MRS-YOLO: AN ENHANCED EFFICIENT OBJECT DETECTION FRAMEWORK FOR POWER SYSTEMS

WENCHONG LI[1*], YAN GAO[2], YAOXUAN SHI[3], ZIHAN WEI[3]

[1]School of Information Technology&Engineering, Guangzhou College of Commerce, Guangzhou, 511363, China
[2]School of Electrical Engineering, Henan Polytechnic University, Jiaozuo, 454150, China
[3]School of Automation, Guangdong University of Technology, Guangzhou, 510006, China
E-MAIL: 305088972@qq.com

Abstract:

With the rapid advancement of deep learning-based object detection techniques, this study identifies critical limitations in the YOLOv11 framework regarding small object detection, multi-scale target handling, and feature representation in complex scenes. To address these challenges, we propose MRS-YOLO, an enhanced framework incorporating two key innovations: (1) a Receptive Field Enhanced Spatial Convolution module that combines dynamic weight allocation with multi-branch dilated convolutions (1×1/3×3/5×5) and spatial attention mechanisms, and (2) a Multi-Scale Feature Fusion module featuring triple-path feature extraction (local/medium/long-range) coordinated with dual-attention mechanisms. Evaluated on the HELMET dataset, MRS-YOLO achieves a 64.2% mAP50, outperforming YOLOv5n (61.7%), YOLOv8n (62.5%), and YOLOv11n (63.3%). The expanded convolutional modules and optimized detection heads significantly improve contextual feature extraction while maintaining computational efficiency. Comparative experiments demonstrate the framework's superior capability in preserving spatial information during down sampling and aligning cross-resolution features, particularly for industrial applications like power system inspection and safety monitoring.

Keywords:

Object-detection; Multi-Scale Feature Fusion; YOLOv11.

1. Introduction

Object detection, as one of the core tasks in the field of computer vision, aims to accurately localize and identify objects of interest in images or videos. With the rapid advancement of deep learning technologies, object detection algorithms have achieved significant breakthroughs in both accuracy and efficiency, finding widespread applications in autonomous driving, intelligent surveillance, industrial quality inspection, medical image analysis, and other domains.

Traditional object detection methods are primarily divided into two-stage detectors (e.g., the R-CNN series) and single-stage detectors (e.g., YOLO, SSD). Two-stage methods achieve high precision through region proposal and classification-regression mechanisms, while single-stage methods, due to their end-to-end architecture, are better suited for real-time scenarios. Examples include the detection of safety helmets and equipment during power system operations, as well as various real-time task applications.

Among single-stage detectors, the YOLO (You Only Look Once) series algorithms have become a research hotspot in both industrial and academic communities due to their exceptional speed-accuracy balance.

Object detection algorithms have achieved remarkable progress in recent years, with the YOLO series attracting significant attention for its efficient real-time detection capabilities. From YOLOv1, which first transformed the detection task into a single grid prediction, to YOLOv3 introducing multi-scale feature fusion and the Darknet-53 backbone network, and further to YOLOv5 improving training efficiency through adaptive anchor boxes and Mosaic data augmentation, each iteration has sought a better trade-off between speed and accuracy [1-3]. The latest YOLOv11 further optimizes performance by incorporating dynamic feature selection mechanisms and mixed-precision training, enhancing small object detection while maintaining real-time processing.

The evolution of this series reflects a technological progression from single-scale prediction to multi-scale fusion, and from fixed parameters to adaptive learning, providing continuously optimized solutions for real-time object detection. However, challenges such as small object detection in complex scenes, occluded object recognition, and computational resource constraints remain critical issues to be addressed.

Building upon the YOLOv11 framework - the latest

evolution in the YOLO series that incorporates advanced designs including C3K2 blocks, SPPF structures and C2PSA attention mechanisms - this study proposes targeted improvements to address critical limitations in complex scenarios. While maintaining the conventional three-stage architecture (Backbone for multi-scale feature extraction, Neck for cross-layer feature fusion and enhancement, and Head for object localization/classification), we identify three key challenges: (1) inadequate preservation of multi-scale features due to spatial information loss during downsampling, (2) difficulties in cross-resolution feature alignment stemming from inter-scale feature misalignment with traditional methods, and (3) constrained small object detection accuracy resulting from insufficient receptive field coverage in existing modules. These limitations significantly impact performance in real-world applications involving diverse object scales and complex backgrounds.

