LETTER

Amodal Instance Segmentation of Thin Objects with Large Overlaps by Seed-to-Mask Extending

Ryohei KANKE[†a)], *Student Member and* Masanobu TAKAHASHI[††b)], *Member*

SUMMARY Amodal Instance Segmentation (AIS) aims to segment the regions of both visible and invisible parts of overlapping objects. The mainstream Mask R-CNN-based methods are unsuitable for thin objects with large overlaps because of their object proposal features with bounding boxes for three reasons. First, capturing the entire shapes of overlapping thin objects is difficult. Second, the bounding boxes of close objects are almost identical. Third, a bounding box contains many objects in most cases. In this paper, we propose a box-free AIS method, Seed-to-Mask, for thin objects with large overlaps. The method specifies a target object using a seed and iteratively extends the segmented region. We have achieved better performance in experiments on artificial data consisting only of thin objects.

key words: *amodal instance segmentation, deep learning, thin object, seed*

1. Introduction

In recent years, due to advancements in deep learning, CNN-based methods have been applied to various tasks in the field of computer vision. Within these tasks, Instance Segmentation (IS) is the basic task that performs pixel-level classification and object detection, distinguishing even homogeneous objects. In addition, while IS divides visible regions, Amodal Instance Segmentation (AIS) aims to segment not only visible regions but also hidden regions occluded by other objects. Figure 1 illustrates the difference between IS and AIS outputs.

Humans are capable of amodal perception; in contrast, the computer is limited to modal perception [1], so many researchers are working on AIS. When an object is occluded by an object of another class, the computer has no visual cue to extend the visible mask to the amodal mask including the hidden part. Furthermore, when an object is occluded by another instance of the same class, category-specific features are present in the surroundings. These characteristics make it difficult to determine how much to extend amodal masks. Therefore, AIS is one of the more challenging tasks.

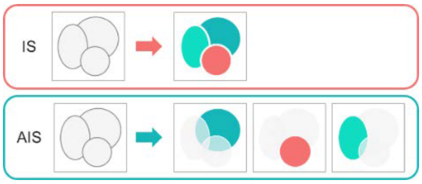

Fig. 1 Comparison of IS and AIS outputs.

Mask R-CNN showed excellent performance on IS [2]. It consists of two stages; the first stage detects bounding boxes as object candidates, and the second stage segments the object regions based on each bounding box. Because Mask R-CNN can be directly applied to AIS [3], many subsequent methods of AIS adopted the same two-stage pipeline and achieved improvements [3]–[5].

However, Mask R-CNN-based methods are not suitable for thin objects with large overlaps (e.g., hair, cords, or thin cells) because of the use of bounding boxes. In the case of objects for which it is difficult to identify the overall shape, incorrect bounding boxes detected in the first stage can cause segmentation failure, regardless of the segmentation performance in the second stage. In addition, the bounding boxes of the overlapping thin objects may be almost identical. In such cases, Non-Maximum Suppression (NMS) frequently excludes them from the pool of object candidates due to their high IoU [6]. Furthermore, the target is not clear because a bounding box can contain many objects in most cases. This leads to missing target and incorrect segmenting of other objects.

We propose a box-free AIS method, Seed-to-Mask, for overlapping thin objects. Unlike the two-stage method based on Mask R-CNN, the proposed method attempts to segment an object from the beginning by specifying the object with a seed. It then generates a new seed from the result and repeatedly segments an object. This repetition process enables the extraction of the entire object by progressively extending the segmented region from an initial seed. In the experiments conducted on artificial data, Seed-to-Mask outperformed existing baseline methods.

2. Related Work

In this paper, we work on AIS for data of thin objects with large overlaps. The conventional mainstream of AIS is Mask R-CNN-based, box-mediated methods.

Since Mask R-CNN can be directly applied to AIS by providing amodal ground truth [3], many Mask R-CNN-based AIS methods have been proposed. ORCNN [3] sep-

Manuscript received October 13, 2023.
Manuscript revised January 8, 2024.
Manuscript publicized February 29, 2024.

[†] The author is with the Graduate School of Systems Engineering and Science, Shibaura Institute of Technology, Saitama-shi, 337–8570 Japan.

