A Unified Framework for Occlusion Reasoning and De-occlusion in Open Scene Understanding

Haoqian Song[1,2,3], Long Cheng[1,3], Xuchen Liu[2], Junjie Wen[2], Jinqiang Cui[2]
[1]*Institute of Automation, Chinese Academy of Sciences, Beijing, China*
[2]*Pengcheng Laboratory, Shenzhen, China*
[3]*School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China*
Email: songhaoqian2022@ia.ac.cn, long.cheng@ia.ac.cn, liuxch@pcl.ac.cn, wenjj@pcl.ac.cn, cuijq@pcl.ac.cn

***Abstract*—The semantic understanding of occlusion relationships is essential for applications such as autonomous driving, robot navigation, and object grasping, yet existing scene understanding methods often lack this capability. Recognizing this limitation, this paper proposes a unified framework of occlusion reasoning and de-occlusion for open scene understanding. This framework consists of three modules: 1) Visible-region-based scene understanding: the framework leverages RAM-BLIP-Grounded-SAM to automatically generate segmentation masks and semantic information for the visible regions of objects. 2) Occlusion reasoning: an improved partial completion network (PCNet) with an occlusion completion prior module is proposed to ensure robust zero-shot generalization in open scenes through a self-supervised manner, and achieves occlusion ordering with semantic information. 3) Automatic de-occlusion: analyzing occlusion ordering results, the framework enables automatically amodal segmentation and completion based on pix2gestalt, eliminating the need for manual prompts. The results on COCOA dataset show that improved PCNet achieves an occlusion ordering accuracy of 91.534%, representing a 4.357% improvement. Moreover, the unified framework for occlusion reasoning and de-occlusion demonstrates outstanding performance in open scenes.**

***Index Terms*—Occlusion Reasoning, Open Scene Understanding, Amodal Segmentation and Completion.**

I. Introduction

Scene understanding is a fundamental aspect of visual tasks and a cornerstone of machine perception [1]. It encompasses various tasks, including object classification, detection, tracking, scene parsing, depth estimation, semantic segmentation, instance segmentation [2]. While recent advancements in vision technologies have significantly improved scene understanding, they primarily focus on the visible regions of objects. This limitation results in incomplete scene understanding due to the lack of processing for invisible regions [3]–[5]. Achieving comprehensive scene understanding requires not only analyzing statistical information, such as object categories and counts, but also grasping the relationships between objects [6]. These relationships often involve occlusion, which can enrich the modal information and improve its overall completeness.

Perceiving and understanding such occlusion relationships pose significant challenges for machines. Achieving complete scene understanding requires determining the occlusion ordering, identifying the positions of occluded regions, and reconstructing occluded pixels. These tasks correspond to three key objectives: occlusion ordering reasoning, which involves reasoning the occlusion ordering between objects; amodal instance segmentation, which focuses on completing the masks of occluded objects; and amodal completion, which reconstructs the invisible region pixels of occluded objects [7].

Current occlusion reasoning primarily relies on reasoning from visible pixels, which necessitates accurate visible masks of both the occludees and occluders, as well as complete masks of the occludees for supervised learning [8]. These complete masks are typically either manually annotated or automatically synthesized, often resulting in certain differences. Moreover, accurate visible masks are usually provided artificially and are limited to specific constrained scenes. Models trained under such data tend to fail when applied to open scenes [9].

Through self-supervised learning, the partial completion network (PCNet) simulates the de-occlusion process, reducing reliance on the complete masks of occludees [10]. With its unique self-supervised and process-learning architecture, PCNet consistently generates partially completed masks of occludees in open scenes with high positional accuracy [11]. In addition, it outperforms other models in occlusion ordering reasoning leveraging the incremental information from occlusion-completed masks, without relying on the complete masks of occludees. Despite these advantages, PCNet encounters both over-completion and under-completion problems, and its focus on modeling only the visible regions inherently constrains its generative capacity [12]. Moreover, PCNet lacks semantic information in occlusion ordering, limiting its ability to achieve semantic understanding of occlusion relationships.

For amodal segmentation and completion, the methods based on diffusion models exhibit strong ability to generate realistic images, implicitly learn internal representations of objects, and reconstruct complete objects with reasonable shapes, geometry, and pixel details [12]–[14]. Recent studies have introduced different diffusion models to de-occlusion tasks by enhancing pre-trained large-scale diffusion models. For instance, pix2gestalt proposed a conditional diffusion model [13], progressive occlusion-aware amodal completion pipeline proposed the mixed context diffusion model [14]. These methods achieve rich amodal completions with accurate

This work was supported by Pengcheng Laboratory under the Cross-Disciplinary Frontier Project No. 2025QYA016. *(Corresponding author: Long Cheng, Xuchen Liu.)*

masks, effectively generalize to diverse zero-shot settings, and outperform state-of-the-art methods. However, these generative methods typically lack occlusion reasoning and rely on manual prompts for occluded objects, resulting in limited autonomy and reduced applicability in automated pipelines.

