Proceedings of the 6th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/95/2024BJ0058

# CN-ViT- Visual Object Detection with VisionTransformer

**Chuan Wang**

School of Software, Taiyuan University of Technology, Shanxi, China

wangchuan5846@163.com

**Abstract.** Object detection has always been an important and challenging task in the field of computer vision. In recent years, Vision Transformers (ViT) have achieved remarkable results on image classification tasks, demonstrating its potential in vision tasks. In this paper, we propose CN-ViT, a novel Vision Transformer based visual object detection model.CN-ViT effectively improves the accuracy and robustness of object detection by combining the advantages of self-attention mechanism and convolutional neural network, and introducing the GCCA (Global Context Block and Coordinate Attention) module. In this paper, CN-ViT is evaluated on the Mini COCO standard dataset. The experimental results suggest that CN-ViT may outperform current mainstream object detection methods in terms of detection accuracy and speed. This study sheds light on the potential of Transformer architectures for complex visual tasks and offers valuable insights for future research in this area.

**Keywords:** CN-ViT, ViT, GCCA, Object detection.

## 1. Introduction

Object detection is one of the core problems in computer vision, whose task is to identify and locate objects in images. Object detection techniques have broad application prospects in many fields such as autonomous driving, surveillance monitoring, and medical image analysis, and have always been a research hotspot [1]. Traditional object detection methods mainly rely on Convolutional Neural Networks (CNNs), such as Faster R-CNN, YOLO, and SSD models, which have achieved significant success in many application scenarios [2-8]. However, as the task complexity increases, traditional methods still face many challenges in dealing with multi-scale targets, complex backgrounds, and occlusion problems.

In recent years, the Transformer architecture has achieved great success in natural language processing (NLP) and has gradually been introduced into computer vision tasks. Transformer uses the self-attention mechanism to capture global information and overcomes the limitations of traditional convolutional neural networks in handling long-distance dependencies and global feature extraction. Vision Transformer (ViT) is an image classification model based on Transformer, which overcomes the problem of the input image size being too large by dividing the image into fixed-size blocks and inputting the features of these blocks into the Transformer for processing [11]. ViT has demonstrated its strong performance in image classification tasks. Despite this, directly applying ViT to the object detection task still faces challenges, such as how to effectively handle high-resolution images and complex target scenes.

DOI: 10.54254/2755-2721/95/2024BJ0058

Traditional object detection methods, such as Faster R-CNN, generate candidate regions through a Region Proposal Network (RPN) [8], and then classify and regress these candidate regions using a convolutional neural network. This method performs well on single-scale targets, but often requires the introduction of a Feature Pyramid Network (FPN) to enhance detection ability when dealing with multi-scale targets. YOLO and SSD achieved real-time detection through a single-stage detection framework, but there is still room for improvement in detection accuracy. As the application of Transformer in visual tasks deepens, researchers have begun exploring the introduction of Transformer architecture into object detection to improve detection performance through the self-attention mechanism.

In this paper, we propose a novel vision-based object detection model called CN-ViT based on Vision Transformer. CN-ViT combines the advantages of self-attention mechanism and convolutional neural networks to design an efficient feature extraction and object detection framework. First, the image is fed into a convolutional neural network for feature extraction. Then, the extracted features are mapped to a high-dimensional space through the ViT (Vision Transformer) structure, where further global feature capture is performed to enhance the detection ability for objects of different scales.

The design of CN-ViT takes advantage of the global feature extraction ability of the Transformer and the local feature extraction ability of the convolutional neural network to form complementary feature representations. First, the input image is divided into fixed-size blocks, and initial features are extracted through convolutional layers. Then, these initial features are fused with global context information through the Global Context Block to further enhance the comprehensiveness of the feature expression. In addition, we apply the Coordinate Attention mechanism to capture channel information sensitive to location, thereby enhancing the feature localization ability. Next, these features are input into the multi-layer Transformer encoder, where global features are extracted through the self-attention mechanism. After being processed by the Transformer encoder, the features are input into the decoder, where target features are decoded through a specific number of learnable queries. Finally, the decoded features are classified into object categories and the object boundaries are refined through regression.

