

HIDA: A Human-Intuition-Guided Depth-Aware Framework for Zero-Shot Amodal Segmentation

Peng Li¹, Kelin Wang¹, Bingchuan Chen², Ya-Li Hou², Mingxia Shen¹, and Bo Li^{1*}

¹ Nanjing Agricultural University, Nanjing, China

² Beijing Jiaotong University, Beijing, China

{li_peng, wangkelin}@stu.njau.edu.cn; {mingxia, libo}@njau.edu.cn
{bcchen, ylhnu}@bjtu.edu.cn

Abstract. Amodal segmentation aims to recover the full shape of objects that are partially occluded; however, most existing approaches are trained and assessed on in-domain benchmarks and exhibit limited generalization in zero-shot scenarios. Human-intuition-guided depth-aware (HIDA) is proposed as a plug-and-play framework that combines two frozen vision foundation models with a lightweight trainable segmentation network to estimate amodal masks using only a visible bounding box. HIDA initially derives a visible prior mask and a pseudo-depth map (PDM) to enable target-centered cropping, which reduces irrelevant context while highlighting occlusion boundaries and hidden regions. An occlusion-aware token-routing soft MoE module is also introduced within a lightweight ViT encoder to adaptively handle varied occlusion patterns and strengthen representations of invisible regions. Evaluations on COCOA-cls, KINS, and D2SA demonstrate that HIDA attains state-of-the-art performance in fully supervised settings and state-of-the-art zero-shot results on COCOA-cls and D2SA while training only a lightweight segmentation network. Code is available at: <https://github.com/lipenghuai/HIDA>.

Keywords: Amodal segmentation · Occlusion-aware Soft MoE · Zero-shot Learning

1 Introduction

Humans possess an inherent ability for amodal perception [16, 31], which enables rapid reconstruction and inference of an object’s complete shape from its visible parts, even when the object is partially occluded. In computer vision, amodal perception has likewise become increasingly important for practical applications, including autonomous driving [3, 15, 35], robotic grasping [14, 37, 42, 47], image enhancement [13, 22, 49], and scene understanding [4, 7, 40]. Occlusion understanding is, therefore, a core capability for visual intelligence systems.

* Corresponding author.



Fig. 1: Zero-shot qualitative results of HIDA. Without any training or fine-tuning on the target dataset, HIDA predicts amodal masks from a visible bounding box by leveraging pseudo-depth guidance and occlusion-aware token-routing.

As interest in amodal perception has increased, amodal segmentation has received substantial attention in recent years, and numerous methods have been explored to support various applications [41]. Many previous studies remain limited to closed-world settings and do not achieve zero-shot generalization, which does not satisfy practical requirements [1]. SAMBA [29] was introduced to address this limitation as a foundation model for zero-shot segmentation. SAMBA is produced by fine-tuning SAM [19] with additional structures added on top of the SAM encoder, leading to strong zero-shot performance. This approach, however, remains bounded by the SAM architecture, indicating a need for a more flexible design to further unlock the potential of vision foundation models.

Human visual intuition suggests that occlusion relationships among objects follow a straightforward geometric principle: only objects that are nearer to the imaging plane in 3D space can occlude those located behind them. Occlusion truncates the visible mask on the image plane but does not disrupt the global 3D continuity of the occluded target [17, 26]. Modal and amodal mask prediction are intrinsically linked. Amodal segmentation uses information from visible regions to form plausible completions, and the accuracy of modal segmentation directly affects amodal segmentation quality [10]. Building on this motivation, amodal segmentation is performed using only a visible bounding box of the target by jointly leveraging a pseudo-depth map (PDM) estimated by Depth Anything V2 [44] and a high-quality visible mask generated by SAM. The PDM provides explicit cues about relative front-back ordering in occluded scenes, allowing more reliable pixel-level separation of occluders from occluded targets. The visible contour simultaneously acts as a geometric anchor that accurately constrains the target’s position and local shape in the image. This formulation shifts the learning challenge toward extrapolating and completing invisible regions, making contour prediction more aligned with human perception: the visible portion is identified first, followed by inference of the complete object shape.

Existing amodal segmentation methods, despite recent progress, remain limited by substantial training costs and still do not reach the desired level of accuracy. Informed by these observations, a two-stage amodal segmentation framework, human-intuition-guided depth-aware (HIDA) modal segmentation,

is developed. HIDA employs two frozen, pretrained foundation models as plug-and-play prior extractors and trains only a lightweight segmentation network, without fine-tuning or architectural changes to the foundation models. Adaptive image cropping is first applied using the visible mask and PDM to concentrate on the target region. An occlusion-aware token-routing soft MoE module [25, 34] is then incorporated into a lightweight ViT encoder [8] to strengthen representations of occlusion boundaries and occluded regions, which supports more stable and accurate amodal contour prediction.

The proposed approach is evaluated on three primary datasets: COCOA-cls [50], KINS [35], and D2SA [10]. The experimental results indicate that HIDA achieves state-of-the-art performance in the fully supervised setting and state-of-the-art zero-shot performance on COCOA-cls and D2SA. HIDA also demonstrates high efficiency in training speed. The contributions are summarized as follows:

- A depth-aware amodal segmentation framework, HIDA amodal segmentation, is introduced, combining two frozen, pretrained vision foundation models with a lightweight, trainable segmentation network. Lightweight interfaces enable interaction among the three modules while keeping the foundation models frozen to support plug-and-play use of off-the-shelf backbones.
- Adaptive image cropping combines visible-mask structure with pseudo-depth occlusion cues to produce informative target regions, while a token-routing soft MoE improves segmentation and visualizes occlusion-aware token dispatch.
- Experiments on COCOA-cls, KINS, and D2SA show that HIDA attains state-of-the-art results under fully supervised training and state-of-the-art zero-shot performance on COCOA-cls and D2SA while requiring only limited computational resources during training.