To address these challenges, this paper proposes two key innovations: (1) a Receptive Field Enhanced Spatial Convolution (RSConv) module featuring dynamic weight allocation through multi-branch dilated convolution (parallel 1×1/3×3/5×5 paths) and spatial attention fusion (7×7 convolution with sigmoid activation); and (2) a Multi-Scale Feature Fusion Enhancement (MSFFusion) module with triple-path extraction (1×1 local, 3×3 medium-range, dilated 3×3 long-range convolutions) and dual-attention coordination (channel & spatial attention).

The proposed RSConv module enhances multi-scale feature preservation capabilities, while the MSFFusion module effectively addresses cross-resolution feature alignment issues, collectively achieving significant improvements in small object detection performance under complex scenarios. This integrated approach provides a more robust solution for industrial inspection, security surveillance, and related applications.

2. Related Works

Object detection, as a core task in computer vision, has research spanning multiple aspects, with detection technologies now penetrating various interdisciplinary fields.

As a classic object detection algorithm, YOLO technology has undergone multiple generations of updates and finds wide applications across domains: industrial automation for product defect detection and assembly line sorting; autonomous driving for real-time vehicle and pedestrian detection to ensure navigation safety; security surveillance systems using face and behavior recognition to enhance warning capabilities; medical image analysis for cell counting and lesion localization (e.g., tumor marking in CT scans); smart agriculture for crop pest monitoring and fruit ripeness assessment; retail sector utilizing shelf product detection to optimize inventory management; drone inspections for power line fault identification and photovoltaic panel defect localization; aerospace supporting ground object classification in satellite imagery (e.g., building and road extraction); and military defense for battlefield target tracking and threat assessment in cutting-edge scenarios [4-9].

Numerous YOLO algorithm improvement studies exist for these domains, such as: [10] proposed an efficient coarse object locating method based on a saliency mechanism. The method could avoid an exhaustive search across the image and generate a small number of bounding boxes, which can locate the object quickly and precisely.[11] proposed the MMI-Det which is a multi-modal fusion method for visible and infrared object detection. The method can provide a good combination of complementary information in the visible-infrared modalities and output accurate and robust object information. [12] proposed an automatic dataset creation method. This method first extracts objects from the source images and then combines them as synthetic images. [13] proposed a novel RSI anchor-free object detection framework that consists of two key components: a cross-channel feature pyramid network (CFPN) and multiple foreground- attentive detection heads. [14] proposed a deblurring dictionary encoding fusion network (DDFN) for infrared and visible image object detection. [15] proposed a novel adaptive object detection system (AdaDet) based on early-exit neural networks. [16] proposed a new probabilistic framework for object detection which is related to the Hough transform.

3. Architecture Design and Key Innovations of MRS-YOLOv11

The proposed model builds upon the robust framework of YOLOv11, a state-of-the-art evolution in the YOLO series renowned for integrating advanced architectural innovations such as C3K2 blocks, SPPF modules, and C2PSA attention mechanisms[17]. Mirroring its predecessors, YOLOv11 employs a three-stage hierarchical structure consisting of a Backbone for multi-scale feature extraction, a Neck for cross-layer feature fusion and enhancement, and a Head for object localization and classification prediction. To address specific challenges prevalent in complex scene understanding, such as preserving fine-grained details across scales and improving feature representation, we introduce targeted refinements to each stage. These modifications encompass: 1) replacing standard convolutions within the Backbone with novel

RSConvFocus and RSConvDownsample modules, designed to optimize spatial information retention during critical downsampling operations;2) integrating MSFFusion blocks into the Neck to significantly enhance cross-resolution feature aggregation， as shown in the overall model Figure 1.

