applied sciences

MDPI

Article

Amodal Segmentation Just Like Doing a Jigsaw

Xunli Zeng †, Xiaoli Liu † and Jianqin Yin *

School of Artificial Intelligence, Beijing University of Posts and Telecommunications No. 10 Xitucheng Road, Haidian District, Beijing 100876, China; zengxunli2021@bupt.edu.cn (X.Z.); liuxiaoli134@bupt.edu.cn (X.L.)

* Correspondence: jqyin@bupt.edu.cn

† These authors contributed equally to this work.

Abstract: Amodal segmentation is a new direction of instance segmentation while considering the segmentation of the visible and occluded parts of the instance. The existing state-of-the-art method uses multi-task branches to predict the amodal part and the visible part separately and subtract the visible part from the amodal part to obtain the occluded part. However, the amodal part contains visible information. Therefore, the separated prediction method will generate duplicate information. Different from this method, we propose a method of amodal segmentation based on the idea of the jigsaw. The method uses multi-task branches to predict the two naturally decoupled parts of visible and occluded, which is like getting two matching jigsaw pieces. Then put the two jigsaw pieces together to get the amodal part. This makes each branch focus on the modeling of the object. And we believe that there are certain rules in the occlusion relationship in the real world. This is a kind of occlusion context information. This jigsaw method can better model the occlusion relationship and use the occlusion context information, which is important for amodal segmentation. Experiments on two widely used amodally annotated datasets prove that our method exceeds existing state-of-the-art methods. In particular, on the amodal mask metric, our method outperforms the baseline by 5 percentage points on the COCOA cls dataset and 2 percentage points on the KINS dataset. The source code of this work will be made public soon.

Keywords: computer vision; amodal segmentation; occlusion context

check for updates

Citation: Zeng, X.; Liu, X.; Yin, J. Amodal Segmentation Just Like Doing a Jigsaw. *Appl. Sci.* **2022**, *12*, 4061. https://doi.org/10.3390/app12084061

Academic Editors: Andrea Prati, Luis Javier García Villalba and Vincent A. Cicirell

Received: 20 March 2022
Accepted: 15 April 2022
Published: 17 April 2022

1. Introduction

When you are walking on the street and about to turn at an intersection, you see a bicycle wheel suddenly appearing in front of you, and you know that there is a cyclist behind the wall at the moment, although you don't see him. Then you stay in place, waiting for the cyclist to pass first. People often witness such scenes in their lives. But this is particularly difficult for robots. Because people have a powerful visual system, they can perceive the overall target object only through some local areas of the target object. In order for the robot to also have the overall visual ability to perceive the object through the local, visible information of the object (shown as Figure 1), the task of amodal segmentation [1] was proposed.

Amodal segmentation is a complex high-level perception task. It needs to segment both the visible part of the target object and the occluded part of the target object. The amodal mask can be considered to be composed of the visible mask and occlusion mask of the instance object. From the perspective of amodal segmentation task, the current research can be roughly divided into two categories. The first category thinks that amodal segmentation is a single task. These models obtain amodal perception ability by learning the amodal mask that people have annotated on the dataset and directly infers the target's amodal mask (ASN [2], SLN [3]). Mask R-CNN [4] is often trained with amodally annotated dataset as the baseline. ASN [2] adds whether there is an occlusion in the branch prediction area and uses the judgment information of whether there is occlusion to assist amodal segmentation. SLN [3] uses a new representation of a semantics-aware distance map instead of the mask as the prediction target to segment the amodal mask of the instance. The second category divides amodal segmentation into two parts(amodal part and visible part). For example, VRS&SP [5] first segment the visible part of the target object and then add shape prior information to infer the amodal mask. ORCNN [6] predicts the amodal part and the visible part separately and subtracts the two to get the occluded part.

