Amodal Instance Segmentation with Diffusion Shape Prior Estimation

Minh Tran[1], Khoa Vo[1], Tri Nguyen[2], and Ngan Le[1]

[1] University of Arkansas, Fayetteville AR, USA
[2] Coupang, Inc., Seattle WA, USA
https://uark-aicv.github.io/AISDiff

Abstract. Amodal Instance Segmentation (AIS) presents an intriguing challenge, including the segmentation prediction of both visible and occluded parts of objects within images. Previous methods have often relied on shape prior information gleaned from training data to enhance amodal segmentation. However, these approaches are susceptible to overfitting and disregard object category details. Recent advancements highlight the potential of conditioned diffusion models, pretrained on extensive datasets, to generate images from latent space. Drawing inspiration from this, we propose AISDiff with a Diffusion Shape Prior Estimation (DiffSP) module. AISDiff begins with the prediction of the visible segmentation mask and object category, alongside occlusion-aware processing through the prediction of occluding masks. Subsequently, these elements are inputted into our DiffSP module to infer the shape prior of the object. DiffSP utilizes conditioned diffusion models pretrained on extensive datasets to extract rich visual features for shape prior estimation. Additionally, we introduce the Shape Prior Amodal Predictor, which utilizes attention-based feature maps from the shape prior to refine amodal segmentation. Experiments across various AIS benchmarks demonstrate the effectiveness of our AISDiff.

1 Introduction

Amodal perception, as described in [18], describe human's remarkable ability to perceive objects in their entirety despite occlusion. Building upon this concept, the pioneering studies by [21,46] introduced amodal instance segmentation (AIS). This approach aims to predict the complete shape of objects, encompassing both their visible and occluded regions. Indeed, AIS exhibits vast potential across various domains, as evidenced by its applications in robot manipulation [2] and autonomous driving [27]. Across various AIS benchmarks [7,27,46], a multitude of approaches addressing the AIS challenge have emerged in the literature. These approaches, as evidenced by numerous studies [7,15,21,23,27,36,37,39], demonstrate the ongoing efforts to tackle this challenge.

Recent research [6,9,15,40,43] highlights the effectiveness of integrating shape prior information in AIS. Indeed, These shape prior AIS methods typically construct shape-prior knowledge from the training dataset, which is later utilized

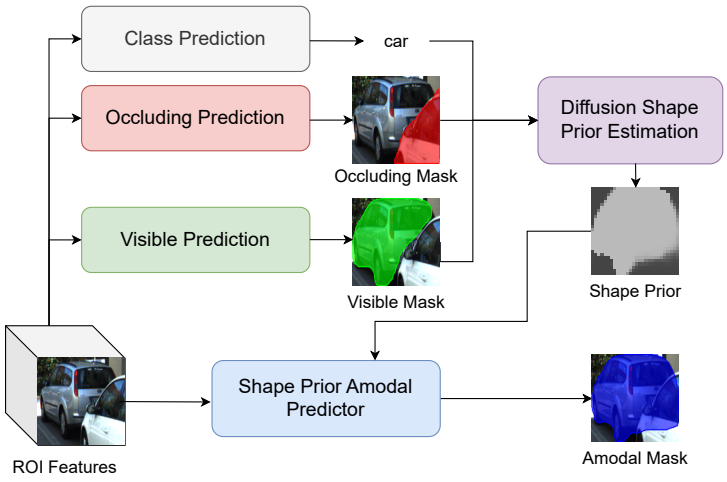

Fig. 1: Overall architecture of AISDiff. AISDiff predicts the visible segmentation mask and the object category while simultaneously addressing occlusion by predicting the occluding mask. Next, these predictions are integrated into the Diffusion Shape Prior Estimation (DiffSP) module to establish the object's shape prior. This shape prior is then utilized by AISDiff to produce the amodal segmentation.

to train the AIS model. In [40], for instance, the authors employ variational autoencoders to reconstruct amodal masks. The concept revolves around using ground truth amodal masks, utilizing autoencoders to reconstruct them, and storing the encoded codebooks as shape priors. Similarly, in [9], the authors also construct a shape prior codebook but employ a vector-quantization variational autoencoder. After establishing the shape prior, these method first predict the coarse amodal segmentation and refine the final amodal segmentation mask using the built shape prior. However, there are limitations to these approaches. Firstly, the shape prior tends to overfit to the training data, consequently leading to overfitting in amodal mask prediction overall. Secondly, since the shape prior is built solely from ground truth amodal masks, it may overlook the object category, which could provide significant supplementary information for deriving the shape prior.

To tackle these issues, we desgin a **AIS** mask head with **Diff**usion Shape Prior Estimation (**AISDiff**). The design of AISDiff is depicted in Figure 1. In essence, AISDiff begins by predicting the visible segmentation mask and the category of the object of interest. Simultaneously, it conducts occlusion-aware processing by predicting the occluding mask, which is the segmentation of occluding elements within the specified ROI. Subsequently, these three pieces of information are fed into the proposed Diffusion Shape Prior Estimation (DiffSP) module to derive the shape prior of the object. Finally, leveraging this shape prior, AISDiff generates the amodal segmentation.

