Micro Object Detection

Tushar Mitra
Computer Science and Engineering
Kalinga Institute of Industrial Technology
Bhubaneswar,India
tushar24.mitra@gmail.com

Apoorva Aanand
Computer Science and Engineering
Kalinga Institute of Industrial Technology
Bhubaneswar,India
apoorvaaanand28@gmail.com

Durjaya Das
Computer Science and Engineering
Kalinga Institute of Industrial Technology
Bhubaneswar,India
durjayadas3@gmail.com

Prachi Sayesha Parida
Computer Science and Engineering
Kalinga Institute of Industrial Technology
Bhubaneswar,India
prachiparida2005@gmail.com

Utkarsh Dubey
Computer Science and Engineering
Kalinga Institute of Industrial Technology
Bhubaneswar,India
dubeyutkarsh021@gmail.com

Sidhant Dash
Computer Science and Engineering
Kalinga Institute of Industrial Technology
Bhubaneswar,India
lostspace747@gmail.com

***Abstract*—Micro object detection remains a significant challenge in computer vision due to limitations in feature extraction and computational efficiency for small-scale targets. This paper introduces YOLOv12s, a deep learning-based architecture optimized for micro object detection, combining advancements in attention mechanisms and lightweight network design. The proposed model integrates a Residual Efficient Layer Aggregation Network (R-ELAN) backbone and an Area Attention Module to enhance feature representation while reducing redundant computations. Key innovations include Flash Attention for accelerated inference on GPU/CPU architectures, 7×7 Separable Convolutions to expand receptive fields without increasing parameters and Feature Map Segmentation for precise localization of micro-objects.**

Experimental results on a small sample of the xView satellite imagery dataset shows a training mAP@0.5 of 0.41 and training mAP@[0.5:0.95] of 0.19. For the validation part mAP@0.5 of 0.194 and mAP@[0.5:0.95] of 0.399 was achieved. Testing mAP@0.5 and mAP@[0.5:0.95] were 0.102 and 0.235 respectively. The model was trained, validated and tested on 4 classes - boat, building, plane and vehicle. It gave good and satisfactory results on 3 of the 4 classes - boat, building and plane. The vehicle class very tiny and therefore it requires further future work for getting better and promising results.

***Index Terms*—Micro Object, YOLOv12s, Deep Learning, Attention Mechanisms, Computer Vision, CNN**

I. Introduction

Detecting objects is a key problem domain in computer vision, which generally entails discovering and localizing objects within a 2D image or 3D video. Object detection is a critical component for many areas where correctly identifying the objects location is essential, such as autonomous vehicles, surveillance, medical imaging, and industrial automation, to name just a few. Although the field has seen significant advancement, especially using deep learning methods, the detection of micro-objects—small-scale objects with fine details—remains a difficult problem area which is weakened due to object size, low resolution and noise [19]; [4].

A. Importance of Micro-Object Detection

Micro-object detection is simply necessary for domains where a high amount of precision and accuracy is required. Some of the most relevant domains where micro-object detection is critical are: Medical Imaging: The detection of tumors, microbleeds, and cellular abnormalities in radiological images (i.e. X-rays, MRIs, and CT scans) [19]. Increasing the detection of those types of medical imaging data is accomplished using various techniques, including deep learning based segmentation models. Remote Sensing: The ability to identify small-scale objects like vehicles, buildings, and wildlife in aerial and satellite imagery; However, the metadata from high resolution satellite images often suffers from scale variance issues that prevents effective micro-object detection. Security and Surveillance: Observing small objects in complicated contexts mainly for security and anomaly detection. Recent advances in small object detection models have greatly improved real-world accuracy in security [18]. Industrial Inspection: Quality control in large and industrial-scale manufacturing situations to ensure defects are detected in products. Micro-object detection models have improved real-world identification of defects for example in semiconductor manufacturing and microscopic assessments [5].

