**Research on human recognition of post disaster search and rescue robot based on YOLOv7 fusionattention mechanism**

Peng Bo Feng1,2, Yu Ming Qi 1,2\*, San Peng Deng 1,2, Chen Zhu1,2, Shao Peng Li1,2

1 Institute of Robotics and Intelligent Equipment, Tianjin Vocational and Technical Normal University; Tianjin , 300222,China;

2 Tianjin Key Laboratory of Intelligent Robot Technology and Application Enterprise, Tianjin, 300350,China.

[a892480499qq.com](mailto:awy875987670@163.com)

Item no.21KPXMRC00040

***Absrtact:****In view of the complex urban environment, the rescue after the earthquake has the problems of not timely discovery and low recognition rate. A small target human detection algorithm model SE-YOLOv7 based on YOLOv7 is proposed to solve this problem. Based on YOLOv7, a human body recognition data set is made for complex urban disaster areas, and the residual network module SE attention mechanism is added to the framework[1]to improve the contribution of positive sample features to better train the model weights. Combined with the characteristics that the affected human body imaging is different from the normal situation, the distribution density of the length and width of the prior box[2]is improved, and the model is deployed on the embedded platform of the post-disaster eight-legged rescue robot[3]to carry out the human target detection experiment in the post-disaster environment. The accuracy of human limb recognition is2.3%map higher than that of voc2012 data set, the recall rate is increased by 2.7%, and the speed is 67.6 FPS, which meets the real-time and accuracy requirements of target detection of rescue robots under disaster conditions.*

***keywords:****YOLO v7;Human body recognition;Attention mechanism;robot*

**preface:** Disasters caused by earthquakes,fires and wars are sudden and destructive. Disaster relief timeliness,high accuracy requirements, but rescue workers immediately into the rescue work is easy to encounter secondary danger. In order to ensure the safety and rescue efficiency of rescuers, rescue robots can be put into the disaster site at the first time after the disaster to collect information and identify and rescue human bodies in the complex environment after the disaster.

In the post-disaster complex environment, the rescue robot wants to completely replace the rescue personnel to complete the search and rescue task independently. Firstly, the robot reaches the safe area at the edge of the affected area through the robot, and puts the robot into the disaster area autonomously. At the same time, the camera is turned on to collect the post-disaster environment image in real time, and then the surrounding environment is analyzed according to the image. Finally, the target detection is carried out with the help of the visual detection algorithm, and the rescue task has been realized. The traditional target detection model will use several sliding windows to generate target candidate boxes from the image, and then extract features in the target candidate boxes[4].After multiple convolutions and pooling, the feature classification is finally performed. The model uses a sliding window to generate candidate boxes, and after pooling, the amount of data becomes larger, resulting in high time complexity, which cannot adapt to more diverse target features, resulting in low recall rate and poor robustness. Target recognition based on deep learning can adaptively learn target features from a large number of pictures, with better accuracy and robustness.

Based on the single-stage target detection model YOLOv7, this paper proposes an improved model for human target recognition. Firstly, the outdoor scene data set is added to the original COV2012 data set, and the number of positive samples is doubled through data enhancement, so that the training model is better. Aiming at the human target recognition of complex scenes in outdoor disasters, this paper chooses the YOLOv7 model with the highest efficiency as the basic framework, and proposes an SE-YOLOv7 detection algorithm based on attention mechanism and multi-scale fusion, which improves the accuracy of human detection in complex backgrounds and can better meet the personnel search and rescue work in emergency situations.