[††] The author is with the College of Systems Engineering and Science, Shibaura Institute of Technology, Saitama-shi, 337–8570 Japan.

a) E-mail: mf22042@shibaura-it.ac.jp
b) E-mail: mtaka@shibaura-it.ac.jp

DOI: 10.1587/transinf.2023EDL8068

arates the mask head of Mask R-CNN into two parts: one for amodal regions and one for visible regions. Based on ORCNN, ASN [4] incorporates an occlusion classification branch and a multilevel coding module to improve reasoning ability for invisible parts. BCNet [5] explicitly models the occlusion relationship with the bilayer structure. This structure naturally decouples the boundaries of both occluding and occluded instances and considers the interaction between them during mask regression. Although these Mask R-CNN-based methods have achieved better performance on data of general objects, they are not suitable for data of thin objects with large overlaps due to use of bounding boxes. Thus, we attempt to segment the final amodal mask from the beginning without bounding boxes.

The repetition of segmentation process for an object is proposed in [7]. However, a bounding box is expanded based on a heatmap. The use of bounding boxes leads to the disadvantages, as described in Sect. 1, such as confusion of the segmentation target caused by multiple objects in a bounding box. AdaptIS [8] specifies the target by a point, whereas it does not specify objects in the input image but in the latent space, that is, it does not specify objects in exact coordinates. Furthermore, it attempts to segment an object at a time, which is not suitable for dense, thin objects. Moreover, AdaptIS is method for not AIS but IS. Therefore, we propose a method that strictly specifies the target in coordinates and repeatedly expands the segmented area without using a bounding box.

3. Proposed Method

The proposed box-free AIS method, See-to-Mask, specifies the target by seed (seed region). Figure 2 illustrates the overall flowchart of the proposed method for thin object data of single class. The method consists of the following five steps.

1. Segmentation of All Object Regions Given an input image, CNN$_1$ segments the regions of all thin objects without distinguishing between instances. The image of the segmented result is referred to as the seed candidate image.

2. Generating the scored candidate image For the seed candidate image, CNN$_2$ predicts a scored candidate image. The pixel value of the scored candidate image is the IoU between the segmentation result using the pixel as a seed and the correct region of the corresponding object. A higher score means that the object can be segmented more accurately. The details of the setup for training CNN$_2$ are described in Sect. 4.1.

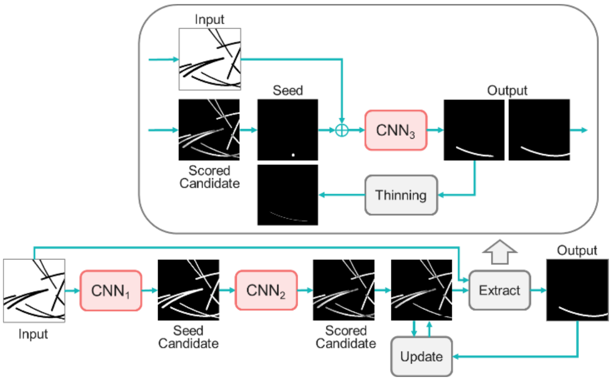

Fig. 2 Overview of the proposed method, Seed-to-Mask.

3. Segmentation of the target object specified by the seed The seed is set at the point with the highest score in the scored candidate image. The seed image is an image in which the five pixels centered on the seed are white and the other pixels are black. CNN$_3$ receives an input image formed by concatenating the seed image with the original image along the channel direction. It then segments the target object region specified by the seed. The output of CNN$_3$ is binarized, expanded, the regions not including the seed are deleted, and shrunk to obtain a temporary segmentation result. Then, this temporary result is thinned to make a line seed image. Finally, CNN$_3$ again performs segmentation with the concatenated image of the new seed image and the original image. By iterating this process, the segmented area is extended so that the entire object is accurately captured. We obtain the final result as a temporary segmentation result after specific iterations.

4. Excluding segmented regions from candidates The segmented region is removed from the scored candidate image.

5. Repeating steps 3 and 4 By repeating steps 3 and 4 until no seed candidates remain in the scored candidate image, all objects are segmented.

By specifying the targets by seeds and segmenting from the beginning, Seed-to-Mask avoids the adverse effects of using bounding boxes. Additionally, by iteratively segmenting regions with the seed created from the temporary segmentation result, the proposed method can extend the segmented regions to spatially distant regions.

