Gesture Recognition Controls Image Style Transfer Based on Improved YOLOV5s Algorithm

Jiangfan Xie, Huilong Jin(✉), Tian Wen, and Ruiyan Du

Hebei Normal University, Shijiazhuang 050024, China
13131145063@163.com

Abstract. With the rapid development of artificial intelligence, human-computer interaction has drawn more researcher's attention. As one of the most important ways of human-computer interaction, Gesture recognition has been widely used in many fields. In this paper, an improved YOLOv5s gesture recognition algorithm is proposed, and the results of gesture recognition are used to carry out interactive experiments with the computer. Different gesture selects corresponding style, then the image style transfer network finishes the image style switch according to the image style. At the same time, PyQt5 is used to design an interactive interface to realize gesture recognition and image style conversion. Compared with YOLOv5s, the recall rate of gesture recognition by the improved algorithm is 94.77%, and the average accuracy is 96.46%, and the average accuracy of the improved YOLOv5s is 2.86% higher than YOLOv5s network, which is meeting the requirements of real-time and accuracy of image style transfer.

Keywords: Gesture recognition · YOLOv5 · Human computer interaction

1 Introduction

1.1 A Subsection Sample

In recent years, gesture recognition technology has been widely used and is a popular topic in computer vision field. Gesture recognition can be divided into contact type and non-contact type according to interaction mode.

Contact gesture recognition usually requires wearing a device with built-in sensors, and its recognition results have high accuracy and are not easily affected by external factors such as occlusion and illumination. Mina I. Sadek et al. divided Arabic Sign Language into multiple categories and used a few smart sensors to design gloves to judge gestures and obtain the meaning of Arabic sign language [1]. This method is simple in design and low in cost. Bin Fang et al. proposed a new type of data glove for capturing finger movements [2]. They carried out the finger motion capture experiment based on it, realizing the remote operation of manipulator through the acquired finger motion characteristics. This algorithm is easy to implement, and obtained more accurate and effective measurement results compared with existing methods. Yiyuan Zhang et al. reviewed the current research on using wearable sensors to identify the activities of the elderly in the bathroom, affirmed the important role of sensors in this research

H. Jin et al. (Eds.): CROWNCOM 2021/WiCON 2021, LNICST 427, pp. 203–212, 2022.
https://doi.org/10.1007/978-3-030-98002-3_15

[3], and advocated the combination with deep learning methods to achieve more accurate detection results.

Non-contact gesture recognition refers to gesture recognition method based on computer vision. This kind of method does not require the experimenter to wear any equipment, which has the characteristics of convenience, fast and easy to operate. Mohammed, A.A.Q et al. proposed an end-to-end method based on deep learning to detect and classify gestures [4]. In this method, the whole image is first extracted through the object detector for hand region, and then Convolutional Neural Networks (CNNs) carry out gesture recognition. The robustness and effectiveness of the proposed method are proved. Hua Li et al. proposed a gesture recognition system based on Leap Motion of the second generation [5], which could be used to recognize static gestures and dynamic gestures. The recognition rate of static gestures ranged from 94% to 100%, and that of dynamic gestures reached more than 90%. According to the experimental environment and the requirements of gesture recognition accuracy and speed, this paper chooses the non-contact gesture recognition method, and adopts YOLOv5s in the popular YOLO series algorithm to detect and classify gestures.

Style transfer is the transformation of general pictures into famous painting style, such as Van Gogh's Starry Night and Monet's Sunrise. Every painter has his own painting style. The purpose of style transfer is to transform the original picture into a specified painting style picture by computer, which can imitate the style. Meijun Sun et al. used MCCH feature selection model and Support Vector Machine (SVM) to describe Chinese painting styles and classify the works of different painters [6]. This paper chooses the algorithm proposed by Gatys et al. for image style transfer [7], which enables neural network model to learn different style images and generate corresponding style models. The model has a fast speed for new images, meeting the real-time requirements of this project.

2 Theoretical Foundation and Key Concepts

2.1 YOLOv5 Algorithms

Object detection algorithms based on deep learning can be roughly divided into two-stage algorithm and one-stage algorithm. In 2016, Redom et al. proposed one-stage object detector YOLO (You Only Look Once) [8] to solve the problem of inefficient two-stage object detection algorithm. In 2020, Jocher proposed YOLOv5, which has four network models: YOLOv5s, YOLOv5m, YOLOv5l and YOLOv5x. This paper selects the YOLOv5s model with the smallest network depth and the smallest feature map.

