YOLOv7-based real-time detection in production workshop scenarios

1st Ziyang Peng
College of Electrical and Information Engineering
Hunan University
Changsha, China
pzy1215281894@hnu.edu.cn

2nd Yuqiao He
College of Electrical and Information Engineering
Hunan University
Changsha, China
s2209w0819@hnu.edu.cn

3rd Jianxu Mao*
College of Electrical and Information Engineering
Hunan University
Changsha, China
maojianxu@hnu.edu.cn

4th Junfei Yi
College of Electrical and Information Engineering
Hunan University
Changsha, China
yijunfei@hnu.edu.cn

5th Ziming Tao
College of Electrical and Information Engineering
Hunan University
Changsha, China
taozimingphd@hnu.edu.cn

6th Xuesan Su
College of Electrical and Information Engineering
Hunan University
Changsha, China
suxuesan@hnu.edu.cn

***Abstract*—Safety helmets and work clothes and other safety equipment are of great significance for ensuring the safety of workers. In view of the complex problems of the construction site scene, this paper proposes a real-time detection algorithm for the production workshop scene based on improved YOLOv7. First, the Dyhead module is added to the Head part of the network architecture, and the multi-head self-attention mechanism is coordinatedly combined in the scale-aware feature layer, the spatial-aware spatial position and the task-aware output channel. Then, the convolution kernel in the main network is replaced with CoordCov, and the input coordinate information is added to the input feature map of the network as an additional channel to help the network learn to process spatial-related information. The experimental results show that compared with the original model, the mAP of the improved model is increased by 1.4%, Recall is increased by nearly 1%, precision is increased by 1.43%, and the improved model can effectively improve the detection performance of workers' safety equipment in the construction site scene.**

Keywords—Safety equipment ; Real-time detection algorithm; YOLOv7 ; Dyhead; CoordCov

I. INTRODUCTION

Hard hats and other safety equipment such as work clothes are common protective equipment in production and life. For workers in construction scenarios, they are essential. However, in actual production, there are often situations where workers forget to wear or wear them incorrectly. This has to some extent buried hidden dangers to the life safety of factory construction workers. On the other hand, accidents will cause the project to fail to be completed normally and smoothly on schedule, damaging economic benefits. At present, most factories in China still use traditional manual detection to detect the normal wearing of safety helmets and other safety equipment such as work clothes. There are serious omissions even under the condition of low efficiency and high labor costs. In recent years, with the development of artificial intelligence visual perception technology, it has achieved certain results in intelligent monitoring, object recognition, traffic safety and other fields. Due to the fact that this technology saves labor costs and is efficient, it is a current research hotspot to think about how to combine it with real-time detection of safety helmet wearing and other safety equipment such as work clothes.

In recent years, many domestic and foreign scholars have conducted research and experiments on this practical problem. HE Min et al. [2] proposed an improved algorithm based on YOLOv7-Tiny for safety helmet and work clothes monitoring due to the complex scenarios, diverse targets, and some occlusion problems in safety control of power field operations. SUN Guodong et al. [3] proposed an object detection algorithm that improves Faster R-CNN by fusing self-attention mechanisms for safety helmet detection, but the algorithm detection rate is low. LI Mingshan et al. [4] increased the feature fusion function of the branch network in the SSD model and improved the default box configuration to improve the accuracy of safety helmet detection in practical application scenarios. Kai-Yang Cao et al. [5] proposed a safety helmet wearing detection algorithm based on YOLOv5 to address the issues of existing safety helmet detection algorithms in detecting small targets, occluded targets, and densely populated targets. F. Zhou et al. [6] proposed a safety helmet detection method based on YOLOv5 is proposed to address worker safety concerns. By utilizing the pre-training weights of a trainable object detector, the mAP of YOLOv5x is improved, demonstrating the effectiveness of safety helmet detection based on YOLOv5.

Although all the above researches are constantly improving and modifying the algorithm security and real-time performance for safety helmets and other safety work equipment, the existing algorithms are still insufficient to meet the safety needs of workers in actual production, and there are still problems of insufficient real-time performance or low accuracy. In order to ensure the real-time and accuracy of visual detection of safety helmets and other safety work equipment, this paper proposes a real-time intelligent detection algorithm for safety risk management based on improved YOLOv7[7], which is used for real-time detection of safety helmets and other safety work equipment while ensuring detection accuracy.