2 Related Work

2.1 Amodal Segmentation

Amodal segmentation seeks to estimate object shapes that include both the visible parts and the regions hidden by occlusion in partially occluded objects. This capability is fundamental to visual intelligence and can also be used as an auxiliary component for 3D reconstruction and novel view synthesis [6]. Amodal segmentation was first introduced by Li and Malik [21], and many methods have since been proposed. Beyond direct end-to-end optimization in fully supervised settings [21, 35, 50], researchers have improved segmentation performance by incorporating various auxiliary cues, such as depth relations [46, 48] and shape priors [6, 11, 22, 43]. Weakly supervised methods rely on cheaper forms of annotation [20, 27, 38], such as bounding boxes and category labels, but their performance is often limited by insufficient supervision signals.

In general, these methods are limited by expensive annotations or category-specific distributions, which reduces their ability to generalize to open-world

scenarios. Recent advances in vision foundation models have motivated multiple approaches that leverage these paradigms for amodal segmentation [29, 46]. Zero-shot methods, in particular, utilize large-scale synthetic datasets or foundation models to effectively predict amodal masks for occluded objects with a certain degree of generalization.

Despite encouraging performance, most existing approaches either depend on a single standalone model or obtain modal masks from a foundation model and then pair a generative model to infer amodal masks, which does not fully leverage the capabilities of vision foundation models. HIDA is, therefore, developed as a two-part amodal segmentation framework that integrates two pretrained vision foundation models with a lightweight trainable segmentation network. This design achieves competitive results under both fully supervised and zero-shot protocols.

2.2 Vision Foundation Models

The rapid progress of vision foundation models in recent years has attracted broad interest within the computer vision community [2]. These foundation models [5, 19, 44] are generally trained on large-scale datasets and exhibit strong generalization ability. For instance, SAM [19] achieves class-agnostic image segmentation by training on massive datasets, substantially improving its performance in zero-shot scenarios. Moreover, Depth Anything V2 [44] provides efficient zero-shot depth estimation by leveraging large-scale pseudo-depth data. Such vision foundation models encode rich implicit priors and generalize well, which is also critical for zero-shot amodal segmentation. Therefore, we aim to fully exploit the priors in SAM and Depth Anything V2 to facilitate amodal segmentation.

Foundation models have motivated early research across a wide range of areas [28, 29, 39], and have also been explored to some extent for amodal segmentation. Existing studies, however, largely focus on fine-tuning a single foundation model or using SAM to extract modal masks for subsequent amodal prediction, which leaves the broader potential of foundation models insufficiently explored.

2.3 Soft MoE

Soft MoE [34] is a deep learning technique that integrates mixture-of-experts with a soft selection mechanism to increase model expressiveness. Conventional MoE [36] typically applies hard selection to determine which experts are activated, whereas soft MoE adopts soft selection so that multiple experts can contribute in a graded manner during each inference pass, mitigating performance limitations associated with hard routing. A K -dimensional routing distribution is generated at each spatial location, providing continuous, normalized weights over all experts rather than selecting a single expert. These weights are conditioned on local appearance and occlusion cues, enabling the network to adjust each expert’s contribution dynamically and to construct a mixture of specialized feature-representation pathways. Consequently, ambiguous areas, including occlusion boundaries, thin structures, and partially visible regions, can be handled

by multiple complementary experts, while more certain regions are typically governed by a smaller subset of experts. This spatially varying soft routing improves robustness under occlusion-related uncertainty and promotes expert specialization across different occlusion patterns and boundary configurations, producing more discriminative and stable representations.

3 Method

3.1 Task Definition

For each instance, the original image I and the visible bounding box B_m as inputs, where B_m is derived from the modal (visible) extent of the target and serves as a spatial prior to localize the region of interest. Conditioned on (I, B_m) , the model outputs a predicted binary amodal mask $\tilde{M}^a \in \{0, 1\}^{H \times W}$, which aims to delineate the complete object silhouette by covering both the visible and occluded parts. Importantly, our formulation is category-agnostic: no category labels are required, and the network learns amodal completion purely from mask supervision, enabling straightforward generalization to novel object categories without class-specific annotations.

3.2 Overall Architecture

An amodal segmentation framework is introduced based on human visual intuition, combining two pretrained vision foundation models with a lightweight trainable segmentation network. As shown in Fig. 2, HIDA is organized into two phases: a trainable phase and a non-trainable phase. Unlike prior methods that fine-tune vision foundation models, we use them strictly as frozen, plug-and-play prior generators, without any fine-tuning or architectural modification. Given an input image $\mathbf{I}$ and a visible bounding box, we first use Depth Anything V2 to generate a pseudo-depth map (PDM) of $\mathbf{I}$, and meanwhile use SAM to produce the visible mask $\tilde{M}^m$ of the target. These two processes can be executed in parallel. We then expand the bounding box based on the foundation model outputs (i.e., the PDM and the visible contour) to obtain a full-object box. Using this expanded box, we crop $\mathbf{I}$, PDM, and the modal mask $\tilde{M}^m$ to obtain cropped $\mathbf{I}$, cropped PDM, and $\tilde{m}^m$.

Cropped $\mathbf{I}$, cropped PDM, and $\tilde{m}^m$ are concatenated along the channel dimension and provided as input to a ViT image encoder. In parallel, the cropped PDM and the modal mask $\tilde{m}^m$ are further processed, and their features are concatenated to construct the routing signal for the soft MoE module. To more effectively leverage the priors encoded in SAM, cropped $\mathbf{I}$ is also passed through the SAM encoder, and the resulting features are concatenated with the ViT image-encoder features along the channel dimension. The fused features are then fed into the decoder to estimate the complete mask $\tilde{m}^a$, which is finally mapped back to the original image to produce the final amodal mask $\tilde{M}^a$.

