

LAC-Net: Linear-Fusion Attention-Guided Convolutional Network for Accurate Robotic Grasping Under the Occlusion

Jinyu Zhang^{1*}, Yongchong Gu^{1*}, Jianxiong Gao¹, Haitao Lin¹, Qiang Sun²,
 Xinwei Sun¹, Xiangyang Xue¹ and Yanwei Fu^{1†}

Abstract—This paper addresses the challenge of perceiving complete object shapes through visual perception. While prior studies have demonstrated encouraging outcomes in segmenting the visible parts of objects within a scene, amodal segmentation, in particular, has the potential to allow robots to infer the occluded parts of objects. To this end, this paper introduces a new framework that explores amodal segmentation for robotic grasping in cluttered scenes, thus greatly enhancing robotic grasping abilities. Initially, we use a conventional segmentation algorithm to detect the visible segments of the target object, which provides shape priors for completing the full object mask. Particularly, to explore how to utilize semantic features from RGB images and geometric information from depth images, we propose a Linear-fusion Attention-guided Convolutional Network (LAC-Net). LAC-Net utilizes the linear-fusion strategy to effectively fuse this cross-modal data, and then uses the prior visible mask as attention map to guide the network to focus on target feature locations for further complete mask recovery. Using the amodal mask of the target object provides advantages in selecting more accurate and robust grasp points compared to relying solely on the visible segments. The results on different datasets show that our method achieves state-of-the-art performance. Furthermore, the robot experiments validate the feasibility and robustness of this method in the real world. Our code and demonstrations are available on the project page: <https://jryryzh.github.io/LAC-Net>.

I. INTRODUCTION

Instance segmentation is crucial for enabling robots to effectively grasp objects in predefined environments. While prior research [1], [2], [3] has advanced visible object segment estimation, understanding occluded object parts is essential for successful grasping.

For example, consider a robot cleaning debris from a beach, as illustrated in Fig. 1. Objects are often partially concealed by sand, making it difficult to identify their complete structure. Traditional visual masking algorithms only estimate visible parts, problematic for objects of various shapes and sizes. Relying solely on observable segments limits the information available for grasping. Therefore, acquiring comprehensive object masks enables well-informed grasp point selections, significantly enhancing grasp success rates, as shown in Fig. 1.

Estimating complete object masks is challenging for several reasons. Firstly, perceiving occluded areas, whether for

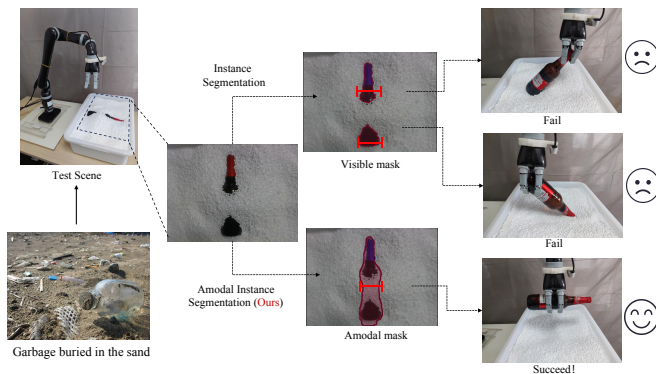


Fig. 1. Comparison between the instance segmentation method and amodal segmentation for robotic grasp. We simulate the scene where the garbage is buried in the sand. Unlike conventional instance segmentation methods that predict solely the visible mask for selecting grasp points, our approach leverages the amodal mask to identify more robust grasp points for the final grasping action.

rigid or non-rigid objects, offers many plausible yet non-unique possibilities. Secondly, real-world scenarios present complex object categories and shapes, making it difficult to pre-learn object shapes accurately. Consequently, estimating the full mask from a single-view image remains an open question.

To address the first challenge, we leverage both RGB and depth information to create robust features, reducing ambiguity through diverse modalities. For the second challenge, we utilize shape priors as initializations to guide amodal mask segmentation. Neural networks are employed to learn the potential distribution of the target object.

Formally, this paper introduces a framework integrating amodal segmentation to enhance robotic grasping in challenging, cluttered environments. Initially, we use off-the-shelf segmentation algorithms like the Segment Anything Model (SAM) [1] or Mask R-CNN [4] to estimate the visible mask of the target object, providing shape priors with the RGB image. Next, our proposed LAC-Net model leverages color and depth information to transform the visible mask into a comprehensive amodal mask. Finally, we determine grasp points using RGB-aligned depth images combined with the full mask of the target object.

In summary, this work makes the following contributions:

- We present a new class-agnostic framework aimed at detecting both visible and amodal masks, enabling accurate and efficient grasping of target objects in cluttered scenes.
- We introduce a novel component, Linear Attention Fusion, designed to fuse multi-modal information with the goal of improving the performance of amodal instance segmen-

*: Equal contribution; †: Corresponding author.

¹Jinyu Zhang, Yongchong Gu, Xinwei Sun and Yanwei Fu are with school of Data Science, Fudan University, China {zhangjy23, yongchonggu22}@m.fudan.edu.cn, yanweifu@fudan.edu.cn; Jianxiong Gao, Haitao Lin and Xiangyang Xue are with Fudan University, China.

²Qiang Sun is with School of statistics and information, Shanghai University of International Business and Economics, China.

tation. Evaluation on the UOAIS-Sim and OSD-amodal benchmarks [5] demonstrates state-of-the-art performance, confirming the effectiveness of our approach.

