IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 25 June 2024, accepted 3 August 2024, date of publication 19 August 2024, date of current version 29 August 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3445950

APPLIED RESEARCH

DETR Novel Small Target Detection Algorithm Based on Swin Transformer

XU FENGCHANG[1,2], **RAYNER ALFRED**[1], **(Member, IEEE), RAYNER HENRY PAILUS**[1], **LYU GE**[2], **DU SHIFENG**[2], **JACKEL VUI LUNG CHEW**[3], **LI GUOZHANG**[4], **AND WANG XINLIANG**[5]

[1]Creative Advanced Machine Intelligence Research Centre, Faculty of Computing and Informatics, Universiti Malaysia Sabah, Jalan UMS, Kota Kinabalu, Sabah 88400, Malaysia
[2]Department of Information Engineering, Shandong Light Industry Vocational College, Zhoucun, Zibo, Shandong 255300, China
[3]Faculty of Computing and Informatics, Universiti Malaysia Sabah Labuan International Campus, Labuan 87000, Malaysia
[4]College of Information Engineering, Hainan Vocational University of Science and Technology, Haikou, Hainan 571126, China
[5]Binzhou Civil Air Defense Engineering and Command Support Center, Bincheng, Binzhou, Shandong 256600, China

Corresponding authors: Xu Fengchang (haiquan2018@sina.com), Rayner Alfred (ralfred@ums.edu.my), and Rayner Henry Pailus (rayner.pailus@ums.edu.my)

This work was supported in part by Shandong Light Industry Vocational College under Grant 2022GB04; and in part by Universiti Malaysia Sabah (UMS), Malaysia.

ABSTRACT A small target object refers to an object whose relative size of the bounding box is very small, usually the ratio of the width of the bounding box to the width and height of the original image is less than 0.1, or the ratio of the area of the bounding box to the area of the original image is less than 0.03, or the absolute size is less than 32*32 pixels. It has important applications in industrial defect detection, medical image processing, intelligent security, unmanned driving, and many other fields. Although great progress has been made in the field of target detection, which is limited to large target objects, due to the challenges of small size, inconspicuous features and insufficient data samples, the accuracy and speed of small target detection are low. To solve this problem, this paper proposes a novel small target object detection algorithm model: Swin Transformer's DETR. In this algorithm, Swin Transformer is used as the backbone to extract the global features and local information of small targets, and a three-layer feature pyramid structure is used for feature fusion at the Neck layer to improve the calculation efficiency and model accuracy. Secondly, the detector is optimized, and the detector is replaced by two stages, and the ReLU activation function of FFN layer is replaced by the latest SwiGLU activation function, to avoid the problems of gradient disappearance and explosion and enhance the nonlinearity of the algorithm model. Large resolution size input is adopted on Tiny Person dataset, and its input value is set to [1400,800]. The above analysis is carried out on VOC and Tiny Person datasets, and the detection rates of small target objects are 88.9% and 48.3% respectively. The results show that the Swin Transformer's DETR algorithm model proposed in this paper performs well on various datasets, and has strong generalization ability, stability and accuracy in different scenarios and datasets, which is higher than other algorithm models.

INDEX TERMS Swin transformer, DETR, small target detection, deep learning.

I. INTRODUCTION

With the rapid development of deep learning, great progress has been made in the field of target detection [1]. Various excellent algorithm models have emerged one after another, including classification-based target detection models focusing on improving accuracy (such as R-CNN [2], Faster R-CNN [3]) and regression-based models focusing on speed (such as YOLOv5 [4], YOLOv8 [5], DSSD [6], and new models based on Transformer (such as LSTR [7], DETR [8], TNT [9]. The above model gets rid of the limitation of traditional manual design features and greatly promotes the development of computer vision. Target detection technology plays an important role in industrial defect detection, human body shape recognition, intelligent security, unmanned driving, and other fields, which provides great convenience for human life and production. Although the detection of large target objects has achieved good results, there are still

The associate editor coordinating the review of this manuscript and approving it for publication was Mouloud Denai.

great challenges and gaps in the detection of small target objects.

Taking the DETR algorithm model as an example, we use ResNet-50 or ResNet-101 backbone network to detect large target objects. DETR has achieved a good balance between accuracy and speed and has excellent real-time processing ability. Its flexibility and generalization make it applicable to various target detection scenarios and datasets. On the COCO [10] dataset, the AP50 value of DETR is basically the same as that of Faster R-CNN, which are 63.1% and 60.5% respectively, while the AP value is 43.3% and 39% respectively. However, on Tiny Person small target dataset, the AP50 of Faster R-CNN is reduced to 43.5%, while that of DETR is less than 10%, which shows that DETR is not adaptable to this type of dataset. The AP value is seriously affected. The performance of DETR is greatly affected by the size of the target object, and the detection ability of small target objects is the main factor.

Small target detection is one of the key challenges in the field of target detection. Compared with large target objects, small target objects are extremely sensitive to image resolution, and factors such as scale imbalance, background information, target density, light interference and shortage of samples all increase the detection difficulty. The current challenge is how to improve the detection accuracy and speed of small target objects while maintaining the detection accuracy of large objects.

Aiming at the above problems, we put forward an innovative small target object detection algorithm model: Swin Transformer's DETR, which can achieve higher detection accuracy and faster detection speed under the same experimental conditions. Compared with the advanced algorithms such as DETR, Faster R-CNN and YOLOv5, the proposed algorithm improves the small target detection accuracy of VOC datasets by 11 AP at the resolution of 800 × 600 pixels. Under the resolution of 1400 × 800 pixels, the small target detection accuracy of Tiny Person dataset is improved by 8 AP, and other indicators are also significantly improved, and the detection speed per hour/epoch is also improved by 2.5 hours.

The main contributions of this paper are summarized as follows:

i. We found the limitation of DETR algorithm in small target object detection, and proposed an innovative small target object detection algorithm, namely Swin Transformer's DETR, to improve the detection performance of small target objects. It is also compared with DETR, YOLOv5 and Faster R-CNN.
ii. The designed new backbone network can effectively extract the global features and local information of small target objects and adopt a three-layer pyramid structure for feature fusion. The diversity and novelty of data samples are enhanced to improve the adaptability of the algorithm model to complex scenes. The latest activation function SwiGLU was introduced to solve the problems of gradient disappearance and explosion.
iii. Three experiments on VOC and Tiny Person datasets, compared Swin Transformer's DETR have been carried out with other most advanced detectors, and proved that the algorithm proposed in this paper is more effective. In terms of accuracy, Swin Transformer's DETR completely surpasses the algorithm models such as DETR, especially on VOC datasets, the detection accuracy of small target objects is improved by 11 AP, and the detection speed is improved by 2.5 hours per epoch.

The rest of this paper is organized as follows: Section II describes our proposed algorithm model in detail and optimizes it in detail. Section III introduces the experimental environment configuration studied in this paper and compares and verifies the proposed algorithm model on VOC and Tiny Person datasets. Section IV compares Swin Transformer's DETR with other target detection algorithms. Finally, section V summarizes the research work of this paper.

II. METHOD

In the foregoing introduction, DETR is equivalent to Faster R-CNN in detecting large target objects. However, when detecting small target objects, because the traditional CNN network is used as the backbone in the DETR algorithm model, CNN mainly relies on local field of view and weight sharing to capture the spatial characteristics of images. In view of the problems of low resolution and inconspicuous local features of small target objects, the detection accuracy of small target objects is low. In this paper, a small target object detection algorithm model based on Swin Transformer's DETR is proposed. The overall structure diagram of the network is shown in Figure 1 that illustrates how Swin Transformer's DETR contributes to detect small target objects.

First, the CNN backbone network in DETR is replaced with the latest Swin Transformer and Swin Transformer is applied as the backbone network to extract the features of small target objects. Swin transformer adopts the hierarchical attention mechanism in stages, which can model the target object on different scales, so it has stronger modeling ability. It provides a more flexible feature extraction mechanism for the target detection task and can better capture the global information and local features of the target object, thus laying a solid foundation for improving the target detection accuracy. Based on the features extracted by Swin Transformer, the subsequent feature extraction processing is carried out to capture the semantic information of small target objects more comprehensively.

Next, to further optimize the performance of neural network and improve the performance and robustness of the algorithm model, in view of the SwiGLU's characteristics of non-monotonicity, smoothness, and universality. This paper adopts the latest activation function SwiGLU instead of ReLU activation function adopted in the first layer.

The purpose of this improvement is to introduce a more elaborate screening mechanism, by generating some candidate regions in advance, and then making specific classification and regression prediction. The two-stage detection method can deal with different scales and shapes of targets more flexibly, thus improving the adaptability of the algorithm model to complex scenes. Through this innovative improvement, the performance of Swin Transformer's DETR algorithm model in small target object detection and other tasks is optimized.

The Hungarian algorithm, through its global optimal matching mechanism, can effectively reduce matching errors. This is particularly useful in detecting small objects, as it minimizes matching errors caused by the weakening of features in small objects.

