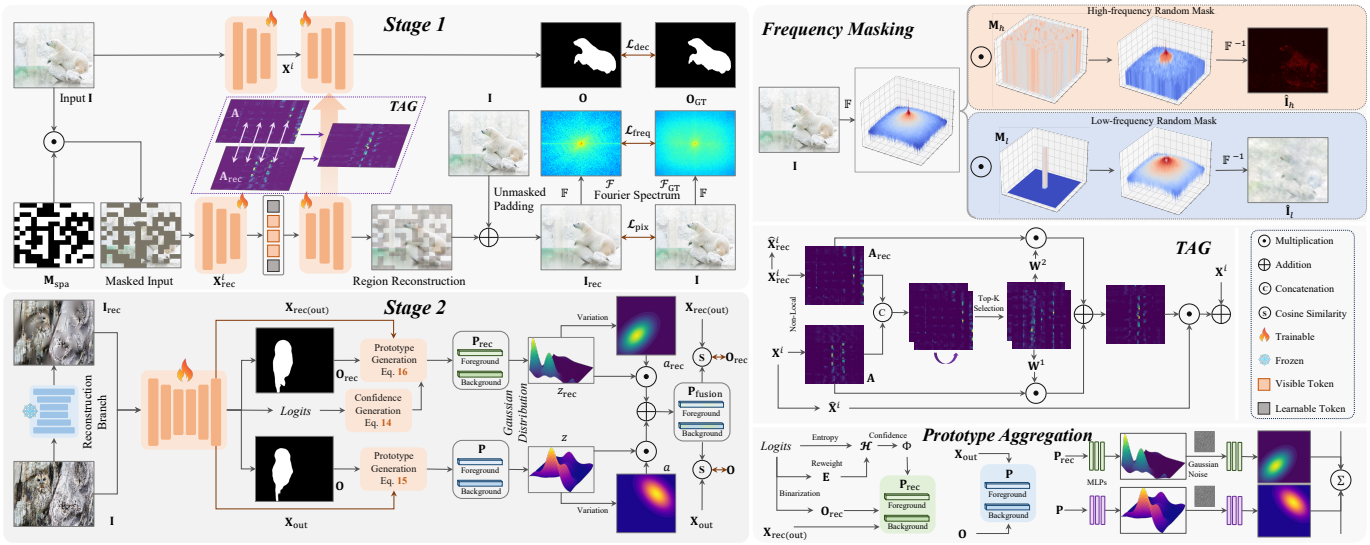

Fig. 2. Stage 1: The input image undergoes masked patch partitioning in the spatial domain, while threshold-based masking is applied to obtain high- and low-frequency components. Unmasked patches are encoded to extract visible tokens, and along with learnable tokens, are decoded in spatial and frequency branches for pixel- and spectrum-level reconstruction. The TAG enables cross-branch knowledge transfer, serving as an explicit constraint. Stage 2: The spatially reconstructed image is passed through the detection network to generate predictions. Uncertainty estimation and edge information are leveraged to construct the confidence map that refines pixel-wise weighting. Finally, prototype integration enforces cross-view consistency, forming a closed-loop from reconstruction to prediction.

D. Task Affinity Guidance

Despite employing reconstruction as auxiliary supervision and implicitly updating the detection model parameters via gradient optimization, our approach may require more training iterations and could lead to suboptimal performance. To address this limitation, we explicitly establish feature associations. Unlike vanilla cross-attention mechanisms with high computational complexity ($\mathcal{O}(H^2W^2)$), we construct a more efficient affinity matrix to reduce computational overhead and suppress noise. Specifically, we adopt channel-wise non-local associations with reduced complexity ($\mathcal{O}(C^2)$) to capture long-range dependencies and enhance informative representations, ultimately generating the affinity map $\mathbf{A}$ and the optimized feature $\hat{\mathbf{X}}$,

$$[\hat{\mathbf{X}}, \mathbf{A}] = \texttt{NonLocal}(\mathbf{X}) \tag{9}$$

For $\mathbf{X}_{\text{rec}}$, the operation follows the same procedure, producing $\hat{\mathbf{X}}_{\text{rec}}$ and $\mathbf{A}_{\text{rec}}$. Due to representation differences, we fuse $\mathbf{A}$ and $\mathbf{A}_{\text{rec}}$ to obtain the weight map $\mathbf{W}$,

$$[\mathbf{W}^1, \mathbf{W}^2] = \texttt{Top-K}(\texttt{Self-Attention}(\texttt{Concat}(\mathbf{A}, \mathbf{A}_{\text{rec}}))) \tag{10}$$

where `Concat` denotes channel concatenation, while `Top-K` selects the top-k weights to filter out low-response elements. In Figure 6, we select the top 70% (or 80%) of elements to achieve a high signal-to-noise ratio. Thus, we can obtain the updated map that incorporates the reconstruction knowledge,

$$\mathbf{A} := \mathbf{W}^1 \odot \mathbf{A} + \mathbf{W}^2 \odot \mathbf{A}_{\text{rec}} \tag{11}$$

Correspondingly, the updated feature is,

$$\hat{\mathbf{X}} := \mathbf{X} + \mathbf{A} \odot \hat{\mathbf{X}} \tag{12}$$

E. Prototype Consistency Calibration

Confidence Map Generation. In practice, the representations of the reconstruction and the original perspective cannot be fully identical. Inevitably, noise disturbances and differentiated modeling can lead to unreliability in certain regions of the predicted map $\mathbf{O}_{\text{rec}}$. To quantify the degree of reliability, we introduce uncertainty estimation. Unlike Monte Carlo strategy that requires multiple samples, we measure based on entropy $\mathcal{H}$, which demands only a single forward pass. For camouflaged targets, the boundary information is more critical than the main body. Therefore, we generate the edge map $\mathbf{E}$ and utilize it as a spatial weighting factor. For any position (i, j) of output-layer feature $\mathbf{X}_{\text{rec(out)}} \in \mathbb{R}^{1\times H\times W}$, we apply sigmoid to obtain the probability $\mathbf{p}$ for class m,

$$\begin{aligned} \mathcal{H}(\mathbf{p}^{i,j}) &= -\sum_{m=0}^{M-1} w(i,j)\mathbf{p}^{i,j}(m)\log \mathbf{p}^{i,j}(m) \\ w(i,j) &= 1 + \alpha \cdot \mathbf{E}(i,j) \end{aligned} \tag{13}$$

where α is a scaling factor to balance pixel-wise weights between edge and non-edge regions, and $\mathbf{E}(i,j) \in \{0,1\}$ denotes the edge indicator function: $\mathbf{E}(i,j) = 1$ if pixel (i,j) lies on generated edge, and 0 otherwise (for COD task, $M = 2$). In information theory, entropy serves as a quantitative measure of uncertainty, where higher entropy values correspond to increased information disorder and probabilistic ambiguity in the prediction confidence. Based on this, we further define the confidence map Φ,

$$\Phi^{i,j} = \frac{1}{H\times W}\left(1 - \frac{\mathcal{H}(\mathbf{p}^{i,j})}{\sum_{i=0}^{H-1}\sum_{j=0}^{W-1}\mathcal{H}(\mathbf{p}^{i,j})}\right) \tag{14}$$

Prototype Generation. The prototype characterizes the overall representation of a region. We generate the raw prototype via masked average pooling (MAP),

$$\mathbf{P}^m = \frac{\sum_{i=0}^{H-1}\sum_{j=0}^{W-1}\mathbf{X}_{\text{out}}^{i,j}\mathbb{1}[\mathbf{O}^{i,j}=m]}{\sum_{i=0}^{H-1}\sum_{j=0}^{W-1}\mathbb{1}[\mathbf{O}^{i,j}=m]} \tag{15}$$

where $\mathbb{1}$ is an indicator function. MAP treats features from

all regions equally. However, the reconstructed features may contain errors (*e.g.,* blurriness and noise), particularly in incomplete or complex areas. Directly using MAP may lead to the smoothing of errors or even obscure the correct representations. We utilize Φ to automatically filter reliable samples, achieving self-calibration,

$$\mathbf{P}_{\text{rec}}^{m} = \frac{\sum_{i=0}^{H-1}\sum_{j=0}^{W-1}\mathbf{X}_{\text{rec(out)}}^{i,j}\Phi^{i,j}\mathbb{1}[\mathbf{O}_{\text{rec}}^{i,j}=m]}{\sum_{i=0}^{H-1}\sum_{j=0}^{W-1}\mathbb{1}[\mathbf{O}_{\text{rec}}^{i,j}=m]} \tag{16}$$

Variational Prototype Fusion. While naive fusion strategies like element-wise summation or channel-wise concatenation between prototypes $\mathbf{P}^m$ and $\mathbf{P}_{\text{rec}}^m$ appear straightforward, both suffer from critical limitations: 1) Direct summation induces linear amplification of noise artifacts: if $\mathbf{P}^m$ contains misclassified background pixels due to blurred object boundaries and $\mathbf{P}_{\text{rec}}^m$ exhibits restoration distortions (*e.g.,* occlusion), error patterns become compounded via accumulation; 2) Concatenation risks feature space misalignment, as $\mathbf{P}^m$ primarily encodes fine-grained representations while $\mathbf{P}_{\text{rec}}^m$ captures structural knowledge, creating distribution shifts that hinder semantic consistency when forcibly combined. We utilize variational fusion to alleviate through probabilistic modeling.

To further explore the intrinsic uncertainty, we project the samples into the latent Gaussian distribution $\mathcal{N}$ to establish semantic associations. For $\mathbf{P}^m$, we have,

$$\begin{aligned} p(z|\mathbf{P}^m) &= \mathcal{N}(z;\mu,(\sigma)^2\mathcal{I}) \\ \mu = \mathcal{M}_\mu(\mathbf{P}^m),\ \sigma &= \mathcal{M}_\sigma(\mathbf{P}^m) \end{aligned} \tag{17}$$

where z, μ, σ, and $\mathcal{I}$ represent the reconstruction vector, mean, variance, and identity matrix, respectively. A larger σ indicates greater uncertainty. $\mathcal{M}$ denotes the multilayer perceptron (MLPs). Similarly, for $\mathbf{P}_{\text{rec}}^m$, we can obtain z_{rec}.

Sampling latent variables from a distribution to generate attention weights results in non-differentiable gradients. To resolve this, we employ the reparameterization trick, which decouples the stochasticity by introducing external noise, allowing gradient flow through a deterministic computational path. Specifically, we introduce standard Gaussian noise η,

$$z = \mu + \eta\cdot\sigma, \quad \eta\in\mathcal{N}(0,1) \tag{18}$$

A similar operation is performed for z_{rec}. Traditional attention mechanisms (like Softmax) generate fixed weights a^i through point estimates, which are essentially deterministic mappings and cannot reflect the model's confidence in the weight values,

$$\begin{aligned} \hat{a}^i &= \mathbf{W}^i z^i + \mathbf{b}^i, \quad i\in\{o,r\} \\ a^i &= \frac{\exp(\hat{a}^i)}{\sum_{j\in\{o,r\}}\exp(\hat{a}^j)} \end{aligned} \tag{19}$$

where o and r denote the original and reconstructed spatial images, respectively. $\mathbf{W}$ and $\mathbf{b}$ are learnable parameters used for projection. Instead of using deterministic weights, we leverage dual uncertainty modeling by treating the weights as probability distributions. *1) Feature-level uncertainty:* The variances σ and σ_{rec} of the prototype features represent local confidence. *2) Weight-level uncertainty:* The variance σ_a of the attention weights captures global confidence.