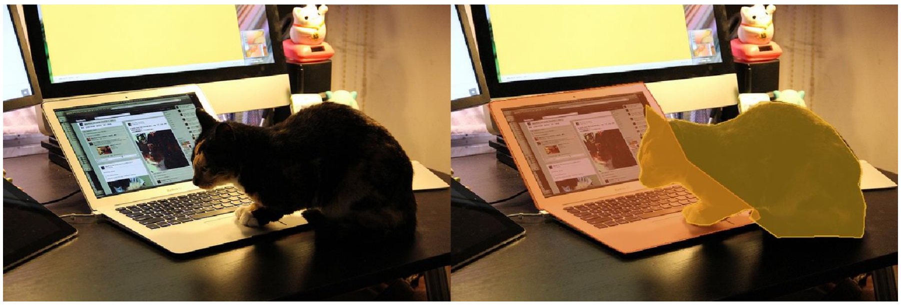

Figure 1. Humans have powerful amodal perception capabilities, just like when seeing the occluded scene in the picture on the left: a cat lying prone in front of the laptop. You can still perceive the complete shape of the laptop, as shown on the right.

Amodal segmentation is a complex task. Dividing it into two parts will help reduce the granularity of the model and improve the prediction effect of each part. In this way, the second category of methods is dominant. When the amodal segmentation is divided into two parts to complete, how to decompose is the key to determining the effect of the model. There is no intersection between the visible mask and the occlusion mask of the same instance, so the predictions of the two parts in the multi-task branch are decoupled, which will make each branch focus on the modeling of the object. What's more, we believe that there are certain rules in the occlusion relationship between the objects in the real world. For example, in the occlusion relationship formed by the dinner plate and bread, it is often that the bread obscures the dinner plate. This is a kind of occlusion context information. The exploration of this context can help the amodal segmentation. These motivate us to propose a multi-task branch and combine it with the occlusion relationship modeling amodal segmentation method. The method firstly uses multi-task branches to first obtain two pieces of the instance (visible mask and occlusion mask). And then, we model the occlusion relationship. Finally, we utilize the modeled occlusion relationship and stitch the two parts to get the complete jigsaw of the instance (amodal mask).

Our contributions could be summarized as the following aspects:

- We propose multi-task branch to obtain two pieces of instance (visible part and occlusion part). The proposed method stitches these parts to get the complete jigsaw of instances, which makes each branch focus on modeling of objects.
- We model the occlusion relationship utilizing the occlusion context information of the visible part and occlusion part, and we apply the occlusion relationship to complete the jigsaw of the instance, which helps the amodal segmentation greatly.
- The experimental results on two widely used datasets (KINS and COCOA cls) show our state-of-the-art performance, proving the effectiveness of our method.

2. Related Work

2.1. Instance Segmentation

As one of the four basic tasks of computer vision (classification, object detection, semantic segmentation, and instance segmentation), predecessors have done a lot of research. Among these works [7–11], the most representative one is Mask R-CNN [4] based on the Faster R-CNN [12] object detection framework, which sends the features extracted by the Backbone into The RPN generates proposals and uses RoIAlign feature pooling to obtain fixed-sized features of each proposal. Because of the fixed-sized features, the accuracy of segmentation is improved. PANet [13] makes the information path between the bottom-up and the top-level features of the deep network shorter by using bottom-up path augmentation. Mask scoring RCNN [14] adds an additional mask head branch to Mask R-CNN to learn MaskIoU consistent Mask score. The combination of Mask R-CNN and MaskIoU Head solves the problem of mismatch between the confidence score and localization accuracy of predicted masks. Ref. [15] uses multi-level feature networks in instance segmentation and proposes an attention-based feature pyramid module, effectively upgrades the performance of the instance segmentation method. These methods have reached state-of-the-art in the field of instance segmentation.

2.2. Amodal Instance Segmentation

The task of amodal segmentation was first proposed by [1]. They use the modally annotated data for object overlap data enhancement to generate amodal data, and then used it to train and validate their methods. They proposed the first method for amodal segmentation, which expands the bounding box of the instance and regenerates the heat map. With the release of some amodal annotation datasets, the research process of amodal segmentation has been accelerated. Ref. [16] uses an amodal annotated dataset to train ShapeMask [17], gets AmodalMask as the baseline. ORCNN [6] can directly predict the amodal mask and visible mask of the instance by adding the mask branch of Mask R-CNN [4]. The former subtracts the latter to get the occluded part. ASN [2] adds a branch to determine whether the instance is occluded and performs multi-level encoding of the determination result with the RoI feature map before predicting the amodal mask and then performs amodal segmentation. SLN [3] introduces a semantic-aware distance map instead of the mask as the prediction target to segment the amodal mask of the instance. VRS&SP [5] proposes to simulate human amodal perception, first roughly estimating the visible mask and amodal mask, and then use the shape prior to refining the amodal mask.

