Received: 1 April 2023 | Revised: 2 July 2023 | Accepted: 11 July 2023

IET Image Processing

DOI: 10.1049/ipr2.12880

The Institution of Engineering and Technology WILEY

ORIGINAL RESEARCH

MaskDis R-CNN: An instance segmentation algorithm with adversarial network for herd pigs

Shuqin Tu[1] | Qiantao Zeng[1] | Haofeng Liu[2] | Yun Liang[1] | Xiaolong Liu[1] | Lei Huang[1] | Zhengxin Huang[1]

[1]College of Mathematics and Informatics, South China Agricultural University, Guangzhou, China

[2]Department of Computer Science and Engineering, Southern University of Science and Technology, Shenzhen, China

Correspondence
Yun Liang, College of Mathematics and Informatics, South China Agricultural University, No. 483, Wushan Road, Tianhe District, Guangzhou City, Guangdong Province 510642, China.
Email: Liang_Yun168@163.com

[Correction added on 8 September 2023, after first online publication: Funding information has been updated in this version].

Funding information
National Natural Science Foundation of China, Grant/Award Numbers: 61772209, 31600591; the Science and Technology Planning Project of Guangdong Province, Grant/Award Number: 2019A050510034; Guangzhou Key R&D Program Project, Grant/Award Numbers: 2023B03J1363, 202206010091; College Students' Innovation and Entrepreneurship Competition, Grant/Award Number: 202110564025

Abstract

The current instance segmentation method can achieve satisfactory results in common scenarios. However, under the overlap or partial occlusion between targets caused by the complex scenes, accurate segmentation of pigs remains a challenging task. To address the problem, the authors propose an instance segmentation method based on Mask Scoring region-based convolutional neural networks (R-CNN) (MS R-CNN), which creates the adversarial network called MaskDis in the head branch of MS R-CNN. The MaskDis is trained as a discriminator using a generative adversarial network, and the MS R-CNN model is used as a generator during model training. The adversarial training enables the generator to learn context information and features at the pixel level, which effectively improves the segmentation quality under pigs' overlapping or dense occlusions scenes. Experimental conducted on the pig object segmentation dataset show that the proposed approach achieves a precision of 92.03%, a recall of 92.18%, and an F1 score of 0.9210. Compared with the basic MS R-CNN model, the approach achieved a 2.25% improvement in precision and 1.18% improvement in F1 score. Furthermore, the improved approach outperformed advanced instance segmentation methods such as YOLACT, Swin Transformer, YOLOv5-seg, and SOLOv2 on COCO evaluation metrics. These experimental results demonstrate the effectiveness of the proposed approach in instance segmentation of pigs in complex scenes, providing technical support for non-contact pig automatic management.

1 | INTRODUCTION

The pig farming industry plays an important role in the national economy, and the main direction of the pig farming industry is large-scale and intelligent pig farming. In large-scale intelligent pig breeding, accurate and timely collection of phenotype information for controlling pig growth is the key technology for precision farming [1–3]. Observing animals on an individual level to assess their health and welfare is necessary. However, on a real-world commercial farm, observing pig contour information by humans is impractical, and subjective human observation can lead to errors [4]. Techniques based on deep learning have been adopted in the last few years, achieving excellent performance in many fields, image inpainting, natural language processing, and so on [5–7]. There have been successful uses of deep learning algorithms for acquiring pig information, providing an efficient, contactless, and non-destructive intelligent method [8–10]. However, objective factors in the pig farm environment, such as light fluctuations, high similarity, and pig adhesion, result in decreased accuracy of detection and segmentation, which fails to meet practical applications in pig farming. Therefore, designing and developing an accurate and efficient segmentation algorithm for large-scale and intelligent pig farming is important. The ability to automatically detect and segment the contour of individual pigs can assist in the early detection of potential health or welfare problems without the need for human observation.

Due to the development of deep learning technology, the performance of instance segmentation has been significantly improved. The classical methods rely on object detectors, such

as Faster region-based convolutional neural networks (R-CNN) [11] by detecting objects, obtaining bounding boxes, and then extracting region features for pixel-wise segmentation using techniques like RoIPooling or RoI-Align. Mask R-CNN [12] is a representative instance segmentation algorithm that adds a mask branch on the Faster R-CNN to predict the mask of the object. Mask Scoring R-CNN [13] modifies the scoring strategy for mask segmentation based on Mask R-CNN and Cascade R-CNN [14] progressively improves object localization using cascade structure to achieve more accurate mask prediction. In addition, a series of instance segmentation algorithms are derived from the single-stage fully convolutional one-stage object detection (FCOS) [15] as the target detection framework. MEInst [16] and CondInst [17] extend FCOS by predicting the encoding mask vector or mask kernel for dynamic convolution [18]. These methods have achieved satisfactory performance in instance segmentation and effectively promote the development of instance segmentation. However, these methods encountered boundary leakage issues, where they failed to properly segment instances of individuals that overlapped within the same class, and the resulting mask segmentation lacked smoothness in its details.

Instance segmentation methods have been widely used in livestock welfare analyses. For example, cattle instances were segmented from real animal farms using these methods, achieving an average accuracy of 92% [19]. Mask R-CNN based on a dual attention guided feature pyramid network was introduced for instance segmentation of group-housed pigs [20], effectively segmenting individual pigs and achieved an average precision (AP) of 93.1%. In addition, our previous study applied Soft-NMS to Mask R-CNN for instance segmentation of pigs with complex backgrounds [21], achieving a harmonic average (F1) of 93.74%. Mask R-CNN combined with a support vector machine (SVM) classifier to identify individual cows and achieved an accuracy of 98.67% for cows [22]. However, the above methods do not consider the diversity of samples and objective factors such as complex light variations, occlusion, adhesion of pigs, and complex backgrounds. Moreover, adversarial networks can be used to significantly improve the performance of the model during training using automatic annotation of samples, which can solve the problems of instance segmentation mentioned above.

Adversarial networks [23] have become popular algorithm because it is capable of learning data distributions without relying on annotations. And its performance can be significantly improved if annotations are used in the training. Adversarial networks were applied to semantic segmentation [24], which detects and corrects higher-order inconsistencies between segmentation maps generated by segmentation networks and the ground truth (GT) segmentation maps. It had also been used for the segmentation of medical images [25], overcoming the limitation of classical adversarial network discriminators, which provide a single scalar true/false output, by generating stable and sufficient gradient feedback for the network. In addition, adversarial networks have been applied to image in painting. Repair network and optimization network (RNON) is an efficient image in painting method consisting of two mutually independent generative adversarial networks, with one network functioning as an image in painting network and the other as an image optimization network [26]. Therefore, it can be widely used in many fields such as image segmentation [27, 28], image classification [29, 30], and so on [31–33].