- Additionally, we validate the real-world applicability of amodal grasping in robotic settings, demonstrating its precision in grasping occluded target objects within complex, cluttered environments.

II. RELATED WORKS

A. Amodal Instance Segmentation

Amodal Instance Segmentation [6] is a more complex task than standard instance segmentation [4], [7], [8], [9], [10], as it involves predicting not only the visible segments of objects but also their occluded shapes. Obtaining accurate masks of target objects is essential for robotic tasks such as grasping [11], [12] and manipulating unseen objects [13], [14]. Previous research [15], [16], [17], [18], [19], [20] has often focused on modeling shape priors based on shape statistics. Notably, VRSP [21] introduced a shape prior module specifically for refining amodal masks. However, adapting these models to open-world scenarios, especially with complex long-tail object category distributions, remains a challenge. AISFormer [22] uses transformers to capture long-range dependencies and employs multi-task training to create a comprehensive segmentation model. SaVos [23] combines spatiotemporal consistency with optical flow for amodal mask prediction, while UOAIS-ent [5] integrates RGB-D images with multi-tasking techniques to improve amodal mask predictions for robotic grasping. Our method, using separate backbones and a linear fusion technique, effectively integrates information from both RGB and depth data, significantly enhancing the performance of amodal mask completion.

B. Target-oriented Grasping in Clutter

Robotic grasping in cluttered environments remains a formidable challenge. Significant strides have advanced the state-of-the-art in cluttered object grasping [24], [25], [26], [27], [28], particularly in generating 6DoF grasp proposals [11], [29], [30], [31]. Target-oriented grasping has also progressed, incorporating target imagery [32], sketches [33], language instructions [34], [35], and guided demonstrations [36] to improve retrieval precision. Methods for target-centric grasping in cluttered contexts have been demonstrated [37], [38], involving navigation through obstacles and pre-grasp operations like strategic pushing [39], [40], [41], [42]. Reinforcement learning has further refined these techniques [43]. Despite advancements, the dynamic and unpredictable nature of cluttered environments—especially those with additional hindrances like snow or sand—poses unique challenges. Traditional obstacle-removal strategies often fail in these conditions. Our approach introduces a novel methodology leveraging amodal segmentation with enhanced generalization capabilities, enabling accurate identification of comprehensive object masks and facilitating the direct grasping of unseen target objects in challenging environments.

III. METHOD

A. Problem Definition

The objective of amodal instance segmentation is to simultaneously segment both the visible and occluded parts of an object. Consider an RGB-D image $I = \{R, U\}$ capturing scene S , where $R \in \mathbb{R}^{W \times H \times 3}$ is the RGB image and $U \in \mathbb{R}^{W \times H}$ represents the depth image. Within this scene, objects are typically comprised of two components: the visible portion and the occluded portion. We represent the mask of the visible part of an object as \mathbf{M}_v and the comprehensive amodal mask, which includes both visible and occluded sections, as \mathbf{M}_a . Our framework utilizes the visible mask \mathbf{M}_v , alongside the RGB-D image I , to accurately infer the amodal mask \mathbf{M}_a , thus achieving a complete representation of the target object.

B. Framework

Our framework consists of two main components: an instance segmentation network and a network for amodal mask completion. The instance segmentation network can be any off-the-shelf model, such as SAM [1] or Mask R-CNN [4], both of which can be trained for optimal results. High-quality instance segmentation provides better visible masks, which in turn leads to higher-quality amodal masks in our model. This relationship is further detailed in the subsequent ablation study.

C. Amodal Mask Completion

Overview. The network comprises two parts: (1) RGB-D fusion backbone, (2) attention-guided completion head. First, we crop the original RGB-D input I using the visible mask to get 256×256 RGB and depth images of the object. These are processed through the RGB-D fusion backbone to generate combined RGB-D features. Next, attention maps are generated using these features and the visible mask to ignore occluders and better capture the object’s details. Finally, the attention map, visible mask, and RGB-D features are merged and fed into the completion head, which outputs the completed amodal mask. This ensures an accurate representation of the object, unaffected by occlusions.

RGB-D Fusion Backbone. RGB images R can provide color and visual texture information, while depth images can U provide spatial distance and three-dimensional structural data. Although both modalities are valuable for predicting amodal masks, efficiently leveraging them together is challenging. To tackle this, we use two separate ResNet-50 networks to independently extract features from RGB and depth inputs, retaining features from layers 1 through 4. We then fuse the corresponding features from each layer using four linear layers, resulting in a new set of features that match the original ResNet feature dimensions. Our experiments show that this approach of using dual independent backbones combined with linear fusion effectively preserves both RGB and depth features, enhancing the model’s generalizability and robustness.

