Occlusion-Aware Diffusion for Amodal Camouflaged Object Completion

Thi Anh Tho Nguyen[1,2], Sinh Van Nguyen[1,2] and Thi-Ngoc-Hanh Le[1,2⋆]

[1] International University, VNU-HCM, Ho Chi Minh City, Vietnam
[2] Vietnam National University Ho Chi Minh City, Viet Nam
itcsiu21236@student.hcmiu.edu.vn, nvsinh@hcmiu.edu.vn, ltnhanh@hcmiu.edu.vn

Abstract. Camouflaged Object Detection (COD) aims to identify targets whose appearance closely resembles the surrounding environment. Although recent methods perform well on visible-region segmentation, they remain limited to modal perception and often produce fragmented masks when objects are partially occluded. Existing amodal segmentation approaches can infer hidden geometry but generally do not reconstruct the missing appearance, while standard inpainting lacks explicit awareness of target identity and foreground occluders. In this paper, we propose a training-free framework for detecting and completing partially occluded camouflaged objects. The framework consists of three phases: semantic-guided visible object extraction, occlusion context construction, and iterative occlusion - aware completion. Our major contribution is in the final phase, which integrates two references with semantic guidance and an occlusion constraint in a diffusion-based completion process. An iterative verification mechanism further refines incomplete boundaries and unresolved regions. Experiments and evaluations show that the proposed framework generates coherent amodal masks and semantically consistent object appearances.

Keywords: amodal, COD, occlusion, camouflage, completion

1 Introduction

Objects that visually blend into their surroundings remain difficult to perceive because their color, texture, and illumination provide only weak cues for separating foreground from background. This challenge motivates Camouflaged Object Detection (COD), which targets objects with low contrast, disrupted contours, and ambiguous boundaries, unlike generic or salient object detection [26]. Reliable camouflage understanding is important for wildlife monitoring, search and rescue, agricultural robotics, and surveillance, where concealed or partially occluded targets may affect counting, localization, and situational awareness. Recent methods such as SINet [3], PFNet [15], and ZoomNet [17] have improved visible-region segmentation through multi-scale reasoning, distraction suppression, and

⋆ Corresponding author: ltnhanh@hcmiu.edu.vn

boundary refinement. However, they remain modal and recover only observable pixels, producing fragmented masks under occlusion. Amodal reasoning models such as InstaOrderNet [11] estimate hidden geometry or depth order but do not reconstruct missing appearance. Generative methods such as LaMa [22] and latent diffusion [20] can synthesize content, yet may propagate occluder textures or alter target identity. Effective completion therefore requires semantic target identification, explicit occlusion reasoning, context-preserving generation, and iterative verification.

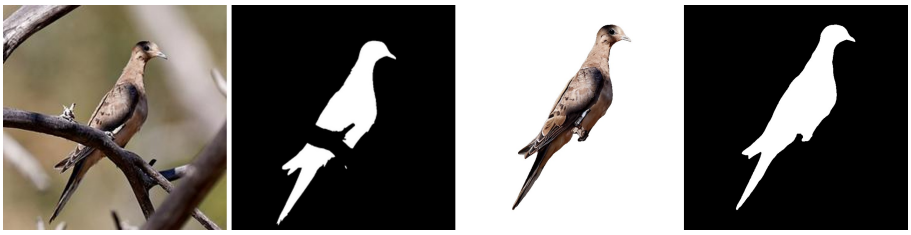


Fig. 1. Our proposed framework can estimate the occluded regions in camouflage image and complete mask. From left to right: *input camouflage object, ground-truth mask, inpainted occluded object, completed occluded mask*

To address these requirements, we propose a training-free unified framework for camouflaged object detection and amodal completion. Instead of training a monolithic network from scarce amodal annotations, the proposed framework coordinates pre-trained foundation models according to three perceptual roles: semantic localization, occlusion-aware context construction, and generative completion. This modular design allows semantic, geometric, and appearance priors to be introduced at the stage where each is most informative. Our proposed framework consists of three phases: The first phase performs *semantic-guided visible object extraction*. The second phase constructs an *occlusion analysis*. And the third phase constitutes the main methodological contribution. An *iterative occlusion-aware completion* strategy combines the clean background, visible object appearance, refined semantic prompt, and occlusion mask within latent diffusion. This design reduces foreground texture leakage and target-identity drift while recovering both the completed image and the corresponding amodal mask.
The main contributions are summarized as follows:

- We formulate camouflaged object understanding as a joint visible detection and amodal completion problem.
- We develop a training-free framework that integrates open-vocabulary localization, promptable segmentation, occlusion ordering, background reconstruction, and generative synthesis.
- We introduce a iterative occlusion-aware completion strategy that combines semantic, object, background, and occlusion cues to reconstruct hidden object regions.

- We employ iterative verification to improve boundary continuity and geometric completeness under severe occlusion.

The remainder of this paper is organized as follows. Section 2 reviews related work on camouflaged object detection, amodal perception, and generative completion. Section 3 presents the proposed three-phase framework. Section 4 reports the experimental results, comparisons, and evaluation. Finally, Section 5 concludes the paper and future directions.