To sum up, current approaches for occlusion reasoning and de-occlusion tasks rely on pre-prepared and accurate masks of the visible regions of occludees, and lack integration with scene semantics. These issues hinder the realization of automatic occlusion reasoning and de-occlusion for online scene semantic understanding. In addition, the occlusion reasoning models struggle to maintain accuracy in open scene. To address these limitations, the main contributions of this paper are summarized as follows:

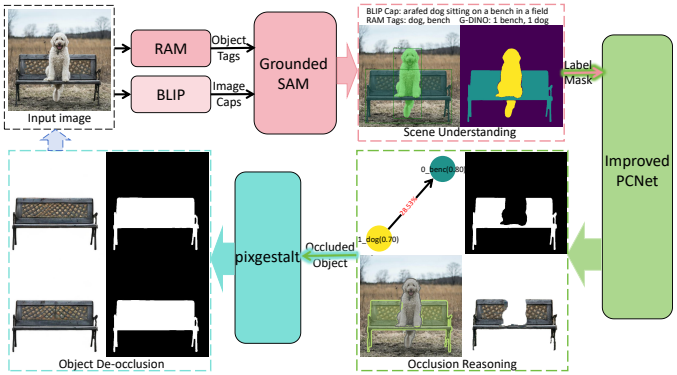


Fig. 1: Unified framework of occlusion reasoning and de-occlusion for open scene understanding.

- We propose a unified framework of occlusion reasoning and de-occlusion for open scene understanding, as illustrated in Fig. 1. It integrates RAM, BLIP, and Grounded SAM to extract visible masks and semantic information, uses improved PCNet to reason occlusion ordering with semantic information, and integrates pix2gestalt to achieve object de-occlusion.
- To tackle the decline in model performance when reasoning occlusion ordering in open scenes, this paper proposes an improved PCNet by incorporating occlusion completion feature derived from stable diffusion. The improved model retains the generalization capabilities of the self-supervised framework, while significantly boosting occlusion reasoning performance in open scenes.
- To eliminate the reliance on manual prompts for occludees and enhance semantic interaction, our framework identifies occludees by analyzing the occlusion ordering aligned with semantic labels, and integrates pix2gestalt to enable end-to-end object de-occlusion in open scenes.

II. Related Work

A. Open Scene Understanding

Scene understanding has become a critical component of advanced robotic systems, particularly in applications like autonomous driving, robotic navigation and grasping. Recently, various approaches to scene understanding have been rapidly developed and improved [15]. Various universal models have emerged, such as recognize anything model (RAM) for object recognition [16], grounding DINO for object detection [17], segment anything model (SAM) for instance segmentation [18], bootstrapping language-image pre-training model (BLIP) for vision-language understanding [19]. However, these methods focus on object itself understanding and do not reason the relationship between objects [20].

The understanding of object relationships primarily focuses on the attribute and action relationships between objects, such as semantic reasoning for attribution [21], [22], and causal reasoning for action [23]. However, these methods operate on visible regions of objects and do not provide understanding of occlusion relationship and the occluded regions. Currently, researches on occlusion are mainly concerned with the amodal segmentation and completion of occluded objects [24]. While these methods address occlusion, they lack a occlusion relationships understanding with the semantic information and exhibit distinct characteristics. The following subsections will provide an introduction to these de-occlusion methods, focusing on their unique characteristics and learning approaches.

B. Supervised De-occlusion Methods

De-occlusion methods with supervised learning are trained with ground truth in a fully supervised manner [25], [26]. The ground truth is typically derived from either synthetic images with complete object masks [27] or real-world images with manually annotated amodal masks [28], both of which have inherent limitations. Synthetic images introduce a domain gap between the test data (real-world) and the training data (synthetic). Meanwhile, manual annotations of real-world images rely on subjective inferences and conjectures by individual annotators to delineate the occluded object boundaries [29].

Supervised de-occlusion methods are inherently restricted by the distribution of training dataset, which confines them to the object categories and scene types present in dataset [30]. As a result, they exhibit limited generalization to open scenes. Additionally, these methods are heavily influenced by the accuracy of manual annotations [31]. Without precise annotations of amodal masks and occlusion ordering as ground truth, supervised de-occlusion methods are unsuitable for occlusion reasoning and amodal segmentation [32]. These limitations are even more pronounced for amodal completion, as the occluded pixels cannot be manually annotated [33]. Consequently, the supervised methods primarily applied to amodal segmentation.

C. Self-supervised De-occlusion Methods

To address the above limitations, the self-supervised de-occlusion methods enable the model to learn the de-occlusion process from the occluded object itself [34]. For example, PCNet employs progressive partial de-occlusion, decomposes the scene into intact and isolated objects and partially completes the objects. This is achieved by intentionally erasing a part of the occluded object and training the model to recover the

unerased object [11]. By using the unerased objects as pseudo-labels, PCNet eliminats the need for annotation of amodal masks, and facilitates the occlusion reasoning [11]. Building on this, ASBU introduced boundary uncertainty estimation to address the problem of difficulty in determining the boundary of occlusion completion [10].

PCNet is effective in amodal segmentation and completion on out-of-distribution data, but struggles with over-completion issues for occluded objects, limiting its performance in amodal segmentation and completion tasks in open scenes [12]. Furthermore, this method relies solely on visible pixels to complete the occluded regions, restricting its generation capability and effectiveness in amodal completion. Therefore, self-supervised de-occlusion methods can work well for open scenes and excel in reasoning about occlusion ordering, orientation and position. However, they are limited by weak generation capability, which result in poor amodal completion performance in open scenes.

D. Diffusion-based De-occlusion Methods

To address the limitations in generation capability and generalization, pre-trained large-scale diffusion models have been integrated into de-occlusion methods [35]. For instance, pix2gestalt introduces a conditional diffusion model built upon a latent diffusion framework, incorporating diffusion representation and category grouping prior to de-occlude occluded objects. This approach delivers detailed image completions with precise masks and demonstrates strong generalization to zero-shot settings [13]. Similarly, a mixed context diffusion model based on stable diffusion has been proposed for progressive occlusion-aware completion. By progressively expanding the contextual conditions, the model can infer the complete object shape from the visible region step by step, avoiding the instability caused by generating it all at once [14].

