Journal of Computing and Electronic Information Management
ISSN: 2413-1660 | Vol. 15, No. 3, 2024

Research Overview of YOLO Series Object Detection Algorithms Based on Deep Learning

Bo Chen

Computer science, south China business college, Guangzhou, Guangdong, 510000, China
1541161220@qq.com

Abstract: In the rapid development of deep learning, YOLO, as the first popular single-stage object detection model, has sparked an innovative storm in the computer vision community with its remarkable architecture and innovative concepts, marking a significant leap forward in object detection technology. Today, it is not only regarded as a milestone in this field but also sets an unparalleled example in the pursuit of the perfect fusion of detection speed and accuracy. YOLO has been widely applied in various fields such as agriculture, industry, pedestrian detection, and more. This research project will first introduce traditional object detection methods, then analyze object detection based on deep learning, and subsequently elaborate on the fundamental concepts of YOLO. It will systematically sort through the YOLO family and its significant improvements. Finally, based on different improvement strategies or application scenarios, the YOLO algorithm will be systematically classified and summarized.

Keywords: YOLO; Target detection; Algorithmic history of YOLO; Deep learning; Convolutional neural networks.

1. Introduction

In recent years, the domestic computer vision field has carried out in-depth exploration and research on the core and very challenging topic of target detection. Thanks to the rapid progress of deep learning technology, the YOLO series of algorithms have won wide attention and in-depth discussions in both academic and industrial circles in China. As an important cornerstone of computer vision technology, the development and improvement of target detection is of great significance in promoting technological innovation and application expansion in related fields. For example, a research team from China University of Mining and Technology proposed a multi-scale feature fusion-based insulator defect degree detection network, which adopts the reconfigured ResNeSt50 as the feature extraction network, and adds the inverse convolution-based feature fusion module and RFB module, which effectively improves the accuracy and efficiency of insulator defect detection.

Geoffrey Hinton first introduced the concept of deep belief network in 2006 [2], which consists of a series of constrained Boltzmann machines [3]. Hinton et al. applied this method to the experiments of handwritten fonts recognition, and achieved good results. Pre-training, as an integral part of the deep belief network, aims to initially set the network parameters to a near-optimal starting state. Subsequently, the entire network is meticulously adjusted through fine-tuning techniques to further promote the optimized performance of the network. This process significantly accelerates the training efficiency of the neural network and effectively alleviates the problem of gradient vanishing, which often occurs in backpropagation algorithms, laying a solid foundation for the stable training and performance improvement of the neural network.

In 2012, AlexNet based on deep learning convolutional neural network CNN won the champion in ImageNet image recognition competition, and object detection algorithms based on deep learning became a research hotspot in this field. [4]

Facebook Artificial Intelligence Institute in conjunction with Allen Institute for Artificial Intelligence and the University of Washington proposed the first real-time single-stage object detector in the deep learning era (You Only Look Once, YOLO) at the 2016 International Conference on Computer Vision and Pattern Recognition, CVPR [5]

2. Deep Learning Based YOLO Target Detection Algorithm

2.1. Difference between traditional and deep learning target detection

Traditional target detection usually divides the target detection process into multiple steps: candidate region generation, feature extraction, and classification, and basically relies on hand-designed features that may not be generalizable and need to be adjusted for different datasets. Traditional methods are computationally intensive and slow, and usually cannot meet the requirements of real-time detection. The detection accuracy of traditional methods in complex scenes is often inferior to deep learning methods.

The YOLO deep learning target detection method treats target detection as a regression problem, directly predicting bounding boxes and categories through a single neural network. Deep learning methods use convolutional neural networks (CNNs) to automatically learn image features without manually designing features, which provides greater adaptability and generalization capabilities. YOLO greatly improves detection speed by dividing the entire image into SxS grids and directly predicting bounding boxes and categories on these grids.

Therefore, deep learning-based target detection methods show higher efficiency and accuracy in most applications and have become mainstream methods in the field of target detection.

2.2. YOLO Basic Idea

The basic idea of YOLO (You Only Look Once) is to divide the input image into an S×S grid system, where each grid cell is regarded as a potential detection unit specifically responsible for detecting those target objects whose centroids

happen to fall inside it. This detection mechanism operates in two main phases: training and inference. In the training phase, the centroid locations of the target objects are determined based on the real bounding box coordinates precisely labeled manually; while in the inference phase, when the model is actually used, these centroids are automatically predicted by the model itself through regression analysis. Once the center point of a target is determined to be located in a grid cell, the cell assumes the responsibility of detecting the target, including predicting the target's category and the exact target bounding box location.