To address the issue of unsatisfactory segmentation performance in scenarios involving pig overlapping and occlusions, this paper proposed an instance segment network model combining adversarial network named MaskDis with the basic MS R-CNN model. Firstly, the MaskDis is used as a discriminator and the mask head of MS R-CNN is used as a generator; they are trained using an adversarial training approach. After adversarial learning between the generator and the discriminator, the generator can learn pixel-level, low-level, and mid-level features, as well as context information for better segmentation performance. Finally, adversarial training makes the prediction mask close to the GT, resulting in improved segmentation quality in complex scenarios.

For this paper, the main contributions are as follows: (1) we proposed an improved instance segmentation algorithm to enhance the segmentation quality by fusing the adversarial network in the MS R-CNN model. (2) We designed an adversarial network (MaskDis) model achieving better performance of instance segmentation under pig overlapping and occluded scenarios. (3) We completed experimental validation with a variety of advanced instance segmentation algorithms on the pig segmentation dataset and proved that our method has better segmentation performance.

2 | METHODS

2.1 | Overall framework of the instance segmentation algorithm

The structure of MaskDis R-CNN (shown in Figure 1) includes two components: MS R-CNN and MaskDis Head. The method firstly extracted the feature maps from the input images using a backbone network of ResNet-101and Feature Pyramid Networks (FPN), and the resulting feature maps are then fed to the Region Proposal Network (RPN) to generate Region of Interests (RoIs). Secondly, the RoiAlign layer fixed the ROIs to the same size and then fed them to the Fully Connected layers (FC layers) used for classification and detection and Fully Convolutional Network (FCN) used for segmentation. The prediction mask, the image of individual pig, and the GT are simultaneously fed into the adversarial network head (MaskDis head) for adversarial training, and then back-propagated to the generator(MS R-CNN) by the multiscale loss function.

The following subsections illustrate the process of the MS R-CNN and MaskDis head model.

2.2 | The basic framework of MS R-CNN

The basic framework of MS R-CNN consists of the following four components:

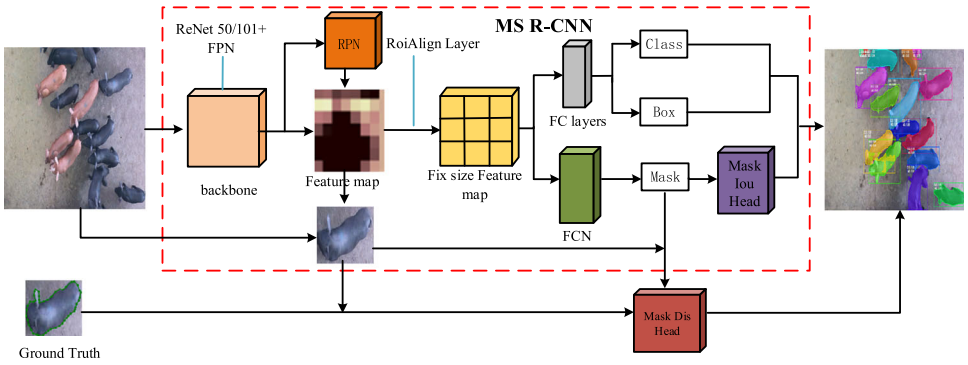

FIGURE 1 The structure of MaskDis R-CNN.

Backbone. ResNet-101 and FPN are a network structure designed for multi-scale feature extraction. The ResNet-101 consists of multiple residual blocks, which effectively reduces the number of parameters in the convolution process and prevents the degradation of the network due to the increase in network depth. ResNet-101 generates five different scale feature maps, which are fed into the FPN. In FPN, the pre-processed images are firstly extracted by the bottom-up forward process to get feature maps of different scales, and then the feature maps P6 are firstly obtained by down-sampling, and then the corresponding feature maps are fused by top-down up-sampling to get multi-scale fused. The features obtained by fusion have more robust semantic information, and also effectively improve the speed and the accuracy of detection.

Region Proposal Networks. RPN are efficient in generating regions of interest because of the advantages of fast candidate region generation and low computational cost, the input of RPN networks is the feature map extracted from the backbone network, and the output is a batch of candidate frames and region scores, enabling end-to-end training. The anchor mechanism was used in the RPN network. The anchor is in the form of a 3×3 sliding window ($n = 3$) over each layer of the input feature map, generating k region boxes of different sizes and proportions at the centre of the sliding window.

Mask Head. The output of Mask Head includes classification prediction, bounding-box regression, and instance segmentation mask prediction. The classification branch and bounding-box regression branch share the features extracted in the first stage, including pooling the RoIs by RoIAlign, followed by extracting the deep features of the RoI feature map by two FC layers, after which they start to divide into two branches and perform one full connection each and then output the results. The principles of classification regression and bounding-box regression are the same as those of classification and border regression in RPN. The RoI feature map used in the segmentation branch of the example is independent of the two branches mentioned above. First, RoI is processed by RoIAlign to obtain a $14 \times 14 \times 256$ RoI feature map, where 14×14 represents the pixels of the feature map and 256 represents the number of channels of the feature map. Then after passing through four convolutional layers comprising the FCN, one layer of deconvolution, and one layer of convolution, the final result of instance segmentation with a scale size of 28×28 is generated.

MaskIoU Head. The input features of the head branch of MaskIoU Head are obtained from Mask Head. Then, the MaskIoU values are obtained through the calculation of four convolutional layers and three FC layers. In the four convolutional layers, the first layer uses a convolutional kernel with a size of $3 \times 3 \times 257$, and the remaining three layers use convolutional kernels with a size of $3 \times 3 \times 256$. In the three FC layers, the output of the first two layers is 1024, and the output of the last layer is the number of categories (set to 2 in the experiment).

2.3 | MaskDis Head

Figure 2 shows the structure of the adversarial network of the herd pigs instance segmentation model. The adversarial network head branch, called MaskDis Head, is added to the MS R-CNN model. And Mask Head is used as the generator and MaskDis Head as the discriminator during model training.

