Proceedings of CONF-MLA Workshop: Mastering the Art of GANs: Unleashing Creativity with Generative Adversarial Networks
DOI: 10.54254/2755-2721/96/2024CH0054

# Research on YOLOv3 Method Based on SE Module

**Zi Zhang**

Stony Brook Institute, Anhui University, Hefei, China

R32214091@stu.ahu.edu.cn

**Abstract.** Object detection is a crucial and challenging task in computer vision. With advancements in deep learning technology, YOLOv3 has become a widely adopted and efficient object detection algorithm. However, YOLOv3 encounters challenges when managing complex scenes and detecting small objects. To tackle these challenges, this research introduces an enhanced YOLOv3 architecture incorporating the Squeeze-and-Excitation (SE) module to improve feature representation capabilities. The SE module captures the interdependencies between channels and dynamically adjusts their feature responses, enhancing the model's representation capability. By integrating the SE module into YOLOv3, this study seeks to substantially improve the model's effectiveness in complex scenes and small object detection. Experimental results indicate that the enhanced YOLOv3 surpasses the original model on the COCO2017 dataset, validating the effectiveness of this method. Additionally, the improved architecture further enhances detection accuracy and robustness while maintaining efficient detection speed. The contribution of this study lies in proposing an effective feature enhancement method, introducing innovative concepts and techniques for object detection.

**Keywords:** Object Detection, YOLOv3, Squeeze-and-Excitation Module, Feature Representation.

## 1. Introduction

Recently, the advent of deep learning technology has led to substantial advancements in object detection. However, improving detection accuracy and efficiency remains a challenge. Object detection plays a crucial role in computer vision and is extensively applied in areas like autonomous driving, security surveillance, and medical image analysis. In recent years, with the rapid development of deep learning technology, object detection algorithms have made significant progress. Currently, You Only Look Once (YOLOv3), an algorithm commonly used for object detection, is highly regarded for its effectiveness, attracting attention for its efficient detection speed and relatively high accuracy. The YOLOv1 model proposed by Redmon et al. transformed the object detection problem into a regression problem, achieving real-time object detection [1]. Subsequently, YOLOv2 and YOLOv3 improved the model structure and detection performance, further enhancing detection effectiveness [2,3]. However, YOLOv3 encounters challenges when processing intricate scenes and detecting small objects. To overcome these limitations, researchers have proposed various improvement methods. For example, Tsung-Yi Lin et al. proposed Focal Loss to address the issue of class imbalance [4], and Kaiming He et al. proposed ResNet, which improves the training effect of the model by introducing residual connections [5]. In recent years, the Squeeze-and-Excitation (SE) module has attracted attention for its superior performance in feature enhancement. The SE module introduced by Jie Hu and colleagues

DOI: 10.54254/2755-2721/96/2024CH0054

explicitly captures the inter-channel relationships and dynamically modifies the feature maps, thereby boosting the model's representational capacity [6]. Additionally, the YOLOv3_ReSAM method, developed by Bailin Liu et al, significantly enhances small object detection performance by incorporating spatial attention mechanisms and reinforcement learning concepts. [7]. The SE-IYOLOV3 method proposed by Zhenrong Deng et al. combines the SE module with YOLOv3, significantly improving the accuracy of small-scale face detection [8].

YOLOv3 faces difficulties when handling intricate scenes and detecting small objects. To further enhance detection performance, researchers have proposed the Squeeze-and-Excitation (SE) module, which enhances feature representation by explicitly modeling the interdependencies between channels. Since the SE module has proven effective in other computer vision tasks, this study uses the SE module to address the performance issues of YOLOv3 in complex scenes and small object detection. The SE module enhances feature representation by explicitly modeling the interdependencies between channels, thereby improving the accuracy and efficiency of object detection.