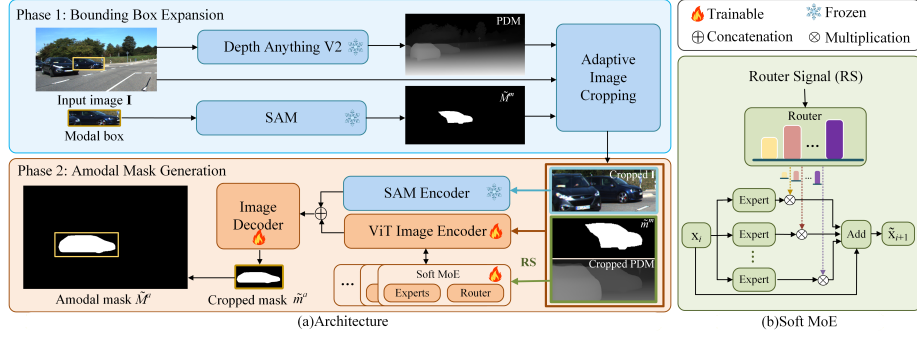


Fig. 2: (a) Overall architecture of HIDA. HIDA employs a plug-and-play architecture that leverages vision foundation models as frozen prior extractors and requires no fine-tuning of the foundation models. Two foundation models are incorporated to obtain rich priors and to guide image cropping (refer to Sec. 3.3). A lightweight ViT-based segmentation model is constructed to integrate the priors provided by the foundation models with the original input, and an occlusion-aware soft MoE is further developed to strengthen the lightweight model’s ability to handle occlusions. (b) Architecture of a single soft MoE block. Soft MoE comprises a router and multiple experts. The router receives an occlusion signal formed using predefined rules and outputs expert weights, whereas all experts take the same feature x_i from the ViT image encoder. A residual design is applied in which expert outputs are weighted by the router’s expert weights and then added to x_i to obtain $\tilde{x}_{i+1}$ (refer to Sec. 3.4). RS denotes the routing signal constructed by combining the cropped PDM and the visible mask $\tilde{m}^m$ through a dedicated fusion operation.

3.3 Adaptive Image Cropping via Visible Masks and Pseudo-Depth

Given the normalized PDM generated by Depth Anything V2, $D \in [0, 255]^{H \times W}$, where larger depth values indicate that the target is closer to the camera, the following intuition is adopted: when pixels outside M^m exhibit larger values than those within the inner border band of M^m , the presence of a foreground occluder on that side is assumed, and a larger expansion is assigned to that side. Further details are provided through the occlusion scoring and expansion allocation strategy described below.

The four sides of M^m are denoted by l, r, t and b . For each side $s \in \{l, r, t, b\}$, an irregular stripe Ω_s is collected outside M^m such that it does not overlap with the visible mask. The stripe width or height is constrained not to exceed the maximum number of expandable pixels permitted by the image boundary on that side. To encourage pixels closer to M^m to contribute more, we apply a linearly decaying weight $w(p) \in (0, 1]$ along the direction perpendicular to that side (where pixels closer to the boundary receive higher weights). The occlusion score for side s is defined as:

$$score_s = \frac{\sum_{p \in \Omega_s} 1[D(p) \geq med(D(band_{in})) + \Delta]w(p)}{\sum_{p \in \Omega_s} w(p)} \quad (1)$$

where band_{in} denotes the inner border band obtained by eroding M^m , with a band width of 5 pixels. $[D(p) \geq \text{med}(D(\text{band}_{\text{in}})) + \Delta]$ indicates the depth polarity, and Δ is the depth margin threshold. If no valid stripe is available on a given side, we set $\text{score}_s = 0$.

For expansion allocation, an expansion ratio $Er \in [Er_{\min}, Er_{\max}]$ is specified while ensuring that the expanded box remains within the image boundaries. The horizontal expansion ratio on the Er_{left} is defined as follows:

$$Er_{\text{left}} = \min \left\{ \max \left(\frac{\text{score}_{\text{left}} \times (Er_{\max} + Er_{\min})}{\text{score}_{\text{left}} + \text{score}_{\text{right}} + \varepsilon}, Er_{\min} \right), Er_{\max} \right\} \quad (2)$$

and similarly for the other sides.

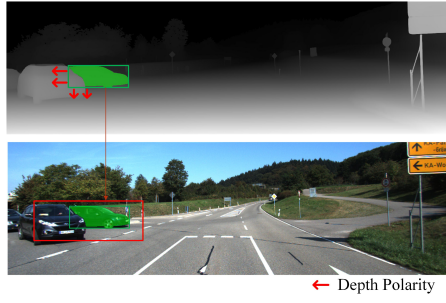


Fig. 3: Visualization of boundary expansion guided by depth polarity.

As illustrated in Fig. 3, the crop is expanded adaptively toward the side that is genuinely more likely to be occluded, based on occlusion cues. This strategy retains local context that is important for completion while largely excluding distant, irrelevant background, thereby increasing the signal-to-noise ratio. The target becomes more centered within the cropped image at a more consistent scale, which reduces scale jitter associated with random enlargement and decreases the excessive blank areas that often occur for elongated objects. Consequently, the encoder’s receptive field is utilized

more effectively on foreground content. In addition, this approach reduces the bias introduced by directly enlarging the visible bounding box, guiding the model to focus more on plausible shape extension into occluded regions, which typically yields more stable prediction performance.

3.4 Occlusion-Aware Token-Routing Soft MoE

Amodal segmentation, unlike modal segmentation, exhibits a substantial representation gap, as texture and depth distributions in visible regions can differ markedly from those in invisible regions. This challenge is addressed through an occlusion-aware token-routing mixture-of-experts module that dynamically assigns expert weights across spatial locations during encoding. Such routing enables discriminative features from visible regions to pass across occlusion boundaries into invisible regions, supporting more stable shape completion.

