# Postprint: Field Maize Kernel Detection and Counting Based on Multiple Deep Learning Algorithms

**Authors:** Liu Xiaohang, Zhang Zhao, Liu Jiaying, Zhang Man, Li Han, Paulo FLORES, Han Xiongzhe, Zhang Zhao

**Date:** 2023-02-17T00:00:00+00:00

## Abstract

To rapidly and accurately obtain information on the number of lost kernels during corn harvest for management tasks such as adjusting harvesting losses, the performance of mainstream single-stage and two-stage object detection networks for field corn kernel counting was comparatively evaluated. First, an RGB camera was used to acquire image data containing different backgrounds and lighting conditions, and a dataset was further generated; second, different object detection networks for kernel recognition were constructed, including Mask RCNN, EfficientDet-D5, YOLOv5-L, and YOLOX-L, and the four constructed networks were trained, validated, and tested using the collected 420 valid images, with 200, 40, and 180 images respectively; finally, kernel counting performance evaluation was conducted based on the recognition results of the test set images. The experimental results showed that the YOLOv5-L network achieved an average precision of 78.3% for test set image detection, with a model size of only 89.3 MB; the detection accuracy, miss rate, and F1-score for kernel counting were 90.7%, 9.3%, and 91.1% respectively, with a processing speed of 55.55 f/s. Its recognition and counting performance were superior to those of Mask R-CNN, EfficientDet-D5, and YOLOX-L networks, and it demonstrated strong robustness to images with different degrees of ground occlusion and kernel aggregation states. The deep learning object detection network YOLOv5-L can achieve real-time monitoring of corn harvest loss kernels in actual operations, with high accuracy and strong applicability.

## Full Text

# Infield Corn Kernel Detection and Counting Based on Multiple Deep Learning Networks

**LIU Xiaohang**[1,2], **ZHANG Zhao**[1,2*], **LIU Jiaying**[1,2], **ZHANG Man**[1,2], **LI Han**[1,2], **Paulo FLORES**[3], **HAN Xiongzhe**[4,5]

[1] Key Laboratory of Smart Agriculture System Integration, Ministry of Education, China Agricultural University, Beijing 100083, China
[2] Key Laboratory of Agricultural Information Acquisition Technology, Ministry of Agriculture and Rural Affairs, China Agricultural University, Beijing 100083, China
[3] Department of Agricultural and Biosystems Engineering, North Dakota State University, Fargo, ND 58012, United States
[4] Department of Biosystems Engineering, Kangwon National University, Chuncheon 24341, Korea
[5] Interdisciplinary Program in Smart Agriculture, Kangwon National University, Chuncheon 24341, Korea

### Abstract

To rapidly and accurately obtain information on lost corn kernels during harvest for adjusting combine settings and managing harvest losses, this study evaluated and compared the performance of mainstream single-stage and two-stage object detection networks for infield corn kernel counting. First, an RGB camera was used to acquire images under varying backgrounds and illumination conditions to construct a comprehensive dataset. Second, four different target detection networks were implemented for kernel recognition: Mask R-CNN, EfficientDet-D5, YOLOv5-L, and YOLOX-L. A total of 420 effective images were collected for training, validation, and testing, with 200 images for training, 40 for validation, and 180 for testing. Finally, kernel counting performance was evaluated based on recognition results from the test set images. Experimental results demonstrated that the YOLOv5-L network achieved the best overall performance among the four models. The average precision (AP) for test set images reached 78.3%, with a model size of only 89.3 MB. The detection accuracy, miss-detection rate, and F1-score for kernel counting were 90.7%, 9.3%, and 91.1%, respectively, with a processing speed of 55.55 f/s. Both recognition and counting performance surpassed those of Mask R-CNN, EfficientDet-D5, and YOLOX-L networks, exhibiting strong robustness across images with varying degrees of ground occlusion and kernel aggregation states. The YOLOv5-L network enables real-time monitoring of corn harvest loss kernels in practical operations with high precision and strong applicability.

**Keywords:** harvest loss; infield corn kernel; deep learning; kernel counting; YOLOv5-L; YOLOX-L; Mask R-CNN; EfficientDet-D5

## 1 Introduction

As one of the traditional staple crops, corn has become the most widely cultivated and traded crop worldwide due to its versatile "grain-feed-forage" attributes. Compared with rice and wheat, corn experiences relatively higher grain loss rates during mechanical harvesting. Conducting research on monitoring corn kernel loss during field harvest is of great significance for evaluating combine harvester performance, enabling autonomous loss adjustment, and ensuring actual grain yield.