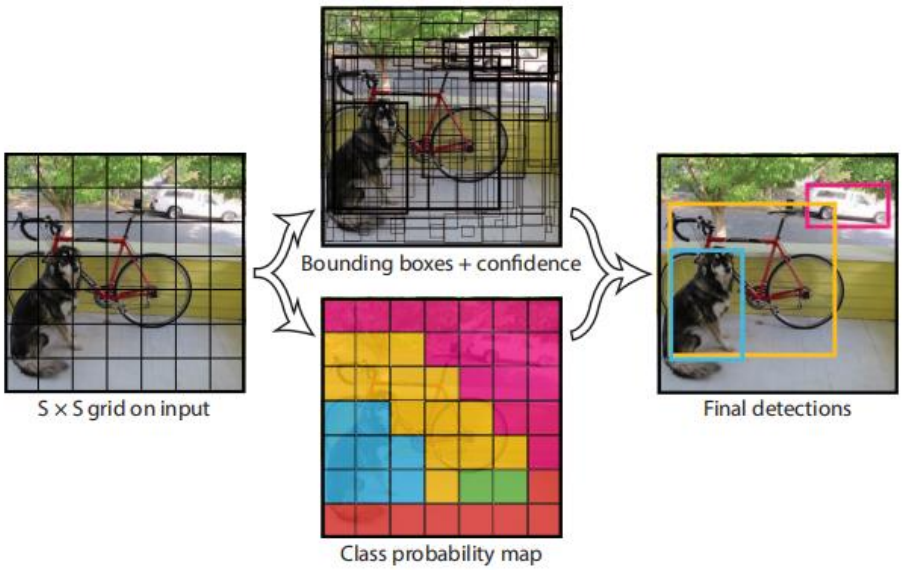

Figure 1. YOLO algorithm for image segmentation

Taking YOLOv1 in Figure 1 [5] as an example, YOLOv1 segments the input image into a 7 × 7 square network, where each cell predicts 2 bounding boxes, and detects a recognition category of 20, then the final network output is 7 × 7 × 30.[6]

2.3. YOLO Family and Important Improvements

The YOLO family consists of YOLOv1-v4, YOLOv5, YOLOR, and YOLOX, among others. As shown in Figure 2.

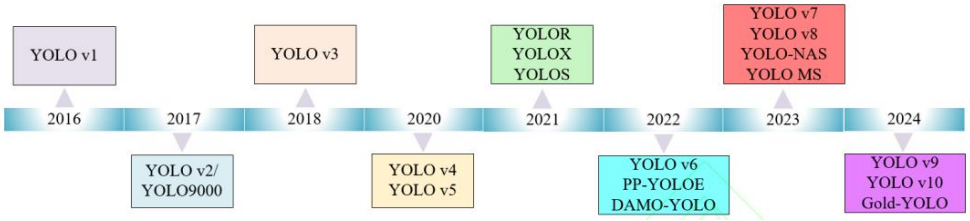

Figure 2. History of the YOLO family [1]

2.3.1. YOLOv1

The YOLOv1 algorithm begins by dividing each of the input images in the form of an S×S grid, which is fed as input to the neural network. The network then passes through a series of convolutional layers to extract image features, followed by a pooling layer for dimensionality reduction, and finally a fully connected layer to process the information in each sub-grid region. Each subgrid is independently responsible for detecting and predicting the class of the target that falls within it, a confidence level indicating the presence or absence of that target, and the exact location of the target bounding box. After completing these calculations, the fully connected layer is able to directly output the detected target categories and their corresponding bounding boxes. Finally, a non-maximal suppression (NMS) technique is used to eliminate those overlaps due to multiple detected bounding boxes, ensuring that each target is labeled only once.The network structure of YOLOv1 is illustrated in Figure 3.

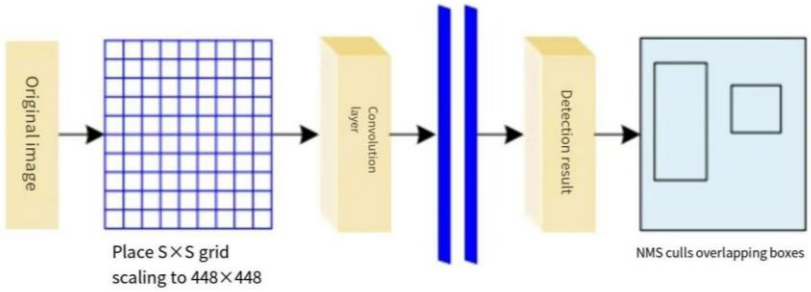

Figure 3. Network structure diagram of YOLOv1[15]

In the training phase of YOLO v1, the system employs a fixed image input size, and the composition of its loss

function L consists of three key components: the localization loss, which evaluates the deviation of the predicted bounding box from the actual target's bounding box; the confidence loss, which measures the confidence accuracy of the detection box in containing the target; and the category loss, which evaluates the correctness of the target's classification within the detection box. The calculation of these losses and the specific parameters involved are shown in Figure 4.

