Journal on Artificial Intelligence
DOI:10.32604/jai.2021.026902
Article

Tech Science Press

Research of Insect Recognition Based on Improved YOLOv5

Zhong Yuan[1], Wei Fang[1,2,*], Yongming Zhao[3,*] and Victor S. Sheng[4]

[1]School of Computer and Software, Engineering Research Center of Digital Forensics, Ministry of Education, Nanjing University of Information Science and Technology, Nanjing, 210044, China

[2]State Key Laboratory of Severe Weather, Chinese Academy of Meteorological Sciences, Beijing, 100081, China

[3]China Meteorological Administration Training Center, Beijing, 100081, China

[4]Department of Computer, Texas Tech University, Lubbock, TX 79409, USA

[*]Corresponding Authors: Wei Fang. Email: Fangwei@nuist.edu.cn; Yongming Zhao. Email: zhaoym@cma.cn

Received: 06 January 2022; Accepted: 21 January 2022

Abstract: Insects play an important role in the natural ecology, it is of great significance for ecology to research on insects. Nowadays, the invasion of alien species has brought serious troubles and a lot of losses to local life. However, there is still much room for improvement in the accuracy of insect recognition to effectively prevent the invasion of alien species. As the latest target detection algorithm, YOLOv5 has been used in various scene detection tasks, because of its powerful recognition capabilities and extremely high accuracy. As the problem of imbalance of feature maps at different scales will affect the accuracy of recognition, we propose that adding an attention mechanism based on YOLOv5. The channel attention module and the spatial attention module are added to highlight the important information in the feature map and weaken the secondary information, enhancing the recognition ability of the network. Through training on self-made insect data sets, experimental results show that the mAP@0.5 value reaches 92.5% and the F1 score reaches 0.91. Compared with YOLOv5, the map has increased by 1.7%, and the F1 score has increased by 0.02, proving the effectiveness of insect recognition based on improved YOLOv5. In conclusion, we provide effective technical support for insect identification, especially for pest identification.

Keywords: YOLOv5; attention mechanism; insect identification

1 Introduction

Insect species account for more than four-fifths of all biological species in nature, and they are an indispensable part of the earth's ecology [1]. On the one hand, insects are beneficial to humans, because the anti-biomass produced in insects has strong bactericidal substances, and genetic engineering can be used to breed pest-resistant varieties. In addition, insect toxins can also be used to treat human various diseases [2]. For example, bee venom can be used to treat rheumatoid arthritis [3], hypertension, and other diseases. Cantharidin has obvious anti-cancer effects [4] and could be used to treat cancer. On another hand, insects often cause great losses to humans. China's annual food loss due to insect pests reaches 50 billion kilograms. It also spreads diseases, some insects carry a large number of germs, and bring them into the body by biting the human body. For example, mosquitoes may cause human diseases such as Japanese encephalitis, malaria, dengue fever, filariasis, and yellow fever, etc. [5].

The timely and accurate identification of insects plays an important role in the protection of the natural environment and the protection of human interests. Invasion of alien species often occurs [6]. In 1987, Platydia saccharin invaded Guangzhou, China from Brazilwood, and then spread to Beijing, threatening crops such as sugarcane, sugarcane, and corn. In 1998, the Coleptera was discovered in Shanxi, China, and then spread to Henan and other places led to the death of a large number of trees in the invaded areas. At present, more than 660 kinds of invasive alien species have been discovered in China, which has brought huge impacts to the environment and greatly harmed the interests of mankind. It is estimated that by 2050, the number of alien species on each continent will increase by 36% [7]. Therefore, timely and accurate identification of insects can protect our ecological environment and protect human interests from harm.

The traditional method of insect identification is mainly based on the human eyes to observe the appearance characteristics, stripe pattern, and color characteristics of the insects. However for ordinary people, their inventory of insect knowledge is not enough to identify invading insects, and they cannot provide timely feedback on the situation. Nowadays, thanks to the continuous advancement of science and technology along with the continuous development of deep learning and machine learning, it is simple and easier to recognize insects. The weights and biases obtained by training on the existing data set from the higher computing power device can be directly applied to the lower computing power device [8].