An occlusion-aware token-routing mixture-of-experts structure is integrated into a lightweight ViT backbone. Occlusion cues are constructed to guide expert selection. Occlusion strength at the visible boundary is characterized by deriving

three low-resolution routing features g , d , m from the depth map and the visible mask, which are then used to form an occlusion-sensitive routing signal as follows:

$$occ = [\sigma(g + d) \| d \| m] \quad (3)$$

where σ denotes the Sigmoid function. The term g represents the depth gradient magnitude, which is computed by applying a Sobel filter to the depth map followed by normalization, which captures the strength of occlusion at depth discontinuities. d denotes the boundary of the visible mask, and m is the down-sampled mask.

This design uses the occlusion-sensitive signal occ to steer tokens toward different experts, producing more discriminative local representations at boundary discontinuities and within occluded regions while preserving computational and representational efficiency in non-occluded areas. Soft routing is applied at each token spatial location (i, j) : first, the backbone feature $x_i \in \mathbb{R}^{B \times C \times H \times W}$ is reduced by a pointwise convolution to $z \in \mathbb{R}^{B \times c \times H \times W}$; after concatenating z with the occlusion signal occ along the channel dimension, the router $R(\cdot)$ produces K expert logits at that location, logits: $r(i, j) = R([z \| occ])$, and a gating vector is obtained along the expert dimension as $gate = \text{softmax}(r(i, j))$. Let the k -th expert be a lightweight convolutional mapping $E_k(\cdot)$; soft routing forms a per-location convex combination of expert outputs using the gating coefficients, yielding the output of occlusion-aware token-routing mixture-of-experts:

$$\mathbf{y}_{i,j} = \sum_{k=1}^K g_{k,i,j} \mathbf{E}_k(\mathbf{x}_{i,j}), \quad \text{s.t. } g_{k,i,j} \geq 0, \sum_{k=1}^K g_{k,i,j} = 1. \quad (4)$$

To prevent the gating from collapsing to a small subset of experts, we introduce a load-balancing regularizer based on batch- and spatially-averaged gating:

$$\mathcal{L}_{LB} = - \sum_{k=1}^K \bar{g}_k \log(\bar{g}_k + \epsilon), \quad \bar{g}_k = \frac{1}{B \times H \times W} \sum_{b=1}^B \sum_{i=1}^H \sum_{j=1}^W g_{b,i,j,k}. \quad (5)$$

Soft routing allocates continuous weights over K experts at each spatial location through a softmax gate conditioned on the estimated occlusion degree occ , allowing features in occluded and boundary regions to adaptively choose more suitable representation pathways. A load-balancing regularization term is also incorporated to regulate expert utilization frequency, which promotes more diverse expert usage and supports stability throughout training.

3.5 Loss Function

The overall loss is composed of two terms: $\mathcal{L}_{seg}$, which measures segmentation quality, and $\mathcal{L}_{LB}$, which constrains expert utilization frequency.

We use a combination of binary cross-entropy loss and Dice loss to reflect the segmentation error of the target:

$$\mathcal{L}_{seg} = \mathcal{L}_{BCE}(\tilde{m}^a, m^a) + \mathcal{L}_{Dice}(\tilde{m}^a, m^a) \quad (6)$$

where $\tilde{m}^a$ and m^a denote the predicted and ground-truth amodal masks of the target in the cropped image, respectively.

A load-balancing regularizer $\mathcal{L}_{LB}$ is further introduced, based on gates averaged across the batch and spatial dimensions, to prevent the gating from collapsing to a small subset of experts. The overall loss is defined as follows:

$$\mathcal{L} = \mathcal{L}_{seg} + \mathcal{L}_{LB} \quad (7)$$

4 Experiments

4.1 Datasets and Metrics

All experiments, including the proposed method and the baseline approaches used for comparison, are performed on the test(or val) splits of three widely-used and challenging amodal segmentation benchmarks: COCOA-cls [50], KINS [35], and D2SA [10]. Unlike the synthetic dataset pix2gestalt [32], both COCOA-cls and KINS are built from real-world natural images with occlusions. COCOA-cls is derived from the COCO [24] dataset by adding amodal annotations to a subset of COCO images, and it contains 2,476 training images and 1,223 val images. The dataset spans 80 object categories. KINS is derived from the KITTI [12] benchmark and focuses on street scenes involving pedestrians and vehicles for autonomous driving applications; it is currently the largest amodal street-scene dataset. It includes seven common on-road categories, such as cars, trucks, and pedestrians, with 7,517 training images and 7,474 test images. Following Bayesian-Amodal [38] and BLADE [27], we restrict evaluation to vehicles whose height is greater than 50 pixels on the test set. The D2SA dataset is built upon the D2S [9] dataset and targets supermarket checkout scenarios, containing 60 categories of diverse objects. D2SA includes 2,000 training images and 3,600 val images. Notably, KINS and D2SA are designed for specific scenarios and thus exhibit strong scene bias.

All methods are assessed using the standard Intersection-over-Union (IoU) metric on these benchmarks.

4.2 Implementation Details

The HIDA framework is implemented in PyTorch. We adopt the ViT-H version of SAM [19] and the Large version of Depth Anything V2 [44] as the foundation models to extract image priors. Considering that the expanded bounding box should cover the full target as much as possible while reducing background redundancy, we set $Er_{\min} = 0.2$, $Er_{\max} = 0.6$. The trainable lightweight segmentation encoder is based on ViT, and we explore three different network depths to study how the encoder capacity affects overall performance.

Training is conducted on the pix2gestalt [32] dataset. This dataset is class-agnostic and contains 849,667 image samples with synthetic occlusions generated based on the SA-1B dataset [19]. We crop images using the expanded bounding boxes; all inputs are resized to 256×256 and then processed, without using

any data augmentation. During training, we jointly update the decoder and the unfrozen lightweight ViT encoder parameters using the AdamW optimizer, with an initial learning rate of 1×10^{-4} , a weight decay of 1×10^{-4} , and 15 epochs. The learning rate is gradually decayed following a cosine annealing schedule [30], with the annealing period set to the total number of training epochs, thereby maintaining larger update steps early to accelerate convergence and smoothly reducing the learning rate later to stabilize optimization [18]. The batch size is set to 32. For zero-shot training, all trainable parameters are randomly initialized, while for fully supervised training we initialize from the zero-shot weights. We precompute and cache all priors offline, and only train the lightweight network. All experiments are conducted on a single NVIDIA RTX 4090 (24 GB).

