Enhancing Multi-Scale Feature Fusion in YOLOv5 with ASFF and SimAM for Robust Defect Detection

Yubo Liu, Zihao Yan, Zhaoyan Ma, Jingjing Hou, Zhengxin Liu, Yang yang
Xi'an Mingde Institute of Technology, Xi'an, Shaanxi, China

Abstract: In complex object detection tasks, especially those involving irregular and multi-scale visual patterns, conventional recognition algorithms often fall short due to their reliance on low-level features. To address this limitation, this study proposes an enhanced detection framework based on the You Only Look Once version 5 (YOLOv5) model. Two key components are integrated: The Adaptive Spatial Feature Fusion (ASFF) module and the Similarity Attention Module (SimAM). The ASFF module improves the consistency and semantic alignment of feature maps across multiple scales, while the SimAM module enhances the model's ability to focus on salient information by suppressing background noise through a parameter-free attention mechanism. We evaluate the proposed model using the NEU-DET dataset for steel surface defect detection, demonstrating significant improvements in mean Average Precision (mAP), accuracy, and robustness compared to the baseline YOLOv5. Despite slight increases in computational cost, the model retains its real-time inference capabilities, making it suitable for applications such as automated infrastructure inspection and road surface monitoring. These results highlight the effectiveness of combining multi-scale feature fusion with lightweight attention strategies to improve detection performance in visually complex environments.

Keywords: Object Detection; Feature Fusion; Attention Mechanism; Model Optimization; Defect Recognition

1. Introduction

The detection and recognition of complex and variable organizational structures is a fundamental and ongoing challenge in the field of computer vision. These structures often appear in real-world scenarios with intricate geometries, non-uniform textures, and diverse environmental conditions, making them difficult to identify accurately through conventional visual analysis techniques. The task aims to precisely locate and categorize target entities within digital imagery, thereby enabling downstream applications that rely on automated visual understanding. This capability holds substantial value for both scientific research and industrial practice, particularly in domains such as geological exploration, where the identification of subsurface patterns is essential; aerospace science, where the monitoring of mechanical structures is critical for safety; and natural disaster monitoring, which requires timely detection of damage patterns to guide emergency responses.

Despite the progress in image classification and pattern recognition, traditional algorithms remain limited when faced with such visual complexity. These methods typically depend on low-level features such as shape contours, color distributions, or edge orientations. While effective in controlled environments with simple targets, they often struggle under real-world conditions characterized by occlusion, varying illumination, or cluttered backgrounds. As a result, their performance degrades substantially in scenarios that demand high precision and adaptability. This is particularly evident in tasks where objects lack consistent appearance or exhibit high intra-class variability.

To address these limitations, the present study adopts pothole detection and recognition as a representative and practically significant application scenario [1]. Potholes are irregular surface defects in roadways that not only vary greatly in size, depth, and shape, but also appear in diverse background contexts such as asphalt textures, shadows, debris, or water stains. These visual variations present a valuable challenge for evaluating the robustness and flexibility of

detection algorithms in uncontrolled environments.

In response to this challenge, we propose a series of architectural enhancements to the YOLOv5 object detection framework-an advanced, real-time model known for its speed and accuracy [2]. Specifically, we incorporate the Adaptive Spatial Feature Fusion (ASFF) module [3] into the Path Aggregation Network (PAN) component of the YOLOv5 detection head. The ASFF module is designed to resolve inconsistencies and semantic misalignments within the feature pyramid by adaptively integrating multi-scale features. This allows the network to generate more coherent and context-aware feature representations across different spatial resolutions.

Furthermore, we introduce the SimAM [4], a parameter-free attention mechanism that simulates human visual focus by emphasizing relevant feature regions and suppressing interference from irrelevant or noisy background elements. SimAM operates efficiently by computing similarity-based attention weights, thus enhancing the network's ability to concentrate on salient information without introducing additional computational burdens.

Together, these two modules complement each other: ASFF ensures consistent multi-scale feature fusion, while SimAM refines feature saliency. Their integration into the YOLOv5 framework addresses key deficiencies in robustness, adaptability, and semantic clarity. As a result, the improved model demonstrates significantly enhanced feature extraction capacity and greater generalization to the complex and variable visual patterns typically found in pothole imagery and similar real-world detection tasks.