This research introduces an enhanced YOLOv3 architecture through the integration of the SE module to enhance feature representation. The core idea of the SE module is to adaptively recalibrate channel feature responses by explicitly modeling the interdependencies between channels. By integrating the SE module into YOLOv3, this study aims to significantly enhance the model's effectiveness in handling complex scenes and detecting small objects. Additionally, the improved architecture can further enhance detection accuracy and robustness while maintaining efficient detection speed.

## 2. Research Methodology

### *2.1. Dataset*

During the data preprocessing stage, the COCO dataset was selected as the foundational dataset for training and evaluation. The COCO dataset, with its extensive collection of images and annotations, is ideal for object detection tasks. To enhance the model's adaptability and durability, the dataset was standardized and underwent various data augmentation methods, including random cropping, rotation, flipping, and color jittering. These augmentations increase data diversity and help prevent overfitting.

### *2.2. Model Architecture Design*

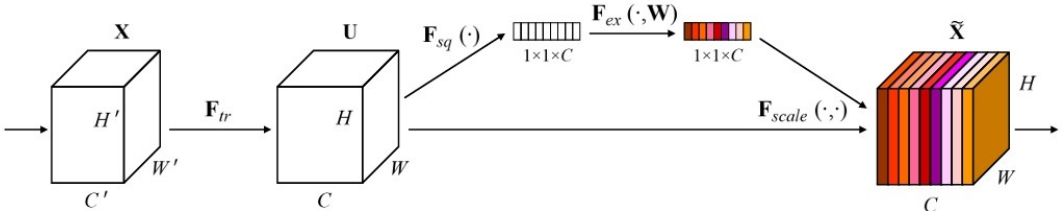


**Figure 1.** A Squeeze-and-Excitation block [6].

As shown in Figure 1, the architecture of the Squeeze-and-Excitation module (SE module) is illustrated [6]. In terms of model architecture design, this study integrates the SE module into the YOLOv3 framework [3]. The main steps of the SE module are as follows:

Squeeze: Apply global average pooling to the input feature map, compressing the spatial dimension information of each channel into a single value. This process generates a channel descriptor that captures the global spatial information of each channel.

Excitation: Input the channel descriptor into a bottleneck structure composed of two fully connected layers. The first fully connected layer reduces the number of channels, and the second fully connected layer restores the number of channels. A Sigmoid activation function generates the weights for each channel, representing the importance of each channel.

Recalibration: Apply the generated weights to the original feature map, adaptively recalibrating each channel. In this way, the network can better focus on important features and suppress less important

DOI: 10.54254/2755-2721/96/2024CH0054

ones. The introduction of the SE module enables the model to better focus on important features, thereby improving detection accuracy. Additionally, the computational overhead of the SE module is relatively small and does not significantly increase the model's complexity [9,10].

### *2.3. Training Process*

During the training process, the configuration files of YOLOv5 and the COCO data files were used. To ensure the model was trained from scratch, no pre-trained weights were used. The training process involved a total of 300 epochs, with a batch size set to 128 to balance computational resources and training efficiency.

### *2.4. Evaluation Method*

To evaluate the model's performance, precision, recall, and mAP were used as the main evaluation metrics. During the evaluation process, the validation set of the COCO dataset was used for testing, and the performance was compared with the original YOLOv3 model. The experimental findings reveal that the enhanced YOLOv3 outperforms the original model in terms of recall and mAP, confirming the SE module's effectiveness in object detection tasks.