4.3 Comparison with Existing Methods

To provide a more comprehensive evaluation, HIDA is compared with seven existing approaches on the COCOA-cls and D2SA validation sets and on the KINS test set. GIN [23], C2F-Seg [11], and SDAmodal [45] are leading fully supervised amodal segmentation methods. SAM [19], SDXL-Inpainting [33], SAMBA [29], and Pix2Gestalt [32] are zero-shot methods. SAM is a general-purpose, promptable foundation model for visible-mask segmentation, while SAMBA further extends SAM to enable zero-shot amodal segmentation. SDXL-Inpainting serves as a general inpainting baseline for appearance completion, and pix2gestalt is a diffusion-based framework specifically designed for zero-shot amodal segmentation and completion.

The framework uses only the modal bounding box as input. For a fair comparison, SAM-predicted modal masks are provided when a method requires a modal mask.

As shown in Table 1, our method achieves the best supervised performance across all three datasets. Under the zero-shot setting, Ours[†] obtains the best results on COCOA-cls and D2SA, outperforming SAMBA, while performing slightly worse than SAMBA on KINS. Interestingly, under the fully supervised setting, the performance gain of our method is more pronounced on scene-biased datasets (D2SA, KINS), highlighting its practical value for scenario-targeted applications.

Fig. 4 further presents qualitative results that visually demonstrate our advantage: compared with other models, HIDA produces more accurate object shapes.

4.4 Ablation Study

The effectiveness of the proposed approach is examined through ablation studies conducted in the zero-shot regime. Training is performed on the pix2gestalt dataset, and evaluation is carried out on the COCOA-cls validation set. Ablations are conducted by varying the encoder data-fusion strategy, the capacity of the ViT encoder in terms of the number of trainable parameters, and the

Table 1: Quantitative results on the three datasets compared with existing methods. † denotes zero-shot methods. Performance is reported as mean IoU (%), where higher values indicate better performance. The best results under the zero-shot setting are shown in bold, and the best results among supervised methods are indicated with underlining.

Method	Venue	COCOA-cls	D2SA	KINS
GIN [23]	TMM 2023	72.46	78.23	68.31
SAM [19]	ICCV 2023	73.10	84.65	75.88
C2F-Seg [11]	ICCV 2023	80.28	89.10	82.22
SDXL-Inpainting† [33]	ICLR 2024	73.65	80.53	76.19
Pix2Gestalt† [32]	CVPR 2024	79.08	81.82	81.45
SDAmodal [45]	CVPR 2024	80.01	-	-
SAMBA† [29]	CVPR 2025	81.82	90.98	88.47
HIDA	-	<u>85.56</u>	<u>93.33</u>	<u>91.50</u>
HIDA†	-	83.16	91.11	86.71

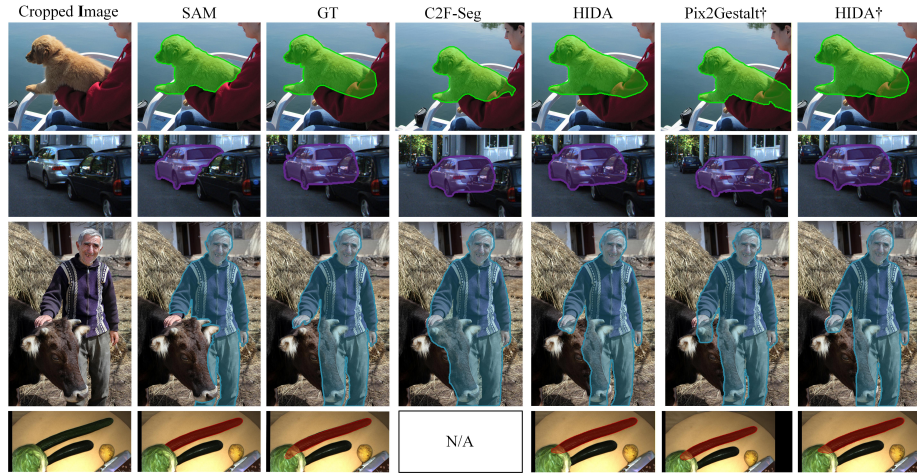


Fig. 4: Qualitative results. Qualitative comparison of predicted amodal masks generated by SAM, Pix2Gestalt, C2F-Seg, and our HIDA under both zero-shot and fully supervised settings. For fair comparison, Pix2Gestalt and C2F-Seg use a visible bounding box enlarged by a factor of two as input to ensure a fair comparison. The official implementation of C2F-Seg does not provide pretrained weights for the D2SA dataset. A zoomed view is recommended for improved clarity.

number of experts, while the per-epoch training time is additionally reported as an efficiency metric.

Table 2 reports zero-shot performance on the COCOA-cls validation set and supports several observations. Decoder feature selection has a substantial impact on performance. Using only SAM encoder features results in relatively low accuracy, whereas incorporating the trainable ViT representation produces a large improvement, indicating that learned global semantics and shape priors are important for extrapolating and completing occluded regions. Further gains, along with more stable performance, are obtained by fusing SAM and ViT features, although this fusion increases per-epoch training time. The contribution of pseudo-depth is assessed by an ablation in which the PDM is set to zero. Under this setting, IoU on COCOA-cls decreases to 81.78%, which confirms that pseudo-depth cues contribute to accurate amodal completion.

Table 2: Ablation study on the COCOA-cls dataset evaluating ViT decoder feature sources with optional depth cues.

