

# Single Point, Full Mask: Velocity-Guided Level Set Evolution for End-to-End Amodal Segmentation

Zhixuan Li

College of Computing and Data  
Science

Nanyang Technological University  
Singapore, Singapore

Yujia Liu

School of Computer Science  
National Key Laboratory for  
Multimedia Information Processing  
National Engineering Research  
Center of Visual Technology  
Peking University  
Beijing, China

Chen Hui

Nanjing University of Information  
Science and Technology  
Nanjing, China

Chenyue Song

Harbin Institute of Technology  
Harbin, China

Weisi Lin\*

College of Computing and Data  
Science  
Nanyang Technological University  
Singapore, Singapore

## Abstract

Amodal segmentation aims to recover complete object shapes, including occluded regions, serving as an essential technique for user-centric multimedia authoring and object-level visual manipulation. Existing methods typically rely on informative prompts, such as bounding boxes or dense visible masks, which heavily degrade the user experience and interaction efficiency in real-world multimedia applications. While recent interactive paradigms (e.g., the Segment Anything Model) support lightweight point-based interactions, they often perform direct mask regression. Crucially, the opaque nature of these direct-regression models offers no visual explainability regarding how occluded structures are inferred, conflicting with the growing demand for interpretable multimedia systems. To address these limitations, we propose VELA, an end-to-end VELOCITY-driven Level-set Amodal segmentation method that enables explicit and transparent contour evolution driven by simple point clicks. VELA constructs an initial level set function from visual features and the user’s point input, which then progressively evolves into the final amodal mask under the guidance of a shape-specific motion field predicted by a fully differentiable network. This mechanism learns to generate evolution dynamics at each step, ensuring that the spatial reasoning process is geometrically grounded, topologically flexible, and visually explainable to the user. Extensive experiments on COCOA-cls, D2SA, and KINS benchmarks demonstrate that VELA outperforms existing methods that use bounding-box or dense visible-mask prompts while requiring only a single-point prompt, validating the effectiveness of explainable geometric modeling for interactive multimedia tasks.

## Keywords

Amodal Segmentation, Occlusion, Level Set

\*Corresponding author.



This work is licensed under a Creative Commons Attribution 4.0 International License.

## 1 Introduction

The conventional visible segmentation task focuses solely on observable areas and is typically formulated as a pixel-wise classification problem, without inferring the shapes of occluded regions. In contrast, the amodal segmentation task aims to predict complete object masks, including both visible and occluded parts, by reasoning about the full object shape from the available visual cues. Despite its inherent challenges, amodal segmentation is rapidly becoming a cornerstone for human-centric multimedia applications, particularly in interactive image editing [49, 59], object-level visual compositing [56, 58], and the structural control of AI-generated content (AIGC) [48, 64]. In everyday multimedia creation, target subjects are frequently obscured by complex foreground elements, scene clutter, or dynamic human activities. Empowering users to easily recover the complete spatial extent of these occluded objects provides the essential geometric foundation required for tasks like seamless background replacement, object removal, and high-fidelity visual composition. Furthermore, facilitating this recovery through intuitive user interactions while ensuring the reasoning process remains visually transparent is a critical requirement for next-generation multimedia editing workflows. More broadly, robust and efficient multimedia perception has been studied under input perturbations, open-vocabulary recognition, domain shifts, and mobile deployment [18, 24, 44, 70, 75].

Existing state-of-the-art amodal segmentation methods [43, 46, 51, 69, 72] mostly rely on informative input prompts, such as dense visible masks [5, 43, 51, 68, 69, 72] or bounding boxes [45, 46], to guide the prediction of amodal masks covering occluded regions. As shown in Fig.1 (a), while these methods have achieved notable performance, obtaining such prompts is often impractical or expensive in real-world scenarios, significantly limiting the applicability of these approaches. Related efforts to reduce data and annotation demands include few-shot multi-view learning, prototype refinement, and monocular 3D perception based on masked autoencoding or weak supervision [19, 20, 73, 74]. Recent models like the Segment Anything Model (SAM) [27, 29, 67] have introduced support for

point-based prompts, offering a more lightweight form of guidance. However, as shown in Fig.1 (b), these approaches typically predict masks directly from these sparse point-based prompts without an explicit mechanism to reason about the shape of occluded regions. As a result, their generalization ability remains limited in complex occlusion scenarios, where point prompts alone provide insufficient guidance for accurately inferring the complete amodal mask.

Another key limitation of existing approaches [3, 37, 40, 60, 68] lies in their black-box nature, which lacks geometric interpretability in the amodal mask prediction process. Specifically, most existing methods adopt CNN-based [4, 10, 38, 39, 55, 66, 69, 71] or Transformer-based [37, 41, 62, 63] architectures that treat amodal mask prediction as a direct regression task, without modeling the underlying morphological processes involved in shape segmentation. Consequently, these models offer little transparency into the internal reasoning behind their predictions, providing limited understanding of how occluded regions are inferred. This absence of explicit geometric reasoning not only limits the interpretability of the model’s predictions but also makes it difficult to analyze or improve model performance, especially in scenarios involving complex occlusions where understanding how the model reasons about fine-grained shapes is crucial for reliable prediction. Motivated by these limitations, we pose the following question: *is it possible to achieve high-quality amodal segmentation using only a single-point prompt by fully leveraging a geometrically interpretable approach?*

To this end, we propose a novel approach named **VELA**, a **VE**LOCITY-driven **Level-set A**modal segmentation method, performing amodal segmentation via Level Set evolution with only a *single point* as input prompt, as shown in Fig.1 (c). Specifically, we begin by embedding the point prompt, which then interacts with the extracted image features to generate an initial level set function, representing the starting state of the object’s amodal mask. This initial function is then progressively updated through a guided evolution process to expand toward the final high-quality amodal mask. To enable precise and adaptive guidance during this evolution, we design a fully differentiable network that learns to predict a shape-specific velocity field, which drives the contour evolution at each step. This velocity prediction network dynamically generates the up-to-date velocity field based on the current state of the evolving contour, providing context-aware guidance that continuously steers the contour toward the complete amodal shape effectively.

This evolution process simulates the gradual expansion of the object’s boundary with the step-by-step guidance generated based on contour status and learned prior knowledge to evolve toward the target shape. In contrast to conventional approaches that directly predict the final mask in a single step, our method models segmentation as a geometrically grounded and iterative process. This step-wise evolution not only enables more interpretable shape reasoning but also provides greater flexibility in handling complex occlusions and diverse object geometries. This is thanks to the Level Set paradigm’s inherent ability to explicitly represent and continuously evolve object contours in a geometrically interpretable and topologically flexible manner. Building upon the unique characteristics of our Level Set-based evolution, VELA introduces a novel paradigm for amodal segmentation with single-point prompts. Moreover, the end-to-end framework allows adaptive modeling of diverse object

shapes and occlusion patterns, alleviating limitations of prior black-box approaches with better transparency.

We extensively evaluate VELA on amodal segmentation benchmarks, including COCOA-cls, D2SA, and KINS. Results show that our Level Set-based approach with a single-point prompt outperforms methods using bounding-box or dense visible-mask prompts, demonstrating the effectiveness of geometry-aware modeling.

Our contributions are summarized as follows: (1) We propose VELA, which, to the best of our knowledge, is the first amodal segmentation method that leverages Level Set evolution with single-point prompting. VELA achieves high-quality amodal segmentation with only a *single point* as input, reducing reliance on bounding-box and dense visible-mask prompts. (2) We introduce a fully differentiable framework that learns a shape-specific velocity field to explicitly guide the contour evolution process, enabling geometrically grounded and interpretable mask prediction. (3) We demonstrate that our method performs well on amodal segmentation datasets using only a single-point prompt, showing its generalization capability in complex occlusion scenarios.