1. **Octapod robot hardware design**

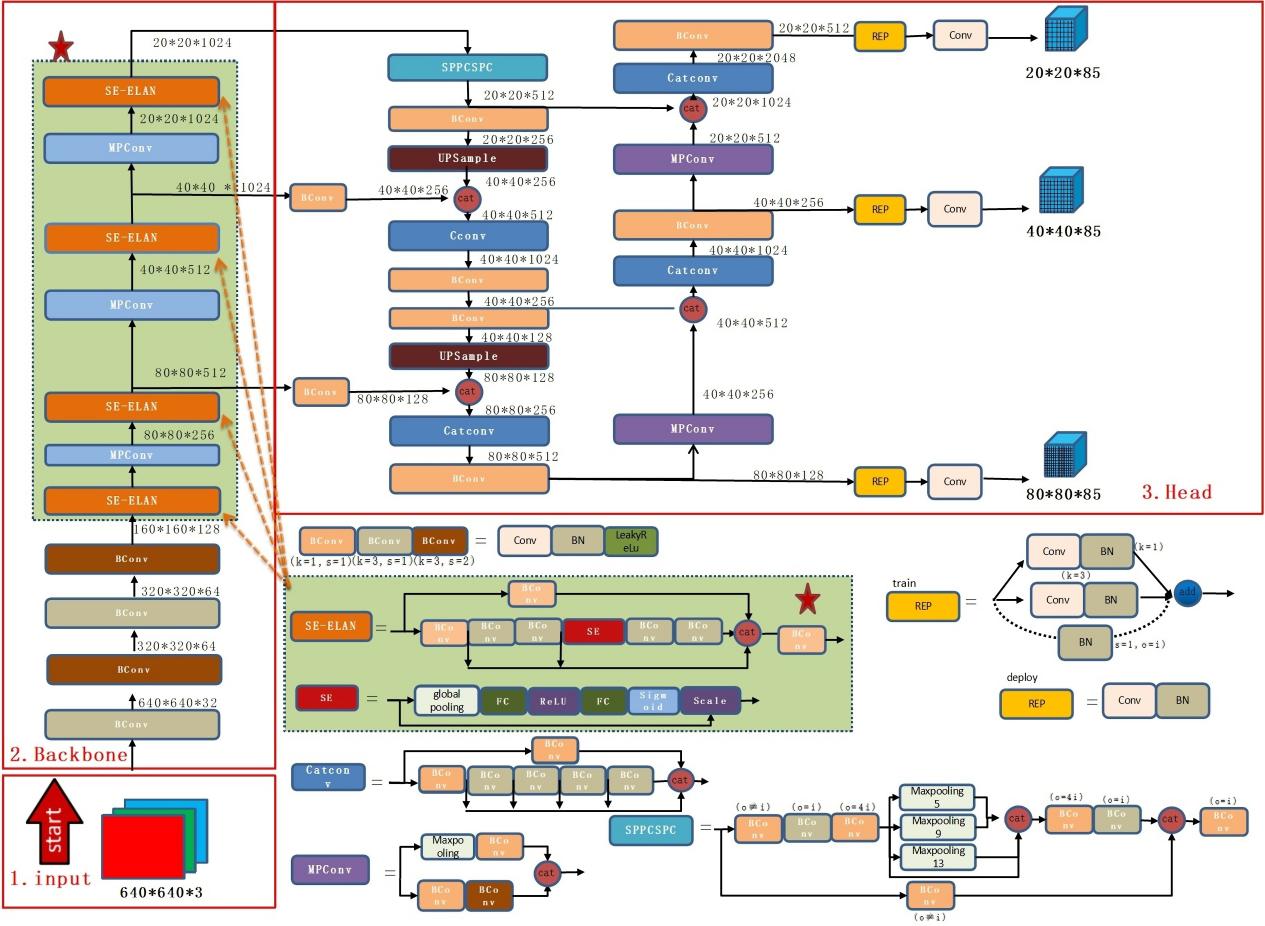
The octopod robot adopts a series leg structure, using 60 kg steering gear as the power, and three steering gears form a leg. Single leg has 3 degrees of freedom, can complete forward, backward, left turn, right turn and other functions, can realize the basic mobile function of rescue robot. A total of 24 steering gears are installed in the graphic position, and three joints are driven by three steering gears, so that the robot has 3 degrees of freedom. Each leg can be lifted to form a hand-foot universal mode to complete more postures and movements. It can cross complex obstacles, and the legs can be equipped with a camera for real-time human target recognition to meet the requirements of post-disaster search and rescue.

|  |  |
| --- | --- |
| 机器人 |  |
| 八足机器人1 | 2e7db29743da637eb59f37ed06708f3 |
| *Figure 1 Design of octopod robot* | |

1. **Research on improved human recognition visual algorithm**

Combined with the situation of this experimental data set, in order to solve the problems of low accuracy, excessive number of negative samples, object occlusion and other problems in human body recognition in post-disaster complex environment, and meet the requirements of accuracy and recall rate, this paper increases the attention mechanism on the basis of YOLOv7, proposes SE-ELAN module to extract the backbone network, enhances the feature extraction ability of the target, and uses K-means++algorithm to cluster the appropriate prior box to improve the detection accuracy of positive samples.

The network sets the size of the input layer to be 640×640×3. The backbone first uses four Bconv convolution kernels[5]to extract features, and then SE-ELAN and MPconv modules sequentially perform cyclic downsampling. The resolutions of the feature maps are 80×80,40×40, and 20×20, respectively. Subsequently, an upsampling and a downsampling are performed in the Head layer, and they are densely connected using the cat module. Finally, three-scale feature outputs are extracted to meet different scales of target detection. The specific network is shown in Figure 2 below.