Diffusion-based de-occlusion methods demonstrate strong generation capability, enabling amodal completion that produces more natural and realistic pixel reconstructions in open scenes. However, these methods do not explicitly encode object boundaries relative to the background, face issues with occlusion relationship reasoning [36]. For instance, pix2gestalt assumes that objects closer to the camera are intact [13]; while the mixed context diffusion model requires depth information based on the visible mask boundaries to identify the occluded object [14]. Additionally, these methods reconstruct occluded pixels by sampling from visible pixels, but fail to fully exploit the visible pixels, often resulting in result in amodal completions that are misaligned with the context of scene.

III. Methods

The automatic occlusion reasoning and de-occlusion framework proposed in this paper mainly includes: semantic perception and segmentation of visible regions for scene understanding, reasoning occlusion ordering for occlusion relationship, and amodal segmentation and completion for object de-occlusion. Therefore, this section provides a detailed introduction to the three modules.

A. Visible-Region-Based Scene Understanding

To achieve scene understanding and automatic extraction of the visible mask and semantic information for the objects, we combine Grounded SAM with the vision-language models to obtain more comprehensive semantic labels and visible mask. Among them, RAM is an image tagging model that excels at accurately recognize common objects, while Tag2Text is a strong model for producing image captions and tags. RAM provides more comprehensive object categories compared to Tag2Text [37]. Therefore, we adopt RAM and integrate it with Grounded SAM to extract object semantic information and segmentation masks for the visible regions of objects.

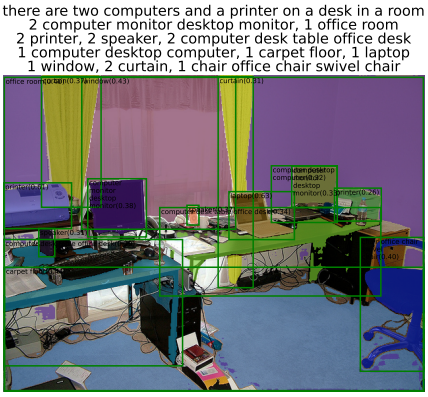


Fig. 2: The scene understanding results of the RAM-BLIP-Grounded-SAM for the visible regions of objects.

To achieve a better understanding of the scene and generate more natural and comprehensive captions, we use BLIP to generate captions, as it is an efficient vision-language model for generating scene descriptions and can produce more fluent and descriptive captions compared to other models [37]. Additionally, we incorporate object statistics into the captions, as shown in Fig. 2. The first line above the image presents the scene description, and the subsequent four lines enumerate the object labels and their associated statistics. Fig. 2 includes a wide range of object categories, and the detection and segmentation instances are comprehensive, particularly for the objects such as houses and the basin in the lower left corner.

B. Occlusion Reasoning with Improved PCNet

1) Improved PCNet: In order to achieve automatic occlusion reasoning, we perform occlusion ordering reasoning based on PCNet framework [11]. PCNet simulates the occlusion process, takes a step back by further trimming down visible mask M randomly to obtain M_{-1} s.t. $M_{-1} \subset M$, and then trains model p_θ via $M_{-1} \xrightarrow{p_\theta} M$. Specifically, the method employs two strategies, as shown in Fig. 3. The partial completion strategy guides PCNet to recover the regions of object A that are occluded by object B, promoting accurate reconstruction of the erasing regions. Object A and B are sampled from the dataset D. The over-completion strategy prevents PCNet from extending object A beyond its authentic geometry, effectively suppressing hallucinated or unnecessary completions. Because

PCNet learns amodal segmentation and completion exclusively from the visible (modal) region of the occluded object, it is inherently unable to model the invisible regions or acquire completion features necessary for de-occlusion.

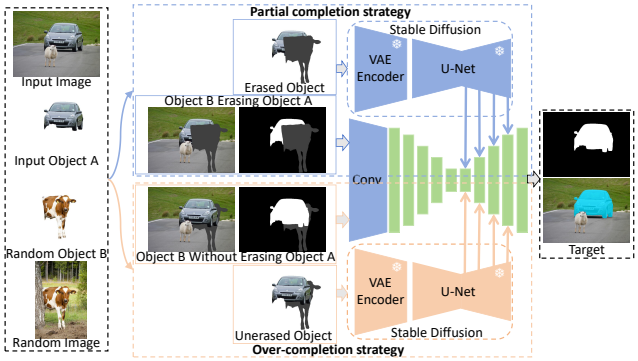


Fig. 3: Network structure of improved PCNet. The object A is provided as input, while object B is randomly sampled from the dataset and positioned arbitrarily. The training alternates between two strategies: (1) Partial completion strategy: encouraging PCNet to complete the regions of object A erased by object B. (2) Over-completion strategy: preventing PCNet from over-completing object A beyond its actual structure.

To enhance the reconstruction of occlusion features in the invisible regions of occluded objects, we leverage stable diffusion to capture and generate the intrinsic feature representation of the invisible regions about occluded objects and integrate it into the decoder of PCNet, as illustrated in Fig. 3. The features extracted from stable diffusion provide holistic shape semantics and category-agnostic priors, which guide PCNet to distinguish occlusion-induced missing regions from non-existent regions, thereby stabilizing completion increment estimation. Stable diffusion has a strong object prior, this approach enhances the model's ability to learn object prior distribution from the visible regions, thereby improving stability and convergence of occlusion ordering reasoning in open scenes.

