International Core Journal of Engineering Volume 9 Issue 11, 2023
ISSN: 2414-1895 DOI: 10.6919/ICJE.202311_9(11).0006

Object Detection Method based on YOLOv5

Hanpeng Ren[1], Lei Dong[1,*], Ruixin Gao[1], Yangang Jin[2], Beiping Zhao[2], Jing Wang[2], Xiaojing Wang[2], and Yuefei Zheng[1]

[1] School of Mechanical Engineering, Tianjin University of Technology and Education, Tianjin 300222, China

[2] Citylong Technology Development Co., Ltd, Tianjin 300300, China

Abstract

In order to automate industrial tasks such as sorting and assembling industrial components, this study utilizes the YOLOv5s model for the identification and localization of industrial parts. We trained our model using a customized dataset of parts and utilized the trained model to extract both the features and positional information of various types of parts, thereby enabling the classification and detection of complex irregular-shaped parts. According to the experimental results, the YOLOv5s-based component classification model was tested and recognized in real-world application scenarios. The model effectively and accurately identifies five distinct components, showcasing its ability, with a detection accuracy of 92.1%. The model also achieves a target detection speed of 110.96 frames per second. The part classification detection model utilizing the YOLOv5s network demonstrates high detection accuracy, excellent robustness, and fast computational speed under varying conditions such as different lighting and viewing angles. Moreover, this approach can be widely applied in various industrial production contexts, providing technical support for the realization of subsequent industrial intelligent production.

Keywords

Object Detection; Machine Learning; YOLOv5.

1. Introduction

Accurately selecting parts is a critical requirement to guarantee precise assembly of industrial components during the process of industrial production and processing. Therefore, strict component inspection is of paramount importance in manufacturing [1]. In the early stages of industrial development, tasks such as part recognition and industrial assembly heavily relied on manual labor, which depended on the skills and experience of assembly workers. Nevertheless, with the rapid development of machine learning, template matching technology has found widespread applications in the industry. Nonetheless, template matching methods that emphasize both the target under test's individual features and their combination impose stricter size and angle requirements on the measured object [2]. Additionally, template matching methods employ manual processing to extract part characteristics by extracting features such as corners, edges, and shapes from images. This approach heavily depends on specific detection environments for identifying structurally complex and visually similar objects. Additionally, the recognition results are contingent upon the designers' experience and application scenarios, making it lack universality [3].

Over the past few years, there has been significant development in target detection algorithms based on convolutional neural networks. These algorithms have demonstrated superior performance and advantages over traditional machine vision techniques, particularly in the areas of detection and

ISSN: 2414-1895 DOI: 10.6919/ICJE.202311_9(11).0006

recognition [4]. Mingtao and Gao Weiwei proposed an enhanced architecture called Feature Pyramid Network and a Region of Interest Align method, which have been successfully incorporated into the Faster R-CNN algorithm for precise detection of brake parts [5]. These improvements have resulted in increased accuracy and robustness of target detection algorithms, particularly in complex backgrounds and scenarios with multiple objects. As a result, they have provided substantial support for automated detection tasks in fields like industrial manufacturing. Qian Yining addressed the application issues of Mask R-CNN and successfully achieved the detection of tiny parts in four types of watches by adjusting the number of anchors and ROIs in a single training image, as well as by considering the feature layer selection rules of FPN along with the prior information [6]. Additionally, Song Shuan Jun, Hou Zhong Yuan, and other researchers made enhancements to YOLOv3. They re-clustered the anchor boxes for custom datasets and incorporated a feature scale to fuse structural information of parts, resulting in a 4.87% increase in accuracy. Zhu Yuanyuan, Peng Lulu, and their colleagues proposed the SE-R-YOLOv4 algorithm, which primarily focuses on channels containing rich information to improve accuracy [6]. Their experiments on a dataset of circular automotive steel parts surface defects demonstrated that the SE-R-YOLOv4 algorithm achieves an accuracy rate of up to 90.5% while processing at a frame rate of 53.1 frames per second. These results indicate that SE-R-YOLOv4 exhibits high accuracy and efficiency in detecting surface defects on circular automotive steel parts, enabling fast and accurate identification and localization of defects. Consequently, it provides effective support for quality control and production processes [8].