4. Experiments

4.1 Experimental Setup

Dataset Experiments were conducted on artificial data of thin objects with large overlaps. Commonly used datasets for benchmarking in AIS, such as COCOA [9] and KINS [4], have small overlaps between objects. Thus, bounding box-mediated methods such as Mask R-CNN have shown sufficient performance [3]–[5]. In contrast, few public datasets exist for thin objects with large overlaps. Therefore, we created experimental data by simulating it with a portion of an elliptical arc.

Figure 3 shows the method for specifying the arcs to be drawn, and Fig. 4 shows examples of the created images. On an image of size of 512×512, 4–16 elliptical arcs specified by random length of major and minor axes (30–512 pixels), rotation angle (−90–90 degrees), starting angle (0–90 degrees), and arc angle (45–180 degrees) were drawn at random locations. Then, central 256×256 pixels were clipped as the thin object image.

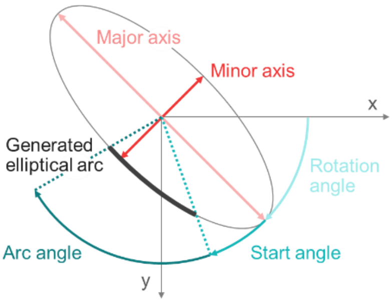

Fig. 3 Method for specifying an elliptical arc as a thin object.

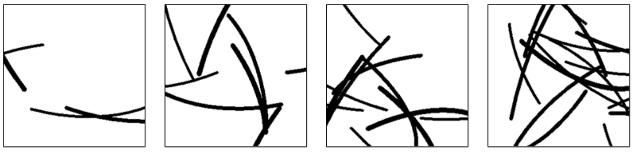

Fig. 4 Samples of simulated thin object images (from left to right: 4, 8, 12 and 16 objects per image).

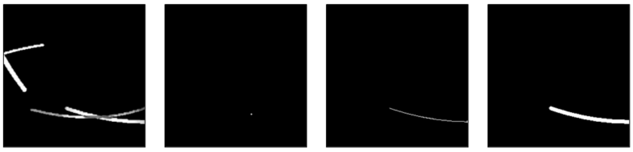

Fig. 5 Samples of a scored candidate image, a point seed image, a line seed image, and the ground truth of the corresponding object from left to right for the leftmost sample of Fig. 4.

Figure 5 shows examples of a scored candidate image, a point seed image, and a line seed image. A point seed image for training was created by selecting random pixels from the object region, and a line seed image was created by shortening the thinned region of the ground truth. After training CNN_1 and CNN_3, the ground truth of the scored candidate image was created by calculating IoUs between (a) the segmented region when each pixel was used as a seed and (b) the ground truth region corresponding to the object.

We used 14,500 images (10,000 for training, 2,000 for validation, and 2,500 for evaluation) for our experiments, with an average number of objects per image of 10.0 and a standard deviation of 4.0.

Implementation Because the artificial data were binary images, segmentation of all regions predicted by CNN_1 in step 1 is not necessary. Thus, we substituted black-and-white inverted images of the original images as seed candidate images. We employed Attention U-Net [10] for CNN_2 to generate scored candidate images in step 2 and CNN_3 to segment objects in step 3. Attention U-Net improves U-Net [11] by incorporating channel and spatial attentions.

The input of CNN_2 is a black-and-white inversion of the original image, and the output is a one-channel scored candidate image of the same size. The input of CNN_3 is a two-channel 256 × 256 pixel image composed of the original grayscale image and the seed image, and the output is a one-channel image of the segmented thin object region of the same size. The other parameters were the same as those in the original paper [10].

Table 1 Performance comparison with all test data.

Method	Bounding Box			Mask		
	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}
Mask R-CNN [2]	80.94	96.96	89.26	52.7	86.96	60.93
ORCNN [3]	87.76	95.75	**92.99**	56.58	88.71	67.34
BCNet [5]	86.21	**97.71**	92.12	61.58	92.53	73.66
Ours	**91.27**	95.46	92.24	**64.97**	**95.69**	**83.69**

Table 2 Performance comparison with test data of constant density. AP(X) denotes the results of AP evaluation on data of density X.