## 2 Related Work

### 2.1 Amodal Segmentation

Since the amodal segmentation task proposed by [31], various approaches have been developed from different perspectives. Several approaches [10, 37, 55, 63, 66, 71] are designed based on existing CNN-based or Transformer-based instance segmentation frameworks to tackle the amodal segmentation problem using only the input image. Meanwhile, many approaches [5, 43, 45, 51, 69, 72] reformulate the task as predicting the amodal mask from the image and additional prompts, such as visible or amodal bounding boxes, or visible masks. Recent work further models shape diversity explicitly through a shape-aware sparse mixture-of-experts framework [36]. However, bounding-box and dense visible-mask prompts for target objects are often impractical or costly to obtain in real-world scenarios. Such high demands on user input reduce the practicality of these methods and limit their ability to generalize to diverse scenarios, particularly in interactive applications. Beyond 2D amodal masks, recent studies extend scene understanding to multimodal 3D segmentation, intention-aware planning, view-robust action generation, camera-robust localization, and geometry-aware vision-language modeling [21–23, 32, 33].

To this end, in this paper, we investigate single-point prompts as guidance for amodal mask segmentation, thereby alleviating the dependence on bounding-box and dense visible-mask prompts. A natural solution is to leverage recent popular models, such as the Segment Anything Model (SAM), which are designed to accept point-based prompts as input. However, SAM focuses on visible mask segmentation with a direct prediction strategy and lacks explicit reasoning for occluded regions, which limits its ability to infer complete amodal masks, especially when only sparse point prompts are available that offer only coarse guidance of the visible parts of target objects. This motivates the need to explore how high-quality amodal segmentation can be achieved using only single-point prompts effectively.

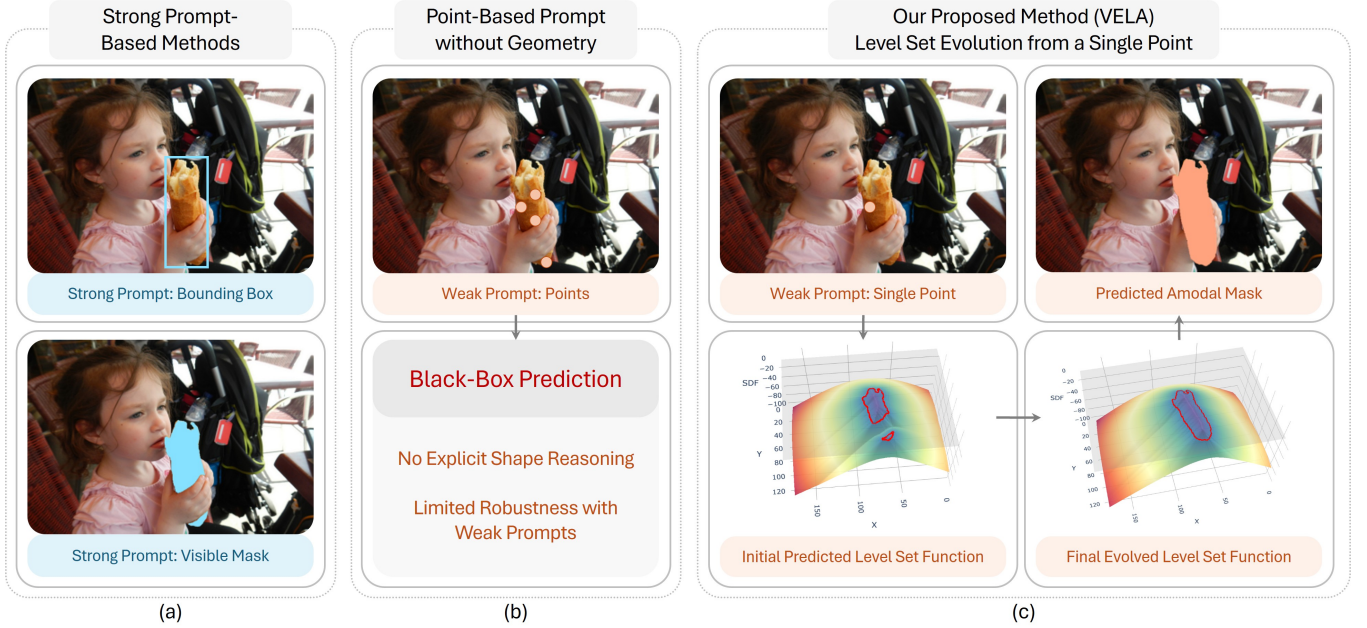


**Figure 1: Comparison of amodal segmentation paradigms. (a) Conventional methods rely on bounding-box or dense visible-mask prompts. (b) Recent approaches support point prompts but directly regress masks without explicit shape reasoning, resulting in limited robustness under complex occlusions. (c) VELA introduces explicit shape evolution via Level Set, effectively exploiting a single-point prompt for interpretable amodal segmentation.**

## 2.2 Level Set Methods for Segmentation

Level set methods offer a powerful framework for image segmentation by implicitly encoding object boundaries as zero-level contours of higher-dimensional functions. Their capacity to represent arbitrary shapes and topological changes, such as merging and splitting, makes them particularly appealing for modeling complex object structures. Traditional approaches [2, 25, 47] typically focus on minimizing hand-crafted energy functions that capture low-level appearance cues, suffering from sensitivity to initialization and limited generalization ability.

To address these limitations, recent methods [8, 28, 35] have integrated the level set with deep learning frameworks. Some methods [50, 65] improve initialization by using networks to predict coarse masks, edge probability maps, or contour priors that serve as starting level-set functions. Several methods utilize the level set from different aspects, including for post-processing for semi-supervised segmentation [61], integrating into the loss function for shape regulation [9], and denoising edge annotations [1].

Different from prior deep level set methods designed for visible-object segmentation, VELA targets amodal completion under a single-point prompt. It predicts a state-dependent velocity field at each iteration rather than relying on a static energy or contour map.

## 3 Methodology

We first define the amodal segmentation task and outline the VELA architecture. We then describe level set initialization, velocity-guided shape evolution, shape-specific velocity generation, and the end-to-end training objective.

### 3.1 Overall Architecture

Given an input image  $I$  and an optional prompt  $P$ , the goal of amodal segmentation is to predict the complete amodal mask  $\hat{M}_a$  of the target object. The prompt  $P$  can take various forms, such as bounding boxes, masks, or points. In this work, we focus on point-based prompts, which offer coarse but efficient guidance for localizing the object to be segmented.

The proposed VELA consists of the following modules, as illustrated in Fig. 2. First, we introduce feature extraction from inputs. Specifically, the backbone network processes the input image  $I$  to extract the global feature representation  $F$  of the entire scene. Simultaneously, the prompt encoder takes the point prompt  $P$  as input and produces a sparse embedding  $e_p$ , which encodes coarse localization information about the target object to be segmented. A *single point* prompt for the target object can provide sufficient guidance for effective amodal segmentation, while VELA also supports multiple point prompts if available.

