Joint C2f and Joint Loss Object Detection Based on YOLOv5

Jinglin Cao[a]
School of Material Science and Engineering, Tongji University, Shanghai, China

Keywords: Object Detection, Complete Union Intersection, YOLOv5, Distributed Focus Loss.

Abstract: Seeking more accurate object detection in densely populated scenes, especially those involving small objects, is crucial for the development of computer vision applications. This study aims to significantly improve the detection capability of the YOLOv5 architecture. Specifically, this article proposes a new combination of C2f modules for enriching feature learning and distributed focus loss with a Complete Union Intersection (CIoU) loss function for improving object localization and class imbalance handling. Specifically, the C2f module helps to achieve better gradient propagation within the network, while the Distributed Focus Loss (DFL)+CIoU loss function improves detection accuracy through advanced boundary box calculations. This study was conducted on the COCO128 dataset. Its rigorous experimental framework confirms that these enhancements significantly improve average accuracy, recall, and bounding box accuracy. Experimental results indicate that the modified YOLOv5 model outperforms the baseline, offering significant improvements in detecting small-scale objects amidst complex backgrounds. The implications of this study are far-reaching, providing a foundation for developing real-time detection systems that are more reliable and effective across varied and challenging visual environments.

1 INTRODUCTION

Object detection, as an integral discipline of computer vision, has been revolutionized with the introduction of deep learning technologies. This field focuses on the identification and localization of objects within images, serving as the foundation for numerous applications such as autonomous driving, security surveillance, and image analysis systems. The significance of object detection lies not only in its ability to enhance the interpretability of visual data but also in its contribution to the advancement of automated systems that require visual comprehension. Recent surveys and reviews in this domain underscore the rapid evolution of methodologies, particularly highlighting the shift from traditional approaches to deep learning-based models such as Region-based Convolutional Neural Networks (R-CNN), Single Shot Detector (SSD), and You Only Look Once (YOLO), each presenting unique solutions to the challenges of accuracy, speed, and scalability in object detection tasks (Girshick, 2014; Liu, 2016; Redmon, 2016).

Initially, the R-CNN approach introduced the concept of region proposals for object detection, enhancing accuracy but at the cost of computational efficiency. To address the limitations of R-CNN, such as speed and scalability, the SSD was developed, providing a balance between speed and accuracy by eliminating the need for explicit region proposal steps and detecting objects in a single pass. Among these developments, the YOLO framework marked a significant departure from traditional object detection methods by integrating the detection process into a single neural network. This integration not only improved detection speeds but also maintained competitive accuracy, making real-time object detection feasible. The subsequent iterations of YOLO, namely YOLOv2 and YOLOv3, introduced incremental improvements aimed at refining accuracy, speed, and the model's ability to detect objects of varying sizes (Redmon, 2017; Redmon, 2018). YOLOv2 focused on enhancing the model's recall and Intersection over Union (IoU) metrics through techniques such as batch normalization and high-resolution classifiers. YOLOv3 further improved upon these enhancements by incorporating

[a] https://orcid.org/0009-0003-6649-6731

Cao, J.
Joint C2f and Joint Loss Object Detection Based on YOLOv5.
DOI: 10.5220/0012950800004508

In *Proceedings of the 1st International Conference on Engineering Management, Information Technology and Intelligence (EMITI 2024)*, pages 452-457
ISBN: 978-989-758-713-9

multi-scale predictions and a deeper, more complex network architecture, significantly boosting the model's performance across a wide range of object sizes and types.

The main objective of this study is to enhance the YOLO framework and enhance the detection accuracy of small objects by introducing new architectural elements and optimization techniques. Specifically, this study aims to enhance the feature extraction capability of YOLO by integrating more gradient-rich modules and modifying its backbone architecture. In addition, this study combines the Distributed Focus Loss (DFL) with the existing Complete Union Intersection (CIoU) loss function (Lin, 2017; Zheng, 2021). This method solves the challenge of precise object detection in complex visual scenes and helps achieve a broader goal of improving the reliability and efficiency of automated systems that rely on visual data. The experimental results demonstrate the effectiveness of these modifications, revealing the optimal model configuration and its impact on detection performance. In summary, this study not only advances the technological foundation of object detection models but also emphasizes their practical significance in real-world applications and the necessity for continuous innovation in this field.