2 Related Work

Camouflaged Object Detection (COD) aims to identify objects that suppress visual distinction from their surroundings through similar color, texture, illumination, or structural patterns [26]. Representative methods include SINet [3], which decomposes detection into search and identification, PFNet [15], which suppresses distracting background regions, and ZoomNet [17], which aggregates multi-scale information. Although these approaches achieve strong visible-region segmentation, they remain modal and cannot recover object regions hidden behind foreground structures.

Amodal perception extends segmentation toward complete-object reasoning. BCNet [1] models overlapping occluder and occludee layers, while CamoFormer [27] employs transformer-based reasoning to estimate complete instance masks. InstaOrderNet [11] further predicts instance-level occlusion and depth ordering. However, these methods primarily recover geometric masks and rely on supervised amodal annotations, which are scarce for natural camouflage scenes. Their dependence on clear contours and stable shape priors also limits performance when visible regions are weak, fragmented, or visually absorbed by the background. Missing-content reconstruction requires more than texture propagation. Artwork inpainting similarly shows that generated regions must preserve local appearance, semantic structure, and global composition [6].

More broadly, multimedia transformation methods preserve key visual structures while modifying spatial, temporal, or stylistic content. Video resizing adapts frame dimensions without distorting salient objects or scene composition [8, 7], while video resequencing maintains action continuity under temporal reorganization [9]. Similarly, style transfer seeks to alter appearance while preserving semantic structure [10, 19, 6, 25]. These shared principles motivate using the visible target as an appearance anchor during amodal completion.

For large-mask image completion, LaMa [23] employs Fast Fourier Convolutions to capture global frequency-domain context and reconstruct coherent background structures. Latent Diffusion Models [20] further enable text-conditioned generation in a compressed latent space. However, standard inpainting does not explicitly model scene layers. When directly applied to an occluded camouflage image, it may reproduce foreground obstacles, propagate their textures into the target, or generate an object inconsistent with the observed instance. Related multimedia reconstruction methods similarly show that uncontrolled synthesis can damage content identity or introduce spatial inconsistency [16, 6, 5, 24].

Recent vision-language foundation models provide complementary semantic and spatial priors. BLIP [12] generates high-level image descriptions, GroundingDINO [14] localizes objects from natural-language queries, and SAM [4] produces class-agnostic masks from spatial prompts. Their combination separates semantic recognition from boundary extraction and supports open-vocabulary targets. CLIP-based alignment further refines the semantic prompt using the visible object appearance, reducing identity drift during generation. Nevertheless, these models still provide only modal evidence and do not independently reconstruct the hidden object.

Existing COD, amodal perception, multimedia completion, and vision-language methods therefore address only separate parts of the problem. COD models recover visible regions, amodal methods infer geometry, and generative models synthesize missing appearance, but none explicitly coordinate target identity, foreground occlusion, and clean environmental context. To bridge this gap, we propose a training-free framework that integrates semantic localization, occlusion reasoning, background reconstruction, and occlusion-aware diffusion. The proposed completion phase jointly exploits an occluder-free background, a visible-object appearance anchor, semantic guidance, and iterative verification to recover coherent hidden object regions.

3 Methodology

3.1 Problem Formulation

Conventional camouflaged object detection recovers only visible target regions, producing fragmented masks under partial occlusion. Direct inpainting is also inadequate because it does not explicitly distinguish the target, foreground occluders, and the hidden region to be reconstructed. We therefore formulate the task as *amodal camouflaged object completion* and solve it using complementary semantic, geometric, and generative priors.

Let $\mathcal{I} \in \mathbb{R}^{H\times W\times 3}$ denote an observed image composed of background B, complete object appearance O, and foreground occluder F:

$$\mathcal{I} = \mathcal{M}_{\text{occ}} \odot F + (1 - \mathcal{M}_{\text{occ}}) \odot [\mathcal{M}_{\text{amo}} \odot O + (1 - \mathcal{M}_{\text{amo}}) \odot B], \tag{1}$$

where $\mathcal{M}_{\text{amo}}$ is the complete object mask, $\mathcal{M}_{\text{occ}}$ is the foreground occlusion mask, and $\odot$ denotes element-wise multiplication. The directly observable region is

$$\mathcal{M}_{\text{vis}} = \mathcal{M}_{\text{amo}} \odot (1 - \mathcal{M}_{\text{occ}}). \tag{2}$$

Recovering $\mathcal{M}_{\text{amo}}$ and the hidden appearance is ill-posed because multiple completions may explain the same visible evidence.