Next, we introduce the evolution process guided by shape-specific velocity fields. A mask decoder first predicts the initial level set function  $\phi_0$ , which is then iteratively evolved over  $T$  steps to produce the final level set function  $\phi_T$ . At each step  $i \in \{0, \dots, T-1\}$ ,  $\phi_{i+1}$  is obtained by evolving  $\phi_i$  under a dynamically generated,

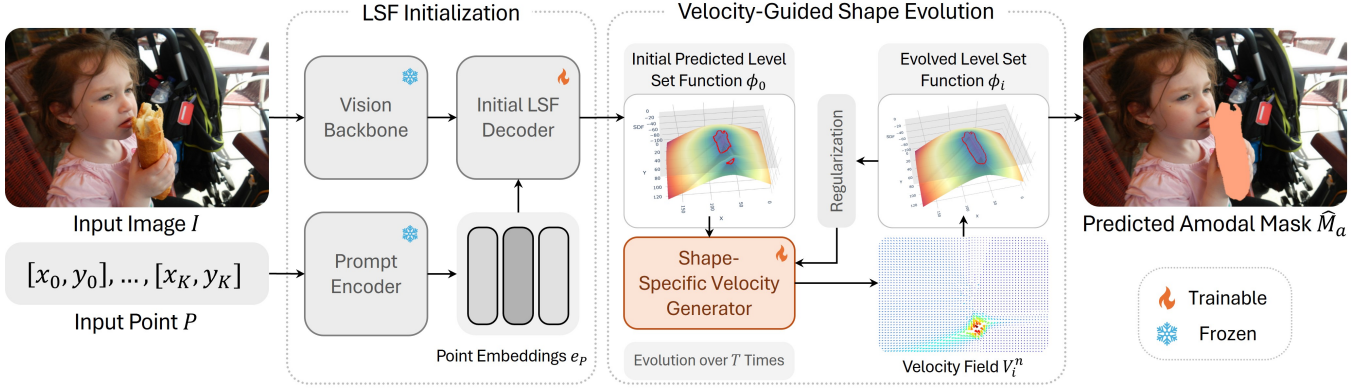


**Figure 2: Overall architecture of the proposed VELA. (a) Level Set Function (LSF) Initialization:** Given an input image  $I$ , the Vision Backbone extracts global image features, while the Prompt Encoder processes the point-level prompt  $P$  to generate point embeddings  $e_p$  that encode coarse localization of the target object. The Initial LSF Decoder then predicts the initial level set function  $\phi_0$  using both the image features and point embeddings. **(b) Velocity-Guided Shape Evolution:** Starting from  $\phi_0$ , the shape evolves over  $T$  steps. At each step  $i$ , the current level set function  $\phi_i$  is fed into the Shape-Specific Velocity Generator to dynamically produce a velocity field  $V_i^n$ , which guides the contour’s deformation towards the final shape. The updated  $\phi_{i+1}$  is then regularized and used for the next step. After  $T$  steps, the final level set function  $\phi_T$  is thresholded to obtain the predicted amodal mask  $\hat{M}_a$ . Best viewed in color.

---

**Algorithm 1** Velocity-Guided Level Set Evolution

---

**Input:** Initial level set function  $\phi_0$ , number of steps  $T$ , time step  $dt$

**Components:** Velocity predictor  $\mathcal{F}_v(\cdot)$ , distance regularizer  $\mathcal{R}(\cdot)$

**Output:** Final evolved level set function  $\phi_T$

- 1: Initialize  $\phi \leftarrow \phi_0$
  - 2: **for**  $i = 0$  to  $T - 1$  **do**
  - 3:   Predict normal velocity  $V_i^n \leftarrow \mathcal{F}_v(\phi)$
  - 4:   Update  $\phi \leftarrow \phi - dt \cdot V_i^n$
  - 5:   Apply regularization:  $\phi \leftarrow \phi + \mu \cdot \mathcal{R}(\phi)$
  - 6: **end for**
  - 7: **return**  $\phi_T \leftarrow \phi$
- 

shape-specific velocity field. This adaptive guidance progressively deforms the zero-level contour toward the target amodal boundary. After  $T$  evolution steps, the final level set function  $\phi_T$  is thresholded at zero to produce the predicted binary amodal mask.

### 3.2 Level Set Function Initialization

The level set function is initialized by predicting  $\phi_0$  using a mask decoder, which takes the global image feature  $F$  and the sparse prompt embedding  $e_p$  as inputs. For a 2D binary mask  $M$  that contains a 2D curve  $C$  of the target object, the level set function  $\phi$  provides a higher-dimensional implicit representation of  $M$ . Specifically,  $\phi$  is a 3D surface whose zero level set corresponds to the object boundary curve  $C$  in the 2D mask  $M$ . Here, we define background regions as negative and foreground regions as positive. Accordingly, the binary mask  $M$  can be obtained by thresholding the level set function  $\phi$  at zero, retaining values greater than zero as the foreground. We can formulate the aforementioned correspondence as follows:

$$M(x, y) = \begin{cases} 1, & \text{if } \phi(x, y) > 0 \\ 0, & \text{otherwise} \end{cases} \quad \text{with } C = \partial M \quad (1)$$

Here,  $x$  and  $y$  represent the 2D spatial coordinates of the image, and  $\partial$  denotes the boundary operator. Accordingly,  $C = \partial M$  defines the object contour corresponding to the boundary of the binary mask  $M$ .

In our design, the predicted initial level set function  $\phi_0$  corresponds to the mask of visible regions for the target object, since the input point prompt only provides very coarse guidance to the visible part of the object. Compared to the traditional level set approach, our initialization is predicted by the network in an end-to-end approach.

### 3.3 Velocity-Guided Shape Evolution

The proposed VELA is designed to fully leverage the information provided by the sparse prompt to achieve high-quality amodal segmentation through a geometrically interpretable approach. A velocity-guided shape evolution process is introduced to drive the morphological transformation of the level set function from the initial state  $\phi_0$ , corresponding to the predicted visible mask, toward the complete object shape represented by the ground-truth amodal mask  $M_a$ . After  $T$  updates, the final predicted state is  $\phi_T$ . The complete evolution procedure is outlined in Alg. 1.

Specifically, the initial level set function  $\phi_0$  is evolved over  $T$  steps to expand the initial contour of the visible mask toward the target amodal mask, which typically encompasses a larger region than the visible mask. Based on the level set theory, for step  $i \in \{0, \dots, T - 1\}$ , the progressively updating process can be formulated as:

$$\phi_{i+1} = \phi_i + dt \frac{\partial \phi}{\partial t}, \quad (2)$$

where  $dt$  denotes the time step, representing the time amount for the contour to move per step, and  $\frac{\partial \phi}{\partial t}$  represents the temporal derivative of the level set function, governing its deformation over time. In classical level set theory, the contour evolves along its normal direction, and the update term is defined as:

$$\frac{\partial \phi}{\partial t} = -V_i^n \|\nabla \phi\|, \quad (3)$$

where  $V_i^n$  denotes the predicted normal velocity field at step  $i$ , which controls the speed of evolution along the normal direction, and  $\|\nabla \phi\|$  is the gradient magnitude that enforces evolution along the normal direction.

To simplify the computation and ensure numerical stability, we apply a regularization term after each evolution step to maintain the level set function  $\phi_i$  as a signed distance function (SDF), which keeps  $\|\nabla \phi\|$  close to 1 throughout the evolution. Under this condition, the evolution equation can be simplified by treating the gradient norm as nearly constant and close to 1. As a result, Eqn. 3 becomes:

$$\frac{\partial \phi}{\partial t} = -V_i^n, \quad (4)$$

leading to a simplified update rule:

$$\phi_{i+1} = \phi_i - dt \cdot V_i^n. \quad (5)$$

This formulation allows the evolving contour to be directly guided by the predicted scalar velocity field  $V_i^n$ , while the distance regularization ensures that the geometric meaning of the level set function is preserved throughout the evolution process.

