



**HAL**  
open science

# SR-YOLO: Small Objects Detection Based on Super Resolution

Biao Hou, Xiaoyu Chen, Shenxuan Zhou, Heng Jiang, Hao Wang

► **To cite this version:**

Biao Hou, Xiaoyu Chen, Shenxuan Zhou, Heng Jiang, Hao Wang. SR-YOLO: Small Objects Detection Based on Super Resolution. 5th International Conference on Intelligence Science (ICIS), Oct 2022, Xi'an, China. pp.352-362, 10.1007/978-3-031-14903-0\_38 . hal-04666410

**HAL Id: hal-04666410**

**<https://hal.science/hal-04666410v1>**

Submitted on 1 Aug 2024

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License



This document is the original author manuscript of a paper submitted to an IFIP conference proceedings or other IFIP publication by Springer Nature. As such, there may be some differences in the official published version of the paper. Such differences, if any, are usually due to reformatting during preparation for publication or minor corrections made by the author(s) during final proofreading of the publication manuscript.

# SR-YOLO: Small Objects Detection Based on Super Resolution

Biao Hou, Xiaoyu Chen, Shenxuan Zhou, Heng Jiang, and Hao Wang

School of Artificial Intelligence, Xidian University, Xi'an 710071, China  
15091892956@163.com

**Abstract.** Since the introduction of convolutional neural networks, object detection based on deep learning has made great progress and has been widely used in the industry. However, because the weak and small object contains too little information, the samples are rich in diversity, and there are different degrees of occlusion, the detection difficulty is too great, and the object detection has entered a bottleneck period of development. We firstly introduce a super-resolution network to solve the problem of small object pixel area being too small, and fuse the super-resolution generator with the object detection baseline model for collaborative training. In addition, in order to reinforce the weak feature of small objects, we design a convolution block based on the edge detection operator Sobel. Experiments show that proposed method achieves mAP50 improvement of 2.4% for all classes and 4.4% for the relatively weak pedestrian class on our dataset relative to the Yolov5s baseline model.

**Keywords:** Small Object Detection · Image Super Resolution · Infrared Image.

## 1 Introduction

Object detection, also known as object category detection or object classification detection, returns the category information and location information of the object of interest in the image. In the past two decades, it has been a research hotspot in the field of computer vision and digital image processing. During the two decades of development of object detection technology, various object detection architecture widely used in the industry have emerged. However, the most state-of-the-art object detection architecture are still a tough act to follow human eyes. The current object detection still faces many obstacles, among which the problem of poor performance of small object detection has not been completely solved.

In the object detection task, the small object contains too little RGB information due to its own definition, it contains too few discriminative features. The existence of the problem is also accompanied by dataset imbalance. As far as the COCO[1] dataset is concerned, only 51.82% of the pictures contain small objects, and that is a serious image-level imbalance. Most of the current mainstream object detection architecture are anchor-based methods, but since the ground truth and anchor of small targets are very small, if there is a slight offset between the two, the IoU will become very low, resulting in the network being misjudged as a negative sample. In addition, there is a difficulty in small object detection that cannot be ignored: for some small object samples, there

are different degrees of occlusion, blurring and incompleteness. Also taking the COCO dataset as an example, 41% of the annotations are small objects. In the case of , the problem of small object detection is still not alleviated, because most of the annotations in these small and difficult objects are difficult to be effectively utilized. Our work can be summarized as follows:

- **Super Resolution Generator** In view of the low detection performance of small object due to their small pixel area, we introduce the generator component of the super-resolution network before the current detection of the backbone network. Unlike the general generative adversarial network, the parameters of the generator network we introduce update relies on the loss return of object detection.
- **Edge-oriented Sobel Block** In order to strengthen the object edge information, a convolution block based on the Sobel edge extraction operator is designed.

