A Real-time UAV Object Detection System Design with FPGA Implementation

Junyu Tang*
Beijing University of Technology

Xin Zheng
Beijing University of Technology

Qiang Wu
Beijing University of Technology

Jinling Cui
Beijing University of Technology

ABSTRACT

This paper proposes Target detection system design and FPGA implementation based on YOLOX algorithm, in order to realize offline real-time image detection in a UAV platform with limited resources and power consumption. First, this paper studied the algorithm of YOLOX convolutional neural network,image fusion mechanism is added to this network, designed and trained the neural network. Secondly, an embedded edge computing system is designed to further speed up the target detection speed at the hardware level.In this paper, the above scheme is implemented at the board level. The test results show that the average recognition speed is 50frame/s on the system, which basically achieves the design goal of real-time detection.

CCS CONCEPTS

• **Computer systems organization**; • **Embedded software**;

KEYWORDS

object detection, Embedded software, UAV, FPGA, MPSoc

ACM Reference Format:
Junyu Tang*, Xin Zheng, Qiang Wu, and Jinling Cui. 2022. A Real-time UAV Object Detection System Design with FPGA Implementation. In *2022 8th International Conference on Computing and Artificial Intelligence (ICCAI '22), March 18–21, 2022, Tianjin, China.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3532213.3532284

1 INTRODUCTION

Since entering the 21st century, with the globalization, modernization and technicalization of the engineering field, the number of UAVs has exploded. These UAV are widely used in various fields of national security and national economic construction. The wide application of UAVs has also brought about a large number of aerial images. How to obtain one's area of interest or target from these images has become a current research hotspot.At present, the most common process of aerial image target detection is:images are collected by uav and sent back to ground base station, and then selected and marked by ground staff through human eyes.This method is not only inefficient but also has the problem of mismarking and missing marks.Therefore, it is urgent to realize the automatization and intelligentization of aerial image target detection.

The aerial image acquisition system is different from the general image acquisition system.As the UAV can take photos from various heights and angles, the camera position is not fixed and it works in different environments.This also makes aerial image target detection face the following severe challenges:

(1) Complex working environment.Uav often works in night, fog, rain, snow, sand and other weather, which tries to single visible or infrared image has significant limitations.

(2) Multi-scale.Because aerial photography is mostly used in overhead imaging, the scale of the target changes significantly in the imaging process due to the influence of the target imaging Angle and imaging distance changes, and the capture of small targets is particularly difficult.

(3) Complex imaging background.As the imaging range of aerial images is wider than that of general cameras, various possible backgrounds will appear in the field of vision, thus causing strong interference to the target detection process.

For the embedded equipment of UAV, most target detection algorithms are convolution neural network, so the selection of embedded equipment needs to be considered.Users need equipment with real-time, stability, low power consumption and high computing power. Therefore, a more advantageous system should be developed for uav target detection scenarios with miniaturization of convolutional neural network.

2 RELATED WORK

2.1 Target detection algorithm

Object detection is an important direction of computer vision, widely used in monitoring, medical, traffic and other fields.Early object detection methods such as Template Matching and Deformable Parts Models (DPM).However, this method is relatively complex and slow in calculation, so it is not effective in selecting and rotating stretched objects.

In 2012, AlexNet algorithm based on CNN made a great breakthrough in the field of image classification and provided a new idea for the field of target detection. After that, target detection algorithms based on deep CNN developed rapidly and were divided into two-stage and one-stage.Two-stage target detection is represented by R-CNN. Candidate regions are generated and CNN is used for feature extraction. Features are sent into multiple SVM classifications, regression correction boundingbox, and finally, NMS and edge detection are used for further correction. SPPNet adds the spatial pyramid pooling layer after the last convolutional layer of

ICCAI '22, March 18–21, 2022, Tianjin, China

ACM ISBN 978-1-4503-9611-0/22/03…$15.00
https://doi.org/10.1145/3532213.3532284

feature extraction to avoid information loss caused by stretching and clipping, establish the mapping relationship between original image and feature extraction, and avoid repeated convolution.