Attention guided completion head Common amodal mask prediction networks often directly concatenate the visible

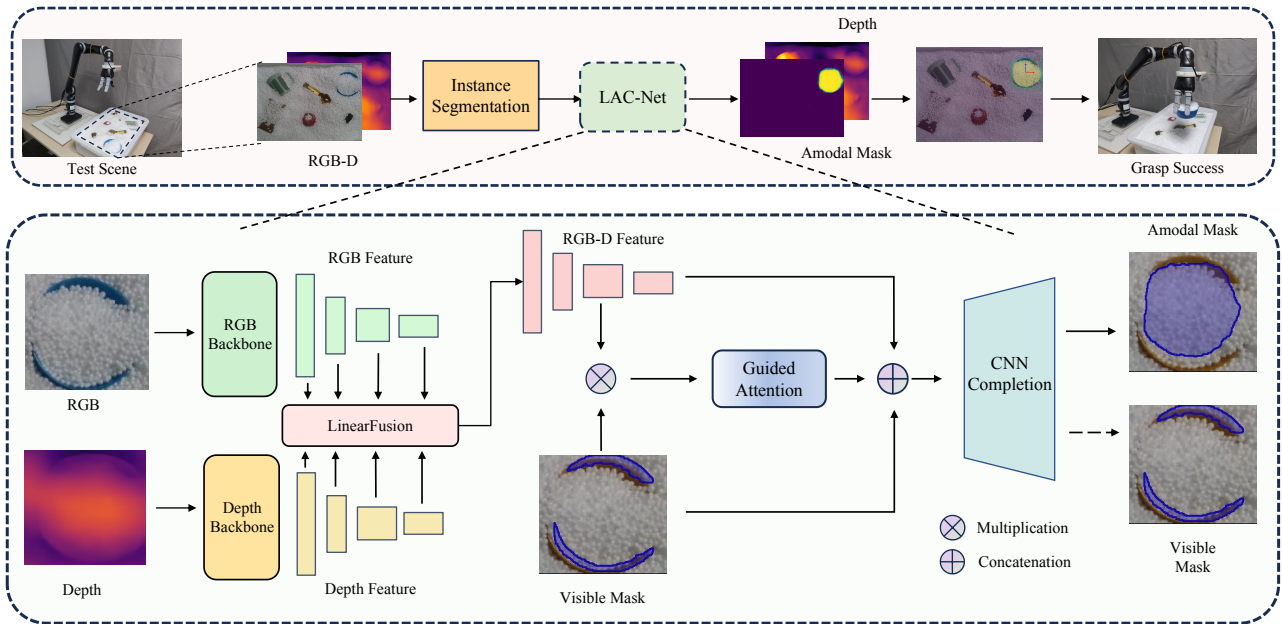


Fig. 2. The workflow of our proposed amodal instance segmentation method for robotic grasping. Given a prompt depicting the target object within the scene, we employ Grounding DINO [44] in conjunction with the Segmentation Anything Model [1] to localize the visible portion of the target object. Subsequently, the visible mask is fed into the coarse-to-fine amodal segmentation module to estimate the complete shape mask of the object. Finally, the robot performs a top-grasp action on the target object using the derived full mask of the target object in conjunction with depth images.

mask with extracted RGB features or RGB-D features, or they perform element-wise multiplication before proceeding with further processing, such as using an upsampling CNN to obtain the amodal mask. However, we propose a more effective way to utilize the visible mask in conjunction with the features. By applying an attention mechanism, we use the visible mask as guidance to calculate attention maps in combination with the RGB-D features. These maps are then concatenated with the visible mask and RGB-D features before being fed into the subsequent completion head. This method refines the feature representation, enabling a more informed and precise completion of the amodal mask. Our completion network is primarily composed of multiple convolutional layers. It progressively interpolates and convolves the composite of previously extracted features, the mask, and the attention map. Ultimately, the network branches into two outputs: one for visible mask and another for amodal mask. The final loss function is simple, as to minimize the binary cross-entropy loss for both the visible and amodal masks simultaneously, $\mathcal{L} := \text{BCE}(\hat{\mathbf{M}}_a, \mathbf{M}_a) + \text{BCE}(\hat{\mathbf{M}}_v, \mathbf{M}_v)$.

D. Grasp Point Generation

Given the ease of capturing RGB-D images with an RGB-D camera, we can conveniently extract the depth pixel corresponding to the target object using the estimated amodal mask. By calculating the center of this mask and utilizing the camera’s intrinsic parameters, we back-project this center into a 3D point. For our grasping scenario, we employ a top-grasp strategy, concentrating solely on the object’s center for robotic grasping.

IV. EXPERIMENT

A. Amodal segmentation

Datasets. We evaluate our methods on UOAI-Sim and OSD-amodal benchmarks. UOAI-Sim [5] features 50,000 RGB-D images from 1,000 cluttered scenes with amodal annotations. Created using photorealistic rendering with BlenderProc, it includes 375 3D textured models of household and industrial items. Objects vary in number (1-40) and are placed on textured surfaces, captured from random camera angles. The dataset is divided into training and test sets in 9:1 and 4:1 ratios, respectively.

The OSD-amodal [45] dataset consists of 111 RGB-D images capturing diverse scenes, such as Boxes, Stacked Boxes, Cylindric Objects, Mixed Objects, and Complex Scenes. It is split into 45 training and 66 validation images, with annotations including amodal masks, visible masks, and occluded masks for all objects in each scene.

Evaluation Metric. For evaluation, we firstly employ the mean Intersection over Union (mean-IoU) to assess the accuracy of the predicted amodal masks, considering both the full mask (denoted as $\text{mIoU}_{\text{full}}$) and the occluded region (mIoU_{occ}). The occluded mIoU is particularly insightful, offering a direct evaluation of the segmentation quality in the occluded portions of the objects, which is a critical aspect in amodal segmentation studies.

