Modern Management based on Big Data III
A.J. Tallón-Ballesteros (Ed.)

doi:10.3233/FAIA220108

Detection Method of Computer Room Personnel Based on Improved Swin Transformer

Qiong-lan NA [a1], Dan SU [a], Hui-min HE [a], Yi-xi YANG [b], Shi-jun ZHANG [a], Yi-fei WANG [a] and Yu-jia MA [c]

[a] *State Grid Jibei Information and Telecommunication Company, Beijing 100053, China*
[b] *State Grid Information and Telecommunication Branch, Beijing 100761, China*
[c] *Shanxi Pengtong Construction Project Management Co., LTD, Taiyuan 030000, China*

Abstract. The accurate detection of computer room personnel can bring great convenience to computer room management and computer room inspection. Swin Transformer is used in object detection and achieves excellent detection performance. In this paper, Swin Transformer is used as the baseline to achieve accurate detection of computer room personnel. This paper mainly makes the following two contributions:1) In this paper, a practical self-attention method is designed. The channel interaction module is used in the self-attention calculation to solve the problem of local window self-attention lacking orientation awareness and location information. Reduce the size of input tokens through depth-wise convolution to reduce the complexity of self-attention calculation. 2) Use a balanced L_1 loss and configure the weights of different stages of loss in the total loss function to solve the problem of imbalance between simple samples and difficult samples. Compared with the original Swin Transformer, the improved method improves the detection accuracy of $mAP_{@0.5}$ by 3.2%.

Keywords. Swin Transformer, Personnel Detection, Self-Attention, Depth-wise Convolution

1. Introduction

At present, the managers of many computer rooms have to be on duty 24 hours a day, relying mainly on manual observation, which will consume a lot of human resources and also cause waste of video surveillance resources. For other intruders, it may pose a threat to the data transmission, storage, and system operation of the computer room. The real-time accurate detection of personnel using video surveillance in the computer room can issue early warnings to administrators promptly, and administrators can remotely record, analyze and communicate the activities of personnel entering the computer room. Therefore, realizing the accurate detection of computer room personnel can improve the

[1] Corresponding Author: Qionglan Na, E-mail: 81885883@qq.com. This work is supported by the Science and Technology Project of State Grid Jibei Power Company Limited (No. B3018E210001).

maintenance efficiency of the unattended computer room, and provide guarantee and convenience for the management of the computer room and the inspection of the computer room.

Recently, Swin Transformer [1] was proposed to build a hierarchical feature structure, which can easily adapt to feature pyramids, etc. And it reduces the complexity from quadratic to linear based on the self-attention computation of local windows. These features make Swin Transformer useful as a general model for various vision tasks. However, Swin Transformer still has two problems in the accurate detection of people: (1) Performing self-attention within non-overlapping windows still has high computational complexity. It will lack orientation awareness and position information, that is, it cannot capture cross-channel information well. (2) During the training process, there is also an imbalance between the simple samples and the difficult samples. When the gradient is back-propagated, the gradient effect of the simple samples is too small[2].

The improved Swin Transformer can improve these two problems: In this paper, a practical self-attention method is designed, Using the channel interaction module in the self-attention calculation can solve the problem that the local window self-attention lacks direction awareness and position information. Reduce the size of input tokens through Dw convolution to reduce the complexity of self-attention calculation. Using a balanced L_1 loss and configuring the weights of different stages of loss in the total loss function to solve the loss imbalance problem.

2. Related Work

Because the scene of personnel detection in the computer room is relatively new, at present, few researchers apply and study-related detection algorithms in this scene. Next, the development status of target detection algorithms will be described from the perspectives of convolution-based target detection algorithms and vision transformer-based target detection algorithms.