YOLOv5 has many advantages in hardware deployment, flexibility, and speed of the model. The overall structure is divided into four parts, including Input, Backbone, Neck, and Prediction. It includes two BottleneckCSP structures. The Cross-stage Partial structure (CSP) [9] is added into Residual Block [10], and BottleneckCSP1 contained Residual Block is mainly applied in Backbone. In the BottleneckCSP2 used in Neck part, the Residual Block is replaced with CBL structure, the structure of them is shown in Fig. 1.

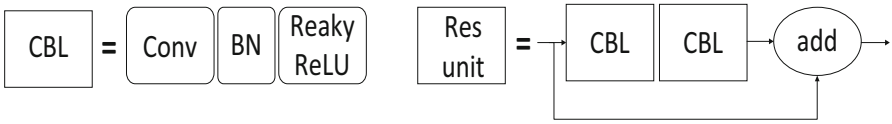

Fig. 1. The structure of CBL and Residual Block

The YOLOv5 network mainly uses the BottleneckCSP structure, as shown in Fig. 2. First, the classical residual structure Bottleneck operation is carried out, and the convolution results are added with the input through a 1×1 and 3×3 convolution operation. The other part is dimensionally reduced through 1×1 convolution, reducing the number of channels by half, and finally combining the two outputs. The number of Bottleneck structures in the LeneckCSP structure is the main point for residual learning and influences the improving of network performance.

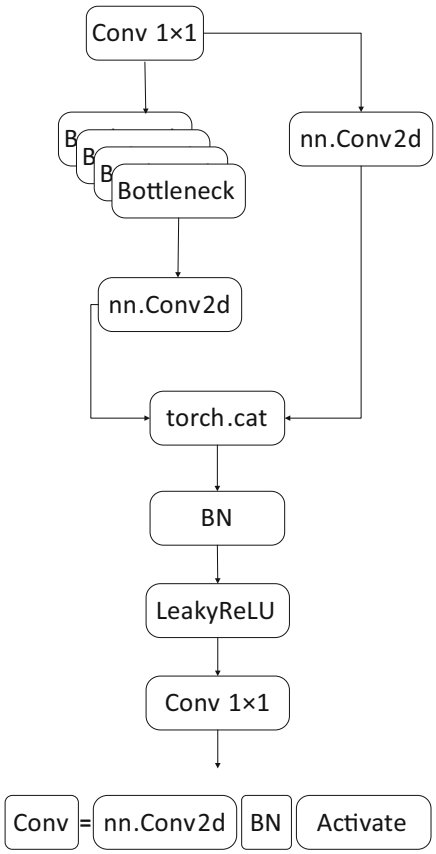

Fig. 2. The structure of BottleneckCSP

2.2 Image Style Transfer

In recent years, image style transfer is a hot research direction in the field of computer vision, the image style transfer keeps the image content and render its color and texture extracted from style image. The Fig. 3 shows the process.

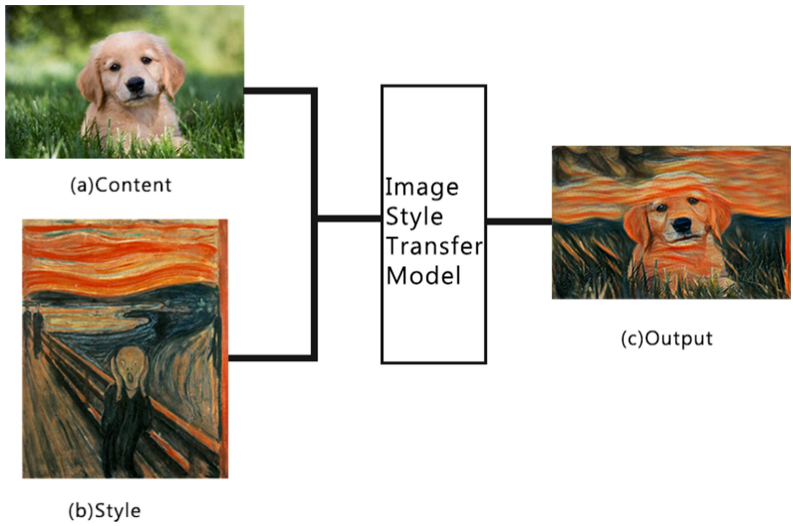

Fig. 3. The image style transfer model receives the content image and style image, then it generates a new image.