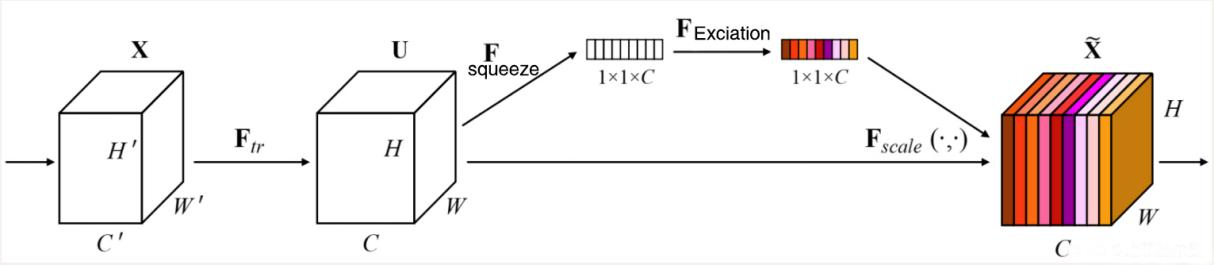


*Figure 2 YOLOv7 network structure with SE attention mechanism*

* 1. ***SE attention module***

Since the downsampling increases the receptive field of the network[6], and at the same time leads to the loss of the feature information of the small target, the SE attention mechanism is introduced to enhance the feature extraction ability of the network. The SE attention mechanism can solve the problem of information loss caused by the difference of channel importance in the downsampling convolution pooling. Therefore, the SE module is introduced to increase the importance of positive sample features in the network.

The SE module structure diagram consists of three parts : Squeeze, Excitation and Scale, as shown in Figure 3. Firstly, the spatial dimension is compressed by squeeze operation, and each feature map is globally pooled and averaged into a real value. The real number has a global receptive field. The excitaton operation generates a weight representing the importance of each feature channel based on the correlation of feature channels, and increases the nonlinear ability of the network. Finally, the Scale operation multiply the weight by multiplicative weighting to the previous feature, and is engaged in the function of improving the positive sample feature and suppressing the negative sample feature.



*Figure 3 SE Module*

* 1. ***Improved SE-ELAN***

The position of the SE module will also affect the representation ability of the network. If the attention mechanism module is placed in the Head layer, it is difficult to distinguish the key features from the useless features in the fused feature map, which will make the redundant features more important and reduce the feature extraction ability of the network. In the backbone of the shallow network has more original image details, more conducive to small target human recognition[7]. As the network deepens, some regional texture information will also be lost. Therefore, the SE attention module is embedded in the residual unit of ELAN.First, the extracted features are used in the attention mechanism to highlight the key features and weaken the redundant information. The residual operation is carried out in-depth extraction, and finally the output is merged.

* 1. ***Optimization design of prior frame***

Selecting the appropriate priori box greatly improves the network optimization effect. Because the human body under the disaster is blocked by the object, the proportion of the human body is different from that of the normal. The original YOLOv7 is based on the COCO data set. The initial prior box obtained cannot meet the detection requirements well, so it is necessary to select its own prior box size. The K-means algorithm is based on the random selection of initial values, and the clustering effect is unstable. In order to reduce the error, the K-means++[8] algorithm is selected to cluster the post-disaster human body tags to generate 9 groups of different sizes Anchor. The clustering results are shown in Table 1.

*Table 1 Prior boxes generated by K-means++*

|  |  |  |
| --- | --- | --- |
| characteristic pattern | reception field | Achor |
| 80×80 | micro | (12,16),(20,36),(41,27) |
| 40×40 | mid | (34,73),(72,55),(69,142) |
| 20×20 | magnum | (132,101),(182,238),(449,397) |

**3 Experiment and result analysis**

***3.1 Platform carrying and model training***

The software environment and hardware environment used for model training and testing in this paper are:win10 operating system, Pytorch deep learning framework, CUDAv11.1, Cudnnv8.2.0, OpenCVv4.5.5, CPU Intel(R)Xeon(R)CPU E5-2678 v3@2.50GHz, 32GBRAM, GPU NVIDIA GeForce GTX 1060.

In this paper, the original data resolution is high, and the input layer input image size is set to 640×640×3. The image is adaptively adjusted during model training to complete scaling and filling. The network performs a total of 300 epochs, the batch size is set to 4, and the learning rate is dynamically changed. It is set to the process of le-2 cosine change decaying to le-3. The learning rate is low at the start and end, and the intermediate process learning rate is high.