Firstly, True/False samples are obtained by dot-multiplying the true mask or the predicted mask with the RoI of the original image, respectively. Then, these True/False samples are used as input data and fed into the network for computation. The input data consists of pixel-level information with dimensions of $28 \times 28 \times 3$. The output of the first layer has dimensions of $14 \times 14 \times 64$, and the output of the second layer has dimensions of $7 \times 7 \times 128$. The network structure comprises two convolutional layers, each with a 5×5 convolutional kernel size, 64 and 128 channels, and a stride of 2. Finally, the pixel-level features extracted from the first convolutional layer output, and the features from the second convolutional layer are combined to form a one-dimensional vector. Hierarchical features are extracted from multiple layers of the MaskDis Head to compute the multi-scale L1 loss. This loss effectively captures both

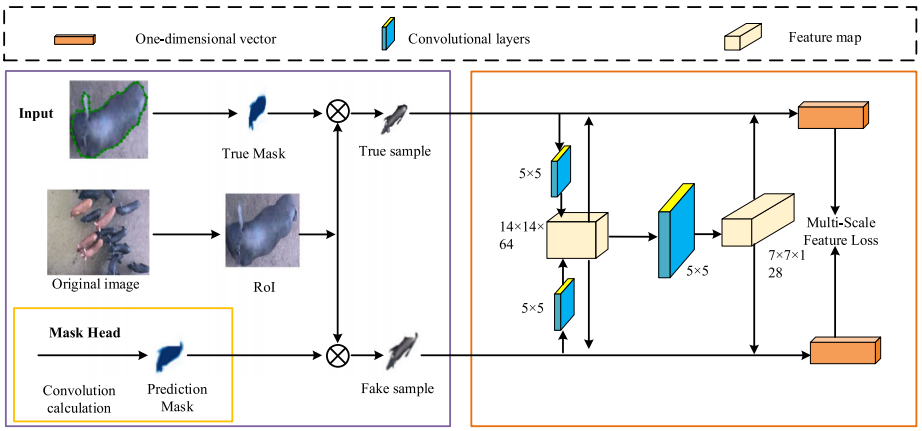

FIGURE 2 The architecture of the MaskDis head.

ALGORITHM 1

Input: *Truemask*,*RoI*,*Predictionmask*; Output: *Loss*

1. $Truesample \leftarrow Truemask \otimes RoI$
2. $Featuremap1\_T \leftarrow Convolution\_layer1(Truesample, (5,5), 64, 2)$
3. $Featuremap2\_T \leftarrow Convolution\_layer2(featuremap1\_T, (5,5), 128, 2)$
4. $Fakesample \leftarrow Predictionmask \otimes RoI$
5. $Featuremap1\_P \leftarrow Convolution\_layer1(Fakesample, (5,5), 64, 2)$
6. $Featuremap2\_P \leftarrow Convolution\_layer2(featuremap1\_P, (5,5), 128, 2)$
7. $Loss \leftarrow \frac{1}{N}\sum_{n=1}^{N} smoothL1_{loss}(fc(Truesample, Featuremap1\_T, Featuremap2\_T), fc(Fakesample, Featuremap1\_T, Featuremap2\_T))$

Return: *Loss*

long- and short-range spatial relations between pixels by utilizing hierarchical features, including pixel-level, low-level, and mid-level features.

The generator and discriminator networks are trained alternately in an adversarial manner: the Mask Head is trained to minimize the multi-scale L1 loss, while the MaskDis Head is trained to maximize the same loss function. This adversarial training improves the quality of segmentation as the generator and discriminator learn pixel-level features, low-level features, and mid-level features. The relevant pseudo-code is presented in the table, and the improvements are described in Algorithm 1.

More details including number of feature maps used in each convolutional layer can be found in Figure 2.

In MaskDis Head, the generator generates n predicted masks denoted as $x_n$and the corresponding original map RoI and true masks denoted as r_n and y_n, respectively, and the Multi-Scale Feature Loss function L_{Dis} is defined as

$$\min_{\theta_G}\max_{\theta_D} L_{Dis}(\theta_G, \theta_D) = \frac{1}{N}\sum_{n=1}^{N} smoothL1_{loss}\left(fc(r_n \cdot x_n), fc(r_n \cdot y_n)\right) \quad (1)$$

θ_G and θ_D represent the parameters for the generator and the discriminator. $r_n \cdot x_n$is the result of the original map RoI and the predicted mask dot product, and $r_n \cdot y_n$ is the result of the original map RoI and the GT mask dot product. The formula $smoothL1_{loss}(T_{pred}, T_{gt})$ is defined as follows:

$$smoothL1_{loss}\left(T_{pred}, T_{gt}\right) = \begin{cases} (T_{pred} - T_{gt})^2/2, if \left|T_{pred} - T_{gt}\right| < 1 \\ \left|T_{pred} - T_{gt}\right| - 1/2, otherwise \end{cases} \quad (2)$$

In the pig instance segmentation algorithm for fusion adversarial networks, the RoI head branch consists of three parts, Mask Head, MaskIoU Head, and MaskDis Head. L_{cls} is the loss value for classification regression, L_{box} is the loss value for border regression, L_{mask} is the loss value for instance segmentation generator, L_{IoU} is the loss value for MaskIoU regression, and L_{Dis} is the loss value for MaskDis Head regression. When training the mask generator, the loss function of the RoI branch is defined as follows:

$$L_{RoI} = L_{\text{cls}} + L_{box} + L_{mask} + L_{IoU} + L_{Dis} \quad (3)$$

3 | RESULTS AND DISCUSSION

In this section, MaskDis R-CNN is compared with Mask R-CNN and MS R-CNN. In addition, it is evaluated in the same experimental configuration: (1) comparison of experimental results in front and top views; (2) comparison of results on COCO evaluation metrics; (3) comparison in segmentation quality and scoring; and (4) comparison with other advanced instance segmentation methods.

3.1 | Experimental parameters and evaluation indicators

A device configuration is established for the implementation of the proposed approach, which consists of Python 3.7

FIGURE 3 Part of the dataset.

TABLE 1 distribution of the dataset.

Dataset	Front view	Top view	Total
Train	105	75	180
Test	65	70	135
Total	170	145	315

software PyTorch running on a PC with AMD Ryzen5 2600X and 3.00 GHz processor, 64 GB RAM and NVIDIA GeForce RTX TITAN X GPU with 12GB GPU VRAM.

Experimental data were collected randomly over 10 days of videos, containing 7 h of video per day, from 9:00 to 16:00 in the 'Lejiazhuang Pig Farm', Foshan City, Guangdong Province. And they were recorded by FL3-U3-88S2C-C cameras. The experimental data were saved in the audio video interleaved format with a video frame rate of 25 fps. To obtain adequate and better images, we focused on the five pens from the top and front view angles. And the size of the pens was 7 m × 5 m × 3 m (m represents the unit of length), respectively, and the number of pigs in each pen ranged from 3 to 20. Part of the training set and test set data are shown in Figure 3.

After obtaining the video data, a total of 315 images were selected according to a certain ratio, and these images were divided into the train dataset and test dataset. The detailed information of the dataset is shown in Table 1. In total, 180 images were selected as the train dataset, including 105 images collected by top view and 75 images by front view. And 135 images were selected as the test dataset, including 65 images collected by front view and 70 images by top view. Labelling the 315 images with including 3423 pig objects cost about 135 person-hours. Moreover, all animal experiments were conducted following the guidelines provided by the Guangdong Provincial Laboratory Animal Welfare and Ethical Review Guidelines and were approved by the Animal Welfare Committee of South China Agricultural University (No: 2021F129).