The main contributions of this paper include proposing a target detection model CN-ViT based on the Vision Transformer. This model combines the advantages of self-attention mechanisms and convolutional neural networks.2. A GCCA module was designed to enhance the comprehensiveness and localization ability of feature representation by fusing global context information and location-sensitive channel features.3. The superior performance of CN-ViT on the Mini COCO standard dataset was verified through experiments, showcasing its advantages in detection accuracy and speed.

## 2. The Method

### *2.1. Problem Definition*

Object detection is a fundamental task in the field of computer vision, aiming to automatically identify and locate the categories and positions of objects in digital images or videos. When performing object detection, it is necessary not only to recognize various objects (such as people, vehicles, and animals) in the image but also to accurately delineate their locations using bounding boxes [6]. Compared to classification tasks, object detection involves more fine-grained features; therefore, detection models often require the ability to detect features at multiple scales and handle complex features in various scenes, posing numerous challenges for accurate object detection.

This study aims to integrate the features of convolutional neural networks (CNNs) and Vision Transformers (ViTs) and propose a novel object detection model that combines the local feature extraction capability of CNNs with ViT's global feature capturing ability. This integration seeks to enhance the model's capacity for detecting multi-scale features and handling high-dimensional features, thereby improving its overall performance in object detection.

DOI: 10.54254/2755-2721/95/2024BJ0058

### 2.2. *Model Overall Structure*

The model structure of CN-ViT includes several key components: input image preprocessing, convolutional feature extraction, GCCA feature modeling and enhancement, Transformer encoder, and decoder. Figure 1 illustrates the overall architecture of CN-ViT.

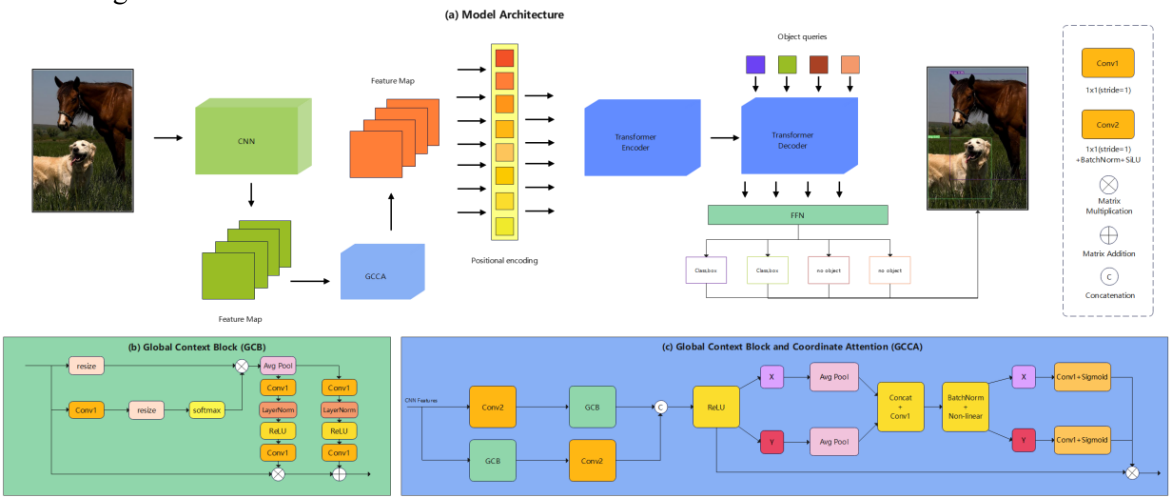


**Figure 1.** CN-ViT overall architecture.

The input image is first processed through a uniform preprocessing, including image size adjustment, normalization, and data augmentation, before being fed into the CNN to extract features, as shown in Figure 2. The features extracted by the CNN are integrated with global information and re-encoded before being fed into the ViT model for further feature extraction and output of object detection results. ViT further analyzes these features by using its self-attention mechanism to pay attention to the relationships between different regions, which is particularly important for understanding the dynamic interactions between multiple objects in complex scenes. Therefore, the CNN serves as a bridge from raw pixels to high-level semantic features in the overall architecture, which is further strengthened by the GCCA module to output high-quality feature representations. The ViT then uses these features to perform fine-grained classification and localization, achieving efficient object detection.