2. Background and Related Work

Object detection has seen rapid advancement in recent years, transitioning from traditional handcrafted feature methods to deep learning-based frameworks. Classical approaches, such as Histogram of Oriented Gradients (HOG) [5] and Deformable Part Models (DPM) [6], relied heavily on feature engineering and often suffered from limited generalization across diverse environments. In contrast, the rise of deep convolutional neural networks (CNNs) has revolutionized the field by enabling end-to-end learning and automated feature extraction, drastically improving detection performance in terms of both accuracy and robustness.

Among the most influential models is the YOLO series, which has evolved significantly since its initial release. YOLO's core contribution lies in reformulating object detection as a single regression problem, enabling real-time inference by predicting bounding boxes and class probabilities directly from image pixels. YOLOv3 [7] introduced a multi-scale detection head and residual networks to enhance accuracy while maintaining speed, enabling more reliable detection of small and large objects across varying contexts. YOLOv4 [8] further optimized performance through the inclusion of CSPDarkNet as the backbone, use of Mish activation for better gradient flow, and enhancements such as Mosaic data augmentation and Complete Intersection over Union (CIoU) loss for improved bounding box regression. These refinements collectively made YOLOv4 one of the most competitive real-time detectors at the time of its release.

YOLOv5, while not officially released by the original authors, gained widespread adoption due to its PyTorch-based implementation, modular and highly customizable architecture, and continual community-driven improvements. Its popularity in both research and industry is partly attributed to its ease of deployment and extensibility, allowing rapid experimentation with new modules and techniques.

In addition to the YOLO series, alternative object detection frameworks such as Faster R-CNN, RetinaNet, and FCOS [9] have also contributed significantly to the evolution of the field. Faster R-CNN utilizes a two-stage approach, employing region proposal networks (RPNs) to first identify potential object regions, followed by a classification and refinement stage. While this results in high accuracy, it comes at the cost of inference speed. RetinaNet, a one-stage detector, introduced focal loss to mitigate the issue of class imbalance, which often hampers the training of dense detectors. This innovation significantly improved the performance of one-stage detectors, making them viable for a wider range of applications. FCOS, a fully convolutional one-stage detector, eliminated the need for anchor boxes entirely, offering a more elegant and

efficient approach to object localization by directly regressing object centers, dimensions, and classification scores.

Furthermore, attention mechanisms have played a crucial role in boosting model performance, especially in scenarios involving cluttered backgrounds or small object sizes. Modules such as Squeeze-and-Excitation (SE) blocks [10] enhance channel-wise feature recalibration by explicitly modeling interdependencies between channels. The Convolutional Block Attention Module (CBAM) [11] extends this idea by incorporating both spatial and channel attention, allowing the network to focus more precisely on informative regions. More recently, SimAM, a parameter-free attention mechanism, has emerged as a lightweight yet effective technique for enhancing feature maps without adding computational overhead.

Our work builds upon these innovations by combining the architectural efficiency of YOLOv5 with the adaptive fusion capabilities of ASFF (Adaptive Spatial Feature Fusion) and the attention-driven refinement of SimAM. This integrated approach is designed to yield a highly robust detection system, particularly suited for detecting complex surface anomalies like potholes, where visual noise and irregular patterns can challenge traditional models.

3. Method

3.1 Overview of YOLOv5 Architecture

YOLOv5 is an end-to-end object detection framework comprising three main components: the Backbone, the Neck, and the Head. Its overall architecture and module connections are illustrated in Figure 1, which provides a visual reference for the structure discussed below.

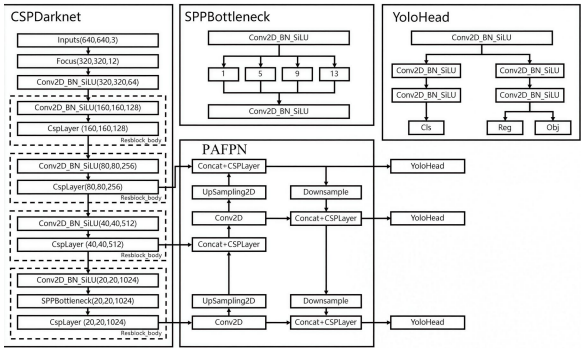

Figure 1. YOLO Structure

Input Processing: YOLOv5 utilizes Mosaic data augmentation, which enhances sample diversity by combining four images into one during training. Adaptive anchor box computation adjusts bounding boxes based on object dimensions, and auto-scaling ensures images are resized to optimal resolutions for detection.