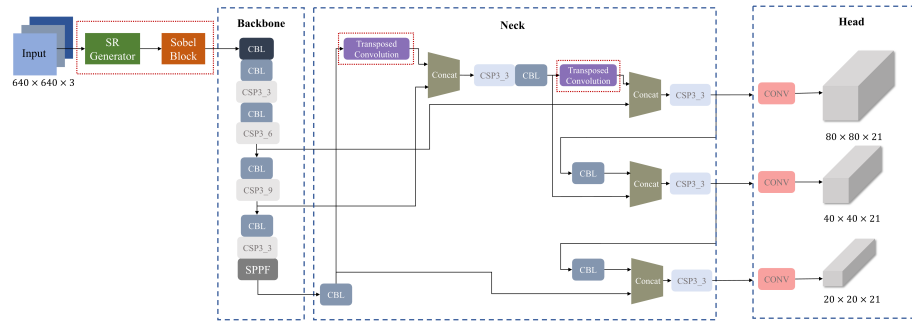


Fig. 1: **Yolov5-Super Resolution Generator.** On the basis of the original yolov5s model, the improvements we have made include: (1) Adding SR Generator and Sobel Block to the front end of the model. (2) Replace Upsampling in the original model with Transposed Convolution.

## 2 Related Work

### 2.1 Object Detection

Object detection has also entered the period of deep learning. According to whether the algorithm needs to generate a pre-selection box, the object detection algorithm based on deep learning can be divided into one-stage detection algorithm and two-stage detection algorithm. Representative networks in one-stage detection algorithms include YOLO[2] series, SSD[3], and RetinaNet[4]. The basic idea of the YOLO series of object detection algorithms is to send the input image into the feature extraction network to obtain a feature map of  $S \times S$  size, and divide the input image into  $S \times S$  cells. Each grid uses logistic regression to predict  $B$  bounding boxes, outputting the location information of the bounding boxes and the probability information of the  $C$  categories. Among them, Yolov5[5] is the latest network structure of the Yolo series.



Data augmentation is the simplest and most effective way to improve the performance of small object detection. From early distortion[6], rotation and scaling, to elastic deformation[7], random cropping[8] and translation[9] derived later, until now Horizontal flip used in Fast R-CNN[10], CutOut[11], MixUp[12], CutMix[13] and other methods used in YOLO. Mosaic data enhancement was further proposed in YOLOv4[14] with reference to the CutMix data enhancement method, and the four pictures were spliced by random scaling, random cropping, and random arrangement. Yolov5 follows the Mosaic enhancement used by Yolov4. Multi-scale learning combining deep semantic features and shallow representation information solves the problem of small object detection to a certain extent. The feature pyramid proposed by Liu[15] et al. is currently the most widely used multi-scale network, and many variants of FPN have also appeared. Context learning is to model the coexistence relationship between objects and scenes and objects and objects, thereby improving the performance of small object detection. Generative adversarial learning[16] generates features that are equivalent to those of high-resolution targets from the feature map of low-resolution small targets, which improves the detection performance of low-resolution targets to a certain extent. In addition, Anchor-free abandons the anchor mechanism and converts the target detection task into the detection of key points, avoiding the small target detection problem caused by the small offset between the anchor and the ground truth.