On the OSD-amodal dataset, we further measured the Overlap P/R/F, Boundary P/R/F, and $F@.75$ for amodal and invisible masks [46], [47]. Overlap P/R/F provides insights into the total area overlap, while Boundary P/R/F reflects the precision of the boundary contours in the predicted instance masks, following Hungarian matching. The $F@.75$ metric quantifies the percentage of segmented objects with an Overlap F-measure above 0.75. For additional insights,

refer to [46], [47]. The accuracy (ACC_O) and F-measure (F_O) of occlusion classification were also calculated, where $ACC_O = \frac{\delta}{\alpha}$, $F_O = \frac{2F_O R_O}{P_O + R_O}$, with $P_O = \frac{\delta}{\beta}$, and $R_O = \frac{\delta}{\gamma}$. Here, α is the total number of matched instances, β and γ represent the count of occlusion predictions and ground truth instances, respectively, and δ signifies the correctly predicted occlusions.

Implementation Details. Our amodal segmentation model is implemented using PyTorch to ensure reproducibility and facilitate comparative analysis. We adopt two segmentation strategies to conduct a fair comparison. The first strategy utilizes dataset-provided ground truth masks to gauge the model’s optimal performance. The second employs the visible bounding boxes and masks detected by UOAIS-net (akin to Mask R-CNN) for a consistent baseline comparison. For preparing RGB-D images and masks, we crop them using bounding boxes of visible areas, expanded by a factor of 2, and standardize these inputs to a resolution of 256×256 . We augment the masks with morphology dilation, erosion, and Gaussian blur. We use the AdamW optimizer with a learning rate of $3e-4$ and batch size of 32, training for 100K iterations on the UOAIS-Sim training set. Performance is evaluated on both the UOAIS-Sim test set and the OSD-amodal dataset.

Performance on Amodal Segmentation. We compared our method with state-of-the-art instance segmentation approaches, all trained on the UOAIS-Sim dataset’s RGB-D data. For our method, we trained separately using depth, RGB, and RGB-D data, and benchmarked all methods on the OSD-amodal dataset. ASN and UOAIS-net directly output occlusion status, for other methods and ours, occlusion is determined based on the ratio of V to A (occlusion $O = 1$ if $V/A < 0.95$). Experimental results show that our method outperforms all baselines across all metrics. Notably, even our models trained with solely RGB or depth data achieved commendable results. When trained with RGB-D data, our method surpassed baselines in nearly all metrics, demonstrating high accuracy without a dedicated occlusion prediction branch.

Ablation study. We analyzed the impact of two key factors: visible mask quality and RGB-D feature fusion methods.

a) Visible mask: We compared results using visible masks detected by UOAIS-net (similar to Mask R-CNN) against those using ground truth visible masks. As shown in Table I, using ground truth masks significantly improved IoU on both UOAIS-Sim and OSD-amodal datasets, highlighting that higher quality visible masks enhance amodal mask completion effectiveness.

b) RGB-D Fusion: We tested three RGB-D feature fusion methods: (1) merging RGB and depth into a 6-channel image for ResNet-50 input, (2) fusing features from ResNet-50 for RGB and depth using 1×1 convolutions, and (3) our final model’s approach, fusing RGB and depth features via linear layers. Linear fusion outperformed the other methods, demonstrating robust training results and superior generalization on the unseen OSD-amodal dataset.

TABLE I
OUR LAC-NET V.S. UOAIS-NET BY USING VARIOUS INPUTS. {*}
DEMOTES USING GROUND-TRUTH VISIBLE MASK AS INPUTS.

Method	In	UOAIS-sim		OSD-amodal	
		IoU full	IoU occ	IoU full	IoU occ
UOAIS	RGB-D	85.51	53.04	81.73	45.31
LAC-Net	RGB	92.74	57.63	81.32	54.34
LAC-Net	D	92.41	50.05	82.10	48.97
LAC-Net	RGB-D	93.21	62.16	83.81	58.03
LAC-Net*	RGB	96.57	73.35	91.41	70.10
LAC-Net*	D	94.23	71.04	94.42	76.98
LAC-Net*	RGB-D	97.01	77.03	95.14	81.30



Fig. 3. Test object collection used in our robotic experiment. We choose 15 instances with different shape for testing.

B. Real-world amodal grasping

Hardware Experimental Setup. In our robotic experiment, we used a KINOVA Gen2 robot with a 6-DoF curved wrist and a KG-3 gripper. An arm-mounted RealSense D435 camera provides RGB-D streams of the scene. The system runs on a desktop equipped with an NVIDIA GTX2080 GPU and leverages pre-trained Grounding DINO and SAM models.

To closely simulate real-world scenarios such as snowy terrains, deserts, and marshlands, we used 5-6mm foam spheres as scene fillers to obscure the objects to be grasped. These foam spheres are highly pliable, allowing us to arrange them manually for intricate occlusion effects, thereby enhancing our ability to create complex occlusion scenarios.

Principle of Human-robot Dialogue. In real-world robotic experiments, experimenters engage in dialogue with the system. The system encapsulates the experimenter’s preferences into target instructions in the format of “[color] [category]”, such as “blue bowl”. These instructions are subsequently conveyed to the visual grounding system.

Test Object Collections. To validate the effectiveness of our method, we deliberately selected fifteen diverse object types with varying shape distributions for grasp, as illus-

TABLE II
PERFORMANCE USING DIFFERENT CROSS-MODAL FUSION METHODS.