To implement this, we adopt a distance regularization term  $\mathcal{R}(\phi)$  inspired by the double-well potential in DRLSE [30], which encourages  $\|\nabla \phi\|$  to stay close to 1. Specifically, at each step, we add a regularization update:

$$\tilde{\phi}_{i+1} = \phi_i - dt \cdot V_i^n, \quad \phi_{i+1} = \tilde{\phi}_{i+1} + \mu \cdot \mathcal{R}(\tilde{\phi}_{i+1}), \quad (6)$$

where  $\tilde{\phi}_{i+1}$  is the intermediate state after the velocity update and  $\mu$  is a weighting factor. As in Alg. 1, regularization is applied to this intermediate state. This regularization supports maintaining the signed distance property of  $\phi$ , thereby enabling the approximation that the gradient norm remains close to 1 throughout the evolution.

Finally, the amodal segmentation mask is obtained by applying a zero threshold to the final level set function  $\phi_T$ , with the predicted foreground amodal mask  $\hat{M}_a$  defined as  $\{(x, y) \mid \phi_T(x, y) > 0\}$ .

### 3.4 Shape-Specific Velocity Generation

To enable accurate and interpretable evolution of the level set function, we introduce the Shape-Specific Velocity Generator that dynamically predicts the normal velocity field  $V_i^n$  at each evolution step. Specifically, we employ a lightweight UNet architecture that takes the current level set function  $\phi_i$  as input and outputs a scalar velocity map that guides the contour deformation. This predicted velocity field encodes the directional cues required to deform the current visible contour toward the final amodal shape, allowing the model to learn shape-aware motion patterns conditioned on the evolving mask geometry dynamically.

*Network Architecture of the Shape-Specific Velocity Generator.* To guide the contour evolution in VELA, we employ a UNet network to dynamically predict the velocity field at each iteration. The input to this module is the current level set function  $\phi_i$ , and the output is a single-channel velocity map  $V_i$  that encodes the scalar speed of contour evolution at each spatial location.

Internally, we adopt a lightweight UNet architecture to extract spatial features from the input  $\phi_i$  and generate the velocity field, enabling flexible and data-adaptive evolution dynamics. By predicting the velocity field in a differentiable manner, this module allows the model to learn how to adjust the evolution process based on intermediate contour states, which is essential for accurate amodal mask recovery.

*Gradient Detachment of Initial Level Set.* We detach the initial level set function  $\phi_0$  before evolution to prevent the amodal supervision applied to later states from directly altering the visible-region initializer. This separates initial localization from subsequent amodal shape evolution.

### 3.5 Training Objective

To supervise the evolution of the level set function toward the desired amodal shape, we define a loss that provides step-wise guidance during training. After each evolution step  $i \in \{0, \dots, T-1\}$ , we apply supervision to the updated level set function  $\phi_{i+1}$ . The loss therefore covers all evolved states  $\phi_1, \dots, \phi_T$ , excluding the initial state  $\phi_0$ . We apply a smooth approximation of the Heaviside function  $H(\phi_{i+1})$  to obtain a differentiable soft mask, which maps positive values to foreground and negative values to background, enabling end-to-end learning. We then compute a binary cross-entropy (BCE) loss between the soft prediction and the ground-truth amodal mask  $M_a$ :

$$\mathcal{L}_{\text{evo}} = \sum_{i=0}^{T-1} \text{BCE}(H(\phi_{i+1}), M_a), \quad (7)$$

This cumulative loss encourages the model to learn meaningful velocity fields that progressively deform the evolving contour toward the complete amodal mask.

## 4 Experiments

To evaluate the effectiveness of the proposed VELA, we conduct experiments on three challenging datasets. Compared with existing methods, VELA achieves state-of-the-art performance. We illustrate the comparison both quantitatively and qualitatively. Finally, extensive ablation studies are conducted to validate the design of VELA.

### 4.1 Datasets and Evaluation Metrics

*Datasets.* We conduct experiments on three publicly available datasets to evaluate the performance of VELA. Specifically, first, the COCOA-cls [4] dataset contains 3,501 images and 10,592 ground-truth masks across 80 categories, which is constructed based on the famous COCO [42] dataset. The COCOA-cls dataset serves as a crucial benchmark due to its diverse and representative daily life scenarios. Second, the D2SA [4] dataset is designed for the shopping scenario, where self-checkout systems are commonly

**Table 1: Comparison with state-of-the-art approaches on the COCOA-cls dataset (validation split), the D2SA dataset (validation split), and the KINS dataset (testing split). Metrics  $mIoU_{full}$  and  $mIoU_{occ}$  are employed and higher values indicate better performance. Zero-shot methods are evaluated without training on the target datasets. Pix2Gestalt [52] uses a synthetic amodal training dataset containing 873K images. For the prompt type, “Mask” denotes the dense visible mask of the target object, while “Point” denotes a single sparse point located within the visible region.**

| Methods                  | Venue       | Prompt | COCOA-cls (Val)        |                       | D2SA (Val)             |                       | KINS (Test)            |                       |
|--------------------------|-------------|--------|------------------------|-----------------------|------------------------|-----------------------|------------------------|-----------------------|
|                          |             |        | mIoU <sub>full</sub> ↑ | mIoU <sub>occ</sub> ↑ | mIoU <sub>full</sub> ↑ | mIoU <sub>occ</sub> ↑ | mIoU <sub>full</sub> ↑ | mIoU <sub>occ</sub> ↑ |
| Zero-shot Methods        |             |        |                        |                       |                        |                       |                        |                       |
| SAM [29]                 | ICCV 2023   | Mask   | 73.10                  | -                     | 84.65                  | -                     | 75.88                  | -                     |
| SDXL-Inpainting [54]     | ICLR 2024   | Mask   | 73.65                  | -                     | 80.53                  | -                     | 76.19                  | -                     |
| Pix2Gestalt [52]         | CVPR 2024   | Mask   | 79.08                  | -                     | 81.82                  | -                     | 81.45                  | -                     |
| Fully-Supervised Methods |             |        |                        |                       |                        |                       |                        |                       |
| BCNet [26]               | CVPR 2021   | -      | 15.10                  | -                     | 74.90                  | -                     | 44.00                  | -                     |
| SLN [71]                 | MM 2019     | -      | 31.50                  | -                     | 31.20                  | -                     | 14.20                  | -                     |
| ORCNN [4]                | WACV 2019   | -      | 57.60                  | -                     | 74.10                  | -                     | 55.10                  | -                     |
| Mask-RCNN [7]            | ICCV 2017   | -      | 63.85                  | -                     | 74.62                  | -                     | 60.13                  | -                     |
| A3D [38]                 | ECCV 2022   | -      | 64.27                  | -                     | 74.71                  | -                     | 61.43                  | -                     |
| GIN [39]                 | TMM 2023    | -      | 72.46                  | -                     | 78.23                  | -                     | 68.31                  | -                     |
| AISFormer [63]           | BMVC 2022   | -      | 72.69                  | 13.75                 | 86.81                  | 30.01                 | 81.53                  | 48.54                 |
| PCNet [69]               | CVPR 2020   | Mask   | 76.91                  | 20.34                 | 80.45                  | 28.56                 | 78.02                  | 38.14                 |
| VRSP [66]                | AAAI 2021   | -      | 78.98                  | 22.92                 | 88.08                  | 35.17                 | 80.70                  | 47.33                 |
| SDAmodal [68]            | CVPR 2024   | Mask   | 80.01                  | -                     | -                      | -                     | -                      | -                     |
| C2F-Seg [5]              | ICCV 2023   | Mask   | 80.28                  | 27.71                 | 89.10                  | -                     | 82.22                  | 53.60                 |
| VELA (Ours)              | ACM MM 2026 | Point  | 80.68                  | 25.00                 | 89.96                  | 39.41                 | 83.06                  | 50.29                 |

used. With 5,600 images and 28,720 ground-truth masks in 60 categories provided, the D2SA dataset can effectively examine the amodal segmentation ability of methods with its intrinsic occlusions constructed by overlapping merchandise. Finally, built upon the KITTI [6] dataset, the KINS [55] dataset comprises 7,474 and 7,517 images in the training and testing splits, respectively. It primarily captures street scenes, making it a valuable benchmark for assessing amodal segmentation performance in complex urban environments.