Specifically, DiffSP leverages the successes of conditioned diffusion models (such as Stable Diffusion [28] and GLIDE [25]), which are pretrained on extensive language vision datasets like LAION [33]. This enables the model to capture rich visual features, making it suitable as prior knowledge for downstream tasks [26, 44]. Building upon this foundation, we feed a trained conditioned diffusion model with an ROI image containing only the visible pixels of the object of interest, expecting the model to generate the missing parts. Additionally, an occluding mask and a textual description of the object category is also feed to condition the mdoel. Subsequently, the denoising process iterates T steps to output the generated image containing the occluded parts. However, rather than relying on the final generated pixels, DiffSP exploits on the attention mechanism between the conditioning information and the image features. This attention map remains relatively stable across time steps, thereby reducing the denoising time needed to obtain the shape prior. Furthermore, we design the Shape Prior Amodal Predictor, which learns the attention-based amodal feature map from the acquired shape prior to predict the amodal mask segmentation.

In summary, our contributions are as follows:

- We present AISDiff, a novel AIS mask head featuring a Diffusion Shape Prior Estimation module. This model predicts the visible segmentation mask and category of the object while considering occlusion. It then uses these predictions to estimate the shape prior of the object before generating the final amodal segmentation mask.
- We propose DiffSP module, harnessing the efficacy of conditioned diffusion models to derive the shape prior of the object of interest.
- We introduce the Shape Prior Amodal Predictor, which learns attention-based amodal feature maps from the obtained shape prior to predict the amodal segmentation.

2 Related Work

2.1 Amodal Segmentation

Amodal instance segmentation involves predicting an object's shape, including both its visible and occluded parts. Li and Malik [21] pioneered a method aimed at addressing AIS. They proposed enlarging the modal bounding box in alignment with high heatmap values and synthesizing occlusions. Following this seminal work, various methodologies have surfaced in literature. Notably, ORCNN [7] introduces instance mask heads for both amodal and visible instances, along with an additional head for predicting occluded masks. ASN [27] builds upon ORCNN by integrating a multi-level coding module for bidirectional feature modeling of visible and amodal aspects. BCNet [17] enhances amodal mask prediction by incorporating a supplementary branch dedicated to predicting occlusion masks within the bounding box. AISFormer [37] introduces

a transformer-based mask head, demonstrating the efficacy of transformer modeling in generating AIS masks. However, their approach, which consolidates all mask relationships into one transformer model, leads to compromised visible segmentation output, consequently affecting the quality of amodal segmentation output due to bidirectional feature relations as mentioned earlier.

Recent studies [9, 15, 40] underscore the benefits of integrating shape priors into AIS. These methods leverage prior knowledge of mask shapes to improve amodal mask predictions. VRSP-Net [40] predicts coarse amodal masks, retrieves shape priors using a simple autoencoder, and then refines the final amodal mask predictions. AmodalBlastomere [15] employs a similar strategy with a variational autoencoder for blastomere and cell segmentation. C2F-Seg [9] constructs a shape prior codebook using a vector-quantization variational autoencoder. After establishing the shape prior, the method first predicts a coarse amodal segmentation. This coarse segmentation is then refined to produce the final amodal segmentation mask using the built shape prior. Despite their progress, these methods often overlook the importance of object categories when utilizing prior shapes. Moreover, their training procedures frequently lead to overfitting of the shape prior model to the training dataset. Additionally, these approaches simply incorporate the shape prior by concatenating it with visible features to refine amodal masks.

2.2 Diffusion Models

The Denoising Diffusion Probabilistic Model (DDPM) [13] has become a widely used generative architecture in computer vision. Its popularity stems from its ability to model multi-modal distributions, training stability, and scalability. The study by [5] first showed that diffusion models outperform GANs [10] in image synthesis. To enhance computational efficiency, Stable Diffusion [28], trained on LAION-5B [32], applied a diffusion model in the latent space of a variational autoencoder [19]. Subsequently, major improvements were made to boost diffusion model performance [14, 34]. With the release of Stable Diffusion [28] as a powerful generative tool, many works have adapted it to tackle tasks in various domains such as image editing [4, 8, 30] and image segmentation [1, 3, 41]. Recently, diffusion models have been applied to the problem of amodal completion. Notably, [26] and [45] leverage diffusion models, such as Stable Diffusion [28], to train on proposed amodal completion datasets with synthetic occlusion. Additionally, [42] utilize pretrained features from Stable Diffusion [28] for their UNet model aimed at amodal mask completion. Unlike these works, AISDiff is an AIS framework designed to amodally detect and segment instances in images. Furthermore, AISDiff takes advantage of the attention maps in diffusion models to build prior knowledge without denoising to the final output.

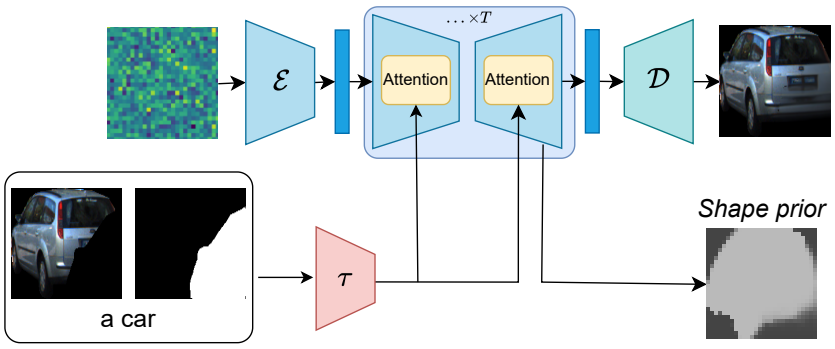

Fig. 2: Overall process of Diffusion Shape Prior Estimation (DiffSP).