Method	Bounding Box			Mask		
	AP(6)	AP(10)	AP(14)	AP(6)	AP(10)	AP(14)
Mask R-CNN [2]	87.45	80.61	74.46	60.20	52.09	45.22
ORCNN [3]	92.57	87.60	82.40	63.29	56.10	49.41
BCNet [5]	91.12	86.00	80.77	65.78	61.25	56.58
Ours	**94.93**	**91.25**	**86.64**	**66.86**	**65.50**	**62.02**

CNN_2 and CNN_3 were trained with Adam as an optimizer, batch size of 4, learning rate starting 10^{-4} and halving every 50 epochs for 100 and 200 epochs, respectively. Although we could have trained two CNN_3's, one with a point seed and the other with a line seed, we trained one CNN_3 with the same ratio of both seed images.

Metrics We evaluated the methods with Mean Average Precision (mAP), a commonly used metric in AIS tasks, for both bounding box and mask. Note that while the mAP is averaged over all classes, the number of classes was only one in this experiment; hence, the evaluation was conducted using AP. AP_{50}, AP_{75}, and AP represent the results at IoU thresholds of 0.50, 0.75, and the average over 0.50–0.95 in 0.05 steps, respectively.

4.2 Performance Comparison and Analysis

In the experiments, we compared the proposed Seed-to-Mask with Mask R-CNN, ORCNN, and BCNet. All the comparison methods used ResNet50+FPN as their backbone. We set the number of iterations of CNN_3 in step 3 of the proposed method to four.

Table 1 shows the results of the evaluation of all test data. The table shows that the Mask R-CNN-based method has the highest precision for bounding box AP_{50}. In contrast, Seed-to-Mask is the best for bounding box AP and all mask APs. The high precisions of the proposed method even at high IoU thresholds indicates that the proposed method can accurately segment objects down to the smallest details.

Table 2 shows the results of the evaluation on subsets of the test data, 500 images each with densities of 6, 10, and 14 objects per image. The table shows that the mask APs of the Mask R-CNN-based methods decrease by 9.20–14.98 as the density increases from 6 to 14 objects per image. By contrast, the proposed method suppresses the decrease to 4.84. In essence, the proposed method exhibits robustness to object overlaps because its precision is less degraded with

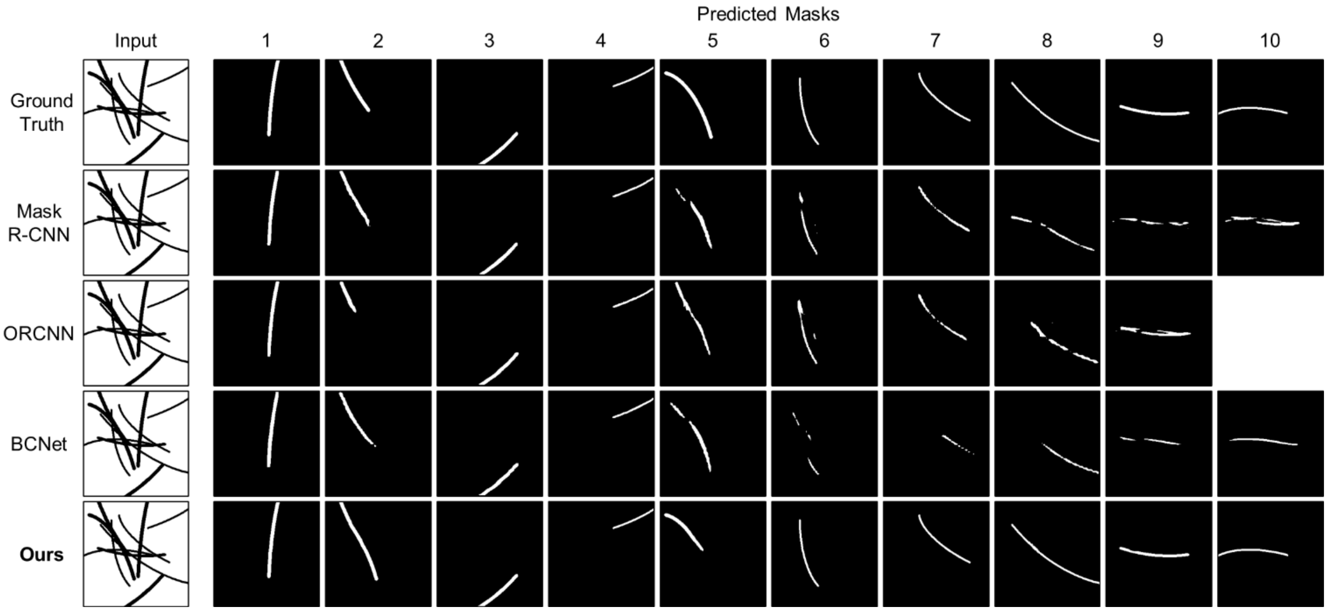

Fig. 6 Qualitative results comparison of amodal mask predictions.

increasing object density.