For each dataset, fully supervised methods are trained on its official training split. Zero-shot methods are evaluated without training on the target datasets. All methods are evaluated on the same official validation or testing split. All experiments of VELA are conducted three times, and the average performance is reported to mitigate the impact of randomness.

*Evaluation Metrics.* Following [5], we adopt mean Intersection over Union ( $mIoU$ ) as the evaluation metric, computed over two regions: the entire amodal mask and the occluded regions. These are denoted as  $mIoU_{full}$  and  $mIoU_{occ}$ , respectively. Larger values indicate better performance.

## 4.2 Implementation Details

We implement our proposed VELA framework based on PyTorch [53]. VELA takes SAM 2 [57] as the base structure, whose pretrained weights are used for initializing VELA. The optimizer for gradient descent is AdamW. For computing infrastructure, all experiments are conducted on a machine equipped with an AMD 9950X3D with 96 GB system memory, an NVIDIA 5090 GPU with 32 GB of

memory, and the Windows 11 operating system. There are no data augmentation techniques employed in the experiments of VELA.

## 4.3 Comparison with State-of-the-Art Methods

We compare the proposed VELA with state-of-the-art approaches, including both zero-shot and fully supervised methods. Specifically, the zero-shot baselines include SAM, SDXL-Inpainting, and Pix2Gestalt, which are evaluated without training on the target datasets. Pix2Gestalt is trained on a synthetic amodal segmentation dataset [52] comprising approximately 873,000 images with simulated occlusions and corresponding ground-truth amodal masks. Fully-supervised methods, including VELA, are trained exclusively on the training splits of the three datasets, without utilizing any additional amodal segmentation data. Several methods use the visible mask of the target object as the input prompt, whereas VELA relies solely on a single point located within the visible region.

As shown in Tab. 1, VELA achieves the highest  $mIoU_{full}$  on all three benchmarks. For  $mIoU_{occ}$ , VELA exceeds all baselines with reported values except C2F-Seg on COCOA-cls and KINS; C2F-Seg uses a dense visible-mask prompt, whereas VELA uses a single sparse point. Compared to the zero-shot method Pix2Gestalt, which is pretrained on a large-scale amodal segmentation dataset, VELA achieves gains of 1.60, 8.14, and 1.61 percentage points in  $mIoU_{full}$  across the three benchmarks. VELA also surpasses the fully supervised C2F-Seg in  $mIoU_{full}$  on all three datasets. Despite using only a single point as the prompt, VELA therefore achieves higher full-mask accuracy than all compared methods that rely on visible-mask prompts. This result demonstrates the effectiveness of our geometrically interpretable Level Set evolution, which progressively

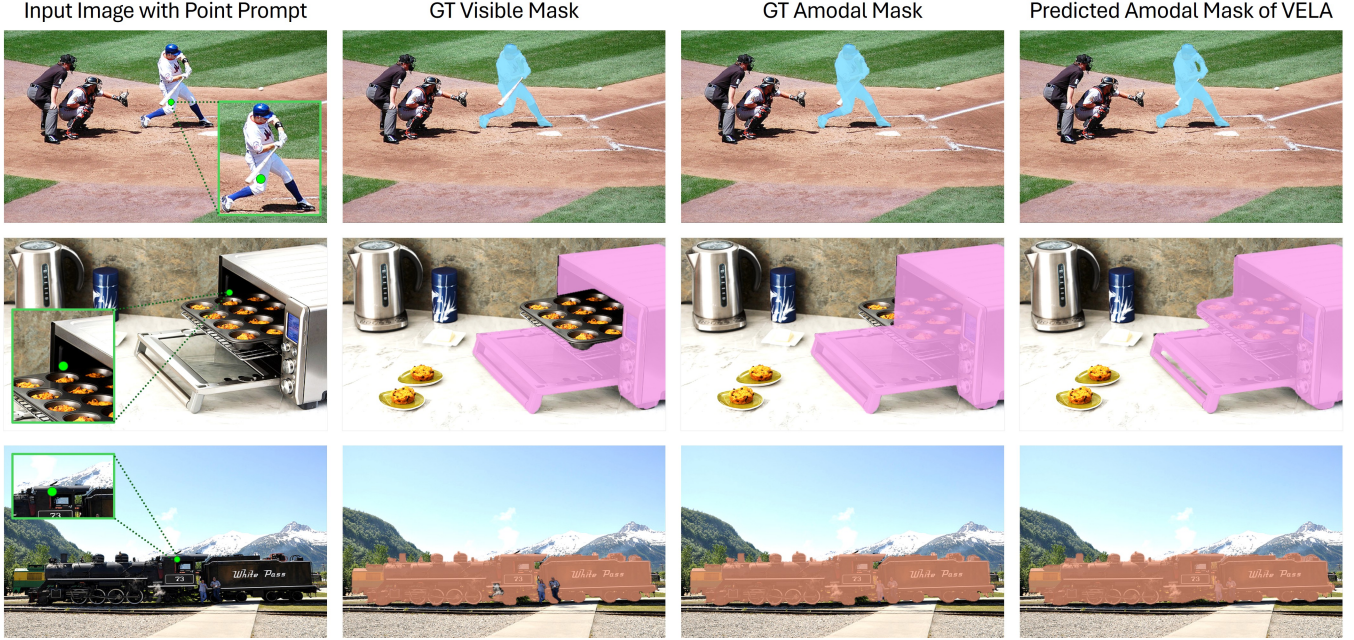


Figure 3: Visualization results of VELA. The green dot denotes the input point prompt. Best viewed in color.

Table 2: Ablation study of the evolution steps.

| Index | Evolution Steps $T$ | COCOA-cls    | D2SA         | KINS         |
|-------|---------------------|--------------|--------------|--------------|
| 1     | 1                   | 58.42        | 62.36        | 56.49        |
| 2     | 2                   | 69.67        | 79.20        | 66.99        |
| 3     | 3                   | <b>80.68</b> | <b>89.96</b> | <b>83.06</b> |
| 4     | 4                   | 78.68        | 83.52        | 80.58        |
| 5     | 5                   | 79.09        | 88.75        | 77.66        |
| 6     | 10                  | <u>79.75</u> | <u>89.65</u> | 78.39        |

transforms the visible region into the complete amodal shape under learned velocity guidance.

Moreover, Fig. 3 and 4 show qualitative results of VELA on the COCOA-cls dataset. The qualitative examples show that VELA can recover coherent amodal shapes across diverse occlusion patterns, although highly irregular local boundaries remain challenging when visual evidence is limited.

#### 4.4 Ablation Study

We conduct ablation studies on VELA to evaluate the effectiveness of its key components, including level set evolution steps, the number of input points, and the segmentation supervision strategy. Unless otherwise stated, all experiments are conducted using three evolution steps and a single-point prompt. Results are reported using the  $mIoU_{full}$  metric on the COCOA-cls, D2SA, and KINS datasets.