Index	SAM Enc.	ViT Enc.	Pseudo-depth	IoU (%)	Training Time / Epoch (h)
1	✓	×	—	60.90	3.5
2	×	✓	✓	81.56	5.3
3	✓	✓	×	81.78	6.1
4	✓	✓	✓	83.16	6.1

Tables 3–4 indicate that enlarging the ViT encoder, reflected by deeper layers, yields consistent performance improvements. Higher encoder capacity, however, is accompanied by increased parameter counts, greater computational cost, and longer per-epoch training time, which reflects the typical trade-off between accuracy and training efficiency. In this implementation, the ViT decoder is fixed at 12.63M parameters, with a computational cost of 17.42G FLOPs. Within the soft MoE design, a moderate increase in the number of experts enhances the model’s ability to capture diverse occlusion patterns and improves performance. Performance saturation or slight degradation occurs when the number of experts becomes too large, which suggests that excessive expert capacity can make routing more difficult to learn and may reduce generalization. The proposed method includes 77.86M trainable parameters, which is slightly higher than SAMBA (72M) and substantially lower than Pix2Gestalt (860M). Training time also does not increase linearly with computational cost, since input-data diversity introduces additional overhead from data loading and preprocessing within the training pipeline.

Table 3: Ablation Study of the Number of Experts in Soft MoE on the COCOA-cls Dataset.

No. Exp.	Enc. Params(M)	Enc. FLOPs(G)	IoU (%)
1	60.29	17.68	81.98
2	61.93	18.10	82.69
4	65.23	18.95	83.16
8	71.87	20.65	82.97

Table 4: Ablation study of ViT encoder depth on the COCOA-cls dataset.

Enc. Layers	IoU	Train Time Epoch (h)	Enc. Param (M)	Enc. FLOPs (G)
4	82.18	4.5	6.43	3.89
6	83.05	5.1	24.81	8.58
8	83.16	6.1	65.23	18.95

4.5 Analysis of Expert Activation Patterns in Soft MoE During Forward Inference

Fig. 5 presents visualizations of the soft MoE routing signals for the same input at Layer 0 and Layer 7. For each layer, routing-weight heatmaps for four experts are shown over a 16×16 patch grid, together with the dominant expert assignment map and an entropy-based routing uncertainty map. Layer 0 shows finer-grained and more localized expert specialization: experts respond differently to interior and boundary regions of the target, while background patches are often consistently dominated by a single expert. With increased depth to Layer 7, routing patterns become smoother and more region-consistent, producing more coherent expert assignments over the target and indicating stronger aggregation of global structure and semantic cues in deeper representations. Higher entropy

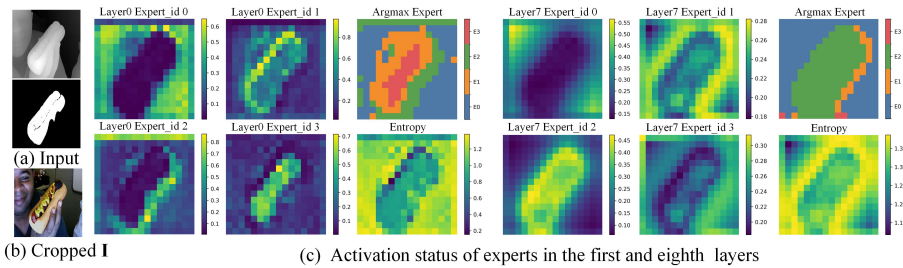


Fig. 5: During inference, we visualize Soft-MoE expert activations. The first column shows the inputs (pseudo-depth and visible mask) with the corresponding RGB image. Columns 2–4 report, for Layer 0, the routing-weight heatmaps of four experts, along with the argmax assignment and entropy maps; Columns 5–7 show the same visualizations for Layer 7. The color legend denotes routing-weight magnitude or expert index.

appears near object boundaries and occlusion-related regions, which reflects increased mixing across experts and greater uncertainty, whereas lower entropy is observed in stable interior regions, indicating more confident routing decisions.

4.6 Limitation



Fig. 6: Typical failure cases.

Fig. 6 illustrates that depth aliasing can arise near contact regions when the elongated boundary of a distant target closely touches a nearer foreground object. Weak or inconsistent depth discontinuities may cause the PDM to assign similar or even inverted depth ordering to boundary pixels, making boundary ownership ambiguous. As a result, HIDA may over-extend into the occluder or insufficiently complete the occluded region. This issue is expected to occur less often under fully supervised training,

where dense amodal supervision helps learn stronger boundary-ownership priors. Although uncommon, such cases reveal limitations of pseudo-depth under severe contact occlusions and motivate future work on more reliable depth cues or explicit boundary reasoning.

HIDA is limited by its inference cost, requiring 0.8 s per frame per instance on a single RTX 4090. Lightweight segmentation and depth backbones will be explored for acceleration.

5 Conclusion

This work introduces HIDA, a modular amodal segmentation framework that integrates two frozen vision foundation models with a lightweight trainable segmentation network. High-quality visible masks and pseudo-depth cues are jointly utilized to perform adaptive, target-centric cropping that suppresses irrelevant context and directs learning toward occlusion boundaries and invisible regions. An occlusion-aware token-routing soft MoE module is also incorporated to more effectively accommodate diverse occlusion patterns. Evaluations on COCOA-cls, KINS, and D2SA show that HIDA achieves state-of-the-art performance in fully supervised settings and strong zero-shot results on COCOA-cls and D2SA, while training only a lightweight segmentation network. Future work will focus on mitigating the depth ambiguity issues described above by investigating improved depth estimation or explicit modeling of occlusion boundaries.

6 Acknowledgements.

This work was supported by Hainan Provincial Natural Science Foundation of China (No. 326QN0702).