Convolution Based Object Detection. The general target detection algorithm can be divided into one-stage and two-stage methods. In the one-stage detection algorithm, Overfeat [3] directly uses convolution feature graph to predict the decision value of classification and location. YOLOV1 [4] and YOLOV3 [5] regress object boundaries and category probabilities directly based on image grids. SSD [6] improves single-stage detection with multi-layer features of various sizes. RetinaNet [7] proposed Focal loss to solve the problem of foreground-background imbalance. In the two-stage detection algorithm, R-CNN [8], Fast R-CNN [9], and Faster R-CNN [10] use the pooling features of the proposed region to predict object scores and boundaries. R-FCN [11] introduced location-sensitive fractional graphs to share each ROI feature calculation. Denet [12] predicts and searches the sparse angular distribution of object boundaries. Cascade R-CNN [13] uses sequential R-CNN stages to gradually refine the detected boxes.

Vision Transformer. Recently, the Transformer-based detector DETR [14] defines target detection as a direct ensemble prediction task and has achieved excellent results. DETR predicts a set of objects by using a converter decoder to participate in the query of the feature graph. The original architecture of DETR is simply based on Transformer [15], which contains multi-layer attention encoders and decoders. The set prediction training in DETR is based on the binary matching between the prediction and the real object. Although DETR is better than the competitive Faster R-CNN baseline, it still has some problems, such as limited spatial resolution, the poor performance of small object

detection, and slow convergence speed of training. There are already several tasks to solve these problems. Deformable DETR [16] considers the shift equivalence in natural images and introduces a series of multi-scale deformable attention operators into the encoders and decoders of DETR.

3. Proposed Method

3.1. Overall Architecture

The network structure of this paper consists of four parts, including the Swin-T backbone network, the feature pyramid (FPN), the region proposal network (RPN), and the cascaded detection head. Swin Transformer is used to extract image features, and FPN is mainly used to extract multi-scale features. RPN is a combination of several convolutional layers that produce regions of interest (ROIs) where objects may be present. The cascaded detection heads classify and localize the region of interest, and output the final detection result. In the cascade detection head, ROI Align is used for regional feature alignment, POOL represents the last part of feature extraction. FC is the fully connected layer, C is the classification probability, and B is the regression of the candidate box.

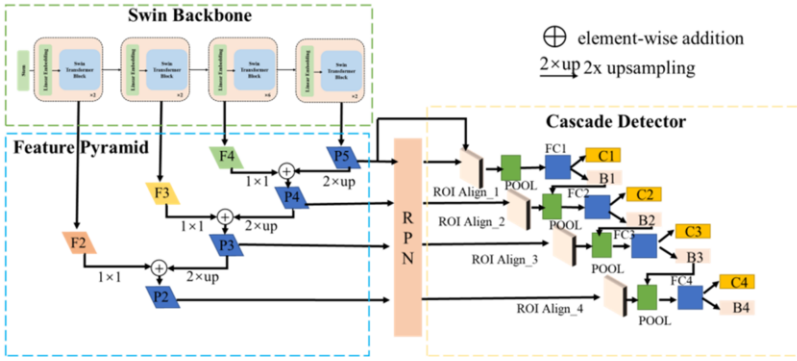

Figure 1. The overall framework of the improved algorithm

3.2. Improved Swin Transformer Backbone Network

Improved self-attention module. The improved attention module in this paper adds two key designs to the standard window-based attention module (W-MSA): (1) Design a self-attention mechanism that requires less computation to reduce the computational complexity of the self-attention mechanism spend. (2) The calculation of the V value uses the channel interaction module to solve the problem of lack of orientation awareness and position information. This paper integrates these two key points and builds an improved self-attention module. Details are described next. As shown in Figure2(c), the query $Q \in \mathbb{R}^{n \times d_k}$ is obtained by linearly projecting the input $X \in \mathbb{R}^{n \times d_m}$, where $n = H \times W$. Reshape the input $X \in \mathbb{R}^{n \times d_m}$ to a spatial vector (d_m, H, W). Reduce the size of input X by Depth-wise Convolution with kernel $s \times s$ and stride s. The size of tokens changed from (d_m, H, W) to $\left(d_m, \frac{H}{s}, \frac{W}{s}\right)$. After linear transformation to get $K \in \mathbb{R}^{n' \times d_k}$. where X is the input token, n is the number of blocks, H is the number of image blocks

in the height direction of the input image, W is the number of image blocks in the width direction of the input image, dand m is the embedding dimension of each image block, the query vector dimension, the key vector The embedding dimension of the sum-value vector is d_k, and n' is the number of blocks.