Fast-RCNN is improved on the basis of RNN by adding ROL pooling layer, changing SVM to Softmax, and placing classification and Boundingbox regression on the same network to avoid repeated convolution, integrate multiple tasks, reduce a lot of computing overhead, and further improve computing efficiency.However, the computational speed is limited by the generation of candidate regions.To solve this problem, the core idea of faster-RCNN is to entrust the generation of candidate regions to the network, which significantly improves the speed and accuracy of target detection.FPN algorithm uses the multi-scale pyramid of deep convolutional neural network to extract the features of each scale image, but it increases the inference time and consumes a lot of memory.R-FCN proposed position sensitive feature map to solve the problem between translation invariance of classification and translation variance of target detection, and reduce the amount of calculation.The continuous progress of two-stage algorithm increases the accuracy and speed of target detection, but it still fails to meet the requirements for scenes requiring real-time detection.If the first phase does not generate candidate boxes, one-stage detection treats each pixel as a potential target and then attempts to classify each region as a background or target.

In 2016, YOLO implemented one-stage target detection, dividing the image into multiple grids, regaining Boundingbox and trust values respectively, and finally, NMS filtered out low-score boxes.Yolo significantly improves the prediction speed, low background error detection rate and strong versatility, but the object location accuracy is poor, and the detection effect is not good for close objects and small objects.To solve this problem, YOLO-v2 and YOLO-v3 use multi-scale prediction to solve the problem of poor detection effect of small targets. YOLO-v3 use better basic classification network and classifier, greatly improve detection speed, lower background error detection rate, strong universality, but also have low accuracy of object location detection.Problems with low recall rates.[1]In YOLO-v4, Mosaic data enhancement was added on the basis of YOLO-v3. CSPDarknet53 and Mish activation function were used in the trunk network. SPP,FPN+PAN and other structures were used in Neck, and CIOU_Lossd and other operations were used in the output end. [2] Yolo-v5 has more excellent detection speed, adaptively calculates the best anchor frame value in different training sets for different data sets, ADAPTS to picture scaling, uses Focus structure to reduce the amount of calculation, including two KINDS of CSP structure in backbone and Neck respectively, which enhances CNN's learning ability.The CSPnet structure in NECK strengthens the ability of network feature fusion by reducing computing bottleneck and memory cost. [3]

2.2 Image fusion algorithm

Image quality has a very important effect on the performance of target detection.Commonly used sensors such as visible light sensor, long wave infrared sensor, medium wave infrared sensor, near infrared sensor and so on.Visible light image generally use RGB three-channel image, through the reflection optical imaging, contains more texture information and contrast, is more suitable for human visual perception, but restricted by environmental influence, insufficient lighting, illumination image, a rainy day, fog and other weather and time factor will greatly affect the quality of the visible light image and imaging effect,It has great influence on the accuracy of target detection algorithm.Obtain the infrared image is formed by infrared radiation image, highlight the hot targets hidden in the background, compared with light contains the details of the information can be less, low contrast and resolution is poor, low signal-to-noise ratio, gray distribution features of target reflection and wireless sexual relationship, but in the visible light is not suitable for the scenario, the infrared image has not affected by light intensity, the advantage of the weather,Widely used in poor light and obstructed vision and other scenarios. [4]

Therefore, giving full play to the advantages of the two images can greatly increase the performance of target detection.In this regard, a method of image fusion has emerged, which integrates infrared image and visible image into the same image to retain as much information as possible, which is of great help to improve the performance of target detection algorithm in complex scenes.In recent years, infrared and visible image fusion methods have developed rapidly, including PCA, ICA, etc.Transform domain fusion methods: pyramid transform, wavelet transform, NSCT and NSST, etc.Feature extraction based fusion methods: Sparse representation (SR), pulse coupled neural network (PCNN);Deep learning based fusion methods: convolutional neural network (CNN) and generative adversarial network fusion methods.The traditional method of image fusion mostly divides the image into different domains, uses the fusion strategy to superimpose two kinds of images in different domains, and then converts them into the fused image through transformation decoding.The image fusion algorithm based on deep learning can extract the information of different images into new images by extracting features.Although the image fusion method can fuse different image information into one image, it will inevitably cause information loss, image distortion, and increase the algorithm time and other problems.

2.3 Embedded computing platform

This paper compares several platforms for deep learning algorithm transplantation, including GPU, ARM, DSP, FPGA and heterogeneous platform.

