ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10096534

OAFORMER: LEARNING OCCLUSION DISTINGUISHABLE FEATURE FOR AMODAL INSTANCE SEGMENTATION

Zhixuan Li[1,2], *Ruohua Shi*[2], *Tiejun Huang*[2], *Tingting Jiang*✉[1,2*]

1. Advanced Institute of Information Technology,
Peking University, Hangzhou, China
2. National Engineering Research Center of Visual Technology,
School of Computer Science, Peking University, Beijing, China

ABSTRACT

The Amodal Instance Segmentation (AIS) task aims to infer the complete mask of occluded instance. Under many circumstances, existing methods treat occluded objects as unoccluded ones, and vice versa, leading to inaccurate predictions. This is because existing AIS methods do not explicitly utilize the occlusion rates of each object as supervision. However, occlusion information is critical for the methods to recognize whether the target objects are occluded. Hence we believe it is vital for the method to *be distinguishable about the degree of occlusion for each instance*. In this paper, a simple yet effective **O**cclusion-**a**ware trans**former**-based model, OAFormer, is proposed for accurate amodal instance segmentation. The goal of OAFormer is to learn the occlusion discriminative features. Novel components are proposed to enable OAFormer to be occlusion distinguishable. We conduct extensive experiments on two challenging AIS datasets to evaluate the effectiveness of our method. OAFormer outperforms state-of-the-art methods by large margins.

Index Terms— Amodal, instance segmentation, occlusion

1. INTRODUCTION

Amodal perception is an innate human ability to imagine the entire shape of an occluded object. Similarly, the amodal instance segmentation (AIS) task aims to predict the complete regions of occluded instances, while the visible instance segmentation task only predicts the visible regions. Occlusion problem exists widely in many computer vision tasks, including pedestrian re-identification [1, 2], robotic-arm grasping [3] and medical image segmentation [4].

At present, the AIS task has drawn great attention from the community. Several datasets [5, 6, 7, 8, 9] and methods [5, 6, 7, 10, 11] have been proposed to tackle the AIS problem from different perspectives, including directly learning methods [5, 6, 7, 8, 9, 10, 11, 12], relative-depth-based methods [13, 14] and shape-prior-based methods [15, 16]. However, all of the existing methods are *confused for distinguishing whether the object is occluded* to some extent. Fig. 1 shows typical errors caused by this *occlusion confusing problem*. As shown in Fig. 1(a), the occluded bottle is falsely recognized as an unoccluded object, leading to the wrong prediction. Besides, as shown in Fig. 1(b), the unoccluded chocolate is falsely recognized as an occluded object. These mistakes can be ascribed to *the ignorance of the occlusion degree* of each object.

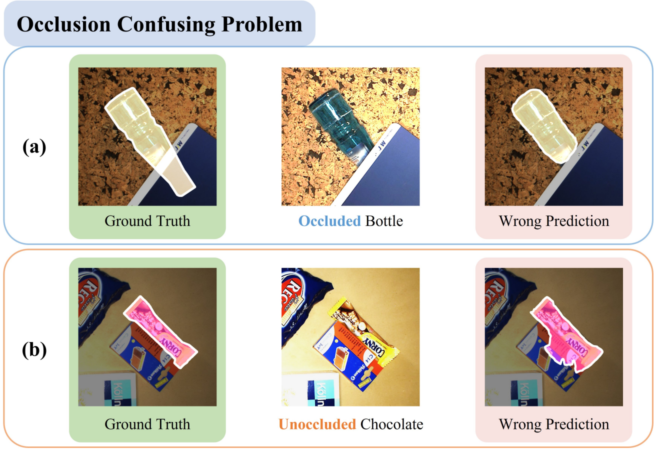

Fig. 1. Typical cases of the occlusion confusing problem. (a) Occlude bottle is regarded as unoccluded and resulting in wrong prediction. (b) Unoccluded chocolate is regarded as occluded and predicted wrongly.

The key to solving the occlusion confusing problem is to figure out the degree of occlusion. Therefore, it is necessary for the model to be aware of the occlusion degree of each object. In this paper, we propose an **O**cclusion-**a**ware trans**former**-based model named OAFormer to alleviate the occlusion confusing problem. OAFormer is based

[*] Corresponding author: ttjiang@pku.edu.cn.