The following are the main contributions of our paper:

- This paper introduces Dyhead[8] into the network architecture of YOLOv7. This method coherently combines multi-head self-attention mechanisms in scale-aware feature layers, spatially-aware spatial positions, and task-aware output channels. After introducing Dyhead, the model's accuracy is

* corresponding author

2023 China Automation Congress (CAC) | 979-8-3503-0375-9/23/$31.00 ©2023 IEEE | DOI: 10.1109/CAC59555.2023.10451838

improved by 3%, significantly improving the representation ability of the target detection head.

- This paper uses CoordConv[9] to replace some convolution kernels in YOLOv7. This method adds the input coordinate information as an additional channel to the input feature map of the network to help the network learn to handle spatially related tasks. After introducing CoordConv, the model's recall rate is increased by 1.27%.
- Finally, the improved YOLOv7 model is applied to the worker operation scene. This model can provide support for monitoring potential safety risks in the worker operation scene, which helps ensure that workers receive effective protection and management during the operation process.

II. METHOD

A. The network structure

We propose a real-time detection algorithm for production workshop scenarios based on improved YOLOv7 model, as shown in the Fig.1. First, due to the real-time problem guaranteed by the selected YOLOv7 baseline model, and to further solve the accuracy problem of safety products such as safety helmets, we first combine DyHead and use a structure with multiple deformable convolution layers or adaptive convolution layers in the network architecture. These layers can adaptively transform the receptive field according to the shape and position of the target to improve the accuracy and robustness of object detection or instance segmentation. Then, using CoordConv, the input coordinate information is added as an additional channel to the input feature map of the network to help the network learn to handle spatially related tasks.

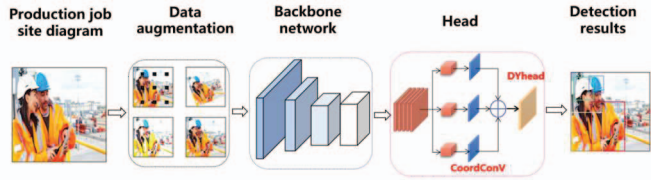

Fig. 1. Real-time detection algorithm for production workshop scenarios based on improved YOLOv7 model

B. The Dynamic Head module

In the field of object detection, various algorithms have emerged due to the complexity of combining classification and localization. These algorithms attempt to enhance performance on detection heads, but they lack a unified perspective on the detection problem. Based on this observation, this paper introduces a novel dynamic head framework that unifies attention mechanisms with object detection heads. The dynamic head framework considers the output of the backbone, which serves as the input to the head, as a three-dimensional tensor: level × space × channel. Here, "level" represents the feature hierarchy, "space" denotes the product of the width and height of the feature map (HW), and "channel" refers to the number of channels.

For a given feature tensor $T\epsilon\delta^{L\times S\times C}$, the generalized attention can be described as shown in (1).

$$\partial(T) = V(T)\cdot T \tag{1}$$

If the above three-dimensional tensor is directly processed using convolutional layers, it would not only pose challenges in optimization but also result in a significant computational burden. To address this problem, a decoupled attention mechanism is employed, transforming the aforementioned attention form into three serialized attentions. Each attention focuses on a single dimension, and its mathematical expression is shown in (2).

$$\partial(T) = V_C(V_S(V_L(T)\cdot T)\cdot T)\cdot T \tag{2}$$

where $V_S(\cdot), V_C(\cdot)$ and $V_L(\cdot)$ represent the attention on the S, C, and L dimensions.

And then integrate the above three types of attention into an efficient attention learning mechanism, as illustrated in Figure 2. The specific implementation methods for the three types of attention will be provided below.

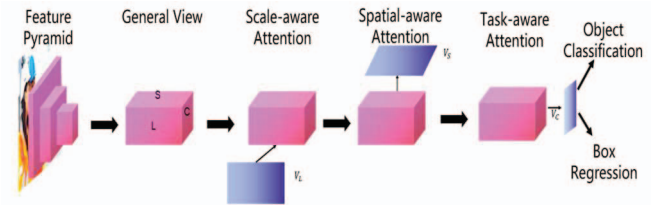

Fig. 2. An illustration of Dynamic Head approach. It contains three different attention mechanisms, each focusing on a different perspective: scale-aware attention, spatial-aware attention, and task-aware attention.