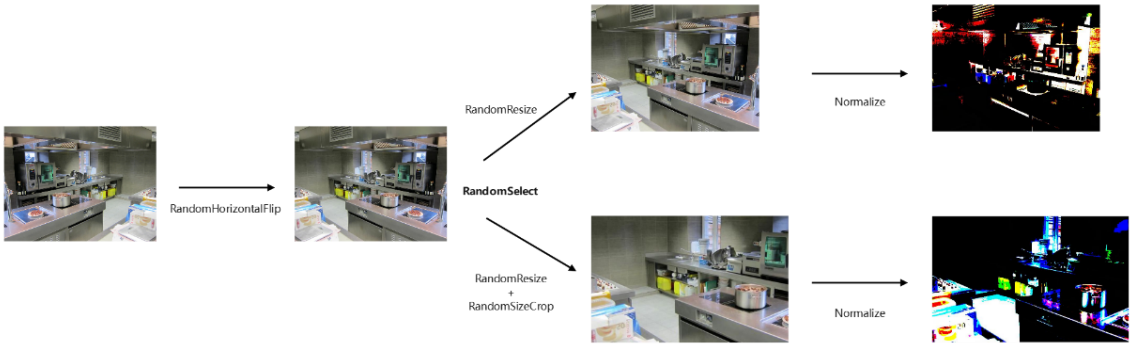


**Figure 2.** Image preprocessing.

#### 2.2.1. *Convolution module*

After preprocessing the input image, it is first fed into a CNN to extract features. Through multi-layer convolution operations, the convolution module can extract image features ranging from low-level features like edge texture to high-level image features of the object's part or overall structure layer by layer. Through the parameter sharing mechanism, the same convolution kernel is allowed to slide over the entire input image to enhance the generalization ability of the model. The CNN firstly processes the

DOI: 10.54254/2755-2721/95/2024BJ0058

input image and extracts rich Feature maps, which not only contain fine details of visual information, but also provide the necessary information basis for high-level tasks through hierarchical feature representation [10]. The processed feature maps are fed into the GCCA module for further processing to fuse global context information and location-sensitive channel features.

### *2.2.2. GCCA*

After the feature map is extracted by the CNN, this paper introduces the GCCA module, namely Global Context Block and Coordinate Attention, to further improve the feature representation ability and generalization performance of the model. The structure of GCCA is shown in Figure 3.

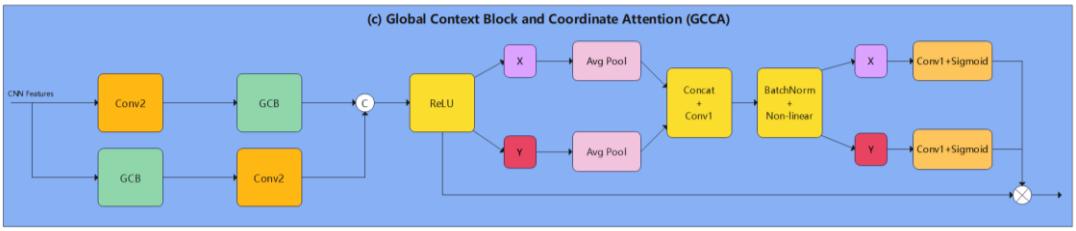


**Figure 3.** Global Context Block and Coordinate Attention structure.

1) Global Context Block (GCB), which is a mechanism to enhance the feature map representation with global information [11]. The GCB structure is shown in Figure 4.GCB extracts global context information from the CNN-processed feature maps through global pooling operations such as global average pooling or global max pooling. This global information can help the model understand the overall structure and background of the input image. Secondly, GCB fuses the global context information with the input feature map and weights each channel through the channel attention mechanism. This approach is able to emphasize channel features that are more important to the task, while suppressing less important channel features. By modeling the global context information and adjusting the channel relationship, GCB can significantly improve the representation ability of the feature map, which is helpful to the performance of the model in complex scenes.