Method	UOAIS-sim		OSD-amodal	
	Iou full	Iou occ	Iou full	Iou occ
RGB-D 6 channel	95.71	73.72	94.96	80.74
RGB-D 1x1conv	96.75	75.17	94.95	80.24
RGB-D Linear Fusion	97.00	77.01	95.05	81.22

TABLE III

OUR MODEL PERFORMANCES ON OSD-AMODAL. {} DENOTES THAT THEY ARE PREDICTED PARALLELLY. → REFERS THE HIERARCHY IN PREDICTION HEADS. OV: OVERLAP F , BO: BOUNDARY F

Method	Input	Hierarchy Order	Amodal Mask (A)			Invisible Mask (IV)			Occlusion (O)	
			OV	BO	F@.75	OV	BO	F@.75	FO	ACCO
Amodal MRCNN	RGB-D	{B, V, A}	82.4	66.6	82.5	50.9	28.4	41.9	74.5	81.9
ORCNN	RGB-D	{B, V, A} → IV	83.1	67.2	84.1	49.2	25.8	33.6	75.5	83.1
ASN	RGB-D	{B, O} → {V, A}	80.7	67.0	84.5	42.0	20.7	38.0	59.1	65.7
UOAIIS-Net	RGB-D	B → V → A → O	82.1	68.7	83.7	55.3	32.3	49.2	82.1	90.9
LAC-Net(Ours)	RGB	V → A → O	87.6	73.4	89.7	55.3	30.3	38.9	79.7	85.2
LAC-Net(Ours)	Depth	V → A → O	85.6	60.2	88.0	51.4	19.7	37.6	67.0	71.6
LAC-Net(Ours)	RGB-D	V → A → O	89.0	79.0	89.4	64.9	40.6	60.1	85.0	90.9

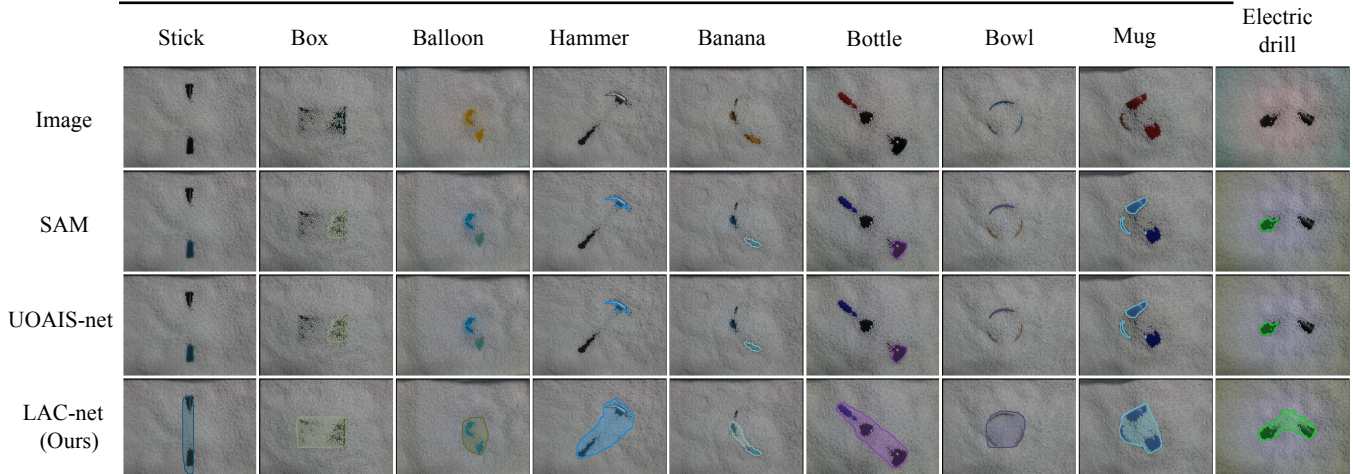


Fig. 4. Qualitative Results of amodal segmentation in real-world Scenes: We present the original RGB image, the visible mask estimated by SAM, and the amodal mask results from both UOAIIS-net and our method, respectively.

trated in Fig. 3. These grasping targets encompass fourteen distinct object classes: stick, box, balloon, hammer, cup, banana, bottle, bowl, mug, telephone model, knife, mouse, screwdriver and electric drill. Among these, eleven classes were employed in the single-object scene, while fourteen classes were utilized in the multi-object scene grasping experiment. It is worth noting that all fourteen classes were employed for qualitative visualization purposes, enhancing the comprehensiveness of our evaluation.

Evaluation Metrics. The object’s full mask is divided into three equal sections along its axis. The middle section, comprising one-third of the full mask area, is designated as Region A. The remaining two sections, occupying two-thirds of the full mask area, are collectively referred to as Region B. Based on this definition, the grasping results are categorized into three types. *Good*: The robot successfully grasps the object stably, with the center of the gripper positioned within Region A. *Poor*: The robot successfully grasps the object stably, but the center of the gripper is located within Region B. *Fail*: The robot fails to grasp the target object, either missing it entirely or being unable to maintain a grip.