The R-CNN (Region-Convolutional Neural Network) proposed in 2014 is a milestone in the application of convolutional networks to target detection [9]. Using the feature extraction and classification performance of CNN, the Region Proposal method is used to achieve target detection. However, the efficiency is low and the images corresponding to multiple candidate regions need to be extracted in advance. The Fast R-CNN (Fast Regions with CNN) proposed later has been greatly improved in speed, but it uses selective search to extract candidate regions [10]. The time is much longer than the extraction feature classification time, and the true end-to-end training mode has not been realized. Then there was Faster R-CNN (Faster Regions with CNN), which became a region suggestion network, and region generation, feature extraction, classifier classification, and regressor regression were all completed by the network, thus getting rid of the traditional manual feature extraction method [11].

The recognition of insects belongs to target detection. At present, a large number of scholars have carried out a series of research on insect recognition. Shen et al. [12] used Faster R-CNN to recognize stored-grain insects, and its average precision reaches 88%. Xia et al. [13] proposed a model based on a convolutional neural network, and improve it by replacing Region Proposal Network is adopted instead of the traditional selective search technique, the average precision reaches 89.2%.

The models proposed by these scholars still have a lot of room for improvement inaccuracy. In this paper, insect recognition is used as a detection task, and 2747 insect pictures collected on the internet are preprocessed to construct a data set of insects, based on the improved YOLOv5 network model, to improve and train. The results show that the model trained by improved YOLOv5 has higher recognition accuracy.

2 Related Work

2.1 YOLOv5

In 2015, a one-stage detection model YOLO (You Only Look Once) was proposed. Compared with Faster R-CNN, which requires repeated training of the RPN (Region Proposal Network) and Fast R-CNN, YOLO only needs one time [14].

YOLO is unified into a regression problem, which is different from R-CNN which divides the detection results into two steps: object category and object position. However, since the output layer is a fully connected layer, during detection, the YOLO training model only supports the same input resolution as the training image. In addition, although there are multiple objects in each detected grid, only one of them can be detected at a time. In YOLOv2, a new training method is adopted, that is, the joint training method, and more objects are recognized, faster and more accurate [15–16]. YOLOv3 can weigh the speed and accuracy by changing the size of the model structure. It is faster and 100 times faster than Fast R-CNN [17–18]. YOLOv4 can reach 43.5% AP and the speed reaches 65FPS, which is faster and more accurate. The size of YOLOv5 is only 27 MB, which is nearly 90% smaller than YOLOv4, and is equivalent to YOLOv4 in accuracy. This shows the powerful capabilities of YOLOv5 [19–20].

There are four versions of YOLOv5, namely YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x. YOLOv5s contains the smallest depth and feature map width and has the fastest training speed, while the other three are based on YOLOv5s, deepened and widened. Fig. 1 is the network model structure of YOLOv5:

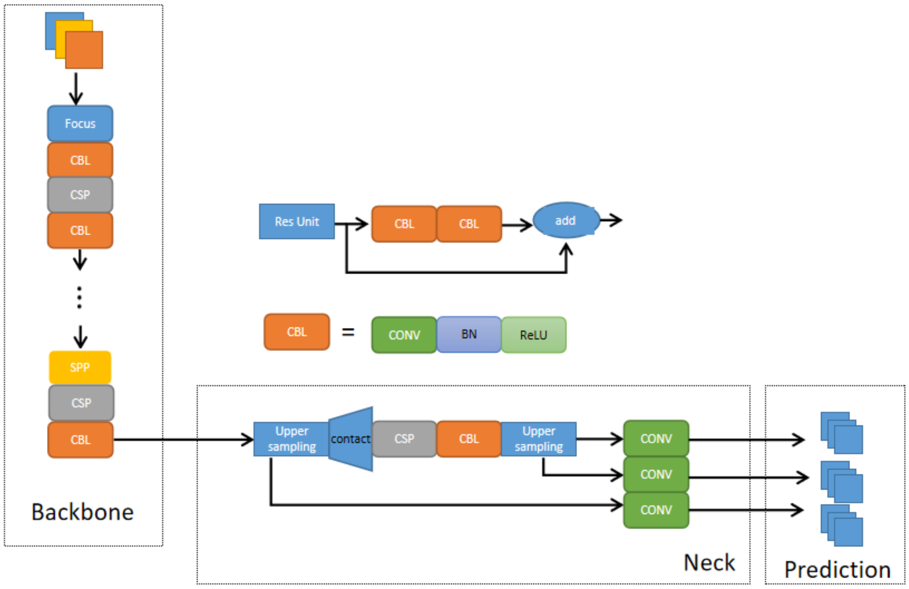

Figure 1: The network model structure of YOLOv5

The YOLOv5 network structure consists of four parts: Input, Backbone, Neck, and Prediction. At the Input, YOLOv5 uses a different image enhancement method from YOLOv4, it is mosaic data enhancement. Mosaic data enhancement is to train by randomly cutting four pictures and synthesizing one picture. Focus structure and CSP structure are used in Backbone to aggregate and form a convolutional neural network of image features on different image granularities. Adding the FPN+PAN (Path Aggregation Network) structure to Neck is a series of network layers that mix and combine image features, and transfer the image features to the prediction layer. In Prediction, the loss function of the anchor box is improved from CIOU (Complete IoU) loss to GIoU (Generalized IoU) loss, as shown in Eq. (1).

$$GIoU = IoU - \frac{|C(A \cup B)|}{|C|} \tag{1}$$

A and B represent two different boxes, C represents a box that frames A and B at the same time.

2.2 Attention Mechanism

Through training, the neural network can understand the areas that need attention in each new image, thereby forming attention. In recent years, more and more scholars have applied the attention mechanism to deep learning, and confirmed that the attention mechanism can improve the performance of the model [21–23].

The Spatial Transformation Network (STN) was proposed by Jaderberg et al. [24] in 2015. STN is a learnable module that is placed in CNN. It can transform itself through the features of the feature map and increase the spatial invariance even if the input is Transformation and slight modification, the model can also recognize and recognize the ability of features.

Squeeze-and-Excitation Networks (SENet) was proposed by Hu et al. [25] in 2017. It is divided into two parts, one is the compression part and the other is the excitation part. The function of the compression part is to compress the channel direction of the input feature map, generally compressed into one dimension. The function of the excitation part is to predict the importance of each channel, and then apply (stimulate) the corresponding channel of the previous feature map after obtaining the importance of different channels, and then perform subsequent operations.

In 2018, Woo et al. [26] proposed the Convolution Block Attention Module (CBAM), which uses the Channel Attention Module (CAM) and the spatial attention module to extract the weight distribution in feature learning, and then use this weight distribution to apply to the original features before, change the original feature distribution, enhance effective features, and suppress ineffective features.

Previous experience has proved that the attention mechanism can make the network pay more attention to the main features, so that the network recognition accuracy rate is higher and the effect is better. So adding attention mechanism to YOLOv5 is a good choice.

3 Method

The attention mechanism comes from the attention distribution of humans when observing things. When humans determine what kind of creature it is, humans will pay more attention to the characteristics of the insect's head and body patterns, while ignoring the green leaves, leaf veins and other details. If you directly input the picture into the network, the network will not weaken the secondary information such as green leaf veins, resulting in poor discrimination. Adding an attention mechanism can weaken the secondary information and emphasize the head, pattern and other important information.

CBAM is an attention mechanism module, which is embedded in the CONV of YOLOv5, so that the network has the ability of attention mechanism. Taking the CBL module in YOLOv5 as an example, the original network structure of the CBL module is as shown in Fig. 2.

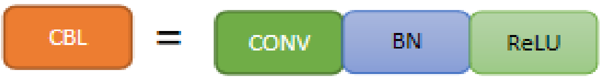

Figure 2: CBL structure

The CBL module is composed of three parts, CONV+BN+ReLU. Combining the CONV among them, the following structure will be obtained, it is shown in Fig. 3.