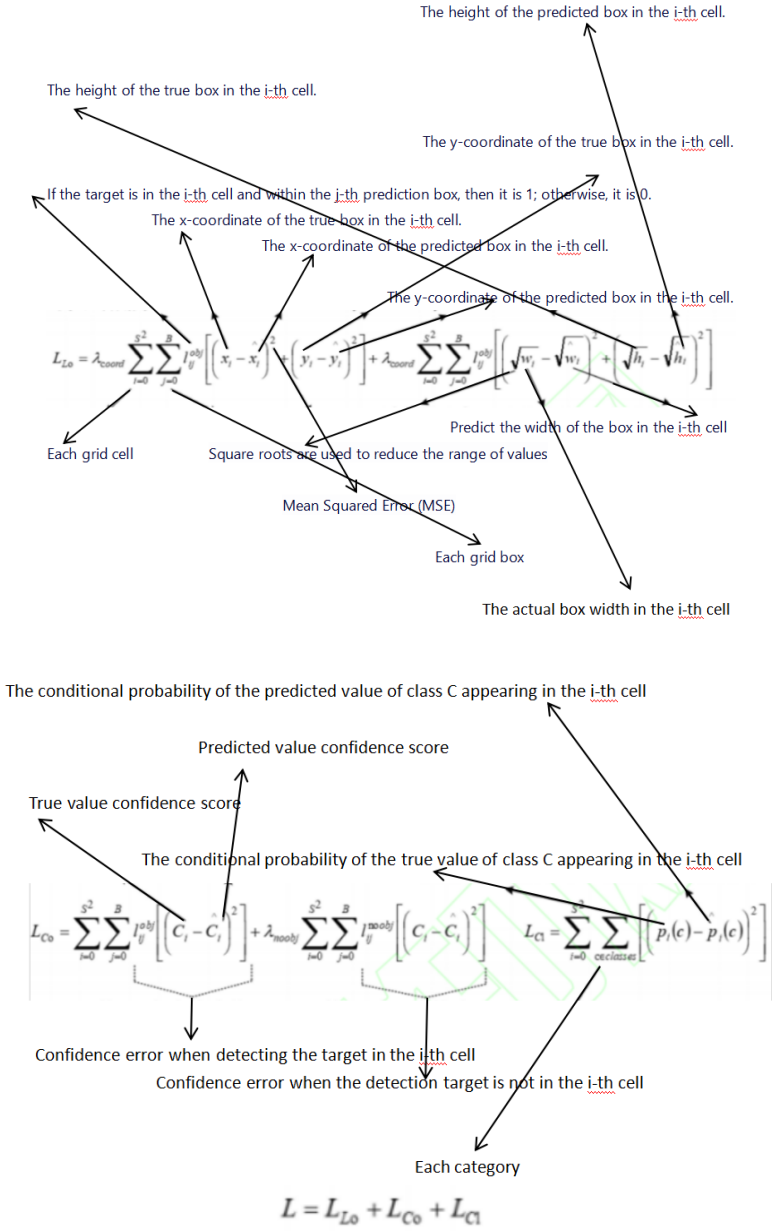

Figure 4. Detailed parameter explanation of YOLO v1 loss function [1]

2.3.2. YOLOv2 (YOLO9000)

YOLOv2 builds on YOLO v1 and constructs a new backbone structure called DarkNet-19, which contains 19 convolutional layers, 5 maximal pooling layers and 1 average pooling layer. [1] Compared with YOLO v1, YOLO v2 improvements mainly include:

1)Batch Normalization. Batch normalization removes the need for additional regularization while producing noticeable improvements in convergence. We gain more than 2% improvement in mAP by using batch normalization to all of the convolutional layers in YOLO. The model is further regularized via batch normalization. We can eliminate dropout from the model without overfitting by using batch normalization.

2)High Resolution Classifier. First, we fine-tune the classification network using ImageNet for 10 epochs at the full 448 × 448 resolution.

This offers the network enough time to modify its filters so they function better on input with higher resolution. Following detection, we adjust the resulting network further. With this high-resolution classification network, our mAP increases by over 4%.

3)Dimension Clusters. We use k-means clustering on the training set bounding boxes to automatically find suitable priors, removing the need for manual prior selection. We use k-means clustering on bounding box dimensions in order to obtain high-quality priors for our model. The average IOU that results from different options for k is displayed in the left image. A good trade-off between recall and model complexity is found at k = 5.

4)Multi-Scale Training.448 x 448 is the input resolution used by the original YOLO. Anchor boxes were added, and

we adjusted the resolution to 416 x 416. On the other hand, our model can be scaled instantly because it simply makes use of convolutional and pooling layers. We train the model to be resilient to YOLOv2 operating on photos of varying sizes.

5.Hierarchical classification. ImageNet labels are pulled from WordNet, a language database that structures concepts and how they relate. [16]

YOLOv2 is cutting edge and outperforms previous detection methods on a range of datasets. Moreover, it may operate at many image proportions to offer a seamless balance between accuracy and speed.

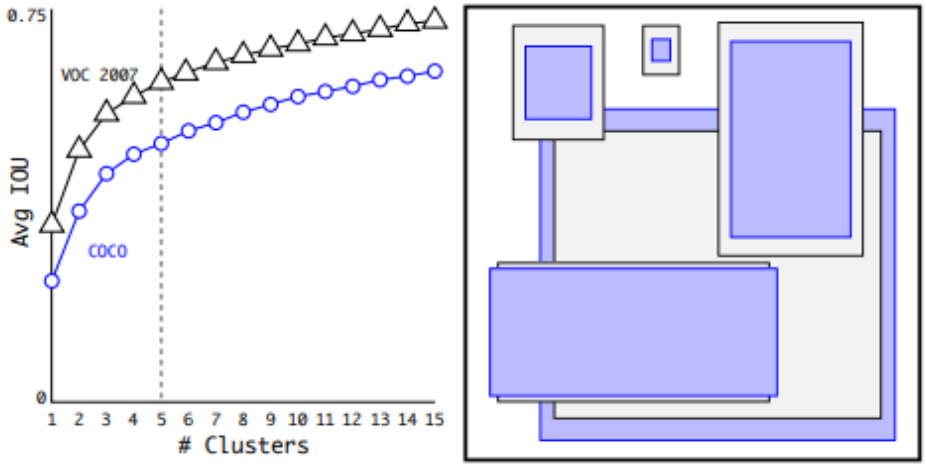

Figure 5. Clustering box dimensions on VOC and COCO [8]

2.3.3. YOLOv3

YOLOv3 has made several improvements over YOLOv1 and YOLOv2, resulting in a significant increase in its performance in target detection tasks. The following are the main improvements of YOLOv3 over YOLOv1 and YOLOv2:

1). Multi-Scale Prediction: YOLOv3 significantly enhances its ability to deal with targets of different sizes by introducing a multi-scale prediction mechanism into its architecture, in particular, target detection on three different levels of Feature Maps. This improvement brings a substantial increase in the recognition efficiency of small targets in particular. In contrast, YOLOv1 and YOLOv2, which are limited by the design of only one output scale, do not perform well in dealing with targets of multiple scales, making it difficult to effectively deal with multi-scale challenges in complex scenes.