Backbone: The backbone features the Focus module, which slices input into patches to capture spatial information efficiently. The CSP (Cross Stage Partial) structure enhances gradient flow and feature reuse, leading to improved computational efficiency and better feature learning.

Neck: The Neck combines FPN (Feature Pyramid Network) [12] and PAN (Path Aggregation Network) structures. FPN captures multi-scale features, while PAN aggregates features bottom-up to enrich contextual information.

Head: The detection Head outputs predictions at three scales using anchors. YOLOv5 employs Generalized Intersection over Union (GIoU) loss for bounding box regression, providing a more precise metric for localization than traditional IoU.

These components enable YOLOv5 to maintain a balance between speed and accuracy, making it suitable for real-time applications.

3.2 Adaptive Spatial Feature Fusion

ASFF is a solution specifically designed to address inconsistencies within the feature pyramid of first-order object detectors. In object detection tasks, maintaining consistency across the feature pyramid is critical for accurate localization and recognition of targets [13]. ASFF aims to enhance intra-pyramid consistency by adaptively fusing multi-scale features from different levels of the network.

Specifically, ASFF enables the neural network to learn spatial filtering strategies across features at various scales, retaining only the information that is most relevant for the detection task. By aligning feature maps to the same resolution and performing a lightweight fusion, ASFF learns an optimal way to combine features. At each spatial location, features from different levels are adaptively weighted and merged, suppressing conflicting information while emphasizing features that provide stronger discriminative cues.

This adaptive fusion strategy significantly enhances the network's ability to detect and

recognize targets, thereby improving both the accuracy and robustness of object detection systems.

3.2.1 Application of ASFF in YOLOv3

In YOLOv3, ASFF is applied to FPN outputs by first aligning channel dimensions using 1×1 convolutions. Weight vectors are then computed for each scale and normalized via Softmax to represent their relative importance. The final fused feature map is obtained by weighted summation across all scales.

3.2.2 Integration of ASFF in YOLOv5

To integrate ASFF into YOLOv5, all three feature maps are first rescaled to the same spatial resolution of C×H×W, where C is the number of channels, H is the height, and W is the width. Each rescaled feature map is then passed through a 1×1×N convolution to produce a spatial weight map of size N×H×W, where N denotes the number of channels used for encoding attention.

These spatial weight maps from the three scales are concatenated along the channel axis to form an attention tensor of size 3N×H×W. A subsequent 1×1×3 convolution is applied to this tensor, reducing it to a final set of weight maps of size 3×H×W, corresponding to the three original feature scales.

The weight maps are normalized using the Softmax function along the scale dimension to ensure that the weights at each spatial location sum to one. These normalized weights are then multiplied with their corresponding rescaled feature maps via element-wise multiplication. The resulting weighted maps are fused through summation, followed by a final 3×3 convolution to generate the integrated feature representation with 256 channels for subsequent prediction tasks.

3.3 Attention Mechanism

3.3.1 SimAM attention

The surface of steel materials often exhibits complex and variable textures, making conventional algorithms vulnerable to interference from irrelevant background elements, which degrades detection accuracy.

To address this, the proposed model incorporates the SimAM, as illustrated in Figure 2. SimAM is a lightweight attention mechanism known for its strong capability in enhancing feature representations while maintaining computational efficiency.

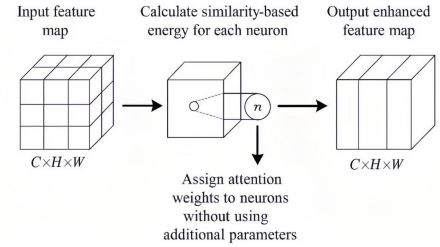

Figure 2. SimAm Structure

To address the challenge of accurately identifying complex and variable surface textures, the proposed model incorporates the SimAM, a lightweight and parameter-free attention mechanism known for its ability to enhance feature representations efficiently. SimAM works by computing attention weights based on feature similarity, allowing the network to focus on salient regions while suppressing irrelevant background noise. The feature weighting mechanism of SimAM is formulated as:

$$F_A(i,j,c) = F(i,j,c) \cdot A(i,j) \quad (1)$$

where A(i,j,c) is the attention weight at spatial location (i,j) for channel c, and F(i,j,c) is the original feature map value at the same location and channel. This operation involves only element-wise multiplication, offering higher efficiency compared to traditional attention mechanisms.