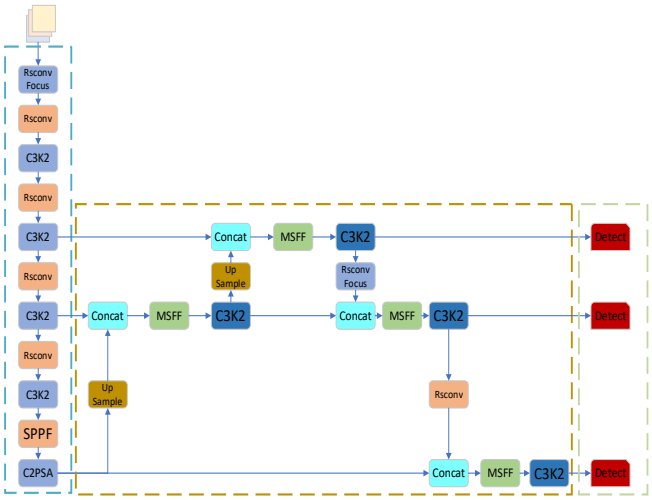

FIGURE 1. Overall Diagram of the MRS-YOLOv11 Model

3.1. Receptive Field Enhanced Spatial Convolution (RSConv)

To address the limitations of standard convolutional operations in handling multi-scale objects and preserving spatial information during downsampling, we propose a novel Receptive Field Enhanced Spatial Convolution (RSConv) architecture. This multi-path approach addresses the fundamental challenge of multi-scale object representation. By integrating parallel feature extractors with hierarchical receptive field coverage, the module overcomes the limitations of traditional single-branch convolutions that struggle with scale variation in complex environments.This module incorporates three key innovations: dynamic weight allocation,multi-branch dilated convolution,and spatial attention mechanisms. Refer to Figure 2 for details.

Dynamic Weight Allocation Mechanism:The core innovation enabling adaptive feature fusion is the DynamicWeightAllocation module, which learns channel-wise weighting coefficients for feature aggregation. This submodule implements a dual-pooling attention mechanism defined as:

$$\mathcal{DWA}(x) = \sigma\left(\mathcal{C}_{1\times1}^{2\to1}\left(\mathrm{SiLU}\left(\mathcal{C}_{1\times1}^{2c\to c/2}[\mathrm{AvgPool}(x) \oplus \mathrm{MaxPool}(x)]\right)\right)\right) \quad (1)$$

Where⊕denotes channel-wise concatenation, $\mathcal{C}_{1\times1}^{2\to1}$ rep resents pointwise convolution, and σ is the softmax act ivation function. As shown in the implementation. This module first extracts complementary channel statistical features through parallel average pooling and max po oling operations. Then, the concatenated features are di mensionally reduced and compressed via a bottleneck c onvolution with a compression ratio of 2. Next, a thre e-dimensional weight vector is generated through softm ax normalization. Finally, it dynamically adjusts the fea ture contribution of different convolution branches.

Multi-branch Dilated Convolution:The RSConv backbone integrates three parallel convolution pathways with progressively expanding receptive fields, embodying a design rationale that includes hierarchical receptive field coverage via 1×1, 3×3, and 5×5 effective kernels. It employs shared convolution parameters (such as kernel size and stride) across branches, coupled with automatic padding computation to maintain consistent feature dimensions. Batch normalization is further incorporated to ensure stable gradient propagation, enabling the module to effectively handle multi-scale feature extraction while preserving spatial information integrity.

RSConvFocus with Spatial Attention:The RSConv framework enhances initial feature extraction through an attention-infused design that integrates multi-scale feature extraction via its RSConv backbone with a 7×7 convolution for broad spatial context modeling. This is complemented by a sigmoid activation function to generate soft attention masks, which are then applied through element-wise multiplication to refine features and selectively amplify salient regions. By synergistically combining these components, the framework substantially improves multi-scale feature representation while maintaining computational efficiency.