Current corn harvest loss detection primarily relies on sensors installed at different positions on the harvester (such as the cleaning sieve or straw outlet), including optical, acoustic, microwave, piezoelectric ceramic, and piezoelectric film sensors. These sensors capture signal characteristics like frequency and amplitude when kernels impact the sensing plate, and machine learning methods are employed to construct predictive models for real-time monitoring of field corn harvest loss rates. However, factors such as impact angle and velocity variations often cause false recognition. Moreover, limitations in sensitive material properties (installation position, sensitivity, effective range), crop conditions, and harvester operating parameters (feed rate, straw-to-grain ratio, travel speed, drum speed) reduce the reliability of loss estimation based on single-stage monitoring (e.g., separation and cleaning). Consequently, sensor-based methods, influenced by multiple complex factors, struggle to meet the precise and efficient requirements for monitoring corn harvest loss rates in practical operations. There is an urgent need for a method that can directly, rapidly, and accurately count lost kernels during harvest.

Machine vision technology has proven feasible for corn kernel quality grading, mass estimation, and damage detection. Building upon this, researchers have proposed image processing methods for grain harvest loss detection through grayscale conversion, denoising, and segmentation, combined with analysis of grain shape, color, and area attributes. However, inconsistent thresholds and image variations compromise the reliability and stability of counting results. These methods also focus on specific harvest stages (threshing, cleaning) while neglecting direct detection of field surface kernels that truly reflect harvest loss. With advances in deep learning-based object detection technology demonstrating great potential for improving detection accuracy, efficiency, and robustness, Monhollen et al. proposed a target detection network for direct identification of field surface kernels, achieving 82% loss detection accuracy. However, this approach required clearing residues before image acquisition, making the operation cumbersome and limiting counting accuracy based on residue clearance effectiveness.

Existing detection methods still fall short of ideal practical requirements in terms of accuracy and applicability. Deep learning offers the potential to further improve corn harvest loss monitoring precision. Therefore, this study aims to evaluate the feasibility and performance of deep learning technology for direct

counting of real surface kernels, thereby simplifying detection steps and achieving comprehensive improvements in monitoring accuracy and applicability. The main contributions include: (1) collecting real ground surface image data after corn harvest using an RGB camera; (2) constructing both a two-stage detection network (Mask R-CNN) and single-stage detection networks (EfficientDet-D5, YOLOv5-L, YOLOX-L) for corn kernel counting; and (3) analyzing the impact of different ground occlusion levels and kernel aggregation states on final counting performance to identify the optimal deep learning model for infield corn harvest loss kernel counting.

## 2 Materials and Methods

### 2.1 Image Acquisition

Test data were collected from corn experimental fields in Grand Forks County, North Dakota, USA, targeting corn kernels left on the ground during harvest. To avoid straw dust affecting image quality during harvesting, image acquisition was performed after harvesting using an EOS Rebel T7i camera (image resolution 2000$\times$2000, frame rate 6 f/s, auto-exposure and auto-focus modes) mounted at a vertical height of approximately 1.3 m above the ground. The field harvest scenario and image acquisition equipment are illustrated in [Figure 1: see original paper]. Sample collection was conducted on November 7, 2020, from 8:00 to 11:00 AM, yielding 500 images.

### 2.2 Technical Route

This study aimed to analyze collected images using deep learning algorithms to achieve automatic detection and counting of infield corn kernels. The overall technical route, shown in [Figure 2: see original paper], comprises three main components: (1) dataset construction—selecting valid image frames, classifying scenes, and annotating kernels to build a target detection dataset; (2) corn kernel counting—constructing and training different network models for real-time counting of harvest loss kernels; and (3) result analysis—visualizing model training processes, evaluating different methods on the test set, and recommending optimal models.