For each latent variable z^i, we employ shared MLPs to estimate the mean μ_a and variance σ_a. Accordingly, we reformulate Eq. 19 as follows,

$$\begin{aligned} \hat{a}^i \sim q(\hat{a}^i \mid z^i) &= \mathcal{N}\left(\mu_a^i,(\sigma_a^i)^2\mathbf{I}\right), \quad i\in\{o,r\} \\ a^i &= \frac{\exp\left(-\gamma(\sigma_a^i)^2\right)}{\sum_{j\in\{o,r\}}\exp\left(-\gamma(\sigma_a^j)^2\right)} \end{aligned} \tag{20}$$

where γ is a learnable temperature coefficient that controls the intensity of variance suppression on the weights. We can obtain the prototype $\mathbf{P}_{\text{fusion}}^m$ after weighted fusion,

$$\mathbf{P}_{\text{fusion}}^m = \sum_{i\in\{o,r\}} a^i z^i \tag{21}$$

To prevent overfitting to noise, we enforce that the prototype distribution approaches the standard Gaussian prior through Kullback–Leibler (KL) divergence,

$$\mathcal{L}_{\text{KL}} = \sum_{i\in\{o,r\}}\frac{1}{2}\left((\mu^i)^2+(\sigma^i)^2-\log((\sigma^i)^2)-1\right) \tag{22}$$

Based on variational fusion, we reduce the weights in high-variance (low confidence) regions to suppress noise propagation. We dynamically adjust the fusion strategy based on input scenarios to achieve adaptive feature complementarity.

Metric Consistency. We use the similarity matrix $\mathbf{S}$ between feature and prototype to estimate the class probability for each pixel. Essentially, this strategy is an implicit form of contrastive learning,

$$\mathbf{S}^{i,j} = \texttt{CosSim}(\mathbf{X}^{i,j},\mathbf{P}^m) = \frac{\mathbf{X}^{i,j}\cdot\mathbf{P}^m}{\|\mathbf{X}^{i,j}\|_2\cdot\|\mathbf{P}^m\|_2+\epsilon} \tag{23}$$

where $\texttt{CosSim}(\cdot)$ and ϵ represent the cosine similarity and the minimum value, respectively. Similarly, we can obtain $\mathbf{S}_{\text{rec}}$ for $\mathbf{P}_{\text{rec}}$. When $\mathbf{P}_{\text{fusion}}$ is sufficiently accurate and robust, it follows that $\mathbf{S}\to\mathbf{O}$ (or $\mathbf{S}_{\text{rec}}\to\mathbf{O}_{\text{rec}}$). Base on cross entropy loss $\mathcal{L}_{\text{CE}}$, we have,

$$\mathcal{L}_{\text{pro}} = \mathcal{L}_{\text{CE}}(\mathbf{S},\mathbf{O}),\ \mathcal{L}_{\text{pro(rec)}} = \sum_{i=0}^{H-1}\sum_{j=0}^{W-1}\Phi^{i,j}\mathcal{L}_{\text{CE}}(\mathbf{S}_{\text{rec}}^{i,j},\mathbf{O}_{\text{rec}}^{i,j}) \tag{24}$$

For the main detection branch, following [42], we employ structure loss $\mathcal{L}_{\text{dec}}$. Note that we apply supervision to the predictions at all stages of the decoder. We can derive the total loss $\mathcal{L}_{\text{total}}$,

$$\mathcal{L}_{\text{totoal}} = \sum_{i\in\text{HRR, KL, pro, proe(rec), dec}}\lambda_i\mathcal{L}_i \tag{25}$$

In the TTA phase, we discard $\mathcal{L}_{\text{dec}}$.

IV. EXPERIMENTS

A. Datasets

We conduct experiments on four standard COD benchmark datasets. CAMO [65] and COD10K [43] contain 1,000 and 3,040 training pairs, with 250 and 2,026 testing pairs, respectively. CHAMELEON [66] and NC4K [47] are used exclusively for testing, consisting of 76 and 4,121 images. Following [31], [54], we aggregate the training pairs from CAMO and COD10K as the full training set, while the remaining data is assigned as the testing set. For the underwater benchmarks, following [5], we select MAS3K [67], RMAS [68], UFO120 [69], and RUWI [70]. MAS3K consists of 1,769 training pairs and 1,141 testing pairs with foreground objects. RMAS and UFO120 contain 3,014/500 and 1,500/120

TABLE I

QUANTITATIVE COMPARISON ON NORMAL BENCHMARK DATASETS. FOUNDATION MODELS INCLUDE, *e.g.,* SAM. EXTRA DATA INCLUDES, *e.g.,* DEPTH MAP AND MORE PIXELS. BEST PERFORMANCE IN **BOLD**, SECOND IN UNDERLINE. ‡ REPRESENTS DATA IS UNAVAILABLE. † DENOTES USING HCL. *: HIGHER/MULTI-SCALE INPUT RESOLUTION. ↑ INDICATES HIGHER VALUES ARE BETTER, WHILE ↓ INDICATES THE OPPOSITE. 'BASE-R' AND 'BASE-P' DENOTE RESNET50 AND PVT-V2 AS BACKBONE, RESPECTIVELY.

Method	Foundation Model	Extra Data	TTA	CHAMELEON (76)				CAMO (250)				COD10K (2026)				NC4K (4121)			
				$S_m\uparrow$	$E_m\uparrow$	$F_\beta^w\uparrow$	MAE↓	$S_m\uparrow$	$E_m\uparrow$	$F_\beta^w\uparrow$	MAE↓	$S_m\uparrow$	$E_m\uparrow$	$F_\beta^w\uparrow$	MAE↓	$S_m\uparrow$	$E_m\uparrow$	$F_\beta^w\uparrow$	MAE↓
PraNet [1] MICCAI'20	✗	✗	✗	0.860	0.898	0.763	0.044	0.769	0.833	0.663	0.094	0.789	0.839	0.629	0.045	0.822	0.876	0.724	0.059
SINet [43] CVPR'20	✗	✗	✗	0.872	0.946	0.806	0.034	0.751	0.771	0.606	0.100	0.771	0.806	0.551	0.051	0.810	0.873	0.772	0.057
MGL [44] CVPR'21	✗	✗	✗	0.893	0.923	0.813	0.030	0.775	0.847	0.673	0.088	0.814	0.865	0.666	0.035	‡	‡	‡	‡
PFNet [45] CVPR'21	✗	✗	✗	0.882	0.931	0.810	0.033	0.782	0.852	0.695	0.085	0.800	0.868	0.660	0.040	0.829	0.887	0.745	0.053
UGTR [46] ICCV'21	✗	✗	✗	0.888	0.940	0.794	0.031	0.785	0.859	0.686	0.086	0.818	0.850	0.667	0.035	‡	‡	‡	‡
LSR [47] CVPR'21	✗	✗	✗	0.893	0.938	0.839	0.033	0.793	0.826	0.725	0.085	0.793	0.868	0.685	0.041	0.839	0.883	0.779	0.053
SINet-v2 [48] TPAMI'21	✗	✗	✗	0.888	0.942	0.816	0.030	0.820	0.882	0.743	0.070	0.815	0.887	0.680	0.037	0.847	0.903	0.769	0.048
PreyNet * [49] MM'22	✗	✓	✗	0.895	0.951	0.844	0.028	0.790	0.842	0.708	0.077	0.813	0.891	0.697	0.034	‡	‡	‡	‡
ZoomNet * [42] CVPR'22	✗	✓	✗	0.902	0.958	0.845	0.023	0.820	0.892	0.752	**0.066**	0.838	0.888	0.729	0.029	0.853	0.896	0.784	0.043
PopNet [14] ICCV'23	✗	✓	✗	**0.917**	**0.965**	**0.875**	**0.020**	0.808	0.859	0.744	0.077	**0.851**	**0.910**	**0.757**	**0.028**	0.861	0.909	**0.802**	0.042
FEDER [50] CVPR'23	✗	✗	✗	0.907	0.964	‡	0.025	0.807	0.873	‡	0.069	0.823	0.900	‡	0.032	0.846	0.905	‡	0.045
RUN [51] ICML'25	✗	✗	✗	0.895	0.952	‡	0.027	0.806	0.868	‡	0.070	0.827	0.903	‡	0.030	0.851	0.908	‡	0.042
Base-R (Ours) †	✗	✗	✓	0.880	0.949	0.840	0.025	**0.827**	**0.898**	**0.765**	0.067	0.842	0.898	0.744	0.030	**0.865**	**0.915**	0.795	**0.042**
SAM [52] ICCV'23	✓	✗	✗	‡	‡	‡	‡	‡	‡	‡	‡	0.778	0.800	0.701	0.050	0.765	0.778	0.696	0.078
HitNet * [53] AAAI'23	✗	✓	✗	0.922	**0.970**	**0.903**	**0.018**	0.844	0.902	0.801	0.057	0.868	0.932	0.798	0.024	0.870	0.921	0.825	0.039
FPNet [54] MM'23	✗	✗	✗	0.914	0.960	0.868	0.022	0.851	0.912	0.802	0.056	0.851	0.909	0.755	0.028	‡	‡	‡	‡
FSPNet [55] CVPR'23	✗	✗	✗	0.908	0.943	0.851	0.023	0.856	0.899	0.799	0.050	0.851	0.895	0.735	0.026	0.879	0.915	0.816	0.035
EVP [56] CVPR'23	✗	✗	✗	0.871	0.917	0.795	0.036	0.846	0.895	0.777	0.067	0.843	0.907	0.742	0.032	0.874	‡	‡	‡
CamoFormer [57] TPAMI'24	✗	✗	✗	0.910	0.957	0.866	0.022	0.872	0.929	0.831	0.046	0.869	0.932	0.786	0.023	0.892	0.939	0.847	0.030
VSCode [25] CVPR'24	✗	✓	✗	‡	‡	‡	‡	0.836	0.892	0.768	0.060	0.847	0.913	0.744	0.028	0.874	0.920	0.813	0.038
Spider [58] ICML'24	✗	✗	✗	0.906	0.951	0.848	0.025	0.855	0.908	0.799	0.053	0.856	0.917	0.756	0.028	0.880	0.925	0.825	0.036
DSAM [59] MM'24	✓	✓	✗	0.853	0.924	0.785	0.045	0.832	0.913	0.794	0.061	0.846	0.921	0.760	0.033	0.871	0.932	0.826	0.040
CamoDiffusion-VAE [60] TPAMI'25	✗	✗	✗	‡	‡	‡	‡	0.868	0.926	0.827	0.047	0.857	0.919	0.759	0.024	0.877	0.926	0.821	0.034
COMPrompter [26] SCIS'25	✓	✗	✗	0.884	0.946	0.830	0.030	0.853	0.919	0.819	0.054	0.861	0.933	0.779	0.026	0.880	0.935	0.840	0.036
USCNet [61] ICCV'25	✓	✓	✗	‡	‡	‡	‡	0.845	0.886	0.790	0.049	0.821	0.869	0.700	0.030	0.839	0.877	0.768	0.039
VLSAM [62] ICCV'25	✓	✓	✗	0.923	0.946	0.853	0.027	0.863	0.901	0.782	0.059	**0.896**	0.907	**0.808**	0.023	0.872	0.918	0.819	0.033
SAMTTT [63] MM'25	✓	✗	✓	‡	‡	‡	‡	0.868	0.935	0.838	0.045	0.874	**0.942**	0.805	0.027	0.884	0.943	0.837	0.031
Base-P (Ours) †	✗	✗	✓	0.893	0.962	0.853	0.023	0.873	**0.940**	0.840	**0.040**	0.860	0.933	0.781	0.022	0.891	**0.950**	0.856	**0.026**
Spider † [58]	✗	✓	✓	0.919	0.963	0.868	0.022	0.869	0.918	0.815	0.049	0.872	0.929	0.774	0.025	0.894	0.938	0.842	0.033
DSAM † [59]	✓	✓	✓	0.868	0.937	0.807	0.041	0.847	0.926	0.814	0.056	0.861	0.934	0.779	0.030	0.885	0.942	0.839	0.037
CamoFormer † [57]	✗	✗	✓	**0.923**	0.966	0.883	0.019	**0.885**	0.940	**0.850**	0.042	0.884	0.941	0.804	**0.020**	**0.907**	0.948	**0.866**	0.027