A. BACKBONE

According to the overall architecture diagram of Swin Transformer network, the input small target image I is firstly partitioned by Patch Partition, and then sent to Linear Embedding module to adjust the number of channels. Secondly, feature extraction and down-adoption are carried out through stage 1, 2, 3 and 4, and finally the prediction results are obtained. After each stage, the image size will be reduced to half of the original size, and the number of channels will be expanded to twice the original number. The Swin Transformer block in each stage consists of two continuous transformer blocks, one of which is based on W-MAS and the other is based on SW-MSA. The calculation performance is improved by using window and translation window mechanism.

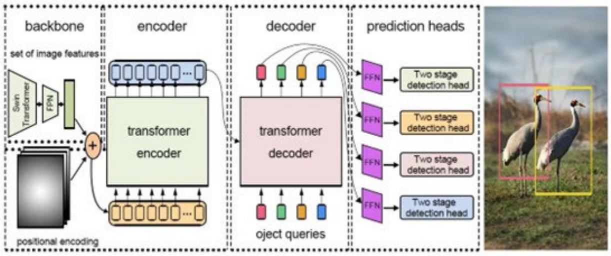


FIGURE 1. Simplified network architecture diagram of Swin Transformer's DETR.

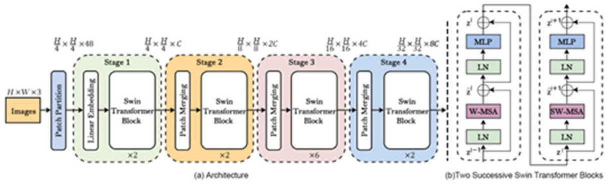


FIGURE 2. (a) The architecture of a Swin Transformer (Swin-T); (b) Two successive Swin Transformer blocks. W-MSA and SW-MSA are multi-head self-attention modules with regular and shifted windowing configurations respectively.

1) PATCH MERGING MODULE

Patch Merging module firstly splices the patch blocks with the size of H×W, concatenate them in the channel dimension to form a new feature map,

$$\frac{H}{2} \times \frac{W}{2} \times 4C, \tag{1}$$

and performs Layer Normalization operation on the output of each stage for regularization. Then, through a linear layer, a feature map with a new size of

$$\frac{H}{2} \times \frac{W}{2} \times 2C, \tag{2}$$

is formed, and the down sampling process of the feature map is completed. In this process, the size is reduced to 1/2 and the channel is expanded to 2 times.

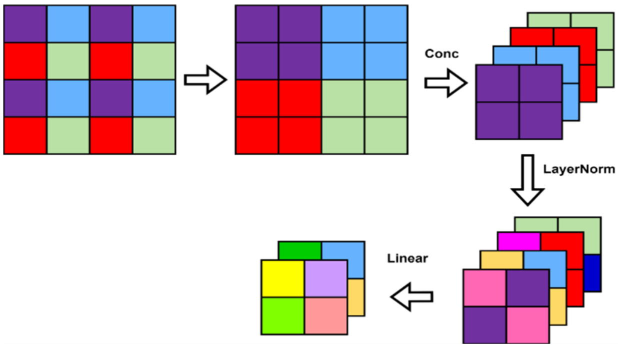


FIGURE 3. Patch merging module.

2) W-MSA MODULE

Swin-T decomposes MSA into several windows with fixed size to form W-MSA. In W-MSA, the pixels in each window can only do inner product operation with other pixels in the window, to obtain the local information of small target objects. This blocking method greatly reduces the calculation amount and improves the operation efficiency of the network structure. The calculation formulas of MSA and W-MSA are as follows,

$$\Omega\,(\mathrm{MSA}) = 4hwC^2 + 2\,(hw)^2\,C, \tag{3}$$

and

$$\Omega\,(\mathrm{W-MSA}) = 4hwC^2 + 2M^2hwC, \tag{4}$$

Here h, w and C represent the height, width, and depth of the feature map respectively, and M represents the size of each window. If h=w=112, M=7 and C=128, it can be calculated that W-MSA saves 40124743680 FL0Ps.

3) SW-MSA MODULE

Although W-MSA can reduce the amount of calculation by dividing windows, it can't interact with each other, which leads to the narrowing of its receptive field and the inability to get more global and accurate information, thus affecting the accuracy of the network. To solve the above problems, information interaction between different windows is realized by sliding windows. The W-MSA of the offset window constitutes the SW-MSA module. Based on the W-MSA, its window is offset by two Patch to the lower right corner to form nine blocks with different sizes. Then, the nine blocks are translated and spliced into four blocks with the same size corresponding to the W-MSA by using cyclic shift, so that the information data can be returned to the original position through reverse cyclic shift. With the help of SW-MSA mechanism, the MSA calculation of pixels in offset windows is completed, and the information exchange of pixels between different windows is realized, thus indirectly expanding the ‘‘receptive field’’ of the network and improving the utilization rate of information.

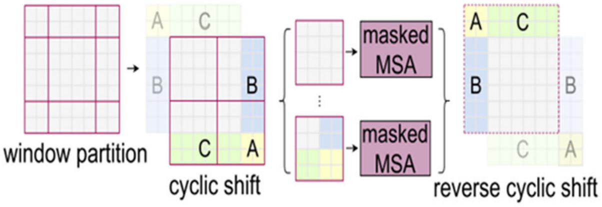


FIGURE 4. SW-MSA module.

4) RELATIVE POSITION BIAS MECHANISM

Swin-T network introduces the relative position offset mechanism in the calculation of Attention to improve the overall accuracy. Through this mechanism, the accuracy can be improved by 1.2%∼2.3%. Taking the 2×2 feature map as an example, firstly, each block in the feature map is numbered in absolute position, and the absolute position index of each block is obtained. Then, the relative position between each block and other blocks is calculated. The calculation method is to subtract the absolute position index of this block from the absolute position index of other blocks to get the relative position index matrix of each block. Finally, the relative position index matrix of each block is flattened and connected to form the relative position index matrix of the whole feature map. The specific calculation process is shown in the figure below.

Swin-T does not use the relative position index matrix in the form of two-dimensional tuples but maps the relative position index in the form of two-dimensional tuples into one-dimensional relative position offset, thus forming the corresponding matrix. The specific mapping method is as follows:

1. Add M-1 to the row index and column index of the corresponding relative position respectively.

2. Multiply the row index by 2M-1.

3. Add the row index and the column index, and then use the corresponding relative position offset table for mapping to get the final relative position offset B.

The specific calculation process is shown in Figure 6.

The following is the calculation formula of Attention with relative position offset mechanism,

$$\mathrm{Attention}\,(\mathrm{Q},\mathrm{K},\mathrm{V}) = \mathrm{Softmax}\left(\frac{\mathrm{QKT}}{\mathrm{d}}+\mathrm{B}\right)\mathrm{V}, \qquad (5)$$

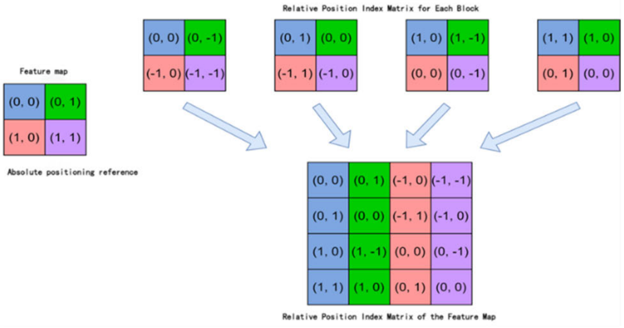


FIGURE 5. Flow chart of relative position index calculation.

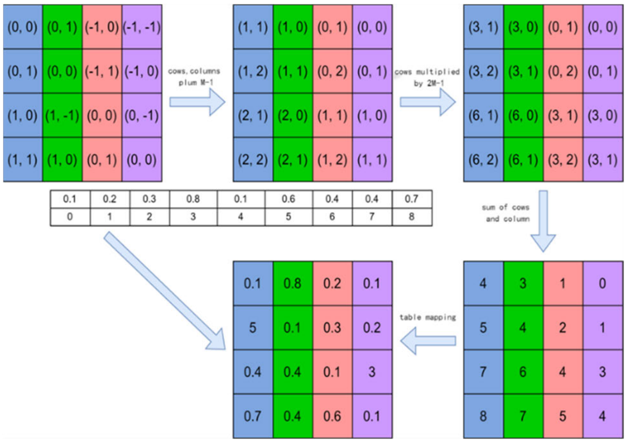


FIGURE 6. Flow chart of relative position offset.

where B is the relative position offset obtained in the above calculation.

Firstly, the global features and local information of small target image are extracted by Swin Transformer backbone network, and then the extracted features are fused by three-layer pyramid feature structure. This three-layer pyramid structure can help integrate shallow network features, as the characteristics of small targets are mostly concentrated in the shallow layers. This enables better regression of small targets.

The two-stage detection head alleviates the issue of sample imbalance by performing an initial round of filtering through ROI. Another challenge in small object detection is sample imbalance, as there are too few pixels representing small objects.