**2.2.1 Dataset Construction** To ensure image validity for model training and testing, 420 images containing clear corn kernels (totaling 6,773 kernels) were manually selected as the complete dataset. These were randomly divided into 200 training images (1,628 kernels), 40 validation images (224 kernels), and 180 test images (4,921 kernels). The validation set was used to tune hyperparameters and prevent overfitting. Labelme software was employed for data annotation, and datasets were constructed in COCO (Common Objects in COntext) format. To accurately assess model applicability and guidance for harvest loss reduction, test set images were categorized into four types based on field straw occlusion levels and kernel aggregation states [Figure 3: see original paper]: bare ground

(60 images, 1,415 kernels), semi-occluded ground (60 images, 1,372 kernels), fully occluded ground (31 images, 218 kernels), and kernel aggregation (29 images, 1,916 kernels). Bare, semi-occluded, and fully occluded ground refer to images where the ground pixel area ratio to total image pixels falls within (0.85, 1), [0.35, 0.85], and (0, 0.35) intervals, respectively, with discrete kernel distribution. Kernel aggregation typically refers to images with more than 12 adhered or stacked kernels. "Occlusion" in scene naming refers to straw covering the ground surface.

**2.2.2 Method Design** To address limitations of traditional object detection algorithms—such as low recognition accuracy, poor model applicability, and strong feature dependency—this study selected deep learning networks with proven advantages and wide application. This approach simplifies feature design and region selection while reducing the impact of manually constructed features on detection accuracy and efficiency, enabling high-precision real-time corn kernel detection. Since deep learning-based object detection algorithms can be categorized into region proposal-based two-stage methods and regression analysis-based single-stage methods, representative networks from both categories were selected to compare their suitability for kernel detection.

For two-stage methods, existing R-CNN, SPP-Net, Fast R-CNN, and Faster R-CNN networks perform inferior to Mask R-CNN, which possesses dual functions of object detection and segmentation. Moreover, Mask R-CNN functions equivalently to Faster R-CNN when segmentation is not considered. Therefore, Mask R-CNN was selected as the two-stage representative. As the mainstream direction, single-stage methods have seen various network models proposed through improvements in feature extraction networks, multi-scale fusion mechanisms, label assignment strategies, and NMS-Free detectors. The YOLO series is the most classic and efficient, leading to the selection of YOLOv5-L and YOLOX-L for comparing model generalization with and without anchor boxes. Additionally, EfficientDet-D5 was chosen to evaluate the feasibility of balancing detection accuracy and efficiency through unified scaling of network depth, width, and resolution under fixed resource constraints.

**Mask R-CNN** is a two-stage detection network that builds upon Faster R-CNN by introducing a parallel mask branch and ROI Align algorithm to eliminate rounding errors and improve accuracy [Figure 4: see original paper]. Capable of pixel-level object mask output, it is often used as a benchmark for evaluating other networks. After image input, the backbone network generates and fuses feature maps with different scales and semantic information. The Region Proposal Network (RPN) and ROI Align layer identify optimal target candidate regions and resolve misalignment between feature maps and the original image, enabling classification and prediction of target categories, positions, and masks.

**EfficientDet** is a single-stage detection model that proposes a weighted bidirectional feature pyramid network (BiFPN) for fast multi-scale feature fusion, based on neural architecture search FPN (NAS-FPN) and path aggregation

network (PANet) principles [Figure 5: see original paper]. This approach significantly improves detection accuracy and efficiency. EfficientDet primarily consists of a backbone feature extraction network, an enhanced feature extraction network, and a prediction network responsible for multi-scale feature extraction, fusion, and target position/category prediction.

**YOLOv5** inherits the architecture of YOLOv4, dividing input images into S×S grids where the grid containing the target center predicts bounding box position, classification probability, and confidence [Figure 6: see original paper]. After input preprocessing, images enter a backbone network based on Cross Stage Partial Network (CSPNet) and Focus architecture for three-scale kernel feature extraction. Features are then aggregated in the Neck module using FPN and PANet structures before entering the Head module, which uses anchor boxes to produce output results with confidence and box coordinates. To enhance detection performance for occluded and overlapping kernels, GIOU_{Loss} was adopted as the bounding box loss function, and weighted non-maximum suppression was applied for filtering.

**YOLOX** optimizes YOLOv3 by incorporating recent deep learning research achievements and training techniques [Figure 7: see original paper]. The model retains the Darknet53+SPP backbone and FPN architecture from YOLOv3 while improving the input and Head modules. The input employs Mosaic and Mixup data augmentation to enrich detection backgrounds. The Head module introduces Decoupled head, Anchor-free, Multi positives, and SimOTA operations to accelerate model convergence, improve computational efficiency, and alleviate positive-negative sample imbalance, achieving optimal sample matching under global information. The added IoU branch in Decoupled head enhances prediction capability, enabling faster and more accurate bounding box regression.