3. Methods

3.1. The Architecture of ARCNN

On the basis of Faster-RCNN [12], Mask-RCNN [4] adds the mask head to generate the mask of the object, and modifies RoI pooling to RoI Align to deal with the problem that the mask is not aligned with the object in the original image. ORCNN [6] can directly predict the amodal mask and visible mask of the object by adding a mask head on the basis of Mask R-CNN [4].

Our Amodal R-CNN (ARCNN) is shown in Figure 2. Inspired by Occlusion R-CNN (ORCNN) [6], we extend Mask R-CNN (MRCNN) [4] with two additional heads to predict amodal masks (amodal mask head) and the occlusion masks (occlusion mask head). As for the original mask head of MRCNN, it's used to predict visible masks(visible mask head). Different from ORCNN, our ARCNN predicts the amodal mask by flattening the visible mask and occlusion mask

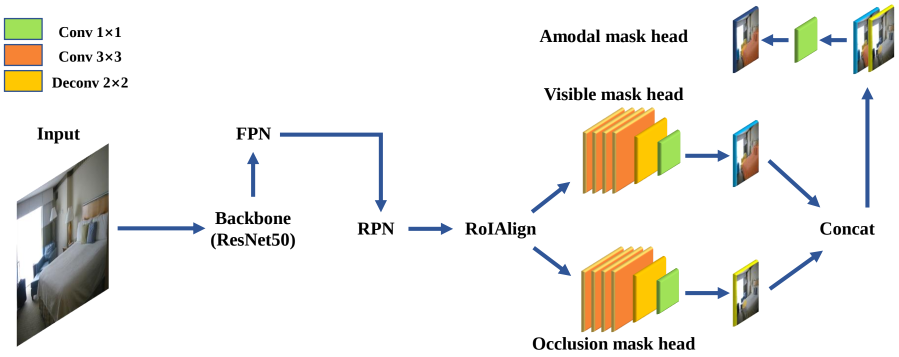

Figure 2. The architecture of ARCNN.

Firstly, image features are extracted by the ResNet50 [18]. In order to take into account the effect of object segmentation of different sizes, we perform multi-scale fusion of features. After these features are sent to FPN [19] for multi-scale fusion, they are input into the RPN to generate proposals. RPN is a network used to generate an area and determine whether there is a possible object in it. For areas where there may be a object, RPN will output it as a proposal. The proposals then are sent to RoIAlign to get RoIs as the input of visible mask head and occlusion mask head. The outputs of these two mask heads correspond to the visible and occlusion masks, respectively. Finally, the output of visible mask head and occlusion mask head are concated, then sent to the amodal mask head to obtain the amodal mask. Among them, in order to make the generated proposals can include the visible mask and occlusion mask of the instance, the RPN is trained with the bounding box of the amodal instance as the ground truth.

The visible mask head and occlusion mask head have the same structure, that is, four cascaded 3 × 3 convolutional layers, a 2 × 2 deconvolutional layer with stride 2, a 1 × 1 convolutional layer. These convolutional layers are used to predict the visible mask and the occlusion mask by using features from RoIs. The amodal mask head is a 1 × 1 convolutional layer, which is used to flatten the visible mask and the occlusion mask.

We propose the method to concat the visible mask and occlusion mask of the instance in a jigsaw-like operation so that we can decompose the prediction of the amodal mask into the prediction of the visible mask and the occlusion mask and make each branch focus on the modeling of the object. So as to better cope with the challenges brought by the complex task of amodal segmentation.