3 Method

3.1 Overall AIS Setup

Given an input image $\mathbf{I}$, we follow most of previous AIS settings [7, 17, 37, 40], utilizing a pre-trained backbone network, such as ResNet [11], RegNet [31] to extract spatial visual representation. An object detector such as FCOS [35], or Faster-RCNN [11], can be subsequently adopted to obtain n regions of interest (RoI) predictions and their corresponding visual features $\{\mathbf{F}^i\}_{i=1}^n$. Follow most of previous works [17, 37, 40], the object detector being chose is Faster R-CNN for fair comparison. Here, each RoI is presented by its visual feature $\mathbf{F}^i \in \mathbb{R}^{C_e \times H_r \times W_r}$, where C_e denotes the feature channel size and $H_r \times W_r$ represents the spatial shape of the pooling feature. In this context, given a RoI, AISDiff takes $\mathbf{F}^i$ as input and aims to predict the amodal mask $\mathbf{M}_a^i$. Moreover, in this case, we also denote the visible mask $\mathbf{M}_v^i$, and the occluding mask $\mathbf{M}_o^i$.

3.2 AISDiff

The overall design of AISDiff is depicted in Fig. 1. Initially, we discuss the prediction process for the visible segmentation of the object of interest, along with its categories, incorporating occlusion-awareness through the prediction of occluding masks (Sec. 3.2). Following this, we introduce the DiffSP method in detail (Sec. 3.2). Lastly, we present the Shape Prior Amodal Predictor (Sec. 3.2).

Occlusion-aware Visible Segmentation Given the ROI feature $\mathbf{F}^i$, AISDiff first aims to predict the visible segmentation mask and the category of the object of interest, while simultaneously conducts occlusion-aware ability by predicting the occluding mask, which is the segmentation of occluding elements within the specified ROI. BCNet [17] is utilized as the foundation for the Occlusion-aware Visible Segmentation module. This module consists of two branches: one for

occluding mask prediction and the other for visible mask prediction. Drawing from the methodology outlined in [17], both branches follow a similar design structure, encompassing two main components: feature extraction and mask prediction. The feature extraction segment comprises a sequence of layers, including a 3×3 convolutional layer with a stride of 1, a Graph Convolutional Network (GCN) [20] block, and another 3×3 convolutional layer with a stride of 1. Subsequently, the mask prediction component is constructed with a 2×2 transposed convolutional layer employing a stride of 2, coupled with a 1×1 convolutional layer using a stride of 1.

Furthermore, to enhance occlusion awareness and subsequently improve visible segmentation accuracy, features extracted from the occluding branch are incorporated into the ROI feature $\mathbf{F}^i$ before being fed into the feature extraction section of the visible branch. Simultaneously, features extracted from the visible branch are utilized for object category prediction. This classification step employs a fully connected layer with an output dimension corresponding to the number of categories present in the datasets under consideration. In summary, the final output of this module comprises the visible mask $\mathbf{M}_a^i$, the occluding mask $\mathbf{M}_o^i$, and the object category c^i.

DiffSP The process depicted in Fig. 2 illustrates the Shape Prior Estimation (DiffSP) module. DiffSP builds upon the successes of conditioned diffusion models, such as Stable Diffusion [28] and GLIDE [25], which are pre-trained on comprehensive language-vision datasets like LAION [33]. This pre-training equips the model with the ability to capture intricate visual features, rendering it suitable as prior knowledge for subsequent tasks [26, 44]. Expanding on this foundation, DiffSP utilizes a trained conditioned diffusion model and inputs a ROI image containing only the visible pixels of the object, an occluding mask and a textual description of the object category under consideration, expecting the model to generate the obscured parts. Subsequently, the denoising process iterates T steps to produce the generated image containing the occluded regions. However, instead of relying solely on the final generated pixels, DiffSP capitalizes on the attention mechanism between the conditioning information and the image features.

Specifically, Stable Diffusion [28] is employed as the pre-trained conditioned diffusion model, leveraging its self and cross-attention layers. Specifically, the random Gausian noise is encoded into latent space and then experiences the denoising process over T time steps to generate the inpainting image. In fact, the ROI image containing only the visible pixels of the object of interest, the occluding mask, and the textual description of the object category serve as conditions and are represented as y, which is projected by τ into an intermediate representation $\tau(y)$. At each denoising step t, a UNet architecture with L layers of self and cross-attention transforms z_t into z_{t-1}. Specifically, at layer l and time step t, the cross-attention layer captures the relationship between z_t and the encoded condition $\tau(y)$, reflecting the entire reconstructed shape of the object. This relationship is formalized as follows: at layer l and time step t, the

