Computer vision

*Detection and classification for pictures

Shaojing Fan
University of New South Wales
z5262042

Yuchuan Deng
University of New South Wales
z5362100

Haotian Luo
University of New South Wales
z5341069

Guangyuan Ma
University of New South Wales
z5307190

Mingrui Ding
University of New South Wales
z5292268

I. Introduction

A. Background

In recent years, the fields of deep learning and machine learning have experienced significant growth and innovation. These technologies have revolutionized various domains, including computer vision, natural language processing, and speech recognition. Specifically, in the area of computer vision, both deep learning and machine learning have shown remarkable capabilities in tasks such as image recognition, object detection, and image classification.

Deep learning, especially convolutional neural networks (CNNs), has emerged as a dominant approach in computer vision. CNNs can automatically learn hierarchical representations of features from raw pixel data, enabling them to achieve state-of-the-art performance in various image-related tasks. On the other hand, traditional machine learning methods have proven effective in some applications, relying on hand-crafted features and well-established algorithms.

The rapid progress in both deep learning and machine learning has led to a growing interest in understanding and comparing their performance in specific applications. One area of particular interest is image detection and classification, where the ability to accurately identify objects or patterns within images is crucial in numerous real-world scenarios.

II. Literature Review

Image detection and classification have been hot and crucial topics in the field of computer vision. With the rise of deep learning technologies, especially convolutional neural networks (CNNs), significant progress has been made in image detection and classification tasks. This literature review summarizes some important advancements in this field in recent years, with a focus on comparing the performance of deep learning and traditional machine learning methods in image detection and classification. It also reviews some relevant previous studies.

1. Application of Deep Learning in Image Detection and Classification

Deep learning, particularly convolutional neural networks (CNNs), has become the dominant approach in computer vision tasks. CNNs can automatically learn hierarchical feature representations from raw pixel data, leading to state-of-the-art performance in various image-related tasks. In image detection tasks, CNNs effectively locate and recognize objects within images, while in image classification, they accurately classify images into different categories. For instance, representative algorithms such as YOLO (You Only Look Once) and Faster R-CNN employ CNNs for image detection.

2. Application of Traditional Machine Learning Methods in Image Detection and Classification

In addition to deep learning methods, traditional machine learning methods have also been widely applied in image detection and classification tasks. These methods often rely on handcrafted features and classical machine learning algorithms. For example, Support Vector Machine (SVM) is a common machine learning method used for image classification, which finds the optimal decision boundary to perform image classification. Moreover, image classification methods based on the Bag of Words (BOW) model have shown promising results in practical applications.

3. Comparison between Deep Learning and Traditional Machine Learning Methods

Deep learning methods and traditional machine learning methods have their own strengths and weaknesses in image detection and classification tasks. Deep learning methods, trained in an end-to-end manner, can automatically learn feature representations from raw data, eliminating the need for manual feature engineering. This makes deep learning perform well on large-scale datasets and complex tasks. However, deep learning methods require more data and computational resources for training, and they may encounter overfitting issues with small datasets. In contrast,

traditional machine learning methods may have an advantage on small datasets and offer stronger interpretability for specific tasks.

4. Relevant Previous Studies

Numerous studies have been conducted in the field of image detection and classification. Some prior research focused on improving the performance and robustness of deep learning models, such as using Attention Mechanisms to enhance the model's attention capacity. Other studies aimed to design more efficient feature representation methods to speed up image processing. Additionally, some research attempts to combine deep learning and traditional machine learning methods to leverage their respective advantages effectively.

Conclusion

In conclusion, image detection and classification are essential problems in the field of computer vision, and both deep learning and traditional machine learning methods play crucial roles. Deep learning methods have advantages on large-scale datasets and complex tasks, while traditional machine learning methods may perform better on small datasets with higher interpretability requirements. Future research can continue to explore improvements and optimizations in deep learning models and combine deep learning with traditional machine learning methods to achieve better image detection and classification results.

III. METHOD

In this section we will discuss the main algorithms we used in our experiment of machine learning and deep learning. We will also give a brief overview of the result after each method is applied.