This work was partially supported by the Natural Science Foundation of China under contracts 62088102. This work was also partially supported by Qualcomm. We also acknowledge High-Performance Computing Platform of Peking University for providing computational resources.

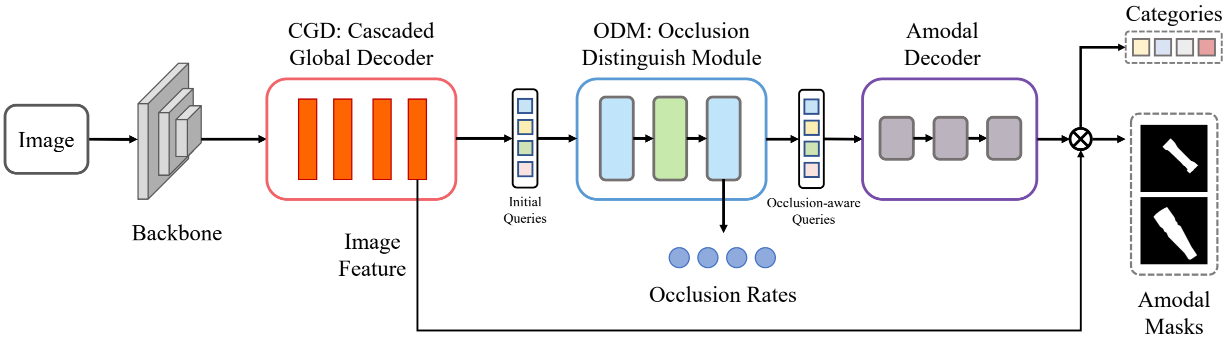

Fig. 2. **Overview of the proposed OAFormer.** OAFormer takes an image as the input. After extracting the features by the *Encoder* and the *Cascaded Global Decoder*, the *Occlusion Distinguish Module* predicts the occlusion rates of each target objects and embeds occlusion information into the attention masks. Finally, the *Amodal Decoder* takes the occlusion-aware attention masks and queries as input, and outputs the predicted amodal masks.

on the transformer-structured network [17], which achieves superior performance on the visible instance segmentation task. The original transformer network [17] uses randomly initialized queries and attention masks, which does not consider learning the occlusion information. In contrast, in this work, we propose to learn the occlusion information of each object instance to make the model aware of the occlusion degree. Specifically, OAFormer incorporates a novel component named Occlusion Distinguish Module (ODM). ODM is designed to enhance the existing transformer model from two aspects: (1) An occlusion-aware input query is proposed to learn and embed the occlusion information of each object in *instance-level*; (2) An occlusion discriminative attention mask is introduced to provide the occlusion information in *spatial-level*. The effectiveness of our method is evaluated on the challenging D2SA dataset [7] and COCOA-cls [7] dataset. Compared with existing methods, OAFormer achieves state-of-the-art performance.

Our contributions are summarized as follows: (1) To our best knowledge, OAFormer is the first method proposed to tackle the occlusion confusing problem. OAFormer is also the *first* transformer-based method for the AIS task. (2) OAFormer outperforms state-of-the-art methods on challenging AIS datasets, including D2SA and COCOA-cls, with a large margin, demonstrating the effectiveness of our method.

2. PROPOSED METHOD

In this section, we first introduce the task definition and the overall architecture. Then the newly proposed components for learning the occlusion-aware queries and attention masks are introduced in detail. Finally, the loss functions of the whole method are described.

2.1. Task Definition

Given an input image $I \in \mathbb{R}^{H\times W\times 3}$ containing K objects, amodal instance segmentation (AIS) task aims to predict the complete amodal mask $M_A \in \{0,1\}^{H\times W}$ and the category label $c \in \{1, 2, ..., C\}$ for each of the instances, including occluded and unoccluded ones. For each instance, the ground-truth amodal mask is defined as M_A^{gt}, and the ground-truth category is defined as c^{gt}.