3.2. Modeling of Occlusion Relationship

The reason for the occlusion in the image is the overlap of two objects. And in the real world, this kind of overlap often contains certain rules. For example, in the occlusion relationship formed by the dinner plate and bread, it is often that the bread obscures the dinner plate. This is a kind of occlusion context information. Therefore, we use a 1 × 1 convolutional layer (amodal mask head) to model the relationship between the masks, thereby improving the model's ability to obtain occlusion context information during the amodal segmentation process. The modeling process is shown in Figure 3. The modeling the relationship is as follows:

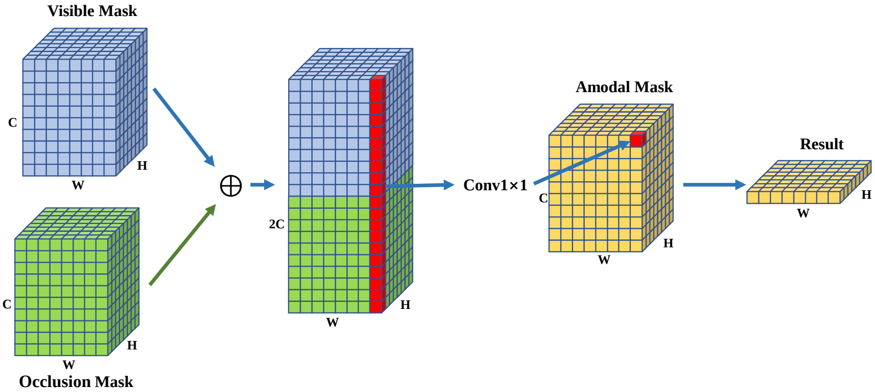

Figure 3. Visible Mask and Occlusion Mask are the output from the visible and occluded mask branches, respectively. The number of channels of the tensor C is the number of categories of instances in the dataset, and W and H are tensors, respectively, Width and height (in this paper, they are both 28). Amodal Mask is the mask predicted by the category of all instances in the dataset. Finally, according to the result of the classification, the mask of the corresponding class is selected as the output result.

$$A_M(:,:,i) = \sum_{j=1}^{n} [V_W(i:j)V_M(:,:,j) + O_W(i:j)O_M(:,:,j)] \tag{1}$$

The mask branch predicts each categorie's mask for the input ROI feature. It needs the classification branch to tell which category's mask is the result. Then we use this category's mask as the final output result. We assume i-th category is the classification's result. We denote V_M, O_M and A_M as Visible, Occlusion and Amodal Mask in Figure 3. $A_M(:,:,i)$ (Result in Figure 3) is the corresponding i-th category's mask in the Amodal Mask. The n is the number of instance category owned by the dataset. $V_M(:,:,j)$, $O_M(:,:,j)$ are respectively the visible mask and the occluded mask of the j-th category output by the mask branch. $V_W(i:j)$, $O_W(i:j)$ are the weights learned by the 1×1 convolutional layer that represents the relationship between the visible and occluded masks of the i-th category and the j-th category.

Due to the modeling of the occlusion relationship, the model can make full use of the occlusion context information.

3.3. Loss Function

The final prediction output of our model includes the bounding box, category, amodal mask, visible mask, and occlusion mask of the instance. These five parts are interrelated, and any part will affect the accuracy of the model. In order to coordinate the model as a whole, we assign the same weight to these five parts of loss.

We follow the settings in [6], L_{box} adopt standard Smooth *L1* loss; L_{cls} adopt standard cross entropy loss; L_{A_M}, L_{V_M}, and L_{O_M} all adopt standard binary cross entropy loss.

The above losses can be described as follows:

$$L_{box} = \begin{cases} 0.5b^2 & if|b|<1 \\ |b|-0.5 & otherwise \end{cases} \tag{2}$$

$$L_{cls} = -\sum_{i=1}^{n} c_i \log c'_i \tag{3}$$

$$L_M = -\frac{1}{N}\sum_{i=1}^{N} m_i \log m'_i + (1-m_i)\log(1-m'_i) \tag{4}$$

Among them, b in L_{box} refers to the difference between the real bbox and the predicted bbox; c in L_{cls} refers to the real category, c' refers to the predicted category; L_M can refer to the L_{A_M}, L_{V_M}, and L_{O_M}, where m_i refers to the real mask, m'_i refers to the predicted mask.