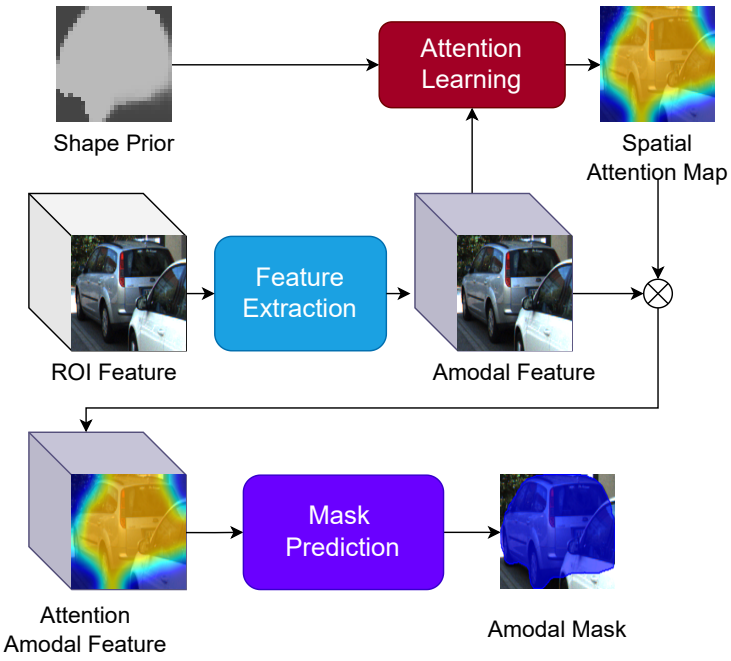

Fig. 3: Overall design of Shape Prior Amodal Predictor.

self-attention map is denoted as $\mathcal{A}_S^{l,t}$, and the cross-attention map is denoted as $\mathcal{A}_C^{l,t}$. Moreover, as demonstrated in [24], the attention map remains relatively stable across time steps. Following the methodology of [24], we average these cross and self-attention maps over layers and time steps, setting $T = 10$. Additionally, as also suggested in [24], although the cross-attention maps $\mathcal{A}_C$ already outline the shape of the reconstructed object, they tend to be coarse-grained and noisy. To refine the precision of object localization, we follow [24], utilizing the self-attention map $\mathcal{A}_S$ to enhance $\mathcal{A}_C$. Consequently, the shape prior is obtained by: $\mathbf{M}_{sp} = (\mathcal{A}_S)^\tau \cdot \mathcal{A}_C$.

Shape Prior Amodal Predictor The design of Shape Prior Amodal Predictor is depicted in Fig. 3. Initially, the feature extraction module utilizes the ROI feature $\mathbf{F}^i$ to generate the amodal feature. This module is constructed using a sequence of 3×3 convolutional layers with a stride of 1. Subsequently, the obtained amodal feature undergoes processing in the attention learning module in conjunction with the shape prior $\mathbf{M}_{sp}$ obtained from DiffSP, aimed at learning the spatial attention map. Specifically, the attention computation involves passing the amodal feature through a sequence of 3×3 convolutional layers with a stride of 1, followed by a sigmoid activation function. This computed attention map is then multiplied with the shape prior $\mathbf{M}_{sp}$. The spatial attention map is further multiplied with the amodal feature to obtain the attention amodal fea-

ture. This feature is then fed into a mask prediction module, which is structured with a 2×2 transposed convolutional layer employing a stride of 2, coupled with a 1×1 convolutional layer using a stride of 1, to derive the amodal mask $\mathbf{M}_a^i$

3.3 Objective Function & Training

Employing AIS protocols, the training adopts a two-stage instance segmentation process similar to Mask R-CNN, facilitating concurrent training of both bounding box and amodal mask prediction heads alongside the object detection framework. In essence, the training procedure optimizes a multi-task loss function $\mathcal{L}$ as follows:

$$\mathcal{L} = \mathcal{L}_{det} + \mathcal{L}_{cls} + \mathcal{L}_v + \mathcal{L}_o + \mathcal{L}_a \tag{1}$$

where $\mathcal{L}_{det}$ is object detection loss, defined similarly to that in Faster R-CNN object detection. The occluding mask loss $\mathcal{L}_o$, the visible mask loss $\mathcal{L}_v$, the amodal mask loss $\mathcal{L}_a$, and the classification loss $\mathcal{L}_{cls}$ are computed using cross entropy loss with the corresponding ground truth.

4 Experiments

4.1 Datasets, Metrics and Implementation Details

Datasets: We benchmark our AISDiff on three AIS datasets, namely KINS [27], COCOA-cls [7], and D2SA [7]. KINS is a large-scale traffic dataset with 95,311 training instances and 92,492 testing instances with 7 categories. COCOA-cls is an AIS dataset that is derived from MSCOCO [22] with 80 categories of 6,763 training instances and 3,799 testing instances. D2SA is an AIS dataset with 60 categories of instances related to supermarket items with 13,066 training instances and 15,654 testing instances.
Metrics: Following existing AIS methods [37, 40], we adopt mean average precision (AP) and mean average recall (AR).
Implementation details: We develop AISDiff utilizing the Detectron2 framework [38]. For the KINS dataset, we employ an SGD optimizer [29] with a learning rate of 0.0025 and a batch size of 1, over $48,000$ iterations. For the D2SA dataset, training is conducted with an SGD optimizer, a learning rate of 0.005, and a batch size of 2 over $70,000$ iterations, while for the COCOA-cls dataset, training involves $10,000$ iterations with a learning rate of 0.0005 and a batch size of 2. All experiments are performed using a Quadro RTX 8000 GPU.