For the value of V, we add the channel interaction module to calculate. Inspired by SE, the channel interaction consists of aan DW_2, a global average pooling layer (GAP). Then there are two consecutive 1×1 convolutional layers, batch normalization (BN) and activation between them (SILU). Finally, we use sigmoid to generate attention in the channel dimension. The formula for calculating V is as follows:

$$V = FC(LN(DW_1(X))).Sigmoid(conv(SILU(BN(conv(GAP(DW_2(x))))))) \quad (1)$$

Among them, FC is fully connected, LN is layer normalization, BN is batch normalization, DW_1 is depth-wise convolution, X is input token vector, conv is 1×1 convolution, GAP is global average pooling, and DW_2 is depth-wise convolution.

where $V \in \mathbb{R}^{n' \times d_k}$, $n' = \frac{H}{s} \times \frac{W}{s}$. where DW_2 is the Depth-wise Convolution with a convolution kernel of 3×3, and the difference between DW_1 and DW_2 needs to be noted here. The input size of DW_1 is reduced by a factor of s, DW_2 does not change the size and number of channels of the input, More channel information is preserved. Then, the self-attention functions of Q, K, and V are calculated by the following formula 2:

$$MSA(Q,K,V) = LN\left(Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)\right)V \quad (2)$$

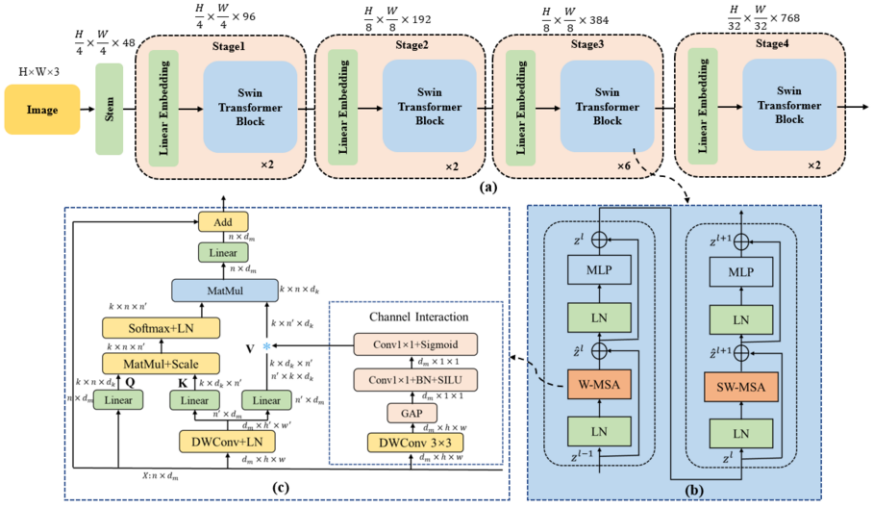

Figure 2. Improved Swin-T backbone network

3.3. Improved Loss Function

RPN classification loss and cascade detection head loss. This paper uses the multivariate cross-entropy loss function, and the goal of bounding box classification assigns C+1class labels to each bounding box, denoted by probability p. Among them, C is all the categories, and 1 is the background. For the training samples x_i and y_i, where

y_i is the true label value of the input x_i, the multivariate cross-entropy loss function is as formula 3:

$$CE(p,y) = -\frac{1}{C+1}\sum_{j=0}^{C+1} W_j log(p_i) \tag{3}$$

Among them, W_j is as formula 4:

$$W_j = \begin{cases} 1, |x| < 1 \\ 0, otherwise \end{cases} \tag{4}$$

RPN bounding box regression loss. Bounding box regression aims to regress the candidate bounding box $b = (b_x, b_y, b_w, b_h)$ to the target bounding box $g = (g_x, g_y, g_w, g_h)$ using the regression function, minimizing the loss function $L_{BIoc}(b_i, g_i)$ as formula 5:

$$L_{BIoc}(b_i, g_i) = \frac{1}{N_{reg}}\sum_i P_i^* L_{Ioc}(b_i, g_i) \tag{5}$$

Among them,

$$L_{Ioc}(b_i, g_i) = \sum_i Smooth_{L_1}(b_i - g_i) \tag{6}$$

$$smooth_{L_1} = \begin{cases} 0.5x^2, |x| < 1 \\ |x| - 0.5, otherwise \end{cases} \tag{7}$$