The final loss function L:

$$L = L_{box} + L_{cls} + L_{A_M} + L_{V_M} + L_{O_M} \tag{5}$$

4. Experiments

4.1. Datasets

Our experiments are conducted on the following two amodal annotated datasets: the KINS dataset [2] and the COCOA cls dataset [6].

The KINS dataset is based on the KITTI dataset [20] for autonomous driving. It consists of 7474 images in the training set and 7517 images in the validation set. The KINS dataset has seven categories of instances. The COCOA cls dataset is based on the COCOA dataset [16] and COCO dataset [21] about the complex everyday scenes. It consists of 2476 training images and 1223 validation images. This dataset has 80 categories of instances.

4.2. Experimental Details

We use detectron2 to build our model. All experiments are done on a GPU with a model of GeForce GTX 1080Ti and memory of 11G.

For a fair comparison, we chose the same hyperparameters as [5]. The main hyperparameters are set as follows: For the KINS dataset, batch size: 4, learning rate: 0.0025, iteration: 48,000. For the COCOA cls dataset, batch size: 2, learning rate: 0.0005, iteration: 10,000. Model training adopts the Stochastic Gradient Descent [22] strategy. The backbone of the model in the experiment is resnet50 [18].

4.3. Evaluation Criterion

In order to make the evaluation of the model in the amodal segmentation task have universal significance, we choose the average precision (AP) and average recall (AR) as metrics that are commonly used in the instance segmentation task. Among them, due to most of the occlusion mask has a small area, the deviation of a few pixels may make a huge difference with the ground truth IoU. Therefore, we calculate the AP of the amodal mask of instances where the occlusion rate exceeds 15% to reflect the model's ability to predict the occlusion. For fair comparisons, We use the evaluation API of the COCO dataset [21].

4.4. Baselines

- **ORCNN** [6] adds a branch to the Mask R-CNN, and the two branches respectively predict amodal mask and visible mask. Subtract the visible mask from the amodal mask to obtain the occlusion mask, thereby completing the task of amodal segmentation.
- **VRS & SP** [5] firstly estimates a coarse visible mask and a coarse amodal mask. Then based on the coarse prediction, it infers the amodal mask by concentrating on the visible region and utilizing the shape prior in the memory.

4.5. Experimental Results

We have completed the experiments of the method we proposed on two datasets and the reproduction of ORCNN [6]. The experimental results of the VRS&SP model are quoted from VRS&SP [5]. The performances of these models are shown in Tables 1 and 2. Occluded AP infers to amodal mask AP of the instances whose occlusion rate is more than 15%. ARCNN-add is the method directly adding the visible and occluded output of the branch.

4.5.1. Quantitative Analysis

We have carried out the following three comparisons and analyses.

ARCNN *vs.* ORCNN. It can be seen from the table that the evaluation indicators of the amodal mask and the occluded mask segmented by ARCNN on the two datasets exceed ORCNN. And for the COCOA cls dataset, ARCNN significantly surpasses ORCNN in the performance of amodal mask and occluded mask. For the visible mask prediction, there is only a small difference (less than 0.3) between the two indicators. This shows that our proposed method surpasses ORCNN in the performance of amodal segmentation.

ARCNN-add *vs.* ORCNN. ARCNN-add is a method of directly adding the output of the visible and occluded branches to get the amodal mask. It has a similar network composition to ORCNN. But it has roughly the same performance as ORCNN on the KINS dataset. On the COCO cls dataset, ARCNN-add's indicators fully exceed ORCNN. This shows that our jigsaw-like idea is effective in improving performance on amodal segmentation tasks.

ARCNN *vs.* VRS & SP. VRS & SP is a state-of-the-art method that introduces shape priors. It can be seen from the experimental results that, except for the occluded mask evaluation indicators of the COCOA cls dataset, the ARCNN we proposed exceeds VRS&SP in all indicators. This shows that our proposed method exceeds the current state-of-the-art methods.