Grasping Experiments in Single-object Scene. In cluttered scenes covered with foam, we conducted grasping experiments on six object categories, comparing our results with UOAIIS-net under identical conditions. Detailed quantitative results in Table IV highlight our system’s performance. Our method achieved 90 successful “Good” grasps, significantly surpassing UOAIIS-net’s 23, a margin of 67. Including both “Good” and “Poor” grasps, our method totaled 96 successful

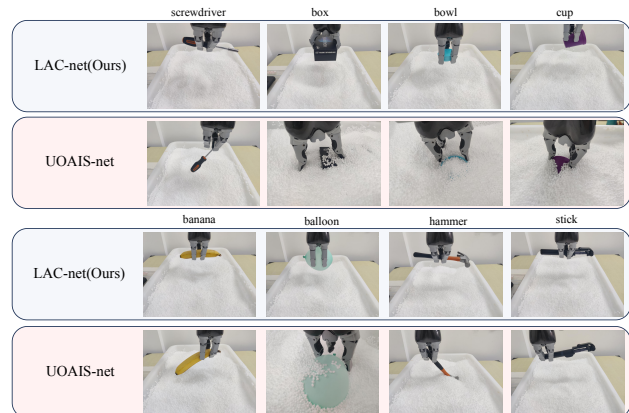


Fig. 5. Comparison results of the robotic grasping between the baseline method and our method in *single-object* scenes. In the comparison of robotic grasping methods, the baseline UOAIIS-net method often tends to grasp the edge portions of the target object, primarily due to the limitations in recovering the amodal mask. In contrast, our approach excels in grasping the object’s center, as our method can estimate an approximate full mask of the target object, thereby ensuring a higher rate of overall success.

grasps, outperforming UOAIIS-net’s 58 by 38 grasps.

The failures in UOAIIS-net can be attributed to several factors related to object characteristics. For instance, grasping sticks and boxes often fails due to imprecision, while heavy objects like filled water bottles and hammers cause tilting and dropping when grasped at one end. Additionally, curved objects like balloons and bowls are prone to slippage, particularly near the edges, highlighting the system’s struggle with varied shapes and weights.

Grasping Experiment in Multi-object Scene Multi-object



Fig. 6. Examples of the robotic grasping between the baseline method and our method in *multiple-object* scenes. (a) Successful grasps. (b) Failed grasps. (c) Instances of visual grounding failures.

TABLE IV

QUANTITATIVE RESULTS ON REAL-WORLD SINGLE-OBJECT SCENE. G, P, F DENOTES GOOD, POOR, AND FAIL GRASPING.

class	UOAI-net			G	Ours	
	G	P	F		P	F
Stick	6/20	14/20	0/20	17/20	3/20	0/0
Box	8/20	11/20	1/20	20/20	0/20	0/20
Bottle	3/20	7/20	10/20	18/20	1/20	1/20
Hammer	1/20	3/20	16/20	11/20	2/20	7/20
Balloon	2/20	0/20	18/20	8/20	0/20	12/20
Bowl	3/20	0/20	17/20	16/20	0/20	4/20
All	23/120	35/120	62/120	90/120	6/120	24/120

scenes present greater challenges than single-object scenes due to potential detection interference and the risk of collisions between the gripper and nearby objects. Table V shows our method achieving 58 successful “good” grasps compared to UOAI-net’s 18, a margin of 40. Overall, we achieve 63 total successful grasps, surpassing UOAI-net’s 43 by 20.

Fig. 6 illustrates varied outcomes of our method. In (a), we show examples of successful object grasping by UOAI-net and LAC-net. In (b), we explore cases where close-proximity objects obstruct the robot’s gripper, leading to failure, despite our method’s accurate amodal mask generation. Our method significantly outperforms UOAI-net by generating central grasp points, reducing peripheral grasp failures.

In (c), we showed some visual grounding errors, such as misidentifying “yellow bottle” when the system focuses on a red mug, highlighting challenges in complex environments. Overall, our method advances robotic grasping by accurately identifying central grasp points, despite issues in object proximity and visual grounding.

Qualitative Results As illustrated in Fig. 4, our amodal segmentation technique consistently yields higher quality masks compared to those produced by UOAI-net. We find that UOAI-net often mistakes the visible mask for the entire object, leading to a propensity for grasping at the extremities of objects. In contrast, our method generates more comprehensive and precise amodal masks. These improved masks better guide the robotic arm to grasp at the object’s center, resulting in a higher success rate and improved grasping quality.

TABLE V

QUANTITATIVE RESULTS ON REAL-WORLD MULTIPLE-OBJECT SCENE. G, P, F DENOTES GOOD, POOR AND FAIL GRASPING.

class	UOAI-net			G	Ours	
	G	P	F		P	F
Stick	4/20	7/20	9/20	9/20	1/20	10/20
Box	3/20	7/20	10/20	15/20	0/20	5/20
Bottle	4/20	3/20	13/20	14/20	0/20	6/20
Mug	5/20	8/20	7/20	13/20	4/20	3/20
Bowl	2/20	0/20	18/20	7/20	0/20	13/20
All	18/120	25/120	57/120	58/120	5/120	37/120

In Fig. 5, we demonstrate the efficacy of our system with examples of single-object grasping in cluttered settings. These examples highlight that successful grasping in such scenarios hinges on accurately targeting the obscured centers of objects, rather than just their visible peripheries.