Among them, N_{reg} represents the number of anchor positions, P_i^* is 1 when the candidate frame is a positive sample, P_i^* is 0 when the candidate frame is a negative sample, b_i represents the bounding box regression parameter for predicting the i-th anchor , g_i represents the ground-truth box corresponding to the i-th anchor.

Cascaded detection head bounding box regression loss. When the weight of the regression loss increases, the model is very sensitive to abnormal values of the regression coordinates. In this paper, $L1_{balanced}$ is used as follows:

$$L1_{balanced}(\hat{x}) = \begin{cases} \frac{\alpha}{b}(b|\hat{x}| + 1)ln(b|\hat{x}| + 1), |\hat{x}| < 1 \\ \gamma|\hat{x}| + C, otherwise \end{cases} \tag{8}$$

Among them,

$$\alpha ln(b + 1) = \gamma \tag{9}$$

Among them, α is the weight coefficient of outliers, γ is used to limit the range of outliers, α and γ are hyperparameters, and the default values are set to 0.5 and 1.5. The parameter b can ensure that the derivative is continuous, and C is a constant. The relationship between hyperparameters is shown in Equation 9.

Total loss. The total loss is defined as follows:

$$L = aL_{RPN} + bL_{stage1} + cL_{stage2} + dL_{stage3} \tag{10}$$

Among them,

$$L_{RPN} = L_{RPN\_cls} + L_{RPN\_reg} \tag{11}$$

$$L_{stage1} = L_{stage1\_cls} + L_{stage1\_reg} \tag{12}$$

$$L_{stage2} = L_{stage2\_cls} + L_{stage2\_reg} \quad (13)$$

$$L_{stage3} = L_{stage3\_cls} + L_{stage3\_reg} \quad (14)$$

a, b, c and d represent the weight coefficients of the loss. For the computer room personnel detection task, the weight coefficients a, b, c, and d are set to [1, 0.75, 0.5, 0.25], respectively. L_{RPN_cls} represents the RPN classification loss, and L_{RPN_reg} is the RPN regression loss. L_{stage1}, L_{stage2} and L_{stage3} represent the total loss of the three stages, L_{stage1_cls}, L_{stage2_cls} and L_{stage3_cls} are the classification losses of each stage, L_{stage1_reg}, L_{stage2_reg} and L_{stage3_reg} are the regression losses for each stage. L_{stage1_reg}, $L_{stage2_reg}$and L_{stage3_reg} apply a balanced L1 loss function.

4. Results and Discussion

4.1. Experiment setup

Constructing a dataset: We extract surveillance videos from multiple computer rooms, cut frames manually, and filter according to lighting conditions, picture integrity, etc. Annotate according to the normalization method to construct a person detection dataset. The dataset contains 5137 images and 14398 labels, and the label's category is person. Most of the image sizes in the dataset are 1980*1080, and a few are 720P.

Experimental setup: The IOU threshold for cascaded detection heads is [0.5, 0.6, 0.7]. The batch size is set to 64, the learning rate is kept at 0.01 during the 1st to 7th epochs, and dropped by a factor of 0.1 after the 8th to 12th epochs. ADAMW optimizer, Data augmentation includes random horizontal flipping, random scaling, and random cropping. In the Swin Transformer module, the Patch Size is set to 2×2, the head is set to 8, and the embedding is set to 64, The experimental equipment in this paper uses two GTX 3090s, and the improved algorithm is implemented based on MMdetection.