References

1. Ao, J., Ke, Q., Ehinger, K.A.: Image amodal completion: A survey. *Computer Vision and Image Understanding* **229**, 103661 (2023). <https://doi.org/10.1016/j.cviu.2023.103661>
2. Awais, M., Naseer, M., Khan, S., et al.: Foundation models defining a new era in vision: a survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **47**(4), 2245–2264 (2025). <https://doi.org/10.1109/TPAMI.2024.3506283>
3. Breitenstein, J., Fingscheidt, T.: Amodal cityscapes: A new dataset, its generation, and an amodal semantic segmentation challenge baseline. In: 2022 IEEE Intelligent Vehicles Symposium (IV). pp. 1018–1025 (2022). <https://doi.org/10.1109/IV51971.2022.9827342>
4. Cao, Y., Zhang, J., Shi, H., Peng, K., Zhang, Y., Zhang, H., Stiefelhagen, R., Yang, K.: Occlusion-aware seamless segmentation. In: Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., Varol, G. (eds.) *Computer Vision – ECCV 2024*. pp. 129–147. Springer Nature Switzerland, Cham (2025). https://doi.org/10.1007/978-3-031-72655-2_8
5. Caron, M., Touvron, H., Misra, I., Jegou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: *Proceedings of the IEEE/CVF international conference on computer vision*. pp. 9650–9660 (2021). <https://doi.org/10.1109/ICCV48922.2021.00951>
6. Chen, J., Niu, L., Zhang, J., Si, J., Qian, C., Zhang, L.: Amodal instance segmentation via prior-guided expansion. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. vol. 37, pp. 313–321 (2023). <https://doi.org/10.1609/aaai.v37i1.25104>
7. Doh, H., Lee, D.I., Chi, S., Huang, P.H., Lee, K., Kim, S., Ramani, K.: Occlusion-aware temporally consistent amodal completion for 3d human-object interaction reconstruction. In: *Proceedings of the 33rd ACM International Conference on Multimedia*. p. 52–61. MM '25, Association for Computing Machinery, New York, NY, USA (2025). <https://doi.org/10.1145/3746027.3754769>
8. Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: *International Conference on Learning Representations* (2021)
9. Follmann, P., Böttger, T., Härtinger, P., König, R., Ulrich, M.: Mytec d2s: densely segmented supermarket dataset. In: *Proceedings of the European conference on computer vision (ECCV)*. pp. 581–597 (2018). https://doi.org/10.1007/978-3-030-01249-6_35
10. Follmann, P., König, R., Härtinger, P., Klostermann, M., Böttger, T.: End-to-end trainable amodal instance segmentation. In: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 1328–1336. IEEE (2019). <https://doi.org/10.1109/WACV.2019.00146>
11. Gao, J., Qian, X., Wang, Y., et al.: Coarse-to-fine amodal segmentation with shape prior. In: *Proceedings of the IEEE/CVF international conference on computer vision*. pp. 1262–1271 (2023). <https://doi.org/10.1109/ICCV51070.2023.00122>

12. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: 2012 IEEE conference on computer vision and pattern recognition. pp. 3354–3361. IEEE (2012). <https://doi.org/10.1109/CVPR.2012.6248074>
13. Gkitsas, V., Sterzentsenko, V., Zioulis, N., Albanis, G., Zarpalas, D.: Panodr: Spherical panorama diminished reality for indoor scenes. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3711–3721 (2021). <https://doi.org/10.1109/CVPRW53098.2021.00412>
14. Inagaki, Y., Araki, R., Yamashita, T., Fujiyoshi, H.: Detecting layered structures of partially occluded objects for bin picking. In: 2019 IEEE/RSJ international conference on intelligent robots and systems (IROS). pp. 5786–5791. IEEE (2019). <https://doi.org/10.1109/IROS40897.2019.8968093>
15. Jin, D., Kim, C.S.: Omr: Occlusion-aware memory-based refinement for video lane detection. In: Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., Varol, G. (eds.) Computer Vision – ECCV 2024. pp. 129–145. Springer Nature Switzerland, Cham (2025). https://doi.org/10.1007/978-3-031-73414-4_8
16. Kanizsa, G., Legrenzi, P., Bozzi, P.: Organization in vision: Essays on Gestalt perception (1979)
17. Kellman, P.J., Shipley, T.F.: A theory of visual interpolation in object perception. *Cognitive psychology* **23**(2), 141–221 (1991). [https://doi.org/10.1016/0010-0285\(91\)90009-D](https://doi.org/10.1016/0010-0285(91)90009-D)
18. Khaled, A., Jin, C.: Tuning-free stochastic optimization. *Proceedings of Machine Learning Research* **235**, 23622–23661 (2024)
19. Kirillov, A., Mintun, E., Ravi, N., et al.: Segment anything. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4015–4026 (2023). <https://doi.org/10.48550/arXiv.2304.02643>
20. Kortylewski, A., Liu, Q., Wang, A., Sun, Y., Yuille, A.: Compositional convolutional neural networks: A robust and interpretable model for object recognition under occlusion. *International Journal of Computer Vision* **129**(3), 736–760 (2021). <https://doi.org/10.1007/s11263-020-01401-3>
21. Li, K., Malik, J.: Amodal instance segmentation. In: European Conference on Computer Vision. pp. 677–693. Springer (2016). <https://doi.org/10.48550/arXiv.1604.08202>
22. Li, Z., Ye, W., Jiang, T., Huang, T.: 2d amodal instance segmentation guided by 3d shape prior. In: European Conference on Computer Vision. pp. 165–181. Springer (2022). https://doi.org/10.1007/978-3-031-19818-2_10
23. Li, Z., Ye, W., Jiang, T., Huang, T.: Gin: Generative invariant shape prior for amodal instance segmentation. *IEEE Transactions on Multimedia* **26**, 3924–3936 (2023). <https://doi.org/10.1109/TMM.2023.3318075>
24. Lin, T.Y., Maire, M., Belongie, S., et al.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014). https://doi.org/10.1007/978-3-319-10602-1_48
25. Liu, T., Blondel, M., Ruiz, C.R., Puigcerver, J.: Routers in vision mixture of experts: An empirical study. In: Transactions on Machine Learning Research (2024)
26. Liu, Z., Liu, Q., Chang, C., et al.: Object-level scene deocclusion. In: ACM SIGGRAPH 2024 Conference Papers. pp. 1–11 (2024)
27. Liu, Z., Li, Z., Jiang, T.: Blade: Box-level supervised amodal segmentation through directed expansion. In: Proceedings of the AAAI conference on artificial intelligence. vol. 38, pp. 3846–3854 (2024). <https://doi.org/10.1609/aaai.v38i4.28176>