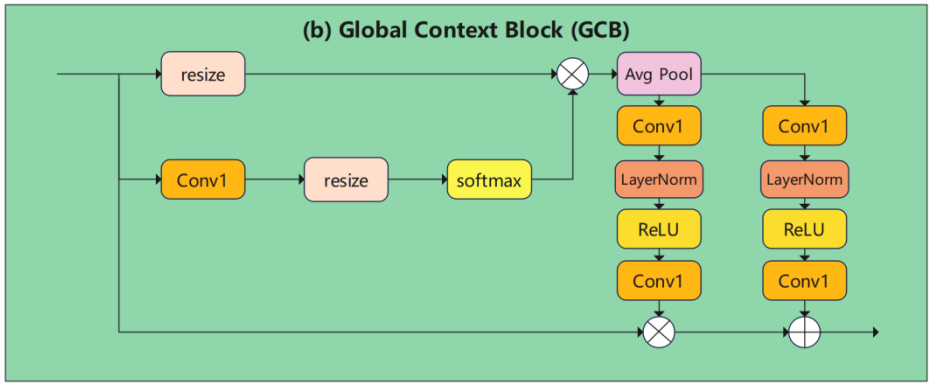


**Figure 4.** Global Context Block structure.

The implementation flow of GCB is as follows:

Assume that the input feature map is represented as (C, H, W), where C is the number of channels, H is the height, and W is the width. First, the context information is extracted by global average pooling.

$$\text{context}_{\text{global}} = spatial\_pool(X) \tag{2.1}$$

where $\text{context}_{global} \in R^{C\times1\times1}$.

DOI: 10.54254/2755-2721/95/2024BJ0058

The global context vector is convolved, normalized, and processed to generate a weight map:

$$W = Conv(Re\,L\,U(LayerNorm(Conv(context_{global})))) \tag{2.2}$$

$$W_{\text{mul}} = \sigma W \tag{2.3}$$

$$W_{\text{att}} = W \tag{2.4}$$

Here, $W_{mul}$ denotes the element-wise product weight matrix and $W_{att}$ denotes the element-wise additive weight matrix.

Finally, we perform feature map weighting and apply the weighting matrix to the input feature map:

$$X_{\text{context}} = X \otimes W_{mul} \oplus W_{att} \tag{2.5}$$

2) Coordinate Attention (CA), which is a lightweight attention mechanism that combines positional encoding and channel attention [12]. CA extracts horizontal and vertical spatial information through horizontal and vertical global pooling operations, respectively, which retains the location information of the feature map and helps the model to understand the spatial structure.

In addition, CA is implemented by simple pooling and convolution operations, which has lower computational cost than the traditional self-attention mechanism, but it can still effectively enhance the feature representation ability. Specifically, the CA implementation flow is as follows:

The input feature map is processed first by extracting spatial information through horizontal and vertical global average pooling:

$$h = \frac{1}{H}\sum_{i=1}^{H} X_{context}(:, i, :) \tag{2.6}$$

$$w = \frac{1}{W}\sum_{j=1}^{W} X_{context}(:, :, j) \tag{2.7}$$

Where $h \in R^{C\times 1\times W}$ and $w = R^{C\times H\times 1}$.

Second, H and W are passed through a convolutional layer to generate attention weights.

$$h_{\text{att}} = \sigma(Conv(h)) \tag{2.8}$$

$$w_{\text{att}} = \sigma(Conv(w)) \tag{2.9}$$

Where $\sigma$ is the activation function, $h_{\text{att}} \in R^{C\times 1\times W}$ and $w_{\text{att}} = R^{C\times H\times 1}$.

Finally, we perform feature map weighting and apply the generated attention weights to the input feature map:

$$X_{\text{GCCA}} = X_{context} \otimes h_{att} \otimes w_{att} \tag{2.10}$$

The combination of GCB and CA can make full use of global context information and spatial location information. Firstly, the feature maps extracted by CNN are enhanced with global context information by the GCB module, and then the spatial attention is enhanced by the CA module. This combination can significantly improve the representation ability of the feature map and the generalization performance of the model, and provide richer and finer feature representations for the subsequent Vision Transformer (ViT). The processed feature maps are subsequently joined with positional encoding and resized to a suitable size for ViT processing before being fed into the ViT for further feature extraction and final object detection.