1.Scale-aware attention （level-wise）$V_L(\cdot)$：Scale-aware attention operates on the level dimension, recognizing that different levels of feature maps correspond to different object scales. By applying attention at the level hierarchy, it enhances the scale-awareness capability of object detection. Let $X_1 = V_L(T)\cdot T$, where the mathematical expression for solving X_1 is shown in (3) and (4)

$$X_1 = \sigma(F_{(S,C)})\cdot T \tag{3}$$

$$F_{(S,C)} = \alpha\left(\frac{1}{SC}\sum_{S,C} T\right) \tag{4}$$

where the function α(·) is a linear function, which is used similarly to a 1x1 convolution. The expression $\sigma(x) = \max\left(0, \min\left(1, \frac{x+1}{2}\right)\right)$ corresponds to the hard-sigmoid function.

2.Spatial-aware attention （spatial-wise）$V_S(\cdot)$：The spatial-aware attention module operates on the spatial dimension and focuses on the discriminative ability of different spatial positions. Considering the high dimensionality of S, this module needs to be decoupled into two steps: first, utilizing deformable convolutions to learn sparsity and then aggregating features at the same spatial position across different levels. So we let $X_2 = V_S(T)\cdot T$, where the mathematical expression for solving X_2 is shown in (5) and (6).

$$X_2 = \frac{1}{L}\sum_{m=1}^{M}\sum_{l=1}^{L} w_{l,m}\cdot T(l;\varpi_{p_k};c)\cdot\Delta n_k \tag{5}$$

$$\varpi_{p_k} = p_k + \Delta p_k \tag{6}$$

where M represents the number of positions sampled sparsely, ϖ_{p_k} introduces position offsets to focus on discriminative regions, and Δn_k refers to the learnable importance measurement factor related to position p_k.

3.Task-aware attention （channel-wise）$V_C(\cdot)$： The task-aware attention operates on the channel dimension, and it dynamically activates or deactivates feature channels to select different tasks. Different channels correspond to different tasks, and by incorporating an attention module in the channel dimension, the task-aware attention enhances the perception of different tasks. This allows for joint learning and generalization of target representations across tasks in object detection. Let $X_3 = V_C(T) \cdot T$, where the mathematical expression for solving X_3 is shown in (7) and (8).

$$X_3 = \max(\omega_T^1, \omega_T^2) \tag{7}$$

$$\omega_T^a = \mathrm{m}^a(T) \cdot T_C + q^a(T), a = 1,2 \tag{8}$$

where $\rho(\cdot) = [m^1, m^2, q^1, q^2]^T$ is a hyperfunction used to learn and control the activation threshold. The usage of $\rho(\cdot)$ is similar to Dynamic ReLU, where it first performs global pooling on the L × S dimensions, followed by two fully connected layers, a normalization layer, and finally normalized output using the shifted sigmoid function.

C. The CoordConv convolutional module

The construction of CoordConv is depicted in Figure 3: Compared to traditional convolution, CoordConv adds two additional channels after the input feature map, one representing the x-coordinate and the other representing the y-coordinate. The subsequent process follows the same steps as a regular convolution.

Upon analysis, traditional convolution possesses three characteristics: low parameter count, computational efficiency, and translation invariance. CoordConv inherits the first two characteristics, while the network itself dynamically maintains or discards translation invariance based on its learning progress. Although it may appear to compromise the model's generalization ability, allocating a portion of the network's capacity to model non-translation invariance actually enhances the model's generalization capability.

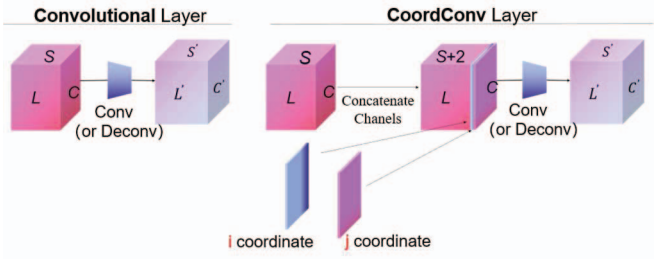

Fig. 3. Comparison of 2D convolutional and CoordConv layers. (left) A standard convolutional layer maps from a representation block with shape S × L × C to a new representation of shape S′ × L′ × C′. (right) A CoordConv layer has the same functional signature, but accomplishes the mapping by first concatenating extra channels to the incoming representation. These channels contain hard-coded coordinates, the most basic version of which is one channel for the i coordinate and one for the j coordinate, as shown above. These coordinates are not unique, and inputting other derived coordinates based on the specific situation is equally effective.