Unlike point-wise concatenation, multi-level features from stable diffusion are progressively injected into the PCNet decoder in a coarse-to-fine manner, providing hierarchical semantic priors. First, four independent 1×1 convolutions project these features into a unified embedding space, yielding learnable multi-granularity semantic representations while suppressing task-irrelevant components. Second, the projected features are aligned via bilinear interpolation to ensure spatial consistency, thereby avoiding misalignment issues caused by implicit resizing operations. Finally, at each upsampling stage of PCNet decoder, the aligned features are concatenated with the skip-connection features from PCNet encoder to enable hierarchical correspondence, followed by two 3×3 convolutions for feature interaction and re-encoding, thereby enhancing local structural representations and reconstruction capability.

The partial completion loss $\mathcal{L}_{part}$ is defined as:

$$\begin{aligned} \mathcal{L}_{part} &= \frac{1}{N} \sum_{A,B \in D} L\left(M_{part}, M_A\right) \\ M_{part} &= P_\theta\left(M_{A\setminus B}, S(M_{A\setminus B}); M_B, I\setminus M_B\right) \end{aligned} \tag{1}$$

where $P_\theta(*)$ is the improved PCNet model with learnable parameters θ, and L is the binary cross-entropy loss, N is the number of training pairs, M_{part} is the predicted mask of object A under the partial completion strategy. M_A and M_B are the modal masks of objects A and B, respectively. $M_{A\setminus B}$ denotes the remaining region of object A after being partially erased by object B. I denotes the input image, $I\setminus M_B$ represents the image where the pixels within the region of mask M_B are removed. $S(*)$ is the occlusion completion feature extracted from a stable diffusion model, and $S(M_{A\setminus B})$ represents the occlusion completion feature within the region defined by $M_{A\setminus B}$.

The over-completion loss $\mathcal{L}_{over}$ is defined as:

$$\begin{aligned} \mathcal{L}_{over} &= \frac{1}{N} \sum_{A,B \in D} L\left(M_{over}, M_A\right) \\ M_{over} &= P_\theta\left(M_A, S(M_A); M_{B\setminus A}, I\setminus M_{B\setminus A}\right) \end{aligned} \tag{2}$$

where M_{over} denotes the predicted mask of object A under the over-completion strategy, $S(M_A)$ is the occlusion completion feature of the region defined by M_A, $M_{B\setminus A}$ denotes the remaining region of object B after being partially erased by object A, and $I\setminus M_{B\setminus A}$ represents the image where the pixels within the region of mask $M_{B\setminus A}$ are removed.

The final loss function is formulated as:

$$\mathcal{L} = \lambda \mathcal{L}_{part} + (1-\lambda)\mathcal{L}_{over} \tag{3}$$

where $\lambda \sim \text{Bernoulli}(\gamma)$, and γ is the probability of selecting the partial completion strategy, which is set to 0.8 following PCNet. The random switching between the two strategies enables the network to learn the occlusion relationships between neighboring instances based on their shapes and boundaries, thereby determining whether completion is required.

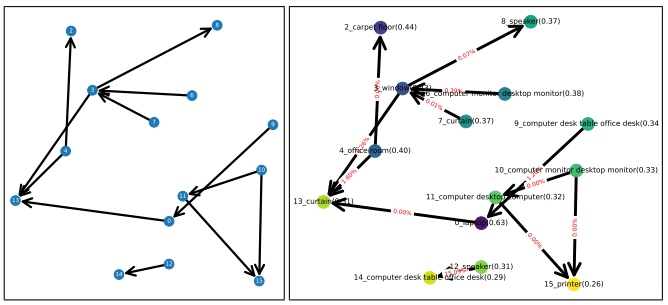
2) Semantic-aware Occlusion Reasoning: The occlusion ordering is determined by computing the increment in mutual completion for all pairs of neighboring objects. Specifically, for each pair, we compare the completion-induced mask increments: the object with the larger increment is identified as the occludee, while the one with the smaller increment is treated as the occluder. However, in complex open scenes, the visible mask of an object is often fragmented into multiple disjoint regions rather than forming a single contiguous area. This fragmentation makes it difficult to reliably determine object adjacency. To address this issue, take the largest connected region among the object's multiple disjoint mask components as the representative region for determining whether two objects are adjacent. This strategy improves both the accuracy and the robustness of occlusion reasoning in open scenes.

To achieve an semantic understanding of occlusion relationship, we enhance the occlusion ordering with semantic labels and occlusion ratio. First, the semantic labels and

segmentation masks obtained from Grounded SAM are aligned with the objects in occlusion ordering. Second, as shown in Fig. 4, object labels and occlusion ratios are integrated into the occlusion ordering directed acyclic graph, with segmentation instance colors used to distinguish different instances. For each pair of neighboring objects, the occlusion ratio is defined as:

$$R = \frac{|M_\Delta|}{|M_\Delta| + |M_{vis}|} \tag{4}$$

where $|M_\Delta|$ and $|M_{vis}|$ denote the set of pixels in predicted occlusion region (increment) and visible region of the object, respectively. Finally, a comprehensive analysis for occlusion completion is conducted based on the number and proportion of occluded objects within the directed acyclic graph. The graph of occlusion ordering is constructed based on all pairwise occlusion relationships between neighboring objects.



(a) (b)

Fig. 4: Occlusion ordering results with semantics information. (a) Occlusion ordering results without incorporating semantic information. (b) Occlusion ordering results with semantic information and occlusion ratios.

The occlusion ordering with semantic information is visualized as a directed acyclic graph. As shown in Fig. 4, the nodes represent the objects and are labeled with their semantic information. The first number denotes the object serial number, used to distinguish different objects, while the middle text indicates the object category. The number in brackets represents the probability of the object belonging to that category. Additionally, the arrow indicate the occlusion direction of neighboring objects, with the number in the middle of the arrow showing the proportion of occluded regions.