### *2.2.3. The Transformer architecture*

In the CN-ViT model, the Transformer encoder is the core component, which processes image patch embeddings using a multi-layer self-attention mechanism to capture global features. The encoder of each layer is composed of a multi-head self-attention mechanism and a feed-forward neural network. The self-attention mechanism divides the image block embedding into multiple heads for independent processing, enhancing the learning ability of the model for different features. The feedforward network performs nonlinear transformation through two fully connected layers and activation function to improve the expression ability. Additionally, each layer includes residual connections and layer normalization, helping to stabilize the training process and prevent gradient problems.

### *2.2.4. CN-ViT input tuning*

In the CN-ViT model, the resizing and processing of the input image are crucial steps. The feature image output by the CNN is three-dimensional, including the height, width, and number of channels of the

DOI: 10.54254/2755-2721/95/2024BJ0058

feature. Before feeding the feature map into ViT, we need to resize the feature map to the appropriate image size of ViT.

The input to the Transformer encoder plays a key role in the whole model training and inference process. In the Vision Transformer (ViT) architecture, the introduction of positional information is crucial because unlike traditional convolutional neural networks, the Transformer architecture does not inherently have the ability to capture positional relationships in the input data. In the original architecture of ViT, position encoding is added to the input feature vector to provide positional context information, which is particularly important for image processing tasks.

*2.2.5. CN-ViT overall forward process*

In the input image, the CNN first extracts features. Assuming the input image is $I$,a series of convolution operations are performed to define a function $f_{CNN}$. This function extracts the feature map of the image $F$,which is recorded as the following formula (2.11):

$$F = f_{CNN}(I) \tag{2.11}$$

During the convolution operation, each convolution layer can be considered as performing the following operations, as shown in Equation (2.12) :

$$F_{l+1} = \sigma(W_l * F_l + b_l) \tag{2.12}$$

Where $F_l$ is the feature map of the layer $l$, $W_l$ and $b_l$ are the convolution kernel and bias of the layer $l$, respectively. $\sigma$ is the activation function ReLu used in this layer, and $*$ means to perform the convolution calculation. The output of this layer is then passed on to the next layer in the neural network.

Then, the feature map $F$extracted by CNN is modeled by GCCA. The global context of the feature map is modeled by the GCB module, and channel weighting is performed to make the expression of the feature map in the global semantics more abundant. Subsequently, the attention distribution in the feature map is further refined by the AC module to generate a more discriminative feature map:

$$F_{\mathrm{GCB}} = GCB(F) \tag{2.13}$$

$$F_{\mathrm{GCCA}} = AC(F_{GCB}) \tag{2.14}$$

Next, we resize the feature map and add positional information to match the input dimensions of the Vision Transformer (ViT), denoted as, and add positional encoding (PE):

$$F^{'} = resize(F_{GCCA}, H^{'}, W^{'}) + PE \tag{2.15}$$

Where, $resize$ is the adjusted size function and PE is the position encoding matrix, which is consistent with the adjusted feature map. Additionally, the adjusted size function plays a crucial role in ensuring that the position encoding matrix aligns perfectly with the feature map.

Feeding this into the system, the output is denoted as Z:

$$Z = f_{ViT}(F^{'}) \tag{2.16}$$

In the ViT model, there are multiple self-attention layers, and each layer operation is represented as follows:

$$Z_{l+1} = Transformer - Block(Z_l) \tag{2.17}$$

Where $Z_l$ is the output of the layer, while $Transformer - Block$ mainly consists of the attention mechanism that helps the model focus on relevant parts of the input data.

Finally, the detection head $f_{out}$ is used to predict the object class and bounding box. This is denoted as:

$$(\mathrm{C,B}) = f_{out}(Z) \tag{2.18}$$

That is, the final result is obtained.

## 3. Numerical experiments

*3.1. The dataset*

In the experiments, the Mini COCO object detection dataset was chosen for this paper. It is considered one of the most challenging and widely used datasets in the field of object detection, containing 80 categories of objects with approximately 123,000 training images and 50,000 validation images. The

DOI: 10.54254/2755-2721/95/2024BJ0058

objects in the COCO dataset are diverse and complex, including target objects at different scales, viewpoints, and backgrounds. The Mini COCO dataset is taken from the COCO 2017 dataset with 5489 images in the training set and 234 images in the validation set.