2). Residual Network: YOLOv3 uses Darknet-53 as the core backbone of its powerful framework, which consists of 53 layers of carefully designed convolutional layers stacked together to form a neural network structure of considerable depth. Darknet-53 cleverly incorporates the design concept of residual blocks, which not only allows the network to expand to a deeper level, but also ensures that the stability of the gradient can be maintained during the network deepening process, thus avoiding the common problem of gradient disappearance in traditional deep networks. Darknet-53 cleverly incorporates the concept of Residual Blocks, which not only allows the network to go deeper, but also ensures that the gradient stability is maintained as the network deepens, thus avoiding the problem of gradient vanishing that is commonly found in traditional deep networks. network structure as its architectural foundation.

3). Class Prediction: YOLOv3 revolutionizes the class prediction strategy by abandoning the traditional Softmax classifier to determine the class of each bounding box individually, and instead employs a set of independent binary classifiers, each of which is responsible for evaluating whether or not a particular class exists in a given bounding box. This shift gives YOLOv3 the ability to recognize multiple objects of different categories within the same spatial location at the same time, making it more adaptable to multi-label classification tasks in complex scenarios.YOLOv1 and YOLOv2 are limited by the use of the Softmax function, where each grid cell can only be assigned to one of the most probable categories, and are unable to efficiently deal with the case of multiple categories coexisting in the same region.

YOLOv3 stands out from the crowd in many environments with limited computational resources by virtue of its superior performance and accuracy, making it the preferred model in the field of target detection. Its fast inference speed and excellent detection performance make it shine in many real-world application scenarios such as drone surveillance and autonomous driving. As an important milestone in the development of YOLO series, YOLOv3 not only stabilizes its position, but also lays a solid foundation for the iterative upgrading of subsequent versions. Its key technologies, such as network architecture design, loss function adjustment, multi-scale detection strategy, and anchor frame mechanism, point the way for subsequent model optimization and performance improvement.

2.3.4. YOLOv4

The design goal of YOLOv4 [9] is to make the training process of large-scale target detection models easier and more efficient, and to improve their usefulness in practical applications. YOLOv4 attaches importance to the following points in its design: 1. Optimization of Speed and Accuracy: optimize the accuracy while ensuring the speed of inference.2. Hardware Compatibility: able to run efficiently on commonly used hardware (e.g., GPU) Efficient operation on commonly used hardware (e.g., GPUs).3. Ease of training: does not require large-scale cluster computing resources for training.

YOLOv4 also introduces two basic concepts:

1)Bag of Freebies: The ability to improve model performance in the training phase without increasing computational overhead in the inference phase. For example: 1.1 Mosaic data augmentation (blending multiple images together) allows the model to learn richer contextual information. 1.2 Self-adversarial training, which makes the

model more robust when encountering noisy or different inputs by augmenting the input images. 1.3 Reducing the risk of overfitting the model by smoothing the labels. 1.4 DropBlock regularization. 1.4 DropBlock regularization: a regularization technique similar to Dropout, especially for convolutional neural networks.

2)Bag of Specials: These techniques may add some computational overhead in the inference phase, but can significantly improve the performance of the model. For example: 2.1 Mish Activation Function: a new activation function with better gradient flow properties than the traditional ReLU and Leaky ReLU.2.2 CSPNet (Cross-Stage Partial Networks): an improved backbone network design that can improve the accuracy of the model while reducing the computational effort.2.3 SAM (Spatial Attention Module) and PAN (Path Aggregation Network): modules for enhanced feature fusion, which improves the efficiency of utilizing features at different scales.

YOLOv4 adopts CSPDarknet53 as its backbone network, which is improved based on Darknet53 of YOLOv3. CSPDarknet53 introduces Cross-Stage Partial Networks (CSP, Cross-Stage Partial Networks), which improves the ability of gradient information flow and effectively reduces the number of the number of model parameters, while enhancing the feature extraction capability.

2.3.5. YOLOv5

YOLOv5 focuses on providing a lightweight and efficient target detection model that can excel in embedded devices and resource-constrained environments. YOLOv5 was developed with the goal of achieving a balance between speed and accuracy, and to significantly increase inference speed by improving the architecture and optimizing the code. [17]

The main features of YOLOv5 are:

YOLOv5's architecture has been optimized to reduce model size for running on edge and mobile devices.2. YOLOv5 is implemented entirely in PyTorch, which makes training, tuning, and deploying the model much easier, whereas YOLOv4 used a Darknet-based C/C++ implementation.3. YOLOv5 can, on a single GPU achieve inference speeds of up to 140 FPS (frames per second), outperforming previous generations of YOLO models in terms of speed. This performance makes YOLOv5 ideal for real-time application scenarios.4. YOLOv5 introduces automatic Anchor Box generation, which automatically adjusts the size and scale of Anchor Boxes based on the dataset, which improves the model's ability to detect targets of different sizes.5. YOLOv5 integrates multiple data enhancement methods (e.g., Mosaic data enhancement) and optimizes the training process.) and optimizes the training process. It defaults to using strategies such as multi-scale training and mixed-precision training, which can effectively improve the model's generalization ability and training speed.

YOLOv5 offers models in several sizes (e.g., YOLOv5s, YOLOv5m, YOLOv5l, YOLOv5x), allowing users to make trade-off choices between performance and computational resources based on their needs.

2.3.6. YOLOR

The YOLOR [7] model is motivated by the concept of multi-task learning and is built on top of the YOLO v4 network architecture. YOLOR is an extended version of the YOLO family that introduces the concept of uniform representation learning on top of YOLO.YOLOR utilizes both explicit and implicit representations to improve the performance of target detection. Explicit representation usually refers to the feature representation learned by traditional deep learning models through convolutional neural networks (CNNs), while implicit representation is a feature enhancement strategy learned through a new optimization approach.