In fact, if the coordinate channels of CoordConv do not learn any meaningful information, CoordConv is equivalent to traditional convolution, fully preserving translation invariance. Conversely, if the coordinate channels learn certain information, CoordConv exhibits a certain degree of translation dependence. Thus, the translation invariance and dependence of CoordConv can be dynamically adjusted based on different tasks. It can perform both identity mappings and learn additional features. Therefore, we can utilize CoordConv when spatial information needs to be captured. On one hand, the model's parameter count is reduced by 10-100 times, which improves detection speed to some extent, making it suitable for real-time detection. On the other hand, translation invariance is not completely eliminated but rather allows the network to learn varying degrees of translation invariance and dependence based on the task at hand.

III. EXPERIMENTS

A. Datasets

This article collected a series of image data related to wo rker job scenarios, totaling 1427 images. These image data a re divided into training set, validation set, and test set, with 9 99 images in the training set, 286 images in the validation se t, and 142 images in the test set. This dataset is sourced from Roboflow's public dataset "workers-safety-equipment".[10]

To effectively analyze and process these images, the detection objects are divided into seven different categories, including 'Helmet', 'Gloves', 'bare-arms', 'Person', 'Vest', 'Non-Helmet', and 'Shoes'. The definitions of these categories cover common situations in job scenarios, mainly involving personal protective equipment. This dataset can provide support for monitoring potential safety risks in worker job scenarios and help ensure effective protection and management for workers during the job process.

B. Experimental environment and training parameters

The hardware configuration for this article consists of a GPU (NVIDIA GeForce RTX 3090) and an Intel(R) Core(TM) i9-12900K CPU. The software configuration includes Ubuntu 18.04, PyTorch 1.7.0, and Python 3.8.3. The algorithm parameters are set as follows: a learning rate of 0.02 and 200 epochs.

C. Experiments

To verify the effectiveness of the method proposed in this paper, the model proposed in this paper is compared with the original YOLOv7, YOLOv7 with only CoordConv added, and YOLOv7 with only DYhead added. The results are shown in TABLE I.

TABLE I. THE IMPACT OF IMPROVED MODULES ON ALGORITHM PERFORMANCE THROUGH ABLATION EXPERIMENTS

Model		Evaluation indicators			
CoordConv	*DYhead*	*P*	*R*	*mAP0.5*	*mAP*
		0.9067	0.7753	0.8584	0.5602
√		0.9109	0.788	0.8657	0.5702
	√	**0.9381**	0.7711	0.8687	0.5812
√	√	0.921	**0.784**	**0.872**	**0.583**

[a.] Note: Bold is the model with the highest accuracy in this module.

The model with the addition of DYhead demonstrates a significant improvement in accuracy, with an increase of approximately 3% compared to the original YOLOv7. DYhead introduces advanced feature extraction and attention mechanisms, effectively enhancing the precision of object detection. On the other hand, the inclusion of CoorConv contributes to the improvement of recall rate. By incorporating CoorConv, the model gains a better understanding of the spatial relationships of objects, thereby enhancing the detection recall rate. Experimental verification shows that the

model with only CoorConv achieves an approximately 1% increase in recall rate compared to the original YOLOv7.

In the proposed model of this paper, the simultaneous integration of DYhead and CoorConv leads to improvements not only in accuracy and recall rate but also in the mAP metric. By comparing the experimental results, the proposed model in this paper achieves a 1.4% enhancement in mAP@0.5, indicating its superior detection capabilities.

D. Comparative experimental analysis

This paper evaluates the performance indicators of four object detection models, namely ATSS[11], PAA[12], YOLOX[13], and Ours, as shown in Table 2. By comparing the accuracy, recall, and mean Average Precision (mAP) among the models, it can be observed that our model performs the best across all metrics. Our model achieves an accuracy of 0.921, a recall of 0.784, an mAP@0.5 of 0.872, and an overall mAP of 0.583. These results demonstrate that our model effectively captures targets in worker operation scenarios and exhibits higher precision and recall compared to the other models, thereby enhancing the safety of worker operations effectively.