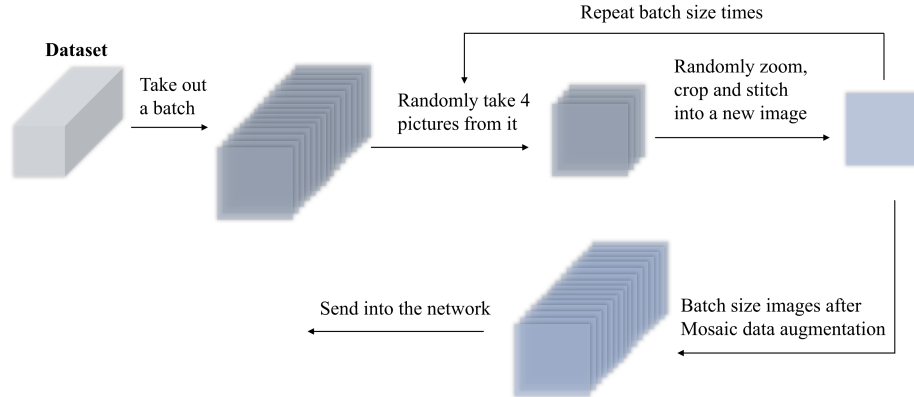


Fig. 2: Mosaic-4 Data Augmentation Process

This paper further uses Mosaic-9 on the basis of Mosaic-4, randomly selects nine pictures in a training batch, randomly scales, randomly crops, and randomly arranges. Figure 2 illustrates the processing of mosaic enhancement. The difference between Mosaic-9 and Mosaic-4 is that Mosaic-9 uses nine images. This approach further expands the number of small objects in the data set, and alleviates the imbalance of the data set. The problem.

## 2.2 Image Super Resolution

High-resolution images tend to contain greater pixel density, richer texture details, and higher reliability than low-resolution images. But usually limited by the limitations of

the acquisition equipment and environment, it is impossible to obtain an ideal image with sharp edges and no blur. Image super-resolution reconstruction technology refers to the restoration of a given low-resolution image into a corresponding high-resolution image through a specific algorithm. It has a wide range of applications and research significance in many fields, such as image compression, medical imaging, remote sensing imaging and video perception, etc. According to time and effect, image super-resolution algorithms can be divided into two categories: traditional algorithms and deep learning algorithms. Traditional super-resolution reconstruction algorithms mainly rely on conventional digital image processing and other technologies. Dong et al. proposed SRCNN[17] (Super-Resolution Convolution Neural Network), the first application of deep learning technology to the field of image super-resolution. In 2017, Christian Ledig et al. proposed SRGAN[18](Super-Resolution Generative Adversarial Networks), which defines a perceptual loss function and uses a generative adversarial network to reconstruct low-resolution images into high-resolution images. Based on the work of the above scholars, we believe that the generator in SRGAN has a structural advantage in image super-resolution, combining it with the object detection backbone network. Using the loss of object detection to update the weights of the generator, the goal is to allow the generator to generate high-resolution images that are more conducive to small object detection.

### 3 Method

#### 3.1 Image Super Resolution Based on Generative Adversarial Networks

SRGAN consists of two parts: generator and discriminator. The generator is used to generate high-resolution images from low-resolution images, and the discriminator is used to compete with the generator, so that the generator can generate more realistic high-resolution images. In this paper, the generator structure in SRGAN is integrated into the original Yolov5s model, which is different from the use of perceptual loss in SRGAN. The loss of Yolov5s is used to update the generator parameters, so that the high-resolution model generated by the generator is more suitable for the object detection task. Essentially alleviates the detection performance problem caused by more small object samples.

Figure 3 illustrates the network structure of the generator part of SRGAN, which is mainly composed of a deep residual module and a sub-pixel convolution module. The generator uses multiple deep residual modules to extract features from images, which ensures the stability of the generation network. Sub-pixel convolution, also known as pixel shuffle is different from upsampling and other methods, avoiding the introduction of too many artificial factors, and using the method of sample learning to enlarge the feature map.

#### 3.2 Feature Extraction Backbone Considering Edge Information

The object information extracted by the feature extractor includes edge information, and the edge information occupies the main position in the feature information, and the human brain also mainly relies on the edge information when distinguishing objects. In this

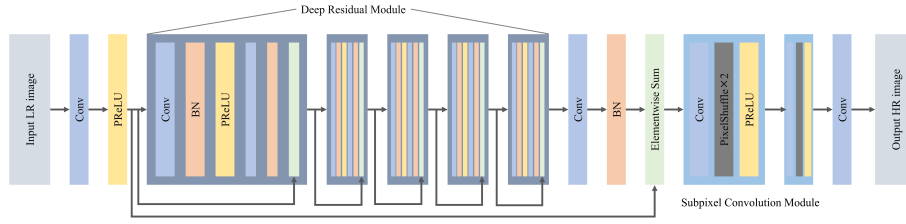


Fig. 3: Super Resolution Generator

paper, the Sobel[20] edge detection operator is used to strengthen the edge information of the target, and the Sobel convolution is designed to fix the value of the convolution kernel to extract the edge information of the target.