2.3.7. YOLOX

YOLOX [10] improves on the following: 1. YOLOX abandons the anchor-based detection mechanism commonly used in the YOLO family in favor of a new anchor-free (anchor-free) design. This has the advantage of simplifying the model architecture, reducing the need for a priori knowledge, and decreasing computational complexity.2. YOLOX introduces more advanced data augmentation strategies, such as MixUp and Mosaic data augmentation methods, which can better enhance the generalization ability of the model and make better use of the data during training. 3. YOLOX adopts a decoupled head design, which can optimize the classification and regression tasks separately. This design can optimize the classification and regression tasks separately, thus improving detection accuracy and speed.4. YOLOX uses SimOTA (an optimized label assignment strategy) to improve training efficiency and model detection by dynamically assigning positive and negative samples.5. YOLOX performs well in several benchmarks, especially on the COCO dataset where it significantly outperforms both YOLOv4 and YOLOv5. It offers faster inference speed and higher accuracy, making it a very competitive choice for real-world applications.

2.3.8. YOLOS

With the rise of the Transformer architecture, the scope of its application has expanded rapidly, from its initial focus on language processing and audio analysis, to its successful foray into computer vision, which led to the birth of Vision Transformer (ViT), a technology that now occupies a central position in the deep learning field, responsible for executing a number of key tasks. The combination of YOLO and ViT technology has become an unstoppable trend in the evolution of technology, and it is a sign that the two will work together to usher in a new era of more efficient and accurate vision processing.

YOLOS [21] employs a direct sequence modeling approach to serialize the pixel values or features of an image into a one-dimensional sequence, which is then processed by a Transformer. This approach avoids image chunking and complex feature extraction steps and simplifies the model structure. The design concept of YOLOS is straightforward and does not rely on complex a priori knowledge (e.g., anchor frames and feature pyramid networks), and directly treats the target detection problem as a sequence-to-sequence mapping problem. By rethinking Transformer's application in vision, YOLOS proposes a new framework for target detection that not only simplifies the structure but also improves the model performance, providing new directions and possibilities for future vision research.

2.3.9. YOLOv6

YOLOv6 [18] provides better performance than its predecessor YOLO model while maintaining real-time performance. Through a series of improvements and optimization strategies, YOLOv6 outperforms other existing single-stage target detectors on COCO and other benchmark datasets.

The network architecture of YOLOv6 has been improved in various ways from the previous YOLO series, including the

following components.1. YOLOv6 uses a lightweight backbone network design optimized based on the Pytorch framework to reduce computational complexity and model size for more efficient inference on a variety of hardware platforms.2. YOLOv6 introduces a more efficient detection head design, which can better handle targets of different scales and shapes and improve the detection accuracy of small targets.3. Improved feature pyramid network structures (e.g., PANet) can more efficiently fuse low-level and high-level feature information, which improves the multi-scale performance of target detection.

YOLOv6 employs a number of new optimization strategies to improve model performance. These include:1. Accelerating the training process through mixed-precision training, while reducing the memory footprint and improving the convergence speed of the model.2. YOLOv6 introduces an improved Mosaic data enhancement method and other data preprocessing strategies to enhance the model's generalization ability.3. Adopting a Dense Anchor Box-based optimization method, which allows the Anchor Box to better adapt to the different sizes and shapes of the targets, significantly improving the accuracy of detection.

YOLOv6, an important member of the YOLO family, has successfully improved the accuracy and speed of target detection while maintaining real-time performance. Its design and optimization strategies are particularly suitable for industrial application scenarios, and it is able to cope with diverse detection needs.

2.3.10. PP-YOLOE

PP-YOLOE is a high-efficiency target detection model developed by Baidu using its own deep learning platform PaddlePaddle, hence the name PP-YOLOE. The model follows the advanced design concepts of the v4 and v6 versions of the YOLO series, and has been comprehensively optimized and enhanced based on the solid foundation of YOLO v3, aiming to improve detection accuracy and speed to meet the needs of more practical application scenarios. This model follows the advanced design concepts of the v4 and v6 versions of the YOLO series, and is fully optimized and enhanced on the solid foundation of YOLO v3, aiming to improve the detection accuracy and speed to meet the needs of more practical application scenarios. [1]

PP-YOLOE introduces dynamic convolution, which is able to adaptively select the optimal convolution kernel and improve the model's ability to adapt to different inputs. This mechanism enhances the flexibility of detection and effectively improves the detection performance. Through the efficient aggregation and distribution mechanism, PP-YOLOE is able to better fuse multi-scale features and improve the performance of the model at various object scales. Especially in scenes with a mixture of small and large targets, the detection accuracy is higher.PP-YOLOE adopts an optimized label assignment strategy (e.g., OTA - Optimal Transport Assignment) to more accurately assign targets to each prediction frame, improving the training efficiency and detection results.PP-YOLOE introduces a variety of data augmentation techniques (e.g. Mosaic, MixUp, etc.), which can improve the generalization ability of the model with limited data sample size.

2.3.11. DAMO-YOLO

DAMO-YOLO is an improved version of the YOLO series of models proposed by Alibaba Dharma Institute, with a focus on optimizing performance and inference speed, especially in real-world application scenarios.DAMO-YOLO adopts more efficient lightweight backbone networks, such as PP-LCNet, which can reduce the number of parameters while maintaining high accuracy. DAMO-YOLO uses Neural Architecture Search (NAS) methodology to automate the design of the network structure, enabling the model to achieve a better balance between accuracy and speed.

DAMO-YOLO introduces a task-aligned label assignment strategy, which dynamically adjusts the allocation of positive and negative samples according to the needs of the target detection task.DAMO-YOLO adopts an Anchor-free strategy, which eliminates the design of anchor points, simplifies the training process, and improves the inference speed and the detection accuracy.DAMO-YOLO synthesizes a variety of optimization techniques, such as data augmentation, model compression, knowledge distillation and other optimization techniques to improve the generality and deployment flexibility of the model.