4.2 Baselines

We compare AISDiff with state-of-the-art (SOTA) AIS methods, including ORCNN [7], BCNet [17], VRSP-Net [40], and C2F-Seg [9]. These methods share the same AIS setup described in Section 3.1. It is important to note that recent works such as [26], [42], and [45] focus on amodal completion with the object

of interest already given, and then extract the amodal mask based on this completion. This differs from the AIS framework mentioned, which involves amodal detection and segmentation of instances within images. Therefore, we exclude these methods from our comparison.

Table 1: *Performance comparison on KINS test set* with various backbones. † indicates our reproduced results.

	Backbones& Methods	Venue	Shape Prior	$AP\uparrow$	$AP_{50}\uparrow$	$AP_{75}\uparrow$	$AR\uparrow$
ResNet-50	Mask R-CNN [16]	ICCV17	✗	30.0	54.5	30.1	19.4
	ORCNN [7]	WACV19	✗	30.6	54.2	31.3	19.7
	ASN [27]	CVPR19	✗	32.2	-	-	-
	AISFormer [37]	BMVC22	✗	33.8	57.8	35.3	21.1
	AmodalBlastomere [15]	TMI20	✓	30.3	-	-	-
	VRSP-Net [40]	AAAI21	✓	32.1	55.4	33.3	20.9
	C2F-Seg [9]	ICCV23	✓	**36.5**	58.2	37.0	**22.1**
	AISDiff (Ours)	-	✓	36.3	**58.8**	**37.2**	22.0
ResNet-101	Mask R-CNN [11] †	ICCV17	✗	30.2	54.3	30.4	19.5
	BCNet [17]	CVPR21	✗	28.9	-	-	-
	BCNet [17] †	CVPR21	✗	32.6	57.2	35.4	21.5
	AISFormer [37]	BMVC22	✗	34.6	58.2	36.7	21.9
	C2F-Seg [9]†	ICCV23	✓	**36.9**	58.9	**37.8**	**23.1**
	AISDiff (Ours)	-	✓	**36.9**	**59.6**	37.5	23.0

4.3 Performance Comparison

Quantitative Results KINS. Tab. 1 depicts the comparison between AISDiff and SOTA AIS methods on the KINS dataset. AISDiff demonstrates consistent improvements across various backbones, including ResNet-50 [12] and ResNet-101 [12]. Specifically, when compared to methods utilizing ResNet-50 as the backbone, AISDiff achieves comparable results with the SOTA method (i.e., and C2F-Seg [9] Simlarly when ResNet-101 is utilized as the backbone, our method achieves great performance, compatible with C2F-Seg.
D2SA. Tab. 2 further validates our approach on D2SA dataset. We achieve best results across all metrics. Specifically, we gains 0.13 on AP and 0.1 AR in comparison with the SOTA method, i.e. C2F-Seg [9].
COCOA-cls. Tab. 3 shows our results on COCOA-cls dataset. AISDiff also outperform other methods on all metrics. In fact, it outperforms the second best by 0.16 AP and 0.03 AR.

Qualitative Results Fig. 4 illustrates the qualitative output of AISDiff. The results are arranged from left to right, encompassing: input ROIs, Visible Masks,

Table 2: *Performance comparison on D2SA test set* with ResNet-50 as backbone. † indicates our reproduced results.

Methods	Venue	Shape Prior	$AP \uparrow$	$AP_{50} \uparrow$	$AP_{75} \uparrow$	$AR \uparrow$
Mask R-CNN [11]	ICCV17	✗	63.57	83.85	68.02	65.18
ORCNN [7]	WACV19	✗	64.22	83.55	69.12	65.25
ASN [27] †	CVPR19	✗	63.94	84.35	69.57	65.20
BCNet [17] †	CVPR21	✗	65.97	84.23	72.74	66.90
AISFormer [37]	BMVC22	✗	67.22	84.05	72.87	68.13
VRSP-Net [40]	AAAI21	✓	70.27	85.11	75.81	69.17
C2F-Seg [40]†	ICCV23	✓	70.88	85.07	75.85	69.19
AISDiff (Ours)	-	✓	**71.01**	**85.12**	**76.23**	**69.29**

Table 3: *Performance comparison on COCOA-cls test set*, ResNet-50 as backbone. † indicates our reproduced results.

Methods	Venue	Shape Prior	$AP \uparrow$	$AP_{50} \uparrow$	$AP_{75} \uparrow$	$AR \uparrow$
Mask R-CNN [11]	ICCV17	✗	33.67	56.50	35.78	34.18
ORCNN [7]	WACV19	✗	28.03	53.68	25.36	29.83
ASN [27] †	CVPR19	✗	35.33	58.82	37.10	35.50
BCNet [17] †	CVPR21	✗	35.14	58.84	36.65	35.80
AISFormer [37]	BMVC22	✗	35.77	57.95	38.23	36.71
VRSP-Net [40]	AAAI21	✓	35.41	56.03	38.67	37.11
C2F-Seg [9]†	ICCV23	✓	35.72	58.80	38.73	37.11
AISDiff (Ours)	-	✓	**35.93**	**58.86**	**38.63**	**37.14**

Occluding Masks, Shape Prior, and Amodal Masks. Fig. 5 visualizes the spatial attention map of the Shape Prior Amodal Predictor on ROIs of the image. The attention maps are well-constrained to the object shape. Moreover, we can see that the decoder typically attends to the visible parts of objects that are similar to the occluded regions when predicting the amodal mask. Fig. 6 shows qualitative comparison between AISDiff and the existing SOTA method C2F-Seg [9]. Example are sampled from D2SA and KINS test sets. As can be seen, AISDiff accurately extracts the amodal mask of the occluded object (i.e. the bag of pasta) (left) and efficiently handles the car and the truck (right).