Algorithm 1 Our HCL Framework

Input: Image $\mathbf{I}$
Stage 1: Hierarchical Reconstruction
$\mathbf{M}_{\text{spa}} \leftarrow \mathsf{SpaMask}(\mathbf{I})$, $[\mathbf{M}_l, \mathbf{M}_h] \leftarrow \mathsf{FreqMask}(\mathbb{F}(\mathbf{I}))$
$\mathbf{X}_{\text{rec}}, \mathbf{X} \leftarrow \mathsf{Enc}(\mathbf{I} \odot \mathbf{M}_{\text{spa}}, \mathbf{I})$ ▷ Encoding Process
$[\mathbf{X}_{\text{rec(l)}}, \mathbf{X}_{\text{rec(h)}}]$, $\mathbf{X} \leftarrow \mathsf{Enc}\big(\mathbb{F}^{-1}(\mathbb{F}(\mathbf{I}) \odot [\mathbf{M}_l, \mathbf{M}_h]), \mathbf{I}\big)$
$\mathbf{X} \leftarrow \mathsf{TAG}(\mathbf{X}_{\text{rec}}, \mathbf{X})$, $\mathbf{X} \leftarrow \mathsf{TAG}([\mathbf{X}_{\text{rec(l)}}, \mathbf{X}_{\text{rec(h)}}], \mathbf{X})$
$\mathbf{O}, \mathbf{I}_{\text{rec}} \leftarrow \mathsf{Dec}(\mathbf{X}_{\text{rec}}, \mathbf{X})$ ▷ Decoding Process
$\mathbf{O}, [\mathbf{I}_{\text{rec(l)}}, \mathbf{I}_{\text{rec(h)}}] \leftarrow \mathsf{Dec}(\mathbf{X}_{\text{rec}}, \mathbf{X})$
Stage 2: Consistency Calibration
$\mathbf{I}_{\text{rec}} \leftarrow \mathsf{FrozenNet}(\mathbf{I})$ ▷ Reconstruction
$[\mathbf{X}_{\text{rec(out)}}, \mathbf{O}_{\text{rec}}, \Phi] \leftarrow \mathsf{TrainableNet}(\mathbf{I}_{\text{rec}})$ ▷ Detection
$[\mathbf{X}, \mathbf{O}] \leftarrow \mathsf{TrainableNet}(\mathbf{I})$
$\mathbf{P}_{\text{rec}}, \mathbf{P} \leftarrow \mathsf{Prototype}([\mathbf{X}_{\text{rec(out)}}, \mathbf{O}_{\text{rec}}, \Phi], [\mathbf{X}, \mathbf{O}])$
$\mathbf{P}_{\text{fusion}} \leftarrow \mathsf{VariationalFusion}(\mathbf{P}_{\text{rec}}, \mathbf{P})$
$\mathbf{S}_{\text{rec}}, \mathbf{S} \leftarrow \mathsf{CosSim}(\mathbf{P}_{\text{fusion}}, [\mathbf{X}_{\text{rec(out)}}, \mathbf{X}_{\text{out}}])$
if NOT INFERENCE **then**
    $\mathcal{L}_{\text{total}} \leftarrow \mathcal{L}_{\text{HRR}} + \mathcal{L}_{\text{KL}} + \mathcal{L}_{\text{pro}} + \mathcal{L}_{\text{pro(rec)}} + \mathcal{L}_{\text{dec}}$
else
    $\mathcal{L}_{\text{total}} \leftarrow \mathcal{L}_{\text{HRR}} + \mathcal{L}_{\text{KL}} + \mathcal{L}_{\text{pro}} + \mathcal{L}_{\text{pro(rec)}}$
end if
Output: Prediction Map $\mathbf{O}$

training/testing pairs, respectively. Following [5], we set the training and testing sets of RUWI to 525 and 175 pairs, respectively. To further evaluate the model's robustness under distribution shifts, inspired by [71], [72], we adopt three common degradation strategies: Gaussian blur (GB), Gaussian noise (GN), and contrast reduction (CR). These perturbations simulate real-world variations and assess the model's adaptability in challenging conditions.

B. Implementation Details.

We implement our model and conduct experiments using the PyTorch framework on NVIDIA A100 GPUs. We utilize PVT-v2 [73] and ResNet50 [74], both pre-trained on ImageNet, as backbone networks. For fair comparison, following [50], [55], we set the input resolution to 384×384. During the training phase, we leverage the AdamW optimizer with a batch size of 16, an initial learning rate of 0.001, and train for 200 epochs. Data augmentation strategies, including horizontal and vertical flipping, are applied to enhance generalization. During the testing phase, we apply the same settings for the TTA stage as in the training phase. Moreover, we do not incorporate any post-processing techniques, such as CRF, to refine predictions. For the reconstruction branch, we set the patch size to 16, the depth of the encoder-decoder and the number of attention heads to 4, with the embedding dimension of 512 to reduce computational cost. To ensure a fair comparison, we compute the evaluation metrics using either the prediction results, pre-trained weights, or source codes provided by the projects.

C. Evaluation Metrics.

We adopt four evaluation metrics: S-measure (S_m), mean E-measure (E_m), weighted F-measure (F_β^w), and Mean Absolute

TABLE II
QUANTITATIVE COMPARISON ON DEGRADED BENCHMARK DATASETS. ATTR. INDICATES THE APPLIED DEGRADATION TYPE: GN (GAUSSIAN NOISE), GB (GAUSSIAN BLUR), AND CR (CONTRAST REDUCTION). GRAY ROWS REPRESENT METHODS EQUIPPED WITH HCL. *: DEPTH MAPS GENERATED VIA [64] FOLLOWING [14], [59].

Method	Attr.	CHAMELEON (76)				CAMO (250)				COD10K (2026)				NC4K (4121)			
		$S_m \uparrow$	$E_m \uparrow$	$F_\beta^w \uparrow$	MAE $\downarrow$	$S_m \uparrow$	$E_m \uparrow$	$F_\beta^w \uparrow$	MAE $\downarrow$	$S_m \uparrow$	$E_m \uparrow$	$F_\beta^w \uparrow$	MAE $\downarrow$	$S_m \uparrow$	$E_m \uparrow$	$F_\beta^w \uparrow$	MAE $\downarrow$
DSAM *	GN	0.805	0.894	0.714	0.059	0.768	0.863	0.704	0.087	0.808	0.897	0.700	0.043	0.828	0.907	0.765	0.054
+ HCL	GN	0.837	0.914	0.758	0.051	0.800	0.892	0.756	0.075	0.833	0.907	0.741	0.038	0.855	0.921	0.797	0.048
Spider	GN	0.877	0.927	0.800	0.034	0.765	0.828	0.668	0.089	0.810	0.881	0.686	0.038	0.831	0.888	0.754	0.052
+ HCL	GN	0.892	0.940	0.831	0.030	0.812	0.873	0.726	0.072	0.838	0.901	0.724	0.033	0.864	0.897	0.788	0.045
CamoFormer	GN	0.872	0.922	0.805	0.036	0.810	0.871	0.731	0.070	0.833	0.902	0.724	0.032	0.854	0.909	0.788	0.043
+ HCL	GN	0.896	0.941	0.844	0.030	0.853	0.896	0.788	0.059	0.850	0.918	0.761	0.027	0.878	0.921	0.823	0.038
SAM-TTT	GN	0.848	0.922	0.798	0.042	0.832	0.910	0.792	0.058	0.842	0.918	0.778	0.034	0.864	0.928	0.808	0.040
+ HCL	GN	0.872	0.939	0.830	0.036	0.855	0.925	0.820	0.050	0.866	0.932	0.812	0.029	0.884	0.941	0.838	0.034
Base (Ours)	GN	0.802	0.887	0.740	0.043	0.753	0.840	0.709	0.075	0.785	0.868	0.686	0.038	0.807	0.875	0.760	0.042
+ HCL	GN	0.863	0.941	0.796	0.031	0.822	0.893	0.762	0.064	0.834	0.911	0.737	0.029	0.866	0.930	0.815	0.035
DSAM *	GB	0.813	0.896	0.731	0.059	0.794	0.883	0.738	0.082	0.813	0.896	0.707	0.042	0.839	0.913	0.779	0.051
+ HCL	GB	0.835	0.911	0.765	0.052	0.812	0.895	0.772	0.072	0.835	0.905	0.738	0.038	0.852	0.922	0.800	0.047
Spider	GB	0.866	0.922	0.785	0.039	0.759	0.808	0.647	0.091	0.802	0.869	0.668	0.044	0.824	0.878	0.739	0.057
+ HCL	GB	0.888	0.936	0.826	0.033	0.803	0.849	0.708	0.077	0.828	0.887	0.710	0.036	0.851	0.899	0.790	0.047
CamoFormer	GB	0.787	0.837	0.659	0.058	0.759	0.811	0.651	0.088	0.789	0.859	0.653	0.044	0.813	0.866	0.721	0.058
+ HCL	GB	0.832	0.872	0.700	0.046	0.797	0.852	0.705	0.072	0.815	0.891	0.696	0.035	0.861	0.903	0.778	0.044
SAM-TTT	GB	0.838	0.915	0.782	0.046	0.825	0.902	0.776	0.061	0.835	0.910	0.760	0.037	0.855	0.920	0.795	0.042
+ HCL	GB	0.858	0.929	0.808	0.041	0.845	0.918	0.802	0.055	0.857	0.926	0.788	0.032	0.874	0.936	0.822	0.037
Base (Ours)	GB	0.786	0.874	0.725	0.047	0.752	0.835	0.699	0.078	0.771	0.843	0.645	0.048	0.795	0.862	0.744	0.046
+ HCL	GB	0.842	0.943	0.773	0.035	0.815	0.896	0.758	0.064	0.827	0.907	0.724	0.029	0.856	0.923	0.799	0.038
DSAM *	CR	0.846	0.923	0.773	0.046	0.828	0.908	0.787	0.064	0.840	0.917	0.750	0.034	0.868	0.930	0.822	0.041
+ HCL	CR	0.856	0.929	0.783	0.044	0.839	0.917	0.801	0.062	0.851	0.929	0.763	0.032	0.879	0.937	0.830	0.040
Spider	CR	0.905	0.951	0.845	0.025	0.850	0.904	0.790	0.055	0.850	0.914	0.746	0.029	0.877	0.923	0.820	0.037
+ HCL	CR	0.913	0.955	0.853	0.024	0.861	0.910	0.799	0.054	0.858	0.920	0.758	0.028	0.883	0.928	0.827	0.036
CamoFormer	CR	0.888	0.936	0.830	0.028	0.852	0.908	0.799	0.054	0.844	0.911	0.745	0.028	0.877	0.924	0.822	0.035
+ HCL	CR	0.895	0.945	0.847	0.025	0.861	0.917	0.818	0.050	0.855	0.920	0.763	0.025	0.884	0.932	0.835	0.032
SAM-TTT	CR	0.882	0.938	0.830	0.031	0.858	0.925	0.815	0.049	0.865	0.930	0.792	0.029	0.876	0.936	0.825	0.035
+ HCL	CR	0.892	0.945	0.842	0.028	0.868	0.933	0.828	0.045	0.876	0.938	0.806	0.026	0.885	0.942	0.838	0.032
Base (Ours)	CR	0.822	0.901	0.754	0.041	0.825	0.888	0.775	0.056	0.802	0.874	0.715	0.033	0.830	0.889	0.783	0.037
+ HCL	CR	0.868	0.945	0.804	0.028	0.863	0.935	0.825	0.044	0.848	0.924	0.761	0.024	0.886	0.945	0.846	0.028

TABLE III
COMPUTATIONAL COST COMPARISON. VALUES IN PARENTHESES DENOTE THE CHANGE AFTER EQUIPPING HCL.