2 METHODOLOGIES

2.1 Dataset Description and Preprocessing

In this research, the COCO128 dataset, a condensed subset of the extensive Common Objects in Context (COCO) dataset, was employed. Developed by Microsoft Research, the COCO dataset is a significant resource for object detection, segmentation, and captioning, designed to enhance scene understanding through diverse and annotated imagery (Lin, 2014; Misra, 2019). The COCO128 variant comprises 128 selectively chosen images that span all 80 categories featured in the original COCO dataset, allowing for rapid prototyping and algorithm validation. The modification leveraged YOLOv5's data preprocessing techniques, notably including image resizing to a dimension of 608x608 pixels, normalization, and the implementation of Mosaic data augmentation. These preprocessing measures are aimed at optimizing the model's performance, ensuring enhanced detection accuracy across varying object sizes and lighting conditions. The utilization of YOLOv5's adaptive anchor box calculations and image scaling methods further augmented the model's efficacy.

2.2 Proposed Approach

This study aims to reimplement and enhance the YOLOv5 object detection algorithm by introducing advanced architecture and functional improvements. The proposed method revolves around the YOLOv5 algorithm architecture. This article revolves around the backbone and output loss function. This method replaces the C3 module with a C2f (Sun, 2021) module rich in gradient flow, optimizes the gradient length to improve accuracy and efficiency, and introduces a combination of DFL and CIoU for the regression loss branch. This method ensures richer gradient flow and more accurate bounding box coordinate representation, aiming to achieve higher object detection performance. The algorithm design includes four general modules: input, backbone network, neck network, and output, each of which is customized to enhance the detection ability of the model. The pipeline is shown in Figure 1. Through meticulous model training and validation processes, including Mosaic data augmentation and adaptive anchor box calculations, the proposed approach seeks to optimize performance across various metrics, such as mean average precision (mAP), box loss, and F1 score, on the COCO128 dataset over 1000 epochs. The results underscore the benefits of the architectural modifications and training strategies, indicating the superior performance of the enhanced model compared to the baseline, especially in scenarios involving complex backgrounds or smaller objects.

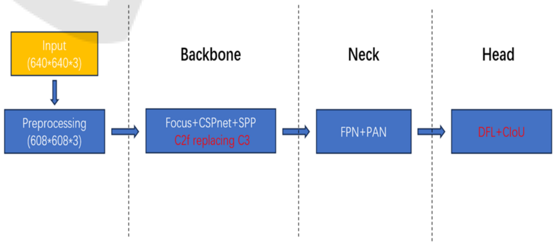

Figure 1: The pipeline of this study (Photo/Picture credit: Original).

2.2.1 Baseline

The YOLOv5 model integrates several advanced architectural techniques to create a robust framework for object detection. It employs a unique input pre-processing step called 'Focus' to reduce spatial dimensions while concatenating feature maps, preparing the input for subsequent layers. The model backbone consists of Cross Stage Partial (CSP) networks, which significantly reduce the computational cost while maintaining feature richness by merging partial features from different stages (Wang, 2020). At the very outset of the backbone, the YOLOv5 employs a distinctive “Focus” mechanism. This novel slicing operation works by splitting the input image into four parts, thereby reducing the spatial dimensions by half while increasing the depth dimension fourfold. Furthermore, the use of the Spatial Pyramid Pooling (SPP) block enhances the receptive field and assists in maintaining spatial hierarchies between features at different scales (He, 2015).