Among the four selected networks, EfficientDet, YOLOv5, and YOLOX have different model series. To maximize detection accuracy and efficiency under fixed resource constraints, EfficientDet-D5, YOLOv5-L, and YOLOX-L were selected. Network training initialization parameters were set as: initial learning rate 0.003, maximum iterations 100, momentum 0.9, batch size 4, and Adam optimizer. Due to its two-stage nature, Mask R-CNN required 60,000 iterations for convergence.

**2.2.3 Evaluation Metrics** To quantitatively analyze network model performance from perspectives of model complexity, detection effectiveness, and mobile deployment potential, standardized object detection evaluation metrics were adopted:

- **Average Precision (AP)**: Measures comprehensive precision-recall performance. AP is calculated as the mean area under P-R curves across 10 IoU thresholds (linearly increasing from 0.50 to 0.95 at 0.05 intervals) to comprehensively reflect model performance.

- **Model parameters, GFLOPs, frame rate (f/s), and model size**: Assess complexity and mobile applicability.

For kernel counting performance evaluation, the following metrics were used: - **Detection Accuracy (DA)**: $\frac{TP}{TP+FN} \times 100\%$ - **Miss-Detection Rate (MDR)**: $\frac{FN}{TP+FN} \times 100\%$ - **False-Detection Rate (FDR)**: $\frac{FP}{TP+FP} \times 100\%$ - **Precision Detection Rate (PDR)**: $\frac{TP}{TP+FP} \times 100\%$ - **F1-score**: $\frac{2\times PDR\times DA}{PDR+DA} \times 100\%$

where $TP$ is true positives (correctly detected kernels), $FP$ is false positives (false detections), and $FN$ is false negatives (missed kernels).

### 2.3 Experimental Configuration

The hardware platform comprised an Intel Core i5-10400F CPU @ 2.90 GHz, 16 GB RAM, 1 TB hard drive, and an 11 GB NVIDIA GeForce GTX 2080Ti GPU. Programming was conducted using PyCharm 2020 Community Edition. The deep learning framework was PyTorch 1.6, with CUDA 11.3 and cuDNN 8.2.0 for parallel computing and GPU acceleration.

## 3 Results and Discussion

### 3.1 Network Model Training

All four networks were trained and tested using the constructed corn kernel dataset under identical experimental configurations. The loss curves during training are shown in [Figure 8: see original paper]. All models demonstrated strong fitting and generalization capabilities, with similar loss trends: rapid decrease in early training, slight reduction with oscillation in mid-training, and stabilization in later stages, indicating convergence. The fast convergence benefited from the Adam optimizer' s computational efficiency, adaptive learning rate, and insensitivity to gradient scaling.

After confirming model convergence, overall detection performance on the test set was compared to identify preliminary optimal models. Table 1 presents the performance comparison. YOLOv5-L achieved optimal values across all metrics, with an AP of 78.3% and frame rate of 55.55 f/s, representing improvements of 11.8 (28.53), 9.5 (46.7), and 27.4 (48.12) percentage points (frames) over YOLOX-L, Mask R-CNN, and EfficientDet-D5, respectively. This indicates YOLOv5-L can provide faster and more accurate detection results. Structural differences in network architecture led to varying attention to corn images and feature learning capabilities, causing performance disparities. Mask R-CNN' s superior AP over EfficientDet-D5 and YOLOX-L may be attributed to its two-stage architecture generating proposals before regression, though this precision gain consumed more computational resources, resulting in much lower recognition speed than single-stage networks.

Considering model complexity and mobile deployment feasibility, YOLOv5-L

showed significant potential, with GFLOPs, model size, and training time at only 37%, 18.6%, and 5.1% of Mask R-CNN' s maximum values, respectively, far outperforming EfficientDet-D5 and YOLOX-L. While mobile deployment limitations can be addressed through memory expansion, detection performance directly affects final counting accuracy and efficiency. Preliminary ranking for kernel detection is: YOLOv5-L > Mask R-CNN > EfficientDet-D5 > YOLOX-L.