The traditional transfer technique uses the method of per-pixel compilation to transform the image, and its speed is slow, so the neural transfer style technique comes into being. Gatys et al. have developed a neural style transfer model based on Visual Geometry Group (VGG) network, which takes advantage of deep neural network and simultaneously extracts the underlying texture information and high-level semantic information of the image, and stylized images are generated by pixel iteration on noise images. Then the stylization method based on statistical parameters is used to match the style according to the global matching information. The method changes the pixel value through the Backpropagation at each pixel point of the image, which makes the composite image have a good visual effect.

3 Gesture Recognition Controls Image Style Transfer Based on Improved Yolov5s Algorithm

3.1 Project Introduction

PyQt5 used in this paper is a Qt application framework binding Python. Qt is a cross-platform framework for creating wonderful user interfaces and powerful native

applications, and it is one of the best choices for human-computer interface development. Qt Designer allows us to design static interfaces that meet our needs. Using the Qt UI plug-in, we can convert interface files into a python callable format.

The interface is divided into two parts. The left part is gesture control area, which has two buttons called “start” and “end”. The start and end button can open and close the camera, and control whether to display the shot picture. The right part is the image style transfer area, which has components to realize the selection and display of pictures in the local folder and the display of five style pictures controlled by five gestures. The project interface is shown in the figure below (Fig. 4).

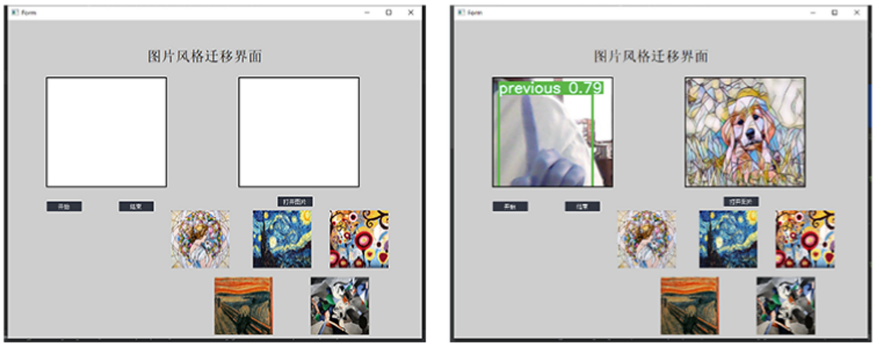

Fig. 4. Program interface

This project is based on YOLOv5s to achieve object detection of hand posture, and combines with the style transfer model, which can achieve real-time image style conversion. The two neural network models interact with each other through the interface designed by PyQt5. On the human-computer interaction interface, the human firstly shows gestures to the computer, which means giving instructions. The neural network model predicts gestures through each frame captured by the camera, and it obtains different instructions to complete the transformation of corresponding image styles. There are more and more applications in our daily life. For example, mobile phone camera can realize screen capture of mobile phone interface by recognizing hand gestures. In the field of smart home, it has been a trend to integrate multi-modal information such as visual information and sound information.

3.2 Improvements

The feature extraction network used by YOLOv5 algorithm is CSPDarknet. The network has a simple structure and low number of parameters.

With the rapid development of attention mechanisms, more and more studies have proved that channel attention mechanisms have great potential in improving the performance of CNNs. In order to balance the performance and complexity of the attention module, Qilong Wang et al. [11]. proposed the Effective Channel Attention (ECA) module. In order to further improve the efficiency and accuracy of gesture detection, we try to improve the feature extraction network structure of YOLOv5 and make it lightweight, and the attention mechanism is introduced to weigh the different channels of the feature map. The structure of the ECA is shown below (Fig. 5).

