A Lightweight Channel Bias for Small Object Detection in Aerial Images

Hongrui Fan*[a], Haoyuan Liu*[a], and Hiroshi Watanabe[a]

[a]Waseda University, Shillman Hall 401, 3-14-9 Okubo, Shinjuku-ku, Tokyo, Japan 169-0072

ABSTRACT

Small-object detection in drone-based imagery is difficult because small targets occupy very few pixels and often appear in cluttered backgrounds. Existing approaches often rely heavily on spatial cues, but their effectiveness is limited when the spatial information itself is unstable or insufficient. This paper introduces a simple *Channel Bias* (CB) mechanism that models relationships along the channel dimension to compensate for weak spatial cues. The method uses global average pooling and a lightweight $1 \times N$ convolution to generate a channel-wise bias, which is broadcast and combined with the original features to highlight useful channels. Experiments on the VisDrone benchmark show that integrating Channel Bias mechanism into YOLO-based detectors yields consistent accuracy gains. These results indicate that channel-level information is a practical way to improve small-object detection when spatial evidence is limited.

Keywords: Small Object Detection, Feature Modulation, Lightweight Model

1. INTRODUCTION

Small-object detection[1] is challenging because small targets occupy only a few spatial cells and often appear in cluttered scenes. As the available spatial information decreases, the features extracted from the spatial dimension become weak and unreliable. This observation naturally motivates us to look for a more stable source of cues. Feature channels in convolutional networks often encode semantic or scale-related information that does not directly depend on object size. This leads to a simple inductive bias: when spatial cues are insufficient, channel-wise relationships are assumed to provide more reliable evidence for refining features. Our method is based on this idea.

We introduce a lightweight Channel Bias mechanism that generates a channel-wise bias using global average pooling and a $1 \times N$ convolution. The bias is then broadcast and concatenated with the original features, providing an additional signal that helps the network emphasize informative channels without changing the spatial structure. The design is simple and intentionally lightweight, and in this work we focus mainly on its effect on accuracy. Integrating the mechanism into YOLO-based detectors yields consistent performance improvements on the VisDrone[2] dataset.

2. RELATED WORK

Detecting small objects is challenging because their limited pixel area provides weak spatial cues. To compensate for this, many studies enhance spatial information using global context or spatial priors. Several studies also combine multi-scale features to recover spatial detail, but their effectiveness is limited when the object size becomes extremely small. In our previous work,[3] we also examined attention-based approaches by adding CBAM to a point cloud classification network, and found that combining spatial and channel cues can help when feature quality is limited. This experience suggests that non-spatial information, such as channel-wise features, may remain useful when spatial evidence is weak. We briefly review non-local and spatial-bias mechanisms here, as they represent two common spatial strategies for handling weak spatial evidence.

Further author information: (Send correspondence to Hongrui Fan)
Hongrui Fan: E-mail: trace@ruri.waseda.jp, Laboratory website: https://www.ams.giti.waseda.ac.jp
Haoyuan Liu: E-mail: liuhaoyuan@akane.waseda.jp, Laboratory website: https://www.ams.giti.waseda.ac.jp
*These authors contributed equally.

2.1 Non-Local Neural Networks

Convolutional networks rely on local receptive fields, which makes it difficult for them to capture long-range dependencies. This limitation becomes important in small-object detection, where spatial features are weak and the available context is limited. To address this issue, a non-local operation[4] was proposed as a way to compute relationships between all spatial positions and obtain global information. It is defined as:

$$y_i = \frac{1}{C(x)} \sum_j f(x_i, x_j)\, g(x_j) \tag{1}$$

where y_i is the response at position i and $f(\cdot)$ measures the similarity between positions i and j. In practice, the feature map is projected into an embedding space, and the similarity between all spatial positions is computed by their dot product:

$$\mathbf{A} = [a_{ij}] \in \mathbb{R}^{hw \times hw} \tag{2}$$

which produces a dense attention map that allows information to be propagated from distant locations, helping the network obtain context that may not be visible from local features alone. Non-local modules have been applied to various detection frameworks to improve contextual reasoning. Our previous study[5] integrated non-local blocks into different levels of the Feature Pyramid Network[6] in YOLO, and showed that adding global context at selected scales can improve small-object detection. These results support the general idea that global dependencies are useful when spatial cues are weak.

2.2 Spatial Bias

Spatial-bias mechanisms[7] provide a lightweight way to introduce global spatial priors into convolutional features. The module reduces the input to a few channels, downsamples it to a low-resolution map, and applies a 1D convolution on the flattened spatial dimension to model global patterns with low cost. The result is reshaped and upsampled to form a set of spatial bias maps, which are later combined with the original feature map. A simplified formulation is:

$$\mathbf{SB} = f_{\mathrm{SB}}(\mathbf{X}), \qquad \mathbf{Y} = \mathrm{BN}\big(\mathrm{Conv}(\mathbf{X}) \oplus \mathbf{SB}\big) \tag{3}$$

where $\mathbf{X}$ is the input feature map, $\mathbf{SB}$ is the upsampled spatial-bias map derived from $\mathbf{X}$, $\oplus$ denotes channel-wise concatenation, and BN is batch normalization. This design adds global spatial information with low overhead and avoids heavy attention computation, making it suitable for real-time detectors such as YOLO.