Specifically, the input features are first subjected to  $1 \times 1$  convolution for channel expansion, and then the output is sent to two Sobel convolution layers Sobel-x and Sobel-y with fixed weights in the vertical direction, respectively. Let  $D_x$  and  $D_y$  denote the convolution kernels used to extract the gradients of features in the x and y directions, respectively.

$$D_x = \begin{bmatrix} +1 & 0 & -1 \\ +2 & 0 & -2 \\ +1 & 0 & -1 \end{bmatrix}, D_y = \begin{bmatrix} +1 & +2 & +1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix} \quad (1)$$

The input is finally summed with the output of the Sobel convolutional layer. The specific structure of the Sobel block is shown in Figure 4.

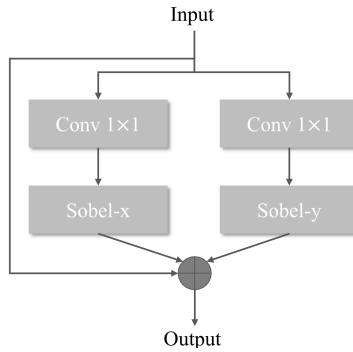


Fig. 4: Edge-oriented Sobel block

## 4 Experiment

### 4.1 Datasets

The infrared data set used in this paper was collected in the spring of 2022, and the infrared video data of Taibai South Road and Xifeng Road Section in Xi'an City, Shaanxi Province were collected from 3:00 p.m. to 5:00 p.m. The total duration is two hours, including 2342 images in the training set, 585 in the validation set, and 1000 in the test set, with an image size of 1920x1080. The objects in the dataset are labeled according to the YOLO format, and the labeled categories include pedestrians and vehicles. The specific data distribution is shown in Fig 5. Fig 6 shows part of the dataset. It can be seen that in addition to the problem of class imbalance, the dataset also has a large number of weak and small objects.

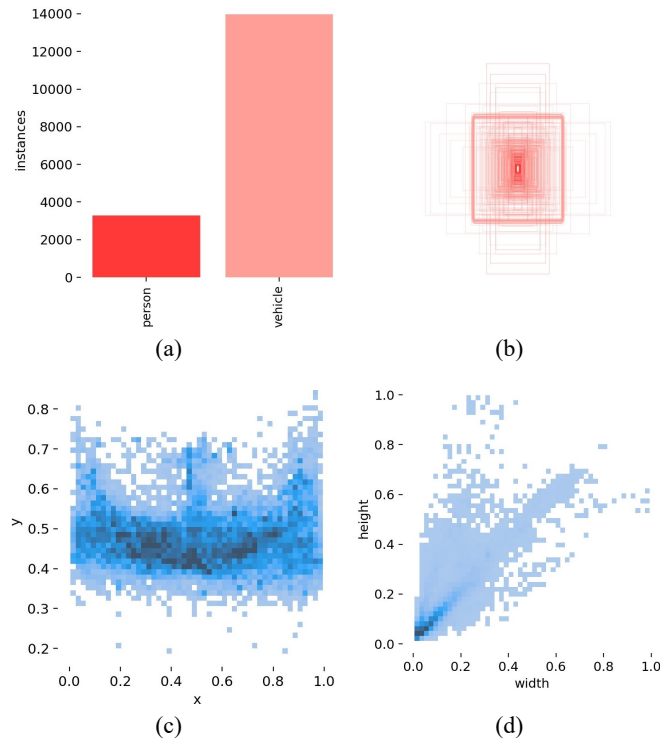


Fig. 5: Illustration of dataset distribution.(a) shows that the dataset has a serious class imbalance problem.(b) draws all the ground truth boxes.(c) and (d) show the distribution of the center point coordinates and width and height of the real box of the dataset, respectively