A. Machine learning

Detection:

a). preprocessing:

For the part of image preprocessing, we experiment with two methods: color enhancement and binarization. The average IOU value we get after color enhancement is about 0.38, and some images with strong background colors cannot extract features correctly. The binary image is more stable, and the average IOU value reaches 0.42. Therefore, we finally choose to preprocess the image into a binary image. We resize all the images to (256,256) to facilitate further processing.

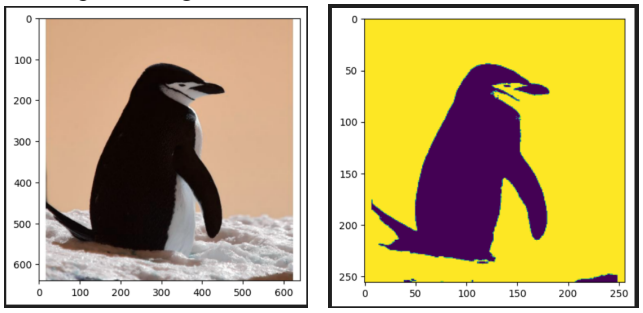

Figure1. Image of picture before preprocessing (left) and after preprocessing (right)

b). Histogram of Oriented Gradients(HOG)

The HOG algorithm in computer vision offers the following advantages: rich feature representation capturing texture and edges, scale-invariance, robustness to lighting changes, and efficient computation for real-time applications. Our dataset contains many images with varying lighting conditions and distinct image edges, making it highly suitable for using HOG for feature extraction.

c). Bag of words(BOW)

The BOW model efficiently represents images by converting them into fixed-length vectors through frequency-based statistics. This reduces the complexity of computation and storage for high-dimensional input data. It is a simple and effective feature representation method, making it suitable for subsequent machine learning algorithms. In our project, using the bow model to convert images to vectors effectively reduces the computational effort.

d). Support Vector Machine(SVC) Regressor

SVC (Support Vector Machine) is a machine learning algorithm used for both classification and regression tasks. Its main advantage lies in finding the optimal decision boundary, which maximizes the margin between classes or minimizes the error in regression. SVC is effective in handling non-linear data and is robust to high-dimensional input. The reason that we chose SVC to be our regressor is that it is highly compatible with HOG, both models are used in computer vision for feature extraction and object detection.

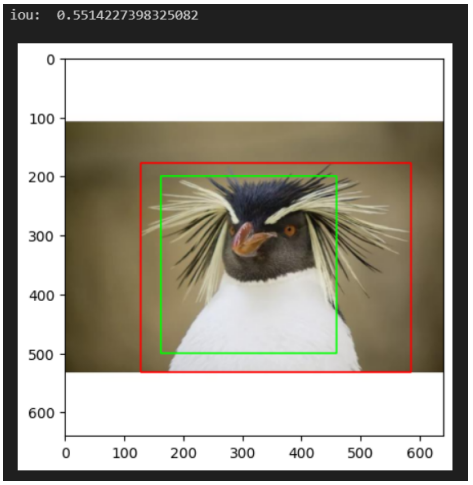

Figure2. Example of feature extraction

Classification:

a). Data enhancement

Data enhancement can effectively improve the generalization ability and robustness of the model. In this experiment, the data set is small since we only have 500 images for training. Therefore, we reversed the image and rotated the image 180 degrees to get more training samples, which effectively improved the accuracy of detection.

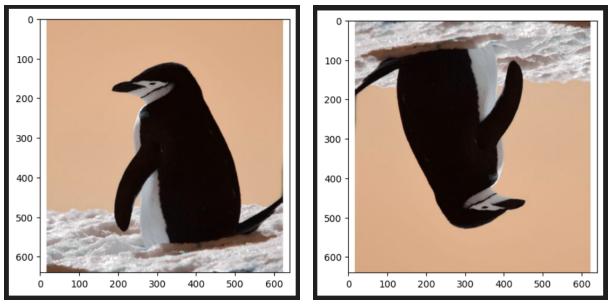

Figure3. Image of picture after reverse (left) and after rotate (right)

b). SVC classifier

The SVC classifier has several advantages, including efficient boundary finding, robustness, and the ability to handle non-linear classification tasks using the kernel trick. Its strong generalization ability and sparsity make it suitable for various tasks like text classification and image classification. Due to these advantages and its compatibility with HOG for feature extraction, we chose the SVC classifier as our classifier of machine learning.