As shown in Fig.2, the proposed training-free framework contains three phases. First, *semantic-guided visible object extraction* derives a semantic query from $\mathcal{I}$, localizes the target, and predicts its visible mask $\mathcal{M}_{\text{vis}}$. Second, *occlusion analysis* estimates $\mathcal{M}_{\text{occ}}$, reconstructs an occluder-free background $\mathcal{I}_{\text{bg}}$, and retains the

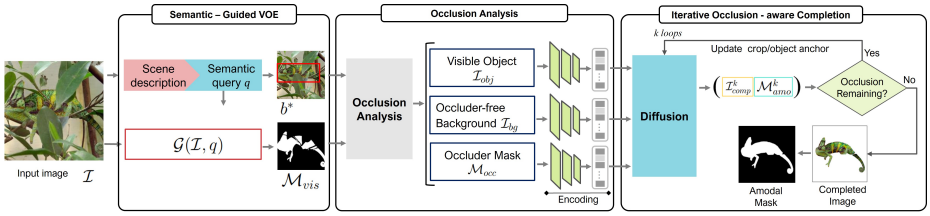


Fig. 2. Overview of the proposed framework.

visible appearance $\mathcal{I}_{\text{obj}}$ as an identity anchor. Third, the proposed *iterative occlusion-aware completion* integrates $\mathcal{I}_{\text{bg}}$, $\mathcal{I}_{\text{obj}}$, $\mathcal{M}_{\text{occ}}$, and a refined semantic prompt to synthesize hidden regions. Iterative verification repeats completion when unresolved occlusion or incomplete boundaries remain. The final outputs are the completed image $\mathcal{I}^*_{\text{comp}}$ and amodal mask $\mathcal{M}^*_{\text{amo}}$.

3.2 Semantic-Guided Visible Object Extraction

Camouflaged objects provide weak local contrast and ambiguous boundaries. Consequently, direct category prediction or segmentation can easily confuse the target with visually similar background regions. We therefore decouple semantic recognition from pixel-level segmentation.

Given $\mathcal{I}$, an image captioning model first produces a scene description $T = \mathcal{C}(\mathcal{I})$, with $\mathcal{C}$ denotes the pre-trained BLIP model [12]. A dependency parser extracts the principal noun phrase from T, yielding a semantic query q. For example, the caption "*a chameleon resting among green leaves*" is reduced to the query "*chameleon*". This operation suppresses background entities that may otherwise compete with the target during grounding. The query q and image $\mathcal{I}$ are then processed by an open-vocabulary grounding model:

$$\mathcal{B} = \mathcal{G}(\mathcal{I}, q) = \{(b_i, s_i) \mid s_i > \tau_b\}, \tag{3}$$

where $\mathcal{G}$ denotes GroundingDINO [13], $b_i \in \mathbb{R}^4$ is a predicted bounding box, s_i is its text-region alignment score, and τ_b is the confidence threshold. Since camouflaged targets commonly receive lower confidence scores than salient objects, a relatively permissive threshold is adopted to avoid discarding weak but semantically relevant detections. From the retained candidate set, the target bounding box is selected as:

$$b^* = \underset{(b_i, s_i) \in \mathcal{B}}{\arg\max}\, s_i. \tag{4}$$

The selected box b^* provides the coarse spatial prior for the subsequent promptable segmentation stage.

A bounding box identifies the approximate target location but cannot describe thin structures or irregular camouflaged boundaries. Each selected box is therefore

supplied as a spatial prompt to the Segment Anything Model (SAM) model [4]:

$$\{\mathcal{M}_j, \rho_j\}_{j=1}^{K} = \mathcal{S}(\mathcal{I}, b^*), \tag{5}$$

where $\mathcal{S}$ denotes SAM, $\mathcal{M}_i$ is a candidate mask, and ρ_i is its predicted mask quality. The final visible mask is obtained from the candidate having the strongest agreement with the grounded region:

$$\mathcal{M}_{\text{vis}} = \mathcal{M}_{j^*}, \qquad j^* = \underset{j \in \{1,\ldots,K\}}{\arg\max}\ \rho_j. \tag{6}$$

This semantic-to-box-to-mask decomposition is important in camouflage scenes. GroundingDINO restricts the search to regions associated with the target identity, while SAM concentrates on class-agnostic boundary recovery within that region. Thus, semantic recognition is not required to produce a precise contour, and segmentation is not required to independently infer the object category.

3.3 Occlusion Analysis

The visible mask alone does not determine where the object should be extended. Naively dilating $\mathcal{M}_{\text{vis}}$ assumes that every neighboring background pixel is a missing object part, whereas directly inpainting all intersecting structures may remove valid object regions. We instead explicitly estimate the scene layers surrounding the target.