It is a good choice to use CPU and GPU for their respective functions, but GPU also has the problem of high power consumption. Considering the limited power carried by UAV platform, if it is used for calculation in large quantities, the endurance of UAV will be greatly reduced, so this is not a suitable solution.ARM is the most widely used uav platform.This kind of chip is similar to CPU. Although it has a high main frequency, it has less computing resources and cannot meet the requirements of large amount of computing.DSP is a microprocessor designed for high speed computing. It adopts Harvard structure and stores data and programs separately.At the same time with special multiplier, so that its computing power is stronger than ARM.However, DSP also has some disadvantages, such as the internal structure has been solidified in the design, can not be modified according to the specific situation, can not adapt to a variety of deep learning network.

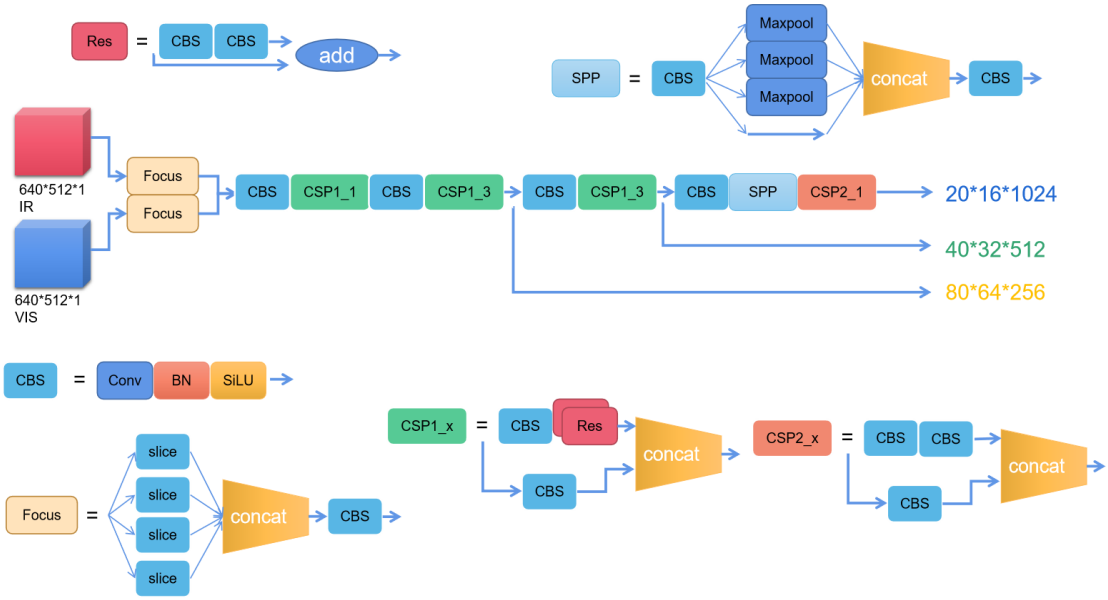

Figure 1: Backbone network with dual channel input

FPGA, also known as programmable logic device, has high flexibility.Its unique hardware programming method makes FPGA has the speed only second to ASIC devices.At the same time, FPGA can be programmed to realize multilevel pipelining, which has better parallelism.However, if a neural network is realized through FPGA programming, a huge amount of programming is needed and the programming is difficult.Heterogeneous platform has attracted more and more attention.For example, DSP and FPGA or ARM and FPGA can be combined on a chip, using on-chip communication instead of off-chip communication, and each part performs its own good things respectively, so as to achieve higher efficiency.Xilinx's ZYNQ MPSoC is a typical example. [5]

3 METHOD

This design proposes to add the image fusion structure into the target detection network. According to the requirements of real-time and accuracy, and taking into account the detection of small targets, we choose to use YOLOX-Tiny to improve the design scene.Because YOLO-v4 and YOLO-v5 are over-optimized, and YOLO-v3 has advantages over YOLOv4 and YOLOv5 for small target detection, YOLOX is based on the version of YOLOV3-SPP with better performance.