28. Liu, Z., Qiao, L., Chu, X., Jiang, T.: Plug: Revisiting amodal segmentation with foundation model and hierarchical focus. arXiv preprint arXiv:2405.16094 (2024). <https://doi.org/10.1145/3641519.3657409>
29. Liu, Z., Qiao, L., Chu, X., Ma, L., Jiang, T.: Towards efficient foundation model for zero-shot amodal segmentation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 20254–20264 (2025). <https://doi.org/10.1109/CVPR52734.2025.01886>
30. Loshchilov, I., Hutter, F.: Sgdr: Stochastic gradient descent with warm restarts. International Conference on Learning Representations (2017)
31. Nanay, B.: The importance of amodal completion in everyday perception. *i-Perception* **9**(4), 2041669518788887 (2018). <https://doi.org/10.1177/2041669518788887>
32. Ozguroglu, E., Liu, R., Suris, D., et al.: pix2gestalt: Amodal segmentation by synthesizing wholes. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3931–3940. IEEE Computer Society (2024). <https://doi.org/10.48550/arXiv.2401.14398>
33. Podell, D., English, Z., Lacey, K., et al.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. The Twelfth International Conference on Learning Representations (2024)
34. Puigcerver, J., Ruiz, C.R., Mustafa, B., Houlsby, N.: From sparse to soft mixtures of experts. In: The Twelfth International Conference on Learning Representations (2024)
35. Qi, L., Jiang, L., Liu, S., Shen, X., Jia, J.: Amodal instance segmentation with kins dataset. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3009–3018 (2019). <https://doi.org/10.1109/CVPR.2019.00313>
36. Shazeer, N., Mirhoseini, A., Maziarz, K., et al.: Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In: International Conference on Learning Representations (2017)
37. Sueyoshi, T., Morimoto, Y.: Dynamic projection mapping for thin plants using a robust tracking method against occlusion. In: ACM SIGGRAPH 2021 Labs. pp. 1–2 (2021). <https://doi.org/10.1145/3450616.3464521>
38. Sun, Y., Kortylewski, A., Yuille, A.: Amodal segmentation through out-of-task and out-of-distribution generalization with a bayesian model. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1215–1224 (2022). <https://doi.org/10.1109/CVPR52688.2022.00128>
39. Tai, W.E., Shih, Y.L., Sun, C., Wang, Y.C.F., Chen, H.H.: Segment anything, even occluded. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 29385–29394 (2025). <https://doi.org/10.1109/CVPR52734.2025.02736>
40. Tang, L., Ruiz, N., Chu, Q., Li, Y., Holynski, A., Jacobs, D.E., Hariharan, B., Pritch, Y., Wadhwa, N., Aberman, K., Rubinstein, M.: Realfill: Reference-driven generation for authentic image completion. vol. 43. Association for Computing Machinery, New York, NY, USA (Jul 2024). <https://doi.org/10.1145/3658237>
41. Tran, M.Q., Vo, K., Yamazaki, K., Fernandes, A., Kidd, M., Le, N.: Aisformer: Amodal instance segmentation with transformer. In: Proceedings of the British Machine Vision Conference (BMVC) (2022)
42. Wada, K., Okada, K., Inaba, M.: Joint learning of instance and semantic segmentation for robotic pick-and-place with heavy occlusions in clutter. In: 2019 international conference on robotics and automation (ICRA). pp. 9558–9564. IEEE (2019). <https://doi.org/10.1109/ICRA.2019.8793783>

43. Xiao, Y., Xu, Y., Zhong, Z., Luo, W., Li, J., Gao, S.: Amodal segmentation based on visible region segmentation and shape prior. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 2995–3003 (2021). <https://doi.org/10.1609/aaai.v35i4.16407>
44. Yang, L., Kang, B., Huang, Z., et al.: Depth anything v2. Advances in Neural Information Processing Systems **37**, 21875–21911 (2024). <https://doi.org/10.48550/arXiv.2406.09414>
45. Zhan, G., Zheng, C., Xie, W., Zisserman, A.: Amodal ground truth and completion in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 28003–28013 (2024). <https://doi.org/10.1109/CVPR52733.2024.02645>
46. Zhang, S., Yang, H., Zhang, J., Wang, X.: Dbnet: A depth-guided and boundary-aware network for amodal instance segmentation. Image and Vision Computing p. 105688 (2025)
47. Zhang, Z., Wang, H., Yu, Z., Cheng, Y., Yao, A., Chang, H.J.: Nl2contact: Natural language guided 3d hand-object contact modeling with diffusion model. In: Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part XXVIII. p. 284–300. Springer-Verlag, Berlin, Heidelberg (2024). https://doi.org/10.1007/978-3-031-73390-1_17
48. Zhang, Z., Chen, A., Xie, L., Yu, J., Gao, S.: Learning semantics-aware distance map with semantics layering network for amodal instance segmentation. In: Proceedings of the 27th ACM international conference on multimedia. pp. 2124–2132 (2019). <https://doi.org/10.1145/3343031.3350911>
49. Zhixuan Li, Ruohua Shi, T.H., Jiang, T.: Oaformer: Learning occlusion distinguishable feature for amodal instance segmentation. In: ICASSP 2023-2023 IEEE international conference on acoustics, speech and signal processing (ICASSP). pp. 1–5. IEEE (2023). <https://doi.org/10.1109/ICASSP49357.2023.10096534>
50. Zhu, Y., Tian, Y., Metaxas, D., Dollár, P.: Semantic amodal segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1464–1472 (2017). <https://doi.org/10.1109/CVPR.2017.320>