Let $\mathcal{M} = \{\mathcal{M}_1, \ldots, \mathcal{M}_n\}$ denote the set of instance masks detected in the target neighborhood, including $\mathcal{M}_{\text{vis}}$. To identify foreground structures that occlude the target, we employ the object-detection variant of InstaOrderNet [11]. Given the input crop and the visible target mask, the model detects neighboring instances and estimates their pairwise depth ordering relative to the target. For each neighboring instance $\mathcal{M}_j$, the predicted relation is

$$r_j = \mathcal{O}\left(\mathcal{I}_{\text{crop}}, \mathcal{M}_j, \mathcal{M}_{\text{vis}}\right), \qquad r_j \in \{\text{front}, \text{behind}, \text{none}\}, \tag{7}$$

where $\mathcal{O}$ denotes InstaOrderNet. The occluder mask is then formed by the union of instances predicted to lie in front of the target:

$$\mathcal{M}_{\text{occ}} = \bigcup_{j:r_j=\text{front}} \mathcal{M}_j. \tag{8}$$

Small inaccuracies along occlusion boundaries may expose foreground colors during generation. We therefore enlarge the occluder mask using morphological dilation:

$$\tilde{\mathcal{M}}_{\text{occ}} = \mathcal{M}_{\text{occ}} \oplus K_d, \tag{9}$$

where K_d is a small structuring element. The resulting mask defines the region in which hidden object content may be synthesized.

A second issue is that the original image contains the very occluders that the completion module must reason behind. If these pixels are retained as generation

context, their textures may leak into the completed object. We construct an occluder-free background reference using LaMa:

$$\mathcal{M}_{\text{rem}} = \mathcal{M}_{\text{vis}} \cup \tilde{\mathcal{M}}_{\text{occ}}, \tag{10}$$

$$\mathcal{I}_{\text{bg}} = \mathcal{L}\left(\mathcal{I} \odot (1 - \mathcal{M}_{\text{rem}}), \mathcal{M}_{\text{rem}}\right), \tag{11}$$

where $\mathcal{L}$ denotes the LaMa inpainting model. Removing both the visible target and its foreground occluders yields a clean estimate of the environmental layer. This background is not treated as the final output; instead, it acts as a structural reference that prevents branches, foliage, or other occluding textures from being copied into the generated object.
The visible object appearance is retained separately as

$$\mathcal{I}_{\text{obj}} = \mathcal{I} \odot \mathcal{M}_{\text{vis}}. \tag{12}$$

Consequently, the generative module receives two contexts with different roles: $\mathcal{I}_{\text{bg}}$ preserves the surrounding scene, whereas $\mathcal{I}_{\text{obj}}$ anchors the target identity, color, and texture.

3.4 Iterative Occlusion-aware Completion

The final phase reconstructs both the hidden appearance and the complete shape of the camouflaged object. A standard diffusion inpainter conditioned only on the occlusion mask may generate plausible content but often changes the target identity. Conversely, using the original image as strong context may preserve the foreground occluder and prevent completion. To address these limitations, the proposed method combines semantic prompt refinement, mixed-context diffusion, and iterative verification within a unified amodal completion process.
Before generation, the semantic query is refined using the visible object appearance. Let E_I and E_T denote the CLIP image and text encoders, respectively. Given a set of candidate descriptions $\mathcal{Q}$, the description most compatible with the visible target is selected as

$$q^* = \arg\max_{q_c \in \mathcal{Q}} \frac{E_I(\mathcal{I}_{\text{obj}})^\top E_T(q_c)}{\|E_I(\mathcal{I}_{\text{obj}})\|_2 \|E_T(q_c)\|_2}. \tag{13}$$

The refined description q^* is then embedded into a photographic prompt and used as the semantic condition for diffusion. This step constrains the generated content to remain consistent with the observed instance rather than only its coarse object category.
Let $\mathcal{E}$ denote the VAE encoder of the latent diffusion model. The clean background and object-anchored contexts are encoded as

$$z_{\text{bg}} = \mathcal{E}(\mathcal{I}_{\text{bg}}), \qquad z_{\text{obj}} = \mathcal{E}\left(\mathcal{I}_{\text{bg}} \odot (1 - \mathcal{M}_{vis}) + \mathcal{I}_{\text{obj}}\right). \tag{14}$$

Here, z_{bg} provides an occluder-free environmental reference, whereas z_{obj} preserves the visible object pixels. At denoising step t, the latent context is formed as

$$z_t^{\text{mix}} = (1 - \mathcal{M}_{\text{occ}}^z) \odot z_{\text{obj},t} + \mathcal{M}_{\text{occ}}^z \odot z_{\text{gen},t}, \tag{15}$$

where $\mathcal{M}^z_{\text{occ}}$ is the dilated occlusion mask resized to the latent resolution, $z_{\text{obj},t}$ is the noised object-anchored latent, and $z_{\text{gen},t}$ is the latent synthesized inside the hidden region. The denoising network predicts

$$\widehat{\epsilon}_t = \epsilon_\theta \left(z_t^{\text{mix}}, t, E_T(q^*), \mathcal{M}^z_{\text{occ}} \right). \tag{16}$$

This formulation preserves the observed target outside the occluded region while allowing new content to be synthesized only where completion is required. The visible object acts as an identity anchor, the clean background maintains environmental consistency, and the refined prompt constrains semantic plausibility. Their combination reduces identity drift and prevents texture leakage from foreground occluders.

After reverse diffusion, the completed crop is decoded as

$$\mathcal{I}^k_{\text{comp}} = \mathcal{D} \left(z_0^k \right), \tag{17}$$

where $\mathcal{D}$ is the VAE decoder and k denotes the completion iteration. The corresponding amodal mask $\mathcal{M}^k_{\text{amo}}$ is obtained by re-segmenting the completed target using the semantic query and updated spatial prompt.