On the input side of the algorithm, Mosaic and Mixup data enhancement methods are used.Mosaic enhancement is an enhancement strategy introduced by YOLO-v3, which can be splicing by random scaling, random clipping and random arrangement. The effect of Mosaic enhancement is significantly improved for small targets, and it is suitable for the scene with many small targets detected in this design.Mixup data enhancement can stably improve detection accuracy by filling images on both sides, up and down, or up and down, left and right, to improve detection accuracy. It has a significant improvement effect on scenes with insufficient data sets. [6]

CSPDarknet is used in the algorithm backbone network, and two kinds of images after registration are input into two Focus modules, which simultaneously play the role of image fusion coding and special zone features, and Focus operation is used to reduce the number of parameters and calculation.After connecting the output of two channels through cat, feature extraction is carried out in the continuous CBS layer and CSP layer in the input trunk network.The three output sizes of the trunk network are shown in Figure 1, which are divided into three scales for subsequent prediction. [7]

In the Neck structure, the benchmark model YOLO-v3 uses FPN structure for fusion, and FPN transmits and fuses the high-level feature information by up-sampling from top to bottom, so as to obtain the feature graph for prediction.In this design YOLOX, FPN+PAN is used to make it easier for the bottom information to be transmitted to the top of the high level.The number of information transfer layers is reduced, thus reducing the amount of computation [8].The Neck structure is shown in Figure 2

The predictive head uses three Decoupled Head, which can significantly increase AP value by one percentage point and increase the convergence rate compared with the end-to-end model.After balancing speed and performance, convolution is used to reduce dimension first, and then 3*3 convolution kernels are used for each of the two branches, which increases the lowest network parameters and significantly improves performance [9].As can be seen from the figure, each Decoupled Head has three branches, corresponding to the type of the target box [10], whether the target box is foreground or background, and the coordinate information of the target.In this design, the Anchor Free method is used, compared with the Anchor Base method in YOLO-v3, v4 and v5, which has the advantage of reducing the number of parameters by about two-thirds [11].The prediction structure with decoupling head is show in Figure 3