4.5.2. Ablation Studies

We designed the method ARCNN-add for ablation experiments. The difference between ARCNN and ARCNN-add is that ARCNN not only stitches visible and invisible masks based on a jigsaw-like idea but also models the occlusion relationship between category instances. In terms of indicators, ARCNN surpasses ARCNN-add in both amodal mask and occluded mask. In terms of visible mask-related evaluation indicators, the gap between the two methods is very small. This shows that our proposed occlusion relationship modeling, using the context information of occlusion, can improve the performance of the model.

Table 1. The results on the KINS dataset.

		Amodal		**Visible**	**Occluded**
Model	**AP**	**AR**	**AP**	**AR**	**AP**
ORCNN	30.57	19.88	30.95	20.6	36.15
VRS & SP	32.08	20.9	29.88	19.88	37.4
ARCNN-add	30.2	19.75	30.89	20.61	36.27
ARCNN	32.94	20.96	30.68	20.56	38.71

Table 2. The results on the COCOA cls dataset.

		Amodal		**Visible**	**Occluded**
Model	**AP**	**AR**	**AP**	**AR**	**AP**
ORCNN	30.75	32.55	34.8	36.78	18.9
VRS & SP	35.41	37.11	34.58	36.42	22.17
ARCNN-add	32.26	34.06	35.46	37.25	19.32
ARCNN	36.29	37.39	35.48	36.69	20.84

4.5.3. Visualization of Amodal Results

Also, we visualized the amodal results of our method and ORCNN, and the results are shown in Figure 4. From the comparison of the pictures, we can see that our proposed method is more complete in the amodal mask predicting. And the ARCNN predicts that the amodal mask is smoother than the ARCNN-add. This also proves the effectiveness of our proposed method from another angle.

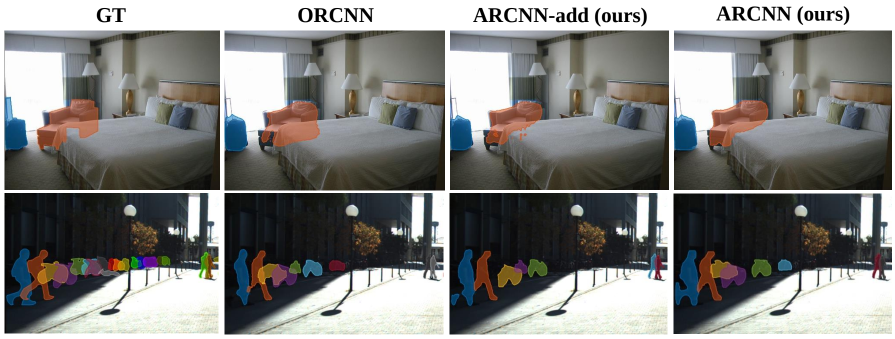

Figure 4. The columns from left to right are the ground-truth amodal masks, prediction of ORCNN and Ours, respectively. The first row is the result from COCOA cls dataset. And the other row is from KINS dataset.

4.6. Visualization of Occlusion Relationship

Interpretation of relational modeling is proved by our experiments and analysis to be independent of categories. For better interpretation of relational modeling through better visualization, we visualize the weights of the 1 × 1 convolutional layer between the modeling cat of the model trained under the COCOA cls dataset and other classes as Figure 5.