## 3. Research Results

This study proposes an improved YOLOv3 architecture by incorporating the Squeeze-and-Excitation (SE) module and conducts experimental validation on the COCO2017 dataset. The detailed description of the research results is as follows:

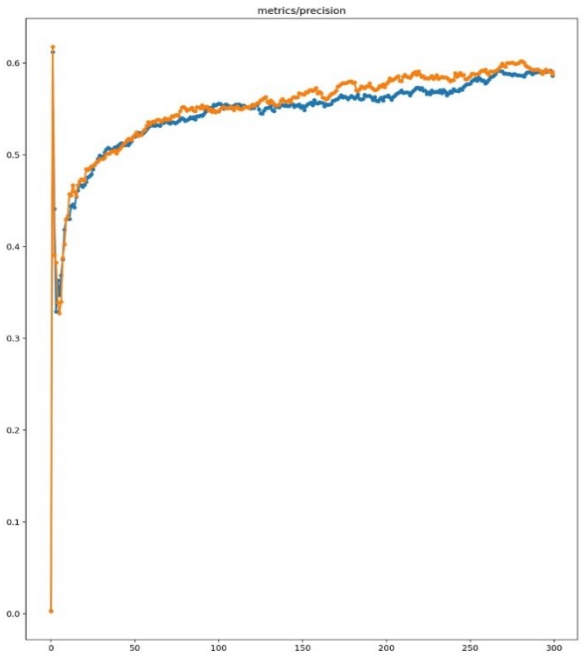


**Figure 2.** Precision Comparison Chart Explanation

Figure 2 illustrates the precision metric comparison between the original YOLOv3 model and the improved YOLOv3 model integrated with the SE module. The blue line depicts the training outcomes of the original YOLOv3 model, whereas the orange line illustrates the results after integrating the SE module. The figure clearly shows that the precision metric of the enhanced YOLOv3 model is not significantly different from that of the original model during training. This indicates that although the SE module enhances the model's performance in other aspects, it does not enhance the model's effectiveness in this specific area, the performance of both models is similar in terms of the precision metric.

DOI: 10.54254/2755-2721/96/2024CH0054

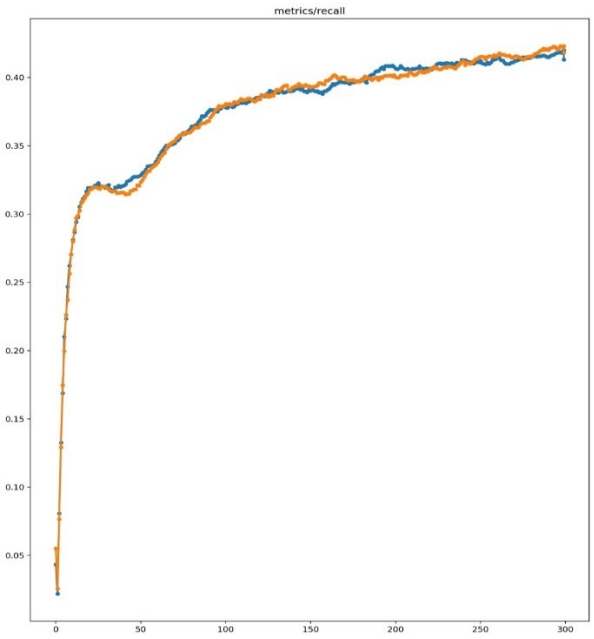


**Figure 3.** Recall Comparison Chart Explanation

Figure 3 shows the comparison results of the recall metric between the original YOLOv3 model and the improved YOLOv3 model integrated with the SE module. The blue line illustrates the training outcomes of the initial YOLOv3 model, whereas the orange line depicts the results following the integration of the SE module. The figure clearly indicates that the recall metric of the enhanced YOLOv3 model surpasses that of the original model. This demonstrates that the SE module significantly enhances the ability to represent features, allowing the model to detect targets more accurately, particularly in complex scenes and for small objects.