Among existing object detection methods, the Faster R-CNN series is known for its accurate recognition and high precision. However, the adoption of the two-stage method in Faster R-CNN results in slower detection speed and a significant computational cost, rendering it unsuitable for real-time monitoring. The two-stage method typically entails generating candidate boxes in the initial stage and subsequently classifying and locating them, thereby increasing computational complexity and time consumption. Hence, in real-time monitoring applications, Faster R-CNN may not be the optimal choice. Single-stage object detection algorithms, represented by the YOLO series, offer a substantial advantage in terms of detection speed while ensuring accuracy, rendering them suitable for practical object detection methods [9]. The YOLO, You Only Look Once: Unified Real-Time Object Detection, algorithm, proposed by scholars such as Joseph Redmon and Ali Farhadi, is a single neural network-based object detection system. YOLO, is similar to the SSD algorithm, is an end-to-end model. In 2017, the authors of YOLO introduced YOLOv2 to enhance prediction accuracy, speed, and expand the number of recognized object categories. YOLOv3 utilizes a deeper network called Darknet53, and YOLOv4 further extends the original Darknet53 backbone network introduced in YOLOv3 [10]. The incorporation of the CSP module in Darknet53 enhances the network's learning capability, accelerates learning speed, and reduces computational and memory consumption. Furthermore, YOLOv4 integrates the SPP mechanism and FPN structure, with the SPP structure partially resolving the challenge of multi-scale targets [11].

2. The YOLOv5s Algorithm Model

YOLOv5s is the fifth-generation object detection algorithm within the YOLO algorithm family. YOLOv5 offers various versions by partitioning the network into submodules of varying depths and widths, resulting in YOLOv5n, YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x. Additionally, official variants like n6, s6, m6, l6, and x6 have been introduced to cater to higher-resolution images, e.g., 1280x1280. These variants also exhibit disparities in their network architectures. The latter undergoes 64x down sampling and utilizes 4 prediction feature layers, whereas the former undergoes 32x down sampling and utilizes 3 prediction feature layers [12]. The design of these distinct variants aims to address target detection tasks across various scenarios and requirements, ultimately delivering enhanced performance and results. Among them, YOLOv5s represents the model with the lowest network depth and width within the YOLOv5 algorithm, as depicted in Figure 1.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202311_9(11).0006