C. Automatic Object De-occlusion

Since current object de-occlusion methods require the visible mask of the occluded object or manual point prompts, they cannot automatically select occluded objects and perform semantic interaction. We can achieve the perception and semantic interaction of selecting occluded objects by analyzing occlusion ordering with semantic information, eliminating the reliance on visible mask or manual prompts to achieve automatic amodal segmentation and completion for object de-occlusion. In detail, the strategy of selecting occluded objects is performed by analyzing the number of occluders that occluding the occluded object and the estimated occlusion ratio. In order to achieve automatic amodal segmentation and completion for the occluded objects, we perform the amodal segmentation and completion based on pix2gestalt model without any fine-tuning. The pix2gestalt delivers detailed amodal completion with precise masks and demonstrates strong generalization to zero-shot settings [13]. In addition, its incorporating diffusion representation and category grouping prior to complete occluded objects, and can achieve realistic object de-occlusion effects in open scenes.

IV. Experiment

A. Experiment Setup

To evaluate the effectiveness of improved PCNet, we conduct experiments on the COCOA dataset [38], a subset of COCO2014 specifically designed for object de-occlusion. COCOA provides annotation for pair-wise occlusion ordering, visible (modal), and complete (amodal) masks. The dataset contains 2,140 categories grouped into two super-classes: things (e.g., human, cat, dog, etc.) and stuff (e.g., sea, grass, sky, etc.). It includes 2,500 training images with 22,163 annotated instances, 1,323 validation images with 12,753 instances, and 1,250 test images. Scenes in COCOA are typically densely populated, with frequent and complex inter-object occlusions.

To assess the generalizability and effectiveness of the unified framework in open scene understanding, we further evaluate the unified framework on publicly available dataset PDD [39]. PDD is a real-world post-disaster image dataset collected in ruin environments, containing images of fire-fighters, disaster victims, rescue personnel, buildings, debris, vehicles, and other relevant objects. The scenes in PDD exhibit diverse occlusion phenomena, ranging from single-object occlusion to complex multi-object interactions. Typical occlusion cases include mutual occlusion among people and occlusions caused by collapsed structures or building debris.

Model training is performed on a server with an Intel Xeon Silver 4210R CPU@2.40GHz, two NVIDIA RTX A40 GPUs. The input image resolution on COCOA is set to 512×512. The model is trained for 56,000 iterations using the SGD optimizer with an initial learning rate of 0.001, a momentum of 0.9, and a weight decay of 0.0001. These parameters used are consistent with those in [11]. We extract four levels of features from the upsampling blocks of the stable diffusion models, with dimensions of 1280, 1280, 640, and 320, respectively. Feature extraction from the stable diffusion model is performed at a fixed diffusion timestep of 181. The effectiveness of occlusion ordering reasoning is quantified using the ordering reasoning accuracy over all instance pairs ($Acc_{allpair}$) and over occluded instance pairs ($Acc_{occpair}$), following the metrics in [11].

B. Occlusion Ordering Results

Table I presents the occlusion ordering results on the COCOA validation set. The original PCNet attains approximately 95% $Acc_{allpair}$ and 85% $Acc_{occpair}$, achieving its best performance at 56,000 iterations, where $Acc_{allpair}$ reaches 95.811% and $Acc_{occpair}$ reaches 86.189%. In contrast, the improved

PCNet consistently outperforms the baseline, yielding around 96% accuracy for all instance pairs and approximately 89% for occluded pairs. At 56,000 iterations, it achieves 96.725% $Acc_{allpair}$ and 90.193% $Acc_{occpair}$, representing improvements of 0.914% and 4.004%, respectively. Moreover, the improved PCNet achieves optimal results at 40,000 iterations, achieving 96.728% $Acc_{allpair}$ and 90.209% $Acc_{occpair}$.

TABLE I: Occlusion ordering results on COCOA validation set.

Model	PCNet		Improved PCNet	
Iterations	$Acc_{allpair}$	$Acc_{occpair}$	$Acc_{allpair}$	$Acc_{occpair}$
10,000	94.643%	80.892%	96.386%	88.65%
20,000	95.756%	85.907%	96.375%	88.603%
30,000	95.686%	85.673%	95.699%	84.453%
40,000	95.749%	85.907%	**96.728%**	**90.209%**
50,000	95.746%	85.876%	96.687%	89.964%
56,000	**95.811%**	**86.189%**	96.725%	90.193%

Table II presents the occlusion ordering results on the COCOA test set. PCNet yields roughly 96% $Acc_{allpair}$ and 87% $Acc_{occpair}$, achieving its best results at 56,000 iterations (96.890% and 87.177%) The improved PCNet consistently outperforms the baseline, achieving around 97% accuracy for all pairs and about 91% for occluded pairs. At 56,000 iterations, the improved PCNet obtains 97.756% $Acc_{allpair}$ and 91.256% $Acc_{occpair}$. Its best results is achieved at 40,000 iterations, with 97.787% $Acc_{allpair}$ and 91.422% $Acc_{occpair}$.

TABLE II: Occlusion ordering results on COCOA test set.