Fig. 6: Part of the Dataset

## 4.2 Experimental Environment and Training Settings

The experimental environment uses the Ubuntu20.04 operating system and a Nvidia GeForce RTX 3090 for computing, its video memory size is 24268MiB, the CPU configuration is Intel(R) Xeon(R) Silver 4216 CPU @ 2.10GHz, the Pytorch version is 1.9.0, and the Python language environment is 3.7.11.

All experiments were trained for 100 epochs, batch-size is set to 32, the optimizer chooses SGD, the initial learning rate of the optimizer is 0.01. The learning rate in the model training phase is warmed up using Warmup, using one-dimensional linear interpolation to update the learning rate of each iteration, and the cosine annealing algorithm is used after the warm-up stage to update the learning rate.

## 4.3 Evaluation Indicators

The indicators used to evaluate the model detection performance in target detection are: **AP** (Average precision), **Precision** and **Recall** and comprehensive evaluation index **F-Measure**.

**Precision and Recall** Precision: The proportion of true targets among the targets detected by the model. Recall: The proportion of all real objects detected as positive. Precision-Recall Curve: The horizontal axis is the Recall, and the vertical axis is the Precision. The two variables are negatively correlated, that is, when the precision is larger, the Recall is smaller, and conversely, when the Recall is larger, the precision is lower. The more ideal the curve, the better the performance of the model and the larger the area enclosed by the abscissa and ordinate.

**AP(Average precision)** Average precision refers to the area under the PR curve, which refers to the average precision for different recall rates. The higher the AP, the better the

recognition performance of the model. mAP is the mean average precision. When the sample has multiple categories, the AP of each category is averaged. mAP measures the detection performance of the model on various samples of the dataset.

#### 4.4 Analysis of Experimental Results

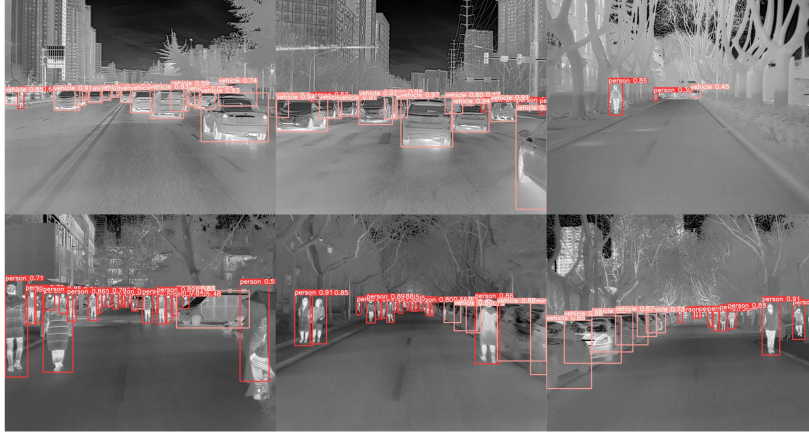


Fig. 7: Partial Inference Results

Table 1: Comparison of the validation results of our method and the baseline method

Method	Layers	Weight(M)	Precision(%)	Recall(%)	mAP@.5(%)	mAP@.5:.95(%)
Baseline	213	6.69	86.6	79.1	86.2	53.2
Sobel	225	6.70(↑0.01)	<b>88.3(↑1.7)</b>	78.7(↓0.4)	85.8(↓0.4)	<b>53.4(↑0.1)</b>
SR	257	7.37(↑0.68)	<b>88.2(↑1.6)</b>	<b>88.5(↑9.4)</b>	<b>88.6(↑2.4)</b>	<b>56(↑2.8)</b>

Table 2: Comparison of the performance difference between our model and the baseline model in the detection of weak pedestrian objects