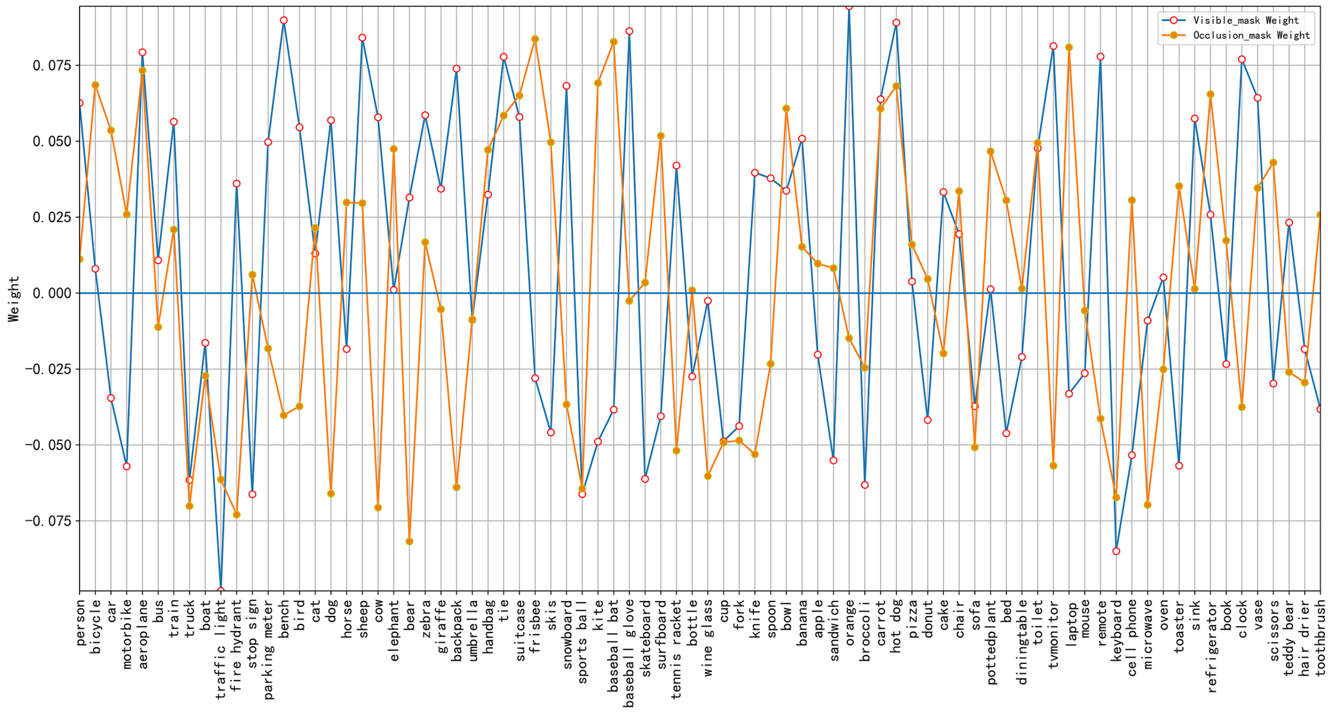

Figure 5. This is the plot of the cat's 1 × 1 conv weight. The abscissa represents the category of the instance in the dataset. The ordinate is the weight value.

The figure contains two types of information. The first category is the correlation between other categories of masks and cat category masks. If the weight is greater than zero, the two are positively correlated, if the weight is less than zero, the two are negatively correlated, and if the weight is equal to zero, the two are uncorrelated. The second category is the relationship between masks of other categories and the occlusion order of the cat category of masks. $V_W(i:j)$, $O_W(i:j)$ respectively represent the possibility of the cat being occluded by the j-th category and cat occluding the j-th category. The relative size between the two reflects the relationship between other categories and the occlusion order of cat on the entire dataset to a certain extent. In the figure, the laptop corresponds to $V_W(i:j) < 0$, $O_W(i:j) > 0$, that is, the occlusion order that appears on the entire dataset is cat occluding the laptop. This is also can be confirmed in the process of visualizing the picture of the dataset. In the figure, the person corresponding to $V_W(i:j) > O_W(i:j) > 0$, that is, the relationship between person occluding cat and cat occluding person has appeared in the entire dataset, but the former appears more often.

5. Conclusions

In this paper, we propose a method of decomposing the task of amodal segmentation into the visible mask and occlusion mask prediction, and finally stitching the two parts to obtain the amodal mask. The predictions of these two parts are naturally decoupled. In this way, the division of labor of the network branches can be clearly realized so as to make each branch focus on the modeling of the object. And we believe that there are certain rules in the occlusion relationship in the real world, so we model it and applied the modeling results to obtain the amodal mask. Experimental results prove that our proposed method is simple and effective. In the case of adding only a small number of parameters, our method on the amodal mask metric outperforms the baseline by 5 percentage points on the COCOA cls dataset and 2 percentage points on the KINS dataset. The performance of our proposed method on amodal segmentation task exceeds the existing state-of-the-art methods.

