IEPet: A Lightweight Multiscale Infrared Environmental Perception Network

Xinhao Jiang*, Wei Cai, Zhiyong Yang, Peiwei Xu and Qingjiang Dong
Xi’an High-Tech Institute, Xi’an, Shaanxi, 710025, China
*Corresponding author’s e-mail: jiangxinhao2020@outlook.com

Abstract. In recent years, the development of unmanned driving technology requires continuous progress in environment perception technology. Aiming at the key research direction of infrared environment perception in unmanned driving technology, a lightweight real-time detection network model for infrared environment perception, IEPet, is proposed. The model backbone adds the BottleneckCSP module and the proposed DCAP attention module, which can significantly improve the detection ability and spatial position perception ability while maintaining light weight. At the same time, the model improves the detection accuracy by using a 3-scales detection head. Comparative experiments on the unmanned driving data set show that compared with the lightweight model YOLOv4-tiny, the model proposed in this paper has an increase in F1 Score by 1.48% and an average detection accuracy by 6.37% to reach 84.31%. And the model is lighter. It shows that the proposed IEPet model can better meet the excellent performance required for infrared environment perception.

1. Introduction
In recent years, the rapid development of artificial intelligence and 5G technology, has promoted the rapid progress of unmanned driving technology integrated with multidisciplinary cutting-edge technologies. Driverless cars have also been tested in individual cities. As a new technology, unmanned driving technology contains greater research value and market value, but at the same time it faces various challenges.

Driverless cars are equipped with on-board multi-dimensional sensors, which transmit the sensed surrounding environmental conditions to the control module to make corresponding decisions. Therefore, the environmental perception sensor can be called the "visual center" of the driverless car, and its detection accuracy and speed directly determine the safety of the driverless car.

At present, there is an imbalance in the development of environment perception of unmanned vehicles, and the research mainly focuses on the visible light imaging perception of daytime scenes. Due to the small number of infrared unmanned driving data sets and the difficulty of target detection in the infrared band, the research and development of night scenes is slow. However, the environmental perception of night scenes based on infrared sensors has extremely high research value. Visible light sensors have limited effects at night. In order to prevent accidents when driverless cars are driving at night, the research on infrared target recognition applied in the field of driverless driving is urgent.

The application of deep learning to the field of unmanned driving target detection is a research hotspot in this direction [1]. The target detection model based on deep learning has been continuously developed into a one-step detection model and a two-step detection model this year. The classic models with good performance and wide application mainly include the YOLO series model [2-3] and the SSD model [4]. For example, Bu et al. [5] improved the SSD algorithm and applied it to pedestrian and
vehicle detection at night. However, the network is more complex, with a large amount of parameters, and the detection speed needs to be improved. The safety-oriented unmanned driving technology requires extremely high detection speed, so this paper uses the infrared unmanned driving data set to study the infrared band environment perception technology of night scenes. A lightweight target detection model for infrared environment perception based on deep learning is proposed. The innovation of this model lies in the use of the BottleneckCSP module [6], the DCAP attention module and the 3-scales detection method.

2. Proposed model

The infrared environment perception lightweight target detection model IEPet proposed in this paper is shown in Figure 1. First, the picture is input to the backbone for feature extraction, and then multi-scale feature fusion is performed, and finally three-scale output is obtained. Each important component of backbone is discussed in detail below.

2.1. BottleneckCSP module

Figure 2 is the structure diagram of BottleneckCSP. From the figure, it can be found that BottleneckCSP is a CSP module nested with a Bottleneck module. The Bottleneck module is the classic residual module shown in the upper right corner of Figure 2. A branch performs two convolution operations successively, and the residual branch is added to the convolutional feature map. The CSP module is a cross-stage partial connection module that divides the input feature map into two branches. One branch passes through the 3*3 convolution layer, Bottleneck, and 3*3 convolution layer, and the other branch directly performs the convolution operation. Then concatenate the two branches, and the superimposed feature map passes through the BatchNorm layer, the LeakyRelu layer and the 3*3 convolutional layer again. The BottleneckCSP structure can enhance the dissemination of gradient information by increasing the number of branches, thereby improving the learning ability of the network. At the same time, the repeated use of gradient information is reduced through continuous splitting and merging, which reduces the memory cost and reduces the amount of calculation.