5) FFN LAYER

In this paper, SwiGLU activation function, which is outstanding in Transformer field, is used to replace ReLU activation function for activation output. SwiGLU activation function is a combination of Swish and GLU activation functions, in which Swish function is used to gate the linear function of GLU, so that SwiGLU can make comprehensive use of the advantages of Swish and GLU and overcome their respective disadvantages. SwiGLU has been proven to perform well in many tasks due to Swish and GLU, including image classification, language modeling and machine translation. The activation function has the characteristics of non-monotonicity, smoothness, and universality, and can selectively activate neurons according to the received input by adopting the gating mechanism. The application of SwiGLU activation function in Swin Transformer's DETR can further optimize the network performance and significantly improve the performance and robustness of the algorithm model.

SwiGLU has been proved to be superior to Swish and GLU in many tasks, including image classification, language modeling and machine translation. The activation function has the characteristics of non-monotonicity, smoothness, and universality, and can selectively activate neurons according to the received input by adopting the gating mechanism. At present, due to the excellent performance of SwiGLU in the field of Transformer, the applicability of other new activation functions such as GLU, GTU, Bilinear, ReGLU and GEGLU is limited. Using SwiGLU activation function in Swin Transformer's DETR can further optimize the network performance and significantly improve the performance and robustness of the algorithm model. Where w, v, b, c and β are trainable parameters.

The mathematical expression of SwiGLU function is as follows,

$$\mathrm{SwiGLU(x) = Swish_{\beta}(xW + b) \otimes (xV + bc)}, \tag{6}$$

and the image of SwiGLU function is shown in the following figure.

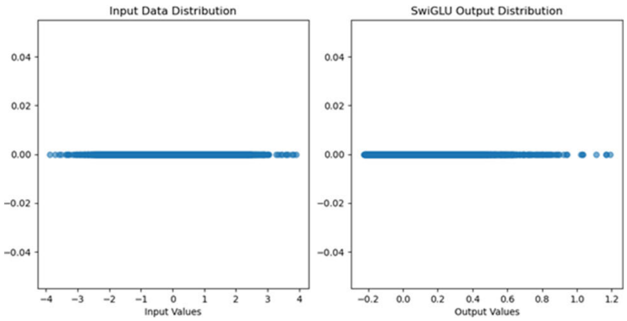


FIGURE 7. **SwiGLU function.**

B. TWO STAGE DETECTION HEAD

Swin Transformer, as the backbone network, extracts the global features and local information of small target objects, and after feature fusion with a three-layer pyramid structure, it enters two-stage detection. Firstly, a small object pre-selection box that may be detected is generated, and then fine-grained object detection is carried out. The two-stage detection head is shown in Figure 8.

C. LOSS FUNCTION

Since DETR infers a fixed size of n prediction sets in a single pass through the decoder. However, the main challenge in the training process is how to accurately predict the category, position, and size of the object. Our loss is to produce an optimal binary distribution between the predicted object and the real object, thus optimizing the object-specific loss. In this paper, an improved Hungarian algorithm is used to allocate loss and BOX loss. The mathematical expressions are as follows,

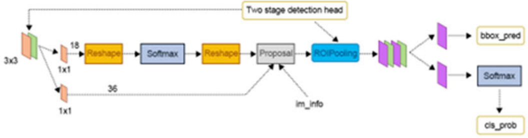


FIGURE 8. **Two stage detection head.**

$$\mathrm{S}\hat{\sigma} = \underset{\sigma \in \ell_N}{argmin} \sum_{i}^{N} L_{match}(y_i, \hat{y}_{\sigma(i)}), \tag{7}$$

and

$$L_{H\mathrm{ungarian}}\left(y, \hat{y}\right) = \sum_{i=1}^{N} \left[-log\, \hat{p}_{\sigma(i)}(c_i) + 1_{\{c_i \neq \emptyset\}} L_{box}(b_i, \hat{b}_{\hat{\sigma}}(i)) \right]. \tag{8}$$

III. EXPERIMENTS

A. DATASET, CONFIGURATION AND EVALUATION MATRIX

1) DATASET

In this paper, two public datasets are used to test and verify the algorithm model: PASCAL VOC [12] and Tiny Person [13]. Image types, input image sizes, and the number of images in the training set and verification set are shown in the following table.

PASCAL VOC dataset originated in 2005, and was originally used for image classification tasks, and gradually extended to target detection, semantic segmentation, and other fields. Among them, PASCLA VOC2007 and PASCAL VOC2012 are two important dataset versions. VOC 2007 dataset contains 20 categories of targets, which are divided into three parts: training, verification, and testing. The VOC 2012 dataset is an upgraded version of VOC 2007, which can be used for image classification, object detection and semantic segmentation tasks.

The image of PASCAL VOC dataset is shown in the following figure.

Tiny Person dataset is published by the University of Chinese Academy of Sciences and is specially designed for tiny target detection. In view of the low resolution of people in sea and beach scenes, this dataset contains two categories: Sea Person and Earth Person. Sea Person includes people who are in the boat, in the water or more than half of their bodies are in the water, while the rest belong to Earth Person. The dataset contains 1610 tagged images and 759 unlabeled images, mainly from the same video set, with a total of 72651 annotations. Tiny Person dataset image is shown in Figure 10.

TABLE 1. Summary of used datasets.

Data name	Input image size	Training set (sheet)	Verification set (sheet)	Image type
Tiny Person	[1400, 800]	3215	493	People: people
VOC2007 and VOC2012	[800, 600]	17416	1936	People: people; Animals: birds, cats, cows, dogs, horses, sheep; Transportation: plane, bicycle, boat, bus, car, motorcycle, train; Indoor objects: bottles, chairs, dining tables, potted plants, sofas, televisions/monitors .

FIGURE 9. VOC dataset.

2) CONFIGURATION

Our experiment is based on the framework of MMDetection3 [14] and trained on a single GPU of Intel Core i9 RTX 3090. Load the weight of deformable-detr-refine-twostage_r50_16xb2-50e_voc/best.pth and fine-tune it. SGD optimizer with weight decay rate of 0.0001 and momentum of 0.999 is adopted. To alleviate the instability and slow convergence of the model, which may be caused by large learning rate, the algorithm model in this paper adopts Cosine Annealing LR learning rate adjustment method, and the maximum learning rate and minimum learning rate are set to be the same during model training, and the value is 0.0001. The training generation of the algorithm model starts from 0, and a total of 200 generations are trained. In this paper, the freezing training mode is not adopted, and the latest SwiGLU activation function is used for output activation. In terms of input image shape, its size must be a multiple of 32. On VOC dataset, the input image size is [800, 600], while on Tiny Person dataset, large-resolution size input is adopted, and the parameter is set to [1400, 800].

The experiment is carried out on the hardware equipment as shown in the table: the operating system is Windows 10, and the computing hardware configuration includes 64GB DDR4 3733MHz memory, RTX 3090 24G discrete graphics card, Intel Core i9 processor, 2TB HDD hard disk and DELL keyboard and mouse. Python programming language, version 3.11, PyTorch deep learning framework. The operation of the program depends on GPU for calculation, without using CPU. On VOC and Tiny Person datasets, the running time is 48 hours and 75 hours respectively. GPU resource utilization is about 40%.

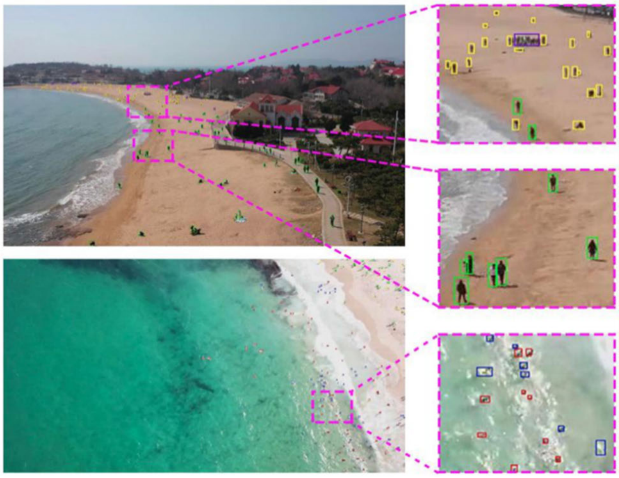


FIGURE 10. Tiny person dataset.

TABLE 2. Specific parameters of hardware equipment.

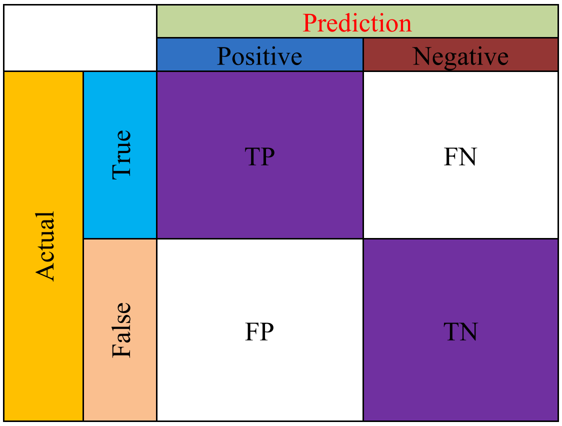
Equipment	Parameter
Processor	Intel Core i9
Internal storage	64GB DDR4 3733MHz
Hard disc	2TB HDD
Display card	RTX 3090 24G solo display
Keyboard	DELL
Operating system	Windows 10
Mouse	DELL

3) EVALUATION MATRIX

In the small target detection task, the evaluation indexes are mainly divided into two categories: detection accuracy and detection speed. The evaluation indexes of detection accuracy include IoU, accuracy, accuracy, recall, F1 value, AP, MAP, ROC curve and AUC value, and the common formulas are shown in (9) to (14). The evaluation indexes of detection speed include forward transmission time, FPS and FLOPS.