Following the backbone, the YOLOv5 model leverages a combination of Feature Pyramid Network (FPN) and Path Aggregation Network (PAN) structures within its neck (Lin, 2017; Lin, 2018). The FPN component aggregates features at different scales, improving the model's ability to detect objects of various sizes. Simultaneously, the PAN part enhances the feature hierarchy by allowing lower-level features to be upsampled and integrated with higher-level ones. The model's head is equipped with convolutional layers that produce final detection predictions, including bounding box coordinates and class probabilities. YOLOv5 uses the “Convolution, Batch Normalization and Leaky ReLU” (CBL) structure throughout its layers, promoting non-linear feature transformation and normalization, which stabilizes training and accelerates convergence. Additionally, the model employs skip connections and upsampling, which ensures that the granularity of the spatial information is preserved throughout the network.

In this study’s framework, the YOLOv5 model serves as the foundation for comparing various improvements. The model is set up using pre-defined hyperparameters and trained on a standard dataset to establish a baseline for object detection performance. This baseline model is pivotal for subsequent experimental iterations where modifications are incrementally introduced to assess their impact on the model's efficiency and accuracy.

2.2.2 C2f Module

The C2f module, central to the enhanced YOLOv5 architecture, is an innovative adaptation of the conventional C3 module, integrated with elements inspired by the ELAN module (Zhang, 2022). This design choice aims to optimize the gradient flow within the network, ensuring that each computational block benefits from richer gradient information. The principal advantage of the C2f module lies in its ability to balance lightweight architecture with high accuracy, an essential trait for real-time object detection tasks. Unlike the more straightforward C3 module, the C2f module is engineered to facilitate a more efficient gradient flow throughout the network. This efficiency is achieved by strategically stacking computational blocks to minimize the length of gradient paths. Such an architecture not only enhances learning efficiency but also reduces latency, making it particularly suited for applications requiring quick and accurate object detection. The C2f module incorporates a dual-focus approach: it retains the compact and efficient nature of its predecessor while significantly amplifying the network's sensitivity to gradient information. This is achieved through a novel stacking method that carefully orchestrates the computational blocks, ensuring optimal gradient flow and minimal information loss.

In the context of this study, the C2f module replaces the traditional C3 component in the backbone network. This replacement not only lightens the model's computational load but also improves the accuracy of object detection. By doing so, it addresses the perennial challenge of balancing speed and performance in object detection models.

2.2.3 Loss Function: DFL+CIoU

The integration of DFL with CIoU Loss marks a significant advancement in the model's ability to precisely localize and classify objects within an image. This hybrid loss function is meticulously designed to optimize the bounding box regression and classification tasks simultaneously, thereby enhancing the overall precision of the object detection process, as:

$$DFL(p_t) = -\alpha_t (1-p_t)^{\gamma \log(p_t)} \tag{1}$$

$$CIoU = 1 - IoU + \frac{\rho^2(b,\hat{b})}{c^2} + \alpha * v \tag{2}$$

DFL modifies the standard focal loss by applying a cross-entropy optimization to the two positions closest to the target label, both to the left and right. This optimization allows the network to concentrate on the distribution surrounding the target area more effectively, bringing the output distribution closer to actual floating-point coordinates. The CIoU loss component further refines this process by taking into account the complete overlap between predicted and ground-truth boxes, thus improving the spatial accuracy of the predictions. In this research, the DFL+CIoU loss function is applied at the output stage of the model, specifically tailored to enhance the precision of bounding box predictions. This approach not only addresses the limitations of conventional loss functions in handling complex object relations and varying scales but also significantly improves the model's ability to distinguish between closely situated objects, a common challenge in dense object detection scenarios. By incorporating the DFL+CIoU loss function, the study aims to establish a more effective training regimen, optimizing both the localization and classification aspects of the detection task. This enhancement is crucial for achieving high accuracy in real-world applications where the precision of object detection can significantly impact performance and outcomes.