B. Recent Developments in Deep Learning for Micro-Object Detection

Traditional object detection methods, such as Region-based Convolutional Neural Network (R-CNN) [16], Feature Pyramid Networks (FPN), and Single Shot MultiBox Detector (SSD) [10], laid foundational work for object detection and class recognition. However, there has been a significant increase in deep learning-based methods, especially convolutional neural network (CNN) and transformer-based approaches, that have greatly improved recognition and detection accuracy for micro-object detection work [3].These methods include (YOLO) You Only Look Once. The YOLO family of models has established accurate real-time/object detection processing. Several YOLO modifications/models have been created that are specialized for micro-object detection: YOLOv3: Incorporates multi-scale feature extraction able to detect small objects specifically [15]. YOLOv4: Builds on previous work with CSPDarknet53 and enhancements with spatial pyramid pooling (SPP) [1]. YOLOv5 is an extremely effective object detection system that used adaptive anchor tuning and a corresponding set of focus layers [6]. YOLOv7: is a state-of-the-art version of YOLO which utilized advanced feature fusion based on the new Extended Efficient Layer Aggregation Networks (E-ELAN) [19]. YOLO for Medical Object Detection: More recent research has examined YOLO for micro-object such as identifying small brain abnormalities or retinal damage, showing great potential for medical applications [13]. Furthermore, super-resolution can improve YOLO object detection of micro-objects [2].Micro-object detection can be performed at low resolution through super-resolution networks situated in the detection pipelines.

C. Challenges in Micro-Object Detection

There are still challenges in detecting micro-objects, although it has advanced. Scale variability: The scale, that is associated with micro-object size often makes distinguishing the micro-object from the background noise very difficult [18]. Low resolution: Object feature extractors that do exist, such as traditional object detection systems, have a limit to the identification of an object occupying a few pixels and will require image enhancement techniques from imaging at higher resolution [5]. Occlusion and Clutter: Identifying distinct patterns among overlapping objects, especially in denser environments, creates obstacles in accurate segmentation and classification. Computational Constraints: For real-time detection, computational efficiency remains a challenge especially in environments where numerous devices could be constrained by factors such as localization or embedded systems. improve [4]. Generalization Across Domains: In most cases, the performance of micro-object detection models is contingent on a specific dataset. On occasion, previous or alternative datasets will require domain adaptation to assist with improved generalization. [7]

D. Future Directions

Researchers are now trying to improve micro-office detection through the use of Hybrid models that combine Convolutional Neural Networks (CNNs) and transformers to further improve features representation layered throughout the decision detect micro-object scenarios or comparable self-identifying dimensional paradigms [3]., Generative Adversarial Network (GAN) models for data-augmentation methods on synthetic data to improve detection performance when capacity is low [18]., Improved research in Edge Computing devices to display improvements in micro object detection in IoT and embedded systems where low latency elements are crucial [14], Self-supervised Learning techniques to mitigate the demands of limited labeled micro-object datasets [7].

II. Related Works

A paper published in 2022 Developed MMOC-Net, which is a two-stage cascade network that combines U-Net and Full-Resolution Network (FRN) with Residual Atrous Spatial Pyramid Pooling (R-ASPP) in favor of cerebral microbleed (CMB) segmentation in SWI-MRI. It has accomplished 87.93% DSC and 90.69% F2-score, which addressed the challenges of small size (¡10mm) and visual similarity to anatomical structures [19].In 2024,research was conducted that Provides research synthesis of YOLO variants that is used in medical detection (2018-2023), which highlights their adaptation for small anatomical structures and restraints in handling class imbalance and low-contrast targets [13].A work in 2021 has been proposed that Analyzed obstacles in optical remote sensing: high intra-class variance, complex backgrounds, and scale variations. Diagnosed multi-scale feature fusion and context-aware networks as fundamental measures [4].In 2024, a group of academics conducted research on In-depth review of deep learning approaches, accentuating feature pyramid networks, super-resolution techniques, and hybrid loss functions to tackle low-resolution and class imbalance [18].In 2023, a paper was published which Emphasized advancements in attention mechanisms and transformer-based architectures for small object localization, with targeted approach on medical and satellite imaging [8].A work published in 2015, Presented Faster R-CNN with Region Proposal Network (RPN), facilitate near real-time detection by sharing undulation features between proposal generation and detection stages [16].A work in 2017 has been proposed that Recommended Feature Pyramid Networks (FPN), utilizing intrinsic CNN hierarchies to create multi-scale feature maps. Upgraded COCO AP by 2.3 points over single-scale baselines [9].In 2016, a paper was published which Outlined Single Shot Multibox Detector (SSD), integrating predictions from multiple feature maps for competent multi-scale detection without region proposals [10].In 2020, a study was conducted on Vision Transformers (ViTs) validated scalability to high-resolution medical images, despite the fact that computational demands remain arduous for 3D datasets [3].A work in 2022 has been proposed that discussed Auto-associative learning approach for microscopy detection, reducing annotation dependency through contrastive learning forthcoming methodology for medical micro-object detection [11].One of the possibilities listed includes YOLOv12 embedded channel-wise attention and dynamic convolutions, achieving 63.8 AP on COCO-small subset - relevant for real-time medical applications, In 2025 [17].