V. CONCLUSIONS

In this work, we introduce a novel framework for amodal instance segmentation. By extracting and linearly fusing RGB and depth features, we obtain robust RGB-D characteristics. Utilizing the visible mask, we calculate guided attention, which, through a convolutional completion network, yields the completed amodal mask. Our experiments demonstrate that our method achieves state-of-the-art (SOTA) results on UOAI-Sim and OSD-amodal benchmarks. Furthermore, we have applied our method to a physical Kinova robot for conducting robotic grasping tasks on objects submerged in white foam. The experimental outcomes underscore our method’s exceptional real-world amodal grasping capabilities, showcasing its particular effectiveness in applications like debris removal by cleaning robots in sandy contexts. Future work aims to delve further into exploiting our method’s potential for beach cleaning scenarios.

ACKNOWLEDGEMENTS

This work was supported in part by Shanghai Platform for Neuromorphic and AI Chip under Grant 17DZ2260900 (NeuHelium). The computations in this research were performed using the CFFF platform of Fudan University.

REFERENCES

- [1] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, “Segment anything,” *arXiv:2304.02643*, 2023.
- [2] C. Xie, Y. Xiang, A. Mousavian, and D. Fox, “Unseen object instance segmentation for robotic environments,” *IEEE Transactions on Robotics*, vol. 37, no. 5, pp. 1343–1359, 2021.
- [3] Y. Xiang, C. Xie, A. Mousavian, and D. Fox, “Learning rgb-d feature embeddings for unseen object instance segmentation,” in *Conference on Robot Learning*. PMLR, 2021, pp. 461–470.
- [4] K. He, G. Gkioxari, P. Dollár, and R. B. Girshick, “Mask R-CNN,” *CoRR*, vol. abs/1703.06870, 2017. [Online]. Available: <http://arxiv.org/abs/1703.06870>
- [5] S. Back, J. Lee, T. Kim, S. Noh, R. Kang, S. Bak, and K. Lee, “Unseen object amodal instance segmentation via hierarchical occlusion modeling,” in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 5085–5092.
- [6] Y. Zhu, Y. Tian, D. Metaxas, and P. Dollár, “Semantic amodal segmentation,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1464–1472.
- [7] R. Girshick, “Fast r-cnn,” in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