In this paper, the performance of the algorithm model is comprehensively evaluated by the evaluation index of detection accuracy, including accuracy, recall rate, mAP and F1 value. The above indicators can effectively evaluate the accuracy, comprehensiveness, stability generalization ability and robustness of the algorithm. Combining the types of model detection with the real situation, TP is the number of positive samples correctly identified, FP is the number of negative samples detected as positive samples, FN is the number of positive samples not detected as negative samples, and TN is the number of negative samples correctly classified, as shown in the following table. Accuracy represents the accuracy of the algorithm relative to the detection results, recall represents the ability of the algorithm to check the whole algorithm, mAP represents the comprehensive performance of target detection algorithms in multiple categories, and F1 value is the harmonic average of accuracy and recall. Swin Transformer's DETR algorithm aims to improve the detection performance of small target objects, so besides the detection accuracy index, it also considers the detection speed indexes such as FPS, APS, APM and APL to comprehensively evaluate the speed of the model and the detection accuracy of multi-scale objects. In addition, we will pay attention to the parameters of the network model to ensure that the actual deployment needs are met.

TABLE 3. Classification result matrix.



		Prediction	
		Positive	Negative
Actual	True	TP	FN
	False	FP	TN

Precision is defined as

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{TF}}. \tag{9}$$

Accuracy is defined as

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}. \tag{10}$$

Recall rate is defined as

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{11}$$

The F1 value is defined as

$$\text{F1} = 2\frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}}. \tag{12}$$

The average accuracy is defined as

$$\text{AP} = \frac{\sum \text{P}_{\text{ri}}}{\sum \text{r}}. \tag{13}$$

Mean precision mean is defined as

$$\text{mAP} = \frac{\sum \text{AP}_{\text{i}}}{\text{n}}. \tag{14}$$

MAP (average precision) standard is adopted in this experiment. When the IoU threshold is 0.5, the precision rate and recall rate under different confidence thresholds are calculated. By calculating the area under the PR curve, the AP value of each category is obtained, and the AP values of all categories are averaged to get the mAP value. The mAP, Recall, F1 and Precision values of Swin Transformer's DETR on VOC and Tiny Person datasets are shown in the following table.

TABLE 4. Experiment results on VOC and tiny person dataset.

Dataset	VOC2007+2012	Tiny Person
mAP	88.9%	48.36%
Recall	97.0%	70.7%
F1	79.0%	79.0%
Precision	95.0%	89.9%

B. EXPERIMENT ANALYSIS

1) RESULTS ON THE VOC DATASET

On the VOC data set, for each type, when IoU=0.5, the algorithm model Swin Transformer's DETR in this paper calculates the Precision, F1, Recall and AP values of each category in the data set under different confidence thresholds, as shown in Table 4.

According to the experimental results, Swin Transformer's DETR performs well on VOC data sets. The average accuracy (AP) of each category is obviously different. For example, the cat category is 99.1%, while the bottle category is only 64.6%, which shows that the models are quite different among different categories. The F1 value of most categories exceeds 0.7, which shows that the model has achieved a good balance between accuracy and recall. However, the F1 values of some categories such as person and diningtable are low, and the algorithm model needs to be further optimized. The recall rate of most categories is over 90%, which shows that the model is highly sensitive to the detection target, but the recall rate of some categories such as bottle and diningtable is low, so it is necessary to further optimize and improve the algorithm model. The accuracy of most categories is over 90%, which shows that the algorithm model has high accuracy in identifying positive samples, but there are still some cases where the accuracy of some categories is low, and there may be false detection, so the algorithm model needs to be further optimized and improved.

To sum up, Swin Transformer's DETR shows high detection accuracy on VOC data sets, but it needs to further optimize and improve the algorithm model according to the performance differences of different categories to improve the overall performance of the model. We draw the training loss, verification loss and mAP value with epoch number in the training process of Swin Transformer's DETR. 11 shows that from a macro point of view, the loss continues to decline during the training process, but there will be some fluctuations at the micro level. After about 5 epoch, the loss decreases slowly and tends to be stable, which shows that the model has strong convergence ability and stability. Fig. 12 shows that the mAP value increases with the increase of the number of training epochs, and basically reaches a stable level in about 5 epochs, which shows that the model has strong learning ability and the ability to identify target features and shows excellent detection accuracy quickly and accurately.

2) RESULTS ON THE TINY PERSON DATASET

On the tiny target object data set of Tiny Person, for each type, when IoU=0.5, the Precision, F1, Recall and AP values

TABLE 5. Prediction performance of the proposed model in terms of Precision, Recall, F1_score, AP on the VOC dataset.

Type	AP	F1	Recall	Precision
tvmonitor	90.90%	0.977	99.40%	99.50%
motorbike	90.90%	0.935	99.40%	99.50%
horse	92.90%	0.912	100%	99.50%
cat	99.10%	0.886	100%	99.50%
bird	90.70%	0.766	96.70%	99.50%
dog	93.20%	0.74	100%	99.50%
chair	88.90%	0.55	99.50%	99.50%
car	85.80%	0.389	96.20%	99.50%
person	82.60%	0.122	92.70%	99.50%
bicycle	90.60%	0.911	98.30%	99.30%
Train	94.10%	0.974	100%	98%
pottedplant	82.60%	0.707	95.10%	98%
aeroplane	90.90%	0.94	96.10%	97.20%
bottle	64.60%	0.72	77.40%	96.00%
sofa	99.20%	0.95	100%	91.10%
boat	87.20%	0.881	97.30%	89.70%
cow	90.90%	0.915	98.40%	87.40%
bus	90.80%	0.892	99.10%	81.80%
diningtable	89.10%	0.876	99.10%	79.60%
sheep	82.70%	0.789	95.30%	76.90%

of Swin Transformer's DETR are calculated under different confidence thresholds, as shown in the following table.

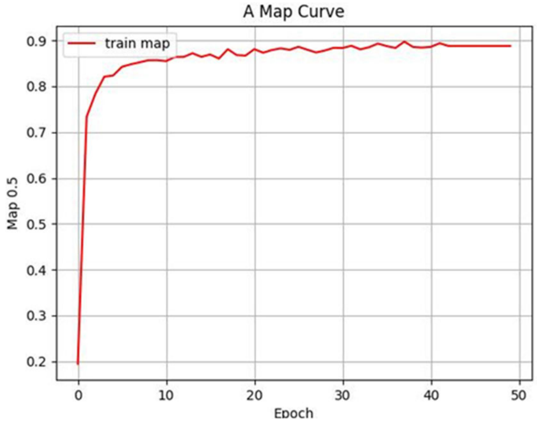


FIGURE 11. Map value curve of Swin Transformer's DETR models on VOC dataset.

According to the experimental results, Swin Transformer's DETR performs well on Tiny Person data set. The average accuracy (AP) of this model on this data set reaches 48.3, which shows that it has good detection ability for small target objects. The F1 value is 0.792, which shows that the model has achieved a good balance between accuracy and recall and has good comprehensive performance. The recall rate reaches 70.7%, which shows that the model has high sensitivity in detecting small target objects and can capture small target objects well. The accuracy rate is 89.9%, which shows that the model has high accuracy in identifying positive samples, reducing false detection, and improving the reliability of detection.

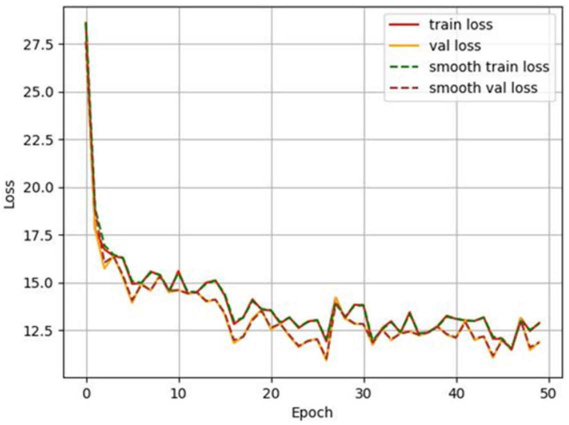


FIGURE 12. Loss curve of Swin Transformer's DETR models on VOC dataset.

TABLE 6. Prediction performance of the proposed model in terms of Precision, Recall, F1_score, AP on the Tiny Person dataset.

Type	AP	F1	Recall	Precision
tiny_person	48.36%	0.792	70.7%	89.9%

In a word, Swin Transformer's DETR shows excellent detection performance on Tiny Person data set, which provides strong experimental support for accurate detection of small target objects. We draw the training loss, verification loss and mAP value with epoch number in the training process of Swin Transformer's DETR model. Fig. 13 shows that, overall, the loss curve fluctuates in the process, indicating that the model has some challenges to the stability and convergence of this data set. Fig. 14 shows that the mAP value gradually increases with the increase of the number of epochs, and almost reaches a high point at about 2 epochs, indicating that the algorithm model has strong learning ability and the ability to accurately identify the target features, and shows excellent detection accuracy.