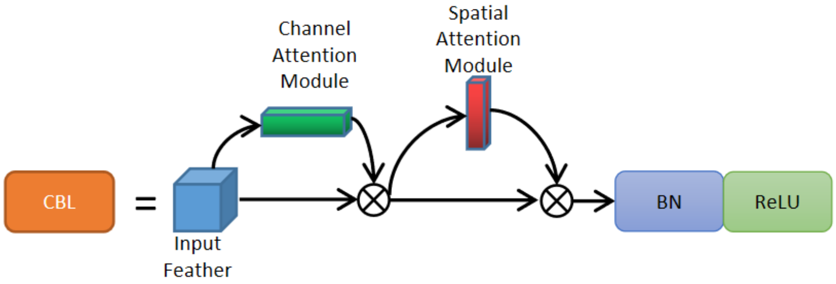

Figure 3: The structure of the CBL after the change

The channel attention module and spatial attention module are added to replace the original CONV, so that the network strengthens important information while reducing the interference caused by unimportant information. Set the module feature map input to the Channel Attention Module, where H represents the length of the feature map, W represents the width of the feature map, and C represents the number of channels in the feature map. The Channel Attention Module process is to pass the input feature map through Max Pooling based on width, then average pooling based on height, and then add and operate through MLP, and finally through sigmoid activation operation to form the final Channel Attention Feature map, as shown in Eq. (2).

$$M_c(F) = \sigma(MLP(AvgPool(F)) + MPL(MaxPool(F))) \tag{2}$$

The formula σ is the sigmoid activation function, and AvgPool(F) and MaxPool(F) are the mean pooling layer and the maximum pooling layer, respectively. As shown in Eq. (3) and Eq. (4). MLP is a simple Multi-Layer Perceptron with three layers.

$$AvgPool\big(F(i,j)\big) = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W}F(i,j) \tag{3}$$

$$MaxPool(F(i,j)) = argmax(\sum_{i=1}^{H}\sum_{j=1}^{W}F(i,j) \tag{4}$$

For the Spatial Attention Module, it takes the Channel Attention feature map as the input feature map, does a channel-based global max pooling and global average pooling, and then performs contact operations on these two results based on the channel. Then after a convolution operation, the dimensionality is reduced to 1 channel. Then generate spatial attention feature through sigmoid. Finally, the feature and the input feature of the module are multiplied to obtain the final generated feature, as shown in Eq. (5).

$$M_s(F) = \sigma(f^{7\times 7}([AvgPool(F); MaxPool(F)])) \tag{5}$$

4 Experiments

4.1 Insect Dataset

The data set in this paper is mainly derived from public data sets, on-site collection, and internet crawling. These data include different environments, different resolutions, and different types of pictures, as shown in Fig. 4. After data cleaning, filtering, and eliminating images with too low resolution, in the end, a total of 17 species of insects were collected in this article, with a total of 2474 pictures.