Method	Input Size	Params (M)	FLOPs (G)	FPS
		Baseline (+Δ)	Baseline (+Δ)	Baseline (-Δ)
DSAM	1024 × 1024	121.93 (+9.24)	1132.15 (+54.86)	8 (-1)
Spider	384 × 384	175.09 (+9.24)	48.59 (+7.58)	38 (-7)
CamoFormer	384 × 384	71.40 (+9.24)	50.12 (+7.58)	32 (-6)
SAM-TTT	1024 × 1024	101.80 (+9.24)	405.76 (+54.86)	12 (-2)
Base (Ours)	384 × 384	40.58 (+9.24)	28.49 (+7.58)	42 (-8)

Error (MAE). Note that for the first three metrics, higher values indicate better performance, whereas for MAE, lower values are preferable.

D. Comparison with State-of-the-Art Methods

1) Quantitative Comparison. In Table I, without relying on foundation models and incorporating depth maps, our method outperforms DSAM by 2.0%, 1.8%, 3.0%, and 1.4% across four metrics on the most challenging NC4K dataset. We analyze that SAM, though pre-trained on large-scale natural images, tends to exhibit biases in camouflaged scenarios and struggles to adapt effectively with limited parameter updates and prior prompts. Moreover, our method also surpasses HitNet by 2.1%, 2.9%, 3.1%, and 1.3%. We argue that although HitNet improves representation capacity by increasing input resolution to preserve discriminative features, it lacks sensitivity to the diversity of camouflage patterns, thereby limiting the adaptability in complex settings. In Table II, under three interference conditions, CR shows the least performance degradation, followed by GN, with GB exhibiting the greatest decline. From the frequency perspective, CR preserves low-frequency structures essential for shape perception, GN adds high-frequency noise that can be partially suppressed, whereas GB removes critical high-frequency details, directly impairing boundary localization. Our method demonstrates consistent improvement of approximately 2%-8% on the NC4K. The higher gains in degraded scenes highlight the effectiveness of representation consistency and self-calibration. We integrate HCL into other methods and observe substantial performance improvements, indicating that train-test distribution shifts inherently exist and require additional customized components. In Table III, we report the parameter count, computational cost (FLOPs), and inference speed (FPS) of the model after integrating HCL, measured on an A100 GPU. While HCL involves multiple auxiliary branches during the adaptation phase (such as spatial and frequency reconstruction), the computational increase remains manageable. For standard resolution inputs (384×384), integrating HCL only adds 9.24M parameters and 7.58G FLOPs. For our base model (Base-P), the inference speed drops slightly from 42 to 34 FPS, which still meets real-time processing requirements. For high-resolution inputs

TABLE IV

QUANTITATIVE ABLATION OF PROPOSED COMPONENTS UNDER DIFFERENT CONDITIONS ON THE COD10K (TOP) AND NC4K (BOTTOM) DATASETS. C1, C2, C3, C4, AND C5 REPRESENT SPATIAL RECONSTRUCTION, LOW-FREQUENCY RECONSTRUCTION, HIGH-FREQUENCY RECONSTRUCTION, CONFIDENCE MAP-GUIDED AND VARIATIONAL FUSION, RESPECTIVELY.

Variant	HRR			TAG	PCC		Clean				Noise				Blur				Contrast			
	C1	C2	C3		C4	C5	$S_m \uparrow$	$E_m \uparrow$	$F_\beta^w \uparrow$	MAE $\downarrow$	$S_m \uparrow$	$E_m \uparrow$	$F_\beta^w \uparrow$	MAE $\downarrow$	$S_m \uparrow$	$E_m \uparrow$	$F_\beta^w \uparrow$	MAE $\downarrow$	$S_m \uparrow$	$E_m \uparrow$	$F_\beta^w \uparrow$	MAE $\downarrow$
I							0.818	0.889	0.747	0.029	0.785	0.868	0.686	0.038	0.771	0.843	0.645	0.048	0.802	0.874	0.715	0.033
II	✓						0.827	0.898	0.755	0.027	0.794	0.874	0.695	0.036	0.780	0.855	0.657	0.044	0.812	0.884	0.720	0.031
III		✓					0.822	0.894	0.753	0.028	0.793	0.870	0.688	0.037	0.781	0.850	0.651	0.045	0.808	0.885	0.718	0.032
IV			✓				0.824	0.891	0.751	0.028	0.789	0.873	0.690	0.036	0.778	0.853	0.648	0.043	0.805	0.880	0.721	0.032
V	✓	✓					0.830	0.903	0.759	0.027	0.801	0.878	0.693	0.035	0.783	0.862	0.670	0.042	0.822	0.889	0.723	0.030
VI	✓		✓				0.831	0.902	0.754	0.026	0.798	0.880	0.698	0.034	0.785	0.864	0.665	0.041	0.817	0.890	0.725	0.030
VII		✓	✓				0.826	0.899	0.755	0.025	0.796	0.875	0.697	0.034	0.784	0.860	0.669	0.040	0.815	0.888	0.724	0.031
VIII	✓	✓	✓				0.839	0.907	0.761	0.024	0.807	0.887	0.705	0.033	0.793	0.875	0.684	0.037	0.834	0.901	0.732	0.029
IX	✓	✓	✓	✓			0.844	0.914	0.765	0.024	0.814	0.895	0.718	0.032	0.803	0.888	0.700	0.035	0.840	0.900	0.741	0.027
X	✓	✓	✓	✓	✓		0.848	0.920	0.771	0.023	0.820	0.902	0.725	0.031	0.810	0.896	0.711	0.032	0.839	0.911	0.748	0.026
XI	✓	✓	✓	✓	✓	✓	**0.860**	**0.933**	**0.781**	**0.022**	**0.834**	**0.911**	**0.737**	**0.029**	**0.827**	**0.907**	**0.724**	**0.029**	**0.848**	**0.924**	**0.761**	**0.024**
XII							0.834	0.900	0.801	0.032	0.807	0.875	0.760	0.042	0.795	0.862	0.744	0.046	0.830	0.889	0.783	0.037
XIII	✓						0.843	0.914	0.811	0.031	0.819	0.886	0.771	0.041	0.811	0.874	0.756	0.044	0.843	0.900	0.795	0.037
XIV		✓					0.838	0.907	0.809	0.032	0.815	0.880	0.766	0.042	0.803	0.868	0.749	0.045	0.835	0.897	0.788	0.037
XV			✓				0.840	0.909	0.805	0.031	0.810	0.883	0.764	0.042	0.804	0.869	0.751	0.045	0.837	0.893	0.790	0.038
XVI	✓	✓					0.847	0.919	0.820	0.030	0.826	0.894	0.776	0.040	0.816	0.882	0.761	0.043	0.851	0.905	0.807	0.035
XVII	✓		✓				0.845	0.921	0.815	0.030	0.824	0.897	0.774	0.041	0.820	0.877	0.763	0.044	0.853	0.907	0.803	0.034
XVIII		✓	✓				0.842	0.912	0.817	0.030	0.820	0.885	0.772	0.040	0.809	0.876	0.759	0.042	0.841	0.902	0.792	0.036
XIX	✓	✓	✓				0.859	0.930	0.835	0.028	0.845	0.909	0.784	0.039	0.829	0.894	0.771	0.042	0.870	0.920	0.828	0.031
XX	✓	✓	✓	✓			0.870	0.935	0.838	0.027	0.842	0.918	0.796	0.038	0.837	0.905	0.783	0.041	0.876	0.925	0.837	0.029
XXI	✓	✓	✓	✓	✓		0.878	0.941	0.844	0.026	0.852	0.923	0.802	0.036	0.843	0.913	0.787	0.040	0.880	0.931	0.833	0.028
XXII	✓	✓	✓	✓	✓	✓	**0.891**	**0.950**	**0.856**	**0.026**	**0.866**	**0.930**	**0.815**	**0.035**	**0.856**	**0.923**	**0.799**	**0.038**	**0.886**	**0.945**	**0.846**	**0.028**

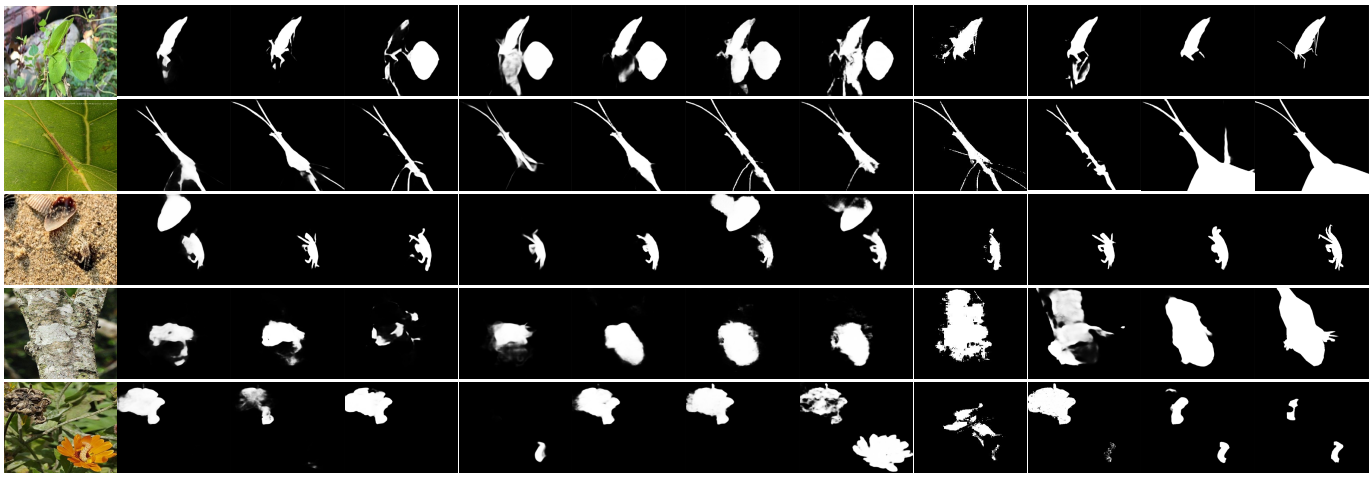

Fig. 3. Qualitative comparison on normal scenarios. Best viewed by zooming in.

(1024×1024), when applied to models like DSAM and SAM-TTT, the parameter overhead remains constant at +9.24M. While the FLOPs naturally scale with the resolution (an increase of 54.86G), the impact on actual inference speed is minimal, experiencing only a negligible drop of 1 to 2 FPS (*e.g.,* DSAM drops from 8 to 7 FPS, and SAM-TTT drops from 12 to 10 FPS). The substantial robustness and generalization gains achieved under severe distribution shifts effectively justify this modest computational trade-off.