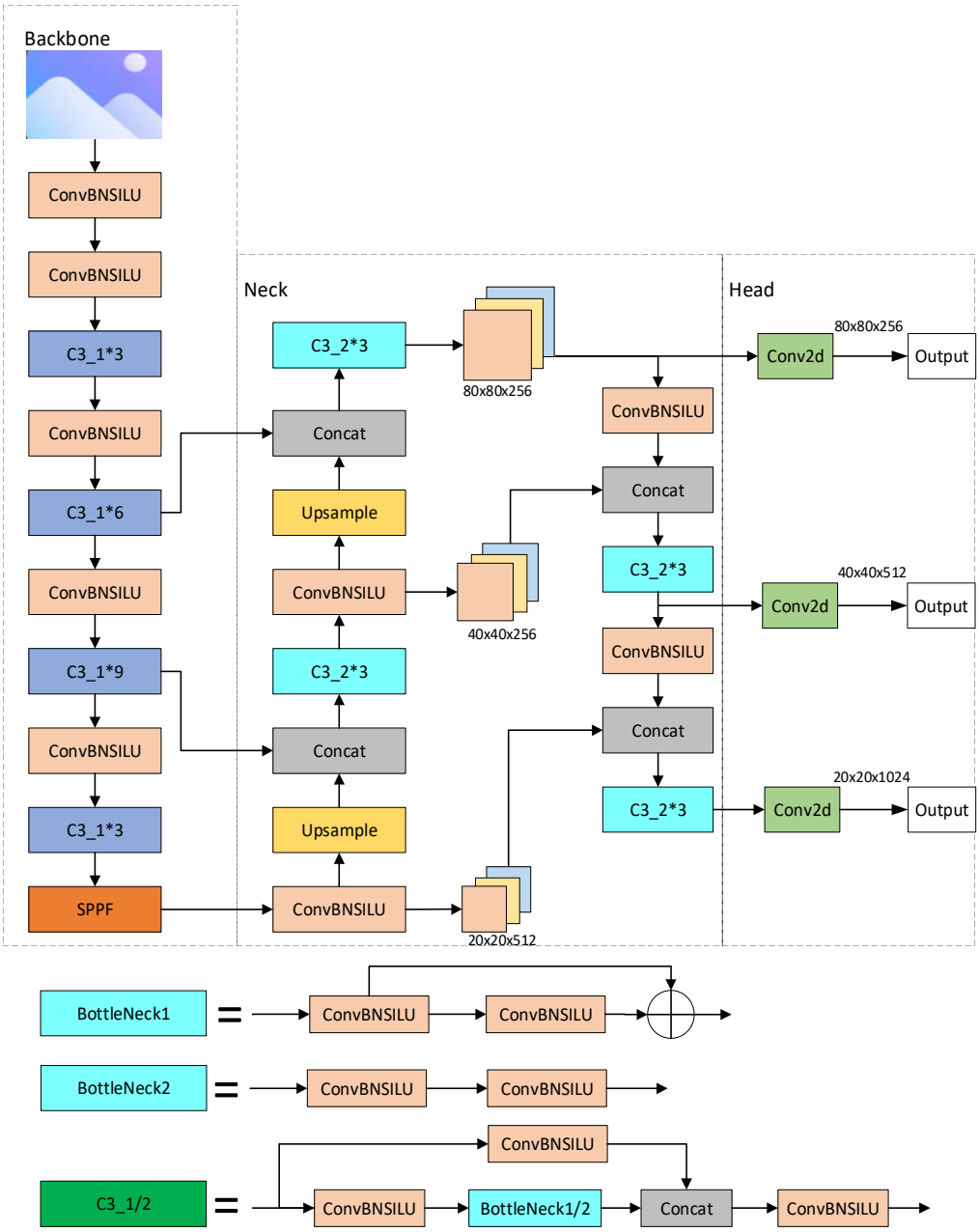

Figure 1. YOLOv5 Network Architecture

Darknet53 is used as the backbone network in YOLOv5. The input image information is processed by a series of Conv BNSILU structures in this network. The Conv BNSILU structure is composed of convolutional layers, Batch Normalization, and the Si Lu function, which are employed for image feature extraction. This structure effectively captures semantic information in the image and exhibits strong non-linear fitting capabilities. YOLOv5 utilizes this structure to extract features from the input image and generate precise feature representations for object detection tasks. YOLOv5 substitutes the original Swish activation function with the Si Lu function. The Swish activation function possesses characteristics such as unboundedness, smoothness, and non-monotonicity. The Swish function is also referred to as the Si Lu function when β=1. To enhance the network's efficiency, YOLOv5 replaced the Focus structure of the original network's first layer with a 6x6 convolutional layer in the Backbone section of its main network. This replacement demonstrates increased efficiency on certain GPU devices. The Spatial Pyramid Pooling structure is substituted with the Spatial Pyramid Pooling with Feature fusion structure in the Neck section. The feature pyramid structure is employed by both the SPP and SPPF structures to extract multi-scale feature information. These replacements and

ISSN: 2414-1895 DOI: 10.6919/ICJE.202311_9(11).0006

structural adjustments enhance the efficiency of YOLOv5 in performing object detection tasks without compromising accuracy. By sequentially passing through multiple 5x5 MaxPool layers, the SPPF structure significantly reduces the computational cost of the system during input processing. The schematic diagrams of the SPP and SPPF structures are presented in Figure 2 and Figure 3, respectively.