Data pre-processing uses data augmentation, which includes random horizontal flip, random brightness adjustment, random contrast adjustment, random saturation adjustment, and random hue adjustment. The Resnet-101 and FPN is used as Backbone, where FPN uses layers 2 to 5 of the Resnet-101 network, and the output of FPN has 256 channels. The number of foregrounds retained after post-processing NMS is adjusted in RPN, and its size is set to 1000. The learning rate is set to 0.0025, the epoch is set to 90, the batch size is set to 2, the scale of degradation is 0.1, and the value of weight decay is set to 0.0001.

To analyze the quality of segmentation results, we used recall, precision, F1 scores, and Precision-Recall (P-R) curves as evaluation metrics. AP is the area under the P-R curve and IoU is the degree of overlap between the predicted bounding box and the GT, which can be used to summarize the performance of an object detection model.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{4}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{5}$$

$$\text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{6}$$

$$\text{AP} = \int_0^1 \text{Precision} \cdot \text{Recalldr} \tag{7}$$

$$\text{IoU} = \frac{\text{detection result} \cap \text{ground truth}}{\text{detection result} \cup \text{ground truth}} \tag{8}$$

where True Positive (TP) is the number of pixels correctly predicted to be pig category, False Positive (FP) is the number of pixels incorrectly predicted to be pig category, False Negative (FN) is the number of pixels predicted to the pig category as the background, and F1 is a comprehensive evaluation

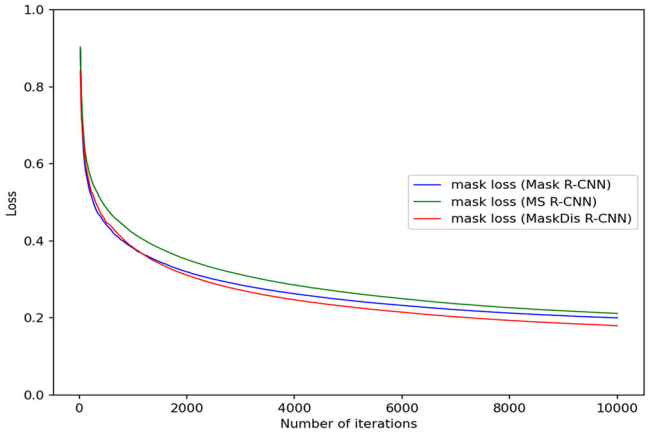

FIGURE 4 Comparison of training loss iteration curves of the model.

metrics of precision and recall rate. In our study, we adopted the standard COCO style AP0.5:0.95 metric, which computes the AP across various IoU thresholds ranging from 0.5 to 0.95 with an interval of 0.05. Additionally, we calculated the AP0.5 and AP0.75 metrics, which provide the AP values for different IoU thresholds. Moreover, we computed AR0.5 and AR0.75, representing the average recall (AR) values for different IoU thresholds, and obtained the mean average recall (mAR) by averaging the AR values at each IoU threshold. The IoU metric serves to assess the accuracy of object detection by measuring the overlap between the predicted bounding boxes and the GT. It is computed as the area of overlap between the two boxes divided by the area of their union.

3.2 | Experimental results

The comparative training loss iteration curves of the three instance segmentation models are shown in Figure 4. The blue, green, and red curves are the segmentation mask loss values of Mask R-CNN, MS R-CNN, and MaskDis R-CNN, respectively. According to Figure 4, each curve starts to converge at approximately 2000 iterations, after which the gap between the loss values of the Mask R-CNN model and the MS R-CNN model gradually decreases, while the gap between the loss values of both and MaskDis R-CNN model gradually increases. In general, the loss value of the MaskDis R-CNN is relatively small, indicating that the segmentation model of the fusion the adversarial network convergence more easily.

Table 2 shows the statistics of image detection results and comparisons. The number of pigs is the total of individual pigs in the test dataset, the test number is the total of individual pigs detected by the model, and the correct number is the total of individual pigs correctly detected by the model in the test number. The MaskDis R-CNN model detected a total of 1204 group-housed pigs, of which 1108 pigs are correctly detected, with a slight improvement in recall to 92.18%, precision increased by 2.25% to 92.03%, and F1 score increased by 0.0118 to 0.9210 when compared with the MS R-CNN model. The model shows a larger improvement on the test set from the front view, where the number of detections is 494, the number of correctly detected is 413. The F1 score increased by 2.56% to 84.63%. In the test set from the top view, the number of pigs detected is 710, the number of correct detections is 695, the recall rate is 96.53%, the precision rate increased by 1.46% to 97.89%, and the F1 score increased only slightly. The MaskDis R-CNN performs better in Precision and F1 values than the MS R-CNN. Thus, the MaskDis R-CNN demonstrates improvements in recall and precision, indicating an improved segmentation quality compared to the Mask R-CNN and MS R-CNN models.

The P-R curve (PR curve) of the three models is shown in Figure 5, with an IoU threshold of 0.75 for the PR curve evaluation criterion. The blue line represents the PR curve of Mask R-CNN, the red line represents the PR curve of MS R-CNN, and the green line represents the PR curve of the MaskDis R-CNN. Figure 5 shows that the green line is closest to the right and covers the largest area, indicating better segmentation performance. Based on the COCO evaluation metric, higher scores for correctly detected results lead to a larger area under the PR curve and higher average accuracy rates. The blue triangle in Figure 5 represents the intersection point of the three curves. To the left of the intersection point, the distance between the

TABLE 2 Statistics of image detection results and comparison.

Model	Image type	Number of pigs	Number of tests	Number of corrects	Recall (%)	Precision (%)	F1	Time (s)
	Front view	482	608	423	87.76	69.57	0.7761	
Mask R-CNN	Top view	720	707	690	96.83	97.60	0.9671	0.284
	Total	1202	1315	1113	**92.60**	84.64	0.8844	
	Front view	482	505	405	84.02	80.20	0.8207	
MS R-CNN	Top view	720	728	702	97.50	96.43	0.9696	0.286
	Total	1202	1233	1107	92.10	89.78	0.9092	
	Front view	482	494	413	85.68	83.60	0.8463	
MaskDis R-CNN	Top view	720	710	695	96.53	97.89	0.9720	0.288
	Total	1202	1204	1108	92.18	**92.03**	**0.9210**	

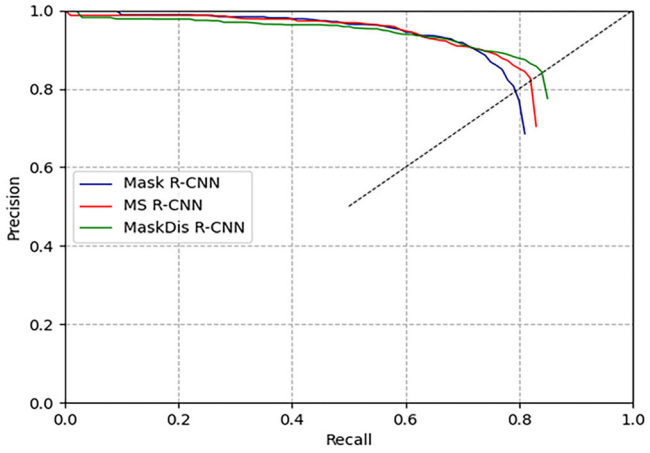

FIGURE 5 Precision-recall curve of herd pig instance segmentation model with fusion adversarial network.

three curves is relatively similar, but to the right of the intersection point, the distance between the curves gradually increases. Due to the fusion of adversarial networks, the quality of detection and segmentation results is improved, resulting in higher recall and precision values.