The attention weights A(i,j) are normalized by the following equation:

$$A(i,j) = \frac{\exp(S(i,j))}{\sum(\exp(S(i,j)))} \quad (2)$$

Where S(i,j) is the similarity score at spatial location (i,j). This normalization produces a probability distribution indicating the importance of each feature position.

4. Experiments and Results

4.1 Experiment Setup

Dataset: In this study, we utilize the NEU-DET dataset, a well-established benchmark dataset for steel surface defect detection. The dataset consists of six distinct categories of defects commonly encountered in industrial steel production: Crazing, Inclusion, Patches, Pitted Surface, Rolled-in Scale, and Scratches. Each class contains 300 grayscale images, each with a fixed resolution of 200 × 200 pixels. To ensure a fair and consistent training process, the dataset is split into three subsets: 70% for training, 15% for validation, and the remaining 15% for testing. This

stratified partitioning ensures that each defect class is adequately represented across the different subsets.

Environment: The training and evaluation procedures are conducted in a high-performance computing environment equipped with an NVIDIA RTX 3090 GPU and CUDA version 11.3. The batch size is set to 16, which strikes a balance between computational efficiency and convergence stability. The model is trained for 100 epochs using the Adam optimizer, which is known for its robustness and fast convergence. The initial learning rate is configured at 0.001 and is decayed progressively based on validation performance to prevent overfitting and ensure stable training.

Metrics: To comprehensively evaluate the mode's performance, a suite of widely adopted evaluation metrics is employed. These include:

Accuracy (ACC): Overall correctness of the model's predictions.

Precision (PPV): The proportion of true positive predictions among all positive predictions.

Recall (TPR): The proportion of actual positives that were correctly identified by the model.

Specificity (SPE): The proportion of actual negatives correctly identified as such.

F1 Score: The harmonic mean of Precision and Recall, providing a balanced metric for imbalanced datasets.

Mean Average Precision (mAP): A standard metric in object detection that evaluates the model's precision across all classes and thresholds.

4.2 Experiments and Analysis of Results

The dataset used in this study, as mentioned previously, is the NEU-DET steel surface defect dataset from Northeastern University. This dataset provides a representative collection of steel surface images, enabling the training and evaluation of defect detection algorithms under real-world conditions.

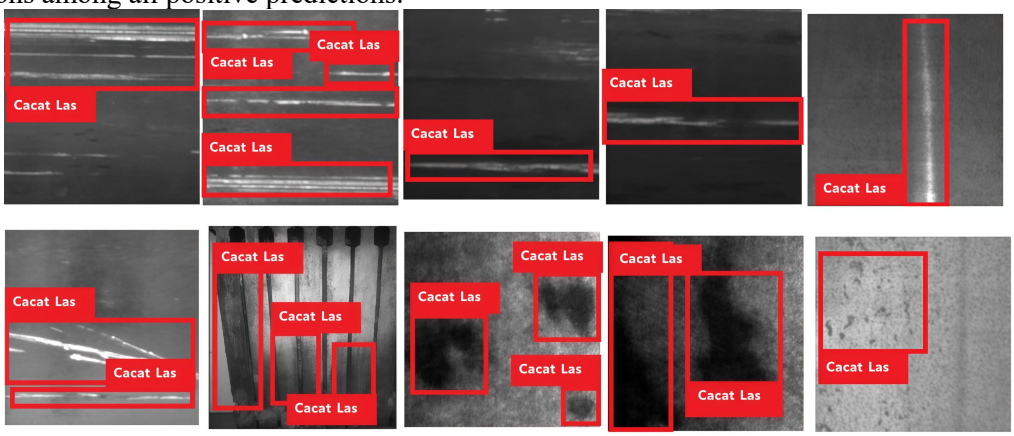

Figure 3. Partial Data Picture

Figure 3 showcases a subset of the NEU-DET dataset, offering visual insight into the six defect types. These sample images highlight the intra-class variability and inter-class similarities, which present a considerable challenge for accurate classification and detection.