2.2. Overall Architecture

The architecture of the proposed method is shown in Fig. 2. OAFormer contains three steps: (1) First, for each input image, the *Encoder* extracts the features through down-sampling. Then the *Cascaded Global Decoder (CGD)* ulteriorly learns and mines the global features of all instances through up-sampling. (2) Next, the **core component** *Occlusion Distinguish Module (ODM)* generates occlusion-aware queries and attention masks, which provides occlusion information in instance-level and spatial-level, respectively. (3) Finally, an *Amodal Decoder* takes the occlusion-aware queries as input and uses the attention mechanism with the predicted occlusion-aware attention masks to obtain the embedded query features of instances. Then the embedded query features are multiplied by the global feature to obtain the prediction of each instance. The basic components, CGD and Amodal Decoder, are replaceable. We instantiate OAFormer on the top of Mask2Former [17]. Moreover, our core component ODM is suitable for being plugged into the general transformer-serials models which take queries as input.

2.3. Encoder and Cascaded Global Decoder

The *Encoder* network aims to comprehend the input image preliminarily, which takes the image as input and outputs the features of the image. The *Cascaded Global Decoder (CGD)* aims to learn the feature of the image globally. A cascaded-style multi-level network [17] is employed to mine the output feature of the backbone network globally and deeply. The input feature of CGD is gradually upscaled to recover structural details like edges and semantic information. The output of CGD is used by two modules, including utilized by the ODM module to predict the occlusion-aware queries, and combined with the output of the amodal decoder to predict the final amodal masks.

2.4. Occlusion Distinguish Module

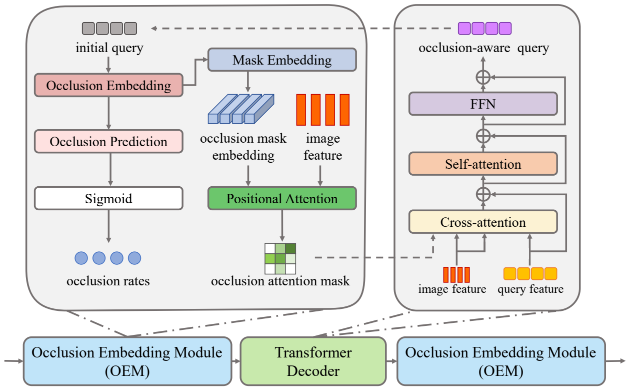

Fig. 3. **Detailed Architecture of Occlusion Distinguish Module.** ODM contains three steps: OEM → Transformer Decoder → OEM. The *Occlusion embedding module (OEM)* takes the randomly initialized queries as inputs, and outputs occlusion rates and attention masks. *Transformer decoder* module takes the occlusion attention mask, image features and randomly initialized query features as inputs, and outputs the occlusion-aware queries. Best viewed in color.

The Occlusion Distinguish Module (ODM) module is designed to learn occlusion-aware queries and attention masks to make the network distinguishable to the occlusion degree of each object, which can alleviate the occlusion confusing problem in the AIS task. The architecture of the ODM is shown in Fig. 3 with two components, occlusion embedding module and transformer decoder module, operating alternately. The ODM contains three steps, as shown below.

First, the initial random queries are fed into the *occlusion embedding module (OEM)* and are converted to the occlusion embedding vectors as shown in Fig. 3 (left). These vectors are then divided into two streams: 1) predicting the *occlusion rates* with the MLP and Sigmoid function, 2) combining the image features of the output of CGD and generating the *occlusion-aware attention masks* of the target objects.

Second, the *transformer decoder* module adopting the same meta-architecture as Mask2Former [17]. As shown in Fig. 3 (right), we modify the standard cross-attention by replacing the binarized mask prediction with the occlusion-aware attention mask. The transformer decoder module predicts the new occlusion-aware queries for the target objects with the query features and image features.

Third, the *occlusion-aware queries* are re-fed into the occlusion embedding module, and the final *occlusion rates* and *occlusion attention masks* are predicted.