Based on Figure 5, MaskDis R-CNN demonstrates superior segmentation effectiveness compared to the other two models. Figure 6 illustrates the segmentation results achieved by MaskDis R-CNN. Each detection box in the upper left corner is labelled with classification (CLS) and MS, representing the classification and segmentation quality scores, respectively. Based on the classification scores in Figure 5, the pigs exhibit a CLS of 1.00, with an average MS score exceeding 0.9. When dealing with densely packed and closely connected pigs, MaskDis R-CNN achieves more comprehensive pig segmentation, displaying smooth segmentation boundaries without fragmentation or missed segments.

3.3 | Comparison of results on the COCO evaluation metrics

The results of the three models' segmentation tasks on the COCO evaluation metrics are shown in Table 3. We compare the three models in the front view, the top view, and the total of front view and top view. According to Table 3, the proposed MaskDis R-CNN method performs better in instance segmentation of objects. The performance of MaskDis R-CNN reaches 96.09(AR_{50}), 85.36(AR_{75}), 73.36(mAR), 92.76(AP_{50}), 80.76(AP_{75}), and 68.55(mAP), has a significant promotion compared to MS R-CNN. Also, compared with MS R-CNN, the MaskDis R-CNN increases by 2.25%, 2.28%, 1.37%, and 1.9% in the metrics of AR_{75}, mAR, AP_{75}, and mAP, respectively. Therefore, our method is validated in improving the segmentation performance.

The ablation experiments of our method are shown in Table 4 under the COCO evaluation metrics. The best performance is achieved using Resnet-101 as the backbone network and adopting Stochastic Gradient Descent (SGD) for the optimizer, as shown in Table 4. Our approach's AP50, AP75, mAP, and mAR values were 92.8%, 80.8%, 68.6%, and 73.4%, respectively. Therefore, we use this configuration to compare it with other advanced instance segmentation methods.

3.4 | Comparison of segmentation quality between MaskDis R-CNN and MS R-CNN

The segmentation result of the MS R-CNN model is shown at the top of Figure 7. The segmentation result of the MaskDis R-CNN model is shown at the bottom of Figure 7. According to Figures 7a and 7b, the segmentation quality of the MS R-CNN model is flawed in the case of dense overlap of pigs, resulting in the segmentation target's MS score falling below 0.7, and

FIGURE 6 The segmentation result of the MaskDis R-CNN model.

TABLE 3 Results of the model under the COCO evaluation metrics.

Model	Image type	AR_{50} (%)	AR_{75} (%)	*mAR* (%)	AP_{50} (%)	AP_{75} (%)	*mAP* (%)
Mask R-CNN	Front view	89.00	64.73	56.64	84.05	55.03	49.86
	Top view	97.78	93.06	78.40	96.60	92.17	75.79
	Total	94.26	81.7	69.68	92.08	77.68	65.67
MS R-CNN	Front view	91.49	65.35	58.42	82.68	55.43	50.35
	Top view	99.03	95.00	79.56	98.31	93.34	76.80
	Total	96.01	83.11	71.08	**93.09**	79.39	66.65
MaskDis R-CNN	Front view	92.12	70.12	61.76	82.84	58.61	52.76
	Top view	98.75	95.56	81.13	97.24	93.98	77.95
	Total	**96.09**	**85.36**	**73.36**	92.76	**80.76**	**68.55**

TABLE 4 ablation experiments under the COCO evaluation metrics.

Model	Backbone	Optimizer	Image type	AP_{50} (%)	AP_{75} (%)	*mAP* (%)	*mAR* (%)
MaskDis R-CNN	R50	Adam	Front view	84.8	47.4	47.1	57.0
			Top view	97.3	91.0	76.3	78.9
			Total	91.5	75.1	65.4	69.7
		SGD	Front view	79.2	49.6	52.4	55.8
			Top view	97.8	91.6	76.9	78.4
			Total	91.2	75.7	68.0	69.3
	R101	Adam	Front view	78.4	49.7	46.1	55.0
			Top view	97.6	93.5	77.2	79.5
			Total	91.9	77.1	65.9	69.7
		SGD	Front view	82.9	58.6	52.8	61.8
			Top view	97.2	94.0	78.0	81.1
			Total	**92.8**	**80.8**	**68.6**	**73.4**

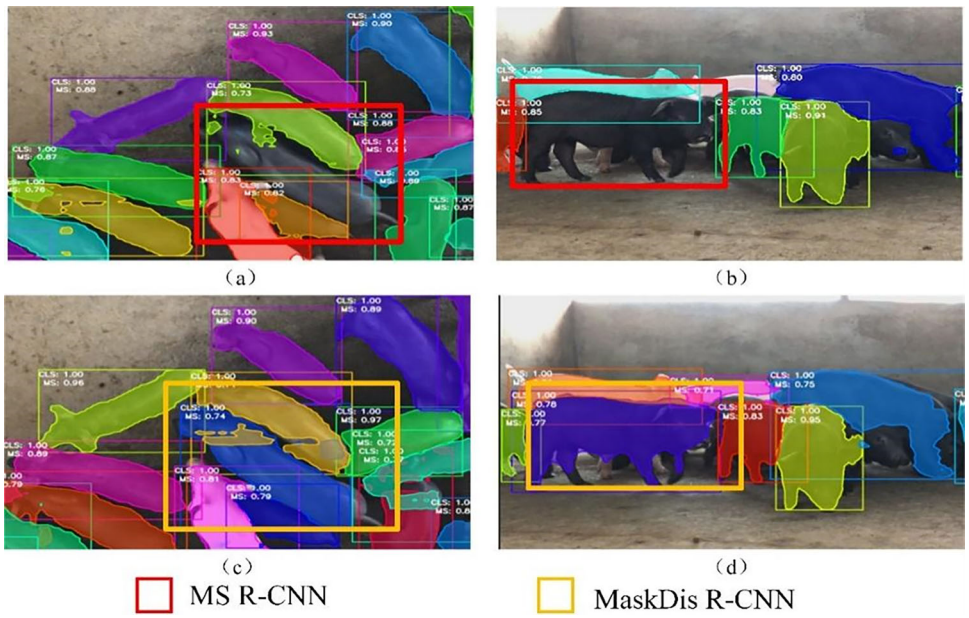

FIGURE 7 Comparison of the segmentation quality score between MaskDis R-CNN model and MS R-CNN model. MS R-CNN, mask scoring R-CNN.