Model	PCNet		Improved PCNet	
Iterations	$Acc_{allpair}$	$Acc_{occpair}$	$Acc_{allpair}$	$Acc_{occpair}$
10,000	95.934%	82.582%	97.361%	89.347%
20,000	96.82%	86.869%	97.414%	89.584%
30,000	96.842%	87.023%	96.716%	85.072%
40,000	96.846%	86.952%	**97.787%**	**91.422%**
50,000	96.875%	87.076%	97.734%	91.149%
56,000	**96.890%**	**87.177%**	97.756%	91.256%

The results in Tables I and II show that the the improved PCNet yields an accuracy gain of roughly 1% for all instance pairs and about 4% for occluded pairs, with these improvements remaining stable across training iterations. Beyond higher accuracy, the improved PCNet exhibits faster convergence, approaching optimal performance at 10,000 iterations and reaching it at 40,000 iterations. In contrast, the original PCNet approaches its optimum at 20,000 iterations and converges fully only at 56,000 iterations.

Table III shows the occlusion ordering results on the COCOA dataset across different methods for occluded pairs ($Acc_{occpair}$). The "evaluated" refers to results obtained using the officially released checkpoints, "reproduced" indicates results obtained by re-training the models with the official code, "reported" corresponds to the results stated in the original papers. The improved PCNet achieves 90.193% on the validation set and 91.256% on the test set at 56,000 iterations. Its best results is observed at 42,000 iterations, achieving 90.360% and 91.534% on the validation and test sets, respectively. These results exceed those of ASBU (90.33% and 90.77% on validation and test sets), and the best validation accuracy of 90.360% is comparable to that of the supervised method HORI (90.90%). As shown in Table III, the improved PCNet consistently outperforms the original PCNet on both validation and test sets, with more substantial gains observed on test sets.

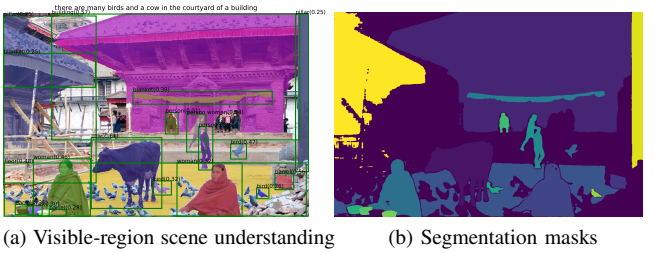
TABLE III: Comparison of occlusion ordering results with state-of-the-art methods on the COCOA dataset.

Learning Paradigm	Models	COCOA	
		Validation	Test
Self-supervised	PCNet(evaluated from [11])	87.112%	88.286%
	PCNet(reproduced by our)	86.189%	87.177%
	PCNet(reproduced by [10])	85.75%	86.73%
	ASBU(reported by [10])	**90.33%**	**90.77%**
	ASBU(reproduced by [40])	88.00%	—
	Improved PCNet(ours 56,000)	**90.193%**	**91.256%**
	Improved PCNet(ours best)	**90.360%**	**91.534%**
Supervised	CSDNet(reported by [7])	84.7%	—
	OrderNet^{M+I}(reported by [38]	88.3%	—
	HORI(reported by [40])	**90.90%**	—

C. Occlusion Reasoning in Open Scene

To demonstrate the model's cross-domain transferability of occlusion reasoning in open scene understanding, we analyze its semantically enriched occlusion ordering results on real post-disaster images from PDD, despite the model never being trained on this dataset. Fig. 5 illustrates the occlusion ordering effect of PCNet and improved PCNet models. Fig. 5(a) presents the visible-region scene understanding effect for a relatively complex scene, and Fig. 5(b) shows the corresponding instance segmentation masks. Fig. 5(c) and 5(d) display the occlusion ordering effect with semantic information produced by PCNet and the improved PCNet, respectively. The arrow direction denotes the occlusion relationship, with each arrow annotated by the estimated occlusion ratio. The two nodes connected by an arrow represent the occluder and occludee. Node colors are aligned with the instance colors in Fig. 5(b), and each node additionally includes semantic labels and detection confidence scores, while numerical identifiers placed at the beginning of each node label are used to differentiate instances with the same object category.

From the perspective of occlusion ordering accuracy, the improved PCNet yields more reliable reasoning results than the original PCNet, as shown in Fig. 5(c) and 5(d). The differences between the two models can be observed in several occlusion relationships: the "9_person woman(0.34)" is occluded by the "8_person(0.36)" in the center of the image, the "18_pillar(0.25)" is occluded by the "16_blanket(0.25)" in the upper-left region of the image, the "4_floor(0.48)" is occluded by "10_bird(0.32)", the "4_floor(0.48)" is occluded "15_bird(0.26)" in the lower part of the image. The improved PCNet correctly reasoned these occlusion relationship, whereas PCNet produces incorrect occlusion relationships. Based on the above analysis of the occlusion relationships reasoning performance in complex open scenes, the PCNet model achieved correct occlusion relationships in half of the evaluated cases, while the improved PCNet achieved correct occlusion relationships in all cases. These results demonstrate



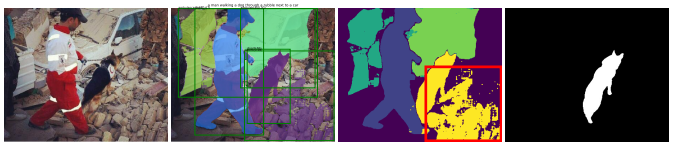
(a) Visible-region scene understanding (b) Segmentation masks

(c) PCNet occlusion ordering (d) ImPCNet occlusion ordering

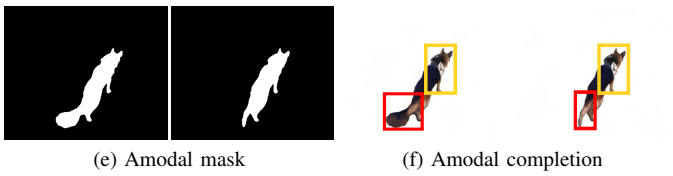
Fig. 5: Comparison of occlusion ordering effect with semantic information between PCNet and improved PCNet.

that the improved PCNet provides more accurate and robust occlusion reasoning for open scene understanding.