4.4 Ablation studies

Effect of DiffSP diffusion models Table 4 compares the performance of using GLIDE [25] and Stable Diffusion [28] models within the DiffSP framework on the KINS and D2SA datasets. For the KINS dataset, GLIDE achieves an AP of 35.16 and an AR of 21.71, while Stable Diffusion shows improved performance with an AP of 36.36 and an AR of 22.02. Similarly, on the D2SA dataset, GLIDE records an AP of 70.18 and an AR of 69.22, whereas Stable Diffusion further excels with an AP of 71.01 and an AR of 69.29. These results indicate that Stable

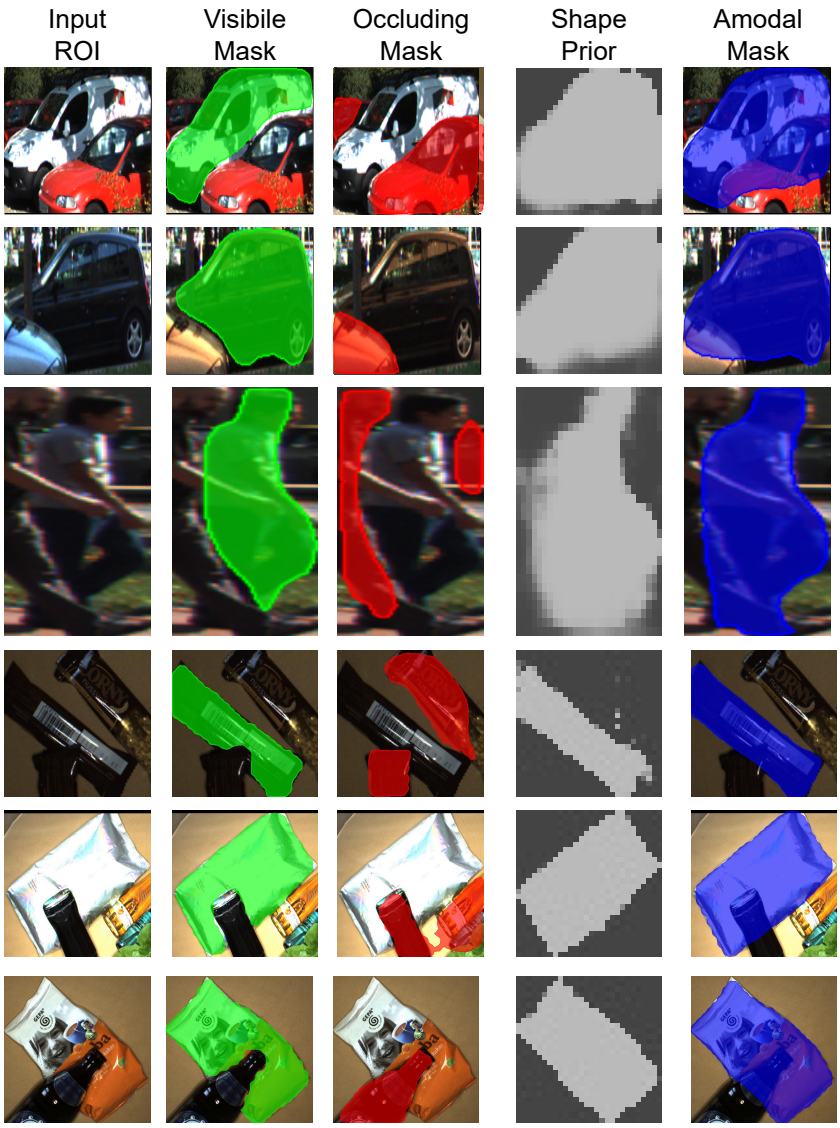

Fig. 4: Qualitative results of AISDiff. Left to right: Input RoI, Visible masks, Occluding masks, Amodal masks. Best viewed in color.

Diffusion consistently outperforms GLIDE, offering better precision and recall in the DiffSP framework.

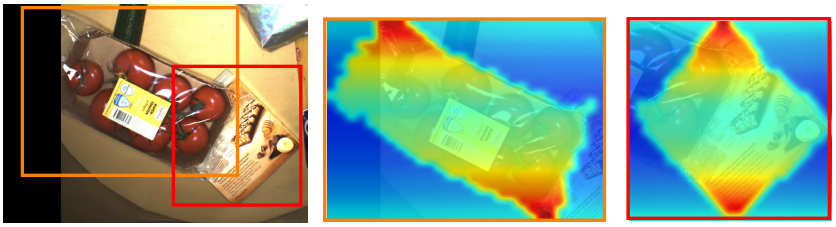

Fig. 5: Spatial attention map of the Shape Prior Amodal Predictor on the each RoI. Best viewed in color.