*Effect of Evolution Steps  $T$ .* As shown in Tab. 2, we evaluate the impact of varying the number of level set evolution steps  $T$  in VELA. The best performance is observed when  $T = 3$ . While fewer steps

Table 3: Ablation study of the number of input points.

| Index | # Input Point(s) | COCOA-cls | D2SA  | KINS  |
|-------|------------------|-----------|-------|-------|
| 1     | 1                | 80.68     | 89.96 | 83.06 |
| 2     | 5                | 80.72     | 90.12 | 83.51 |
| 3     | 10               | 80.79     | 90.23 | 83.94 |
| 4     | 20               | 80.83     | 90.51 | 83.96 |
| 5     | 50               | 79.86     | 89.11 | 83.02 |

Table 4: Ablation study of the supervision design.

| Index | Supervision Design      | COCOA-cls    | D2SA         | KINS         |
|-------|-------------------------|--------------|--------------|--------------|
| 1     | Only final prediction   | 79.12        | 88.64        | 81.47        |
| 2     | All evolved predictions | <b>80.68</b> | <b>89.96</b> | <b>83.06</b> |

may limit occlusion recovery, a large number of steps can introduce mild contour smoothing or deformation, which may have a certain impact on the quality of predicted amodal masks.

*Effect of Input Points Number.* As shown in Tab. 3, we investigate the impact of using different numbers of point prompts. Increasing the number of input points brings modest improvements up to 20 points, but further increases offer little benefit and may introduce noise that affects performance. Notably, VELA performs well even with a single-point prompt, demonstrating its ability to effectively utilize sparse guidance for high-quality amodal segmentation.

*Effect of Supervision Design.* We investigate two supervision strategies by either applying the loss only to the final evolved



Figure 4: Visualization results of VELA. The green dot denotes the input point prompt. Best viewed in color.

prediction  $\phi_T$  or supervising all updated states  $\phi_1, \dots, \phi_T$ , including the final prediction and excluding the initial state  $\phi_0$ . As shown in Tab. 4, incorporating supervision at all steps consistently improves performance across datasets. This design enables step-wise supervision for the velocity field, allowing the model to refine the shape evolution more effectively and maintain stability throughout the process.

## 5 Discussions

Level sets provide geometry-aware contour evolution and accommodate topology changes under occlusion. VELA combines semantic completion from learned velocity fields with contour continuity from regularization, although weak local cues can lead to over-smoothed fine boundaries. A single point minimizes interaction while supporting accurate amodal recovery; multiple points can provide more precise localization when available. Future extensions may incorporate upstream detections, temporal consistency, or language conditioning. Complementary multimedia components include adaptive acquisition and reconstruction [13, 15–17], while spatio-temporal modeling, watermarking, and source-camera identification offer further system-level capabilities [11, 12, 14, 34].

## 6 Conclusion and Limitation Analysis

VELA formulates single-point amodal segmentation as explicit contour evolution driven by a differentiable velocity generator. Unlike direct mask regression, VELA initializes a level set function from a

single-point prompt and progressively evolves its zero-level contour using shape-specific velocity fields. This formulation combines learned semantic completion with geometry-aware contour evolution and exposes the intermediate deformation process before the final mask is obtained. Experiments on COCOA-cls, D2SA, and KINS demonstrate strong performance with sparse guidance, showing that a single point can provide an effective initialization for recovering complete object shapes under occlusion. Future work will investigate instance-adaptive stopping criteria and closed-loop interaction, allowing additional corrective prompts to refine uncertain contours while preserving the efficiency of sparse guidance.

## Acknowledgments

This research was supported by the Ministry of Education of Singapore Tier3 grant MOET32022-0006. This research was partially funded by National Natural Science Foundation of China (62506009) and China Postdoctoral Science Foundation (2026M791599). This work was supported by the Natural Science Foundation of Jiangsu under Grant BK20250731, the General Program of China Postdoctoral Science Foundation under grant 2025M784450, the Startup Foundation for Introducing Talent of Nanjing University of Information Science and Technology under grant 2025r029.