### *3.2. Experimental Setup*

The model performance largely depends on the feature extraction ability of the embedding function. As shown in Table 1, ResNet-50 is used as the backbone network in this paper. In the middle layers of these backbone networks, max pooling, ReLU nonlinear activation function and batch normalization operation are used. It is worth noting that in the implementation of this paper, the batch normalization layer is replaced by a frozen batch normalization layer to ensure that the model is more stable during training. In addition, to prevent the network from overfitting, a dropout layer is added to ResNet-50.

All experiments were performed using the deep learning framework PyTorch 1.13.1, and the hardware configuration was AMD EPYC 7543 32-core processor and Nvidia A40 (48GB) ×1 graphics card. During the experiment, the learning rate of the model was set to 1e-4 and the learning rate of the backbone network was set to 1e-5.The number of samples per batch (batch size) is 4, the weight decay factor is 1e-4, and the total number of training rounds (epochs) is 200. When training reaches the 150th round, the learning rate will decrease. To prevent gradient explosion, we set the maximum norm of gradient clipping (clip_max_norm) to 0.1. In the testing phase, 234 images are selected for evaluation in this paper, and the evaluation index of the model is the accuracy under 95% confidence.

### *3.3. Analysis of experimental results*

In order to verify the performance of the CN-ViT model proposed in this paper, we conduct comparative experiments on the Mini COCO dataset, including metric-based FPN and end-to-end detection network DETR. The average accuracy comparison of the model in medium size object detection under the Mini COCO dataset is shown in Figure 5. The average precision and average recall of the model under the Mini COCO dataset are shown in Tables 1 and 2. By analyzing the experimental results, this paper draws the following conclusions:

1) CN-ViT shows excellent performance for medium and large object detection using the same skeleton network on the Mini COCO dataset, as shown in Table 1. Specifically, for medium-size object detection, CN-ViT has an average precision improvement of 6.10 percentage points over FPN, and DETR has an average precision improvement of 4.60 percentage points over FPN. In terms of large-size object detection, CN-ViT has the average precision improved by 20.22 percentage points compared with FPN, and DETR has the average precision improved by 14.62 percentage points compared with FPN. As a result, CN-ViT performs much better in the detection of objects above medium size.

According to the experimental results in Tables 1 and 2, CN-ViT achieves the highest detection accuracy on the Mini COCO dataset compared to the control methods. CN-ViT shows a significant improvement over FPN in terms of average precision on this dataset. When IoU=0.50:0.95 and maxDets=100, the AP of CN-ViT is 12.7, significantly higher than DETR's 8.60 and FPN's 4.60, as shown in Table 1. In the same range, the AR of CN-ViT is 23.9, surpassing FPN's 7.50 and DETR's 17.1, as shown in Table 2.FPN method learns multi-scale context information of sample features by constructing feature pyramid structure. It combines bottom-up and top-down methods to obtain strong semantic features and improve the performance of object detection and instance segmentation in multiple datasets. It may lead to the reduction of the global capture ability of the model. By modeling object detection as an end-to-end sequence-to-sequence prediction problem and utilizing the global modeling ability of the Transformer structure, DETR realizes the direct prediction of the target, breaking away from the traditional anchor-based methods. DETR does not require predefined bounding boxes and classifiers, but directly predicts the category and bounding box coordinates of the target. At the same time, it enhances the global modeling ability of the target location and category by introducing a dynamic target-query matching mechanism. However, DETR also has some shortcomings, such as slow training convergence, weak detection performance, high memory usage, and limited generalization ability.

DOI: 10.54254/2755-2721/95/2024BJ0058

CN-ViT combines the features of CNN and Transformer, improves the multi-scale representation ability of the model in high-dimensional features, effectively captures long-range dependence information, and enhances the learning ability of the model on long sequences. At the same time, the introduction of GCCA provides the model with a deep fusion of global context information, thereby enhancing the diversity and robustness of the model in different application scenarios. Through the attention mechanism, CN-ViT can better capture the context information between features and strengthen the category features related to object detection, thus significantly improving the accuracy and stability of medium and large scale object detection.