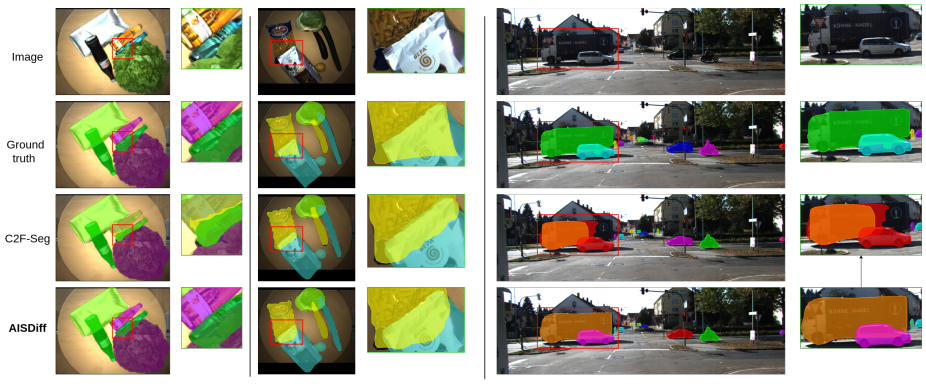

Fig. 6: Qualitative comparison between AISDiff and the SOTA method C2F-Seg [9]. Examples are from D2SA (left) and KINS (right) test set.

Table 4: Effect of diffusion models in DiffSP

Diffusion Model	KINS				D2SA			
	$AP\uparrow$	$AP_{50}\uparrow$	$AP_{75}\uparrow$	$AR\uparrow$	$AP\uparrow$	$AP_{50}\uparrow$	$AP_{75}\uparrow$	$AR\uparrow$
GLIDE [25]	35.16	57.97	37.11	21.71	70.18	85.11	74.96	69.22
Stable Diffusion [28]	36.36	58.84	37.24	22.02	71.01	85.12	76.23	69.29

Effect of denoising time steps Table 5 presents the impact of varying diffusion timesteps (T = 10, 50, 100) on the performance of the DiffSP framework, evaluated using the Average Precision (AP) and Average Recall (AR) metrics on the KINS and D2SA datasets. For the KINS dataset, the AP remains relatively stable across different timesteps, with values of 36.36, 36.33, and 36.43 for $T = 10$, $T = 50$, and $T = 100$, respectively. Similarly, the AR values are consistently around 37.24 for all timesteps. In the D2SA dataset, the AP values are 71.01, 70.99, and 71.11 for $T = 10$, $T = 50$, and $T = 100$, respectively, while the AR values remain steady at approximately 69.29 across all timesteps. Overall, the results indicate that varying the number of diffusion timesteps has minimal impact on both AP and AR metrics for both datasets, suggesting that the DiffSP

framework performs robustly regardless of the diffusion timestep settings. Thus, we opt for $T = 10$ for efficiency.

Table 5: Effect of difusion timesteps

Dataset	$T = 10$		$T = 50$		$T = 100$	
	AP	AR	AP	AR	AP	AR
KINS	36.36	37.24	36.33	37.23	36.43	37.24
D2SA	71.01	69.29	70.99	69.27	71.11	69.29

Effect of object category and occluding mask We conducted an ablation study on the inputs of DiffSP, which utilizes three inputs: visible pixels of the object of interest, the occluding mask, and the object category. The first input is essential for reconstructing the missing parts and cannot be omitted. Therefore, our study focuses on the other two inputs: the object category and the occluding mask, as shown in Table 6. The results demonstrate that including either the object category or the occluding mask improves performance, with the best results achieved using both (the default DiffSP configuration).

Table 6: Effect of object category and occluding mask

Object Category	Occluding Mask	KINS		D2SA	
		AP↑	**AR↑**	**AP↑**	**AR↑**
✗	✗	31.12	20.10	64.87	67.09
✓	✗	35.23	21.83	67.92	68.13
✗	✓	35.81	21.87	69.13	69.21
✓	✓	36.36	22.02	71.01	69.29

5 Conclusion

In conclusion, we propose AISDiff, an AIS mask head with a Diffusion Shape Prior Estimation module. This module, termed DiffSP, leverages pre-trained conditioned diffusion models on extensive datasets to extract nuanced visual features for deriving the shape prior of the object. Furthermore, we present the Shape Prior Amodal Predictor, which utilizes attention-based feature maps from the shape prior to enhance amodal segmentation. Through extensive experimentation across diverse AIS benchmarks, we affirm the efficacy of AISDiff.

Acknowledgments. This work is sponsored by the National Science Foundation (NSF) under Award No OIA-1946391.