III. Proposed Model

A. Architecture

YOLOv12 small [17] consists of three primary structural components.

Backbone: R-ELAN (Residual Efficient Layer Aggregation Network) - It processes input images through multiple convolutional layers. This layer features specialized C3k2 blocks with kernel size 2 for efficient feature extraction. It incorporates Area Attention modules(A2C2f) that distribute computational resources selectively. It also has Progressive spatial dimension reduction (640→320→160→80→40→20) with corresponding channel expansion (64→128→256→512→1024).

Neck: Feature Fusion with Area Attention - This layer integrates features across multiple scales through bidirectional pathways. It employs Flash Attention for memory-efficient feature processing. It uses depth wise separable convolutions(7x7) to reduce computational demands. The Neck creates robust multi-scale representations through skip connections and feature aggregation.

Head: Detection and Prediction - This is the final layer of YOLOv12 small architecture. It generates final predictions using feature maps from three different scales (P3, P4, P5). This layer produces bounding box coordinates, class probabilities, and confidence scores. It implements multi-scale feature fusion to handle objects of varying sizes. It utilizes specialized loss functions that balance localization and classification objectives.

B. Algorithm

The algorithmic process of YOLOv12 small follows a streamlined detection pipeline that emphasizes both accuracy and computational efficiency.

Input Processing and Feature Extraction - The Input images (640×640×3) are processed through the backbone's initial convolutional layers. The backbone progressively extracts multi-scale features through R-ELAN blocks. Area Attention modules (A2C2f) selectively focus computational resources on informative regions. Multiple feature maps at different resolutions are generated, representing various semantic levels.

Feature Fusion and Enhancement - The neck combines features from different scales through upsampling and concatenation.Flash Attention mechanism processes segmented feature maps more efficiently:

$$\text{Attention}(Q, K, V) = \text{SoftMax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V$$

Where Q, K, V are query, key, and value matrices derived from segmented feature maps. Depthwise separable convolutions refine spatial information with minimal parameter overhead. Skip connections between backbone and neck preserve gradient flow and spatial information.

Object Detection and Prediction - The head processes refined feature maps at three different scales (P3, P4, P5). For each position in these feature maps, the model predicts Bounding box coordinates relative to grid cells, Objectness scores indicating likelihood of object presence and Class probabilities across all defined categories.

Multi-scale feature fusion enhances detection performance across object sizes. The model's loss function integrates localization, classification, and confidence components:

$$L = \lambda_{\text{coord}} \sum \left((\hat{x} - x)^2 + (\hat{y} - y)^2\right) + \lambda_{\text{obj}} \sum (\hat{C} - C)^2 + \dots$$

Where coordinates $(\hat{x}, \hat{y})$ and confidence $\hat{C}$ represent predictions.

C. Block Diagram

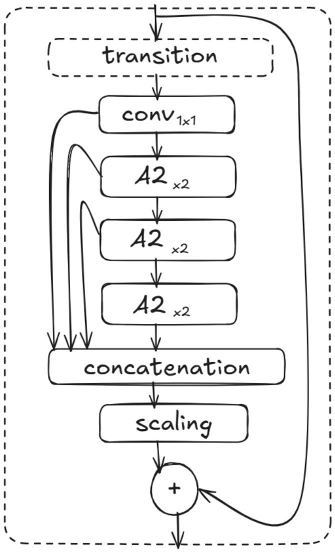

Fig. 1: Block Diagram

D. Performance

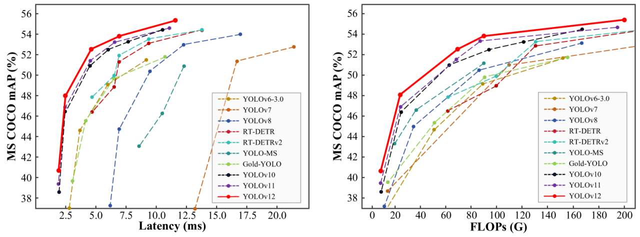

Fig. 2: Latency(ms) and FLOPs(G)