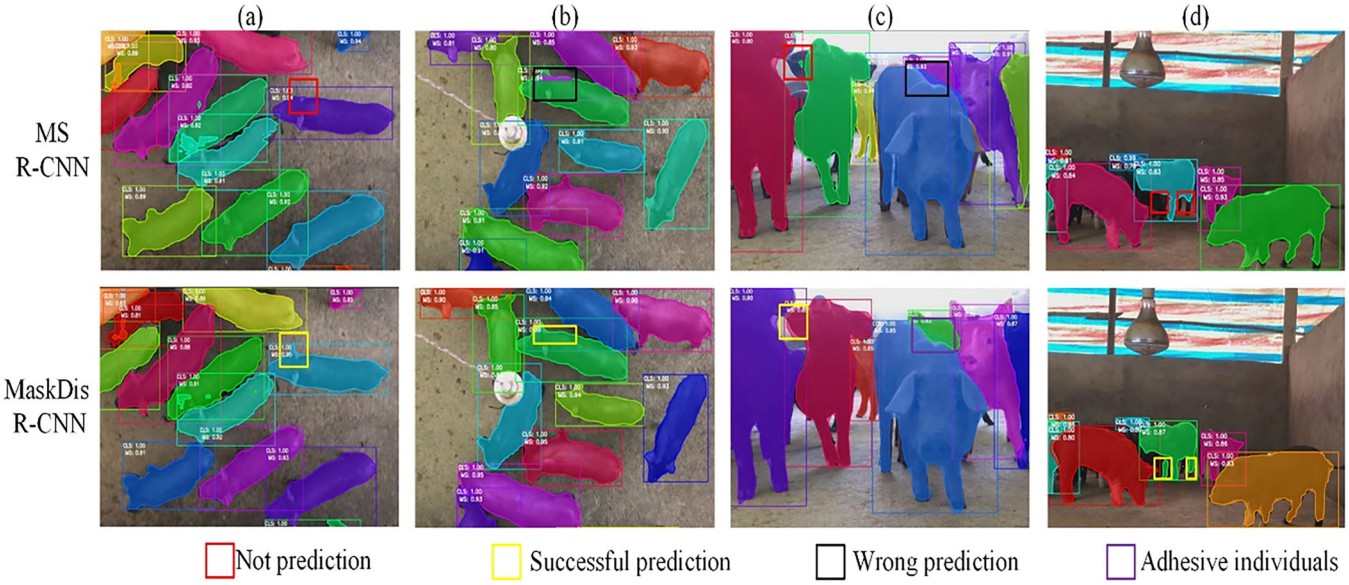

FIGURE 8 Comparison of front view and top view segmentation quality between MaskDis R-CNN model and MS R-CNN model. MS R-CNN, mask scoring R-CNN.

therefore the model has missed detection. According to Figures 7c and 7d, the model with the fused adversarial network has improved segmentation quality, increasing the MS score of the segmentation results, thus avoiding missed detections to some extent and improving the recall rate.

To further demonstrate the effectiveness of our proposed MaskDis R-CNN, we compare the segmentation quality of the MaskDis R-CNN model and the MS R-CNN model from the top and front views. The segmentation result of the MS R-CNN model is shown at the top of Figure 8, and the segmentation result of the MaskDis R-CNN model is shown at the bottom of Figure 8. According to Figure 8, the main problem of MS R-CNN is the incomplete segmentation of pig bodies (Figures 8a, 8c, 8d), and a small number of segmented fragments (Figure 8b) appear on other pig bodies in the case of dense and sticky pig populations. And the segmentation boundary is not smooth enough (Figure 8c). However, there are improvements in the segmentation results of the MaskDis Mask method with the addition of the adversarial network, such as improved quality of both segmentation details and segmentation boundaries, more complete segmented pigs, and fewer segmented fragments.

3.5 | Results comparison with other advanced instance segmentation methods

We used the proposed model to conduct comparison experiments with the four advanced instance segmentation methods, including YOLACT [34], Swin Transformer [35],YOLOv5-seg, and SOLOv2 [36] on the same dataset. The YOLACT is a real-time instance segmentation model that combines detection and segmentation, achieving 33.5 fps on the MS COCO dataset. The Swin Transformer is a transform-based instance segmentation method that introduces a hierarchical Transformer whose representation is computed with Shifted windows to better capture global contextual information. The YOLOv5-seg is a high-speed instance segmentation algorithm that keeps accuracy while achieving real-time performance. The SOLOv2 is a one-stage instance segmentation method that combines detection and segmentation, which uses deformable convolution and a feature selection module with an attention mechanism that can better adapt to the shape and scale variations of the instances.

The comparison results of our approach with other advanced instance segmentation methods are shown in Table 5. Our approach achieved the best performance on the metrics of COCO. The AP_{50}, AP_{75}, mAP, and mAR values of our approach were 92.8%, 80.8%, 68.6%, and 73.4%, respectively. Compared with the YOLACT method, our approach improved by 4.9%, 12.1%, 8.9%, and 5.9% in AP_{50}, AP_{75}, mAP, and mAR. Compared with the Swin Transformer method, our approach improved by 2.7%, 20.7%, 16.3%, and 10.6% in AP_{50}, AP_{75}, mAP, and mAR. Compared with the SOLOv2 method, our approach improved by 0.5%, 6.9%, and 2.9% in AP_{50}, AP_{75}, mAP. Compared with the YOLOv5-seg method, our approach improved by 2.7%, 2.5%, and 5.5% in AP_{50}, AP_{75}, and mAP. These comparison results demonstrate that our approach can effectively improve the pig segmentation performance.

4 | DISCUSSION

In this paper, we propose MaskDis R-CNN, a deep learning algorithm by fusing adversarial networks, for the detection and instance segmentation of herd pigs in complex scenarios. The key innovations include designing adversarial networks, integrating them into MS R-CNN, and then proving their

TABLE 5 Results comparison with other advanced instance segmentation methods.