## References

- [1] David Acuna, Amlan Kar, and Sanja Fidler. 2019. Devil is in the edges: Learning semantic boundaries from noisy annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11075–11083.
- [2] Vicent Caselles, Ron Kimmel, and Guillermo Sapiro. 1997. Geodesic active contours. *International journal of computer vision* 22, 1 (1997), 61–79.
- [3] Junjie Chen, Li Niu, Jianfu Zhang, Jianlou Si, Chen Qian, and Liqing Zhang. 2023. Amodal instance segmentation via prior-guided expansion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 313–321.
- [4] Patrick Follmann, Rebecca König, Philipp Härtinger, Michael Klostermann, and Tobias Böttger. 2019. Learning to see the invisible: End-to-end trainable amodal instance segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. IEEE, 1328–1336.
- [5] Jianxiong Gao, Xuelin Qian, Yikai Wang, Tianjun Xiao, Tong He, Zheng Zhang, and Yanwei Fu. 2023. Coarse-to-fine amodal segmentation with shape prior. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1262–1271.
- [6] Andreas Geiger, Philip Lenz, and Raquel Urtasun. 2012. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 3354–3361.
- [7] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask R-CNN. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2961–2969.
- [8] Namdar Homayounfar, Yuwen Xiong, Justin Liang, Wei-Chiu Ma, and Raquel Urtasun. 2020. Levelset r-cnn: A deep variational method for instance segmentation. In *Proceedings of the IEEE/CVF European Conference on Computer Vision*. Springer, 555–571.
- [9] Ping Hu, Bing Shuai, Jun Liu, and Gang Wang. 2017. Deep level sets for salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2300–2309.
- [10] Yuan-Ting Hu, Hong-Shuo Chen, Kexin Hui, Jia-Bin Huang, and Alexander G Schwing. 2019. SAIL-VOS: Semantic amodal instance level video object segmentation—a synthetic dataset and baselines. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3105–3115.
- [11] Chen Hui, Feng Jiang, Shaohui Liu, and Debin Zhao. 2022. Source Camera Identification with Multi-Scale Feature Fusion Network. In *2022 IEEE International Conference on Multimedia and Expo*. 1–6. doi:10.1109/ICME52920.2022.9859965
- [12] Chen Hui, Shaohui Liu, Wenxue Cui, Jinghua Zeng, Feng Jiang, and Debin Zhao. 2021. Adaptive Flexible 3D Histogram Watermarking. In *2021 IEEE International Conference on Multimedia and Expo*. 1–6. doi:10.1109/ICME51207.2021.9428143
- [13] Chen Hui, Shaohui Liu, and Feng Jiang. 2022. Multi-Channel Adaptive Partitioning Network for Block-Based Image Compressive Sensing. In *2022 IEEE International Conference on Multimedia and Expo*. 1–6. doi:10.1109/ICME52920.2022.9859846
- [14] Chen Hui, Shaohui Liu, Wuzhen Shi, Feng Jiang, and Debin Zhao. 2022. Spatio-Temporal Context Based Adaptive Camcorder Recording Watermarking. *ACM Transactions on Multimedia Computing, Communications, and Applications* 18, 3s (2022), 1–25. doi:10.1145/3503160
- [15] Chen Hui, Shengping Zhang, Wenxue Cui, Shaohui Liu, Feng Jiang, and Debin Zhao. 2023. Rate-Adaptive Neural Network for Image Compressive Sensing. *IEEE Transactions on Multimedia* 26 (2023), 2515–2530. doi:10.1109/TMM.2023.3301213
- [16] Chen Hui, Debin Zhao, Weisi Lin, Shaohui Liu, and Feng Jiang. 2025. Image Compressive Sensing With Scale-Variable Adaptive Sampling and Hybrid-Attention Transformer Reconstruction. *IEEE Transactions on Multimedia* 27 (2025), 4333–4347. doi:10.1109/TMM.2025.3535114
- [17] Chen Hui, Haiqi Zhu, Shuya Yan, Shaohui Liu, Feng Jiang, and Debin Zhao. 2024. S<sup>2</sup>-CSNet: Scale-Aware Scalable Sampling Network for Image Compressive Sensing. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 2515–2524. doi:10.1145/3664647.3681323
- [18] Xueying Jiang, Jiaxing Huang, Sheng Jin, and Shijian Lu. 2023. Domain Generalization via Balancing Training Difficulty and Model Capability. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 18993–19003.
- [19] Xueying Jiang, Sheng Jin, Lewei Lu, Xiaoqin Zhang, and Shijian Lu. 2024. Weakly Supervised Monocular 3D Detection with a Single-View Image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10508–10518.
- [20] Xueying Jiang, Sheng Jin, Xiaoqin Zhang, Ling Shao, and Shijian Lu. 2024. Mono-MAE: Enhancing Monocular 3D Detection through Depth-Aware Masked Autoencoders. In *Advances in Neural Information Processing Systems*, Vol. 37. 11392–11411. doi:10.52202/079017-0362
- [21] Xueying Jiang, Wenhao Li, Quanhao Qian, Deli Zhao, Shijian Lu, Gongjie Zhang, and Ran Xu. 2026. Towards Camera-Robust 3D Localization: Equation-Anchored Tool-Use for MLLMs. *arXiv preprint arXiv:2605.19528* (2026).
- [22] Xueying Jiang, Wenhao Li, Xiaoqin Zhang, Ling Shao, and Shijian Lu. 2025. Exploring 3D Reasoning-Driven Planning: From Implicit Human Intentions to Route-Aware Activity Planning. *arXiv preprint arXiv:2503.12974* (2025).
- [23] Xueying Jiang, Lewei Lu, Ling Shao, and Shijian Lu. 2024. Multimodal 3D Reasoning Segmentation with Complex Scenes. *arXiv preprint arXiv:2411.13927* (2024).
- [24] Sheng Jin, Xueying Jiang, Jiaxing Huang, Lewei Lu, and Shijian Lu. 2024. LLMs Meet VLMs: Boost Open Vocabulary Object Detection with Fine-grained Descriptors. In *International Conference on Learning Representations*. <https://openreview.net/forum?id=usrChqw6yK>
- [25] Michael Kass, Andrew Witkin, and Demetri Terzopoulos. 1988. Snakes: Active contour models. *International journal of computer vision* 1, 4 (1988), 321–331.
- [26] Lei Ke, Yu-Wing Tai, and Chi-Keung Tang. 2021. Deep occlusion-aware instance segmentation with overlapping bilayers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4019–4028.
- [27] Lei Ke, Mingqiao Ye, Martin Danelljan, Yifan Liu, Yu-Wing Tai, Chi-Keung Tang, and Fisher Yu. 2023. Segment Anything in High Quality. In *Advances in Neural Information Processing Systems*, Vol. 36. 29914–29934. doi:10.52202/075280-1303
- [28] Boah Kim and Jong Chul Ye. 2019. Mumford–Shah loss functional for image segmentation with deep learning. *IEEE transactions on image processing* 29 (2019), 1856–1866. doi:10.1109/TIP.2019.2941265
- [29] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. 2023. Segment Anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4015–4026.
- [30] Chunming Li, Chenyang Xu, Changfeng Gui, and Martin D Fox. 2010. Distance regularized level set evolution and its application to image segmentation. *IEEE transactions on image processing* 19, 12 (2010), 3243–3254.
- [31] Ke Li and Jitendra Malik. 2016. Amodal instance segmentation. In *Proceedings of the IEEE/CVF European Conference on Computer Vision*. Springer, 677–693.
- [32] Wenhao Li, Xueying Jiang, Quanhao Qian, Deli Zhao, Shijian Lu, Gongjie Zhang, and Ran Xu. 2026. From Fixed to Free Cameras: Calibration-Free View-Robust Vision-Language-Action Model. *arXiv preprint arXiv:2607.05396* (2026).
- [33] Wenhao Li, Xueying Jiang, Quanhao Qian, Deli Zhao, Ran Xu, Shijian Lu, and Gongjie Zhang. 2026. 3D-Aware VLMs with Implicit and Explicit Geometries. *arXiv preprint arXiv:2607.21595* (2026).
- [34] Wenhao Li, Xueying Jiang, Gongjie Zhang, Xiaoqin Zhang, Ling Shao, and Shijian Lu. 2026. STS-Mixer: Spatio-Temporal-Spectral Mixer for 4D Point Cloud Video Understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Findings*. 8185–8194.
- [35] Wentong Li, Wenyu Liu, Jianke Zhu, Miaomiao Cui, Xian-Sheng Hua, and Lei Zhang. 2022. Box-supervised instance segmentation with level set evolution. In *European conference on computer vision*. Springer, 1–18.
- [36] Zhixuan Li, Yujia Liu, Chen Hui, Jeonghaeng Lee, Sanghoon Lee, and Weisi Lin. 2025. Shape Distribution Matters: Shape-specific Mixture-of-Experts for Amodal Segmentation under Diverse Occlusions. *arXiv preprint arXiv:2508.01664* (2025). doi:10.48550/arXiv.2508.01664
- [37] Zhixuan Li, Ruohua Shi, Tiejun Huang, and Tingting Jiang. 2023. OAFormer: Learning occlusion distinguishable feature for amodal instance segmentation. In *International Conference on Acoustics, Speech, and Signal Processing*. IEEE, 1–5.
- [38] Zhixuan Li, Weining Ye, Tingting Jiang, and Tiejun Huang. 2022. 2D amodal instance segmentation guided by 3D shape prior. In *Proceedings of the IEEE/CVF European Conference on Computer Vision*. Springer, 165–181.
- [39] Zhixuan Li, Weining Ye, Tingting Jiang, and Tiejun Huang. 2023. GIN: Generative INvariant Shape Prior for Amodal Instance Segmentation. *IEEE Transactions on Multimedia* 26 (2023), 3924–3936. doi:10.1109/TMM.2023.3318075
- [40] Zhixuan Li, Weining Ye, Juan Terven, Zachary Bennett, Ying Zheng, Tingting Jiang, and Tiejun Huang. 2023. MUVA: A new large-scale benchmark for multi-view amodal instance segmentation in the shopping scenario. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 23504–23513.
- [41] Zhixuan Li, Hyunse Yoon, Sanghoon Lee, and Weisi Lin. 2025. Unveiling the Invisible: Reasoning Complex Occlusions Amodally with AURA. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 21927–21937. doi:10.1109/ICCV51701.2025.02036
- [42] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Proceedings of the IEEE/CVF European Conference on Computer Vision*. Springer, 740–755.
- [43] Huan Ling, David Acuna, Karsten Kreis, Seung Wook Kim, and Sanja Fidler. 2020. Variational amodal object completion. *Advances in Neural Information Processing Systems* 33 (2020), 16246–16257.
- [44] Yujia Liu, Chenxi Yang, Dingquan Li, Tingting Jiang, and Tiejun Huang. 2025. A Norm Regularization Training Strategy for Robust Image Quality Assessment Models. *International Journal of Computer Vision* 133, 9 (2025), 5883–5897. doi:10.1007/s11263-025-02458-8
- [45] Zhaochen Liu, Zhixuan Li, and Tingting Jiang. 2024. BLADE: Box-Level Supervised Amodal Segmentation through Directed Expansion. *Proceedings of the AAAI Conference on Artificial Intelligence* 38, 4 (2024), 3846–3854. doi:10.1609/aaai.v38i4.28176
- [46] Zhaochen Liu, Limeng Qiao, Xiangxiang Chu, Lin Ma, and Tingting Jiang. 2025. Towards Efficient Foundation Model for Zero-shot Amodal Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