2.3.12. YOLOv7

YOLOv7 [19] is a major release of the YOLO family, which significantly improves the performance of target detection by introducing a trainable “Bag-of-Freebies” strategy. yolov7 outperforms the original real-time target detection model on several benchmark datasets, reaching a new performance standard. YOLOv7 outperforms the original real-time target detection model on several benchmark datasets and achieves a new performance standard.

YOLOv7 further optimizes the architecture of the YOLO series by adopting a neural network structure that was not previously available to improve the performance of target detection. The main architectural improvements include: 1. YOLOv7 introduces the E-ELAN module, which enhances the multi-scale representation of features by introducing a more efficient feature fusion strategy, and E-ELAN uses a deeper feature aggregation mechanism to improve the expressiveness and computational efficiency of the network. 2. YOLOv7 introduces a new combination of convolutional modules and an improved design of activation functions to improve the nonlinear representation and feature capture of the model. 3. model's nonlinear expressiveness and ability to capture features.3. A more lightweight and efficient detection head design is introduced to better handle target detection tasks of different scales and complexity.

YOLOv7 proposes unprecedented “Bag-of-Freebies” strategies, which can improve the generalization ability and accuracy of the model during the training process without increasing the inference time. The main innovations include: 1. Dynamic label assignment: a new target label assignment strategy that dynamically adjusts the label assignment according to the difficulty of the target to improve the training effect and generalization ability of the model. 2. YOLOv7 uses an adaptive anchor frame optimization algorithm that dynamically adjusts the size and shape of the anchor frames according to the characteristics of different datasets to improve the accuracy of the target detection. 3. Introducing a number of new data enhancement methods, such as self-supervised learning and hybrid data enhancement strategies, which make the data enhancement process more adaptive and intelligent.

By introducing a trainable “Bag-of-Freebies” strategy and an improved network architecture design, YOLOv7 successfully achieves a balance between speed and accuracy, and establishes an unprecedented performance benchmark in the field of target detection. Its innovative design provides a strong guidance and reference for later target detection

models.

2.3.13. YOLOv8

YOLOv8 [20] is an Ultralytics release that inherits the efficient speed and accuracy of the YOLO series and adds some new features: 1. YOLOv8 is architecturally optimized to enable it to better capture and process complex feature information. 2. Support for OBB detection: YOLOv8 is able to output a bounding box that includes position, size, and angle information to directed bounding boxes, enabling it to handle more detection tasks in real-world scenarios.

The steps to train an OBB model using YOLOv8 with Roboflow and Ultralytics include:

1)Data Preparation: Collecting and labeling data: select an appropriate labeling tool (e.g., Roboflow Annotate) to create oriented bounding box labels for the target objects. The OBB labels contain the centroid coordinates, width, height, and the rotation angle.1.2 Dataset management: process the dataset using Roboflow's data management tools, such as data augmentation, dividing the training set and validation sets, etc.

2)Model Configuration: Use the Ultralytics YOLOv8 configuration file to set up the model's parameters, such as input size, batch size, learning rate, etc. 2.2 For the OBB task, specific configuration options need to be adjusted to ensure that the model correctly predicts rotation angle information.

3)Model Training: Use a Python script to run the training command for the YOLOv8 model and specify the dataset path and configuration file. During the training process, Ultralytics provides tools to monitor the model's performance metrics such as loss value, mAP, etc. in real time.3.2 The model training process may take from a few hours to a few days depending on the size of the dataset and the strength of the computational resources.

The trained YOLOv8 OBB model can be deployed into various real-time detection systems, such as self-driving vehicles. These application scenarios usually require high-precision target localization and rotation information, and the OBB model has significant advantages in this regard.

2.3.14. YOLO-NAS

YOLO-NAS automates the design of network structures using neural architecture search techniques, which allows the model to generate optimal architectures for different hardware and application scenarios. NAS is able to better balance the accuracy of the model with the speed of inference, avoiding the limitations that occur when manually designing models. Compared to other YOLO versions, YOLO-NAS performs better when dealing with complex backgrounds and multi-scale targets, and its performance optimization makes it more suitable for real-time applications on edge devices. YOLO-NAS employs more advanced data augmentation and regularization techniques (e.g., MixUp, CutMix, Mosaic, etc.), which significantly improves the model's generalization ability. These techniques are effective in preventing model overfitting and provide more robust performance especially on small sample datasets. YOLO-NAS employs an Anchor-free design, i.e., it does not require predefined anchor frames, which simplifies the training process, improves the training efficiency, and reduces the need for hyper-parameter tuning.

2.3.15. YOLOMS

YOLOMS [24] is a product of further development based on YOLO v8, and its core design essence is rooted in an in-depth study of the specific impact of different sizes of convolutional kernels on the target detection efficacy at each level. To achieve this goal, YOLOMS innovatively incorporates MS-Block into its backbone network structure, which employs a hierarchical feature fusion strategy to extract features in a parallel manner by constructing multiple branches. In particular, MS-Block is integrated with an inverted bottleneck block containing deep convolution, which effectively optimizes the use of large convolutional kernels and promotes efficient feature extraction.

To complement the design philosophy of MS-Block, YOLOMS also elaborates a heterogeneous selection protocol that flexibly employs multiple sizes of convolutional kernels at different stages of network development. This strategy aims to capture and fuse more extensive and detailed multi-scale feature information to further enhance the comprehensiveness and accuracy of target detection. YOLOMS focuses on improving multi-scale feature representation learning to enhance the model's performance on both small and large target detection through more effective feature fusion and enhancement strategies. The multi-scale learning strategy enables the model to better cope with target objects of different scales and densities in complex scenes, improving the accuracy and robustness of detection.