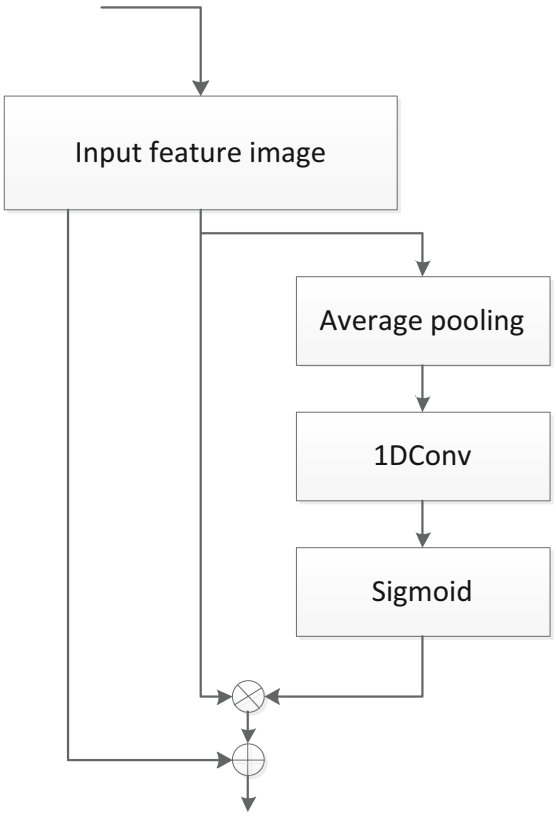

Fig. 5. The structure of ECA

As a result of the algorithm combined with the ECA layer design feature extraction network, the module has only a small increase in parameters, but has a significant performance gain. Adding the ECA structure into the Residual Block of the network can effectively fuse the information between different channels of the input feature graph, and improve the sensitivity of the algorithm to channel information. The structure of the algorithm in this paper is shown in the figure below (Fig. 6).

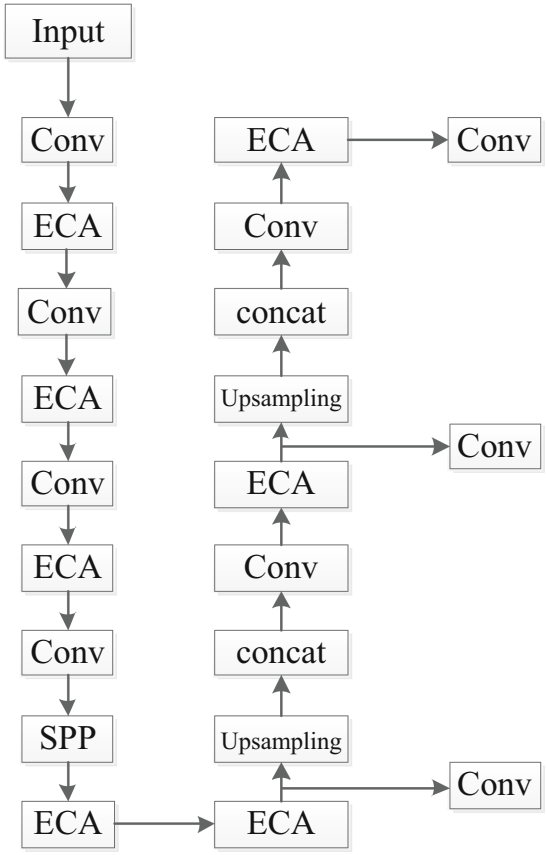

Fig. 6. The structure of the improved YOLOv5s

4 Experiments and Results

4.1 Dataset Description

The dataset used in this paper is a self-made dataset, including 1,597 training images and 400 test images. We use an application named LabelImg to label the dataset and label the minimum enclosing rectangle of the gesture in the image and save the annotations as XML. Then the training set and test set were randomly generated in the ratio of 4:1.

In order to enrich the data set and obtain better training results, the background is divided into two types when taking pictures, including pure color background and chaotic background. At the same time, the shooting distance is about 60 cm and the hand is in the center of the picture. Five gestures are selected as representatives, which represent the function of switching different styles in the picture style transfer.