3. PROPOSED METHOD

The proposed Channel Bias mechanism is designed to provide a lightweight way to incorporate channel-wise contextual cues into convolutional features. Unlike spatial-based methods that focus on enhancing location-dependent information, Channel Bias mechanism aims to compensate for the weak spatial evidence of small objects by introducing an additional bias derived purely from channel statistics. Fig. 1 illustrates the overall structure of the module.

3.1 Channel Bias Mechanism

Let $\mathbf{X} \in \mathbb{R}^{C \times H \times W}$ be the input feature map. The Channel Bias mechanism is motivated by a practical limitation of small-object detection: small objects occupy only a few pixels, causing their spatial features to become sparse, unstable, and easily overwhelmed by background activations. In such cases, relying on spatial cues is inherently unreliable because the spatial dimension simply does not contain enough informative evidence. This observation suggests a natural shift in perspective. Instead of strengthening a dimension that fundamentally lacks information, it is more effective to exploit the channel dimension, where global activation statistics remain relatively stable even when spatial details are lost. Based on this insight, the Channel Bias mechanism introduces a lightweight channel-derived bias to complement the weak spatial evidence and guide feature refinement in a stable manner. To implement this idea, the feature map is first summarized by global average pooling to obtain a compact descriptor representing the overall activation strength of each channel. This descriptor is then transformed by a 1×1 convolution to produce a reduced channel-bias vector of dimension k, enabling the

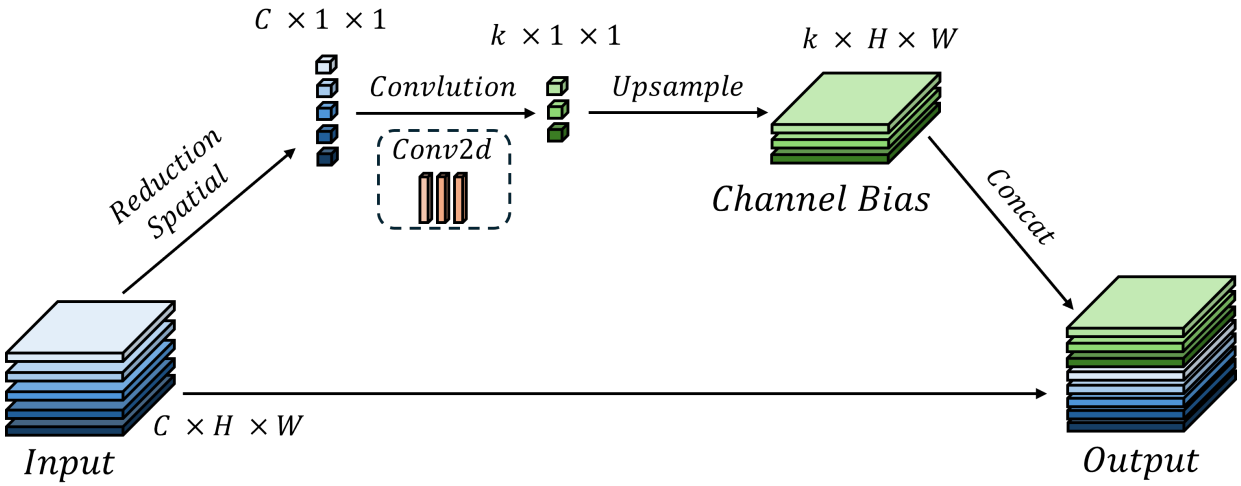

Figure 1. Structure of the proposed Channel Bias mechanism. A global channel descriptor is extracted by global average pooling, transformed by a 1×1 convolution into a compact bias vector, expanded spatially, and fused with the original feature map using channel concatenation followed by a 1×1 convolution.

model to capture simple inter-channel dependencies without introducing heavy computation. The resulting bias is spatially broadcast to match the resolution of the original feature map and concatenated with $\mathbf{X}$. A final 1×1 convolution restores the channel dimension and yields the refined output:

$$\mathbf{Y} = \mathrm{Conv}_{1\times1}([\mathbf{X}, \mathbf{CB}]) \tag{4}$$

which integrates the original convolutional response with the channel-derived bias. This fusion provides an additional cue when spatial evidence is weak, while keeping the overall computation minimal.

3.2 Computational Complexity

The Channel Bias mechanism consists of global average pooling, two 1×1 convolutions, and tensor broadcasting. Since none of these operations depend on the spatial resolution (H, W), the computational cost is dominated by the channel-reduction step:

$$\mathcal{O}_{CB} = \mathcal{O}(Ck) \tag{5}$$

where k is a small constant (e.g., k=3). The spatial expansion is implemented through broadcasting and incurs constant-time cost. As a result, the overall complexity is effectively independent of the feature-map resolution, making the Channel Bias mechanism suitable for lightweight and real-time detection frameworks.