B. YOLO5

a)Introduction of YOLO5

YOLO5 is the latest version of the YOLO series, which uses a deep learning network called the "Darknet" architecture, which is fast and efficient. YOLO5 also employs multiscale prediction, i.e., object detection at multiple different scales, which allows it to better handle objects of various sizes. sizes of objects. In addition, YOLO5 introduces a new loss function, the CIOU loss, which takes into account not only the overlap between the predicted and real frames, but also their centroid distances and aspect ratios, thus providing a more accurate measure of the quality of the prediction.

b)Preprocess

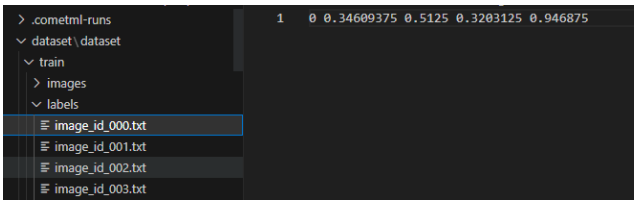

Figure4. Image of labels

Before using YOLO5, we preprocess our dataset. Our dataset consists of images of turtles and penguins, as well as labels and bounding boxes for the objects in those images. We do this by reading the data in json and converting it into YOLO's format. (<class> <x_center> <y_center> <width> <height>)

- For <class> we specify 0 for penguins and 1 for turtles.
- <x_center> and <y_center> are the coordinates of the center of the bounding box. The origin of the coordinates (0,0) is the upper left corner of the image.
- <width> and <height> are the width and height of the bounding box. All values are scaled relative to the width and height of the image, the coordinates of the bounding box should be normalized to a range of 0 to 1, and the coordinates and width and height of the center point are used.

c)Why we choose YOLO5

- Speed and accuracy: YOLO5 is very fast, without sacrificing performance, and still delivers accurate results even for real-time object detection tasks.
- Adaptability: The YOLO5 is able to handle a wide range of sizes and shapes, which is important for animals with different forms like turtles and penguins.
- Ease of use: YOLO5 provides easy-to-use training and inference tools, we can implement a powerful object detection system with a small amount of code.

d) Conclusion

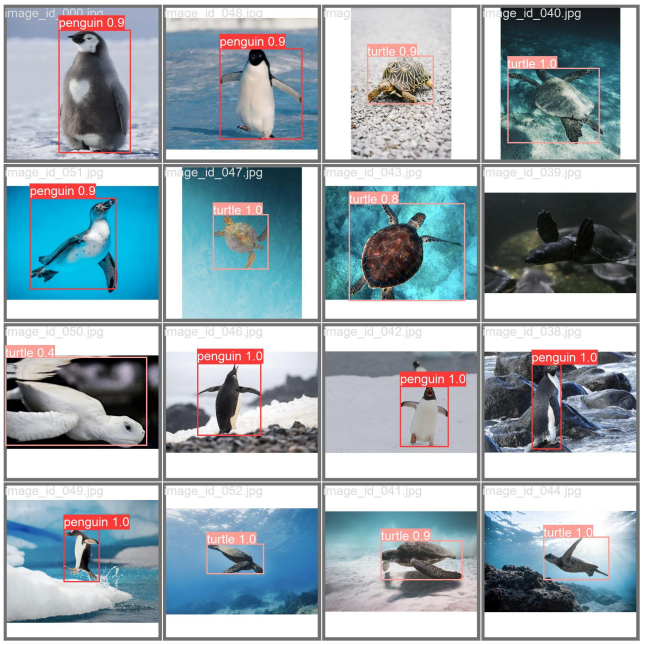

Figure5. Image of yolov5

Overall, due to the advantages of YOLO5 's speed, accuracy, adaptability, and ease of deployment, YOLO5 is an ideal choice for our turtle and penguin recognition task. We use the ordinary YOLO5 model as a basic reference, and later we will employ data augmentation, attention mechanisms, and other means to further improve the accuracy of the yolo model.