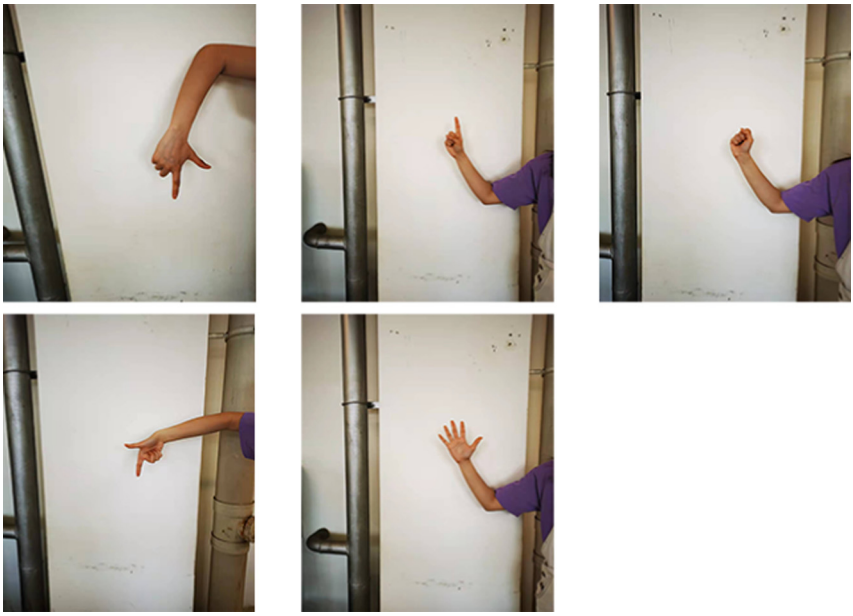

Fig. 7. The picture shows the five gestures.

4.2 Results

When training the network model, we set several hyperparameters. The iteration batch size is set as 16, the learning rate is 0.01, the attenuation coefficient is 0.005, and the total iteration number is 500 Epochs. Before training, images are processed to the same size 640 * 640. The training data set based on YOLOv5 model has a variety of image sizes, but due to the internal structure of the network, the image size is required to be a multiple of 32.

The experiment of this paper is carried out in windows10 system, equipped with Nvidia GeForce RTX 2080Ti graphics card, whose memory is 11 GB. We use Pytorch1.7 framework to set up the YOLOv5s model.

The gesture recognition accuracy based on the original algorithm has reached a high level. The recall rate of gesture recognition by the improved algorithm is 94.77%, and the average accuracy is 96.46%, which is 2.86% higher than the average accuracy of Yolov5s network. By comparing the mAP of the improved algorithm and the original algorithm in the training stage, we find that the convergence speed of the algorithm with ECA module is slightly faster than the original algorithm. The AP analysis of various gesture detection results shows that the improved algorithm has high detection accuracy for gesture.

Figure 7 shows gesture recognition using the improved YOLOv5s model. It can be seen from the figure if the color and shape of hands differ greatly from the background, the model will have a high accuracy in gesture prediction results. The recognition accuracy of the same gesture is slightly different under the background of pure color and cluttered background, the former is better than the latter (Fig. 8).