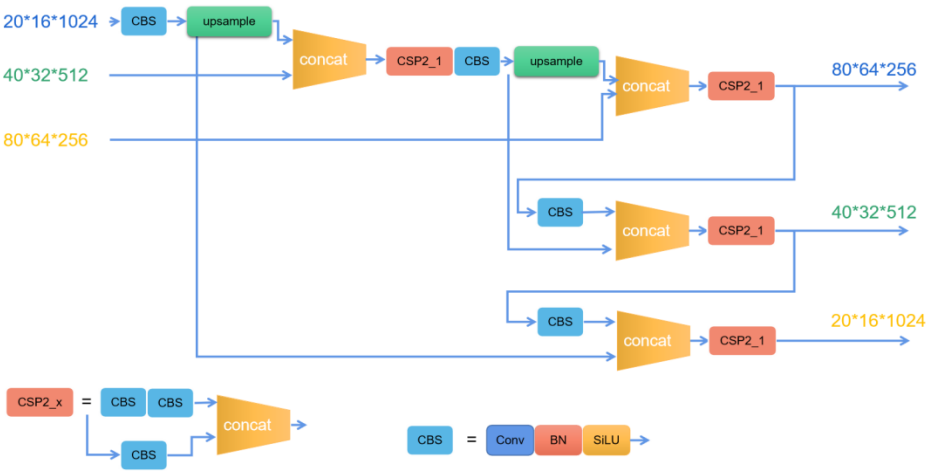

Figure 2: The neck structure of this algorithm

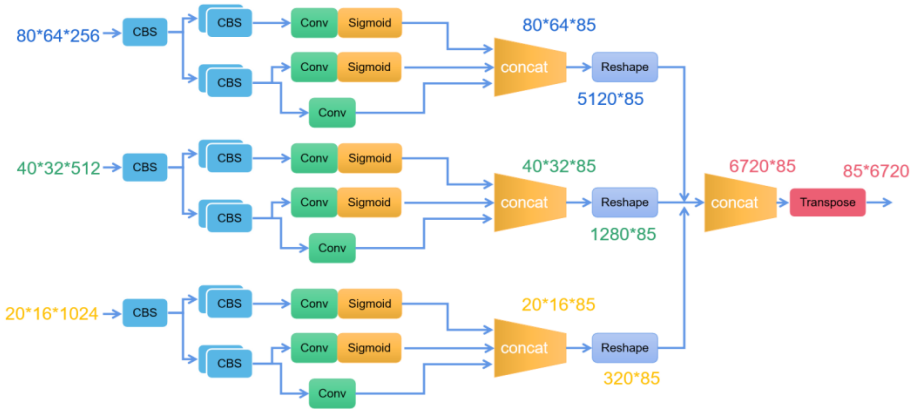

Figure 3: Prediction structure with decoupling head

4 SYSTEM IMPLEMENTATION

4.1 module design

UAV embedded system chip needs to provide powerful parallel computing capability, rich image interface, as well as low power consumption and heat as possible.XCZU9EG belongs to Xilinx ZYNQ UltraScale+ series heterogeneous multi-processor SoC platform.Support current image sensor mainstream interface Camera Link, HDMI, LVDS, etc.FPGA contains a large number of operation units such as LUT and DSP, a variety of high-speed communication interfaces and rich IO pins [12].High-speed transceivers are included to bridge high-speed serial buses such as PCIe, SRIO and HyperLink.External support EMMC card or SD card for system storage, and use SATA interface for real-time acquisition of a large number of image data.DDR can meet the requirements of intermediate data and image cache in the operation of neural network.The figure 4 below shows FPGA interface design [13].

In this design, the sensor uses Camera Link interface for image acquisition. The sensor uses TRIGGER signal for synchronization to ensure that each frame of image is collected at the same time to facilitate subsequent image registration and fusion processing.After the image is collected, the image is preprocessed, including image denoising, image motion blur elimination and image registration [14]. Image cache to DDR, image transmission using image dedicated VDMA scheduling, flexible and stable use. The FPGA interface design is show in figure 5

4.2 Neural network deployment

Edge calculation is performed using Xilinx's official Vitis AI tool, which can perform a secondary optimization of the trained model: prune and quantize.

When the model is determined, the number of parameters of the model has been determined, but obviously not all models need the same parameters for characterization, which means that there is a great possibility of redundancy of parameters in the model.That is, there are some parameters whose existence has little effect on the accuracy of the model.Pruning is usually done in a trained model.In the process of pruning, it is often necessary to first determine the evaluation index of each parameter, which represents the importance of each parameter.Pruning is to cut out unimportant