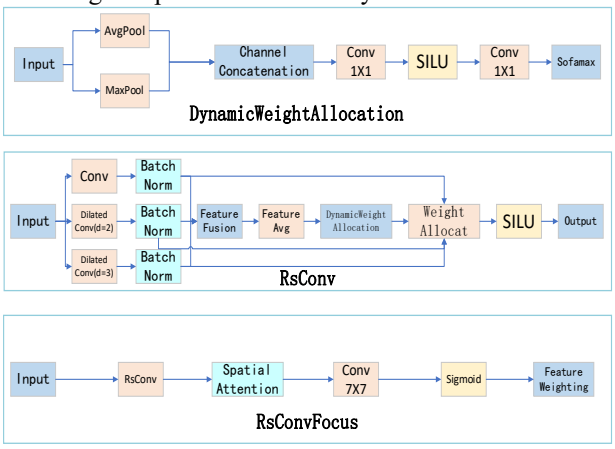

FIGURE 2. Receptive Field Enhanced Spatial Convolution Module

3.2. Multi-Scale Feature Fusion Enhancement Module

The Multi-Scale Feature Fusion (MSFFusion) module represents a sophisticated advancement in multi-scale feature integration within convolutional neural networks, incorporating three specialized components that operate synergistically to enhance feature representation: multi-scale feature extraction layers with hierarchical receptive fields, complementary channel and spatial attention mechanisms, and residual learning pathways for feature refinement. This architecture demonstrates particular efficacy in addressing feature misalignment across scales—a persistent challenge in multi-scale object detection that traditional frameworks often struggle to resolve. By harmonizing hierarchical feature extraction with attention-driven refinement, the module enables more coherent cross-scale feature aggregation, thereby improving detection performance for objects of varying sizes in complex scenes. Refer to Figure 3 for details.

Multi-Scale Feature Extraction: The Scale Feature Aggregation component of the multi-scale feature extraction module implements parallel feature extraction pathways: a 1×1 convolution branch for local feature extraction and channel compression, a standard 3×3 convolution with dilation=1 to capture medium-range spatial contexts, and a dilated 3×3 convolution with dilation=2 to extend the effective receptive field. Feature maps from these branches are concatenated and fused through a 1×1 convolution layer. This architecture enables comprehensive scale coverage while maintaining computational efficiency via strategic channel reduction in each branch, ensuring that the model can effectively handle objects of varying sizes without incurring excessive computational overhead.

Dual-Attention Integration: Channel Attention Component implements dual-pooling excitation where both average-pooled and max-pooled features are processed through shared bottleneck layers (reduction ratio=8). The resultant channel attention weights are computed as:

$$\mathcal{CA}(\mathrm{x}) = \sigma\left(\mathcal{C}_{1\times1}\left(\mathrm{ReLU}\left(\mathcal{C}_{1\times1}(\mathrm{AvgPool}(\mathrm{x}))\right)\right) + \mathcal{C}_{1\times1}\left(\mathrm{ReLU}\left(\mathcal{C}_{1\times1}(\mathrm{MaxPool}(\mathrm{x}))\right)\right)\right) \quad (2)$$

Spatial Attention Component constructs position-wise significance maps through depthwise convolution, preserving spatial relationships while reducing computational overhead. The spatial weighting function employs 3×3 depthwise convolution with channel-wise normalization and sigmoid activation, generating spatially-variant attention masks that dynamically adjust feature importance across the image plane.

Residual Feature Refinement:The attention-weighted features are integrated with the original input through a residual pathway:

$$\mathrm{Output} = \mathcal{P}(\mathcal{W}_c \circ \mathcal{W}_s \circ \mathrm{MSFA(x)}) + \mathrm{x} \quad (3)$$

Where $\mathcal{P}$ denotes the projection convolution, $\mathcal{W}_c$ represents channel attention weights, $\mathcal{W}_s$ denotes spatial attention weights, $\mathrm{MSFA(x)}$ signifies the multi-scale feature aggregation output, and $\circ$ indicates element-wise multiplication. This residual structure ensures stable gradient propagation while preventing information degradation during feature fusion.

The forward propagation process implements a multi-stage feature integration framework: input features first undergo channel alignment via dimension-preserving convolution, followed by multi-scale feature extraction through parallel processing branches. Channel attention then recalibrates feature significance across channels, while spatial attention identifies and emphasizes positionally critical regions. The dual-attention weighted features are projected through a 1×1 convolution layer, and original features are preserved via skip connections. This residual learning pathway ensures feature fidelity while enabling progressive refinement of multi-scale representations, allowing the model to effectively balance contextual information and computational efficiency.

This strategy synergistically combines dimensional consistency, scale diversity, and attention-driven feature selection, preventing information loss during multi-stage fusion. The integration of residual connections further stabilizes gradient flow, making it particularly suitable for tasks requiring fine-grained feature alignment across varying scales.