2.3 Implementation Details

The enhanced system, utilizing Python 3.8.0 and the PyTorch framework, operates on an NVIDIA Tesla T4 GPU, ensuring efficiency in processing and model training. Advanced data augmentation techniques, including Mosaic augmentation alongside standard practices such as flipping and scaling, significantly enhance the training dataset's diversity. This approach improves the model's robustness and accuracy across various object detection scenarios. The experiment environment is finely tuned with selected hyperparameters to optimize the training process: a learning rate of 0.01 facilitates a balanced convergence speed, while momentum at 0.937 helps navigate local minima. The image size is set to 640x640 pixels, optimizing detail preservation and computational load. A batch size of 16 ensures efficient data processing without compromising system responsiveness. These configurations contribute to a balanced framework, aiming for superior performance in the enhanced YOLOv5 object detection model.

3 RESULTS AND DISCUSSION

The third chapter delves into the results derived from implementing strategic architectural improvements and a novel loss function within the YOLOv5 framework. It ties the enhancements directly to the methodologies described in the Methods section, emphasizing the incorporation of the C2f module and the DFL function.

3.1 mAP Analysis

Figure 2. demonstrates the mAP scores for both the enhanced model (blue line) and the original YOLOv5 (green line) across epochs. The C2f module's inclusion, which optimizes gradient flow and enriches gradient information for computational blocks, has contributed to a steady increase in the mAP scores of the modified model. This architectural modification balances the need for a lightweight structure with accuracy, resulting in improved precision, particularly in real-time detection tasks.

3.2 Recall Analysis

In Figure 3, the enhanced model's recall rate improves and stabilizes above the original YOLOv5. This can be accredited to the C2f module's efficiency in facilitating gradient flow, which in turn influences the sensitivity of the model to detect true positives, thereby enhancing recall rates.

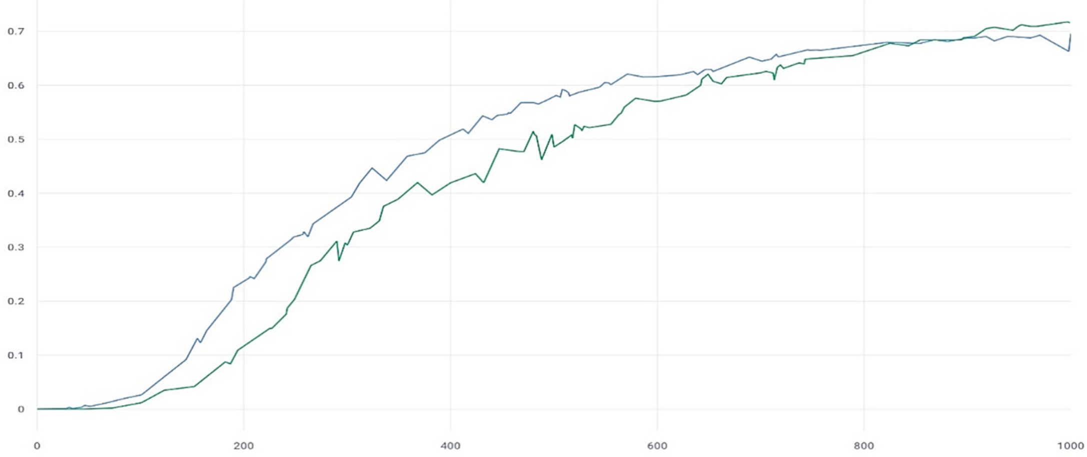

Figure 2: mAP progression over epochs, showcasing the impact of the C2f module (Photo/Picture credit: Original).