![IEPet network structure diagram.](image)
2.2. DCAP module
Considering the importance of target spatial position information in the unmanned target detection task, this paper introduces the Coordinate Attention module [7], and embeds it in the depthwise separable convolution, and adds it to the last layer of Backbone. This paper refers to it as the DCAP attention module for short, and the specific structure is shown in Figure 3. Among them, the Coordinate Attention module plays the role of enhancing position information, the depthwise separable convolution module plays the role of information fusion, and the number of channels at the end of the Backbone is large, and the use of depth separable convolution can effectively reduce the amount of calculation and the amount of parameters. The feature map of the previous output layer first performs a depthwise convolution operation, and then outputs three branches. The two branches on the right first use two one-dimensional global pooling and aggregate direction-aware feature maps, and then split the cascaded feature map convolution For the two sets of tensors, two attention maps are generated that store position information, and the two sets of attention maps are multiplied by the feature map of the leftmost branch to generate a new feature map containing direction and position information. Finally, input the new feature map into the point convolution module for further fusion of information. The DCAP module can fully perceive spatial position information while ensuring lightweight, and improve the detection capability of the network.

3. Experimental results and analysis
3.1. Experiment preparation
3.1.1. Data set and experimental platform. The data set used in this paper is the FLIR Thermal Datasets [8] issued by FLIR Company. The data set is the road driving sequence thermal images collected by FLIR Company using thermal sensors equipped on the car. The labeled targets include people, bicycles and cars. The scene includes highways and streets, with 60% in the daytime and 40% at night, with a total of 10228 infrared images of 640 pixel×512 pixel. FLIR Company open sourced the data set to promote research on driverless car systems. The quantitative statistics of the three targets in this data set are shown in Table 1.
Table 1. Precision and efficiency of different detection methods.

| Target category     | Number of targets |
|---------------------|-------------------|
| FLIR ThermalDatasets|                   |
| Person              | 28151             |
| Bicycle             | 4457              |
| Car                 | 46692             |

The hardware platform graphics card used in the experimental training and testing phase is NVIDIA Quadro GV100, the video memory is 32G, the CUDA version is 11.0, the CPU is Inter Xeon Silver 4210, the operating system is Win 10, and the Pytorch deep learning development framework is used.

3.1.2. Experimental parameter configuration and evaluation indicators. The experimental parameters of this paper are set as follows. The optimization algorithm adopts the random gradient descent method, the initial learning rate is set to 0.01, the momentum factor is introduced and set to 0.937, and the weight decay is set to 0.0005. In the training phase, batchsize is set to 64 according to the data set size and device performance, the epoch is set to 500, and the input image size is set to 512×512. In the test phase, batchsize is set to 1, and the input image size is set the same as in the training phase.

The evaluation indicators in this paper select the commonly used deep learning target detection evaluation indicators: precision, recall, precision and recall's equal-weight harmonic average index F1 Score, mean average precision(mAP).

3.2. Comparative experiment on FLIR Thermal Datasets

In order to verify the detection performance of the IEPet network model proposed in this paper on the infrared driverless data set, a comparative experiment was carried out on the FLIR Thermal Datasets. Comparison algorithms include the widely used YOLOv3-tiny lightweight network model and the newly proposed YOLOv4-tiny high-performance lightweight network model. The experimental results are shown in Table 2.

Table 2. Comparison of experimental results.

| Method   | Precision | Recall | F1 Score | mAP   | Params/10^6 | Model size/MB |
|----------|-----------|--------|----------|-------|-------------|---------------|
| YOLOv3-tiny | 0.6049   | 0.7600 | 0.6714   | 0.7264 | 8.67        | 16.60         |
| YOLOv4-tiny | 0.5966   | 0.8129 | 0.6881   | 0.7794 | 6.69        | 12.85         |
| IEPet    | 0.5868   | 0.8764 | 0.7029   | 0.8431 | 4.99        | 9.66          |

It can be seen from Table 2 that the IEPet model proposed in this paper is better than the comparison models except precision is lower than YOLOv3-tiny and YOLOv4-tiny. Compared with the better-performing YOLOv4-tiny model, the IEPet model Recall proposed in this paper has increased by 6.35%, the average index F1 Score has increased by 1.48%, and mAP has increased by 6.37%. At the same time, the IEPet model has better lightweight characteristics. Compared with YOLOv4-tiny, the amount of model parameters is reduced by 25.4%, and the model size is reduced by 24.8%.