D. Object De-occlusion in Open Scene



(a) Input (b) Understanding (c) Masks (d) Visible Mask



(e) Amodal mask (f) Amodal completion

Fig. 6: The effect of object de-occlusion in open scene.

To verify the effectiveness and generalization of the unified framework for open scene understanding, we conduct an object de-occlusion analysis on the real post-disaster images from PDD, despite the fact that the model has never been trained on the PDD dataset. Fig. 6 visualizes the object de-occlusion pipeline and illustrates the effect of object de-occlusion. Fig. 6(b) shows the open scene understanding results produced by RAM-BLIP-Grounded-SAM, while Fig. 6(c) shows the instance segmentation masks, where the yellow debris mask covered the dog's mask. As highlighted by the red box in Fig. 6(c), the visible mask of an object in a complex environment is discontinuous and composed of multiple disjoint regions. In this case, our framework can also effectively perform occlusion reasoning and object de-occlusion. Fig. 6(d) provides the visible mask of the occluded object, Fig. 6(e) shows the amodal (complete) mask of the occluded object, and Fig. 6(f) displays the amodal completion results with reconstructed pixels. The framework can accurately detect occluded areas, such as the dog's tail (red box), and effectively recover the occluded region of the dog's tail, achieving accurate mask completion and pixel reconstruction with different tail states. In addition, other unoccluded regions remain unaffected, so the framework exhibit good de-occlusion performance. However, the visible regions in the amodal completion result appear blurred (yellow box), and this issue is particularly evident in the facial features of the person highlighted in Fig. 7(c). This is because pix2gestalt performs object de-occlusion by sampling the visible region and directly generating the entire object, without fully exploiting the detailed information contained in the visible region.

(a) Understanding (b) Masks (c) Amodal completion

Fig. 7: Amodal completion effect in person de-occlusion.

The proposed framework enables end-to-end occlusion reasoning and de-occlusion for open scene understanding, however, several limitations remain.

- Grounded SAM requires enhancements in the object detection, instance segmentation performance, and generalization capability in open scenes. For instance, in Fig. 7(b), the occluded person in the background is neither detected nor segmented.
- The framework is heavily dependent on the detection and segmentation models, necessitating accurate visible masks for occlusion reasoning and de-occlusion.
- While the pix2gestalt amodal completion model, exhibits strong generation capability and achieves remarkable de-occlusion effect, it struggles in scenarios where the reconstructed object morphology deviates from the scenes. For example, the completed legs of the person highlighted by the cyan box in Fig. 7(c) remain significantly misaligned with the ideal state depicted in Fig. 7(a).

V. Conclusion

This paper proposes a novel framework of occlusion reasoning and de-occlusion for open scene semantic understanding, incorporating an improved PCNet based on stable diffusion for occlusion reasoning in open scenes. The improved PCNet is evaluated on the COCOA dataset, achieving the highest accuracy of 91.534% in a self-supervised manner and yielding an approximate 4.357% improvement in occlusion ordering. Furthermore, the effectiveness of our framework is validated

on the images from a proprietary real-world disaster-scene dataset PDD. The framework effectively captures occlusion semantics and delivers notable object de-occlusion effect in open scenes. These advancements pave the way for applications in autonomous driving, augmented reality, and robotics by enabling more reliable occlusion reasoning and object de-occlusion, thereby enhancing scene understanding and interaction in dynamic and complex environments.

REFERENCES

[1] J. Zhang, Y. Rao, X. Huang, G. Li, X. Zhou, and D. Zeng, "Frequency-aware multi-modal fine-tuning for few-shot open-set remote sensing scene classification," *IEEE Transactions on Multimedia*, vol. 26, pp. 7823–7837, 2024.

[2] J. Shi, M. Wang, H. Duan, and S. Guan, "Language embedded 3D gaussians for open-vocabulary scene understanding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 5333–5343.

[3] P. He, L. Jiao, F. Liu, X. Liu, R. Shang, and S. Wang, "Cross-domain scene unsupervised learning segmentation with dynamic subdomains," *IEEE Transactions on Multimedia*, vol. 26, pp. 6770–6784, 2024.

[4] R. Ding, J. Yang, C. Xue, W. Zhang, S. Bai, and X. Qi, "Lowis3D: language-driven open-world instance-level 3D scene understanding," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 12, pp. 8517–8533, 2024.

[5] B. Jia, Y. Chen, H. Yu, Y. Wang, X. Niu *et al.*, "SceneVerse: Scaling 3D vision-language learning for grounded scene understanding," in *European Conference on Computer Vision*, 2024, pp. 289–310.

[6] J. Yang, R. Ding, W. Deng, Z. Wang, and X. Qi, "RegionPLC: Regional point-language contrastive learning for open-world 3D scene understanding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 19 823–19 832.

[7] C. Zheng, D.-S. Dao, G. Song, T.-J. Cham, and J. Cai, "Visiting the invisible: Layer-by-layer completed scene decomposition," *International Journal of Computer Vision*, vol. 129, no. 12, pp. 3195–3215, 2021.

[8] J. Gao, X. Qian, Y. Wang, T. Xiao, T. He *et al.*, "Coarse-to-fine amodal segmentation with shape prior," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 1262–1271.