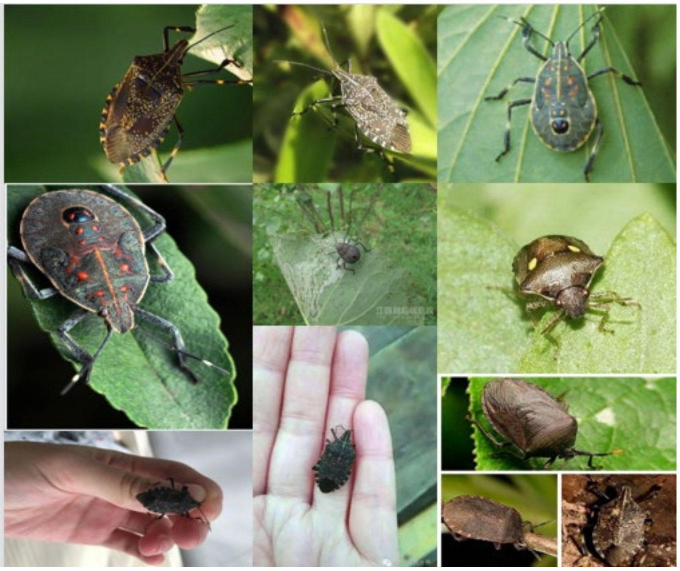

Figure 4: Sample image of data set

4.2 Experiments Environment

This experiment is based on the Windows platform and uses the deep learning framework PyTorch to build the network model. The number of iterations of the experimental sample is 50 times, the batch size is set to 4, and the image resolution is set to 640 × 640. In addition, GPU (Graphics Processing Unit) acceleration is also based on CUDA (Compute Unified Device Architecture) to improve computer graphics computing capabilities. The specific experimental environment is shown in Table 1.

Table 1: Experiments environment

Configuration	Version
System	Windows10
GPU	GeForce RTX 2080 Ti
CPU	AMD Ryzen 7 3800X 8-Core
Python	3.8
Pytorch	1.9
CUDA	10.2

4.3 Experimental Results and Evaluation Indicators

4.3.1 Evaluation Indicators

This model uses the evaluation indicators of precision, recall, F1 score, and average accuracy to evaluate the optimization of the model.

The detection results are divided into four categories, namely True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). TP means that the result is the correct identification of the category of the insects on the picture, TN means that the result is that the insects on the picture do not belong to a certain category, and FP means that the result is that the insects that are not of this category are mistakenly identified as this insect of the genus, FN indicates that the result is that the insects belonging to this genus are mistakenly regarded as not belonging to this genus.

Precision is used to reflect the classification ability of the model. It is obtained by dividing the number of correct samples in the test results by the number of all predicted samples. The precision rate calculation formula is shown in Eq. (6).

$$Precision = \frac{TP}{TP+FP} \tag{6}$$

Recall is used to reflect the model's ability to detect the target. It is obtained by dividing the correct number of samples in the result by the number of all samples, not all predicted samples, but all true samples. The calculation formula is shown in Eq. (7):

$$Recall = \frac{TP}{TP+FN} \tag{7}$$

AP stands for average precision. There are two methods to calculate the value of AP. The first method AP is obtained by averaging the precision of each class in multi-class prediction. The second method is to use Precision and Recall first. Draw a curve as the ordinate and abscissa respectively, that is, the PR curve. The area under the PR curve is the AP value. mAP is a commonly used evaluation index to evaluate the detection accuracy of the network, and its value reflects the effect of the network detection. As shown in Eq. (8).

$$mAP = \frac{1}{|Q_R|}\sum_{q\epsilon Q_R} AP(q) \tag{8}$$

F1 score is defined as the harmonic average of precision and recall, which combines the results of Precision and Recall. Generally, if you look at Precision or Recall alone, you cannot judge the effect of a model. If Precision is very high but Recall is very low, this model cannot be considered effective. The formula for the F1 score is shown in Eq. (9).

$$F1\ score\ = 2 \times \frac{precision \times recall}{precision \times recall} \tag{9}$$

4.3.2 Results and Analysis

As shown in Fig. 5, the average precision, recall rate, and F1 score are used as the evaluation indicators of model performance to measure the target detection ability and practical application ability of this model. During the training process, we can see that in the first 20 epochs, the training accuracy has improved significantly. After 40 epochs, the accuracy was maintained at a high level. In the first 20 epochs, mAP@0.5 is steadily increasing, and the rate of increase is fast. When it comes to 40 epochs, it is relatively stable. We can see that our model has a high enough precision in both Precision, Recall and mAP@0.5. mAP stands for mean precision, which is a more accurate and effective interpretation of the effect of the model, our model has achieved a high mAP value, which can meet the needs of most scenarios.