References

1. Amit, T., Shaharbany, T., Nachmani, E., Wolf, L.: Segdiff: Image segmentation with diffusion probabilistic models. arXiv preprint arXiv:2112.00390 (2021)
2. Back, S., Lee, J., Kim, T., Noh, S., Kang, R., Bak, S., Lee, K.: Unseen object amodal instance segmentation via hierarchical occlusion modeling. In: ICRA. pp. 5085–5092. IEEE (2022)
3. Baranchuk, D., Rubachev, I., Voynov, A., Khrulkov, V., Babenko, A.: Label-efficient semantic segmentation with diffusion models. arXiv preprint arXiv:2112.03126 (2021)
4. Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18392–18402 (2023)
5. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in neural information processing systems **34**, 8780–8794 (2021)
6. Duncan, J.: Selective attention and the organization of visual information. Journal of experimental psychology: General **113**(4), 501 (1984)
7. Follmann, P., König, R., Härtinger, P., Klostermann, M., Böttger, T.: Learning to see the invisible: End-to-end trainable amodal instance segmentation. In: WACV. pp. 1328–1336. IEEE (2019)
8. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-Or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion. arXiv preprint arXiv:2208.01618 (2022)
9. Gao, J., Qian, X., Wang, Y., Xiao, T., He, T., Zhang, Z., Fu, Y.: Coarse-to-fine amodal segmentation with shape prior. In: ICCV. pp. 1262–1271 (2023)
10. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Advances in neural information processing systems **27** (2014)
11. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: ICCV. pp. 2961–2969 (2017)
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
13. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
14. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)
15. Jang, W.D., Wei, D., Zhang, X., Leahy, B., Yang, H., Tompkin, J., Ben-Yosef, D., Needleman, D., Pfister, H.: Learning vector quantized shape code for amodal blastomere instance segmentation. arXiv preprint arXiv:2012.00985 (2020)
16. Ke, L., Danelljan, M., Li, X., Tai, Y.W., Tang, C.K., Yu, F.: Mask transfiner for high-quality instance segmentation. In: CVPR. pp. 4412–4421 (2022)
17. Ke, L., Tai, Y.W., Tang, C.K.: Deep occlusion-aware instance segmentation with overlapping bilayers. In: CVPR. pp. 4019–4028 (2021)
18. Kellman, P.J., Shipley, T.F.: A theory of visual interpolation in object perception. Cognitive psychology **23**(2), 141–221 (1991)
19. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
20. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907 (2016)

21. Li, K., Malik, J.: Amodal instance segmentation. In: ECCV. pp. 677–693. Springer (2016)
22. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV. pp. 740–755. Springer (2014)
23. Mohan, R., Valada, A.: Amodal panoptic segmentation. In: CVPR. pp. 21023–21032 (2022)
24. Nguyen, Q., Vu, T., Tran, A., Nguyen, K.: Dataset diffusion: Diffusion-based synthetic data generation for pixel-level semantic segmentation. Advances in Neural Information Processing Systems **36** (2024)
25. Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models. arXiv preprint arXiv:2112.10741 (2021)
26. Ozguroglu, E., Liu, R., Surís, D., Chen, D., Dave, A., Tokmakov, P., Vondrick, C.: pix2gestalt: Amodal segmentation by synthesizing wholes. arXiv preprint arXiv:2401.14398 (2024)
27. Qi, L., Jiang, L., Liu, S., Shen, X., Jia, J.: Amodal instance segmentation with kins dataset. In: CVPR. pp. 3014–3023 (2019)
28. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
29. Ruder, S.: An overview of gradient descent optimization algorithms. arXiv preprint arXiv:1609.04747 (2016)
30. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 22500–22510 (2023)
31. Schneider, N., Piewak, F., Stiller, C., Franke, U.: Regnet: Multimodal sensor registration using deep neural networks. In: 2017 IEEE intelligent vehicles symposium (IV). pp. 1803–1810. IEEE (2017)
32. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al.: Laion-5b: An open large-scale dataset for training next generation image-text models. Advances in Neural Information Processing Systems **35**, 25278–25294 (2022)
33. Schuhmann, C., Vencu, R., Beaumont, R., Kaczmarczyk, R., Mullis, C., Katta, A., Coombes, T., Jitsev, J., Komatsuzaki, A.: Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. arXiv preprint arXiv:2111.02114 (2021)
34. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
35. Tian, Z., Shen, C., Chen, H., He, T.: Fcos: Fully convolutional one-stage object detection. In: ICCV. pp. 9627–9636 (2019)
36. Tran, M., Bounsavy, W., Vo, K., Nguyen, A., Nguyen, T., Le, N.: Shapeformer: Shape prior visible-to-amodal transformer-based amodal instance segmentation. arXiv preprint arXiv:2403.11376 (2024)
37. Tran, M., Vo, K., Yamazaki, K., Fernandes, A., Kidd, M., Le, N.: Aisformer: Amodal instance segmentation with transformer. arXiv preprint arXiv:2210.06323 (2022)
38. Wu, Y., Kirillov, A., Massa, F., Lo, W.Y., Girshick, R.: Detectron2. https://github.com/facebookresearch/detectron2 (2019)

39. Xiao, Y., Xu, Y., Zhong, Z., Luo, W., Li, J., Gao, S.: Amodal segmentation based on visible region segmentation and shape prior. arXiv preprint arXiv:2012.05598 (2020)
40. Xiao, Y., Xu, Y., Zhong, Z., Luo, W., Li, J., Gao, S.: Amodal segmentation based on visible region segmentation and shape prior. In: AAAI. vol. 35, pp. 2995–3003 (2021)
41. Xu, J., Liu, S., Vahdat, A., Byeon, W., Wang, X., De Mello, S.: Open-vocabulary panoptic segmentation with text-to-image diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2955–2966 (2023)
42. Xu, K., Zhang, L., Shi, J.: Amodal completion via progressive mixed context diffusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9099–9109 (2024)
43. Yao, J., Hong, Y., Wang, C., Xiao, T., He, T., Locatello, F., Wipf, D.P., Fu, Y., Zhang, Z.: Self-supervised amodal video object segmentation. NeurIPS **35**, 6278–6291 (2022)
44. Zhan, G., Zheng, C., Xie, W., Zisserman, A.: Amodal ground truth and completion in the wild. arXiv preprint arXiv:2312.17247 (2023)
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)
46. Zhu, Y., Tian, Y., Metaxas, D., Dollár, P.: Semantic amodal segmentation. In: CVPR. pp. 1464–1472 (2017)