3.2 Evaluation indicators

In this experiment, the threshold of the ratio of the intersection of the candidate box and the target real box (IOU) is greater than 0.5 as the standard evaluation target detection, AP is the average accuracy, mAP is the average accuracy and fps is the number of frames per second. P-Accuracy represents the ratio of correct prediction samples to total detection targets, R-Recall represents the ratio of correct prediction samples to all targets, and AP is the area surrounded by P-R curve and coordinate axis.



****

****

****

3.3 Ablation experimental results

In order to better analyze the superiority of the improved module in SE-YOLOv7 for human target detection, the ablation experiment is designed. The improved network is added to the original YOLOv7 and compared with the original algorithm. The specific experimental contents are shown in Table 2. Through Table 2, the contribution of each improvement strategy to the model is analyzed. From the experimental results, it is concluded that each improvement strategy has different improvements to the overall performance of the model.

*Table 2 Module ablation experiment of SE-YOLOv7*

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| model | SE | K-means++ | Algorithm | P / % | R / % | mAP / % | rate/(frames/s) |
| 1 | × | × | YOLOv7 | 94.3 | 90.2 | 92.1 | 69.7 |
| 2 | √ | × | SE-YOLOv7 | 95.4 | 92.1 | 93.8 | 67.9 |
| 3 | × | √ | SE-YOLOv7 | 95.0 | 92.6 | 93.6 | 69.3 |
| 4 | √ | √ | SE-YOLOv7 | 96.2 | 92.7 | 94.5 | 67.6 |

The SE attention mechanism is added to the model residual network, highlighting the key features. In order to further verify the superiority of the mechanism, the class activation graph is introduced, as shown in Figure3.The red region contributes the most to the image classification, while the blue region contributes the least, which successfully optimizes the recognition performance of the network and improves the algorithm mAP to 94.5%.

|  |  |  |
| --- | --- | --- |
| 2022_00008 | 2022_00008 | 2022_00008SE-v7 |
| 2022_00005原yolov7 | 2022_00006# | 2022_00005SE-yolov7 |
| (a)YOLOv7 | (b)Embedded SE module class activation diagram | (c)result of survey |

Figure 3 Comparison before and after introducing attention mechanism

4 Conclusion

Aiming at the problems of missing and wrong detection of small targets in YOLOv7 model,this paper proposes a post-disaster human target detection algorithm based on attention mechanism and multi-scale fusion,which realizes the accurate recognition of human body and residual limbs in complex background.

The experimental results show that the mAP value of the improved model in the self-built data set is 94.5%,which can accurately detect human targets at various angles after various disasters. Compared with other popular models,the YOLO model in this experiment has great advantages in comprehensive performance and meets the needs of personnel search and rescue tasks in complex post-disaster backgrounds. Subsequently,the network should be lightweight and the generalization ability should be improved while ensuring the accuracy of the algorithm, so that it can be applied to various targets and not limited to human target recognition.

**References**

*[1]Xu Qin,Liang Yulian,Wang Dongyue,Luo Bin.Hyperspectral image classification based on SE-Res2Net and multi-scale spatial-spectral fusion attention mechanism[J].Journal of Computer Aided Design and Graphics,2021,33(11):1726-1734.*

*[2] Li Sitai,Zhang Rongfen,Liu Yuhong.Multi-scale detection of traffic signs based on improved prior boxes and loss [J].Intelligent computer and application,2021,11 (06):56-62.*

*[3] Wu Huiting,Zhou Haijing,Dong Shitao.Coal mine rescue robot embedded real-time operating system design and analysis [J].Coal technology,2022,41(02):213-215.*

*[4] Chen Bing, Jiang Xingguo. Convolutional neural network for facial feature extraction [J].Modern electronic technology,2022,45(18):182-186.DOI:10.16652/j.issn.1004-373x.2022.18.035.*

*[5] Xu Tingting.Research and application of gesture estimation based on convolutional neural network [D].Yangzhou University,2022.DOI:10.27441/d.cnki.gyzdu.2022.000697.*

*[6] Zheng Wei,Zhao Jinfang,Zhang Yijing,Liu Shuaiqi,Zhang Xiaodan,Ma Zepeng.U-Net Brain Tumor MR Image Segmentation Based on Receptive Field Amplification and Attention Mechanism [J/ OL].Journal of Hebei University(Natural Science Edition):1-10.*

*[7] Chen Puqiang,Guo Lijun,Zhang Rong,Zhao Jieyu.Target Human Recognition Based on Global Spatial Constraint Block Matching [J].Computer Research and Development, 2015,52(03):596-605.*

*[8] Chen Tianyang,Qian Zheng,Jing Bo,Han Miaoquan.Research on Short-term Wind Power Prediction Model Based on K-means++and ELM[J/OL].Electrical Measurement and Instrumentation:1-7.*