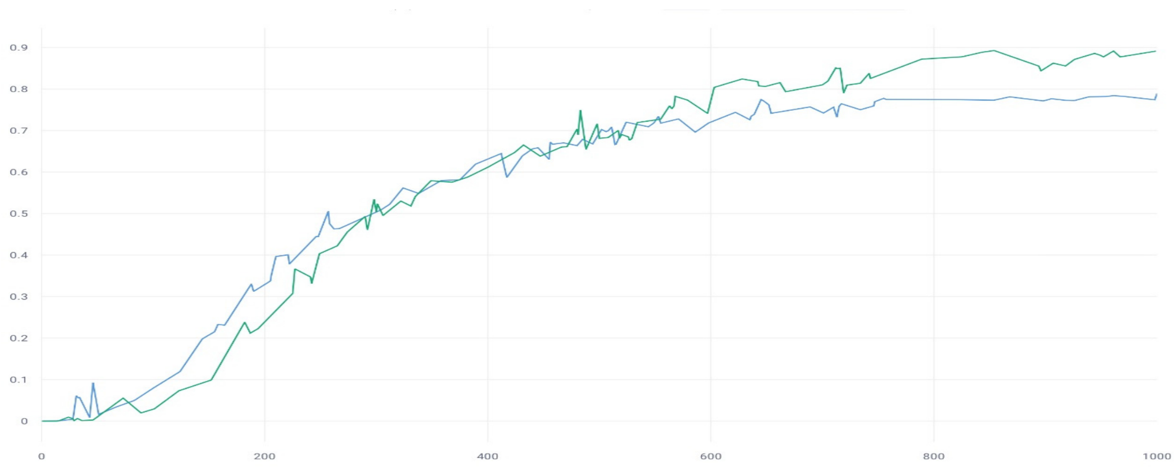

Figure 3: Recall rate comparison over epochs between the enhanced and original YOLOv5 (Photo/Picture credit: Original).

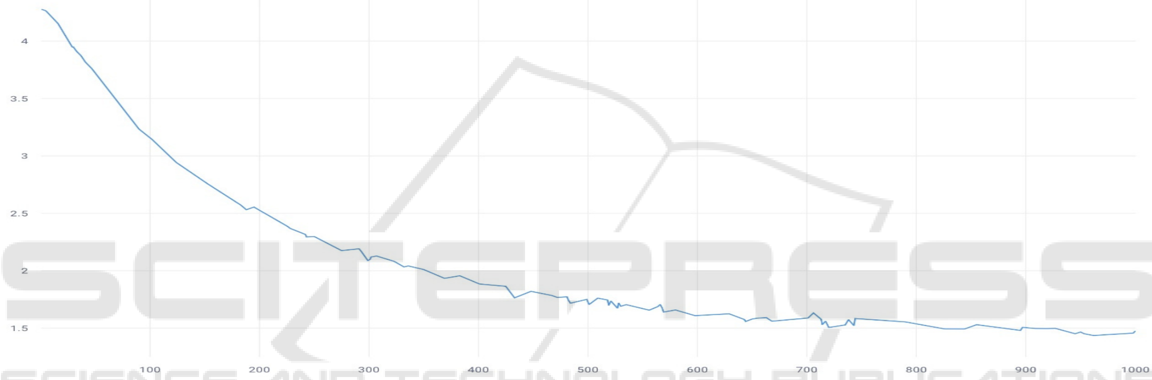

Figure 4: Training loss over epochs with DFL, underscoring localization and classification improvements (Photo/Picture credit: Original).

3.3 Loss Analysis

Figure 4. depiction of the loss trend displays the significant impact of the DFL function, showing a more pronounced decrease in loss values over epochs for the modified model. The DFL, which adjusts the focal loss to concentrate on the distribution surrounding the target area, coupled with the CIoU component, has optimized the model's localization and classification capabilities. These adjustments address complex object relations and scale variations, reducing the model's loss and enhancing its detection accuracy, especially for densely packed objects.

3.4 F1-Confidence Analysis

Figure 5. incorporates the F1-Confidence curve analysis, highlighting the optimal confidence threshold identified through the study. The curve peaks at an F1 score of 0.81 at a confidence level of 0.621, suggesting the modified model's precise balance between precision and recall. This peak performance is a testament to the methods employed, particularly the DFL and CIoU loss function which has been pivotal in enhancing bounding box accuracy.