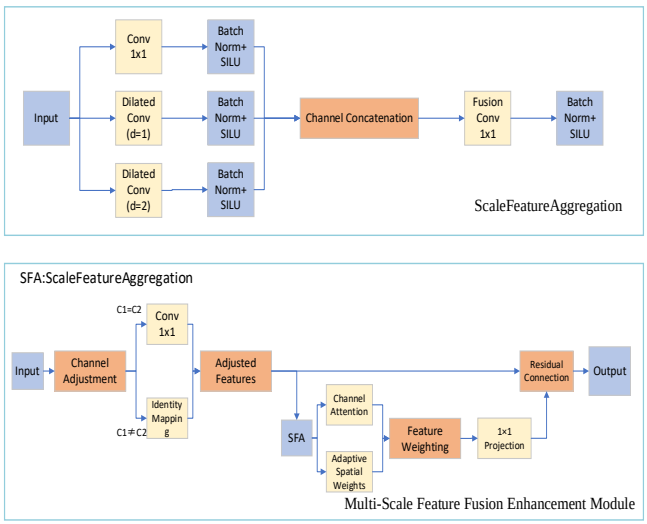

FIGURE 3. Multi-Scale Feature Fusion Enhancement Module

4. Experiments

To further validate the detection performance of the MRS - YOLOv11 model in complex detection environments, this paper conducts performance testing on the self -built HELMET dataset. Additionally, comparative experiments with classic object detection algorithms are performed.

Experimental Conditions: The operating system is Ubuntu 22.04, the deep learning framework is PyTorch 1.9.0, the CPU is an Intel® Core™ i3 - 7350K, the memory is 16GB, the GPU is a GeForce RTX 3070 Ti, and the CUDA version is 11.1.

TABLE 1. Experimental conditions

Experimental conditions					
Operating system	*Deep learning framework*	*CPU*	*Memory*	*GPU*	*CUDA*
Ubuntu 22.04	Pytorch 1.9.0	Intel i3 - 7350K	16G	NVIDIA 3070Ti	CUDA11.1

As shown in Table 1, to ensure the fairness of comparative experiments, both the MRS-YOLOV11 algorithm and classical algorithms were tested on the helmet dataset containing 3 categories of objects. The dataset includes 4,500 training samples, 500 validation samples, and 25,502 target instances. The experiment was configured with 100 training epochs, a batch size of 8, using the SGD optimizer for training parameters, a learning rate of 0.01, and a momentum of 0.937.

TABLE 2. Comparison of Experimental Results

Model	HELMET dataset.			
	YOLOv5n	*YOLOv8n*	*YOLOv11n*	*ours(MRS-YOLOv11)*
Map50(%)	61.7	62.5	63.3	**64.2**

As shown in TABLE 2,the experimental results indicate that the MRS-YOLOv11model (ours) achieves a mAP50(%) of 64.2 on the HELMET dataset. It outperforms the YOLOv5n model by 2.5 percentage points, surpasses the YOLOv8n model by 1.7 percentage points, and exceeds the YOLOv11n model by 0.9 percentage points. These results demonstrate that the improvement strategies adopted in MRS - YOLOv11, namely the Receptive Field - Enhanced Spatial Convolution and the Multi - Scale Feature Fusion Enhancement Module, can effectively boost the model's detection performance, leading to more accurate detection results.

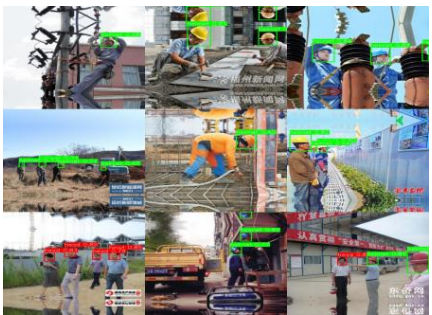

FIGURE 4. Graph of Detection result

As shown in Figure 4 ,the experimental results demonstrate that the model is capable of recognizing helmets and unhelmeted objects with high accuracy. This indicates that the incorporation of the Receptive Field-Enhanced Spatial Convolution and the Multi-Scale Feature Fusion Enhancement Module can improve the overall detection performance of the model.

5. Conclusions

The study successfully addresses key challenges in power system object detection through the MRS-YOLO framework. By integrating RSConv's hierarchical receptive field coverage and MSFFusion's attention-driven multi-scale fusion, the model significantly improves detection accuracy for small objects and complex scenes.

