Difference Explanation Between Our Work and Sun's Paper

Dear Editors and Reviewers,

We appreciate the opportunity to clarify the core differences between our submitted work "*AI-Native Multi-Scale Attention Fusion for Ubiquitous Aerial Sensing: Small Object Detection in UAV Imagery*" **(MAF-Net)** and the previously referenced paper:

*[R1] Sun, G.; Wang, S.; Xie, J. An Image Object Detection Model Based on Mixed Attention Mechanism Optimized YOLOv5. Electronics **2023**, 12, 1515. https://doi.org/10.3390/electronics12071515*

These two studies address distinct challenges in object detection with targeted technical designs and application scenarios, and the key differences are elaborated as follows. In the following, we refer to the core contributions proposed in [R1] as **MAO-YOLOv5**.

1. Research Background and Core Challenges

MAO-YOLOv5 focuses on general-purpose object detection scenarios, which can cover conventional fields such as traffic monitoring and public security surveillance. Its core challenge lies in improving detection performance under complex environmental interference, including insufficient light, partial occlusion, and background clutter. The model aims to handle multi-scale targets with moderate size differences, without specific constraints on target pixel occupancy.

In contrast, our proposed MAF-Net is dedicated to small object detection in high-altitude UAV visible light remote sensing imagery, a critical component of AI-native ubiquitous aerial sensing. The unique characteristics of UAV imagery pose unprecedented challenges: extremely low pixel occupancy (e.g., 97.96% of targets in the AI-TOD dataset are smaller than 32×32 pixels, and some targets have a width or height of less than 10 pixels), severe scale variation caused by altitude changes and viewpoint shifts, and high similarity between background and targets. Mathematically, these challenges manifest as an extremely low feature signal-to-noise ratio (SNR $\ll 1$), ambiguous spatial localization due to the sparse spatial support set of small targets, and insufficient channel feature decoupling. These issues are specific to aerial sensing and

not addressed by MAO-YOLOv5's general-purpose design.

2. Dataset Adaptation

MAO-YOLOv5 utilizes the PASCAL VOC2007+2012 dataset, a general-purpose object detection benchmark containing 20 categories of targets with balanced size distribution and conventional image dimensions. This dataset is suitable for verifying the performance of general-purpose models but lacks the characteristics of aerial imagery.

By contrast, our proposed MAF-Net employs three specialized aerial remote sensing datasets: AI-TOD, DOTA, and RSOD. AI-TOD consists of 14,018 images with an extreme small target distribution; DOTA includes 2,806 high-resolution images (800×800 to 4000×4000 pixels) covering 15 target categories with multi-scale variations; RSOD provides 976 images with a tri-modal size distribution (small, medium, and large targets). These datasets are carefully selected to match the practical application scenarios of UAV sensing, ensuring the model is evaluated under realistic aerial imaging conditions, which is fundamentally different from the general-purpose dataset used in MAO-YOLOv5.

3. Model Architecture and Technical Innovations

The two models adopt different baseline frameworks and technical routes based on their respective objectives. Specifically, MAO-YOLOv5 is built on YOLOv5, optimizing the feature extraction process by advancing feature map extraction to preserve spatial information of small targets and adjusting the convolution stride to achieve multi-scale feature fusion. It integrates the conventional CBAM mixed attention mechanism (sequential channel and spatial attention) to enhance feature discrimination.

In contrast, MAF-Net proposed in our work is developed on YOLOv7, leveraging its ELAN-W and SPPCSPC modules for deep feature extraction. To address the unique challenges of UAV small object detection, MAF-Net introduces three key innovations: first, a dedicated 160×160 small object detection layer with a receptive field of 4×4, which effectively captures extremely small targets that cannot be detected by traditional 20×20/40×40/80×80 feature maps; second, a Hybrid-Attention Encoder (HAE) that incorporates variance pooling to capture high local contrast of small targets (e.g., edges and textures), mitigating feature dilution caused by global pooling; third, an Attention-

Guided Decoder (AGD) that performs row/column 1D pooling to preserve x/y coordinate information, decoupling channel attention from positional information to better support bounding box regression. Additionally, MAF-Net adopts a density-adaptive anchor optimization strategy based on K-means++ clustering with an IoU-based distance metric, which adapts to the dynamic density-scale distribution of aerial targets, whereas MAO-YOLOv5 only adjusts the anchor scale range without considering density adaptation.

4. Loss Function and Performance Optimization

MAO-YOLOv5 uses binary cross-entropy (BCE) for classification loss and GIoU loss for regression loss, focusing on general loss optimization without targeted design for class imbalance or small target localization.

Our proposed MAF-Net employs Focal Loss to address the severe class imbalance in aerial imagery (small targets are easily misclassified as background), reducing the weight of easy samples and focusing on hard samples at the foreground-background boundary. For regression loss, MAF-Net combines GIoU loss with the coordinate-decoupled features from AGD, resulting in higher IoU between predicted and ground-truth boxes and more accurate localization.

In terms of performance, MAO-YOLOv5 achieves a mAP@0.5 of 92.30% on the general PASCAL VOC dataset with an FPS of 23.07.

Our proposed MAF-Net achieves mAP@0.5 gains of 14.1%, 11.28%, and 22.09% on the specialized AI-TOD, DOTA, and RSOD datasets, respectively. The obvious difference in mAP values stems from the distinct difficulty levels of the datasets, i.e., general datasets have larger targets and simpler backgrounds, while aerial datasets feature extremely small targets and complex clutter. MAF-Net maintains real-time inference efficiency suitable for edge deployment on UAVs, balancing accuracy and computational cost for aerial sensing applications.

5. Applications and Contributions

MAO-YOLOv5 provides a balanced solution for general-purpose object detection, improving accuracy and speed for multi-scale targets in conventional environments, and its application is limited to scenarios without extreme size constraints or specialized imaging conditions.

MAF-Net addresses the long-standing challenges of missed and false detections of small targets in UAV aerial imagery, supporting critical applications such as infrastructure inspection, precision agriculture, and public safety monitoring. Its innovations in shallow feature reinforcement, variance-augmented attention, and coordinate-decoupled feature modeling offer new technical insights for aerial sensing, filling the gap between general-purpose object detectors and specialized aerial detection requirements.

In summary, MAF-Net and MAO-YOLOv5 differ fundamentally in research objectives, application scenarios, technical designs, and dataset adaptation. MAF-Net is a targeted innovation for UAV-based small object detection, while MAO-YOLOv5 is a general-purpose optimization. These differences ensure the novelty and uniqueness of our submitted work, which contributes specifically to the field of ubiquitous aerial sensing.

We hope this explanation clarifies the distinctions between the two studies. Please feel free to contact us if further details are required.

Sincerely,
Ke Ma