2.3.16. YOLOv9

YOLOv9 [23] introduces an innovative gradient control mechanism called “programmable gradient information”. This approach allows the user or algorithm to autonomously adjust the direction and strength of the model's gradient updates to optimize for the goals of a particular task or dataset. This approach helps the model learn specific target features more accurately, thus improving target detection performance. In YOLOv9, the model can dynamically adjust the weights and structure of the feature extraction layer based on the feedback of the gradient information. This enables the model to capture important features more efficiently and ignore irrelevant noise under different data distributions and complex scenarios. YOLOv9 optimizes inference efficiency through quantitative awareness training (QAT), pruning, and knowledge distillation, making the model more suitable for running on low-resource devices while maintaining high performance.

2.3.17. YOLOv10

Tsinghua University's latest open source YOLOv10 [22] series of models, covering multiple versions from n to x, forms a complete scale spectrum, of which YOLOv10n is the most streamlined version, while YOLOv10x is the most powerful version, which is designed to flexibly adapt to the needs of diversified application scenarios. This series of models are highly consistent in architecture, built on three core components: Backbone, Neck, and Detection Head, with the main difference being the depth and breadth of model design. The Backbone module focuses on the efficient extraction of features, and its core components include Conv, C2f, and SPPF, which work together in the initial processing of the feature information. The Neck part skillfully combines the output of the Backbone with the multi-scalable output from the Backbone, and the Neck part with the Neck. The Backbone module focuses on efficient feature extraction, and its core components include Conv, C2f, and SPPF, which work together for the initial processing of feature information, while the Neck part skillfully integrates the multi-scale feature maps from the Backbone output to promote the organic combination of shallow detail information and deep semantic information.

In terms of optimization strategy, YOLOv10 introduces CIoU as a regression loss function, which not only measures

the degree of overlap between the predicted and real frames, but also additionally takes into account the deviation of the centroid distances and the matching of the aspect ratios, which significantly improves the positioning accuracy. Meanwhile, by implementing the dual-label assignment strategy, the label assignment process is effectively simplified and the reliance on the traditional non-maximum suppression (NMS) is eliminated, an improvement that directly reduces the computational overhead during model inference and accelerates the overall detection speed.

2.3.18. Gold-YOLO

Gold-YOLO significantly enhances the fusion capability of multi-scale features by introducing the innovative GD (Gather-and-Distribute) mechanism. The GD mechanism realizes the unified aggregation and fusion of features from different layers on the global view through convolution and self-attention operations, and injects global information into different layers, thus constructs a more adequate and efficient information interaction fusion mechanism. This mechanism enables Gold-YOLO to utilize multi-scale features more effectively when detecting objects of different sizes and improve the accuracy of detection.

Additionally Gold-YOLO has implemented MAE (Masked Autoencoder for self-supervised learning) style pre-training for the first time in the YOLO series. MAE pre-training is a self-supervised learning method that learns the intrinsic of the input data by training a model to reconstruct randomly masked portions of the input data to representation. This pre-training method improves the learning efficiency and accuracy of the model, allowing Gold-YOLO to converge faster and improve performance in target detection tasks.

2.4. YOLO algorithm and deep learning in different scenarios

2.4.1. Agriculture

We have developed a lightweight model, YOLOv4-CA, based on the YOLOv4 architecture, for real-time detection of apples.By deploying this optimized model on an embedded system platform, we have achieved a significant improvement in the detection speed, and the model's generalization ability, i.e., detection accuracy, has also been enhanced under different environments and conditions. detection accuracy, is also enhanced. This result provides strong technical support for fast and efficient detection of apples and other fruits. [12] CNN convolutional neural network is good at processing image data and extracting features in the image through convolutional layers. By regularly taking images of farmland and analyzing them with CNN, the growth status of crops can be monitored in real time, and problems such as abnormal growth and nutritional deficiencies can be detected in a timely manner.

2.4.2. Traffic

We made innovative improvements to YOLOv4 using spatial pyramid pooling techniques as a way to expand the model's perceptual range (i.e., perceptual field), which is specifically optimized for the traffic sign detection task. This improvement allows the model to demonstrate excellent performance in traffic sign detection, with the average recognition accuracy soaring to 99.0%, while maintaining an efficient detection speed of only 0.449 seconds per detection on average. [13] The traffic monitoring system uses Faster R-CNN to monitor traffic flow, detect traffic violations (e.g., red-light running, driving against traffic), and identify traffic accidents. By analyzing video streams captured by cameras, these systems can provide real-time traffic conditions and trigger alerts.

2.4.3. Industry

By applying the optimized YOLOv3 framework and combining it with the k-means clustering algorithm to accurately generate a priori frames, we further enhance the adaptability of the model. Meanwhile, the residual module is introduced to construct the feature fusion detection layer, a strategy that significantly improves the detection accuracy of the model. In the electrical connector defect detection task, the model exhibits higher detection accuracy compared to Faster R-CNN. [11] Target detection techniques in the field of deep learning have been effectively applied to real-time monitoring of work environments, which has the powerful ability to detect and recognize possible safety hazards, such as the ability to promptly detect workers who do not wear the required safety equipment, thus significantly improving the safety and protection level of the factory operating environment.

2.4.4. Pedestrian Detection

In pedestrian detection, a key area of intelligent autonomous driving technology, especially in infrared sensing environments, we have conducted an in-depth study using the U-FOV infrared pedestrian dataset. Through the integration of a well-designed feature pyramid structure and attention mechanism, we successfully constructed an optimized detection model. Compared with the classic YOLOv3, the model achieves a significant improvement in detection performance, as evidenced by a 26.49% increase in accuracy, which provides stronger technical support for the safe driving of self-driving vehicles at night or in low-light environments. [14]. YOLOv3 is used to automate the identification and localization of human bodies in thermal imaging images, in addition, the recognition results of the algorithm in distinguishing human bodies from animals in thermal imaging are also demonstrated, which further validates its powerful image analysis and classification capabilities. Self-driving cars use SSDs to detect pedestrians in real time and take collision avoidance measures, such as slowing down or emergency braking, to ensure pedestrian safety.