Because a single pass may leave unresolved occlusions or truncate the object at the crop boundary, the completion is embedded in an iterative perception-generation-verification loop. At iteration k, boundary closure is measured as

$$C^k_{\text{bd}} = \sum_{p \in \partial\Omega_k} \mathcal{M}^k_{\text{amo}}(p), \tag{18}$$

where $\partial\Omega_k$ denotes the crop boundary. A positive value indicates that the predicted object may extend beyond the current crop. The remaining occlusion ratio is computed as

$$C^k_{\text{occ}} = \frac{\left| \tilde{\mathcal{M}}^k_{\text{occ}} \cap \mathcal{M}^k_{\text{amo}} \right|}{|\mathcal{M}^k_{\text{amo}}| + \varepsilon}. \tag{19}$$

The procedure terminates when

$$C^k_{\text{bd}} = 0 \quad \text{and} \quad C^k_{\text{occ}} < \delta, \tag{20}$$

or when the maximum number of iterations $K_{\max}$ is reached. Otherwise, the crop is expanded, the current amodal mask becomes the object anchor for the next iteration, and occlusion analysis and mixed-context generation are repeated:

$$\left(\mathcal{I}^{(k+1)}_{\text{comp}}, \mathcal{M}^{(k+1)}_{\text{amo}} \right) = \mathcal{F} \left(\mathcal{I}^k_{\text{comp}}, \mathcal{M}^k_{\text{amo}}, q^* \right). \tag{21}$$

Unlike unconstrained repeated inpainting, each iteration verifies the reconstructed boundary and remaining occlusion before generating additional content. The process therefore expands the object only when the current reconstruction remains incomplete. The final outputs are the completed image $\mathcal{I}^*_{\text{comp}}$ and the amodal mask $\mathcal{M}^*_{\text{amo}}$ obtained from the last valid iteration.

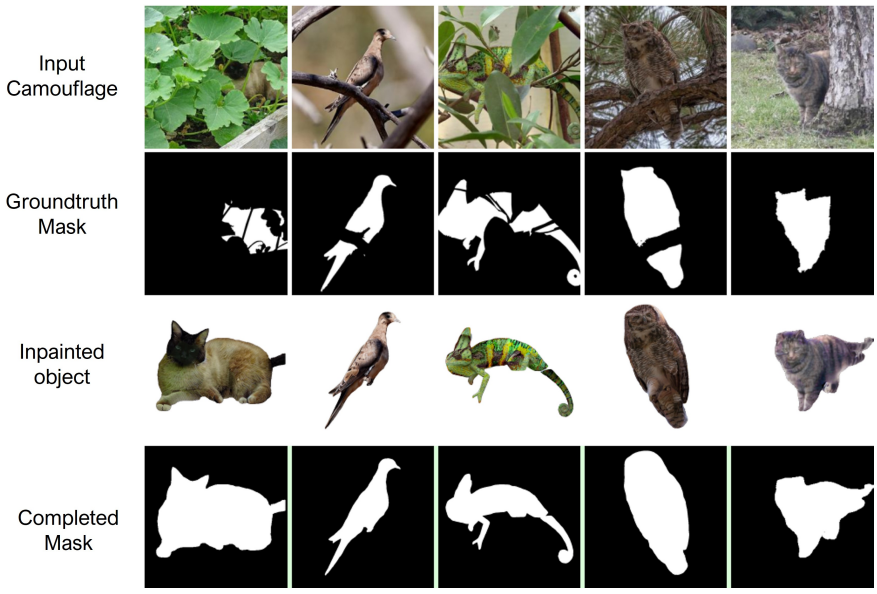


Fig. 3. Our model performs on various input camouflage images

4 Experimental Results

4.1 Implementation Details

Our framework was implemented in PyTorch and evaluated on an NVIDIA A100 GPU using half-precision inference where supported. GroundingDINO with a Swin-T backbone and SAM with ViT-H were used for localization and visible-mask extraction, respectively. InstaOrderNet-OD estimated instance-level occlusion order, while `big-lama` reconstructed the clean background. Completion employed Stable Diffusion Inpainting v1.5 with DDIM sampling, 50 denoising steps, and a guidance scale of 7.5. All components were used without task-specific training or fine-tuning. In terms of data, our experiments used 100 challenging images curated from COD10K, CAMO, and MS-COCO. The subset includes diverse camouflaged objects, strong foreground-background similarity, and controlled occlusion cases. Only partially occluded instances were retained to emphasize ambiguous boundaries and missing object regions.

4.2 Evaluation and Comparison

Qualitative results and comparisons. Figure 3 shows representative results across diverse object categories and occlusion patterns. The method converts fragmented visible regions into coherent silhouettes for heavily concealed targets, while preserving body continuity and thin structures in elongated objects such as birds. It also reconstructs large missing regions under severe block occlusion, as

observed for the owl and the cat near the tree trunk. Although the synthesized content is only a plausible estimate of the hidden appearance, the results demonstrate consistent preservation of category identity, topological connectivity, and complete object geometry across challenging camouflage conditions.