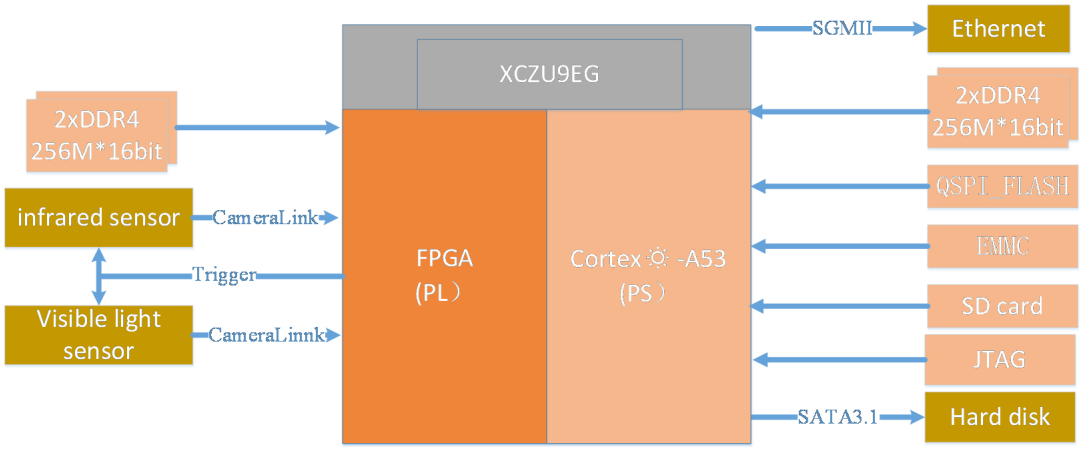

Figure 4: FPGA interface design

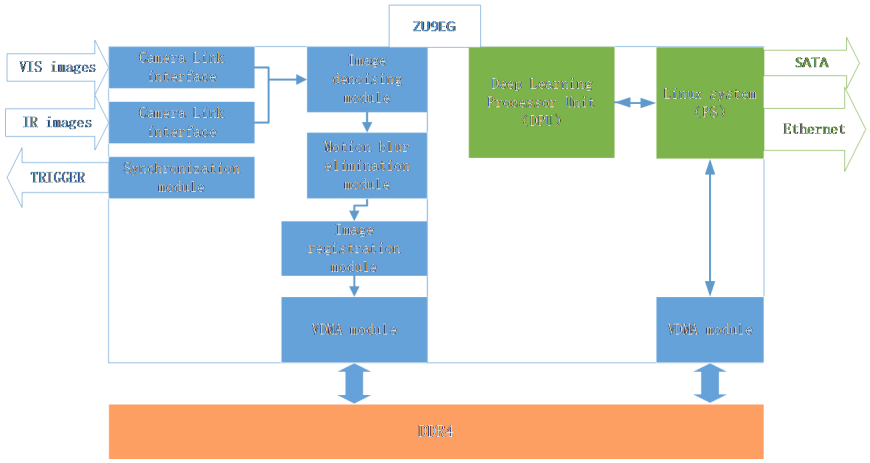

Figure 5: FPGA interface design

connections and convolution kernels according to the importance degree, so as to achieve the purpose of reducing parameters. [15]

Due to the need to capture small gradient changes in the training process of neural networks, this means that a complex and high-precision model is required.Therefore, in the process of model training, the data type of parameters is mostly set as 32 bit floating point type [16].In the process of model deduction, speed will become the first priority in the case of ensuring accuracy, so it is necessary to convert the data type of parameters into int8 and other low-precision types [17].

5 RESULT

5.1 Algorithm results

As shown in Table 1, the algorithm in this paper is based on Yolox-Tiny. Compared with Yolox-x, although mAP is decreased, the number of parameters and computation amount are greatly reduced.Therefore, we optimize the algorithm based on Yolox-Tiny.

Table 1: Parameter quantity and calculation quantity

	Params(M)	FLOPs(G)
YOLOX-x	99.1	281.9
YOLOX-tiny	5.06	6.45
YOLOX-tiny (Algorithm in this paper)	5.03	5.91

The algorithm in this paper reduces the number of parameters and the amount of calculation, improves the calculation speed, and optimizes the detection effect of small targets.

As shown in Table 2, for the training set in the uav detection scene, using the basic Yolox-Tiny to detect single visible or infrared images, mAP can reach 0.636 and the calculation amount is 6.45. In order to make the algorithm practical in different environments, we tried to extract features using image fusion algorithm and detect Yolox-Tiny in the fusion image. We found that mAP was

Table 2: Performance of mAP on different YOLOX-tiny

	mAP (0.5:0.95)	FLOPs(G)
YOLOX-tiny (VIS)	0.636	6.45
YOLOX-tiny (IR)	0.653	6.45
YOLOX-tiny (Fusion + detection)	0.667	6.45+0.6
YOLOX-tiny (Algorithm in this paper)	0.678	5.914

Table 3: Comparison with CPU and GPU

	CPUI7-7700	GPUGTX-1060	FPGAZU9EG
Max power consumption (W)	65	120	20
Frequency(GHz)	2.9	1.5	0.33
FLOPs(G)	149	4000	4100

significantly improved, but the amount of computation was also increased.Finally, our algorithm has the highest MAP-0.678 and significantly reduces the amount of computation to 5.914, which has the balance advantage of detection accuracy and speed.

Table 3 shows some performance parameters of single-channel target detection using the basic algorithm Yolox-Tiny, as well as the performance parameters of the proposed algorithm.Compared with yolox-Tiny algorithm using infrared images alone or visible images alone, the accuracy of the algorithm in this paper has been improved. Under IoU= 0.50:0.95, the mAP has been improved by 0.03∼0.04, and it has a more significant improvement in complex scenes.And when IoU= 0.50:0.95, recall rate increased by 0.02∼0.03.When area is small, medium and large, the algorithm in this paper also has obvious performance improvement.