2.5. Amodal Decoder

The Amodal Decoder is employed to predict the final predictions. The input contains the occlusion-aware queries and attention masks, and the output contains the feature embeddings of all queries. Finally, these feature embeddings are multiplied with the image features to obtain the final predictions, including amodal masks and class labels.

2.6. Loss Functions

There are three kinds of loss functions used: *occlusion loss*, *mask loss* and *classification loss*. The *occlusion loss* $\mathcal{L}_{\text{occlusion}}$ optimizes the occlusion rates predicted by ODM through the smooth L1 loss [18]. The *mask loss* $\mathcal{L}_{\text{mask}}$ is used to supervised the prediction of amodal masks. The *mask loss* consist of the Cross Entropy loss [19] $\mathcal{L}_{\text{ce}}$ and the Dice loss [20] $\mathcal{L}_{\text{dice}}$. The *classification loss* $\mathcal{L}_{\text{cls}}$ is the Cross Entropy loss [19]. When computing the three losses, the ground truth of the occlusion rate, amodal mask and category label of each instance is used respectively. The final loss function is:

$$\mathcal{L}_{\text{all}} = \lambda_1 \mathcal{L}_{\text{occlusion}} + \lambda_2 \mathcal{L}_{\text{ce}} + \lambda_3 \mathcal{L}_{\text{dice}} + \lambda_4 \mathcal{L}_{\text{cls}} \qquad (1)$$

where $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ are empirically set to 2, 0.5, 0.5, 0.2 to balance between different losses.

3. EXPERIMENTS

In this section, we first demonstrate the effectiveness of our proposed method quantitatively on three popular datasets compared with state-of-the-art methods. Then extensive ablation study is introduced.

3.1. Datasets and Experimental Settings

D2SA [7] and COCOA-cls [7] datasets are used for experiments. The D2SA dataset contains 60 categories and 5600 images for automatic checkout. COCOA-cls contains daily-life scene and contains 80 categories and 3501 images. The ground truth of the occlusion rate is contained in the official annotation of above datasets. The metric for evaluation is the same with COCO, including AP_{avg}, AP_{50} and AP_{75}. All reported results are rounded to the first decimal.

For fairness, all methods are trained on the same training dataset and validated on the same validation dataset. There is no extra dataset used. The ground-truth occlusion rates are provided in the official annotation. All methods use the same *Encoder* network ResNet-50-FPN [21].

Table 1. Comparison with state-of-the-art methods on the D2SA and COCOA-cls datasets. For supervision, “bbox” means amodal bounding box, “mask” means amodal masks, and “cls” means class labels. For each metric, the **bold** performance is the best, and the second-best is underlined.

Method	Publication	Supervision	D2SA			COCOA-cls		
			AP_{avg}	AP_{50}	AP_{75}	AP_{avg}	AP_{50}	AP_{75}
Mask-RCNN [22]	ICCV’19	bbox, mask, cls	63.6	83.9	68.0	33.7	**56.5**	35.8
ORCNN [7]	WACV’19	bbox, mask, cls	64.2	83.6	69.1	28.0	53.7	25.4
SLN [13]	ACM MM’19	bbox, mask, cls	25.1	30.8	29.4	14.4	23.6	15.8
BCNet [11]	CVPR’21	bbox, mask, cls	50.9	66.9	57.2	22.1	32.3	24.5
ShapeDict [15]	AAAI’21	bbox, mask, cls	70.3	85.1	75.8	35.4	56.0	38.7
A3D [16]	ECCV’22	bbox, mask, cls	68.5	N/A	N/A	34.9	N/A	N/A
Ours (w/o ODM)	N/A	mask, cls	61.7	78.7	63.3	33.9	45.0	35.8
Ours (w/ ODM)	N/A	mask, cls	**72.5**	**86.5**	**76.1**	**37.4**	49.7	**40.5**

3.2. Comparison to Previous Methods

Our method is compared with state-of-the-art methods on the D2SA and COCOA-cls datasets, as shown in Tab. 1. Mask-RCNN [22] and BCNet [11] are two visible instance segmentation methods, trained with amodal annotations for comparison. The ORCNN [7], SLN [13], ShapeDict [15] and A3D [16] are amodal instance segmentation methods.