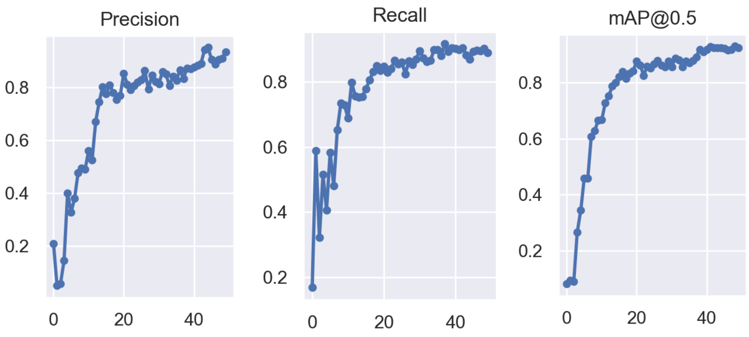

Figure 5: Evaluation indicators

In addition, when the confidence is 0.546, the F1 score reaches the maximum value of 0.90, and when the confidence is 0.901, the maximum Precision value is 1.00. The maximum Recall value is 0.98 when the confidence level is 0.0, and the maximum mAP value is 0.925 when the threshold IoU is 0.5. F1 score balances Precision and Recall, reflecting the overall effect of the model. The higher the F1 score score, the higher the precision and recall of the model, and the better the model effect. Our model F1 score has reached 0.901, it is high enough to meet the needs of most scenarios. The following table is a comparison of the two models, it is shown in Table 2.

Table 2: Comparison of result of detection models

Model	Precision%	Recall%	mAP@0.5/%	F1 score
YOLOv5l	92.3	87.6	90.8	0.89
Ours	93.3	88.9	92.5	0.91

Taken together, the use of a model based on YOLOv5l+CABM is better than pure YOLOv5l. The accuracy rate is increased by 1%, the recall rate is increased by 1.3%, the mAP@0.5 value is increased by 1.7%, and the F1 score is increased by 0.02. See that the improved model works better than pure YOLOv5l.

5 Conclusion

This paper proposes an insect recognition algorithm based on YOLOv5+CABM, which can be used in customs inspections and daily recognition, and can accurately identify the types of insects. This is also important for the prevention of invasive species and the protection of the ecological environment. Since the YOLOv5 model is still large and cannot be directly applied to the mobile terminal, in the future, it can be considered to combine MobileNet with YOLOv5 to greatly improve the detection speed with a certain loss of accuracy. In addition, you can also consider combining the CABM module with the residual network to improve accuracy.

Funding Statement: This work is supported by the National Natural Science Foundation of China (Grant No. 42075007), the Open Grants of the State Key Laboratory of Severe Weather (No. 2021LASW-B19).

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

Reference

[1] R. K. Didham, Y. Basset and C. M. Collins, “Interpreting insect declines: Seven challenges and a way forward,” *Insect Conservation and Diversity*, vol. 13, no. 2, pp. 103–104, 2020.

[2] H. S. Kachel, S. D. Buckingham and D. B. Sattelle, “Insect toxins–selective pharmacological tools and drug/chemical leads,” *Current Opinion in Insect Science*, vol. 30, pp. 93–98, 2018.

[3] R. Wehbe, J. Frangieh and M. Rima, “Bee venom: Overview of main compounds and bioactivities for therapeutic

interests," *Molecules*, vol. 24, no. 16, pp. 2997, 2019.

[4] G. Wang, W. Wang and C. Wu, "The new developments of cantharidin and its analogues," *Journal of the Chemical Society of Pakistan*, vol. 39, no. 4, pp. 599–609, 2017.

[5] H. Gao, C. Cui and L. Wang, "Mosquito microbiota and implications for disease control," *Trends in parasitology*, vol. 36, no. 2, pp. 98–111, 2020.

[6] P. Pyšek, P. E. Hulme and D. Simberloff, "Scientists' warning on invasive alien species," *Biological Reviews*, vol. 95, no. 6, pp. 1511–1534, 2020.

[7] H. Seebens, S. Bacher and T. M. Blackburn, "Projecting the continental accumulation of alien species through to 2050," *Global Change Biology*, vol. 27, no. 5, pp. 970–982, 2021.