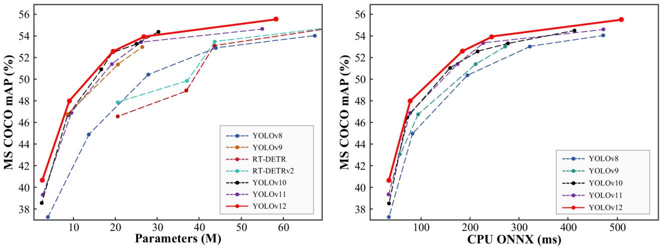

Fig. 3: Parameters(M) and CPU ONNX(ms)

IV. Results Analysis

A. Table

Class	Images	Instances	Box (p)	r	mAP50	mAP50-95
all	8	2020	0.446	0.386	0.399	0.194
boat	4	159	0.475	0.239	0.264	0.067
plane	6	513	0.580	0.678	0.679	0.380
building	3	48	0.545	0.604	0.582	0.310
vehicle	5	1300	0.1838	0.0238	0.0714	0.0191

validation metrics for different classes

Class	Images	Instances	Box (p)	r	mAP50	mAP50-95
all	8	2020	0.446	0.386	0.399	0.194
boat	4	159	0.475	0.239	0.264	0.067
plane	6	513	0.580	0.678	0.679	0.380
building	3	48	0.545	0.604	0.582	0.310
vehicle	5	1300	0.1838	0.0238	0.0714	0.0191

testing metrics for different classes

	mAP		Precision				Recall			
Dataset	@[0.5:0.95]	@0.5	Boat	Building	Plane	Vehicle	Boat	Building	Plane	Vehicle
Validation	0.194	0.399	0.47541	0.58021	0.54543	0.18279	0.23899	0.67836	0.60417	0.023846
Testing	0.102	0.235	0.38824	0.20476	0.69798	0.35511	0.19666	0.44949	0.48148	0.013749