TABLE II. PERFORMANCE COMPARISON OF DIFFERENT MODELS

Model		Evaluation indicators			
Name	Pub	P	R	mAP0.5	mAp
ATSS	2020CVPR	0.820	0.606	0.8200	0.5160
PAA	2020ECCV	0.869	0.642	0.8687	0.5640
YOLOX	2021CVPR	0.854	0.630	0.8535	0.5699
Ours		**0.921**	**0.784**	**0.872**	**0.583**

[b.] Note: The bolded part is the accuracy of the student model after distillation

The actual training process and detection results of our model are shown in Fig. 4 and Fig. 5. In this paper, a Dynamic Head module is introduced after the neck detection layer module in YOLOv7. It unifies the three types of attention, scale-aware attention, spatial-aware attention, and task-aware attention, into an efficient attention learning mechanism on the level × space × channel dimensions. This attention reconstruction on the feature maps enables the network to fully utilize the information from the three types of attention for the detection of safety equipment and other targets. Consequently, the model achieves better detection performance in training on safety equipment and similar objects. Additionally, the CoordConv layer is incorporated into the convolutional layers to further enhance the model's generalization capability. The visualizations of the detection results after adding these modules are presented in Figure 5, demonstrating the effectiveness of introducing the CoordConv layer and the Dyhead module in improving the model's performance.

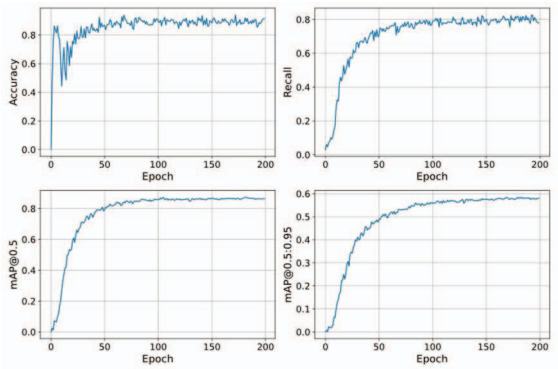

Fig. 4. The training process of our model

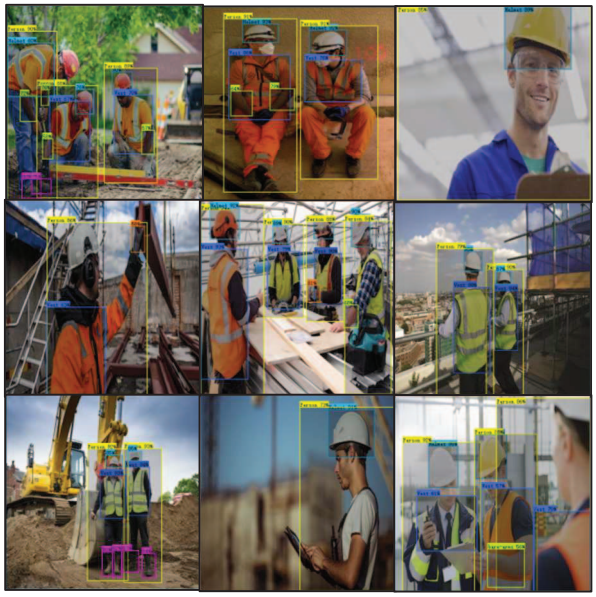

Fig. 5. The actual detection results of the our model

IV. CONCLUSION

In order to better protect the safety of workers in production scenarios, this paper proposes a real-time detection algorithm based on YOLOv7 for detecting whether safety equipment such as safety helmets is worn correctly. Corresponding improvements are made to the performance of YOLOv7 in safety equipment detection. Under the original attention mechanism, a novel dynamic head framework is added, which uses attention mechanisms to unify different object detection heads. It can significantly improve the expression ability of the model's object detection head, thereby improving the accuracy and performance of the model. At the same time, CoordConv layer is introduced on the convolutional layer to process object detection, further improving the generalization ability of the model. Good detection results can be achieved in different safety equipment detections. Finally, a series of experiments were conducted on the above improvements. The experimental results show that this algorithm has high accuracy in safety equipment detection and has been compared with the original model. On the basis of the original model, the accuracy rate has increased by 1.43%, the recall rate has increased by 0.83%, and mAP@0.5 has increased by 1.3%. At the same time, our model was compared with three current excellent object detection models (ATSS, PAA, YOLOX) for relevant performance indicators. The experimental results show that our model performs better in accuracy rate, recall rate and mAP. All these results show that our model can effectively capture safety equipment in worker production scenarios and has higher accuracy and recall rate than other detection models, which can effectively improve worker operation safety.