- 2025a–2026a.
- [47] Ravi Malladi, James A. Sethian, and Baba C. Vemuri. 1995. Shape Modeling with Front Propagation: A Level Set Approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 17, 2 (1995), 158–175. doi:10.1109/34.368173
- [48] Hayk Manukyan, Andranik Sargsyan, Barsegh Atanyan, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. 2025. HD-Painter: High-Resolution and Prompt-Faithful Text-Guided Image Inpainting with Diffusion Models. In *The Thirteenth International Conference on Learning Representations*. <https://openreview.net/forum?id=6lB5qtYdYAg>
- [49] Oscar Michel, Anand Bhattad, Eli VanderBilt, Ranjay Krishna, Aniruddha Kembhavi, and Tanmay Gupta. 2023. Object 3dit: Language-guided 3d-aware image editing. *Advances in Neural Information Processing Systems* 36 (2023), 3497–3516.
- [50] Tuan Anh Ngo and Gustavo Carneiro. 2013. Left ventricle segmentation from cardiac MRI combining level set methods with deep belief networks. In *IEEE International Conference on Image Processing*. IEEE, 695–699.
- [51] Khoi Nguyen and Sinisa Todorovic. 2021. A weakly supervised amodal segmenter with boundary uncertainty estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 7396–7405.
- [52] Ege Ozguroglu, Ruoshi Liu, Didac Suris, Dian Chen, Achal Dave, Pavel Tokmakov, and Carl Vondrick. 2024. pix2gestalt: Amodal Segmentation by Synthesizing Wholes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3931–3940.
- [53] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems* 32 (2019).
- [54] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2024. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *International Conference on Learning Representations* (2024).
- [55] Lu Qi, Li Jiang, Shu Liu, Xiaoyong Shen, and Jiaya Jia. 2019. Amodal instance segmentation with KINS dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3014–3023.
- [56] Yansong Qu, Shaohui Dai, Xinyang Li, Yuze Wang, You Shen, Shengchuan Zhang, and Liujuan Cao. 2026. Deocc-1-to-3: 3d de-occlusion from a single image via self-supervised multi-view diffusion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 40. 8677–8685. doi:10.1609/AAAI.V40I11.37820
- [57] Nikhila Ravi, Valentim Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Juntang Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. 2025. SAM 2: Segment Anything in Images and Videos. In *The Thirteenth International Conference on Learning Representations*. <https://openreview.net/forum?id=Ha6RTeWmD0>
- [58] Yizhi Song, Zhifei Zhang, Zhe Lin, Scott Cohen, Brian Price, Jianming Zhang, Soo Ye Kim, and Daniel Aliaga. 2023. Objectstitch: Object compositing with diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18310–18319.
- [59] Jingxiang Sun, Xuan Wang, Yichun Shi, Lizhen Wang, Luca Wang, and Yebin Liu. 2022. Ide-3d: Interactive disentangled editing for high-resolution 3d-aware portrait synthesis. *ACM Transactions on Graphics (ToG)* 41, 6, Article 270 (2022), 10 pages. doi:10.1145/3550454.3555506
- [60] Wei-En Tai, Yu-Lin Shih, Cheng Sun, Yu-Chiang Frank Wang, and Hwann-Tzong Chen. 2025. Segment Anything, Even Occluded. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 29385–29394.
- [61] Min Tang, Sepehr Valipour, Zichen Zhang, Dana Cobzas, and Martin Jagersand. 2017. A deep level set method for image segmentation. In *International Workshop on Deep Learning in Medical Image Analysis*. Springer, 126–134.
- [62] Minh Tran, Winston Bounsavay, Khoa Vo, Anh Nguyen, Tri Nguyen, and Ngan Le. 2024. ShapeFormer: Shape Prior Visible-to-Amodal Transformer-Based Amodal Instance Segmentation. In *2024 International Joint Conference on Neural Networks*. 1–8. doi:10.1109/IJCNN60899.2024.10650837
- [63] Minh Tran, Khoa Vo, Kashu Yamazaki, Arthur Fernandes, Michael Kidd, and Ngan Le. 2022. Aisformer: Amodal instance segmentation with transformer. In *33rd British Machine Vision Conference 2022, BMVC 2022*.
- [64] Su Wang, Chitwan Saharia, Ceslee Montgomery, Jordi Pont-Tuset, Shai Noy, Stefano Pellegrini, Yasumasa Onoe, Sarah Laszlo, David J Fleet, Radu Soricut, et al. 2023. Imagen editor and editbench: Advancing and evaluating text-guided image inpainting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 18359–18369.
- [65] Zian Wang, David Acuna, Huan Ling, Amlan Kar, and Sanja Fidler. 2019. Object instance annotation with deep extreme level set evolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7500–7508.
- [66] Yuting Xiao, Yanyu Xu, Ziming Zhong, Weixin Luo, Jiawei Li, and Shenghua Gao. 2021. Amodal segmentation based on visible region segmentation and shape prior. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 2995–3003.
- [67] Yunyang Xiong, Bala Varadarajan, Lemeng Wu, Xiaoyu Xiang, Fanyi Xiao, Chenchen Zhu, Xiaoliang Dai, Dilin Wang, Fei Sun, Forrest Iandola, et al. 2024. EfficientSAM: Leveraged masked image pretraining for efficient segment anything. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 16111–16121.
- [68] Guanqi Zhan, Chuanxia Zheng, Weidi Xie, and Andrew Zisserman. 2024. Amodal Ground Truth and Completion in the Wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 28003–28013.
- [69] Xiaohang Zhan, Xingang Pan, Bo Dai, Ziwei Liu, Dahua Lin, and Chen Change Loy. 2020. Self-supervised scene de-occlusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3784–3792.
- [70] Jingyi Zhang, Jiaxing Huang, Xueying Jiang, and Shijian Lu. 2023. Black-Box Unsupervised Domain Adaptation with Bi-Directional Atkinson-Shiffrin Memory. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 11771–11782.
- [71] Ziheng Zhang, Anpei Chen, Ling Xie, Jingyi Yu, and Shenghua Gao. 2019. Learning semantics-aware distance map with semantics layering network for amodal instance segmentation. In *ACM International Conference on Multimedia*. 2124–2132.
- [72] Chuanxia Zheng, Duy-Son Dao, Guoxian Song, Tat-Jen Cham, and Jianfei Cai. 2021. Visiting the invisible: Layer-by-layer completed scene decomposition. *International Journal of Computer Vision* 129 (2021), 3195–3215.
- [73] Yuan Zhou, Yanrong Guo, Shijie Hao, and Richang Hong. 2022. Hierarchical Prototype Refinement With Progressive Inter-Categorical Discrimination Maximization for Few-Shot Learning. *IEEE Transactions on Image Processing* 31 (2022), 3414–3429. doi:10.1109/TIP.2022.3170727
- [74] Yuan Zhou, Yanrong Guo, Shijie Hao, Richang Hong, and Jiebo Lu. 2023. Few-Shot Partial Multi-View Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 10 (2023), 11824–11841. doi:10.1109/TPAMI.2023.3275162
- [75] Yuan Zhou, Qingshan Xu, Jiequan Cui, Junbao Zhou, Jing Zhang, Richang Hong, and Hanwang Zhang. 2025. CARE Transformer: Mobile-Friendly Linear Visual Transformer via Decoupled Dual Interaction. In *Proceedings of the IEEE/CVF*