3. Conclusion

This paper focuses on the basic idea of YOLO target detection algorithm based on deep learning, overviews the family of YOLO series target detection algorithms and important improvements, analyzes the difference between traditional and deep learning target detection and finally enumerates its applications in agriculture, industry, etc. The YOLO series target detection algorithms have not only achieved a steady increase in the detection speed and accuracy in the continuous optimization, but also demonstrated excellent stability. However, the performance of the algorithms is not sufficient for the detection of small-sized targets and densely distributed targets. In the future, more in-depth research and development to address these challenges will undoubtedly greatly broaden the application value and potential of the YOLO series of algorithms. [6] Combined with the current state of research on target detection algorithms, deep learning-based YOLO target detection can be applied to relatively cold life scenarios, thus making the YOLO target detection algorithms innovative, for

example, ancient cultural relics can be detected and analyzed by using YOLO, which is conducive to the archaeological team's human archaeology. It can also be applied to the wild environment, using YOLO detection technology to survey the living status of wild animals.

References

[1] Mizen & Lian Zhe. A Research Review of YOLO Methods for Generalized Target Detection. Computer Engineering and Applications 1-19.

[2] HINTON G, OSINDERO S, TEH Y. A fast learning algorithm for deep belief nets[J]. Neural Computation, 2006, 18(7): 1527-1554.

[3] SALAKHUTDINOV R, MNIH A, HINTON G. Restricted Boltzmann machines for collaborative filtering[C]//Proceedings of the 24th international conference on Machine learning. ACM, 2007: 791-798.

[4] Shao, Y. H., Zhang, D., Chu, H. Y., Zhang, X. Q. & Rao, Y. B... (2022). A review of YOLO target detection based on deep learning. Journal of Electronics and Information (10), 3697-3708.

[5] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: Unified, real-time object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 779-788.

[6] Xinjie Wang & Jiping Wang. (2024). A review of YOLO target detection algorithms. Guangxi Physics (02),50-53.

[7] Joseph Redmon,Santosh Kumar Divvala,Ross B. Girshick & Ali Farhadi.(2015).You Only Look Once: Unified, Real-Time Object Detection..CoRR

[8] Joseph Redmon & Ali Farhadi. (2016).YOLO9000: Better, Faster, Stronger..CoRR

[9] Bochkovskiy, A., Wang, C. Y., & Liao, H. Y. M. (2020). Yolov4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934.

[10] Ge, Z., Liu, S., Wang, F., Li, Z., & Sun, J. (2021). Yolox: Exceeding yolo series in 2021. arXiv preprint arXiv:2107.08430.

[11] Wu Weihao, Li Qing. Electrical connector defect detection based on improved Yolo v3 [J]. Journal of Sensing Technology, 2020, 33(2):299-307.

[12] Wang Zhuo, Wang Jian, Wang Lingxiong, et al. A lightweight detection method for apples in natural environment based on improved YOLO v4 [J]. Journal of Agricultural Machinery, 2022, 53(8): 294-302

[13] Huiping Pan, Minqin Wang, Fuquan Zhang. Traffic sign detection and recognition method based on optimized YOLO-V4 [J]. Computer Science, 2022, 49(11): 179-184.

[14] B. Zhao, C. Wang, Q. Fu. A multi-scale infrared pedestrian detection method for saliency background awareness[J]. Journal of Electronics and Information, 2020, 42(10): 2524-2532. doi: 10. 11999/JEIT190761.

[15] Zou, Jun, Zhang, S. Y. & Li, J.. (2023). A review of deep learning-based target detection algorithms. Sensor World (08), 9-15. doi: 10.16204/j.sw.issn.1006-883X.2023.08.002.

[16] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. J. Miller. Introduction to wordnet: An on-line lexical database. International journal of lexicography, 3(4):235–244, 1990.

[17] JOSEPH NELSON, JACOB SOLAWETZ. YOLOv5 is Here: State-of-the-Art Object Detection at 140 FPS [EB/OL]. [2020-06-10]. https://blog.roboflow.com/yolo v5-is-here/.

[18] LI C, LI L, JIANG H, et al. YOLOv6: A single-stage object detection framework for industrial applications[J]. arxiv preprint arxiv:2209.02976, 2022.

[19] WANG C Y, BOCHKOVSKIY A, LIAO H Y M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 7464-7475.

[20] JAMES GALLAGHER. How to Train an Ultralytics YOLOv8 Oriented Bounding Box (OBB) Model. [2024-02-06]. https://blog.roboflow.com/train-yolov8-obb-model/.

[21] FANG Y, LIAO B, WANG X, et al. You only look at one sequence: Rethinking transformer in vision through object detection[J]. Advances in Neural Information Processing Systems, 2021, 34: 26183-26197.

[22] Huang Y, Zhou Chun, Liu XJ & Chen Q. YOLOv10-based multi-scale detection model for UAVs in complex backgrounds. Optical Communication Research 1-8.

[23] WANG C Y, YEH I H, LIAO H Y M. YOLOv9: Learning What You Want to Learn Using Programmable Gradient Information[J]. arxiv preprint arxiv:2402.13616, 2024.

[24] CHEN Y, YUAN X, WU R, et al. YOLO-MS: rethinking multi-scale representation learning for real-time object detection[J]. arxiv preprint arxiv:2308.05480, 2023.