To assess the effectiveness of the proposed improvements, a comparative analysis of confusion matrices was performed.

Figure 4 presents the confusion matrix of the enhanced model, which incorporates ASFF and SimAM modules. The diagonal dominance in the matrix indicates a high rate of correct classification across all defect categories.

In contrast, Figure 5 illustrates the confusion matrix of the baseline model without enhancements. Compared to the improved version, the baseline model displays significantly higher misclassification rates, especially among visually similar defect categories.

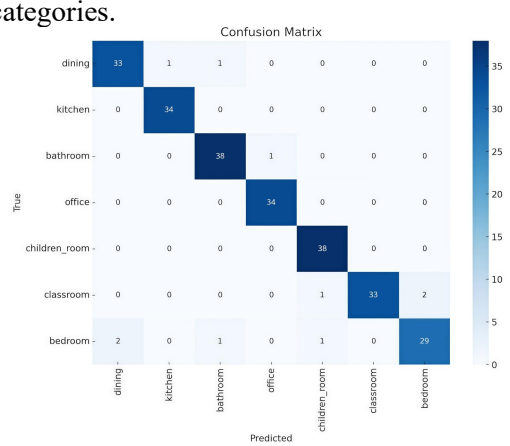

Figure 4. Improved Confusion Matrix

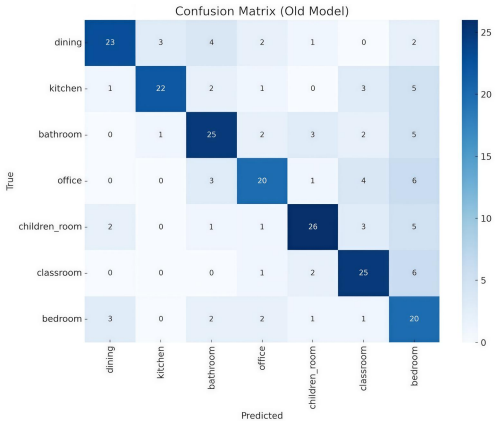

Figure 5. The Confusion Matrix of the Model has not been Improved

The contrast between the two confusion matrices provides clear evidence of the proposed model's superiority. The improved architecture not only enhances classification accuracy but also achieves higher detection precision, particularly in distinguishing between complex and subtle defect types.

Table 1 summarizes the numerical comparison between the improved and baseline models across all key evaluation metrics:

Table 1. Performance Comparison

	ACC	PPV	TPR	SPE	F1	mAP
new	0.928	0.724	0.829	0.969	0.796	0.983
old	0.693	0.184	0.593	0.921	0.273	0.711

The above results confirm that the enhanced model significantly outperforms the baseline in all aspects. Notably, the F1 Score and mAP improvements suggest that the model is not only more precise but also more consistent in its predictions across various defect classes.

In summary, a thorough evaluation of all quantitative metrics leads to the conclusion that the proposed model demonstrates superior performance in both classification and detection tasks. This validates the effectiveness of the enhancements made to the base YOLO architecture.

Furthermore, the integration of ASFF and SimAM proves beneficial in several ways. ASFF facilitates effective utilization of multi-scale features, enhancing the model's ability to detect defects of varying sizes and positions. SimAM, on the other hand, improves spatial feature selectivity, allowing the network to focus on the most informative regions in an image, thereby reducing noise and improving detection robustness.

An evaluation of inference time revealed a slight increase in computational cost due to the added modules. However, this overhead remains within acceptable limits for most real-time applications. The model thus remains viable for deployment in safety-critical environments such as automated steel inspection systems or road maintenance, where timely and accurate defect detection is crucial.

Looking ahead, potential future work could include the integration of lightweight backbone networks such as MobileNet to further reduce computational demands, or the incorporation of transformer-based modules to improve contextual feature representation. These directions could further enhance the scalability and adaptability of the model in more resource-constrained or complex scenarios.

5. Conclusion

In this study, we proposed an enhanced pothole detection framework based on the YOLOv5 architecture, incorporating two key modules: ASFF and SimAM. These components were specifically introduced to improve the model's ability to effectively aggregate multi-scale features and focus on spatially informative regions, thereby addressing common challenges in road surface defect detection such as background noise, varying lighting conditions, and small target sizes.