Model	Image type	AP_{50} (%)	AP_{75} (%)	*mAP* (%)	*mAR* (%)
YOLACT	Front view	77.8	47.0	44.7	55.9
	Top view	93.8	84.5	70.5	75.3
	Total	87.9	68.7	59.7	67.5
Swin transformer	Front view	83.9	40.8	42.4	54.6
	Top view	94.9	77.3	61.2	68.3
	Total	90.1	60.1	52.3	62.8
YOLOv5-seg	Front view	80.2	66.8	53.6	–
	Top view	96.9	86.7	73.1	–
	Total	90.1	78.3	63.1	–
SOLOv2	Front view	83.5	54.8	51.8	63.3
	Top view	97.7	86.7	75.3	80.6
	Total	92.3	73.9	65.7	**73.7**
MaskDis R-CNN	Front view	82.8	58.6	52.8	61.8
	Top view	97.24	94.0	78.0	81.1
	Total	**92.8**	**80.8**	**68.6**	73.4

effectiveness, for instance, segmentation tasks of group-housed pigs. The advantage of the MaskDis R-CNN method is the ability to detect and segment instances of heavily obscured and overlapping pigs, which can be further developed to perform tasks such as pig welfare monitoring [37–39].

Previous instance segmentation studies in pigs were challenged by occlusions, light variations, and background factors [21]. To address this issue, adding adversarial network can enhance the quality of instance segmentation, as the training process uses adversarial training to make the prediction mask closer to the GT in order to achieve enhanced segmentation quality. In our work, the adversarial network head branch is designed and added to the MS R-CNN, called MaskDis Head, which is used as a discriminator, and Mask Head is used as a generator during model training. Through the adversarial training of the generative network, the generator learns pixel-level, low-level, and middle-level features. The segmentation quality is improved as the function of the segmentation mask loss converges more easily during model training.

Mask R-CNN and MS R-CNN are efficient methods in both the top and front views. However, it does not achieve the expected results in the situation of the overlapped pigs. When there is a problem caused by dense overlap and severe occlusion, the problem is more severe in front views. To get better segmentation results, the MaskDis R-CNN instance segmentation model creates the adversarial network branch for guiding the segmentation mask training of the model. The improved model is compared with the MS R-CNN model for experiments and analysis, and the results show that MaskDis R-CNN improves the segmentation quality. In the comparative analysis of the segmentation results between MaskDis R-CNN and the MS R-CNN model, it is found that the MaskDis R-CNN improves the quality of instance segmentation on the situation of the overlapped pigs, which indicates that the adversarial network branch of MaskDis R-CNN improves the ability to handle the detail information of instance segmentation.

The final experimental results have shown that the MS R-CNN model can achieve detection accuracy with a recall of 92.10% and a precision of 89.78%. And this method sometimes misses the detection and does not work well for segmentation of herd pigs in the case of dense and adhesive. However, the MaskDis R-CNN can void this situation and improve the quality of intensive herd raised piglets with a precision of 92.03% and an F1 score of 92.10%. In conclusion, the MaskDis R-CNN significantly outperforms MS R-CNN for instance segmentation in a dense environment of pig pens.

5 | CONCLUSION

To solve the problem of unsatisfactory segmentation performance under overlapping or partial occlusion between pig's scenes, we propose an approach named MaskDis R-CNN by fusing the MS R-CNN model and the adversarial network. The new method solves the problem that MS R-CNN does not achieve the expected results in specific situations such as pig dense and occlusion situations. Mask Head is used as a generator, and MaskDis Head as a discriminator during model training. Adversarial training of the generated network keeps the prediction mask close to GT for improving segmentation quality during model testing.

We conducted comparative experiments between Mask R-CNN, MS R-CNN, and our proposed approach to demonstrate our method's effectiveness in the overlapped pigs and occlusion situation. Our method outperforms the other two approaches, which can achieve a recall of 92.18%, a precision of 92.03%, and an F1 score of 0.9210. In addition, by adopting the COCO evaluation metrics, our approach achieves an mAP of 73.36% and an mAR of 68.55%, which are higher than the other two algorithms. In addition, our method performance is better than other advanced instance segmentation methods. Considering these results, it is demonstrated that the proposed method improves the problem of poor segmentation quality due to the dense and occlusive herd of pigs.

Although this paper achieves accurate segmentation, there are still some constraints: (1) the proposed model cannot achieve the characteristics of lightweight, fast speed, and strong portability, which requires operations such as convolution and ROI pooling on each candidate region. This leads to higher computational and memory requirements. (2) In the case of severe occlusion, the proposed model still suffers from missed and false detections, resulting in segmentation performance that does not meet practical needs. Future work will optimize the network architecture of the proposed approach, which can reduce the computational requirements and increase the inference speed. The improved model also provides a theoretical basis for the intelligent development of pig farming and has great significance for improving pig welfare and guiding the production.

AUTHOR CONTRIBUTIONS

Shuqin Tu:Writing—Review & Editing, Methodology, Conceptualization, Funding Acquisition, Resources, Formal Analysis, **Qiantao Zeng**:Writing—Original Draft, Validation, Methodology, Visualization, Writing—Review & Editing, **Haofeng Liu**:Conceptualization, Formal Analysis, Funding Acquisition, **Yun Liang**:Supervision, Funding Acquisition, Project Administration, **Xiaolong Liu**: Resources, Visualization, Investigation, **Lei Huang**: Resources, Data Curation, Visualization, **Zhengxin Huang**: Resources, Data Curation.

ACKNOWLEDGEMENTS

This research was supported by the National Natural Science Foundation of China (61772209 and 31600591), the Science and Technology Planning Project of Guangdong Province (Grant No. 2019A050510034), Guangzhou Key R&D Program Project (2023B03J1363, 202206010091), and College Students' Innovation and Entrepreneurship Competition (202110564025).