To understand which input image regions contributed most to YOLOv5-L' s recognition decisions, class activation heatmaps were generated for four randomly selected images [Figure 9: see original paper]. Different colors represent different weights, with darker colors indicating higher weights. The model suppressed background regions and focused on learning kernel-related information for recognition decisions. Within kernel regions, the embryo and axis areas contributed most to classification decisions. Compared to kernels annotated as shriveled, severely occluded, broken, or in shadow, plump kernels with full color contributed more to classification decisions. The bias toward higher-weight regions may cause kernels with smaller weights to be missed or falsely detected.

### 3.2 Recognition Result Analysis

The optimal YOLOv5-L model was applied to infer test set images, examining kernel target recognition effectiveness in complex backgrounds. To avoid label interference with result observation, labels were removed from visualizations [Figure 10: see original paper]. The YOLOv5-L model correctly localized nearly all corn kernels across different scenes, demonstrating stable and reliable detection robust to changes in image clarity, illumination, and ground surface conditions. However, performance varied by scene. Recognition accuracy for bare ground and semi-occluded ground exceeded that for other scenes. The slight performance degradation in the former two scenes primarily involved false detection of shriveled kernels, likely due to discrete kernel distribution and fewer occlusions causing the model to overlearn kernel features and overfit, corroborating the finding that shriveled kernel regions with smaller decision weights are prone to false detection.

For fully occluded ground and kernel aggregation scenes, performance losses mainly stemmed from missed detections due to kernel stacking or severe occlusion, and false detections where a single bounding box contained two kernels. Missed detections likely occurred because the critical embryo and axis regions for classification decisions were covered, depriving the model of key feature information. False detections may have resulted from unreasonable threshold settings when filtering low-confidence boxes using non-maximum suppression. Further analysis revealed that network reliability may be enhanced by strong contrast transitions between kernel color/shape features and surrounding regions, providing more useful information for target recognition.

### 3.3 Counting Result Comparison

To verify YOLOv5-L' s reliability and stability for kernel counting, comprehensive evaluation using DA, MDR, FDR, PDR, and F1-score was performed on the test set and compared with YOLOX-L, Mask R-CNN, and EfficientDet-D5 [Figure 11: see original paper]. The four networks exhibited different counting performances. YOLOv5-L achieved optimal DA (90.7%) and MDR (9.3%), outperforming the other three networks by 5–19 percentage points. However, when considering FDR and PDR, the ranking reversed, with YOLOv5-L performing worst and falling 7 percentage points below the optimal Mask R-CNN. For the comprehensive F1-score metric, only YOLOv5-L and Mask R-CNN swapped positions, with values of 91.1% and 91.6%, respectively—a 0.5% difference. The superior YOLOv5-L and Mask R-CNN networks showed different strengths. In practical production, kernel loss counting should follow the principle of "better false detection than missed detection" to avoid greater harvest losses from overestimating combine performance. YOLOv5-L' s lower PDR may be caused by extensive false detection of shriveled kernels. While not counted in this study, detecting shriveled kernels provides potential for improving subsequent loss yield estimation accuracy. Additionally, YOLOv5-L's comprehensive advantages in detection efficiency, model complexity, and application potential offer convenience for addressing practical production requirements and accelerating commercialization.

To clarify each network' s counting performance across scene types, metrics were calculated for each network-scene combination (Table 2 ). YOLOv5-L and Mask R-CNN consistently achieved higher DA, MDR, and F1 values across scenes, with YOLOv5-L showing optimal DA and MDR. YOLOv5-L' s much lower PDR and FDR compared to the other three networks were the main reason its F1 values were slightly lower than Mask R-CNN' s in most scenes (except kernel aggregation). Overall, YOLOv5-L and Mask R-CNN demonstrated superior counting performance across different scenes, with YOLOv5-L' s high DA in bare ground, semi-occluded ground, and kernel aggregation scenes (containing more kernels) being the primary reason for its superior final ranking over Mask R-CNN. All four networks showed significantly lower performance for fully occluded ground and kernel aggregation scenes, providing direction for future research.

## 4 Conclusion

This study developed and evaluated a deep learning-based approach for direct counting of real surface kernels to assess corn harvest losses. The feasibility of deep learning technology for this task was verified through implementation of different target detection networks. Results indicated that among the four models, YOLOv5-L achieved the best performance with DA and MDR of 90.7% and 9.3%, respectively, outperforming Mask R-CNN, EfficientDet-D5, and YOLOX-L. With a processing speed of 55.55 f/s, YOLOv5-L meets the requirements for real-time loss monitoring and rapid harvest quality assessment, making it suit-

able as the core algorithm for developing precision control information systems and detection devices for corn combine harvesters.