C. Data enhancement

a) Data enhancement is a technique used to expand the size of our dataset by applying various transformations to the original images. In Computer Vision, this technique is incredibly valuable and plays a significant role in enhancing the performance of our models.

b) The primary objectives of data enhancement in Computer Vision are to increase the quantity of available

data for training, improve the generalization ability of our models, enhance their robustness to handle real-world variations, and balance class distribution.

c) Data enhancement offers several advantages. It helps us avoid overfitting by making our models less prone to memorizing training data. Additionally, it enables our models to handle real-world variations and challenges, making them more reliable in practical applications. When we have limited data, data enhancement allows us to make the most efficient use of that data by generating additional training samples. As a result, data enhancement often leads to improved performance and accuracy of our models compared to training on the original dataset alone.

d) In Computer Vision, there are several common data enhancement techniques, such as rotation, flip, scaling, translation, and brightness/contrast adjustment. These techniques introduce diversity and variability into the dataset, making our models more robust.

e) In our project, we implemented a function called "rotate4," which generated four rotated versions of each input image along with their corresponding bounding boxes. Additionally, we apply the horizontal flip techniques and save the enhanced images to the destination folder with the format which yolov5 required.

Figure6. Image of rotate4

f) In conclusion, data enhancement is a critical technique in Computer Vision tasks. It helps us improve model performance, generalization, and robustness. Our implementation of data enhancement demonstrates the effectiveness of these techniques.

D. Attention Mechanisms

Attention mechanism is a technique used in deep learning to enhance a model's attention and selection of important information.

The goal of attention mechanisms is to automatically select relevant information and ignore irrelevant information given an input.

In deep learning tasks, the input data is typically high-dimensional, such as images, text, or sequences. The attention mechanism's primary objective is to assign different weights to various parts or elements within these inputs, enabling the model to focus more on crucial information. This process of weighting can be achieved through learning, where the model automatically learns the importance of different elements, or it can be manually designed based on the specific characteristics of the problem.

In our project, we intend to use the SE attention to add attention mechanisms.The application of SE attention in convolutional neural networks is very simple and effective. It does not increase the computational complexity of the network, but can significantly improve the performance of the model. The SE attention mechanism has been widely used in various computer vision tasks, such as image classification, object detection, image segmentation, etc., and achieved significant performance improvements.

The SE model is mainly divided into two parts:

- Squeeze: In this step, the SE model converts the feature map of each channel into a numerical value through a global average pooling operation, thereby obtaining a channel descriptor. This channel descriptor reflects the global information of the features within that channel.
- Excitation: In this step, the SE model processes the channel descriptors through two fully connected layers to generate a channel attention vector. Each element in this vector represents the importance weight of the corresponding channel.

Through the above steps, SE attention obtains a channel attention vector, which contains the importance weight of each channel. Then, SE attention multiplies this vector with the original feature map to achieve channel-adaptive scaling of the feature map. This means that the model can adjust the contribution of different channels in the feature map according to the importance of each channel.

IV. Experimental Result

When the yolo model finishes running, a latest exp folder will be generated under the run folder. The exp folder contains the results of the last run of the model. Among them, result.png shows the changes of 10 data as the epoch runs. As shown below:

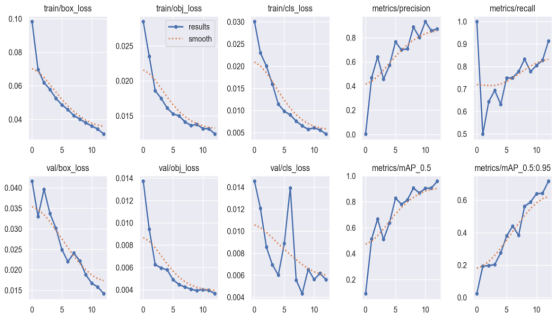

Figure7. Image of result.png

1) IOU

IOU stands for Intersection over Union, also known as Jaccard index. It is a metric used to measure the degree of overlap between predicted bounding boxes or segmentation masks and ground-truth labels in object detection and segmentation tasks.

Table 1 shows the performance of IOU of different models.