- [8] P. Guo, W. Zhang, X. Li, and W. Zhang, "Adaptive online mutual learning bi-decoders for video object segmentation," *IEEE Transactions on Image Processing*, vol. 31, pp. 7063–7077, 2022.
- [9] W. Li, J. Fan, P. Guo, L. Hong, and W. Zhang, "Hfvos: History-future integrated dynamic memory for video object segmentation," *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.
- [10] P. Guo, T. Huang, P. He, X. Liu, T. Xiao, Z. Chen, and W. Zhang, "Openvis: Open-vocabulary video instance segmentation," *arXiv preprint arXiv:2305.16835*, 2023.
- [11] M. Sundermeyer, A. Mousavian, R. Triebel, and D. Fox, "Contact-graspnet: Efficient 6-dof grasp generation in cluttered scenes," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 13 438–13 444.
- [12] H. Lin, Z. Liu, C. Cheang, Y. Fu, G. Guo, and X. Xue, "Sar-net: Shape alignment and recovery network for category-level 6d object pose and size estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 6707–6717.
- [13] A. Murali, A. Mousavian, C. Eppner, C. Paxton, and D. Fox, "6-dof grasping for target-driven object manipulation in clutter," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 6232–6238.
- [14] H. Lin, Y. Fu, and X. Xue, "Pourit!: Weakly-supervised liquid perception from a single image for visual closed-loop robotic pouring," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 241–251.
- [15] Z. Zhang, A. Chen, L. Xie, J. Yu, and S. Gao, "Learning semantics-aware distance map with semantics layering network for amodal instance segmentation," in *Proceedings of the 27th ACM International Conference on Multimedia*, 2019, pp. 2124–2132.
- [16] L. Ke, Y.-W. Tai, and C.-K. Tang, "Occlusion-aware video object inpainting," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 14 468–14 478.
- [17] J. Yang, Z. Ren, M. Xu, X. Chen, D. J. Crandall, D. Parikh, and D. Batra, "Embodied amodal recognition: Learning to move to perceive objects," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 2040–2050.
- [18] H. Ling, D. Acuna, K. Kreis, S. W. Kim, and S. Fidler, "Variational amodal object completion," *Advances in Neural Information Processing Systems*, vol. 33, pp. 16 246–16 257, 2020.
- [19] J. Gao, X. Qian, Y. Wang, T. Xiao, T. He, Z. Zhang, and Y. Fu, "Coarse-to-fine amodal segmentation with shape prior," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 1262–1271.
- [20] J. Gao, X. Qian, L. Liang, J. Han, and Y. Fu, "Hyper-transformer for amodal completion," *arXiv preprint arXiv:2405.19949*, 2024.
- [21] Y. Xiao, Y. Xu, Z. Zhong, W. Luo, J. Li, and S. Gao, "Amodal segmentation based on visible region segmentation and shape prior," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 4, 2021, pp. 2995–3003.
- [22] M. Tran, K. Vo, K. Yamazaki, A. Fernandes, M. Kidd, and N. Le, "Aisformer: Amodal instance segmentation with transformer," *arXiv preprint arXiv:2210.06323*, 2022.
- [23] J. Yao, Y. Hong, C. Wang, T. Xiao, T. He, F. Locatello, D. Wipf, Y. Fu, and Z. Zhang, "Self-supervised amodal video object segmentation," *arXiv preprint arXiv:2210.12733*, 2022.
- [24] L. Pinto and A. Gupta, "Supersizing self-supervision: Learning to grasp from 50k tries and 700 robot hours," in *2016 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2016, pp. 3406–3413.
- [25] J. Mahler and K. Goldberg, "Learning deep policies for robot bin picking by simulating robust grasping sequences," in *Conference on robot learning*. PMLR, 2017, pp. 515–524.
- [26] J. Mahler, J. Liang, S. Niyaz, M. Laskey, R. Doan, X. Liu, J. A. Ojea, and K. Goldberg, "Dex-net 2.0: Deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics," *arXiv preprint arXiv:1703.09312*, 2017.
- [27] D. Kalashnikov, A. Irpan, P. Pastor, J. Ibarz, A. Herzog, E. Jang, D. Quillen, E. Holly, M. Kalakrishnan, V. Vanhoucke, *et al.*, "Scalable deep reinforcement learning for vision-based robotic manipulation," in *Conference on Robot Learning*. PMLR, 2018, pp. 651–673.
- [28] T. Wang, Y. Li, H. Lin, X. Xue, and Y. Fu, "Wall-e: Embodied robotic waiter load lifting with large language model," *arXiv preprint arXiv:2308.15962*, 2023.
- [29] H.-S. Fang, C. Wang, M. Gou, and C. Lu, "Graspnet-1billion: A large-scale benchmark for general object grasping," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 444–11 453.
- [30] C. Wang, H.-S. Fang, M. Gou, H. Fang, J. Gao, and C. Lu, "Graspness discovery in clutters for fast and accurate grasp detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 15 964–15 973.
- [31] D. Son, "Grasping as inference: Reactive grasping in heavily cluttered environment," *IEEE Robotics and Automation Letters*, vol. 7, no. 3, pp. 7193–7200, 2022.
- [32] M. Sun and Y. Gao, "Gater: Learning grasp-action-target embeddings and relations for task-specific grasping," *IEEE Robotics and Automation Letters*, vol. 7, no. 1, pp. 618–625, 2021.
- [33] H. Lin, C. Cheang, Y. Fu, and X. Xue, "I know what you draw: Learning grasp detection conditioned on a few freehand sketches," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 8417–8423.
- [34] C. Cheang, H. Lin, Y. Fu, and X. Xue, "Learning 6-dof object poses to grasp category-level objects by language instructions," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 8476–8482.
- [35] Q. Sun, H. Lin, Y. Fu, Y. Fu, and X. Xue, "Language guided robotic grasping with fine-grained instructions," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2023, pp. 1319–1326.
- [36] M. Laskey, J. Lee, C. Chuck, D. Gealy, W. Hsieh, F. T. Pokorny, A. D. Dragan, and K. Goldberg, "Robot grasping in clutter: Using a hierarchy of supervisors for learning from demonstrations," in *2016 IEEE international conference on automation science and engineering (CASE)*. IEEE, 2016, pp. 827–834.
- [37] E. Jang, S. Vijayanarasimhan, P. Pastor, J. Ibarz, and S. Levine, "End-to-end learning of semantic grasping," *arXiv preprint arXiv:1707.01932*, 2017.
- [38] A. Zeng, S. Song, K.-T. Yu, E. Donlon, F. R. Hogan, M. Bauza, D. Ma, O. Taylor, M. Liu, E. Romo, *et al.*, "Robotic pick-and-place of novel objects in clutter with multi-affordance grasping and cross-domain image matching," *The International Journal of Robotics Research*, vol. 41, no. 7, pp. 690–705, 2022.
- [39] Y. Yang, H. Liang, and C. Choi, "A deep learning approach to grasping the invisible," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 2232–2239, 2020.
- [40] M. Kiatos and S. Malassiotis, "Robust object grasping in clutter via singulation," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 1596–1600.
- [41] A. Kurenkov, J. Taglic, R. Kulkarni, M. Dominguez-Kuhne, A. Garg, R. Martín-Martín, and S. Savarese, "Visuomotor mechanical search: Learning to retrieve target objects in clutter," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 8408–8414.
- [42] K. Xu, H. Yu, Q. Lai, Y. Wang, and R. Xiong, "Efficient learning of goal-oriented push-grasping synergy in clutter," *IEEE Robotics and Automation Letters*, vol. 6, no. 4, pp. 6337–6344, 2021.
- [43] K. Xu, S. Zhao, Z. Zhou, Z. Li, H. Pi, Y. Zhu, Y. Wang, and R. Xiong, "A joint modeling of vision-language-action for target-oriented grasping in clutter," *arXiv preprint arXiv:2302.12610*, 2023.
- [44] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, C. Li, J. Yang, H. Su, J. Zhu, *et al.*, "Grounding dino: Marrying dino with grounded pre-training for open-set object detection," *arXiv preprint arXiv:2303.05499*, 2023.
- [45] A. Richtsfeld, T. Mörwald, J. Prankl, M. Zillich, and M. Vincze, "Segmentation of unknown objects in indoor environments," in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2012, pp. 4791–4796.
- [46] C. Xie, X. Yu, A. Mousavian, and D. Fox, "The best of both modes: Separately leveraging rgb and depth for unseen object instance segmentation," *Cornell University - arXiv, Cornell University - arXiv*, Jul 2019.
- [47] A. Dave, P. Tokmakov, and D. Ramanan, "Towards segmenting anything that moves," *arXiv: Computer Vision and Pattern Recognition, arXiv: Computer Vision and Pattern Recognition*, Feb 2019.