As shown in Tab. 1, the performance of our method outperforms all previous AIS methods. Moreover, our method also beats the VIS methods, including Mask-RCNN and BCNet. OAFormer outperforms the secondary best method Shapedict by 2.2% AP_{avg} on D2SA dataset and 2.0% AP_{avg} on COCOA-cls dataset. It is worth noticing that OAFormer only needs the ground truth of the amodal masks and class labels as supervision signals, while all the other methods need the ground-truth amodal bounding boxes additionally.

3.3. Ablation Study

To quantitatively analyze the effect of the proposed components of our OAFormer and verify the effectiveness of different factors, ablation study is conducted on the D2SA dataset. **Occlusion Distinguish Module (ODM).** The ODM is the core component in OAFormer. As shown in Tab. 1, our method with the proposed ODM can outperform our method without the ODM by 10.8% AP_{avg} on the D2SA dataset and 3.5% AP_{avg} on the COCOA-cls dataset. The results confirm the effectiveness of the proposed ODM.

Number of queries. The number of queries in the OAFormer represents the maximum amount of instances in each image. As shown in Tab. 2, the performances are improving while the number of queries increases from 10 to 200. This result implies that in a particular range, the more queries used in the OAFormer, the better performance can be obtained. However, the performance drops from 72.5% AP to 71.2% AP when the number of queries improves from 200 to 300. The result denotes that using 200 queries is more suitable than using 300 queries for the OAFormer on the D2SA dataset, because using 300 queries can cause the matching between predictions and ground truths being difficult.

Table 2. Ablation experiments of using the different number of queries in OAFormer on the D2SA dataset.

Number of queries	AP_{avg}	AP_{50}	AP_{75}
10	61.8	77.4	64.7
50	64.6	80.4	67.0
100	72.5	86.5	76.1
200	**72.5**	**86.9**	**76.1**
300	71.2	85.7	74.3

Table 3. Ablation experiments of three groups for using different combinations of loss functions for supervision.

Group	$\mathcal{L}_{ce}$	$\mathcal{L}_{dice}$	$\mathcal{L}_{occlusion}$	AP_{avg}	AP_{50}	AP_{75}
1		✓		59.5	77.6	61.4
		✓	✓	**62.9**	**80.5**	**65.2**
2	✓			70.0	85.5	72.7
	✓		✓	**70.7**	**85.6**	**73.6**
3	✓	✓		71.5	85.9	75.4
	✓	✓	✓	**72.5**	**86.5**	**76.1**

Loss functions. Ablation experiments are conducted for three losses. The prediction of amodal masks are supervised by $\mathcal{L}_{ce}$ and $\mathcal{L}_{dice}$, and at least one of them needs to be used. The prediction of occlusion rates are supervised by $\mathcal{L}_{occlusion}$. As shown in Tab. 3, in all three groups, the performance when using $\mathcal{L}_{occlusion}$ are better than no using $\mathcal{L}_{occlusion}$. The best performance can be obtained when using all three losses.

4. CONCLUSION

In this paper, we have proposed an end-to-end transformer-based method named OAFormer, which aims to handle the occlusion confusing problem in the AIS task. OAFormer contains two novel components that learn and embed each instance’s occlusion information to make the OAFormer occlusion distinguishable. Experiments show that OAFormer can achieve state-of-the-art performance on the D2SA and COCOA-cls datasets. We hope that OAFormer will serve as a strong baseline for the community to handle the occlusion confusing problem in the AIS task.

5. REFERENCES

[1] Jin Xie, Yanwei Pang, M. H. Khan, Rao Muhammad Anwer, Fahad Shahbaz Khan, and Ling Shao, "Mask-guided attention network and occlusion-sensitive hard example mining for occluded pedestrian detection," *IEEE TIP*, vol. 30, pp. 3872–3884, 2021.

[2] Xiaolin Song, Kaili Zhao, Wen-Sheng Chu, Honggang Zhang, and Jun Guo, "Progressive refinement network for occluded pedestrian detection," in *ECCV*, 2020.