5.2 System results

Xilinx DPU is used as a neural network calculator in this paper. The deployment size is B4096, which can process 8 pixels in parallel, with a maximum of 4096 operations per cycle. The clock frequency can reach 333Mhz, and the GOPs can reach 4100.For visible and infrared images, the size is 640*512*8bit, and the processing capacity can meet 50FPS.As can be seen from Table 3, COMPARED with CPU, FPGA has higher calculation examples and lower power consumption; compared with GPU, FPGA has basically the same calculation examples and lower power consumption, but its disadvantage is higher price.

6 SUMMARY

In this paper, a target detection system based on YOLOX algorithm and FPGA implementation are proposed to realize offline real-time image detection on uav platform with limited resources and power consumption. Experiments show that the author's research basically achieves the design goal of real-time detection.

REFERENCES

[1] Detection and Tracking for High Resolution Video: Overview and State-of-the-Art," 2019 5th International Conference On Computing, Communication, Control And Automation (ICCUBEA), 2019, pp. 1-9, doi: 10.1109/ICCUBEA47591.2019.9128812.pp

[2] Fang W,Wang L,Ren P.Tinier-YOLO: A Real-time Object Detection Method for Constrained Environments[J]. IEEE Access, 2019, PP(99):1-1.

[3] Bi F, Yang J.Target Detection System Design and FPGA Implementation Based on YOLO v2 Algorithm. IEEE, 2019

[4] Wu X,Sahoo D,Hoi S . Recent Advances in Deep Learning for Object Detection[J]. Neurocomputing, 2020, 396..

[5] Gao Xinbo,Mo Mengjingcheng,Wang Haitao,et al.Recent Advances in Small Object Detection[J].Journal of Data Acquisition and Processing,2021,36(03):391-417.

[6] X Zhu, Lyu S,X Wang, *et al.* TPH-YOLOv5: Improved YOLOv5 Based on Transformer Prediction Head for Object Detection on Drone-captured Scenarios[J]. 2021.

[7] Wang X, Wang S,Cao J,et al.Data-driven based tiny-YOLOv3 method for front vehicle detection inducing SPP-net[J]. IEEE Access, 2020, PP(99):1-1

[8] Olugboja A, Wang Z, Sun Y. Parallel Convolutional Neural Networks for Object Detection[J]. Journal of Advances in Information Technology Vol, 2021, 12(4).

[9] Zhou Zhili, Li Yujiang, Zhang Yulan *et al.* Residual visualization-guided explainable copy-relationship learning for image copy detection in social networks[J] Knowledge-Based Systems, 2021, 228

[10] Qiu Zhi-cheng, Huang Zi-qian A shape reconstruction and visualization method for a flexible hinged plate using binocular vision[J] Mechanical Systems and Signal Processing, 2021, 158

[11] Kim Jin-Kook, Jung Sunghoon, Park Jinwon *et al.* Arrhythmia detection model using modified DenseNet for comprehensible Grad-CAM visualization[J] Biomedical Signal Processing and Control, 2022, 73

[12] Babu P , Parthasarathy E . Optimized Object Detection Method for FPGA Implementation[C]// 2021 Sixth International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET). 2021.

[13] Kim S , Na S , Kong B Y , *et al.* Real-Time SSDLite Object Detection on FPGA[J]. IEEE Transactions on Very Large Scale Integration (VLSI) Systems, 2021, PP(99):1-14.

[14] Attarmoghaddam N , Li K F . An Area-Efficient FPGA Implementation of a Real-Time Binary Object Detection System[M]. 2021.

[15] J. Wang and S. Gu, "FPGA Implementation of Object Detection Accelerator Based on Vitis-AI," 2021 11th International Conference on Information Science and Technology (ICIST), 2021, pp. 571-577, doi: 10.1109/ICIST52614.2021.9440554.

[16] E. Rzaev, A. Khanaev and A. Amerikanov, "Neural Network for Real-Time Object Detection on FPGA," 2021 International Conference on Industrial Engineering, Applications and Manufacturing (ICIEAM), 2021, pp. 719-723, doi: 10.1109/ICIEAM51226.2021.9446384.

[17] H. Zhang, J. Jiang, Y. Fu and Y. Chang, "Yolov3-tiny Object Detection SoC Based on FPGA Platform," 2021 6th International Conference on Integrated Circuits and Microsystems (ICICM), 2021, pp. 291-294, doi: 10.1109/ICICM54364.2021.9660358.