2) Qualitative Comparison. In Figure 3 and Figure 4, we present visualization results across various scenarios. Existing methods are prone to false negatives and false positives due to intrinsic physical characteristics and external disturbances, particularly in cases involving significant scale variations and multiple targets. In contrast, our approach effectively mitigates background clutter and imaging artifacts, accurately and comprehensively locating camouflaged regions.

E. Ablation Study

We quantitatively and qualitatively validate the impact of the proposed components, method variants, and critical hyper-parameter settings.

1) Effect of the crucial components. In Table IV, we integrate components through one-step process. We find that: 1) For the HRR, directly using spatial reconstruction (*i.e.,* vanilla MAE) does not yield significant results, while combining with frequency reconstruction achieves complementary effects, especially in degraded scenarios. 2) Integrating the TAG generally improves performance, although some cases show a decline. 3) The confidence map guidance and variational

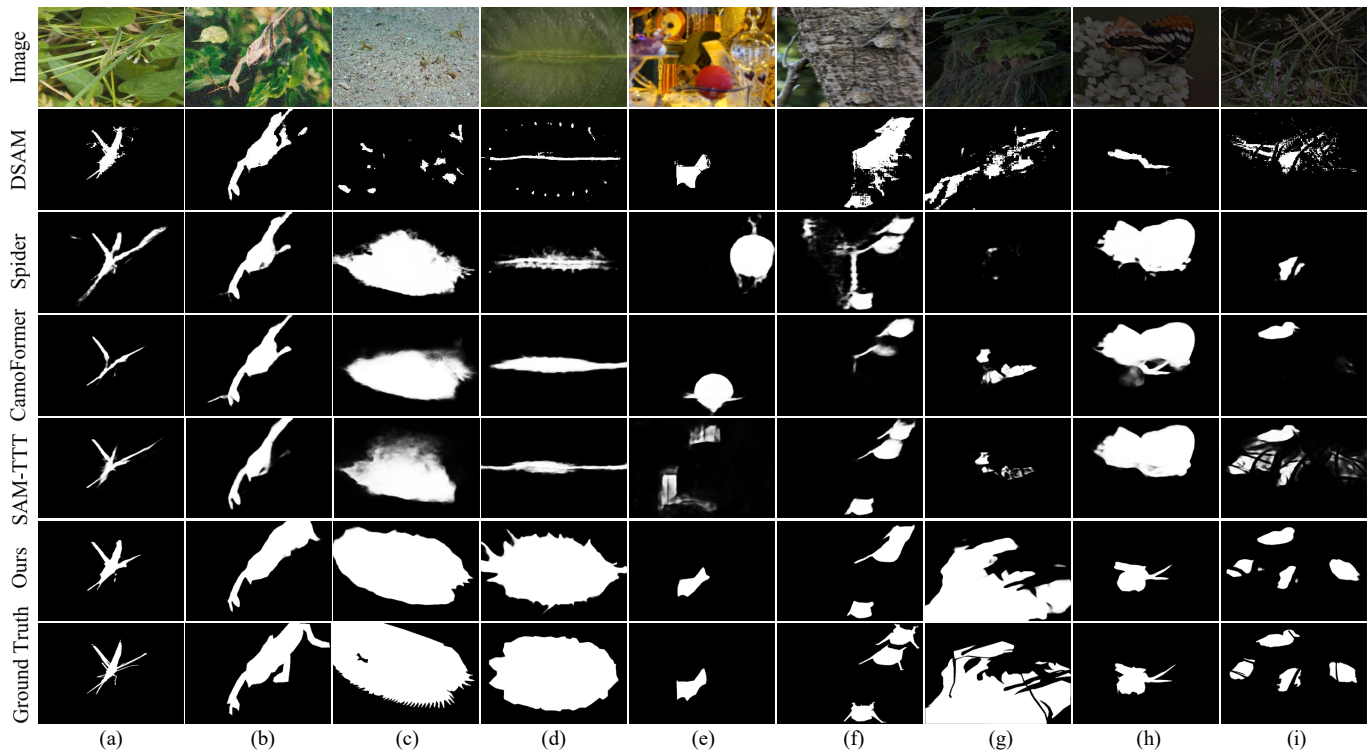

Fig. 4. Qualitative comparison on degraded camouflaged scenarios. (a)-(c): GN setting, (d)-(f): GB setting, (g)-(i): CR setting.

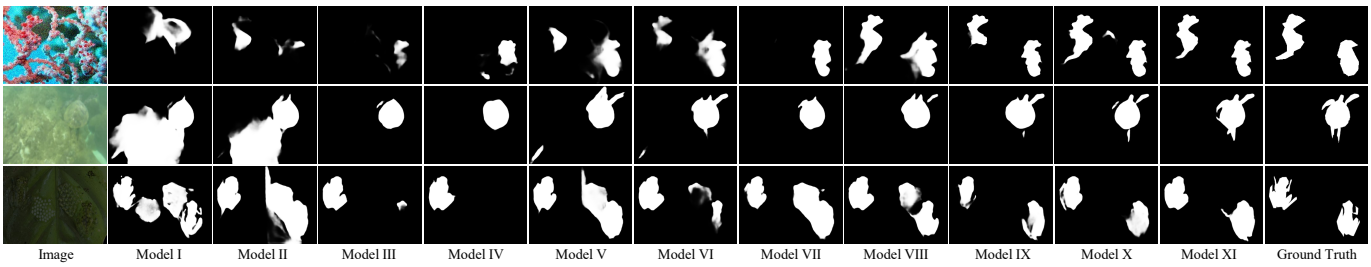

Fig. 5. Qualitative ablation of the proposed components. Variants corresponding to Table IV.

fusion components exhibit progressive coupling. We analyze: 1) Spatial and frequency domains emphasize different aspects of information recovery, and the variations in dynamic scenes and camouflage patterns necessitate more comprehensive clue extraction. 2) We utilize the TAG as a bridge to inject prior knowledge, enhancing detection representation when reconstruction results are ideal; conversely, it may weaken or even obscure when results are suboptimal. 3) The confidence map provides the robust prototype, while the variational component addresses internal instabilities and weight biases, transforming point estimates into probability distributions to mitigate the turbulence. In Figure 5, as the proposed components are gradually integrated, the detection areas approach the ground truth. In Table II, embedding the HCL into other methods can also effectively boost performance by about 3%. The unified principle driving the HCL framework is the maximization of the Evidence Lower Bound (ELBO) [75] for the joint distribution of camouflaged object detection and structural reconstruction during test-time, which treats adaptation as an optimization of the posterior distribution of the latent representation conditioned on the observed image. To achieve this under severe distribution shifts, each component plays a theoretically necessary and complementary role: first, the HRR acts as an unbiased structural anchor that minimizes the uncertainty of the latent manifold by enforcing physical consistency across spatial and spectral domains; second, the TAG ensures task isomorphism by explicitly aligning the discriminative feature manifold with the structural reconstruction manifold to mitigate semantic drift; and finally, the PCC functions as a probabilistic filter that minimizes prediction risk by mapping deterministic prototypes into a probabilistic latent space to suppress high-variance noise. Together, these three components systematically prevent the model from collapsing into trivial solutions and ensure robust feature recalibration.

2) Analysis of TTA strategies. In Table V, we consider: 1) Dynamically adjusting the statistics or parameters of the Batch Normalization layer to adapt to the distribution of the target domain [34], [76]; 2) Updating parameters through

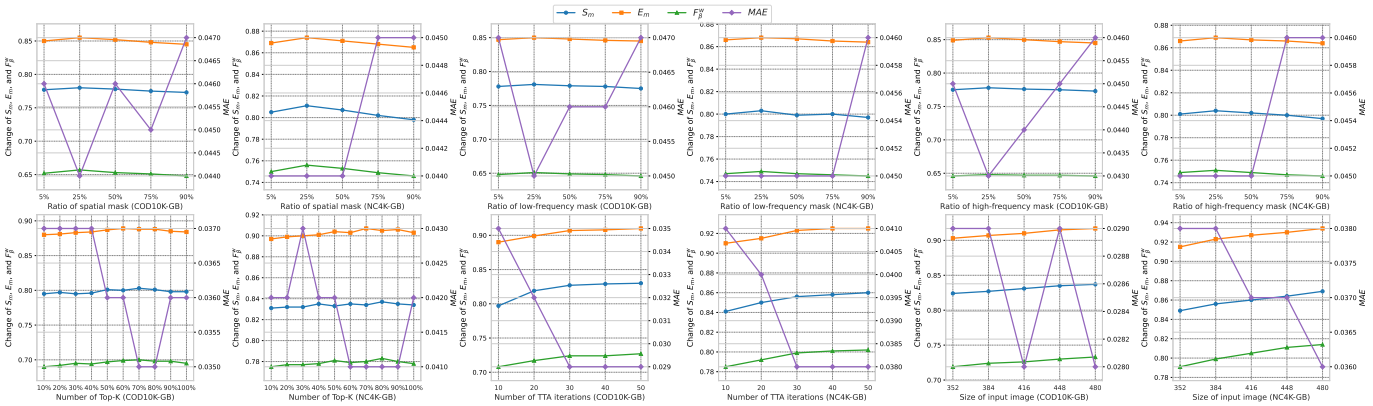

Fig. 6. Quantitative ablation on the ratios of spatial, low-frequency, and high-frequency masks (with optimal spatial ratio), as well as the effects of Top-K selection, TTA iterations, and input image resolution

TABLE V
QUANTITATIVE COMPARISON ON DEGRADED BENCHMARK DATASETS UNDER DIFFERENT ADAPTATION STRATEGIES.

Method	COD10K-GB				NC4K-GB			
	$S_m \uparrow$	$E_m \uparrow$	$F_\beta^w \uparrow$	MAE $\downarrow$	$S_m \uparrow$	$E_m \uparrow$	$F_\beta^w \uparrow$	MAE $\downarrow$
DSAM [59]	0.813	0.896	0.707	0.042	0.839	0.913	0.779	0.051
+ TTT-Rot [2]	0.820	0.900	0.721	0.040	0.845	0.917	0.788	0.049
+ TENT [34]	0.817	0.898	0.724	0.039	0.842	0.915	0.784	0.049
+ UDA [76]	0.818	0.899	0.718	0.040	0.843	0.916	0.782	0.050
+ GH [77]	0.823	0.900	0.716	0.040	0.845	0.915	0.785	0.049
+ TTT-MIM [78]	0.822	0.902	0.726	0.039	0.847	0.919	0.790	0.048
+ HCL	0.835	0.905	0.738	0.038	0.852	0.922	0.800	0.047
Spider [58]	0.802	0.869	0.668	0.044	0.824	0.878	0.739	0.057
+ TTT-Rot [2]	0.810	0.876	0.682	0.042	0.832	0.885	0.748	0.054
+ TENT [34]	0.806	0.872	0.690	0.041	0.829	0.881	0.744	0.054
+ UDA [76]	0.807	0.873	0.679	0.042	0.830	0.882	0.742	0.055
+ GH [77]	0.808	0.874	0.675	0.043	0.828	0.885	0.745	0.055
+ TTT-MIM [78]	0.812	0.879	0.694	0.041	0.835	0.888	0.752	0.053
+ HCL	0.828	0.887	0.710	0.036	0.851	0.899	0.790	0.047
CamoFormer [57]	0.789	0.859	0.653	0.044	0.813	0.866	0.721	0.058
+ TTT-Rot [2]	0.798	0.869	0.668	0.042	0.822	0.874	0.735	0.055
+ TENT [34]	0.794	0.863	0.676	0.041	0.818	0.870	0.731	0.055
+ UDA [76]	0.796	0.865	0.664	0.042	0.820	0.872	0.728	0.056
+ GH [77]	0.795	0.869	0.677	0.042	0.832	0.875	0.733	0.055
+ TTT-MIM [78]	0.801	0.872	0.681	0.040	0.826	0.879	0.742	0.053
+ HCL	0.815	0.891	0.696	0.035	0.861	0.903	0.778	0.044
Base (Ours)	0.771	0.843	0.645	0.048	0.795	0.862	0.744	0.046
+ TTT-Rot [2]	0.795	0.874	0.694	0.033	0.827	0.875	0.764	0.045
+ TENT [34]	0.782	0.845	0.700	0.035	0.816	0.867	0.752	0.041
+ UDA [76]	0.800	0.866	0.687	0.036	0.824	0.879	0.748	0.044
+ GH [77]	0.795	0.858	0.691	0.037	0.811	0.882	0.759	0.043
+ TTT-MIM [78]	0.792	0.879	0.695	0.034	0.821	0.890	0.762	0.043
+ HCL	0.827	0.907	0.724	0.029	0.856	0.923	0.799	0.038

TABLE VI
QUANTITATIVE COMPARISON OF MASK STRATEGIES.