Figure 6 shows the prediction results for an image with large overlaps. Seed-to-Mask method segments the objects more accurately than the other comparison methods. For instance, for the long object in the eighth row, the comparison methods fail to segment the entire object. However, the proposed method correctly segments the entire object. Furthermore, in the case of rows 9 and 10, where the bounding boxes of the two objects are almost identical, the comparison methods fail to segment them correctly. In contrast, the proposed method can distinguish and segment them separately. This qualitative comparison underscores the effectiveness of the proposed method in scenarios with substantial bounding box overlaps.

5. Conclusion

In this paper, we have proposed a box-free AIS method, Seed-to-Mask, which specifies a part of the target object by a seed in the input image and extends the segmented region from the seed by iterating the segmentation process. This method can identify the same region even when they are spatially distant from each other. Importantly, it also eliminates the need to detect object candidates using bounding boxes. The effectiveness of this method was confirmed in experiments using artificial data consisting only of thin objects.

Seed-to-Mask successfully performs segmentation even for high-density data. Therefore, we anticipate that the proposed method will exhibit even more substantial improvements when applied to more complex and crowded data. In the future, we plan to apply the proposed method to real data such as hair images. In addition, because the data consisted only of thin objects, we plan to conduct additional experiment to confirm whether the proposed method also works with data containing both large objects and thin objects.

References

[1] A. Valada, A. Dhall, and W. Burgard, “Convoluted mixture of deep experts for robust semantic segmentation,” IEEE/RSJ Int. Conf. on Intelligent Robots and Syst. (IROS) Workshop, State Estimation and Terrain Perception for All Terrain Mobile Robots, 2016.
[2] K. He, G. Gkioxari, P. Dollár, and R.B. Girshick, “Mask R-CNN,” Proc. IEEE Int. Conf. on Computer Vision, pp.2980–2988, 2017.
[3] P. Follmann, R. König, P. Härtinger, M. Klostermann, and T. Böttger, “Learning to see the invisible: End-to-end trainable amodal instance segmentation,” 2019 IEEE Winter Conf. on Applications of Computer Vision, pp.1328–1336, 2019.
[4] L. Qi, L. Jiang, S. Liu, X. Shen, and J. Jia, “Amodal instance segmentation with KINS dataset,” Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition, pp.3014–3023, 2019.
[5] L. Ke, Y.-W. Tai, and C.-K. Tang, “Deep occlusion-aware instance segmentation with overlapping bilayers,” Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition, pp.4019–4028, 2021.
[6] S. Liu, D. Huang, and Y. Wang, “Adaptive NMS: Refining pedestrian detection in a crowd,” Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition, pp.6459–6468, 2019.
[7] K. Li and J. Malik, “Amodal instance segmentation,” Proc. 14th European Conf. on Computer Vision, pp.677–693, 2016.
[8] K. Sofiiuk, O. Barinova, and A. Konushin, “AdaptIS: Adaptive instance selection network,” Proc. IEEE/CVF Int. Conf. on Computer Vision, pp.7355–7363, 2019.
[9] Y. Zhu, Y. Tian, D. Metaxas, and P. Dollár, “Semantic amodal segmentation,” Proc. IEEE Conf. on Computer Vision and Pattern Recognition, pp.1464–1472, 2017.
[10] O. Oktay, J. Schlemper, L.L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N.Y. Hammerla, B. Kainz, B. Glocker, and D. Rueckert, “Attention U-Net: Learning where to look for the pancreas,” arXiv preprint, arXiv:1804.03999, 2018.
[11] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional networks for biomedical image segmentation,” Proc. 18th Int. Conf. on Medical Image Computing and Computer Assisted Intervention, pp.234–241, 2015.