[8] S. Dargan, M. Kumar and M. R. Ayyagari, "A survey of deep learning and its applications: A new paradigm to machine learning," *Archives of Computational Methods in Engineering*, vol. 27, no. 4, pp. 1071–1092, 2020.

[9] R. Girshick, J. Donahue and T. Darrell, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *CVPR*, pp. 580–587, 2014.

[10] R. Girshick, "Fast R-CNN," in *Proc. CVPR*, Boston, MA, pp. 1440–1448, 2015.

[11] S. Ren, K. He and R. Girshick, "Faster R-CNN: Towards real-time object detection with region proposal networks," *Advances in Neural Information Processing Systems*, vol. 28, pp. 91–99, 2015.

[12] Y. Shen, H. Zhou and J. Li, "Detection of stored-grain insects using deep learning," *Computers and Electronics in Agriculture*, vol. 145, pp. 319–325, 2018.

[13] D. Xia, P. Chen and B. Wang, "Insect detection and classification based on an improved convolutional neural network," *Sensors*, vol. 18, no. 12, pp. 4169, 2018.

[14] J. Redmon, S. Divvala and R. Girshick, "You only look once: Unified, real-time object detection," in *Proc. CVPR*, Las Vegas, NV, USA, pp. 779–788, 2016.

[15] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. CVPR*, Honolulu, HI, USA, pp. 7263–7271, 2017.

[16] C. B. Murthy, M. F. Hashmi, G. Muhammad and S. A. AlQahtani, "YOLOv2pd: An efficient pedestrian detection algorithm using improved YOLOv2 model," *Computers, Materials & Continua*, vol. 69, no. 3, pp. 3015–3031, 2021.

[17] Y. M. Wang, K. B. Jia and P. Y. Liu, "Impolite pedestrian detection by using enhanced YOLOv3-Tiny," *Journal on Artificial Intelligence*, vol. 2, no. 3, pp. 113–124, 2020.

[18] Q. Liu, S. Lu and L. Lan, "YOLOv3 attention face detector with high accuracy and efficiency," *Computer Systems Science and Engineering*, vol. 37, no. 2, pp. 283–295, 2021.

[19] M. Kasper-Eulaers, N. Hahn and S. Berger, "Detecting heavy goods vehicles in rest areas in winter conditions using YOLOv5," *Algorithms*, vol. 14, no. 4, pp. 114, 2021.

[20] A. H. Ashraf, M. Imran, A. M. Qahtani, A. Alsufyani, O. Almutiry *et al.,* "Weapons detection for security and video surveillance using cnn and YOLO-v5s," *Computers, Materials & Continua*, vol. 70, no. 2, pp. 2761–2775, 2022.

[21] H, Fukui, T. Hirakawa and T. Yamashita, "Attention branch network: Learning of attention mechanism for visual explanation," in *Proc. CVPR*, Long Beach, CA, USA, pp. 10705–10714, 2019.

[22] K. Prabhu, S. Sathish Kumar, M. Sivachitra, S. Dineshkumar and P. Sathiyabama, "Facial expression recognition using enhanced convolution neural network with attention mechanism," *Computer Systems Science and Engineering*, vol. 41, no. 1, pp. 415–426, 2022.

[23] G. Hou, J. Qin, X. Xiang, Y. Tan and N. N. Xiong, "AF-net: A medical image segmentation network based on attention mechanism and feature fusion," *Computers, Materials & Continua*, vol. 69, no. 2, pp. 1877–1891, 2021.

[24] M. Jaderberg, K. Simonyan and A. Zisserman, "Spatial transformer networks," *Advances in Neural Information Processing Systems*, vol. 28, pp. 2017–2025, 2015.

[25] J. Hu, L. Shen and G. Sun, "Squeeze-and-excitation networks," in *Proc. CVPR*, Salt Lake, UT, USA, pp. 7132–7141, 2018.

[26] S. Woo, J. Park and J. Y. Lee, "Cbam: Convolutional block attention module," in *Proc. ECCV*, Munich, GER, pp. 3–19, 2018.