Method	COD10K-GB				NC4K-GB			
	$S_m \uparrow$	$E_m \uparrow$	$F_\beta^w \uparrow$	MAE $\downarrow$	$S_m \uparrow$	$E_m \uparrow$	$F_\beta^w \uparrow$	MAE $\downarrow$
Round	0.820	**0.909**	0.713	0.030	0.849	**0.925**	0.789	0.039
Grid	**0.829**	0.900	0.715	**0.028**	0.853	0.920	0.791	0.039
Block	0.818	0.904	0.718	0.029	0.848	0.918	0.792	0.038
Random	0.827	0.907	**0.724**	0.029	**0.856**	0.923	**0.799**	**0.038**

TABLE VII
QUANTITATIVE COMPARISON OF INTERACTIVE STRATEGIES.

Method	COD10K-GB				NC4K-GB			
	$S_m \uparrow$	$E_m \uparrow$	$F_\beta^w \uparrow$	MAE $\downarrow$	$S_m \uparrow$	$E_m \uparrow$	$F_\beta^w \uparrow$	MAE $\downarrow$
Spatial Cross	0.797	0.880	0.693	0.036	**0.840**	0.898	0.774	0.042
Channel Cross	0.795	**0.889**	0.690	0.037	0.833	0.900	0.780	0.042
TAG	**0.803**	0.888	**0.700**	**0.035**	0.837	**0.905**	**0.783**	**0.041**

rotation predictions to construct consistency [2]. Our HCL shows the best performance. We argue that: 1) Both TENT and UDA struggle to handle confusable visual space and structural ambiguity; 2) Due to the high coupling between foreground and background, camouflaged regions can be positioned anywhere while preserving characteristics. In other words, rotation is inadequate for capturing camouflage and degradation cues. In contrast, image reconstruction explores the intrinsic relationships between regions, and the representations learned through self-supervised learning align better with the detection branch. Mainstream TTA methods (*e.g.,* TENT) are primarily confidence-driven. In COD, targets inherently mimic the background. Using confidence-driven updates leads to confirmation bias, where the model inadvertently amplifies the confidence of erroneous background predictions. Domain adaptation methods (*e.g.,* GH) operate at the domain level, requiring access to target domain datasets, and fail to resolve instance-level feature entanglement. HCL is a sample-specific TTA method that dynamically adapts during inference on a single image. Moreover, HCL is fundamentally consistency-driven. By utilizing hierarchical representation reconstruction and prototype consistency to capture global structures, texture residuals, and robust semantics independently of classification logits, HCL provides an unbiased anchor.

3) Analysis of mask strategies. In Table VI, random masking achieves superior performance compared to fixed geometric patterns (round, grid, block). We attribute to: 1) Fixed masking strategies consistently eliminate specific frequency components, creating persistent blind spots in feature space and learning. 2) Random masking dynamically varies the suppressed frequency bands across spatial locations, ensuring no single spectral component is entirely excluded during reconstruction. This encourages the model to holistically integrate multi-spectrum cues rather than over-relying on localized patterns.

TABLE VIII
QUANTITATIVE COMPARISON OF PROTOTYPE INTEGRATION STRATEGIES.

Method	COD10K-GB				NC4K-GB			
	$S_m \uparrow$	$E_m \uparrow$	$F_\beta^w \uparrow$	MAE $\downarrow$	$S_m \uparrow$	$E_m \uparrow$	$F_\beta^w \uparrow$	MAE $\downarrow$
Sum	0.823	0.901	0.717	0.032	0.848	0.914	0.795	0.040
Concat	0.815	0.899	0.718	0.032	0.840	0.917	0.790	0.039
Cross Attention	0.820	0.902	0.715	0.031	0.847	0.915	0.792	0.039
Variational Fusion	**0.827**	**0.907**	**0.724**	**0.029**	**0.856**	**0.923**	**0.799**	**0.038**

3) The stochastic characteristic of random masks enhances robustness against diverse corruption types by preventing overfitting to artificial geometric artifacts.

4) Analysis of interactive strategies. In Table VII, our TAG outperforms other cross-attention-based strategies, which can be attributed to: 1) Instead of direct feature-level interaction, we refine internal features along the channel dimension and construct an affinity map, encouraging coherent and compact representations while mitigating noise entanglement; 2) A Top-K selection operator is introduced to filter out low-response values, enabling entropy compression and enhancing the quality of information propagation.

5) Analysis of prototype integration strategies. In Table VIII, variational fusion demonstrates superior performance. We analyze that: 1) Unlike fixed fusion schemes, variational fusion learns distribution-based representations, allowing to dynamically assign confidence-aware weights to different features based on reliability; 2) By explicitly incorporating uncertainty, variational fusion suppresses noisy or redundant information and highlights informative cues, improving robustness in challenging conditions such as low saliency or distribution shifts; 3) Fusion in the probabilistic latent space facilitates smoother integration of complementary cues, enabling more consistent and discriminative representations than simple arithmetic operations or attention that may amplify noise.

6) Analysis of mask ratios. In Figure 6, spatial and frequency masking both achieve optimal performance at 25% mask ratio, as moderate masking balances feature stability and contextual integrity. Excessive spatial masking (>25%) removes critical local details, while excessive frequency masking (>25%) disproportionately disrupts high-frequency textures or low-frequency structures, creating irrecoverable information gaps. Besides, frequency masking shows sharper performance decay due to the entangled spectral components increasing reconstruction ambiguity, whereas spatial masking preserves more structural coherence at higher ratios.

7) Analysis of Top-K selection operator. In Figure 6, selecting the top 70% (80%) values yields the best performance. When the threshold is lower, performance improves gradually; when it exceeds 70% (80%), a performance drop is observed. We argue that: 1) Retaining too few connections (low Top-K ratio) limits the model's ability to capture sufficient contextual cues, especially long-range dependencies essential for complex scenes; 2) Retaining too many connections introduces noisy or less informative interactions, which may overwhelm the discriminative patterns and degrade the effectiveness of the feature refinement process. The optimal threshold achieves a balance between preserving informative associations and filtering out noise, optimizing both precision and generalization.

8) Analysis of TTA iterations. In Figure 6, we observe a turning point at 30 adaptation iterations: performance improves steadily when the iteration count is below 30, and then stabilizes beyond that. We analyze that: 1)During the early testing phase, a clear distribution mismatch exists between the test samples and the model's learned feature space, as indicated by the significant discrepancy between prototypes $\mathbf{P}$ and $\mathbf{P}_{\text{rec}}$ distributions; 2) Through iterative refinement (*e.g.,* progressive calibration guided by confidence map Φ), the model gradually aligns its internal parameters with the distribution of samples, reaching a relatively optimal state. Further updates yield diminishing returns, resulting in performance convergence.

9) Analysis of input size. In Figure 6, the performance generally exhibits a positive correlation with the input resolution. We attribute this to: 1) Higher-resolution inputs provide denser spectral sampling in the Fourier domain (expanded along the u and v dimensions), enabling the model to more precisely separate critical high-frequency details and structured low-frequency components; 2) The larger pixel space offers greater capacity to preserve discriminative features of the target, mitigating information loss during the encoding process.

F. Broader Impacts

We evaluate state-of-the-art methods on four underwater benchmarks under the same three degradation settings, and further integrate HCL to explore its performance gains.

1) Quantitative Analysis. In Table IX, we present results under GN, GB, and CR conditions. Across all datasets, GB causes the most severe degradation, followed by GN, while CR shows the least impact. By integrating HCL, both DualSAM and MASSAM consistently achieve performance gains across metrics and datasets. Notably, for DualSAM on the MAS3K, HCL improves F_β^w by 4.2% and reduces MAE by 1.3% under GB, highlighting its ability to recover fine details. Under GN and CR, HCL still brings steady gains of 1.2%–4.4%

2) Qualitative Analysis. In Figure 7, we present qualitative comparisons under various challenging scenarios. Integrating HCL boosts comprehensive perception. In multi-object scenes (columns a, b, and e), it preserves complete instance integrity and avoids omissions by leveraging prototype consistency and fusion to suppress interference between adjacent objects. For objects with highly irregular shapes (columns c, d, and h), the variant captures long-range dependencies and preserves boundary details through non-local affinity modeling. In low-saliency settings (column f), where foreground-background contrast is weak, our hierarchical frequency reconstruction enhances subtle discriminative cues, effectively improving foreground-background separation. In camouflaged instances (columns f, g, and h), our method leverages task-guided reconstruction and entropy-based confidence weighting to highlight critical boundaries and suppress false detections.

3) Why can HCL work under underwater conditions? 1) HCL leverages spatial- and frequency-decoupled reconstruction for TTA, which is not limited to camouflage scenarios and can be extended to other binary segmentation tasks; 2) Existing underwater benchmarks and learning paradigms also suffer

TABLE IX
QUANTITATIVE COMPARISON ON DEGRADED UNDERWATER BENCHMARK DATASETS. ATTR. INDICATES THE APPLIED DEGRADATION TYPE: GN (GAUSSIAN NOISE), GB (GAUSSIAN BLUR), AND CR (CONTRAST REDUCTION). N/A INDICATES NO DEGRADATION. GRAY ROWS REPRESENT METHODS EQUIPPED WITH HCL.