[9] Y. Sun, A. Kortylewski, and A. Yuille, "Amodal segmentation through out-of-task and out-of-distribution generalization with a bayesian model," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 1215–1224.

[10] K. Nguyen and S. Todorovic, "A weakly supervised amodal segmenter with boundary uncertainty estimation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 7396–7405.

[11] X. Zhan, X. Pan, B. Dai, Z. Liu, D. Lin, and C. C. Loy, "Self-supervised scene de-occlusion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3784–3792.

[12] G. Zhan, C. Zheng, W. Xie, and A. Zisserman, "Amodal ground truth and completion in the wild," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 28 003–28 013.

[13] E. Ozguroglu, R. Liu, D. Surís, D. Chen, A. Dave, P. Tokmakov, and C. Vondrick, "pix2gestalt: Amodal segmentation by synthesizing wholes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 3931–3940.

[14] K. Xu, L. Zhang, and J. Shi, "Amodal completion via progressive mixed context diffusion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 9099–9109.

[15] Y. Xue, Y. Cao, X. Feng, M. Xie, K. Li *et al.*, "Towards handling sudden changes in feature maps during depth estimation," *IEEE Transactions on Multimedia*, vol. 25, pp. 4002–4012, 2023.

[16] Y. Zhang, X. Huang, J. Ma, Z. Li *et al.*, "Recognize anything: A strong image tagging model," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 1724–1732.

[17] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang *et al.*, "Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection," in *European Conference on Computer Vision*, 2024, pp. 38–55.

[18] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland *et al.*, "Segment anything," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4015–4026.

[19] J. Li, D. Li, C. Xiong, and S. Hoi, "BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *International Conference on Machine Learning*, 2022, pp. 12 888–12 900.

[20] Y. Chen, K. Peng, A. Roitberg, D. Schneider, J. Zhang *et al.*, "Exploring self-supervised skeleton-based action recognition in occluded environments," in *Proceedings of the International Joint Conference on Neural Networks*, 2025, pp. 1–9.

[21] C. Zhu, F. Chen, U. Ahmed, Z. Shen, and M. Savvides, "Semantic relation reasoning for shot-stable few-shot object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 8782–8791.

[22] S. Kim, D. Jung, and M. Cho, "Relational context learning for human-object interaction detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 2925–2934.

[23] Y. Liu, G. Li, and L. Lin, "Cross-modal causal relational reasoning for event-level visual question answering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 10, pp. 11 624–11 641, 2023.

[24] Z. Wang, Y. Fan, and B. Ye, "Towards reliable X-ray security: SCE-DETR for detection and occlusion handling," in *Proceedings of the International Joint Conference on Neural Networks*, 2025, pp. 1–9.

[25] Z. Li, W. Ye, T. Jiang, and T. Huang, "2D amodal instance segmentation guided by 3D shape prior," in *European Conference on Computer Vision*, 2022, pp. 165–181.

[26] Y. Xiao, Y. Xu, Z. Zhong, W. Luo, J. Li *et al.*, "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.

[27] Y. Hu, H. Chen, K. Hui, J. Huang, and A. G. Schwing, "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*, 2019, pp. 3105–3115.

[28] P. Follmann, R. König, P. Härtinger, M. Klostermann, and T. Böttger, "Learning to see the invisible: End-to-end trainable amodal instance segmentation," in *IEEE Winter Conference on Applications of Computer Vision*, 2019, pp. 1328–1336.

[29] L. Ke, Y.-W. Tai, and C.-K. Tang, "Deep occlusion-aware instance segmentation with overlapping bilayers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 4019–4028.

[30] 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.

[31] R. Mohan and A. Valada, "Amodal panoptic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 21 023–21 032.

[32] Z. Li, W. Ye, T. Jiang, and T. Huang, "GIN: Generative invariant shape prior for amodal instance segmentation," *IEEE Transactions on Multimedia*, vol. 26, pp. 3924–3936, 2024.

[33] M. Tran, K. Vo, K. Yamazaki, A. Fernandes, M. Kidd, and N. Le, "Aisformer: Amodal instance segmentation with transformer," *arXiv:2210.06323*, 2022.

[34] W. Zhao, L. Yang, W. Zhang, Y. Tian, W. Jia, W. Li, M. Yang, X. Pan, and H. Yang, "Learning what and where to learn: A new perspective on self-supervised learning," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 8, pp. 6620–6633, 2023.

[35] Q. Fan, Z. Li, J. Li, and C. Cao, "MoDE: Mixture of diffusion experts for any occluded face recognition," in *Proceedings of the International Joint Conference on Neural Networks*, 2025, pp. 1–8.

[36] R. Birkl, D. Wofk, and M. Müller, "MiDaS v3.1–a model zoo for robust monocular relative depth estimation," *arXiv:2307.14460*, 2023.

[37] T. Ren, S. Liu, A. Zeng, J. Lin *et al.*, "Grounded SAM: Assembling open-world models for diverse visual tasks," *arXiv:2401.14159*, 2024.

[38] Y. Zhu, Y. Tian, D. Metaxas, and P. Dollár, "Semantic amodal segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1464–1472.

[39] H. Song, W. Song, L. Cheng, Y. Wei *et al.*, "PDD: Post-disaster dataset for human detection and performance evaluation," *IEEE Transactions on Instrumentation and Measurement*, vol. 73, p. 5501214, 2023.

[40] B. Zhang, Q. Liu, J. Zhang, Y. Wang, L. Liu *et al.*, "Amodal scene analysis via holistic occlusion relation inference and generative mask completion," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 7, 2024, pp. 6997–7005.