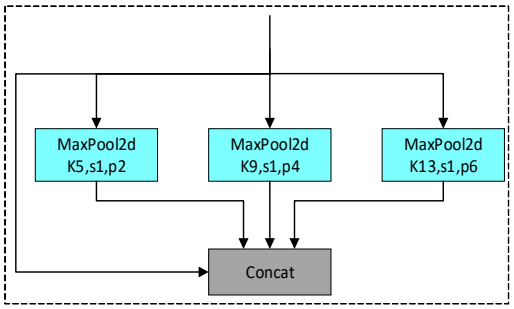

Figure 2. SPP structure

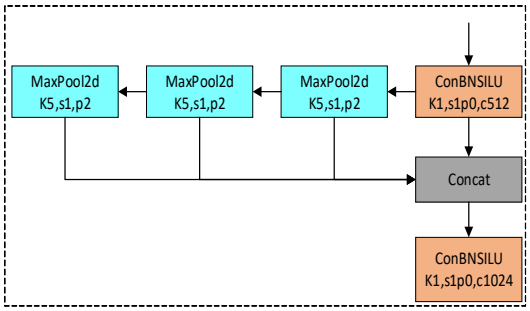

Figure 3. SPPF structure

In YOLOv5, the Neck layer combines Feature Pyramid Network and Path Aggregation Network structures for integrating multi-scale feature information. FPN merges features in a top-down pyramid structure from different levels, generating feature maps of different scales. PAN enhances the expressive power of features through bottom-up information fusion. This combination enables effective integration of multi-scale feature information in the Neck layer and enhances the performance of object detection. Through the combination of FPN and PAN structures, different backbone output layers can aggregate semantic and localization features for different detection layers. Additionally, the Neck incorporates the Cross Stage Pariti Network structure into each C3 module, utilizing CSP1_X in the Backbone Back bone network and CSP2_X in the Neck. The structure diagram of C3-1/2 is depicted in Figure 4. Through the utilization of the CSP module, the feature maps of the base layer can be partitioned and merged using a cross-stage hierarchical structure, achieving accuracy while reducing computational complexity. Predictions are generated for targets of various sizes, small, medium, and large in the Head layer.

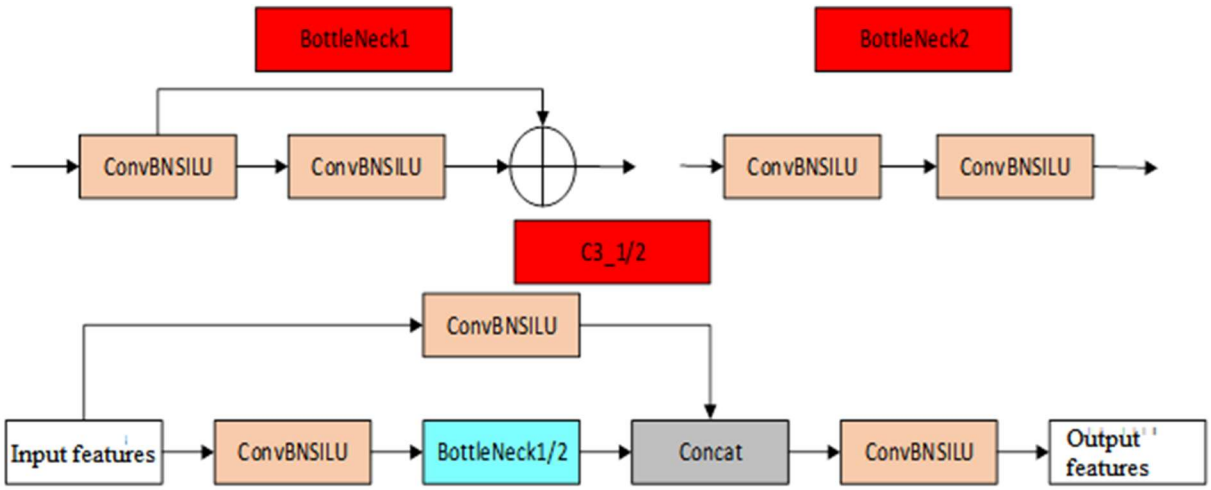

Figure 4. C3-1/2 Figure

By conducting fine-tuning training on the original YOLOv5s algorithm model, one can achieve effective recognition of detected objects.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202311_9(11).0006