B. Graph

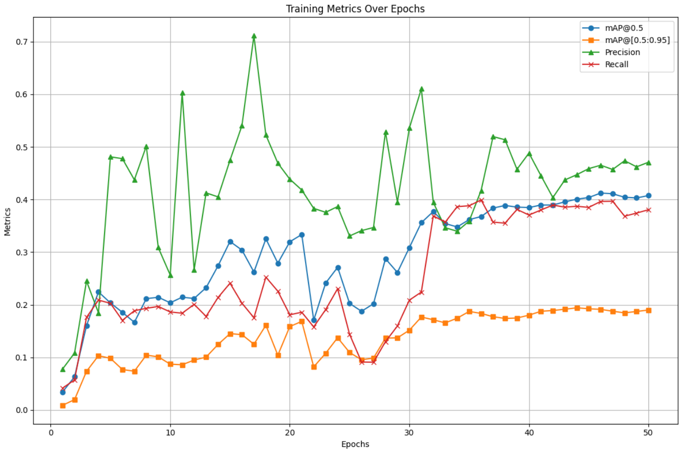

Fig. 4: Graph for change in metrics with epochs

C. Sample Dataset

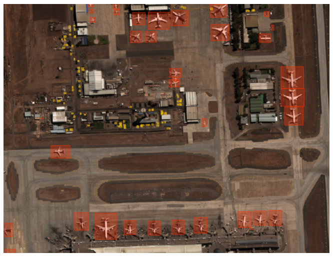

Fig. 5: Train Image 1

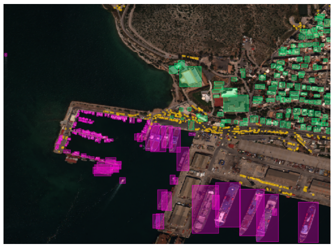

Fig. 6: Train Image 2

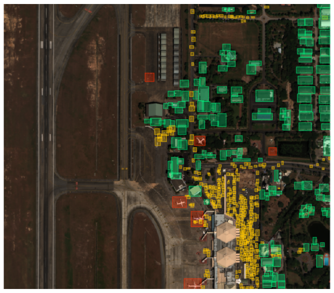

Fig. 7: Train Image 3

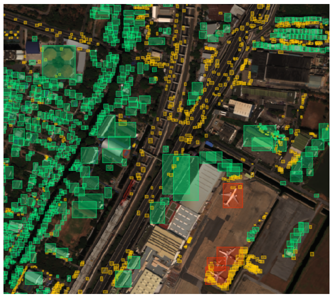

Fig. 8: Train Image 4

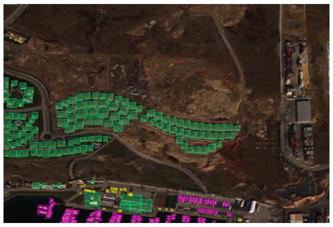

Fig. 9: Validation Image 1

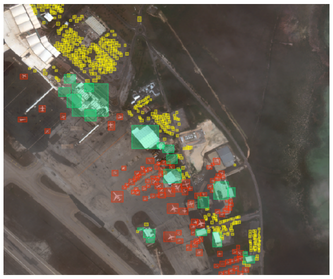

Fig. 10: Validation Image 2

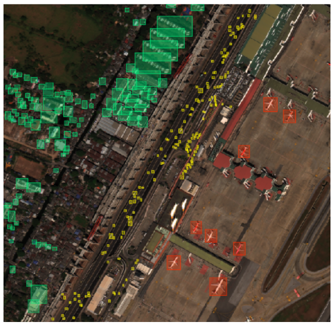

Fig. 11: Validation Image 3

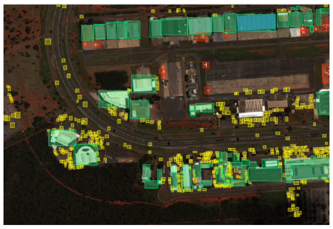

Fig. 12: Validation Image 4

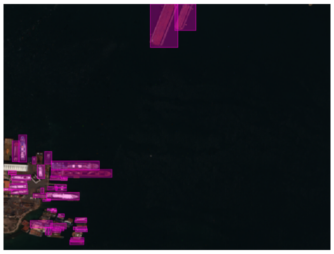

Fig. 13: Testing Image 1

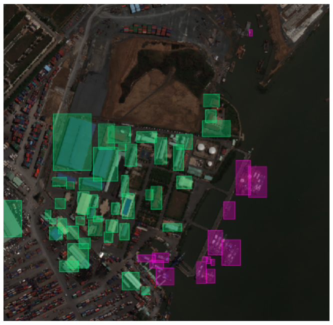

Fig. 14: Testing Image 2

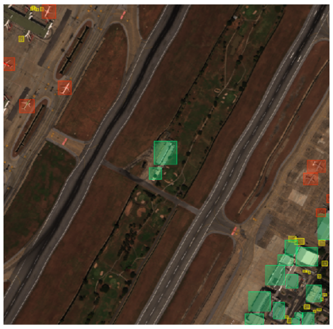

Fig. 15: Testing Image 3

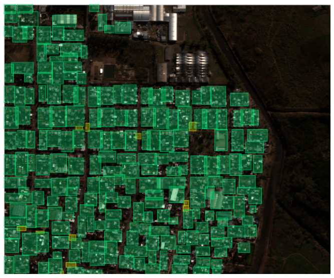

Fig. 16: Testing Image 4

D. Comparison with other people's works

Micro object detection has emerged as a challenging yet crucial subfield of computer vision, with distinctive characteristics that differentiate it from general object detection. This section compares our approach with existing methods in the literature, highlighting key innovations, limitations, and performance metrics across different methodologies for detecting extremely small objects.