Method	Layers	Weight(M)	Precision(%)	Recall(%)	AP@.5(%)	AP@.5:.95(%)
Baseline	213	6.69	80.9	67	76.5	36.8
SR+Sobel	257	7.37(↑0.68)	<b>83.7(↑2.8)</b>	<b>70.9(↑3.9)</b>	<b>80.9(↑4.4)</b>	<b>41.8(↑5)</b>

Table 1 and Table 2 illustrate the detection performance of the baseline model and our model on the dataset, respectively. The results described in Table 1 are for the entire data set. It can be seen that the Sobel module can improve the precision of the model with almost no increase in computational cost. Unsurprisingly, the SR module has improved various indicators of the model. The validity of the Sobel module and the SR module can be demonstrated. As mentioned above, the dataset has severe class imbalance, and there are many weak objects in pedestrian samples. To verify the improvement of our method for weak and small object detection, Table 2 records the results of the two models

that are only validated for the pedestrian category. Obviously, our model has achieved greater improvements in Precision, Recall, and mAP than the overall dataset relative to the baseline model, which shows that the performance improvement in Table 1 is mainly due to the performance improvement of the model for weak and small target detection.

## 5 Conclusion

In this paper, we propose a framework for small target detection: SR-YOLO, which combines the image super-resolution network SRGAN and the target detection baseline network Yolov5s, and uses the structural advantages of the super-resolution network generator to effectively improve the detection accuracy on self-collected dataset, especially in relatively weak categories.

## 6 Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant 62171347, 61877066, 61771379, 62001355, 62101405; the Foundation for Innovative Research Groups of the National Natural Science Foundation of China under Grant 61621005; the Key Research and Development Program in Shaanxi Province of China under Grant 2019ZDLGY03-05 and 2021ZDLGY02-08; the Science and Technology Program in Xi'an of China under Grant XA2020-RGZNTJ-0021; 111 Project.

## References

1. Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.
2. Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.
3. Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016.
4. Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.
5. Glenn Jocher, Alex Stoken, Jirka Borovec, NanoCode012, Ayush Chaurasia, TaoXie, Liu Changyu, Abhiram V, Laughing, tkianai, yxNONG, Adam Hogan, lorenzomamma, AlexWang1900, Jan Hajek, Laurentiu Diaconu, Marc, Yonghye Kwon, oleg, wanghaoyang0106, Yann Defretin, Aditya Lohia, ml5ah, Ben Milanko, Benjamin Fineran, Daniel Khromov, Ding Yiwei, Doug, Durgesh, and Francisco Ingham. ultralytics/yolov5: v6.0 - YOLOv5-P6 1280 models, AWS, Supervise.ly and YouTube integrations, May 2022.
6. Larry Yaeger, Richard Lyon, and Brandyn Webb. Effective training of a neural network character classifier for word recognition. *Advances in neural information processing systems*, 9, 1996.

7. Patrice Y Simard, David Steinkraus, John C Platt, et al. Best practices for convolutional neural networks applied to visual document analysis. In *Icdar*, volume 3, 2003.
8. Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.
9. Li Wan, Matthew Zeiler, Sixin Zhang, Yann Le Cun, and Rob Fergus. Regularization of neural networks using dropconnect. In *International conference on machine learning*, pages 1058–1066. PMLR, 2013.
10. R Girshick. Fast r-cnn. *IEEE international conference on computer vision*. *IEEE*, 2015:1440–1448.
11. Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017.
12. Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.
13. Sangdoon Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032, 2019.
14. Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020.
15. Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016.
16. Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.
17. Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2):295–307, 2015.
18. Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017.
19. Alexander Neubeck and Luc Van Gool. Efficient non-maximum suppression. In *18th International Conference on Pattern Recognition (ICPR'06)*, volume 3, pages 850–855. IEEE, 2006.
20. Irwin Edward Sobel. *Camera models and machine perception*. Stanford University, 1970.