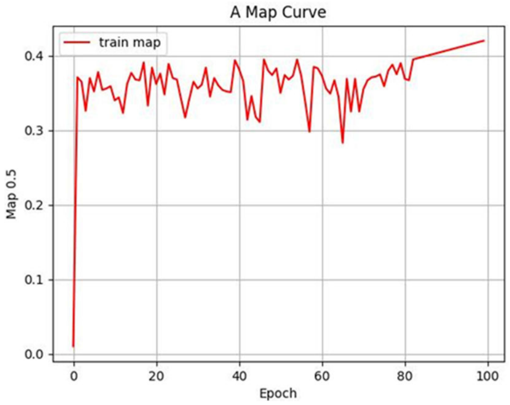


FIGURE 13. Map value curve of Swin Transformer's DETR model on tiny person dataset.

3) ABLATION EXPERIMENT

The experimental results of four algorithms on VOC and Tiny Person data sets. Through the ablation experiment,

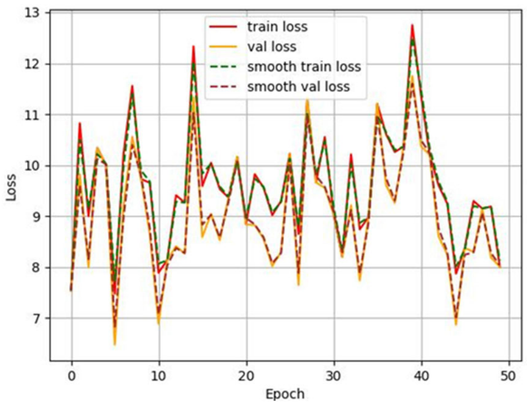


FIGURE 14. Loss curve of Swin Transformer's DETR model on tiny person dataset.

the necessity and effectiveness of Swin Transformer's DETR in the target detection task are verified. On the data sets of VOC and Tiny Person, YOLOv5, Faster R-CNN and DETR algorithm models are selected for experiments. The experimental results are listed in Tables 7 and 8 respectively.

Table 7 compares the experimental results of four algorithm models on VOC data sets. Swin Transformer's DETR achieved the highest mAP value of 88.9% when the input image size was 800 × 600, which was 11.6% higher than that of Faster R-CNN. Its Recall value is as high as 97%, which is 15.4% higher than that of Faster R-CNN, while its F1 value is 0.79, which is 0.06 higher than that of YOLOv5. In addition, Precision is as high as 95%, which is 10.8% higher than YOLOv5. Overall, the accuracy of our proposed algorithm model is improved by 11.6%.

Table 8 compares the experimental results of four algorithm models on Tiny Person data set. Swin Transformer's DETR achieved remarkable results when the input image size was 1400 × 800, and the mAP reached 48.36%, which was 8.36% higher than YOLOv5. The other three detection accuracy indexes are obviously better than other algorithm models, which are improved by 39.5%, 0.33 and 2.1% respectively. Overall, the accuracy of our proposed algorithm model is improved by 8.36%. The experimental results show that the Swin Transformer's DETR algorithm model can improve the accuracy of target detection.

TABLE 7. Comparison of Prediction performance of YOLOv5, Faster R-CNN, DETR, Swin Transformer's DETR in terms of Precision, Recall, F1_score, AP on the VOC dataset.

Method	Input image	mAP	Recall	F1	Precision
YOLOv5	640x640	74.3%	66%	0.73	84.2%
Faster R-CNN	640x640	77.3%	83.6%	0.61	47.7%
DETR	800x800	76.1%	80%	0.70	62.2%
Swin Transformer's DETR	**800x600**	**88.9%**	**97.0%**	**0.79**	**95%**

4) STABILITY AND ROBUSTNESS

We comprehensively analyze the stability and robustness of Swin Transformer's DETR algorithm model. On the data sets of VOC and Tiny Person, the curves of the mAP value and Loss function of YOLOv5, Faster R-CNN, DETR and Swin Transformer's DETR varying with the increase of epoch are shown in Figs. 15, 16, 17 and 18 respectively. The above graph can clearly present the training process and performance of each algorithm model.

TABLE 8. Prediction performance of the proposed model in terms of Precision, Recall, F1_score, AP on the Tiny Person dataset.

Method	Input image	mAP	Recall	F1	Precision
YOLOv5	640x640	40.1%	31.29%	0.46	87.8%
Faster R-CNN	600x600	2.1%	7.56%	0.11	22.03%
DETR	800x800	0.1%	2.35%	0.02	2.40%
Swin Transformer's DETR	**1400x800**	**48.36%**	**70.7%**	**0.792**	**89.9%**

Fig. 15 shows the curves of mAP of four algorithm models varying with epoch number on VOC data set. With the increase of epoch number, the mAP of the four algorithm models all showed an upward trend, and the YOLOv5 curve was quite stable. Compared with the other three algorithm models, the mAP value of Swin Transformer's DETR increases the fastest and the mAP value is the largest, and the curve tends to be stable, which shows its strong learning ability, accurate identification of target features and excellent detection accuracy.

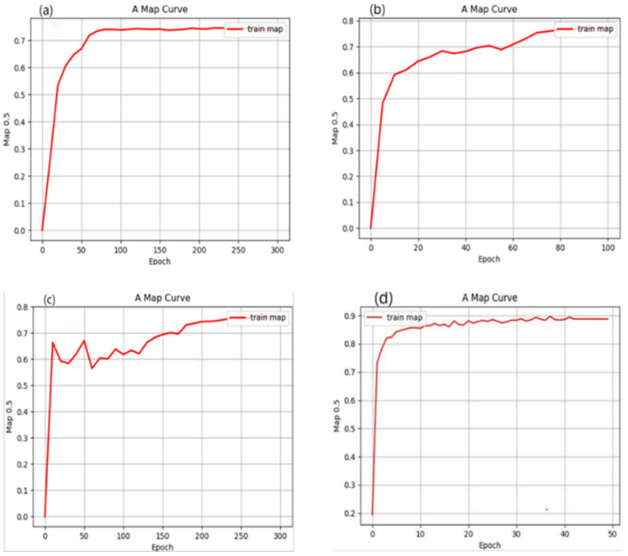


FIGURE 15. Map value curves of four algorithm models on VOC data set: (a) YOLOv5 (b) Faster R-CNN (c) DETR (d) Swin Transformer's DETR.

Fig. 16 shows that the curves of the mAP of the four algorithm models with the number of epoch on the Tiny Person data set all show an upward trend. The model of DETR algorithm has a serious shock, and the curve of YOLOv5 is the most smooth and stable. Compared with the other three algorithm models, the mAP value of Swin Transformer's DETR grows fastest and has the largest value, but it fluctuates slightly during the training process. It shows that Swin

Transformer's DETR has strong learning ability and the ability to accurately identify target features, showing excellent detection accuracy.

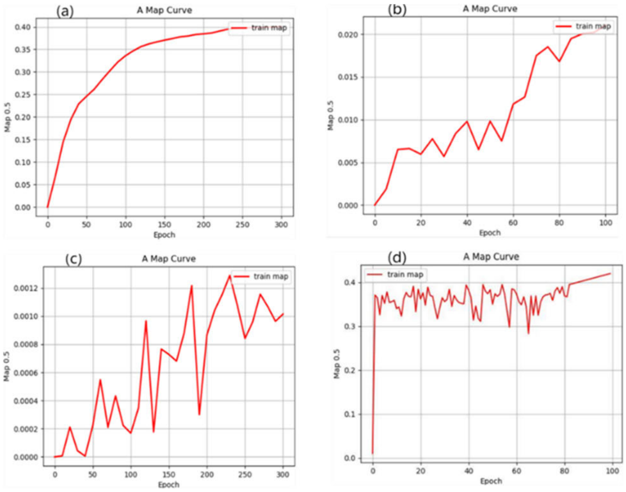


FIGURE 16. Map value curves of four algorithm models on tiny person data set: (a) YOLOv5 (b) Faster R-CNN (c) DETR (d) Swin Transformer's DETR.

Fig. 17 shows the curves of the Loss loss function of the four algorithm models with the epoch number on the VOC data set, all of which show a downward trend. The Loss curve of YOLOv5 has the fastest decline and the lowest Loss value, and its value is close to 0, and the curve is the smoothest and most stable. The Loss of Swin Transformer's DETR gradually decreases with the increase of epoch number, which shows the effectiveness of the model. However, compared with YOLOv5, the stability of the algorithm model in this paper has not reached the ideal level, and further optimization and improvement are needed to improve the stability and generalization ability of the model.