**Figure 5.** Comparison of average accuracy of medium and large size object detection

**Table 1.** Average accuracy comparison under Mini COCO dataset.

| Method | Network | AP@[IoU=0.50:0.95] | $AP_s$ | $AP_m$ | $AP_l$ |
|---|---|---|---|---|---|
| FPN | ResNet-50 | 4.61 | 2.26 | 4.31 | 9.18 |
| DETR | ResNet-50 | 8.60 | 1.90 | 8.90 | 23.8 |
| CN-ViT | ResNet-50 | 12.7 | 5.00 | 15.0 | 29.4 |

**Table 2.** Comparison of average recall under Mini COCO dataset.

| Method | Network | AR@[IoU=0.50:0.95] | $AR_s$ | $AR_m$ | $AR_l$ |
|---|---|---|---|---|---|
| FPN | ResNet-50 | 7.50 | 3.30 | 7.20 | 14.4 |
| DETR | ResNet-50 | 17.1 | 2.90 | 18.5 | 41.1 |
| CN-ViT | ResNet-50 | 23.9 | 8.90 | 27.5 | 43.3 |

## 4. Conclusions

In this paper, we introduce CN-ViT, a novel visual object detection model based on Vision Transformer. The model effectively combines the strengths of the self-attention mechanism and CNN. Additionally, it incorporates Global Context Block and Coordinate Attention to address the challenges posed by multi-scale objects, intricate backgrounds, and object occlusion. Through this design, the experimental results show that CN-ViT performs significantly better than the current mainstream object detection methods on the Mini COCO standard dataset, especially in the detection accuracy and processing speed of medium and large scale objects. This paper provides a research idea of multi-scale feature fusion for the study of target detection, but there is still some room for improvement. In future work, the representation of multi-scale features and the improvement of attention mechanism can be studied to further improve the performance and adaptability of the model.

## References

[1] Srivastava S, Divekar A V, Anilkumar C, et al. Comparative analysis of deep learning image detection algorithms[J]. Journal of Big data, 2021, 8(1): 66.

[2] Jiang P, Ergu D, Liu F, et al. A Review of Yolo algorithm developments[J]. Procedia computer science, 2022, 199: 1066-1073.

DOI: 10.54254/2755-2721/95/2024BJ0058

[3] Liu W, Anguelov D, Erhan D, et al. Ssd: Single shot multibox detector[C]//Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14. Springer International Publishing, 2016: 21-37.
[4] Desnoyers P. Analytic models of SSD write performance[J]. ACM Transactions on Storage (TOS), 2014, 10(2): 1-25.
[5] Redmon J, Farhadi A. Yolov3: An incremental improvement[J]. arXiv preprint arXiv:1804.02767, 2018.
[6] Fang Y, Guo X, Chen K, et al. Accurate and automated detection of surface knots on sawn timbers using YOLO-V5 model[J]. BioResources, 2021, 16(3): 5390.
[7] Wang C Y, Yeh I H, Liao H Y M. YOLOv9: Learning What You Want to Learn Using Programmable Gradient Information[J]. arXiv preprint arXiv:2402.13616, 2024.
[8] Girshick R. Fast r-cnn[C]//Proceedings of the IEEE international conference on computer vision. 2015: 1440-1448.
[9] DosoViTskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[J]. arXiv preprint arXiv:2010.11929, 2020.
[10] Chauhan R, Ghanshala K K, Joshi R C. Convolutional neural network (CNN) for image detection and recognition[C]//2018 first international conference on secure cyber computing and communication (ICSCCC). IEEE, 2018: 278-282.
[11] CAO Y, XU J, LIN S, et al. GCNet: Non-Local Networks Meet Squeeze-Excitation Networks and Beyond[C/OL]//2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), Seoul, Korea (South). 2019.
[12] HOU Q, ZHOU D, FENG J. Coordinate Attention for Efficient Mobile Network Design[C/OL]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA. 2021.