Method	Attr.	MAS3K (1141)				RMAS (500)				UFO120 (120)				RUWI (175)			
		$S_m \uparrow$	$E_m \uparrow$	$F_\beta^w \uparrow$	MAE $\downarrow$	$S_m \uparrow$	$E_m \uparrow$	$F_\beta^w \uparrow$	MAE $\downarrow$	$S_m \uparrow$	$E_m \uparrow$	$F_\beta^w \uparrow$	MAE $\downarrow$	$S_m \uparrow$	$E_m \uparrow$	$F_\beta^w \uparrow$	MAE $\downarrow$
DualSAM [79]	N/A	0.884	0.933	0.838	0.023	0.860	0.944	0.812	0.022	0.856	0.914	0.864	0.064	0.903	0.959	0.939	0.035
MASSAM [80]	N/A	0.887	0.938	0.840	0.025	0.865	0.948	0.819	0.021	0.861	0.914	0.864	0.063	0.894	0.961	0.941	0.035
DualSAM [79]	GN	0.814	0.870	0.755	0.045	0.749	0.781	0.613	0.041	0.798	0.865	0.793	0.085	0.783	0.872	0.827	0.056
+ HCL	GN	0.853	0.901	0.799	0.032	0.785	0.833	0.651	0.033	0.827	0.892	0.831	0.075	0.832	0.903	0.869	0.045
MASSAM [80]	GN	0.833	0.866	0.748	0.042	0.733	0.766	0.587	0.047	0.784	0.850	0.779	0.083	0.773	0.853	0.812	0.053
+ HCL	GN	0.844	0.900	0.777	0.036	0.770	0.801	0.629	0.038	0.814	0.870	0.794	0.077	0.837	0.898	0.867	0.048
DualSAM [79]	GB	0.795	0.857	0.746	0.050	0.740	0.769	0.615	0.044	0.778	0.848	0.758	0.082	0.795	0.863	0.834	0.051
+ HCL	GB	0.844	0.895	0.788	0.037	0.784	0.808	0.648	0.035	0.811	0.875	0.799	0.073	0.838	0.912	0.873	0.045
MASSAM [80]	GB	0.785	0.844	0.750	0.046	0.745	0.781	0.620	0.043	0.773	0.835	0.739	0.079	0.763	0.849	0.830	0.055
+ HCL	GB	0.839	0.875	0.770	0.037	0.769	0.800	0.649	0.037	0.810	0.874	0.786	0.071	0.803	0.891	0.878	0.047
DualSAM [79]	CR	0.844	0.898	0.807	0.035	0.833	0.909	0.778	0.029	0.827	0.876	0.820	0.074	0.873	0.941	0.918	0.042
+ HCL	CR	0.867	0.921	0.819	0.031	0.840	0.930	0.790	0.026	0.844	0.899	0.840	0.071	0.888	0.945	0.938	0.038
MASSAM [80]	CR	0.849	0.908	0.803	0.036	0.828	0.911	0.772	0.031	0.820	0.869	0.823	0.073	0.865	0.935	0.911	0.044
+ HCL	CR	0.864	0.916	0.824	0.032	0.844	0.924	0.796	0.027	0.834	0.883	0.833	0.069	0.882	0.940	0.924	0.038

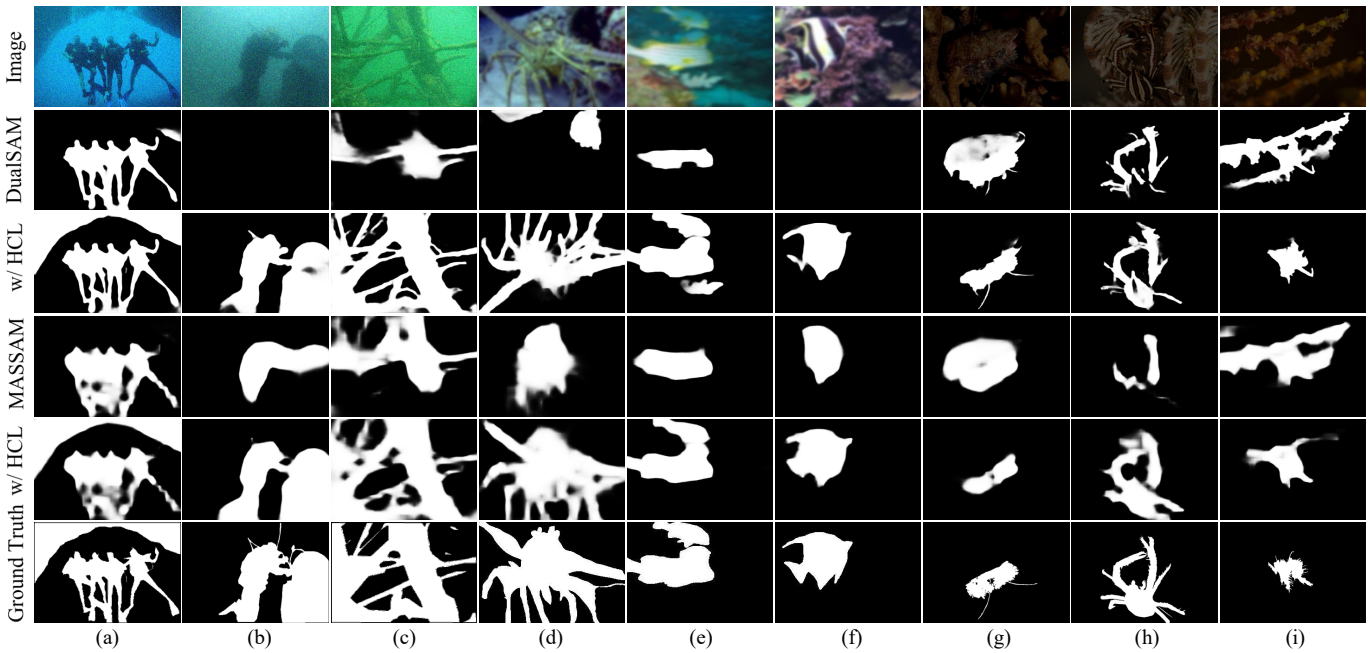

Fig. 7. Qualitative comparison on degraded underwater scenarios. (a)-(c): GN setting, (d)-(f): GB setting, (g)-(i): CR setting.

Fig. 8. Failure cases under degradation conditions.

from train-test distribution shifts and fixed model parameters, making TTA essential, especially when degradation further amplifies distributional discrepancies.

G. Limitation and Future Work

In Figure 8. the HCL framework remains challenged in extreme scenarios: highly camouflaged targets (*e.g.,* insects perfectly mimicking leaves) may evade detection as global feature homogenization overshadows subtle discriminative cues, while highly irregular backgrounds (*e.g.,* forest floors with fragmented shadows) disrupt frequency-based analysis, causing localization errors. These limitations reflect the inherent difficulty in balancing structural preservation and noise suppression under environmental uncertainty. We will focus on: 1) Integrating attention with dynamic sparsity to enhance feature disentanglement by focusing on sample-critical frequency bands; 2) Extending the HCL framework to specific scenes (*e.g.,* remote sensing [81] and agricultural protection [82]),

multimodal learning [83]–[85].

V. CONCLUSION

This paper tackles the challenge of COD under distribution shifts, where static models struggle to generalize. We propose HCL, a dynamic TTA framework that enables sample-specific self-calibration via spatial-frequency decoupled reconstruction. Key components such as TAG and entropy-guided prototype fusion enhance structural consistency and feature robustness. These findings highlight the importance of dynamic adaptation and suggest HCL's potential for broader binary segmentation tasks.

REFERENCES

[1] D.-P. Fan, G.-P. Ji, T. Zhou, G. Chen, H. Fu, J. Shen, and L. Shao, "Pranet: Parallel reverse attention network for polyp segmentation," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2020, pp. 263–273.

[2] Y. Sun, X. Wang, Z. Liu, J. Miller, A. Efros, and M. Hardt, "Test-time training with self-supervision for generalization under distribution shifts," in *International conference on machine learning*. PMLR, 2020, pp. 9229–9248.

[3] S. Niu, J. Wu, Y. Zhang, Y. Chen, S. Zheng, P. Zhao, and M. Tan, "Efficient test-time model adaptation without forgetting," in *International conference on machine learning*. PMLR, 2022, pp. 16 888–16 905.

[4] M. Zha, F. Fu, Y. Pei, G. Wang, T. Li, X. Tang, Y. Yang, and H. T. Shen, "Dual domain perception and progressive refinement for mirror detection," *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.

[5] M. Zha, G. Wang, Y. Pei, T. Li, X. Tang, C. Li, Y. Yang, and H. T. Shen, "Heterogeneous experts and hierarchical perception for underwater salient object detection," *IEEE Transactions on Image Processing*, 2025.

[6] M. Zha, Y. Pei, G. Wang, T. Li, Y. Yang, W. Qian, and H. T. Shen, "Weakly-supervised mirror detection via scribble annotations," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 7, 2024, pp. 6953–6961.

[7] Y. Zhao, L. Zhao, Q. Yu, L. Sheng, J. Zhang, and D. Xu, "Distortion-aware transformer in 360 salient object detection," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 499–508.

[8] M. Zha, G. Wang, T. Li, W. Dong, P. Wang, and Y. Yang, "Seeing beyond illusion: Generalized and efficient mirror detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026.

[9] F. Liu, Y. Liu, K. Xu, S. Ye, G. P. Hancke, and R. W. Lau, "Language-guided salient object ranking," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 29 803–29 813.

[10] H. Guan, J. Lin, and R. W. Lau, "A contrastive-learning framework for unsupervised salient object detection," *IEEE Transactions on Image Processing*, 2025.

[11] Y. Sun, S. Wang, C. Chen, and T.-Z. Xiang, "Boundary-guided camouflaged object detection," *arXiv preprint arXiv:2207.00794*, 2022.

[12] J. Zhu, X. Zhang, S. Zhang, and J. Liu, "Inferring camouflaged objects by texture-aware interactive guidance network," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, no. 4, 2021, pp. 3599–3607.

[13] Y. Sun, C. Xu, J. Yang, H. Xuan, and L. Luo, "Frequency-spatial entanglement learning for camouflaged object detection," in *European Conference on Computer Vision*. Springer, 2024, pp. 343–360.

[14] Z. Wu, D. P. Paudel, D.-P. Fan, J. Wang, S. Wang, C. Demonceaux, R. Timofte, and L. Van Gool, "Source-free depth for object pop-out," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 1032–1042.

[15] Y. Zhang, J. Zhang, W. Hamidouche, and O. Deforges, "Predictive uncertainty estimation for camouflaged object detection," *IEEE Transactions on Image Processing*, vol. 32, pp. 3580–3591, 2023.

[16] H. Zhang, Y. Lyu, Q. Yu, H. Liu, H. Ma, D. Yuan, and Y. Yang, "Unlocking attributes' contribution to successful camouflage: A combined textual and visual analysis strategy," in *European Conference on Computer Vision*. Springer, 2024, pp. 315–331.

[17] J. Du, J. Wu, D. Kong, W. Liang, F. Hao, J. Xu, B. Wang, G. Wang, and P. Li, "Upgen: Unleashing potential of foundation models for training-free camouflage detection via generative models," *IEEE Transactions on Image Processing*, 2025.

[18] H. Chen, D. Shao, G. Guo, and S. Gao, "Just a hint: Point-supervised camouflaged object detection," in *European Conference on Computer Vision*. Springer, 2024, pp. 332–348.

[19] M. Zha, G. Wang, Y. Pei, T. Li, X. Tang, J. Ma, Y. Yang, and H. T. Shen, "Think twice before determining: Towards scene-aware visual reasoning for mirror detection," *IEEE Transactions on Circuits and Systems for Video Technology*, 2026.

[20] X. Lai, Z. Yang, J. Hu, S. Zhang, L. Cao, G. Jiang, Z. Wang, S. Zhang, and R. Ji, "Camoteacher: Dual-rotation consistency learning for semi-supervised camouflaged object detection," in *European Conference on Computer Vision*. Springer, 2024, pp. 438–455.

[21] H. Li, C.-M. Feng, Y. Xu, T. Zhou, L. Yao, and X. Chang, "Zero-shot camouflaged object detection," *IEEE Transactions on Image Processing*, vol. 32, pp. 5126–5137, 2023.

[22] Z. Wang, Y. Li, Y. Yang, Y. Li, and G. Liu, "Few-shot camouflaged object segmentation," in *2024 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2024, pp. 1–10.

[23] J. Zhao, X. Li, F. Yang, Q. Zhai, A. Luo, Z. Jiao, and H. Cheng, "Focusdiffuser: Perceiving local disparities for camouflaged object detection," in *European Conference on Computer Vision*. Springer, 2024, pp. 181–198.

[24] H. Chen, P. Wei, G. Guo, and S. Gao, "Sam-cod+: Sam-guided unified framework for weakly-supervised camouflaged object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.