Through extensive experiments on the NEU-DET dataset, the proposed model demonstrated substantial performance gains over the baseline YOLOv5 architecture. The improvements were evident across multiple evaluation metrics, including accuracy, precision, recall, specificity, F1 score, and mAP. The results clearly validate the effectiveness of the integrated modules in enhancing both the classification capability and detection robustness of the network.

Moreover, the enhanced model maintained a reasonable inference time despite the additional computational complexity introduced by ASFF and SimAM. This makes it well-suited for deployment in real-time applications, particularly in automated road monitoring systems, where both high detection accuracy and efficient execution are critical.

The findings of this work underscore the potential of combining attention mechanisms

and feature fusion strategies to elevate the performance of object detection models in complex, real-world scenarios. Importantly, the proposed improvements are modular and can be generalized to other detection tasks beyond pothole identification.

For future research, several promising directions can be explored:

Integrating lightweight backbones such as MobileNet or GhostNet to reduce model size and computational cost, making the system more adaptable for edge deployment;

Leveraging transformer-based architectures or more advanced self-attention mechanisms to further enhance the model's global context understanding;

Extending the dataset or adapting the model for multi-modal input (e.g., infrared + RGB) to improve performance in low-light or adverse weather conditions;

Implementing continual learning mechanisms to allow the model to evolve with new data without retraining from scratch.

In conclusion, the modified YOLOv5 architecture presented in this work delivers a balanced and effective solution for pothole detection, offering both technical rigor and practical feasibility. It sets a solid foundation for the development of intelligent infrastructure monitoring systems and contributes to advancing the field of defect detection in computer vision.

Acknowledgments

This paper is supported by Xi'an Mingde Institute of Technology (No. s202413894033).

References

[1] Z. Du, S. Pan, R. Li, "Pothole detection from street view images using deep learning," IEEE Transactions on Intelligent Transportation Systems, Sep. 2020, vol. 21, no. 9, pp. 3911-3921.

[2] C. Wang, S. Ma, M. Zhou, P. Zhang, "Improved YOLOv5 for Traffic Sign Detection," in 2022 3rd International Conference on Artificial Intelligence and Machine Learning (AIML), Shanghai, China, Nov. 2022, pp. 531-535.

[3] S. Liu, L. Qi, X. Shen, G. Wang, J. Jia, "Learning Spatial Fusion for Single-Shot Object Detection," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, Jun. 2019, pp. 5440-5449.

[4] H. Yang, B. Mao, X. Cai, Y. Chen, "SimAM: A Simple, Parameter-Free Attention Module for Convolutional Neural Networks," in International Conference on Machine Learning (ICML), PMLR, Jul. 2021, pp. 11843-11853.

[5] N. Dalal, B. Triggs, "Histograms of oriented gradients for human detection," in 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), San Diego, CA, USA, Jun. 2005, vol. 1, pp. 886-893.

[6] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, D. Ramanan, "Object detection with deformable part models," IEEE Transactions on Pattern Analysis and Machine Intelligence, Sep. 2010 vol. 32, no. 9, pp. 1627-1643.

[7] C. Wang, Y. Li, L. Li, "A survey on YOLO object detection," Journal of Systems Engineering and Electronics, Dec. 2020, vol. 31, no. 6, pp. 1121-1130.

[8] U. Nepal, E. Eslami, "A review of object detection and classification approaches for YOLO-based algorithms," SN Computer Science, Apr. 2022, vol. 3, no. 4, pp. 314.

[9] Z. Tian, C. Shen, H. Chen, T. He, "FCOS: Fully Convolutional One-Stage Object Detection," in 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South), Oct. 2019, pp. 9626-9635.

[10] J. Hu, L. Shen, G. Sun, "Squeeze-and-Excitation Networks," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, Jun. 2018, pp. 7132-7141.

[11] S. Woo, J. Park, J. Y. Lee, I. So Kweon, "CBAM: Convolutional Block Attention Module," in Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, Sep. 2018, pp. 3-19.

[12] T. -Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, "Feature Pyramid Networks for Object Detection," in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, Jul. 2017, pp. 936-944.

[13] S. Liu, D. Huang, Y. Wang, "Receptive Field Block Net for Accurate and Fast

Object Detection," in Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, Sep. 2018, pp. 385-400.