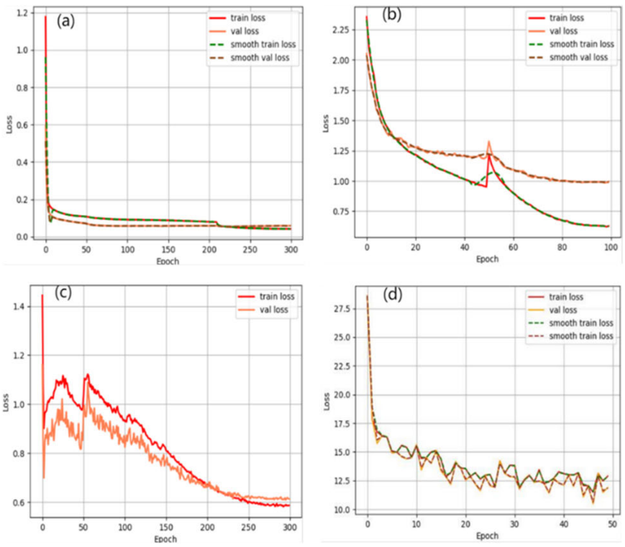


FIGURE 17. Loss curves of four algorithm models on VOC data set: (a) YOLOv5 (b) Faster R-CNN (c) DETR (d) Swin Transformer's DETR.

Fig. 18 shows the curves of the Loss of four algorithm models on Tiny Person data set with the increase of epoch number. The Loss of the other three algorithm models all showed a downward trend, among which the Loss of YOLOv5 decreased the fastest, and finally approached zero, and the curve was the most smooth and stable. However, the Loss curve of Swin Transformer's DETR algorithm model shows obvious oscillation trend, and there are different degrees of over-fitting. It shows that the algorithm model has shortcomings in stability, generalization ability and robustness, and it needs to be further optimized and improved.

In a word, Swin Transformer's DETR algorithm model is excellent in learning ability and target feature recognition. However, there are some shortcomings in stability, generalization ability and robustness, and the algorithm model needs to be further optimized and improved.

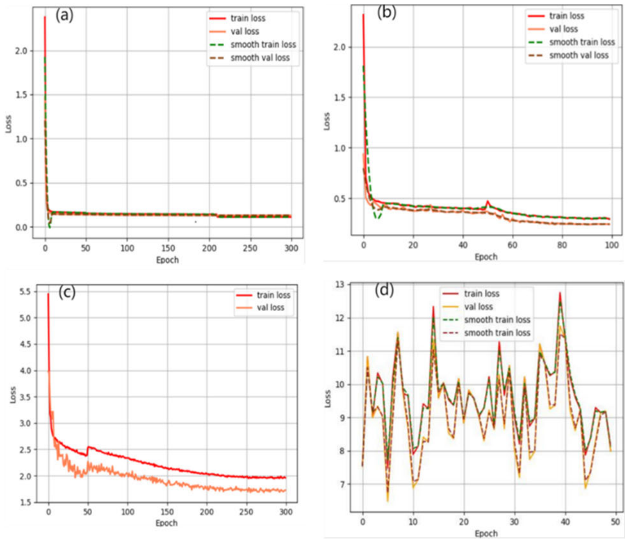


FIGURE 18. Map value curves of four algorithm models on Tiny Person data set: (a) YOLOv5 (b) Faster R-CNN (c) DETR (d) Swin Transformer's DETR.

IV. DISCUSSION

In this study, a novel method is proposed to detect small target objects. We use Swin Transformer as the backbone network to extract the global features and local information of small target objects and use a three-layer pyramid structure for feature fusion. Then, we use a two-stage detection head to generate a pre-selection box for small target objects, and then carry out fine-grained object detection. Finally, we use the most advanced SwiGLU activation function in the field of Transformer to activate the output, and optimize the algorithm model structure, data enhancement, training methods and testing methods. The evaluation results on PASCAL VOC and Tiny Person data sets show that our method outperforms previous studies in detection accuracy.

Table 9 compares the model detection results of our research and previous research on PASCAL VOC data sets. Tian et al. [16] proposed FCOS, which is a single-stage

TABLE 9. Comparison of detection results in previous work and in the proposed method on the VOC dataset.

Method	Backbone	Dataset	AP50
Faster R-CNN [15]	ResNet-50	VOC2007+2012	79.6
FCOS [16]	ResNet-50	VOC2007+2012	68.9
FSAF [17]	ResNet-50	VOC2007+2012	77
RetinaNet [18]	ResNet-50	VOC2007+2012	80
YOLOv5 [4]	CSPDarknet53	VOC2007+2012	77.9
YOLOF [19]	ResNet-50	VOC2007+2012	79.8
YOLOFs [20]	ResNet-50	VOC2007+2012	81
Swin Transformer's DETR	**Swin Transformer**	**VOC2007+2012**	**88.9**

target detector with full convolution. The target detection problem is solved by pixel-by-pixel prediction like semantic segmentation, and the accuracy rate reaches 68.9%. Zhu et al. [17] put forward the feature selection anchor-free module FSAF, which is a simple and effective basic building module of single-stage target detector, with an accuracy rate of 77%. Jocher et al. [5] proposed YOLOv5, a real-time object detection algorithm based on Deep Convolutional Neural Network (CNN), with an accuracy rate of 77.9%. Ren et al. [15] proposed Faster R-CNN, a target proposal network for real-time object detection, with an accuracy rate of 79.6%. Chen et al. [19] proposed a You Only Look One-level Feature (YOLOF) algorithm model for target detection, and its accuracy rate reached 79.8%. Lin et al. [18] designed and trained a simple dense detector named RetinaNet to evaluate the effectiveness of the algorithm model loss, and its detection accuracy reached 80%. In contrast, our proposed method achieves a detection accuracy of 88.9% on PASCAL VOC data sets, which exceeds Rudong's [20] best performance of 7.9%.

TABLE 10. Comparison of detection results in previous work and in the proposed method on the Tiny Person dataset.

Method	Backbone	Dataset	AP50
Faster R-CNN [15]	ResNet-50	Tiny Person	43.55
FPN [21]	VGG16	Tiny Person	47.35
FCOS [16]	ResNet-50	Tiny Person	17.9
RetinaNet [18]	ResNet-50	Tiny Person	33.53
Grid R-CNN [22]	ResNet-50	Tiny Person	47.14
DSFD [23]	VGG/ResNet	Tiny Person	31.15
FreeAnchor [24]	ResNet-50	Tiny Person	41.36
Li-RCNN [25]	ResNet-50-FPN	Tiny Person	44.68
Swin Transformer'DETR	Swin Transformer	Tiny Person	48.36

Table 10 shows our detection results on Tiny Person data set using Swin Transformer's DETR and lists the previous research results. Tian et al. [16] proposed FCOS, a single-stage object detector, which is like semantic segmentation and solves the problem of object detection by pixel-by-pixel prediction and achieves 17.9% detection accuracy on Tiny Person data set. Li et al. [23] proposed a network for face detection called Double Detection Face Detector (DSFD), which achieved a detection accuracy of 31.15% on Tiny Person data set. Lin et al. [18] designed a simple dense detector RetinaNet and trained it to evaluate the effectiveness of the algorithm model loss and achieved a detection accuracy of 33.53% on Tiny Person data set. Zhang et al. [24] proposed a learning matching method to match anchor points in a flexible way, which is called FreeAnchor. On Tiny Person data set, this method achieved an accuracy of 41.36%. Pang et a. [25] proposed a simple but effective balanced learning framework, called Libra R-CNN, and their method achieved a detection accuracy of 44.68% on Tiny Person data sets. Lu et al. [22] proposed Grid R-CNN, which is a new target detection framework with grid-guided positioning mechanism. The detection accuracy is 47.14% on Tiny Person data set. Our method achieves the highest detection accuracy of small target objects on Tiny Person data set, reaching 48.36%. It is 1% higher than the best detection algorithm FPN proposed by Liu et al. [21].

We classified and summarized the experimental data of YOLOv5, Faster R-CNN, DETR, and Swin Transformer's DETR on the VOC, and Tiny Person datasets, including F1 value, AP0.5 value, Recall value, and Precision value. Then, T-test analysis was conducted on them separately.

Table 11 shows that the F1 value of Swin Transformer's DETR test results group is higher than that of Faster R-CNN test results group (the average difference is 0.23), and the analysis of variance shows that both sides have P < 0.05. The F1 values of the two algorithms are significantly different, so to some extent, Swin Transformer's DETR algorithm is more effective.

TABLE 11. Comparison results of F1 values of different algorithms.

(I) type	(J) type	Mean diff (I-J)	Std error	P value	95% conf interval	
					lower limit	upper limit
YOLOv5 test results	Faster R-CNN test results	0.15	0.05	0.036	0.006	0.292
	DETR test results	0.09	0.05	0.523	-0.051	0.235
	Swin Transformer's	-0.08	0.05	0.871	-0.221	0.065
Faster R-CNN test results	YOLOv5 test results	-0.15	0.05	0.036	-0.292	-0.006
	DETR test results	-0.06	0.05	1.000	-0.201	0.086
	Swin Transformer's	-0.23	0.05	0.000	-0.370	-0.084
DETR test results	YOLOv5 test results	-0.09	0.05	0.523	-0.235	0.051
	Faster R-CNN test results	0.06	0.05	1.000	-0.086	0.201
	Swin Transformer's	-0.17	0.05	0.011	-0.313	-0.027
Swin Transformer's DETR test results	YOLOv5 test results	0.08	0.05	0.871	-0.065	0.221
	Faster R-CNN test results	0.23	0.05	0.000	0.084	0.370
	DETR test results	0.17	0.05	0.011	0.027	0.313

*P＜0.05

The F1 value of YOLOv5 test results group was higher than that of Faster R-CNN test results group (the average difference was 0.15), and the analysis of variance showed that both sides had p < 0.05. The F1 values of the two algorithms were significantly different, so to some extent, the YOLOv5 algorithm was more effective.