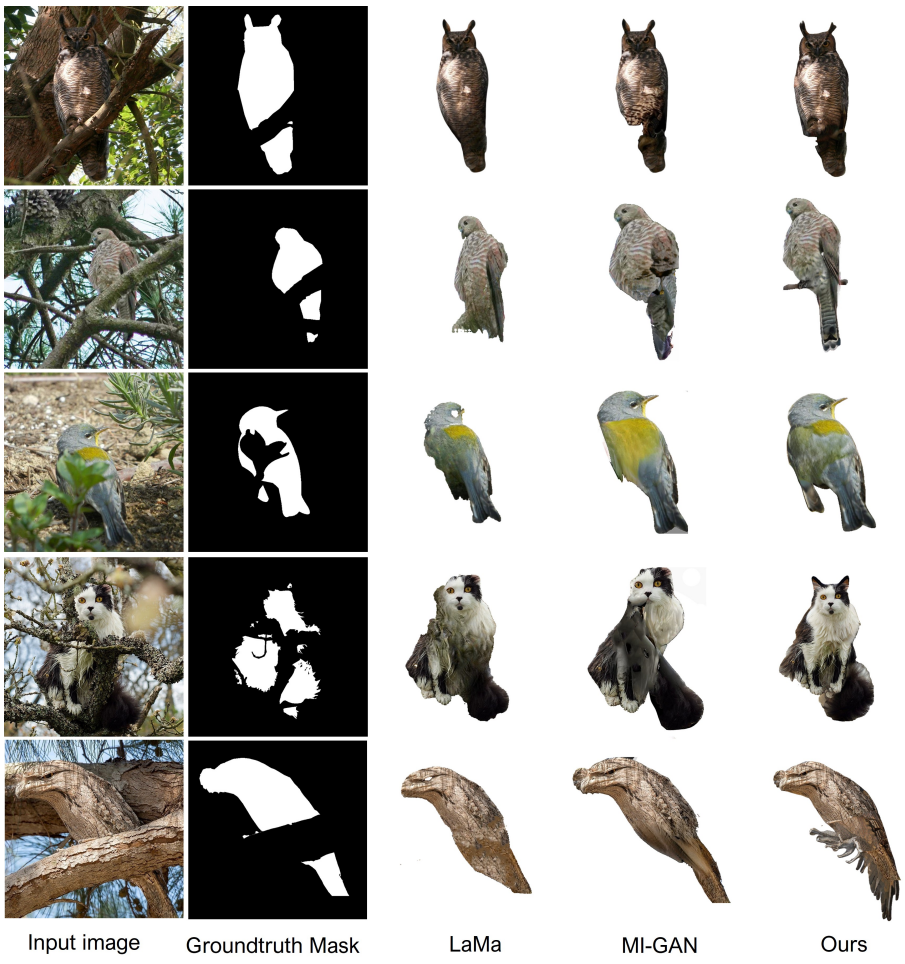


Fig. 4. Our method competes with prior amodal completion approaches

Figure 4 compares our method with LaMa [23] and MI-GAN [21] across diverse object shapes and occlusion patterns. LaMa produces smooth completions but often oversimplifies distinctive geometry, while MI-GAN preserves more local detail yet introduces incomplete boundaries, distorted proportions, and residual artifacts. In contrast, the proposed framework better maintains category-specific structure and visible-instance identity, particularly for elongated tails, limbs, and