[25] Z. Luo, N. Liu, W. Zhao, X. Yang, D. Zhang, D.-P. Fan, F. Khan, and J. 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*, 2024, pp. 17 169–17 180.

[26] X. Zhang, Z. Yu, L. Zhao, D.-P. Fan, and G. Xiao, "Comprompter: reconceptualized segment anything model with multiprompt network for camouflaged object detection," *Science China Information Sciences*, vol. 68, no. 1, p. 112104, 2025.

[27] Z. He, C. Xia, S. Qiao, and J. Li, "Text-prompt camouflaged instance segmentation with graduated camouflage learning," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 5584–5593.

[28] W. Hui, Z. Zhu, S. Zheng, and Y. Zhao, "Endow sam with keen eyes: Temporal-spatial prompt learning for video camouflaged object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 19 058–19 067.

[29] X. Zhang, T. Xiao, G.-P. Ji, X. Wu, K. Fu, and Q. Zhao, "Explicit motion handling and interactive prompting for video camouflaged object detection," *IEEE Transactions on Image Processing*, 2025.

[30] L. Wang, J. Yang, Y. Zhang, F. Wang, and F. Zheng, "Depth-aware concealed crop detection in dense agricultural scenes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 17 201–17 211.

[31] C. Hao, Z. Yu, X. Liu, J. Xu, H. Yue, and J. Yang, "A simple yet effective network based on vision transformer for camouflaged object and salient object detection," *IEEE Transactions on Image Processing*, 2025.

[32] Y. Pang, X. Zhao, J. Zuo, L. Zhang, and H. Lu, "Open-vocabulary camouflaged object segmentation," in *European Conference on Computer Vision*. Springer, 2024, pp. 476–495.

[33] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[34] D. Wang, E. Shelhamer, S. Liu, B. Olshausen, and T. Darrell, "Tent: Fully test-time adaptation by entropy minimization," *arXiv preprint arXiv:2006.10726*, 2020.

[35] M. Jang, S.-Y. Chung, and H. W. Chung, "Test-time adaptation via self-training with nearest neighbor information," *arXiv preprint arXiv:2207.10792*, 2022.

[36] D. Brahma and P. Rai, "A probabilistic framework for lifelong test-time adaptation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 3582–3591.

[37] B. Zhao, C. Chen, and S.-T. Xia, "Delta: degradation-free fully test-time adaptation," *arXiv preprint arXiv:2301.13018*, 2023.

[38] J. Ma, "Improved self-training for test-time adaptation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 23 701–23 710.

[39] W. Liu, X. Shen, H. Li, X. Bi, B. Liu, C.-M. Pun, and X. Cun, "Depth-aware test-time training for zero-shot video object segmentation," in

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 19 218–19 227.

[40] J. Hu, J. Lin, S. Gong, and W. Cai, "Relax image-specific prompt requirement in sam: A single generic prompt for segmenting camouflaged objects," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 11, 2024, pp. 12 511–12 518.

[41] L. Jiang, B. Dai, W. Wu, and C. C. Loy, "Focal frequency loss for image reconstruction and synthesis," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 13 919–13 929.

[42] Y. Pang, X. Zhao, T.-Z. Xiang, L. Zhang, and H. Lu, "Zoom in and out: A mixed-scale triplet network for camouflaged object detection," in *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, 2022, pp. 2160–2170.

[43] D.-P. Fan, G.-P. Ji, G. Sun, M.-M. Cheng, J. Shen, and L. Shao, "Camouflaged object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2777–2787.

[44] Q. Zhai, X. Li, F. Yang, C. Chen, H. Cheng, and D.-P. Fan, "Mutual graph learning for camouflaged object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 12 997–13 007.

[45] H. Mei, G.-P. Ji, Z. Wei, X. Yang, X. Wei, and D.-P. Fan, "Camouflaged object segmentation with distraction mining," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 8772–8781.

[46] F. Yang, Q. Zhai, X. Li, R. Huang, A. Luo, H. Cheng, and D.-P. Fan, "Uncertainty-guided transformer reasoning for camouflaged object detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 4146–4155.

[47] Y. Lv, J. Zhang, Y. Dai, A. Li, B. Liu, N. Barnes, and D.-P. Fan, "Simultaneously localize, segment and rank the camouflaged objects," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 11 591–11 601.

[48] D.-P. Fan, G.-P. Ji, M.-M. Cheng, and L. Shao, "Concealed object detection," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 10, pp. 6024–6042, 2021.

[49] M. Zhang, S. Xu, Y. Piao, D. Shi, S. Lin, and H. Lu, "Preynet: Preying on camouflaged objects," in *Proceedings of the 30th ACM international conference on multimedia*, 2022, pp. 5323–5332.

[50] C. He, K. Li, Y. Zhang, L. Tang, Y. Zhang, Z. Guo, and X. Li, "Camouflaged object detection with feature decomposition and edge reconstruction," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 22 046–22 055.

[51] C. He, R. Zhang, F. Xiao, C. Fang, L. Tang, Y. Zhang, L. Kong, D.-P. Fan, K. Li, and S. Farsiu, "Run: Reversible unfolding network for concealed object segmentation," *arXiv preprint arXiv:2501.18783*, 2025.

[52] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4015–4026.

[53] X. Hu, S. Wang, X. Qin, H. Dai, W. Ren, D. Luo, Y. Tai, and L. Shao, "High-resolution iterative feedback network for camouflaged object detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 1, 2023, pp. 881–889.

[54] R. Cong, M. Sun, S. Zhang, X. Zhou, W. Zhang, and Y. Zhao, "Frequency perception network for camouflaged object detection," in *Proceedings of the ACM International Conference on Multimedia*, 2023, pp. 1179–1189.

[55] Z. Huang, H. Dai, T.-Z. Xiang, S. Wang, H.-X. Chen, J. Qin, and H. Xiong, "Feature shrinkage pyramid for camouflaged object detection with transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 5557–5566.

[56] W. Liu, X. Shen, C.-M. Pun, and X. Cun, "Explicit visual prompting for low-level structure segmentations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 19 434–19 445.

[57] B. Yin, X. Zhang, D.-P. Fan, S. Jiao, M.-M. Cheng, L. Van Gool, and Q. Hou, "Camoformer: Masked separable attention for camouflaged object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[58] X. Zhao, Y. Pang, W. Ji, B. Sheng, J. Zuo, L. Zhang, and H. Lu, "Spider: A unified framework for context-dependent concept understanding," *arXiv preprint arXiv:2405.01002*, 2024.

[59] Z. Yu, X. Zhang, L. Zhao, Y. Bin, and G. Xiao, "Exploring deeper! segment anything model with depth perception for camouflaged object detection," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 4322–4330.

[60] K. Sun, Z. Chen, X. Lin, X. Sun, H. Liu, and R. Ji, "Conditional diffusion models for camouflaged and salient object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 47, no. 4, pp. 2833–2848, 2025.

[61] Z. Zhou, Y. Li, C. Zhong, J. Huang, J. Pei, H. Li, and H. Tang, "Rethinking detecting salient and camouflaged objects in unconstrained scenes," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 22 372–22 382.

[62] G. Ren, H. Liu, M. Lazarou, and T. Stathaki, "Multi-modal segment anything model for camouflaged scene segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 19 882–19 892.

[63] Z. Yu, L. Zhao, G. Xiao, and X. Zhang, "Sam-ttt: Segment anything model via reverse parameter configuration and test-time training for camouflaged object detection," in *Proceedings of the 33rd ACM International Conference on Multimedia*, 2025, pp. 4030–4038.

[64] R. Ranftl, A. Bochkovskiy, and V. Koltun, "Vision transformers for dense prediction," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 12 179–12 188.

[65] T.-N. Le, T. V. Nguyen, Z. Nie, M.-T. Tran, and A. Sugimoto, "Anabranch network for camouflaged object segmentation," *Computer vision and image understanding*, vol. 184, pp. 45–56, 2019.

[66] P. Skurowski, H. Abdulameer, J. Błaszczyk, T. Depta, A. Kornacki, and P. Kozieł, "Animal camouflage analysis: Chameleon database," *Unpublished manuscript*, vol. 2, no. 6, p. 7, 2018.

[67] L. Li, E. Rigall, J. Dong, and G. Chen, "Mas3k: An open dataset for marine animal segmentation," in *International Symposium on Benchmarking, Measuring and Optimization*. Springer, 2020, pp. 194–212.

[68] Z. Fu, R. Chen, Y. Huang, E. Cheng, X. Ding, and K.-K. Ma, "Masnet: A robust deep marine animal segmentation network," *IEEE Journal of Oceanic Engineering*, 2023.

[69] M. J. Islam, P. Luo, and J. Sattar, "Simultaneous enhancement and super-resolution of underwater imagery for improved visual perception," *arXiv preprint arXiv:2002.01155*, 2020.

[70] P. Drews-Jr, I. d. Souza, I. P. Maurell, E. V. Protas, and S. S. C. Botelho, "Underwater image segmentation in the wild using deep learning," *Journal of the Brazilian Computer Society*, vol. 27, pp. 1–14, 2021.

[71] D. Hendrycks and T. Dietterich, "Benchmarking neural network robustness to common corruptions and perturbations," *arXiv preprint arXiv:1903.12261*, 2019.

[72] K. Huang, L. Fang, and C. Tian, "Learning to adapt using test-time images for salient object detection in optical remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[73] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pvt v2: Improved baselines with pyramid vision transformer," *Computational Visual Media*, vol. 8, no. 3, pp. 415–424, 2022.

[74] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[75] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," in *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

[76] M. J. Mirza, J. Micorek, H. Possegger, and H. Bischof, "The norm must go on: Dynamic unsupervised domain adaptation by normalization," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 14 765–14 775.

[77] F. Huang, S. Song, and L. Zhang, "Gradient harmonization in unsupervised domain adaptation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 12, pp. 10 319–10 336, 2024.

[78] Y. Mansour, X. Zhong, S. Caglar, and R. Heckel, "Ttt-mim: Test-time training with masked image modeling for denoising distribution shifts," in *European Conference on Computer Vision*. Springer, 2024, pp. 341–357.

[79] P. Zhang, T. Yan, Y. Liu, and H. Lu, "Fantastic animals and where to find them: Segment any marine animal with dual sam," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 2578–2587.

[80] T. Yan, Z. Wan, X. Deng, P. Zhang, Y. Liu, and H. Lu, "Mas-sam: Segment any marine animal with aggregated features," *arXiv preprint arXiv:2404.15700*, 2024.

[81] M. Zha, W. Qian, W. Yang, and Y. Xu, "Multifeature transformation and fusion-based ship detection with small targets and complex backgrounds," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[82] M. Zha, W. Qian, W. Yi, and J. Hua, "A lightweight yolov4-based forestry pest detection method using coordinate attention and feature fusion," *Entropy*, vol. 23, no. 12, p. 1587, 2021.

[83] M. Zha, T. Li, G. Wang, P. Wang, Y. Wu, Y. Yang, and H. T. Shen, "Implicit counterfactual learning for audio-visual segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 22 349–22 360.

[84] Y. Pei, J. Tang, Q. Tang, M. Zha, D. Xie, G. Wang, Z. Liu, N. Xie, P. Wang, Y. Yang *et al.*, "Emotion recognition in hmds: A multi-task approach using physiological signals and occluded faces," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 5977–5986.

[85] Y. Pei, R. Huang, M. Zha, G. Wang, P. Wang, Q. Kang, Y. Yang, and H. T. Shen, "Attentionar: Ar adaptation and warning for real-world safety via attention modeling and mllm reasoning," in *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology*, 2025, pp. 1–19.