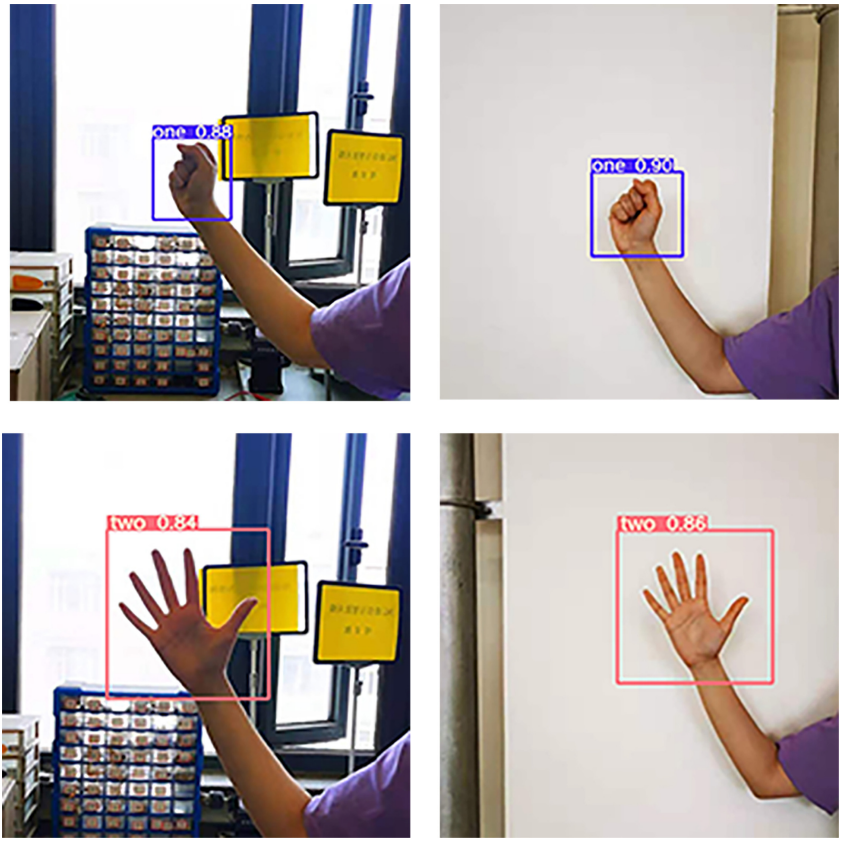

Fig. 8. The four images show the results of gesture recognition. The accuracy is more than 83%, regardless of whether the hand is on a solid or cluttered background.

5 Conclusion

This paper proposes an improved YOLOv5s gesture recognition network based on the existing YOLOv5 model and combined with practical applications, which can recognize hand posture more quickly and accurately. In this paper, the ECA module is added into the BottleNeckCSP, which improves the performance of the network architecture and the generalization capability of the model. Through the comparison experiment of different network models on self-made datasets, it shows that the MAP of the improved network has been improved to some extent, achieving the expected effect, laying a good foundation for the next image style transfer. Finally, in the static interface, the camera monitors gestures in real time to change the style of the imported image. The gesture recognition is accurate and fast, and the image style can be changed correctly, which proves the effectiveness of the algorithm.

In subsequent experiments, more abundant gesture datasets will be collected to verify the effectiveness of the proposed algorithm through further training. With experience of the combination of gesture recognition and style transfer, we will continue to explore more practical applications in the future, and integrate artificial intelligence more closely with life.

References

1. Sadek, M.I., Mikhael, M.N., Mansour, H.A.: A new approach for designing a smart glove for Arabic sign language recognition system based on the statistical analysis of the sign language. In: 2017 34th National Radio Science Conference (NRSC), pp. 380–388. IEEE (2017)
2. Fang, B., Sun, F., Liu, H., Guo, D.: A novel data glove for fingers motion capture using inertial and magnetic measurement units. In: 2016 IEEE International Conference on Robotics and Biomimetics (ROBIO), December 2016, pp. 2099–2104. IEEE (2016)
3. Zhang, Y., D'Haeseleer, I., Coelho, J., Vanden Abeele, V., Vanrumste, B.: Recognition of bathroom activities in older adults using wearable sensors: a systematic review and recommendations. Sensors **21**(6), 2176 (2021)
4. Mohammed, A.A., Lv, Q., Islam, M.D.: A deep learning-based End-to-End composite system for hand detection and gesture recognition. Sensors **19**(23), 52–82 (2019)
5. Li, H., Wu, L., Wang, H., Han, C., Quan, W., Zhao, J.: Hand gesture recognition enhancement based on spatial fuzzy matching in leap motion. IEEE Trans. Industr. Inf. **16** (3), 1885–1894 (2019)
6. Sun, M., Zhang, D., Wang, Z., Ren, J., Jin, J.S.: Monte Carlo convex hull model for classification of traditional Chinese paintings. Neurocomputing **171**, 788–797 (2016)
7. Gatys, L.A., Ecker, A.S., Bethge, M.: Image style transfer using convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2414–2423 (2016)
8. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 779–788 (2016)
9. Wang, C.Y., et al.: CSPNet: a new backbone that can enhance learning capability of CNN. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 390–391 (2020)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
11. Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., Hu, Q.: ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks. arXiv preprint: arXiv:1910.03151 (2020)
12. Redmon, J., Farhadi, A.: Yolov3: an incremental improvement. arXiv preprint: arXiv:1804.02767 (2018)