Table 12 shows that the AP0.5 value of Swin Transformer's DETR test results group is higher than that of DETR test results group (the average difference is 0.18), and the analysis of variance shows that both sides have $P < 0.05$, and the AP0.5 values of the two algorithms are significantly different, which shows that Swin Transformer's DETR algorithm is more effective, to some extent.

TABLE 12. Comparison results of AP0.5 values of different algorithms.

(I) type	(J) type	Mean diff (I-J)	Std error	P value	95% conf interval	
					lower limit	upper limit
YOLOv5 test results	Faster R-CNN test results	0.01	0.06	1.000	-0.148	0.162
	DETR test results	0.05	0.06	1.000	-0.109	0.201
	Swin Transformer's	-0.14	0.06	0.113	-0.293	0.018
Faster R-CNN test results	YOLOv5 test results	-0.01	0.06	1.000	-0.162	0.148
	DETR test results	0.04	0.06	1.000	-0.116	0.194
	Swin Transformer's DETR test results	-0.14	0.06	0.083	-0.300	0.011
DETR test results	YOLOv5 test results	-0.05	0.06	1.000	-0.201	0.109
	Faster R-CNN test results	-0.04	0.06	1.000	-0.194	0.116
	Swin Transformer's DETR test results	-0.18	0.06	0.012	-0.338	-0.028
Swin Transformer's DETR test results	YOLOv5 test results	0.14	0.06	0.113	-0.018	0.293
	Faster R-CNN test results	0.14	0.06	0.083	-0.011	0.300
	DETR test results	0.18	0.06	0.012	0.028	0.338

*P <0.05

Table 13 shows that the Recall value of Swin Transformer's DETR test results group is higher than that of YOLOv5 test results group (the average difference is 0.31), and the analysis of variance shows that both sides have $P < 0.05$. The Recall values of the two algorithms are significantly different, which shows that Swin Transformer's DETR algorithm is more effective to some extent. The Recall value of Swin Transformer's DETR test results group is higher than that of Faster r-CNN test results group (the average difference is 0.17), and the analysis of variance shows that both sides have $P < 0.05$. The Recall values of the two algorithms are significantly different, which shows that Swin Transformer's DETR algorithm is more effective to some extent.

The Recall value of Swin Transformer's DETR test results group is higher than that of DETR test results group (the average difference is 0.23). The analysis of variance shows that both sides have $P < 0.05$, and the Recall values of the two algorithms are significantly different, which shows that Swin Transformer's DETR algorithm is more effective, to some extent.

TABLE 13. Comparison results of Recall values of different algorithms.

(I) type	(J) type	Mean diff (I-J)	Std error	P value	95% conf interval	
					lower limit	upper limit
YOLOv5 test results	Faster R-CNN test results	-0.14	0.05	0.060	-0.285	0.003
	DETR test results	-0.08	0.05	0.785	-0.226	0.063
	Swin Transformer's DETR test results	-0.31	0.05	0.000	-0.455	-0.166
Faster R-CNN test results	YOLOv5 test results	0.14	0.05	0.060	-0.003	0.285
	DETR test results	0.06	0.05	1.000	-0.085	0.204
	Swin Transformer's DETR test results	-0.17	0.05	0.012	-0.314	-0.025
DETR test results	YOLOv5 test results	0.08	0.05	0.785	-0.063	0.226
	Faster R-CNN test results	-0.06	0.05	1.000	-0.204	0.085
	Swin Transformer's DETR test results	-0.23	0.05	0.000	-0.373	-0.085
Swin Transformer's DETR test results	YOLOv5 test results	0.31	0.05	0.000	0.166	0.455
	Faster R-CNN test results	0.17	0.05	0.012	0.025	0.314
	DETR test results	0.23	0.05	0.000	0.085	0.373

*P <0.05

Table 14 shows that the Precision value of Swin Transformer's DETR test results group is higher than that of Faster R-CNN test results group (the average difference is 0.49), and the analysis of variance shows that both sides have $P < 0.05$, and the Precision values of the two algorithms are significantly different, so to some extent, Swin Transformer's DETR algorithm is more effective.

TABLE 14. Comparison results of Precision values of different algorithms.

(I) type	(J) type	Mean diff (I-J)	Std error	P value	95% conf interval	
					lower limit	upper limit
YOLOv5 test results	Faster R-CNN test results	0.39	0.04	0.000	0.288	0.498
	DETR test results	0.29	0.04	0.000	0.187	0.396
	Swin Transformer's DETR test results	-0.10	0.04	0.087	-0.202	0.008
Faster R-CNN test results	YOLOv5 test results	-0.39	0.04	0.000	-0.498	-0.288
	DETR test results	-0.10	0.04	0.065	-0.206	0.004
	Swin Transformer's DETR test results	-0.49	0.04	0.000	-0.594	-0.385
DETR test results	YOLOv5 test results	-0.29	0.04	0.000	-0.396	-0.187
	Faster R-CNN test results	0.10	0.04	0.065	-0.004	0.206
	Swin Transformer's DETR test results	-0.39	0.04	0.000	-0.493	-0.284
Swin Transformer's DETR test results	YOLOv5 test results	0.10	0.04	0.087	-0.008	0.202
	Faster R-CNN test results	0.49	0.04	0.000	0.385	0.594
	DETR test results	0.39	0.04	0.000	0.284	0.493

*P <0.05

The Precision value of Swin Transformer's DETR test results group is higher than that of DETR test results group (the average difference is 0.39), and the analysis of variance shows that both sides have $P < 0.05$, and the Precision values of the two algorithms are significantly different, which shows that Swin Transformer's DETR algorithm is more effective to some extent.

The Precision values of YOLOv5 test results group were all higher than those of Faster R-CNN test results group (the average difference was 0.39), and the analysis of variance showed that both sides had P < 0.05. The difference of Precision values between the two algorithms was significant, so the YOLOv5 algorithm was more effective, to some extent.

The Precision values of YOLOv5 test results group were all higher than those of DETR test results group (the average difference was 0.29). The analysis of variance showed that both sides had P < 0.05, and the Precision values of the two algorithms were significantly different, which showed that YOLOv5 algorithm was more effective, to some extent.

V. CONCLUSION

This paper proposed a novel small target object detection algorithm model, Swin Transformer's DETR, which uses Swin Transformer as the backbone to extract the global features and local information of small target objects and uses a three-layer feature fusion gold structure to carry out subsequent fusion processing on its features. Secondly, the output of FFN is activated by the latest SwiGLU activation function. Finally, we comprehensively evaluate the proposed algorithm model on two data sets, VOC, and Tiny Person, and achieve the accuracy of 88.9% and 48.36% respectively. The experimental results show that the Swin Transformer's DETR small target object detection algorithm model can improve the accuracy of target detection, reduce the overall calculation amount, and thus improve the calculation efficiency. Its flexible network structure design makes it outstanding in task detection in different application scenarios. The stability, generalization ability and robustness of the algorithm model are comprehensively evaluated by drawing the curves of mAP value and Loss loss function with the increase of epoch. In a word, Swin Transformer's DETR algorithm model performs well on small target data sets, which improves the accuracy and speed of detection.

The impacts of this work can be summarized based on each contribution as follows:

i. Applying Swin Transformer as the backbone network: This significantly improves the discrimination ability of the algorithm model to the target object. This improvement not only reduces the dimension and computational complexity, but also retains the key information, making the algorithm model lighter and more efficient. Finally, the above optimization measures improve the accuracy, robustness, and stability of the small target object detection task.

ii. Applying three-layer pyramid structure: This design makes the network more flexible and can better adapt to targets with different scales and complexity, thus significantly improving the accuracy and robustness of detection.

iii. Applying two-stage detection head: The two-stage detection method can deal with different scales and shapes of targets more flexibly, thus improving the adaptability of the algorithm model to complex scenes.

iv. Loss Function: The loss calculation and matching mechanism of the Hungarian algorithm in object detection have a significant impact on the detection of small objects. By reasonably designing and adjusting the loss function, the accuracy and robustness of small object detection can be significantly improved.

Future research direction: focus on improving the performance of the algorithm model on Tiny Person data set to solve the problems of low mAP value, shock, and over-fitting. We will continue to optimize the algorithm model, improve the accuracy and speed of detection, and solve the challenges of stability, generalization ability and robustness. It involves parameter adjustment, training strategy improvement and data enhancement technology optimization to ensure that the algorithm model can perform stably and well in all situations.

ACKNOWLEDGMENT

The authors would like to thank Universiti Malaysia Sabah for their support to publish this article.

REFERENCES

[1] S. Agarwal, J. O. Du Terrail, and F. Jurie, "Recent advances in object detection in the age of deep convolutional neural networks," 2018, *arXiv:1809.03193*.