[3] Seunghyeok Back, Joosoon Lee, Taewon Kim, Sangjun Noh, Raeyoung Kang, Seongho Bak, and Kyoobin Lee, "Unseen object amodal instance segmentation via hierarchical occlusion modeling," in *ICRA*, 2022, pp. 5085–5092.

[4] Yan Xu, Yang Li, Yipei Wang, Mingyuan Liu, Yubo Fan, Maode Lai, I Eric, and Chao Chang, "Gland instance segmentation using deep multichannel neural networks," *T-BME*, vol. 64, no. 12, pp. 2901–2912, 2017.

[5] Ke Li and Jitendra Malik, "Amodal instance segmentation," in *ECCV*, 2016, pp. 677–693.

[6] Yan Zhu, Yuandong Tian, Dimitris Metaxas, and Piotr Dollár, "Semantic amodal segmentation," in *CVPR*, 2017, pp. 1464–1472.

[7] Patrick Follmann, Rebecca König, Philipp Härtinger, Michael Klostermann, and Tobias Böttger, "Learning to see the invisible: End-to-end trainable amodal instance segmentation," in *WACV*, 2019, pp. 1328–1336.

[8] Lu Qi, Li Jiang, Shu Liu, Xiaoyong Shen, and Jiaya Jia, "Amodal instance segmentation with KINS dataset," in *CVPR*, 2019, pp. 3014–3023.

[9] Yuan-Ting Hu, Hong-Shuo Chen, Kexin Hui, Jia-Bin Huang, and Alexander G Schwing, "SAIL-VOS: Semantic amodal instance level video object segmentation - a synthetic dataset and baselines," in *CVPR*, 2019, pp. 3105–3115.

[10] Kiana Ehsani, Roozbeh Mottaghi, and Ali Farhadi, "SeGAN: Segmenting and generating the invisible," in *CVPR*, 2018, pp. 6144–6153.

[11] Lei Ke, Yu-Wing Tai, and Chi-Keung Tang, "Deep occlusion-aware instance segmentation with overlapping bilayers," in *CVPR*, 2021, pp. 4019–4028.

[12] Xiaosheng Yan, Feigege Wang, Wenxi Liu, Yuanlong Yu, Shengfeng He, and Jia Pan, "Visualizing the invisible: Occluded vehicle segmentation and recovery," in *ICCV*, 2019, pp. 7618–7627.

[13] Ziheng Zhang, Anpei Chen, Ling Xie, Jingyi Yu, and Shenghua Gao, "Learning semantics-aware distance map with semantics layering network for amodal instance segmentation," in *ACM MM*, 2019, pp. 2124–2132.

[14] Xiaohang Zhan, Xingang Pan, Bo Dai, Ziwei Liu, Dahua Lin, and Chen Change Loy, "Self-supervised scene de-occlusion," in *CVPR*, 2020, pp. 3784–3792.

[15] Yuting Xiao, Yanyu Xu, Ziming Zhong, Weixin Luo, Jiawei Li, and Shenghua Gao, "Amodal segmentation based on visible region segmentation and shape prior," in *AAAI*, 2021, vol. 35, pp. 2995–3003.

[16] Zhixuan Li, Weining Ye, Tingting Jiang, and Tiejun Huang, "2D amodal instance segmentation guided by 3D shape prior," in *ECCV*, 2022.

[17] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar, "Masked-attention mask transformer for universal image segmentation," in *CVPR*, 2022, pp. 1290–1299.

[18] Arief Rachman Sutanto and Dae-Ki Kang, "A novel diminish smooth L1 loss model with generative adversarial network," in *IHCI*. Springer, 2020, pp. 361–368.

[19] Ma Yi-de, Liu Qing, and Qian Zhi-Bai, "Automated image segmentation using improved PCNN model based on cross-entropy," in *MVSP*. IEEE, 2004, pp. 743–746.

[20] Carole H Sudre, Wenqi Li, Tom Vercauteren, Sebastien Ourselin, and M Jorge Cardoso, "Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations," in *DLMIA ML-CDS*, pp. 240–248. Springer, 2017.

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[22] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick, "Mask R-CNN," in *ICCV*, 2017, pp. 2961–2969.