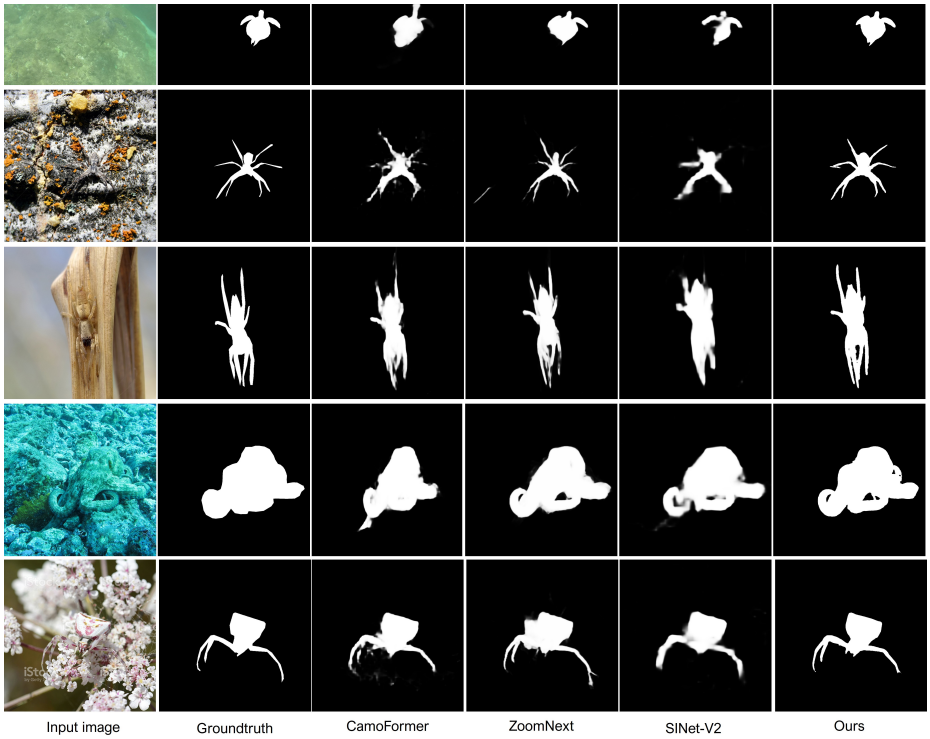


Fig. 5. Visual comparison of visible camouflaged object extraction with representative COD methods.

partially hidden bodies. Although the reconstructed regions are only plausible estimates of the unobserved content, our method achieves a stronger balance among structural completeness, semantic consistency, and texture fidelity.

A part from above, we show that the proposed framework remains competitive in the visible COD extraction stage in Fig. 5. Across small targets, thin structures, and highly cluttered backgrounds, the predicted masks are generally sharper and more complete than those of CamoFormer [27], ZoomNext [18], and SINet-V2 [2]. In particular, the method better preserves fine object parts such as legs, antennae, tails, and irregular contours, while suppressing background noise. These results indicate that, although the framework is designed primarily for amodal completion, its semantic-guided localization and promptable segmentation stages also provide reliable modal COD masks as a strong basis for subsequent occlusion reasoning and reconstruction.

Quantitative comparisons. To complement the qualitative comparisons, we evaluate both reconstructed geometry and synthesized appearance. Amodal mask quality is measured using S_α and MAE, where higher S_α and lower MAE indicate better structural agreement and pixel-wise accuracy. Since the hidden content is not directly observable, the generated regions are regarded as semantically plau-

sible completions rather than exact ground-truth reconstructions. Table 1 shows that the proposed training-free framework achieves the best amodal segmentation performance, obtaining the highest S_α and lowest MAE. This indicates that explicit occlusion reasoning and completion recover more coherent full-object geometry than the compared supervised baselines, particularly in hidden regions.

Table 1. Amodal segmentation comparison on the synthetically occluded evaluation subset.

Method	Learning setting	$S_\alpha \uparrow$	MAE$\downarrow$
LaMa [22]	Supervised modal	0.856	0.143
MI-GAN [21]	Supervised modal	0.840	0.159
Proposed	Training-free amodal	**0.859**	**0.112**

Ablation study. Figure 6 illustrates the contribution of each component to completion quality. We can see that without InstaOrderNet, the model cannot reliably distinguish foreground occluders from the target, resulting in large missing regions and disconnected object parts. Removing LaMa causes the completion to rely on contaminated context, leading to incomplete structure and residual artifacts. A single refinement round improves object continuity but still leaves local inconsistencies around the reconstructed body. The full configuration produces the most coherent silhouette and appearance, confirming that occlusion ordering, background cleaning, and iterative refinement are complementary components for recovering complete object geometry.



Fig. 6. Ablated results

Limitation. Figure 7 illustrates a failure under extremely dense occlusion. The initial segmentation captures only sparse visible regions, providing an unreliable structural anchor. Consequently, the completion stage hallucinates an incorrect object shape and retains occluder-related artifacts. This highlights the framework's sensitivity to inaccurate visible masks and insufficient target evidence.



Fig. 7. A sample of our failure

5 Conclusion

In this work, we address the limitation of conventional camouflaged object detection under partial occlusion. Existing COD methods primarily recover visible pixels and therefore produce fragmented masks when foreground structures hide parts of the target. To move beyond modal perception, we proposed a training-free framework that integrates semantic localization, visible-object segmentation, occlusion reasoning, background reconstruction, and generative completion. Experimental results showed that the proposed framework can recover coherent object shapes and plausible hidden appearance across different camouflage patterns, object categories, and occlusion configurations. Quantitative comparisons and ablation studies confirmed that occlusion ordering improves structural reconstruction, background cleaning reduces texture leakage, and iterative refinement contributes most strongly to amodal geometric completeness. The main limitation is that the reconstructed hidden content is semantically plausible rather than guaranteed to match the true unobserved appearance. In addition, the use of multiple foundation models increases inference cost and may propagate errors between phases. Future work may therefore focus on building larger amodal camouflage datasets, improving computational efficiency, and introducing lightweight task-specific adaptation. Extending the framework to multiple interacting objects and video sequences would also support more realistic ecological, robotic, and surveillance applications.

Acknowledgement

We acknowledge International University, VNU-HCM for supporting this study.