3. Experiment and Results

3.1 Experimental Data Set

To enhance the model validation and assess its applicability in various industrial production processes, this study employs a dataset consisting of commonly encountered irregular parts. These parts encompass Connector A, Connector B, Connector C, Vertical Component, and Vibration Component. The dataset is compiled by capturing photos under diverse environmental conditions, backgrounds, and occlusions. Initially, there were 1207 original photos taken with a resolution of 3000×4000. To ensure the dataset's comprehensiveness in terms of data diversity, sample complexity, and feature variation, data augmentation techniques were employed to expand the pool of collected images, resulting in a total of 4677 images. Among these, 3682 images are allocated for the training set, while the test set comprises 981 images. The dataset was meticulously annotated using the Label me annotation software to obtain the final dataset.

3.2 Train YOLOV5s

The experimental environment of this study is shown in Table 1.

Table 1. Experimental Environment Table

Configuration Name	Version/parameter
Operating system	64-bit Ubuntu 20.04
CPU	12 vCPU Intel(R) Xeon(R) Platinum
	8255C CPU @ 2.50GHz
GPU	Ge Force RTX 2080TI
Random Access Memory (RAM)	40G
CUDA	11.3
PY Torch deep learning framework	1.10.0

In this study, YOLOv5s was utilized as the foundational training model for the irregular components. To mitigate the impact of the model's original weights on the training outcomes, we opted not to employ the original training weights. The network was initially assigned a learning rate of 0.01, accompanied by a momentum factor of 0.937. The training employed the HYP scratch-low hyperparameters. The input dimensions were defined as 640×640. The training was performed for 200 iterations, utilizing a batch size of 32 and 8 data loading threads.

Table 2. Comparison of experimental results

Network model	Map_0.5/%	Model size/M	Frames per second
YOLOv4	89	13.7	34.96
YOLOv5s	92.1	12.8	110.96

In scenarios where similar model sizes and high real-time requirements are involved, the table below highlights that the detection speed of YOLOv5s reaches 110.96 frames per second, demonstrating a significant accuracy advantage over the YOLOv4 algorithm. To further scrutinize the YOLOv5s and YOLOv4 algorithms, experiments were conducted on a self-made parts dataset, and the resulting training function curves were plotted, refer to Figure 5. Analyzing the training curves reveals that YOLOv5s exhibits a faster training speed, stabilizing after approximately 200 epochs, compared to

ISSN: 2414-1895 DOI: 10.6919/ICJE.202311_9(11).0006

YOLOv4 which gradually stabilizes after around 300 epochs. This observation suggests that YOLOv5 possesses more reasonable initial parameter values and converges at a swifter rate.

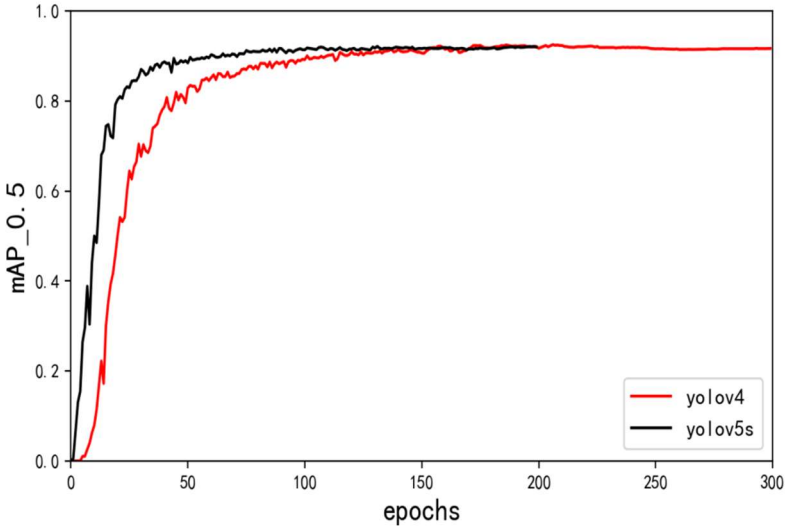

Figure 5. Training curve

4. Test Results and Analysis

We evaluated the algorithm's detection performance on the dataset of irregular parts through analysis of color variations, angles, and multi-object detection. The experimental results are presented in the following figure, where the numerical values indicate the confidence of the detection results. The results clearly demonstrate that both YOLOv4 and YOLOv5 accurately detect various types of irregular industrial parts. However, YOLOv5 clearly outperforms YOLOv4 in terms of accuracy and overall performance.