Figure 4 shows the visualization results of infrared targets in different scenes in the IEPet model detection data set. It can be found that the IEPet network model has a high detection confidence and has good detection capabilities for small targets. Therefore, combined with the analysis of comparative experiments and visualization results, the IEPet model proposed in this paper is lighter and has better infrared target sensing performance.
3.3. Ablation experiment on FLIR Thermal Datasets

In order to verify the impact of each key module in the IEPet network model on the detection performance compared to the baseline network, an ablation experiment was performed on FLIR Thermal Datasets. The experimental results are shown in Table 6.

It can be seen that by using the BottleneckCSP structure, adding the designed DCAP attention module, and increasing to three-scale detection, the comprehensive indicators F1 Score and mAP have been significantly improved, and the detection performance has been gradually strengthened.

| Baseline | With Bottleneck CSP | With DCAP module | With 3 detection heads | Precision | Recall | F1 Score | mAP |
|----------|---------------------|------------------|------------------------|----------|--------|----------|-----|
| √        |                     |                  |                        | 0.5633   | 0.7866 | 0.6565   | 0.7440 |
| √        | √                   |                  |                        | 0.5711   | 0.7969 | 0.6654   | 0.7530 |
| √        | √                   | √                |                        | 0.5830   | 0.7937 | 0.6722   | 0.7624 |
| √        | √                   |                  | √                      | 0.5618   | 0.8916 | 0.6893   | 0.8418 |
| √        | √                   | √                | √                      | 0.5868   | 0.8764 | 0.7029   | 0.8431 |

However, the table cannot reflect the increase in the utilization of spatial position information after the DCAP attention module is added to the model. Therefore, the detection results of the network model with the DCAP attention module added and the network model without the DCAP attention module added are visualized, and the visualization results are shown in Figure 5. It can be found from the figure that after adding the DCAP attention module, the confidence of model detection is increased, and it can detect incomplete and occluded targets that are difficult to detect.

![Figure 4. Visualization of IEPet model test results.](image)
4. Conclusion

Aiming at the problem of infrared environment perception in unmanned driving technology, this paper proposes a more powerful IEPet network model. The model increases the infrared target perception ability and detection confidence by adding the BottleneckCSP module and the DCAP attention module, as well as increasing the detection scale. The FLIR Thermal Datasets test results show that F1 Score can reach 0.7029, mAP can reach 0.8431, and the model parameter amount is only 4.99 million. This paper also sets up an ablation experiment to prove the promotion effect of each key module of the model. Therefore, the IEPet model proposed in this paper can be better applied in the field of infrared environment perception.

References

[1] Ghose, D., Desai, S. M., Bhattacharya, S., et al. (2019). Pedestrian Detection in Thermal Images Using Saliency Maps. (2019) IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 988-997.

[2] Qi, H. Y., Xu, T. H., Wang, G., et al. (2020) MYOLOv3-Tiny: A new convolutional neural network architecture for real-time detection of track fasteners. Computers in Industry., 123.

[3] Wang, C. Y., Bochkovskiy, A., Liao, H. (2020) Scaled-YOLOv4: Scaling Cross Stage Partial Network. https://arxiv.org/pdf/2011.08036.pdf.

[4] Liu, W., Anguelov, D., Erhan, D., et al. (2016) SSD: Single Shot MultiBox Detector. In: Computer Vision – ECCV 2016, pp, 21-37.

[5] Bu, D. F., Sun, Z. Y., Huang, R., et al. (2021) Unmanned night object detection based on improved SSD. Journal of Donghua University(Natural Science), 47(01): 63-69.

[6] Glenn. J. (2020) yolov5: v3.1 - Bug Fixes and Performance Improvements (Version v3.1). http://doi.org/10.5281/zenodo.4154370.

[7] Hou, Q., Zhou, D., Feng, J., (2021) Coordinate Attention for Efficient Mobile Network Design. https://arxiv.org/abs/2103.02907

[8] FLIR Systems, Inc. (2019) FLIR Thermal Dataseted. https://www.flir.cn/oem/adas/adas-dataset-form/