4.2. Experimental results

To verify the performance of the improved detection network, this paper will compare with the current popular target detection algorithms, test DETR based on ResNet50, Deformable DETR based on ResNet50, YOLOX-x, Retinanet based on ResNext101, and using FPN, YOLOF algorithm based on ResNet50 and The detection performance of the cascaded Swin Transformer-T and other algorithms is shown in Table 1. First, considering the mAP value of *IoU=0.5,* the detection accuracy of Swin Transformer is better than that of DETR, Deformable DETR, and YOLOF, and the detection accuracy of Swin Transformer is 0.05 points higher than that of the one-stage algorithm Retinanet. This seems to indicate that the two-stage algorithm has higher detection accuracy than the end-to-end and one-stage algorithms in the field of computer room personnel detection. Of course, this does not include YOLOX-x, because YOLOX is mainly a collection of various techniques. The detection accuracy of YOLOX-x is better than that of Swin Transformer. Deformable DETR is 1.4 points higher than DETR in detection accuracy. The detection network in this paper is based on the improved Swin

Transformer detection network $mAP_{@0.5}$ is 89.8%, which is 3.2 points higher than the detection accuracy of Swin-T. The detection accuracy of the improved algorithm in this paper is 0.6 points higher than that of YOLOX-x. Secondly, in terms of detecting small objects, the improved algorithm in this paper is second only to YOLOX in detecting small objects, and the DETR series algorithms have the worst performance in detecting small objects. Deformable DETR has a higher detection performance than DETR for small objects. In addition to YOLOX, the two-stage detection algorithm is significantly better than the one-stage algorithm in detecting small objects. Finally, in terms of model complexity, to ensure fairness, the input size of all target detection networks is set to (3, 1280, 800). Compared with Swin Transformer, the GFLOPS in this paper is reduced by 5.43G and the number of parameters is increased by 3.42M. The improved method in this paper has a huge improvement in the detection accuracy of the Swin Transformer.

Table 1. Results of different detection methods

Method	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	GFLOPs	Par
DETR[14]	0.469	0.794	0.408	0.238	0.438	0.574	91.64	41.3
De_DETR[16]	0.486	0.808	0.426	0.288	0.445	0.559	195.26	39.84
Retinanet[17]	0.502	0.816	0.452	0.319	0.488	0.593	315.39	56.74
YOLOX-x[18]	0.608	0.893	0.652	0.384	0.525	0.702	352.42	99.07
YOLOF[19]	0.504	0.818	0.443	0.287	0.452	0.583	99.98	43.88
Swinr[1]	0.588	0.866	0.627	0.378	0.521	0.715	263.78	47.79
Improved Swin	0.614	**0.898**	0.645	0.381	0.505	0.722	258.35	51.21

4.3. Ablation experiment

We use Swin-T as the baseline algorithm, and then test each improved detection performance on this basis, and measure the detection performance by $mAP_{@0.5}$. The experimental results are shown in Table 2, and the detection accuracy is improved by 1.83 points using the improved Swin Transformer module (ISTB). Using a balanced loss function (BLOSS), the detection accuracy improves by 1.4 points. Therefore, the comprehensive use of the improved method in this paper can improve the detection accuracy of Swin Transformer.