Specifically, when it comes to detecting rotating objects, the figure reveals that the detection results of YOLOv4 are somewhat unsatisfactory. However, despite its ability to recognize the parts, the confidence level is only 64%, indicating shortcomings in accurately sorting complex irregular parts. Conversely, YOLOv5 demonstrates a confidence level of up to 94% in this aspect, clearly outperforming YOLOv4.

The experimental results strongly demonstrate the superiority of YOLOv5 in handling rotation object detection. Although YOLOv4 performs well in general cases, YOLOv5 excels particularly in dealing with rotation objects, providing powerful support for accurate sorting of complex irregular parts.

5. Conclusion

The present study trained a neural network based on YOLOv5s using a custom dataset of irregular parts, resulting in the creation of corresponding weight files. Eventually, an efficient detection model was obtained. Several methods were employed to ensure the dataset's relevance to real-world applications, and part object detection was achieved through video stream processing using computer vision techniques. The incorporation of the YOLOv5s algorithm not only achieved outstanding precision in detecting irregular industrial parts but also demonstrated a detection speed of 115 frames per second. The implementation of YOLOv5s offers significant support for facilitating efficient assembly and processing on automated production lines, considering the urgent demand for real-time and efficiency in industrial production. In future work, the dataset will be expanded to encompass more diverse part data, aiming to improve the performance of the model. A specific focus will be given to model lightweighting to enhance its suitability for industrial production domains. This enhancement will improve production line efficiency and ensure precise part detection and assembly, meeting the high standards of industrial production.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202311_9(11).0006

Acknowledgments

This research was funded by Development of special structure technology and simulation technology of high precision servo motor for aerospace application [Project No. 2022ZD027], and Research on visual inspection technology of complex special-shaped parts of satellites [Project No. 2022SKYZ294].

References

[1] Jia, D., Zhu, N., Yang, N., et al. (2019). Research review on image matching methods. Journal of China Image and Graphics, 24(5), 677-699.

[2] Ding, X., Zhao, Q., Li, Y., et al. (2018). Improved target recognition algorithm based on template matching. Journal of Shandong University (Engineering Science), 48(2), 1-7.

[3] Shan, M., Gao, W., Yang, Y., Fan, B., & Jiang, X. (2022). Brake parts defect detection algorithm based on improved Faster R-CNN. Foreign Electronic Measurement Technology, 41(04), 22-28.

[4] Song, S., Hou, Z., Wang, Q., Ni, Y., & Huang, Q. (Year). Application of Improved YOLOV3 Algorithm in Component Recognition. Journal Name, Pages 1-9.

[5] 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, 37(9), 1904-1916.

[6] Gir Shick, R. (2015). Fast R-CNN. In Proceedings of the International Conference on Computer Vision (pp. 1440-1448).

[7] Ren, S., He, K., Gir Shick, R., & Sun, J. (2015). Faster R-CNN: Towards Real-time Object Detection with Region Proposal Networks. In Proceedings of the Neural Information Processing Systems (pp. 91-99).

[8] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., & Berg, A. C. (2016). SSD: Single Shot Multi Box Detector. In Proceedings of the European Conference on Computer Vision (pp. 21-37). Springer International Publishing.

[9] Redmon, J., Div Vala, S., Gir Shick, 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 (pp. 779-788).

[10] Redmon, J., & Farhadi, A. (2017). YOLO9000: Better, Faster, Stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 6517-6525).

[11] Redmon, J., & Farhadi, A. (2018). YOLOv3: An Incremental Improvement. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 1-6).

[12] Lin, T., Dollar, P., Gir Shick, 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 (CVPR) (pp. 936-944).