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

ORCID

Yun Liang https://orcid.org/0000-0003-0799-0054

REFERENCES

1. Wongsriworaphon, A., Arnonkijpanich, B., Pathumnakul, S.: An approach based on digital image analysis to estimate the live weights of pigs in farm environments. Comput. Electron Agr. 115, 26–33 (2015)
2. Wang, K., Guo, H., Ma, Q., Su, W., Chen, L., Zhu, D.: A portable and automatic Xtion-based measurement system for pig body size. Comput. Electron. Agric. 148, 291–298 (2018)
3. Bhoj, S., Tarafdar, A., Chauhan, A., Singh, M., Gaur, G.K.: Image processing strategies for pig liveweight measurement: Updates and challenges. Comput. Electron. Agric. 193, 106693 (2022)
4. Nasirahmadi, A., Edwards, S.A., Sturm, B.: Implementation of machine vision for detecting behaviour of cattle and pigs. Livestock Sci. 202, 25–38 (2017)
5. Chen, Y., Xia, R., Zou, K., Yang, K.: FFTI: Image inpainting algorithm via features fusion and two-steps inpainting. J. Visual Commun. Image Represent. 91, 103776 (2023)
6. Chen, Y., Xia, R., Yang, K., Zou, K.: MFFN: Image super-resolution via multi-level features fusion network. Vis. Comput. (2023)
7. Patil, R., Boit, S., Gudivada, V.N., Nandigam, J.: A survey of text representation and embedding techniques in NLP. IEEE Access 11, 36120–36146 (2023)
8. Kim, J., Chung, Y., Choi, Y., Sa, J., Kim, H., Chung, Y., Park, D., Kim, H.: Depth-based detection of standing-pigs in moving noise environments. Sensors (Basel) 17(12), 2757 (2017)
9. Dominiak, K.N., Pedersen, L.J., Kristensen, A.R.: Spatial modeling of pigs' drinking patterns as an alarm reducing method I. Developing a multivariate dynamic linear model. Comput. Electron Agr. 161, 79–91 (2019)
10. Pan, X., Zhu, J., Tai, W., Fu, Y.: An automated method to quantify the composition of live pigs based on computed tomography segmentation using deep neural networks. Comput. Electron Agr. 183, 105987 (2021)
11. Ren, S., He, K., Girshick, R.B., Sun, J.: Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell. 39(6), 1137–1149 (2017)
12. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. In: 2017 IEEE International Conference on Computer Vision (ICCV 2017), Venice, Italy, pp. 2980–2988 (2017)
13. Huang, Z., Huang, L., Gong, Y., Huang, C., Wang, X.: Mask scoring R-CNN. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 6409–6418 (2019)
14. Cai, Z., Vasconcelos, N.: Cascade R-CNN: High quality object detection and instance segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 43(5), 1483–1498 (2021)
15. Tian, Z., Shen, C., Chen, H., He, T.: FCOS: Fully convolutional one-stage object detection. In: International Conference on Computer Vision (ICCV 2019), Seoul, South Korea, pp. 9626–9635 (2019)
16. Zhang, R., Tian, Z., Shen, C., You, M., Yan, Y.: Mask encoding for single shot instance segmentation. In: Conference on Computer Vision and Pattern Recognition (CVPR 2020), Seattle, WA, USA, pp. 10223–10232 (2020)
17. Tian, Z., Shen, C., Chen, H.: Conditional convolutions for instance segmentation. In: Conference on Computer Vision and Pattern Recognition (CVPR 2020), Seattle, WA, USA, pp. 282–298 (2020)
18. Chen, Y., Dai, X., Liu, M., Chen, D., Yuan, L., Liu, Z.: Dynamic convolution: Attention over convolution kernels. In: Conference on Computer Vision and Pattern Recognition (CVPR 2020), Seattle, WA, USA, pp. 11027–11036 (2020)
19. Qiao, Y.L., Truman, M., Sukkarieh, S.: Cattle segmentation and contour extraction based on Mask R-CNN for precision livestock farming. Comput. Electron Agr. 165, 10458 (2019)
20. Hu, Z., Yang, H., Lou, T.: Dual attention-guided feature pyramid network for instance segmentation of group pigs. Comput. Electron Agr. 186, 106140 (2021)
21. Tu, S., Yuan, W., Liang, Y., Wang, F., Wan, H.: Automatic detection and segmentation for group-housed pigs based on PigMS R-CNN. Sensors (Basel) 21(9), 3251 (2021)
22. Xiao, J., Liu, G., Wang, K., Si, Y.: Cow identification in free-stall barns based on an improved Mask R-CNN and an SVM. Comput. Electron Agr. 194(3), 106738 (2022)
23. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., David Warde-Farley, S., Ozair, A.C., CourvilleBengio, Y.: Generative adversarial nets. In: Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems, pp. 2672–2680 (2014)
24. Luc, P., Couprie, C., Soumith ChintalaVerbeek, J.: Semantic segmentation using adversarial networks. CoRR 2016, abs/1611.08408, (2016)
25. Xue, Y., Xu, T., Zhang, H., Long, L.R., Huang, X.: SegAN: Adversarial network with multi-scale L1 loss for medical image segmentation. Neuroinformatics 16(3), 383–392 (2018)
26. Chen, Y., Xia, R., Zou, K., Yang, K.: RNON: Image inpainting via repair network and optimization network. Int. J. Mach. Learn Cybern. 1–17 (2023)
27. Manohara Pai, M.M., Mehrotra, V., Ujjwal VermaPai, R.M.: Improved semantic segmentation of water bodies and land in SAR images using generative adversarial networks. Int. J. Semant Comput. 14, 55–69 (2020)
28. Khaled, A., Han, J.-J., Ghaleb, T.A.: Multi-model medical image segmentation using multi-stage generative adversarial networks. IEEE Access 10, 28590–28599 (2022)
29. Man, R., Yang, P., Xu, B.: Classification of breast cancer histopathological images using discriminative patches screened by generative adversarial networks. IEEE Access 8, 155362–155377 (2020)
30. Minagi, A., Hokuto HiranoTakemoto, K.: Natural images allow universal adversarial attacks on medical image classification using deep neural networks with transfer learning. J. Imaging 8(2), 38 (2022)
31. Choi, S.H., Shin, J.-M., Peng LiuChoi, Y.-H.: ARGAN: Adversarially robust generative adversarial networks for deep neural networks against adversarial examples. IEEE Access 10, 33602–33615 (2022)

32. Guo, X., Hiroyuki OkamuraDohi, T.: Automated software test data generation with generative adversarial networks. IEEE Access 10, 20690–20700 (2022)
33. Nistal, J.: Exploring generative adversarial networks for controllable musical audio synthesis. (Synthèse audio musicale contrôlable à l'aide de réseaux adverses génératifs) (2022)
34. Bolya, D., Zhou, C., Xiao, F., Lee, Y.J.: YOLACT: Real-time instance segmentation. In: International Conference on Computer Vision (ICCV 2019), pp. 9156–9165, Seoul, South Korea (2019)
35. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: International Conference on Computer Vision (ICCV 2021), Montreal, QC, Canada, pp. 9992–10002 (2021)
36. Wang, X., Zhang, R., Kong, T., Li, L., Shen, C.: SOLOv2: Dynamic and fast instance segmentation. In: Annual Conference on Neural Information Processing Systems (NeurIPS 2020) (2020)
37. Valletta, J.J., Torney, C., Kings, M., Thornton, A., Madden, J.: Applications of machine learning in animal behaviour studies. Anim. Behav. 124, 203–220 (2017)
38. Wathes, C.M., Kristensen, H.H., Aerts, J.M., Berckmans, D.: Is precision livestock farming an engineer's daydream or nightmare, an animal's friend or foe, and a farmer's panacea or pitfall? Comput. Electron Agr. 64(1), 2–10 (2008)
39. Alameer, A., Kyriazakis, I., Bacardit, J.: Automated recognition of postures and drinking behaviour for the detection of compromised health in pigs. Sci. Rep. 10(1), 13665 (2020)

How to cite this article: Tu, S., Zeng, Q., Liu, H., Liang, Y., Liu, X., Huang, L., Huang, Z.: MaskDis R-CNN: An instance segmentation algorithm with adversarial network for herd pigs. IET Image Process. 1–12 (2023). https://doi.org/10.1049/ipr2.12880