[2] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," 2013, *arXiv:1311.2524*.

[3] R. Girshick, "Fast R-CNN," *Comput. Sci.*, vol. 18, no. 3, pp. 5–7, 2015.

[4] G. Jocher, K. Nishimura, and T. Mineeva. *YOLOv5*. [Online]. Available: https://github.com/ultralytics/yolov5

[5] Ultralytic. (2023). *YOLOv8*. [Online]. Available: https://github.com/ultralytics/yolov8

[6] C. Y. Fu, W. Liu, A. Ranga, A. Tyagi, and A. C. Berg, "DSSD: Deconvolutional single shot detector," in *Proc. Comput. Vis. Pattern Recognit.*, 2017, pp. 6659–6669.

[7] R. Liu, Z. Yuan, T. Liu, and Z. Xiong, "End-to-end lane shape prediction with transformers," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 3693–3701.

[8] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Comput. Vis. Pattern Recognit.*, 2020, pp. 12872–12898.

[9] H. Zhao, J. Gao, T. Lan, C. Sun, B. Sapp, B. Varadarajan, Y. Shen, Y. Shen, Y. Chai, C. Schmid, C. Li, and D. Anguelov, "TNT: Target-driveN trajectory prediction," 2020, *arXiv:2008.08294*.

[10] T. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. CVPR*, 2015, pp. 1405–1420.

[11] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002.

[12] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, Jun. 2010.

[13] X. Yu, Y. Gong, N. Jiang, Q. Ye, and Z. Han, "Scale match for tiny person detection," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 1246–1254.

[14] K. Chen, J. Wang, J. Pang, Y. Cao, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Xu, and Z. Zhang, "MMDetection: Open MMLab detection toolbox and benchmark," 2019, *arXiv:1906.07155*.

[15] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[16] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: Fully convolutional one-stage object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9626–9635, doi: 10.1109/ICCV.2019.00972.

[17] C. Zhu, Y. He, and M. Savvides, "Feature selective anchor-free module for single-shot object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 840–849, doi: 10.1109/CVPR.2019.00093.

[18] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2999–3007.

[19] Q. Chen, Y. Wang, T. Yang, X. Zhang, J. Cheng, and J. Sun, "You only look one-level feature," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 13034–13043, doi: 10.1109/CVPR46437.2021.01284.

[20] R. Jing, W. Zhang, Y. Liu, W. Li, Y. Li, and C. Liu, "An effective method for small object detection in low-resolution images," *Eng. Appl. Artif. Intell.*, vol. 127, Jan. 2024, Art. no. 107206.

[21] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, and A. C. Berg, "SSD: Single shot MultiBox detector," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 21–37, doi: 10.1007/978-3-319-46448-0_2.

[22] X. Lu, B. Li, Y. Yue, Q. Li, and J. Yan, "Grid R-CNN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7355–7364, doi: 10.1109/CVPR.2019.00754.

[23] J. Li, Y. Wang, C. Wang, Y. Tai, J. Qian, J. Yang, C. Wang, J. Li, and F. Huang, "DSFD: Dual shot face detector," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5055–5064, doi: 10.1109/CVPR.2019.00520.

[24] X. Zhang, F. Wan, C. Liu, X. Ji, and Q. Ye, "Learning to match anchors for visual object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 6, pp. 3096–3109, Jun. 2022.

[25] J. Pang, K. Chen, J. Shi, H. Feng, W. Ouyang, and D. Lin, "Libra R-CNN: Towards balanced learning for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 821–830, doi: 10.1109/CVPR.2019.00091.

XU FENGCHANG received the M.Sc. degree from the School of Information Engineering, Nanchang University, in 2013. He is currently pursuing the Ph.D. degree with the Faculty of Computing and Informatics, University of Malaysia.

He is currently a Senior Lecturer with the Department of Information Engineering, Shandong Light Industry Vocational College. His research interests include machine learning, deep learning, natural language processing, and big data analytics.

RAYNER ALFRED (Member, IEEE) received the master's degree in computer science from Western Michigan University, Kalamazoo, USA, the degree in computer science from the Polytechnic University of Brooklyn, New York, USA, and the Ph.D. degree in computer science from York University, U.K., in 2008. He is currently a Technologist and a Professor in computer science with Universiti Malaysia Sabah, Malaysia. He is also looking at intelligent techniques using machine learning to model and optimize the dynamic and distributed processes of knowledge discovery for structured and unstructured data. He has also been a Visiting Scholar to various universities around the world. He has been a Visiting Scholar with Korea Institute of Advanced Study, Seoul, South Korea, in 2020, and Japan Advanced Institute of Science and Technology, Japan, in 2018. He was recently appointed as a Visiting Professor with the Faculty of Information Engineering, Hainan Vocational University of Science and Technology, China, from 2020 to 2023. Besides that, he has also been serving as a member of board of directors for Malaysian Institute of Microelectronic Systems (MIMOS), since 2019. He is also appointed as the Subject Expert Matter in the field of Evaluation of Generators-KPT Career Advancement Program for Private Higher Education Institutions at the level of the Department of Higher Education (JPT), Ministry of Higher Education Malaysia, from March 2021 to February 2023. He has authored or co-authored more than 150 journals/book chapters and conference papers, editorials, and served on the program and organizing committees of numerous national and international conferences and workshops. He has been the Chair for computational science and machine learning conferences, such as Computational Science and Technology (ICCST), Computational Science and Engineering (ICCSE), and Advanced Information System and Knowledge Management (AISKM). He is a Certified Software Tester (CTFL) from the International Software Testing Qualifications Board (ISTQB) and a certified IBM DB2 Academic Associate (IBM DB2 AA). He leads the Creative Advanced Machine Intelligence (CAMI) Research Center, UMS. He was also a recipient of multiple GOLD and SPECIAL awards at national and international research exhibitions in Data Mining and Machine Learning-Based Solutions (Face Recognition and Knowledge Discovery), that include International Trade Fair Ideas in Nuremberg, Germany (iNEA2018 and iENA2019) International Invention Innovation Competition in Toronto, Canada, in 2022, 2021, 2020, and 2018 (iCAN2022, iCAN2021, iCAN2020, and iCAN2018), Seoul International Invention Exhibition in Seoul, South Korea (SIIF 2010), International Conference and Exposition on Inventions by Institutions of Higher Learning (PECIPTA2010 and 2019), and International Invention, Innovation & Technology Exhibition, Malaysia (ITEX2019, ITEX2018, and ITEX2010). He was a recipient of the Myron M. Rosenthal Academic Achievement Award for the Outstanding Academic Achievement in Computer Science from the Polytechnic University of Brooklyn, in 1994.

RAYNER HENRY PAILUS is currently a Data Scientist with over ten years of experience in architecting technology solutions to increase the efficiency, accuracy, and utility of internal data processing. Highly experienced in biometric security data science and machine learning. Highly skilled in creating data regression models, using predictive data modeling, and analyzing data mining algorithms to deliver insights and implement action-oriented solutions to complex business problems. He consistently provides high-quality technical and knowledge in biometric security using facial recognition, robust face detection, and developing software, making a significant impact in the biometric patient authentication system, including telecommunication industry. He is currently serving as the BIMP EAGA Business Council, Malaysia (Sabah Chapter) as a Technical Advisory Panel Communication and Engineering Technology for Business Communication & Information Exchange.

LYU GE received the B.S. degree in English language and literature and the M.S. degree in law from East China Normal University, Shanghai, China. She is currently pursuing the Ph.D. degree in management with Semyung University.

She is good at English Chinese translation in the field of computer science, computer information management, and interdisciplinary research; and has published many academic articles in China.

DU SHIFENG was born in Yantai, Shandong, China, in 1970. She received the M.S. degree in engineering from the Ocean University of China, in 2007.

Since 1991, she has been teaching with Shandong Light Industry Vocational College, holding the title of a Professor. She has published over 20 articles. Her research interests include artificial intelligence and new generation information technology and its applications.

JACKEL VUI LUNG CHEW was born in Kota Belud, Sabah, Malaysia, in 1993. He received the B.S. degree in mathematics with economics and the Ph.D. degree in mathematics from Universiti Malaysia Sabah, Sabah, Malaysia, in 2015 and 2019, respectively.

Since 2019, he has been a Senior Lecturer with the Faculty of Computing and Informatics, Universiti Malaysia Sabah Labuan International Campus, Labuan, Malaysia. He is the author of more than 40 articles. His research interests include numerical analysis particularly for solving PDE and FDE, statistical modelling and analysis, data visualization, and machine learning.

LI GUOZHANG was born in Anhui, China, in 1986. He received the M.S. degree from Jiangxi Normal University, China, in 2014.

He has been the Vice President of Hainan Vocational University of Science and Technology, since 2020. He is the author of more than 20 articles. His research interests include big data and e-commerce.

WANG XINLIANG was born in Shandong, China, in 1987. He received the B.S. degree from Yantai Nanshan University, China, in 2010.

Since 2013, he has been engaged in work related to civil air defense emergency command communication and network security. He obtained the qualification of a Senior Engineer, in 2017, and holds two national software copyright certificates, one national invention patent, and one local research project.