Methodological Approaches to Micro Object Detection

Scale-Aware Methods - Scale-aware approaches represent a significant category of methods addressing micro object detection. These techniques typically modify feature pyramid networks (FPNs) to better handle extremely small-scale objects. Standard FPNs can negatively impact tiny object detection, leading researchers to introduce statistically estimated fusion factors to control how deep and shallow features are combined [8].Other works have introduced more sophisticated scale-aware mechanisms. The misalignment between deep and shallow features was addressed by developing an Image Pyramid Transformation Module (IPGT) [8]. Some researchers have combined feature-fusion with additional techniques such as attention mechanisms to enhance detection of extremely small objects across different feature map layers [8].

Attention-Based Methods - Attention mechanisms have proven particularly effective for micro object detection by suppressing background noise and highlighting relevant features. Multiple approaches incorporate attention modules to enhance feature representation of small objects.

Recurrent Neural Networks (RNNs) were utilized with attention to focus on relevant image areas [8]. The Attention-Guided Balanced Pyramid (ABP) adaptively fuses features at different pyramid levels using a two-part attention-based sub-network [8]. Several works have adapted channel attention mechanisms based on Squeeze-and-Excitation (SE) blocks to highlight channels relevant to detection while suppressing noise [8].

In the context of micro-scale crack detection, a self-attention mechanism was introduced after each pooling step to allow the model to focus on specific regions of feature maps, particularly helping identify micro-cracks that might otherwise be missed [12]. This approach demonstrated an average IOU of 0.511 for all micro cracks and 0.631 for larger micro cracks ($> 4\ \mu$m) [12].

Focus-and-Detect Methods - Several approaches employ a two-stage detection strategy where the first stage identifies regions of interest and the second stage performs detailed analysis. This focus-and-detect methodology has shown promising results for micro object detection by allowing the model to concentrate computational resources on areas likely to contain microscopic objects.

A common implementation divides the original image into tiles and then selects specific tiles for fine detection. However, regular grid sampling can lead to errors when objects span multiple tiles [8]. More sophisticated approaches use coarse cluster proposals to guide fine detection, with selected tiles being adjusted before resizing to maintain scale consistency [8].

Data Augmentation Strategies - Data augmentation techniques specifically designed for micro object detection address the challenge of limited training data. These methods typically involve duplicating images with small objects and performing copy-paste operations to increase the representation of micro instances in the training set [8].However, simple duplication can introduce noise, requiring semantic masks to precisely crop objects. Furthermore, in domains such as aerial imagery, objects tend to occupy specific areas, making random placement counterproductive [8]. More advanced approaches incorporate semantic segmentation networks to extract environmental context and properly handle object scale [8].

Self-Supervised Learning Approaches - A notable innovation in the field is the development of self-supervised learning techniques that can detect microscopic objects without extensive labeled datasets. LodeSTAR (Localization and detection from Symmetries, Translations And Rotations) exploits roto-translational symmetries to enable training on extremely small datasets—down to a single image—without ground truth [11].LodeSTAR achieves sub-pixel root mean square error (RMSE) and outperforms traditional methods in accuracy, particularly when analyzing challenging experimental data containing densely packed cells or noisy backgrounds [11]. This represents a significant advancement for micro object detection in fields like microscopy, where labeled data is scarce and difficult to generate.

V. Conclusion

This research introduces YOLOv12s, a novel deep learning-based architecture optimized for micro-object detection. By integrating the Residual Efficient Layer Aggregation Network (R-ELAN) backbone and Area Attention Module, the model effectively enhances feature representation and computational efficiency. Key innovations, such as Flash Attention and 7×7 separable convolutions, enable precise localization of small-scale objects while maintaining real-time performance. Experimental results on the xView satellite imagery dataset demonstrated satisfactory performance on three out of four object classes (boat, building, and plane), achieving a training mAP@0.5 of 0.41 and testing mAP@[0.5:0.95] of 0.235. However, the vehicle class posed challenges due to its extremely small size, underscoring the need for further refinement. Overall, YOLOv12s bridges the gap between accuracy and efficiency in micro-object detection, making it suitable for applications in domains like remote sensing, medical imaging, and industrial inspection.