Table 2. Results of different detection methods

Method	$mAP_{@0.5}$
Swin-T	0.866
+ISTB	0.884
+ISTB+BLOSS	0.898

4.4. Visualization

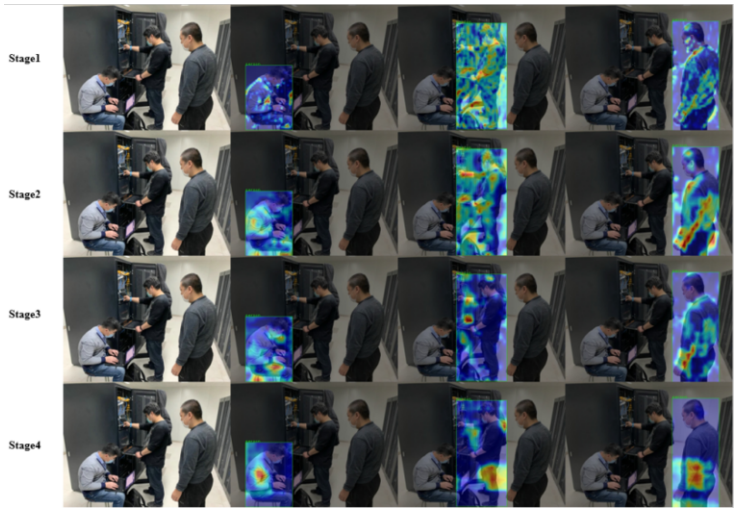

Figure 3. Multi-self-attention visualization of different stages of Swin-T

Visualize the multi-self-attention of different stages of Swin-T to illustrate how the channel interaction module collects global information. As shown in Figure 3, the first and second columns represent the Stage and the original image of the backbone network; while the other columns represent the heatmap visualization of the last multi-head self-attention feature of the same target in different Stages, The visualization shows that the channel interaction module seems to be able to select globally important regions and suppress background regions for better detection.

5. Results and Discussion

In this paper, Swin Transformer is used for the first time in computer room personnel detection. In this paper, a practical self-attention method is designed. The use of channel interaction module in self-attention calculation can solve the problem of limited self-attention receptive field of local windows. Depth-wise Convolution is used to reduce the size of the input tokens to reduce the self-attention calculation complexity. In addition, a balanced L_1 loss is adopted and the weights of different stage losses are configured in the total loss function to solve the gradient imbalance problem. Experiments show that The improved method can realize the accurate detection of the personnel in the computer room, improve the maintenance efficiency of the unattended computer room, and provide guarantee and convenience for the management of the computer room and the inspection of the computer room. Our research on detecting small targets is not enough, and improving the detection accuracy of small targets will be our next research direction.

References

[1] Liu Z, Lin Y, Cao Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 10012-10022.
[2] Pang J, Chen K, Shi J, et al. Libra r-cnn: Towards balanced learning for object detection[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 821-830.
[3] Sermanet P, Eigen D, Zhang X, et al. Overfeat: Integrated recognition, localization and detection using convolutional networks[J]. arXiv preprint arXiv:1312.6229, 2013.
[4] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 779-788.
[5] Redmon J, Farhadi A. Yolov3: An incremental improvement[J]. arXiv preprint arXiv:1804.02767, 2018.
[6] Liu W, Anguelov D, Erhan D, et al. Ssd: Single shot multibox detector[C]//European conference on computer vision. Springer, Cham, 2016: 21-37.
[7] Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection[C]//Proceedings of the IEEE international conference on computer vision. 2017: 2980-2988.
[8] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2014: 580-587.
[9] Girshick R. Fast r-cnn[C]//Proceedings of the IEEE international conference on computer vision. 2015: 1440-1448.
[10] Ren S, He K, Girshick R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[J]. Advances in neural information processing systems, 2015, 28.
[11] Dai J, Li Y, He K, et al. R-fcn: Object detection via region-based fully convolutional networks[J]. Advances in neural information processing systems, 2016, 29.
[12] Tychsen-Smith L, Petersson L. Denet: Scalable real-time object detection with directed sparse sampling[C]//Proceedings of the IEEE international conference on computer vision. 2017: 428-436.
[13] Cai Z, Vasconcelos N. Cascade r-cnn: Delving into high quality object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 6154-6162.
[14] Carion N, Massa F, Synnaeve G, et al. End-to-end object detection with transformers[C]//European conference on computer vision. Springer, Cham, 2020: 213-229.
[15] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.
[16] Zhu X, Su W, Lu L, et al. Deformable detr: Deformable transformers for end-to-end object detection[J]. arXiv preprint arXiv:2010.04159, 2020.
[17] Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection[C]//Proceedings of the IEEE international conference on computer vision. 2017: 2980-2988.
[18] Ge Z, Liu S, Wang F, et al. Yolox: Exceeding yolo series in 2021[J]. arXiv preprint arXiv:2107.08430, 2021.
[19] Chen Q, Wang Y, Yang T, et al. You only look one-level feature[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 13039-13048.