References

1. Chen, T., Xiao, J., Hu, X., Zhang, G., Wang, S.: Boundary-guided context-aware network for camouflaged object detection. Neural Computing and Applications (2023)

2. Fan, D.P., Ji, G.P., Cheng, M.M., Shao, L.: Concealed object detection. IEEE transactions on pattern analysis and machine intelligence **44**(10), 6024–6042 (2021)
3. Fan, D.P., Ji, G.P., Sun, G., Cheng, M.M., Shen, J., Shao, L.: Camouflaged object detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
4. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4015–4026 (2023)
5. Le, K.Q.B., Le, T.N.H.: Clip-based dual-encoder approach for sketch-to-image retrieval. In: Asian Conference on Intelligent Information and Database Systems. pp. 507–518. Springer (2026)
6. Le, T.N.H., Chen, Y.H., Lee, T.Y.: Structure-aware video style transfer with map art. ACM Transactions on Multimedia Computing, Communications and Applications **19**(3s), 1–25 (2023)
7. Le, T.N.H., Huang, H., Chen, Y.R., Lee, T.Y.: Retargeting video with an end-to-end framework. IEEE Transactions on Visualization and Computer Graphics **30**(9), 6164–6176 (2024). https://doi.org/10.1109/TVCG.2023.3327825
8. Le, T.N.H., Lee, T.Y., Lin, S.S., Dong, W.: Deep learning-based importance map for content-aware media retargeting. Multimedia Tools and Applications **83**(30), 74301–74322 (2024)
9. Le, T.N.H., Yao, S.Y., Wu, C.T., Lee, T.Y.: Regenerating arbitrary video sequences with distillation path-finding. IEEE Transactions on Visualization and Computer Graphics **30**(7), 3622–3635 (2024). https://doi.org/10.1109/TVCG.2023.3237739
10. Le, T.N.H., Yeh, C.K., Lin, Y.C., Lee, T.Y.: Animating still natural images using warping. ACM Transactions on Multimedia Computing, Communications and Applications **19**(1), 1–24 (2023)
11. Lee, H., Park, J.: Instance-wise occlusion and depth orders in natural scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2022)
12. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International conference on machine learning. pp. 12888–12900. PMLR (2022)
13. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., et al.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In: European conference on computer vision. pp. 38–55. Springer (2024)
14. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., et al.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. arXiv preprint arXiv:2303.05499 (2023)
15. Mei, H., Ji, G.P., Wei, Z., Yang, X., Wei, X., Fan, D.P.: Camouflaged object segmentation with distraction mining. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8772–8781 (June 2021)
16. Nguyen, T.H., Le, T.N.H.: Cross-quantized hyperbolic representations for enhancing cartoon image retrieval. In: Asian Conference on Intelligent Information and Database Systems. pp. 384–398. Springer (2026)
17. Pang, Y., Zhao, X., Xiang, T.Z., Zhang, L., Lu, H.: Zoom in and out: A mixed-scale triplet network for camouflaged object detection (2022)
18. Pang, Y., Zhao, X., Xiang, T.Z., Zhang, L., Lu, H.: Zoomnext: A unified collaborative pyramid network for camouflaged object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence (2024). https://doi.org/10.1109/TPAMI.2024.3417329

19. Pham, K.G., Van Nguyen, S., et al.: Material-aware style transfer for lacquer paintings. In: 2025 RIVF International Conference on Computing and Communication Technologies (RIVF). pp. 102–107. IEEE (2025)
20. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models (2021)
21. Sargsyan, A., Navasardyan, S., Xu, X., Shi, H.: Mi-gan: A simple baseline for image inpainting on mobile devices. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7335–7345 (2023)
22. Suvorov, R., Logacheva, E., Mashikhin, A., Remizova, A., Ashukha, A., Silvestrov, A., Kong, N., Goka, H., Park, K., Lempitsky, V.: Resolution-robust large mask inpainting with fourier convolutions. arXiv preprint arXiv:2109.07161 (2021)
23. Suvorov, R., Logacheva, E., Mashikhin, A., Remizova, A., Ashukha, A., Silvestrov, A., Kong, N., Goka, H., Park, K., Lempitsky, V.: Resolution-robust large mask inpainting with fourier convolutions. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 2149–2159 (2022)
24. Van Nguyen, S., Nguyen, V.X., Le, T.N.H.: An improved method for enhancing images quality based on convolution efficient transformer. SN Computer Science **7**(4), 347 (2026)
25. Wu, K., Tang, F., Liu, N., Deussen, O., Le, T.N.H., Dong, W., Lee, T.Y.: Lighting image/video style transfer methods by iterative channel pruning. In: ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 3800–3804. IEEE (2024)
26. Xiao, F., Hu, S., Shen, Y., Fang, C., Huang, J., He, C., Tang, L., Yang, Z., Li, X.: A survey of camouflaged object detection and beyond (2024), https://arxiv.org/abs/2408.14562
27. Yin, B., Zhang, X., Fan, D.P., Jiao, S., Cheng, M.M., Van Gool, L., Hou, Q.: Camoformer: Masked separable attention for camouflaged object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence **46**(12), 10362–10374 (2024)