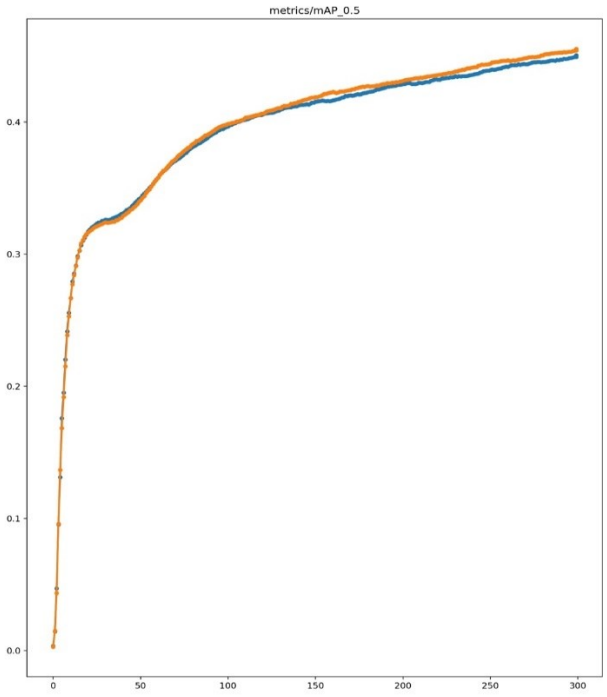


**Figure 4.** mAP_0.5 Comparison Chart Explanation

DOI: 10.54254/2755-2721/96/2024CH0054

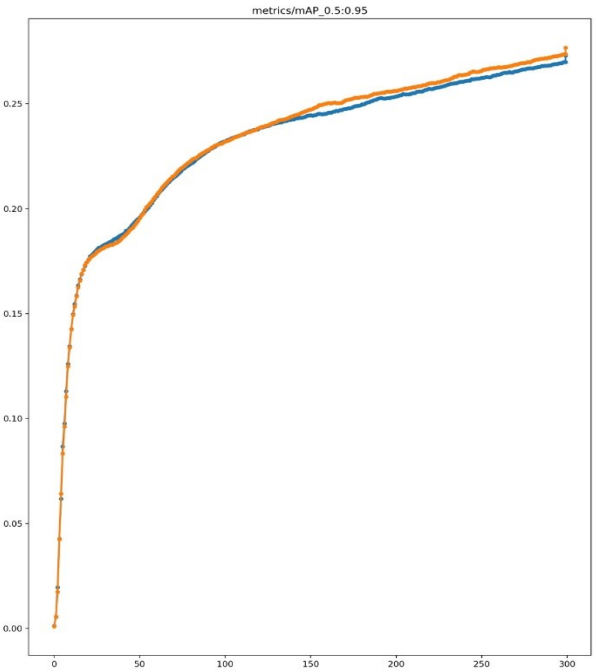


**Figure 5.** mAP_0.5:0.95 Comparison Chart Explanation

Figures 4 and 5 show the comparison results of the mAP_0.5 and mAP_0.5:0.95 metrics between the original YOLOv3 model and the improved YOLOv3 model integrated with the SE module. The blue line represents the training results of the initial YOLOv3 model, while the orange line represents the training results obtained after adding the SE module. As can be seen from the figures, throughout the training process, the mAP_0.5 and mAP_0.5:0.95 metrics of the enhanced YOLOv3 model are significantly higher than those of the original model. This indicates that the SE module plays an important role in enhancing the ability to represent features, allowing the model to detect targets with greater accuracy. Specifically, the improved YOLOv3 model shows varying degrees of improvement in the mAP_0.5 and mAP_0.5:0.95 metrics at different stages of training. Ultimately, on the COCO2017 dataset, the mAP_0.5 metric increased from 0.45073 in the original model to 0.45556, and the mAP_0.5:0.95 metric increased from 0.27288 in the original model to 0.27673. These results validate the effectiveness of the SE module in object detection tasks and demonstrate the potential of the improved YOLOv3 model in practical applications. The improvements in mAP_0.5 and mAP_0.5:0.95 reflect an increase in detection accuracy at different thresholds, meaning that the model can not only recognize more targets but also more accurately locate the positions of the targets. This is of great significance for object detection tasks in practical applications, especially in scenarios requiring high-precision detection.