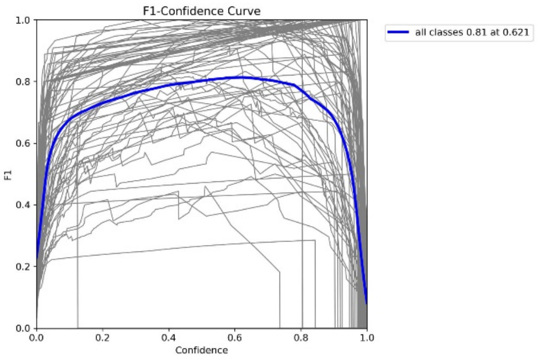

Figure 5: F1-Confidence Curve illustrating the model's precision-recall balance (Photo/Picture credit: Original).

In essence, the experimental findings presented in this chapter are a direct consequence of the C2f module and DFL function integrated into the YOLOv5 architecture. The results not only validate the proposed methods but also highlight the significance of these enhancements in achieving high accuracy in complex object detection scenarios, a crucial factor in the applicability of such models in real-world settings.

4 CONCLUSIONS

This study enhances object detection by innovatively modifying the YOLOv5 architecture. This article introduces the C2f module for improving gradient flow and feature learning and integrates the distribution focus loss with the CIoU loss function for fine localization and classification. This fusion solves the inherent class imbalance and complex spatial relationships in dense object scenes. Additionally, this article rigorously evaluates the enhanced model using the COCO128 dataset, demonstrating substantial improvements compared to the original YOLOv5. Key indicators such as mAP, recall, and bounding box prediction accuracy demonstrate enhanced effectiveness. Emphasis was placed on the role of C2f modules in ensuring lightweight yet accurate architecture, as well as the contribution of DFL+CIoU loss functions in addressing changes in scale and complex object relationships. Future work will aim to scale the model for broader real-world applications, focusing on enhancing its robustness across varying object sizes and environmental conditions. The objective is to optimize performance through advanced data augmentation and expanded dataset diversity, ensuring the model's adaptability and effectiveness in real-world deployments. This research lays the groundwork for future advancements in object detection, paving the way for more nuanced and robust detection models.

REFERENCES

Girshick, R., Donahue, J., Darrell, T., & Malik, J., 2014. Rich feature hierarchies for accurate object detection and semantic segmentation, CVPR, pp. 580–587.

He, K., Zhang, X., Ren, S., & Sun, J. 2015. Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE transactions on pattern analysis and machine intelligence, vol, 37(9), pp: 1904-1916.

Lin, T. Y., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. 2017. Feature pyramid networks for object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition pp. 2117-2125.

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. pp. 2980-2988.

Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., & Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In Computer Vision–ECCV 2014: 13th European Conference, pp. 740-755.

Liu, S., Qi, L., Qin, H., Shi, J., & Jia, J. 2018. Path aggregation network for instance segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 8759-8768.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, Y.C., & Berg, A.C., 2016. Ssd: Single shot multibox detector. ECCV. pp. 21–37.

Misra, D., 2019. Mish: A self regularized non-monotonic activation function. arXiv:1908.08681.

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A., 2016. You only look once: Unified, real-time object detection. CVPR, pp. 779–788.

Redmon, J., & Farhadi, A., 2017. YOLO9000: better, faster, stronger. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7263-7271.

Redmon, J., & Farhadi, A., 2018. Yolov3: An incremental improvement. arXiv:1804.02767.

Sun, Y., Chen, G., Zhou, T., Zhang, Y., & Liu, N. 2021. Context-aware cross-level fusion network for camouflaged object detection. arXiv:2105.12555.

Wang, C. Y., Liao, H. Y. M., Wu, Y. H., Chen, P. Y., Hsieh, J. W., & Yeh, I. H. 2020. 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.

Zhang, X., Zeng, H., Guo, S., & Zhang, L. 2022. Efficient long-range attention network for image super-resolution. In European conference on computer vision, pp. 649-667.

Zheng, Z., Wang, P., Ren, D., Liu, W., Ye, R., Hu, Q., & Zuo, W., 2021. Enhancing geometric factors in model learning and inference for object detection and instance segmentation. IEEE transactions on cybernetics, vol. 52(8), pp: 8574-8586.