Model	The performance of IOU

Machine Learning	0.42
YOLOv5	0.72
YOLOv5 with Data Enhancement	0.86
YOLOv5 with Attention Mechanisms	0.80

Table.1. IOU of each models

2) Classification Accuracy

All models will give a confusion metric at the end, and we will give a classification accuracy through the calculation of the confusion metric. The higher the accuracy, the more accurate our model is. The expression are shown below:

$$Accuracy = \frac{The\ total\ number\ of\ correct\ predictions}{The\ total\ number\ of\ images},$$

Table 2 shows the Classification Accuracy of different models.

Model	Classification Accuracy
Machine Learning	0.63
YOLOv5	0.82
YOLOv5 with Data Enhancement	0.92
YOLOv5 with Attention Mechanisms	0.87

Table.2.Classification Accuracy of each models

V. DISCUSSION

The result and performance could be divided into two parts for discussion, detection and classification.

1. Detection:

The above results show that the accuracy of detection images in deep learning is much higher than that in machine learning. The highest IOU we achieved was 0.42 for machine learning and 0.86 for deep learning. Machine learning handles image detection tasks with hand-designed feature extractors and classifiers, while deep learning uses neural networks as models without the need to manually design feature extractors. In the case of large-scale data and sufficient computing resources, deep learning models generally have better performance. In the case of small-scale data and resource constraints, traditional machine learning methods are an effective choice, but we can increase the sample size through data enhancement and other means, so deep learning is still suitable for small-scale data, and has higher accuracy.

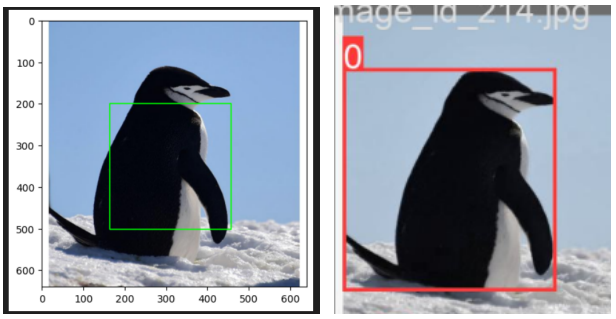

Figure8. Image of picture detected by machine learning (left) and deep learning (right)

When it comes to deep learning models, data enhancement performs better than attention mechanisms in our experiments. Data enhancement generates more training samples and improves robustness. For some data sets with an uneven number of categories, it is also possible to smooth the distribution of different categories. However, at the same time, unreal samples may be introduced, which will reduce the performance of the model. The attention mechanism can automatically learn the key parts and features of the input data to improve the performance of the model. Our sample picture has more noise, and the data enhancement makes our data more robust. We only trained 500 images in total, which is not enough for the attention mechanism, which is why the IOE value obtained by the attention mechanism is lower than the data enhancement.

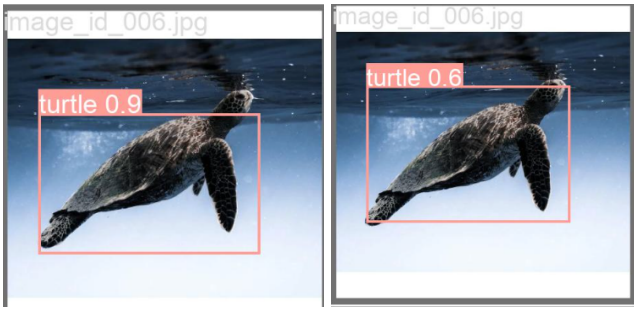

Figure9. Image detected by using YOLOv5 with Attention Mechanisms (left) and YOLOv5 WITH Data enhancement(right)

2. Classification:

From our results, deep learning performs better than machine learning in achieving classification. We speculate that there may be several reasons:

a. Network Depth

Deep learning models usually have a deeper network structure and can learn more levels of abstract features, thereby improving the expressiveness and generalization capabilities of the model. In contrast, traditional machine learning methods usually only have shallow networks or linear models, which cannot make full use of the information in the data.

b. Automatic feature learning

The deep learning model has the ability of automatic feature learning when processing data, without manual feature engineering. This is one of

the notable differences from traditional machine learning methods. Traditional machine learning methods need to rely on domain experts to manually design and select features, and this process may miss some important information or introduce redundant features, thus limiting the performance of the model. The deep learning model can learn higher-level and more abstract features through a multi-layer neural network to better represent the complexity of the data.

c. Overfitting and underfitting

Machine learning has the problem of overfitting or underfitting. When overfitting, the model learns the noise in the data rather than the true pattern. When underfitting, the model fails to capture patterns in the data.