4. EXPERIMENT

We evaluate the proposed Channel Bias mechanism on the VisDrone detection benchmark, which contains many small and densely distributed objects. All experiments are performed on a single NVIDIA RTX 2080 Ti GPU, and the training settings follow the default configurations of the original YOLOv11[8] and YOLOv12[9] implementations. To ensure a fair comparison, the Channel Bias mechanism is inserted at the same stage of the backbone for both detectors, and no other modifications are introduced to the network or training schedule.

Table 1 summarizes the results. For both YOLOv11 and YOLOv12, adding the Channel Bias mechanism leads to consistent improvements across the standard evaluation metrics. In YOLOv11, the overall gain is modest but stable, showing a small increase in $mAP_{50\text{-}95}$. This suggests that even a lightweight channel-derived bias can provide additional cues when spatial evidence is limited.

The improvement is more noticeable in YOLOv12, where both AP_{75} and $mAP_{50\text{-}95}$ increase. This indicates that the Channel Bias mechanism is able to enhance the feature representation without requiring architectural

changes or additional computation. Since the mechanism operates independently of the spatial resolution, it integrates smoothly into both detectors and provides benefits under the same training conditions.

Overall, the results show that the Channel Bias mechanism works as a simple and effective plugin for improving accuracy on small-object detection tasks. Its low computational cost and ease of integration make it practical for real-time detection frameworks and resource-constrained environments.

Table 1. Performance Evaluation of YOLO Architectures with Channel Bias on VisDrone

Method	AP_{50} ↑	AP_{75} ↑	$mAP_{50\text{-}95}$ ↑
YOLOv11	36.0	21.7	21.4
YOLOv11 + Channel Bias	36.5	21.7	**21.5**
YOLOv12	35.7	21.2	21.0
YOLOv12 + Channel Bias	36.4	22.0	**21.6**

5. CONCLUSIONS

This paper presented a lightweight Channel Bias mechanism designed to improve small-object detection by introducing a simple channel-derived bias into convolutional features. The method focuses on channel-wise contextual cues, which remain stable even when spatial evidence is limited, and can be integrated into existing detectors without architectural modifications. Experiments on the VisDrone dataset show consistent accuracy improvements on both YOLOv11 and YOLOv12, while keeping the computational overhead negligible. These results suggest that channel-level information provides a practical complement to spatial features for small-object detection. As future work, it would be valuable to evaluate the Channel Bias mechanism on additional aerial-image datasets such as UAVDT[10] to further verify its generality across different scenarios.

REFERENCES

[1] Nikouei, M., Baroutian, B., Nabavi, S., Taraghi, F., Aghaei, A., Sajedi, A., and Moghaddam, M. E., "Small object detection: A comprehensive survey on challenges, techniques and real-world applications," *arXiv preprint arXiv:2503.20516* (2025).

[2] Zhu, P., Wen, L., Du, D., Bian, X., Fan, H., Hu, Q., and Ling, H., "Detection and tracking meet drones challenge," *IEEE Transactions on Pattern Analysis and Machine Intelligence* , 1–1 (2021).

[3] Fan, H. and Watanabe, Y., "A study on point cloud classification using residual mlp and attention mechanism," in [*Proceedings of the IEICE General Conference*], D–12B–13 (Mar. 2024). (in Japanese).

[4] Wang, X., Girshick, R., Gupta, A., and He, K., "Non-local neural networks," *in Proceedings of the IEEE conference on computer vision and pattern recognition* , 7794–7803 (June 2018).

[5] Fan, H., Liu, H., and Watanabe, H., "Enhancing object detection with non-local modules in the feature pyramid network," in [*2024 IEEE 13th Global Conference on Consumer Electronics (GCCE)*], 870–873, IEEE (2024).

[6] Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., and Belongie, S., "Feature pyramid networks for object detection," *in Proceedings of the IEEE conference on computer vision and pattern recognition* , 2117–2125 (Jul. 2017).

[7] Go, J. and Ryu, J., "Spatial bias for attention-free non-local neural networks," *Expert Systems with Applications* **238**, 122053 (2024).

[8] Khanam, R. and Hussain, M., "Yolov11: An overview of the key architectural enhancements," *arXiv preprint arXiv:2410.17725* (2024).

[9] Tian, Y., Ye, Q., and Doermann, D., "Yolov12: Attention-centric real-time object detectors," *arXiv preprint arXiv:2502.12524* (2025).

[10] Du, D., Qi, Y., Yu, H., Yang, Y., Duan, K., Li, G., Zhang, W., Huang, Q., and Tian, Q., "The unmanned aerial vehicle benchmark: Object detection and tracking," in [*Proceedings of the European conference on computer vision (ECCV)*], 370–386 (2018).