VI. Future Works

Future research will focus on addressing the limitations observed in detecting extremely small objects, such as vehicles in this study. Potential directions include:

1) Hybrid Architectures - Exploring hybrid models combining convolutional neural networks (CNNs) with transformers to enhance feature representation for micro-scale objects.
2) Self-Supervised Learning - Incorporating self-supervised techniques to mitigate reliance on extensive labeled datasets, particularly for challenging classes like vehicles.
3) Domain Adaptation - Enhancing model generalization across different datasets by leveraging transfer learning and domain adaptation techniques to tackle diverse micro-object detection scenarios.

These advancements aim to further improve the robustness and scalability of YOLOv12s for real-world applications involving micro-object detection.

References

[1] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection, 2020.

[2] Christoph Borel-Donohue and S Susan Young. Image quality and super resolution effects on object recognition using deep neural networks. In *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications*, volume 11006, pages 596–604. SPIE, 2019.

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[4] Wei Han, Jia Chen, Lizhe Wang, Ruyi Feng, Fengpeng Li, Lin Wu, Tian Tian, and Jining Yan. Methods for small, weak object detection in optical high-resolution remote sensing images: A survey of advances and challenges. *IEEE Geoscience and Remote Sensing Magazine*, 9(4):8–34, 2021.

[5] Rudong Jing, Wei Zhang, Yanyan Liu, Wenlin Li, Yuming Li, and Changsong Liu. An effective method for small object detection in low-resolution images. *Engineering Applications of Artificial Intelligence*, 127:107206, 2024.

[6] Glenn Jocher, Alex Stoken, Ayush Chaurasia, Jirka Borovec, NanoCode012, Yonghye Kwon, and ChristopherSTAN. Yolov5 by ultralytics. https://github.com/ultralytics/yolov5, 2021.

[7] Asifullah Khan, Anabia Sohail, Umme Zahoora, and Aqsa Saeed Qureshi. A survey of the recent architectures of deep convolutional neural networks. *Artificial intelligence review*, 53:5455–5516, 2020.

[8] Aleksandra Kos, Dominik Belter, and Karol Majek. Deep learning for small and tiny object detection: A survey. *Pomiary Automatyka Robotyka*, 27(3), 2023.

[9] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[10] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 21–37. Springer, 2016.

[11] Benjamin Midtvedt, Jesús Pineda, Fredrik Skärberg, Erik Olsén, Harshith Bachimanchi, Emelie Wesén, Elin K Esbjörner, Erik Selander, Fredrik Höök, Daniel Midtvedt, et al. Single-shot self-supervised object detection in microscopy. *Nature communications*, 13(1):7492, 2022.

[12] Fatahlla Moreh, Yusuf Hasan, Bilal Zahid Hussain, Mohammad Ammar, and Sven Tomforde. Deep learning for micro-scale crack detection on imbalanced datasets using key point localization. *arXiv preprint arXiv:2411.10389*, 2024.

[13] Mohammed Gamal Ragab, Said Jadid Abdulkader, Amgad Muneer, Alawi Alqushaibi, Ebrahim Hamid Sumiea, Rizwan Qureshi, Safwan Mahmood Al-Selwi, and Hitham Alhussian. A comprehensive systematic review of yolo for medical object detection (2018 to 2023). *IEEE Access*, 2024.

[14] Rajeev Ranjan, Swami Sankaranarayanan, Ankan Bansal, Navaneeth Bodla, Jun-Cheng Chen, Vishal M Patel, Carlos D Castillo, and Rama Chellappa. Deep learning for understanding faces: Machines may be just as good, or better, than humans. *IEEE Signal Processing Magazine*, 35(1):66–83, 2018.

[15] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[16] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

[17] Yunjie Tian, Qixiang Ye, and David Doermann. Yolov12: Attention-centric real-time object detectors. *arXiv preprint arXiv:2502.12524*, 2025.

[18] Wei Wei, Yu Cheng, Jiafeng He, and Xiyue Zhu. A review of small object detection based on deep learning. *Neural Computing and Applications*, 36(12):6283–6303, 2024.

[19] Zeliang Wei, Xicheng Chen, Jialu Huang, Zhenyan Wang, Tianhua Yao, Chengcheng Gao, Haojia Wang, Pengpeng Li, Wei Ye, Yang Li, et al. Construction of a medical micro-object cascade network for automated segmentation of cerebral microbleeds in susceptibility weighted imaging. *Frontiers in Bioengineering and Biotechnology*, 10:937314, 2022.