In deep learning models, data augmentation works better than attention mechanisms. Attention mechanism is better than data augmentation.

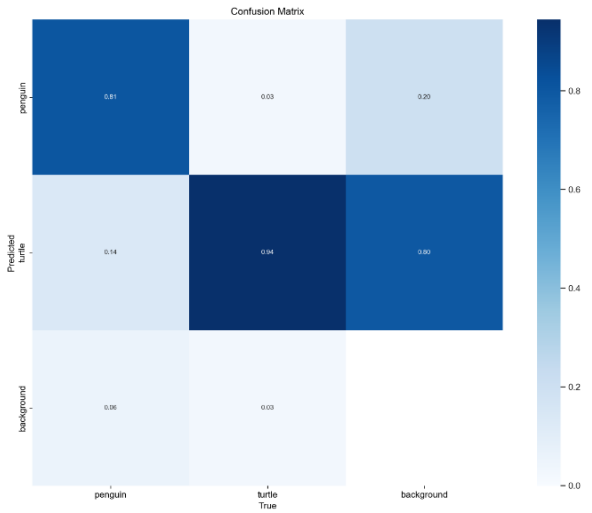

Figure9. Image of yolov5 confusion matrix

Compared with traditional deep learning, data enhancement is mainly reflected in improving data diversity and alleviating overfitting. Data enhancement technology can generate diverse samples by performing operations such as rotation, flipping, translation, and scaling on the original data. This can increase the sample size of the data set, enable the model to learn the distribution and characteristics of the data more fully, and improve the generalization ability of the model. At the same time, in deep learning tasks, the model is prone to overfitting due to too many parameters, that is, it performs well on the training data, but performs poorly on the test data. Data augmentation can introduce randomness in the training process and perturb the distribution of the original data, thereby reducing the model's overfitting of the training data and improving the performance of the model on unseen samples.

3. Limitation and improvement

Data Augmentation Strategy: For the penguin and turtle recognition task, you can explore various data augmentation techniques such as scale transformation, brightness adjustment, and color transformation. Diversifying data augmentation can enhance the model's adaptability to different scenes and conditions.

Attention Mechanism Optimization: If attention mechanisms are used, you can further optimize their design. Consider using different attention modules or adjusting the weight allocation of attention to improve the model's focus on key regions, thereby enhancing detection and classification performance.

Model Fusion: For the different methods employed, you can try model fusion to combine their strengths. For instance, ensemble learning methods can be used to combine predictions from multiple models, resulting in a more robust and accurate recognition outcome.

Feature Engineering: For traditional machine learning methods, you can explore optimizing the feature engineering process by selecting more effective feature representations or using automatic feature learning methods to replace manual feature engineering, thereby improving the performance of machine learning models.

Data Quality and Sample Balancing: Ensuring data quality and sample balance are crucial for model performance. Data cleaning and filtering can be performed to remove low-quality samples, while balancing the number of samples in each class can prevent model biases toward certain classes.

Cross-Validation and Hyperparameter Tuning: In experimental design, it is recommended to use cross-validation to evaluate model performance, overcoming randomness caused by data splitting. Additionally, for each method, perform careful hyperparameter tuning to find the optimal parameter configuration and improve model performance.

VI. Conclusion

As for detection, the best iou result of machine learning is inferior to the worst iou result of deep learning, so we can say that deep learning is more suitable for our data set than machine learning. As for the two optimization schemes of deep learning, data enhancement has achieved better results. The result of our analysis is that the amount of data in this task is less, and the difference between penguins and turtles is large, so data enhancement will perform better than the attention mechanism. A similar result is found in the Classification section, where machine learning accuracy is lowest and data enhancement performance is higher than attention mechanisms. Overall, for images with small amounts of data and large feature differences, a data-enhanced deep learning model is the best choice.

VII. Contribution of group members

Yuchuan Deng: Wrote the detection and classification sections of machine learning as well as being responsible for the corresponding sections in presentations and reports.

Guangyuan Ma: Responsible for data preprocessing, training and analysis of the yolo5 base model, and corresponding presentations and reports.