V. DISCUSSION

Despite the promising results and contributions of our study, it is important to acknowledge its limitations. One notable limitation is the constrained computational resources, encompassing hardware limitations in terms of processor and GPU, as well as limited computational time. These constraints may have influenced the scale and complexity of our experiments and the size of our dataset. While we have optimized our research within these constraints, future studies with more substantial computational resources could extend

our research and improve the robustness and scalability of our proposed methods.

ACKNOWLEDGMENT

This work was supported in part by the National Key Research and Development Program of China under Grant 2020YFB1712600; in part by the National Natural Science Foundation of China under Grant 62133005, Grant 62293510/62293512, Grant 62293510/62293515, Grant 62027810; in part by the Special funding support for the construction of innovative provinces in Hunan Province under Grant 2021GK1010 (Corresponding author is Jianxu Mao)

REFERENCES

[1] R. Zhao, H. Liu, P. Liu et al., "Helmet detection algorithm based on improved YOLOv5s," Journal of Beijing University of Aeronautics and Astronautics, pp. 1-16, 2021.

[2] M. He, L. Qin, F. Zhao et al., "Intelligent Detection Algorithm of Security Risk Management and Control for Power System On-site Operation," High Voltage Engineering, vol. 49, no. 06, pp. 2442-2457, 2023.

[3] G. Sun, C. Li, H. Zhang et al., "Safety Helmet Wearing Detection Method Fused with Self-Attention Mechanism," Computer Engineering and Applications, vol. 58, no. 20, pp. 300-304, 2022.

[4] M. Li, Q. Han and T. Zhang et al., "Safety Helmet Detection Method of Improved SSD," Computer Engineering and Applications vol. 57 no. 08 pp. 192-197,2021.

[5] K. -Y. Cao, X. Cui and J. -C. Piao, "Smaller Target Detection Algorithms Based on YOLOv5 in Safety Helmet Wearing Detection," 2022 4th International Conference on Robotics and Computer Vision (ICRCV), Wuhan, China, 2022, pp. 154-158

[6] F. Zhou, H. Zhao and Z. Nie, "Safety Helmet Detection Based on YOLOv5," 2021 IEEE International Conference on Power Electronics, Computer Applications (ICPECA), Shenyang, China, 2021, pp. 6-11.

[7] C.Y.Wang et al., "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle WA USA , June 19th -24th ,2023 , pp .7464 -7475 .

[8] X. Dai, Y. Chen, B. Xiao, et al., "Dynamic head: Unifying object detection heads with attentions," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 7373-7382.

[9] Liu R, Lehman J, Molino P, et al. An intriguing failing of convolutional neural networks and the coordconv solution[J]. Advances in neural information processing systems, 2018, 31.

[10] wyhil, "workers-safety-equipment Dataset," Roboflow Universe, 2023. [Online]. Available: https://universe.roboflow.com/wyhil-ru2ds/workers-safety-equipment-z1mra. [Accessed: July 11, 2023].

[11] S. Zhang et al., "Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA (virtual), Jun. 14-19, 2020, pp. 975

[12] K. Kim and H. S. Lee., "Probabilistic anchor assignment with iou prediction for object detection," in Proceedings of the European Conference on Computer Vision (ECCV), Glasgow (virtual), Scotland (UK), Aug. 23-28, 2020.

[13] Z. Ge et al., "Yolox: Exceeding yolo series in 2021," arXiv preprint arXiv:2107.08430v1 [cs.CV], Jul. 17, 2021.

[14] J. Terven and D. Cordova-Esparza., "A comprehensive review of yolo: From yolov1 and beyond," arXiv preprint arXiv:2304.00501v1 [cs.CV], Apr. 1st ,2023.