Future work will address current limitations through: (1) introducing attention and feature enhancement mechanisms to improve counting accuracy for fully occluded ground and kernel aggregation scenes; (2) investigating the influence of kernel color, shriveling degree, and size variations on recognition, and developing secondary discrimination algorithms to reduce false detection rates; (3) incorporating more diverse data categories (different blur levels, brightness, angles, corn varieties) and expanding the training dataset to improve model robustness and stability; and (4) addressing dust issues during harvesting by developing protective enclosures to enable real-time loss detection during operation.

## References

[1] ZHANG W, LIU X, LI Q, et al. General situation analysis of world corn production and trade[J]. World Agriculture, 2014(3): 111-114.

[2] WU Y, LI X, MAO E, et al. Design and development of monitoring device for corn grain cleaning loss based on piezoelectric effect[J]. Computers and Electronics in Agriculture, 2020, 179(12): ID 105793.

[3] XU L, WEI C, LIANG Z, et al. Development of rapeseed cleaning loss monitoring system and experiments in a combine harvester[J]. Biosystems Engineering, 2019, 178: 118-130.

[4] VALIENTE-GONZÁLEZ J M, ANDREU-GARCÍA G, POTTER P, et al. Automatic corn (Zea mays) kernel inspection system using novelty detection based on principal component analysis[J]. Biosystems Engineering, 2014, 117: 94-103.

[5] ORLANDI G, CALVINI R, FOCA G, et al. Automated quantification of defective maize kernels by means of multivariate image analysis[J]. Food Control, 2018, 85: 471-480.

[6] LI X, DAI B, HONG S, et al. Corn classification system based on computer vision[J]. Symmetry, 2019, 11(4): 591-591.

[7] ZHANG T, ZHAO D, ZHOU T. Application of image processing on combine harvester attachment loss[J]. Journal of Agricultural Mechanization Research, 2009, 31(4): 70-72.

[8] XIN B, WU T, CHEN C, et al. A real-time online detection method for grain harvesting and cleaning loss based on image processing: CN107123115A[P]. 2017-09-05.

[9] WELLINGTON C K, BRUNS A J, SIERRA V S, et al. Grain quality monitoring: US10664726B2[P]. 2017-05-16.

[10] SUN H, LI S, LI M, et al. Research progress of image sensing and deep learning in agriculture[J]. Transactions of the CSAM, 2020, 51(5): 1-17.

[11] MONHOLLEN N S, SHINNERS K J, FRIEDE J C, et al. In-field machine vision system for identifying corn kernel losses[J]. Computers and Electronics in Agriculture, 2020, 174: ID 105496.

[12] LYU L, CHENG H, ZHU H, et al. Progress of research and application of object detection based on deep learning[J]. Electronics & Packaging, 2022, 22(1): 72-80.

[13] DONG L, ZENG Z, YI S, et al. Research on a YOLOv5-Based remote sensing image target detection[J]. Journal of Hunan University of Technology, 2022, 36(3): 44-50.

[14] XIE F, ZHU D. Survey on deep learning object detection[J]. Computer Systems & Applications, 2022, 31(2): 1-12.

[15] KOU D, QUAN J, ZHANG Z. Research on progress of object detection framework based on deep learning[J]. Computer Engineering and Applications, 2019, 55(11): 12-18.

[16] BAO X, WANG S. Survey of object detection algorithm based on deep learning[J]. Transducer and Microsystem Technologies, 2022, 41(4): 5-9.

[17] HE K, GKIOXARI G, DOLLÁR P, et al. Mask r-cnn[C]//The IEEE International Conference on Computer Vision. Piscataway, New York, USA: IEEE, 2017: 2961-2969.

[18] TAN M, PANG R, LE Q. Efficientdet: Scalable and efficient object detection[C]//The IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, New York, USA: IEEE, 2020: 10781-10790.

[19] BOCHKOVSKIY A, WANG C Y, LIAO H Y M. YOLOv4: Optimal speed and accuracy of object detection[J/OL]. arXiv:2004.10934[cs.CV], 2020.

[20] GE Z, LIU S, WANG F, et al. YOLOx: Exceeding yolo series in 2021[J/OL]. arXiv:2107.08430, 2021.

*Note: Figure translations are in progress. See original paper for figures.*