Jingfan Shao: Responsible for the algorithms for data enhancement based on yolo5, as well as training and

improvement, and the corresponding presentation and reports sections.

Haotian Luo: Responsible for the algorithm of the attention mechanism based on yolo5, and the corresponding presentation and reports sections. Also provided major arithmetic support for training.

Mingrui Ding: Responsible for the collection and collation of results as well as the summary part of the corresponding presentations and reports.

VIII. References

[1] W. Fang, L. Wang and P. Ren, "Tinier-YOLO: A Real-Time Object Detection Method for Constrained Environments," in IEEE Access, vol. 8, pp. 1935-1944, 2020, doi: 10.1109/ACCESS.2019.2961959.

[2] R. Huang, J. Pedoeem and C. Chen, "YOLO-LITE: A Real-Time Object Detection Algorithm Optimized for Non-GPU Computers," 2018 IEEE International Conference on Big Data (Big Data), Seattle, WA, USA, 2018, pp. 2503-2510, doi: 10.1109/BigData.2018.8621865.

[3] X. Sun, X. Jia, Y. Liang, M. Wang and X. Chi, "A Defect Detection Method for a Boiler Inner Wall Based on an Improved YOLO-v5 Network and Data Augmentation Technologies," in IEEE Access, vol. 10, pp. 93845-93853, 2022, doi: 10.1109/ACCESS.2022.3204683.

[4] J. Sun, H. Ge and Z. Zhang, "AS-YOLO: An Improved YOLOv4 based on Attention Mechanism and SqueezeNet for Person Detection," 2021 IEEE 5th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), Chongqing, China, 2021, pp. 1451-1456, doi: 10.1109/IAEAC50856.2021.9390855.

[5] Li, R.; Wu, Y. Improved YOLO v5 Wheat Ear Detection Algorithm Based on Attention Mechanism. Electronics 2022, 11, 1673. https://doi.org/10.3390/electronics11111673

[6] Kim, M.; Jeong, J.; Kim, S. ECAP-YOLO: Efficient Channel Attention Pyramid YOLO for Small Object Detection in Aerial Image. Remote Sens. 2021, 13, 4851. https://doi.org/10.3390/rs13234851

[7] Liu, K.; Peng, L.; Tang, S. Underwater Object Detection Using TC-YOLO with Attention Mechanisms. Sensors 2023, 23, 2567. https://doi.org/10.3390/s23052567

[8] Sekharamantry, P.K.; Melgani, F.; Malacarne, J. Deep Learning-Based Apple Detection with Attention Module and Improved Loss Function in YOLO. Remote Sens. 2023, 15, 1516. https://doi.org/10.3390/rs15061516

[9] C. Liu, D. Li and P. Huang, "ISE-YOLO: Improved Squeeze-and-Excitation Attention Module based YOLO for Blood Cells Detection," 2021 IEEE International Conference on Big Data (Big Data), Orlando, FL, USA, 2021, pp. 3911-3916, doi: 10.1109/BigData52589.2021.9672069.

[10] Yu, L.; Zhu, J.; Zhao, Q.; Wang, Z. An Efficient YOLO Algorithm with an Attention Mechanism for Vision-Based Defect Inspection Deployed on FPGA. Micromachines 2022, 13, 1058. https://doi.org/10.3390/mi13071058

[11] L. Jiang et al., "MA-YOLO: A Method for Detecting Surface Defects of Aluminum Profiles With Attention Guidance," in IEEE Access, vol. 11, pp. 71269-71286, 2023, doi: 10.1109/ACCESS.2023.3291598.

[12] W. Lan, J. Dang, Y. Wang and S. Wang, "Pedestrian Detection Based on YOLO Network Model," 2018 IEEE International Conference on Mechatronics and Automation (ICMA), Changchun, China, 2018, pp. 1547-1551, doi: 10.1109/ICMA.2018.8484698.

[13] R. Laroca et al., "A Robust Real-Time Automatic License Plate Recognition Based on the YOLO Detector," 2018 International Joint Conference on Neural Networks (IJCNN), Rio de Janeiro, Brazil, 2018, pp. 1-10, doi: 10.1109/IJCNN.2018.8489629.