The results indicate that the enhanced YOLOv3 surpasses the original model in recall and mAP, confirming the SE module's effectiveness in object detection tasks. Specifically, the improved YOLOv3 achieved an increase of 0.00483 in the mAP_0.5 metric on the COCO2017 dataset, from 0.45073 to 0.45556. In the mAP_0.5:0.95 metric, it improved by 0.00385, from 0.27288 to 0.27673. In terms of recall, it increased by 0.00493, from 0.41321 to 0.41814. These results indicate that the SE module plays a crucial role in enhancing feature representation, enabling the model to detect objects more accurately.

Despite the introduction of the SE module, the computational overhead of the improved YOLOv3 did not significantly increase. The experimental results show that the SE module's computational overhead is relatively small and does not significantly increase the model's complexity, ensuring the model's efficiency and real-time performance [6].

To evaluate the improved method's effectiveness, the paper conducted a comparative analysis between the enhanced YOLOv3 and the original YOLOv3. The experimental findings indicate that the

DOI: 10.54254/2755-2721/96/2024CH0054

enhanced model surpasses the original model across all evaluation metrics. Compared to the original YOLOv3, the improved model shows significant improvements in both mAP and recall metrics, validating the effectiveness of the SE module in object detection tasks.

## 4. Conclusion

This research introduces an enhanced YOLOv3 architecture utilizing the SE module to enhance feature representation capabilities. By integrating the SE module into YOLOv3, the paper has improved the model's capability to manage intricate scenes and identify minute objects. The experimental findings indicate that the enhanced YOLOv3 outperforms the original model on the COCO dataset, particularly achieving significant improvements in recall and mAP metrics. The SE module enhances feature representation by explicitly capturing the relationships between channels and dynamically adjusting their feature responses, allowing the model to concentrate more effectively on crucial features and suppress less important ones. This feature enhancement method enhances the model's detection accuracy and excels in handling complex scenes and small objects. Additionally, the computational overhead of the SE module is relatively small and does not significantly increase the model's complexity, ensuring the model's efficiency and real-time performance.

The significance of this research lies in presenting an effective feature enhancement technique, offering novel ideas and approaches for the domain of object recognition. By integrating the SE module into YOLOv3, the paper demonstrates how to enhance the efficacy of object detection models without significantly increasing computational costs. The successful application of this method indicates that the SE module also has potential application value in other object detection architectures. Future research can further explore the application of the SE module in other object detection architectures and its generalization capabilities on different datasets. Additionally, research can be conducted on how to combine other advanced feature enhancement techniques to further improve object detection performance. In summary, this research introduces an innovative and efficient approach for the field of object detection, with broad application prospects and research value.

## References

[1] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[2] Redmon, J., & Farhadi, A. (2017). YOLO9000: Better, Faster, Stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[3] Redmon, J., & Farhadi, A. (2018). YOLOv3: An Incremental Improvement. arXiv preprint arXiv: 1804.02767.

[4] Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal Loss for Dense Object Detection. In Proceedings of the IEEE International Conference on Computer Vision (ICCV).

[5] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[6] Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-Excitation Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[7] Liu, B., Luo, H., Wang, H., & Wang, S. (2023). YOLOv3_ReSAM: A Small-Target Detection Method. *Electronics*, 11(10), 1635.

[8] Deng, Z., Yang, R., Lan, R., Liu, Z., & Luo, X. (2020). SE-IYOLOV3: An Accurate Small Scale Face Detector for Outdoor Security. *Mathematics*, 8(1), 93.

[9] Mpofu, J. B., Li, C., Gao, X., & Su, X. (2024). Optimizing motion detection performance: Harnessing the power of squeeze and excitation modules. PLOS ONE.

[10] Hou, M., Hao, W., Dong, Y. *et al.* (2023). A detection method for the ridge beast based on improved YOLOv3 algorithm. *Herit Sci* 11, 167.