Experimental results validate a 0.9–2.5% mAP50 gain over baseline YOLO variants, with robust performance in helmet detection tasks. The proposed innovations—context aggregation and dedicated small-object detection heads—demonstrate generalizability for industrial use cases requiring fine-grained feature alignment. Future work will explore lightweight deployment for edge devices and extension to multi-modal data to further enhance real-world applicability.

References

[1] Bochkovskiy A, Wang C Y,Liao H YM. Yolov4Optimal speed and accuracy of object-detection[JarXiv preprint arXiv:2004.10934, 2020.

[2] Jocher G, Stoken A, Chaurasia A, et al. ultralytics/yolov5: v6. 0-YOLOv5n'Nano'models, Roboflow integration, TensorFlow export, OpenCV DNN support[J]. Zenodo, 2021

[3] R. Varghese and S. M., "YOLOv8: A Novel Object Detection Algorithm with Enhanced Performance and Robustness," 2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS), Chennai, India, 2024.

[4] A. Mohan, H. K. Meena, M. Wajid and A. Srivastava, "FPGA-Based Real-Time Road Object Detection System Using mmWave Radar," in *IEEE Sensors Letters*, vol. 9, no. 4, pp. 1-4, April 2025.

[5] W. Chen, H. Wang, H. Li, Q. Li, Y. Yang and K. Yang,

"Real-Time Garbage Object Detection With Data Augmentation and Feature Fusion Using SUAV Low-Altitude Remote Sensing Images," in *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1-5, 2022.

[6] J. He and K. L. E. Law, "Deep Learning Models for Rotated Object Detection in Aerial Images: Survey and Performance Comparisons," in *IEEE Access*, vol. 12, pp. 180436-180457, 2024.

[7] T. Ye, Z. Zhang, X. Zhang and F. Zhou, "Autonomous Railway Traffic Object Detection Using Feature-Enhanced Single-Shot Detector," in *IEEE Access*, vol. 8, pp. 145182-145193, 2020.

[8] M. G. Ragab *et al*., "A Comprehensive Systematic Review of YOLO for Medical Object Detection (2018 to 2023)," in *IEEE Access*, vol. 12, pp. 57815-57836, 2024.

[9] S. -C. Huang, Q. -V. Hoang and T. -H. Le, "SFA-Net: A Selective Features Absorption Network for Object Detection in Rainy Weather Conditions," in *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 8, pp. 5122-5132, Aug. 2023.

[10] W. Diao, X. Sun, X. Zheng, F. Dou, H. Wang and K. Fu, "Efficient Saliency-Based Object Detection in Remote Sensing Images Using Deep Belief Networks," in *IEEE Geoscience and Remote Sensing Letters*, vol. 13, no. 2, pp. 137-141, Feb. 2016.

[11] Y. Zeng, T. Liang, Y. Jin and Y. Li, "MMI-Det: Exploring Multi-Modal Integration for Visible and Infrared Object Detection," in *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 11, pp. 11198-11213, Nov. 2024.

[12] S. Zhou et al., "A Method to Automatic Create Dataset for Training Object Detection Neural Networks," in IEEE Access, vol. 10, pp. 80505-80517, 2022.

[13] W. Cai, B. Zhang and B. Wang, "Scale-Aware Anchor-Free Object Detection via Curriculum Learning for Remote Sensing Images," in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 14, pp. 9946-9958, 2021

[14] J. Lai, J. Geng, X. Deng and W. Jiang, "DDFN: Deblurring Dictionary Encoding Fusion Network for Infrared and Visible Image Object Detection," in IEEE Geoscience and Remote Sensing Letters, vol. 20, pp. 1-5, 2023.

[15] L. Yang, Z. Zheng, J. Wang, S. Song, G. Huang and F. Li, "AdaDet: An Adaptive Object Detection System Based on Early-Exit Neural Networks," in IEEE Transactions on Cognitive and Developmental Systems, vol. 16, no. 1, pp. 332-345, Feb. 2024.

[16] O. Barinova, V. Lempitsky and P. Kholi, "On Detection of Multiple Object Instances Using Hough Transforms," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 34, no. 9, pp. 1773-1784, Sept. 2012.

[17] M. F. Riftiarrasyid, F. L. Gaol, H. Soeparno and Y. Arifin, "Suitability of Latest Version of YOLOv11 in Drone Development Studies," 2024 7th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI), Yogyakarta, Indonesia, 2024.