Author Contributions: Conceptualization, X.Z. and X.L.; methodology, X.Z.; writing—original draft, X.L.; writing—review and editing, J.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded partly by the National Natural Science Foundation of China (Grant No. 62173045, 61673192), partly by the Fundamental Research Funds for the Central Universities (Grant No. 2020XD-A04-2), and partly by BUPT Excellent Ph.D. Students Foundation (CX2021314).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Li, K.; Malik, J. Amodal instance segmentation. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 October 2016; Springer: New York, NY, USA, 2016; pp. 677–693.
2. Qi, L.; Jiang, L.; Liu, S.; Shen, X.; Jia, J. Amodal instance segmentation with kins dataset. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3014–3023.
3. Zhang, Z.; Chen, A.; Xie, L.; Yu, J.; Gao, S. Learning semantics-aware distance map with semantics layering network for amodal instance segmentation. In Proceedings of the 27th ACM International Conference on Multimedia, New York, NY, USA, 21–25 October 2019; pp. 2124–2132.
4. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2961–2969.
5. Xiao, Y.; Xu, Y.; Zhong, Z.; Luo, W.; Li, J.; Gao, S. Amodal Segmentation Based on Visible Region Segmentation and Shape Prior. In Proceedings of the AAAI Conference on Artificial Intelligence, Online, 2–9 February 2021.
6. Follmann, P.; König, R.; Härtinger, P.; Klostermann, M.; Böttger, T. Learning to see the invisible: End-to-end trainable amodal instance segmentation. In Proceedings of the 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE, Waikoloa, HI, USA, 7–11 January 2019; pp. 1328–1336.
7. Hariharan, B.; Arbeláez, P.; Girshick, R.; Malik, J. Simultaneous detection and segmentation. In Proceedings of the European Conference on Computer Vision, Zürich, Switzerland, 6–12 September 2014; Springer: New York, NY, USA, 2014; pp. 297–312.
8. Dai, J.; He, K.; Sun, J. Instance-aware semantic segmentation via multi-task network cascades. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 3150–3158.
9. Li, Y.; Qi, H.; Dai, J.; Ji, X.; Wei, Y. Fully convolutional instance-aware semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2359–2367.
10. Chen, L.C.; Hermans, A.; Papandreou, G.; Schroff, F.; Wang, P.; Adam, H. Masklab: Instance segmentation by refining object detection with semantic and direction features. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4013–4022.
11. Pinheiro, P.O.; Lin, T.Y.; Collobert, R.; Dollár, P. Learning to refine object segments. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 October 2016; Springer: New York, NY, USA, 2016, pp. 75–91.
12. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *39*, 1137–1149. [CrossRef] [PubMed]
13. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path aggregation network for instance segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8759–8768.
14. Huang, Z.; Huang, L.; Gong, Y.; Huang, C.; Wang, X. Mask scoring r-cnn. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 6409–6418.
15. Sun, Y.; Gao, W.; Pan, S.; Zhao, T.; Peng, Y. An Efficient Module for Instance Segmentation Based on Multi-Level Features and Attention Mechanisms. *Appl. Sci.* **2021**, *11*, 968. [CrossRef]
16. Zhu, Y.; Tian, Y.; Metaxas, D.; Dollár, P. Semantic amodal segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1464–1472.
17. Pinheiro, P.O.; Collobert, R.; Dollár, P. Learning to segment object candidates. *arXiv* **2015**, arXiv:1506.06204.
18. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
19. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.

20. Geiger, A.; Lenz, P.; Urtasun, R. Are we ready for autonomous driving? the kitti vision benchmark suite. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, Providence, RI, USA, 16–21 June 2012; pp. 3354–3361.
21. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; Springer: New York, NY, USA, 2014; pp. 740–755.
22. Zinkevich, M.; Weimer, M.; Smola, A.J.; Li, L. Parallelized Stochastic Gradient Descent. In Proceedings of the NIPS, Citeseer, Vancouver, QC, Canada, 6–11 December 2010; Volume 4, p. 4.