Adaptive Sparse Convolutional Networks with Global Context Enhancement for Faster Object Detection on Drone Images
Supplementary Material

Bowei Du[1,2†], **Yecheng Huang**[1,2†], **Jiaxin Chen**[2], **Di Huang**[1,2,3*]
[1] State Key Laboratory of Software Development Environment, Beihang University, Beijing, China
[2] School of Computer Science and Engineering, Beihang University, Beijing, China
[3] Hangzhou Innovation Institute, Beihang University, Hangzhou, China
{boweidu, ychuang, jiaxinchen, dhuang}@buaa.edu.cn

In this document, we provide more implementation details, additional ablation studies as well as more visualization results.

A. More Implementation Details

In Table 1 of the main body, we evaluate the performance of our approach with various base detectors, including RetinaNet, FSAF, Faster-RCNN besides GFL V1. Since both RetinaNet and FSAF are one-stage detectors as GFL V1, we adopt the same setting as used by GFL V1. Regarding the two-stage Faster-RCNN detector, we follow [4], modifying its RPN head to 4 Convolution-GN-ReLU layers instead of 1 convolution layer and using 256 feature channels, which proves effective in balancing the accuracy and efficiency [4].

In Table 8 of the main body, we compare our network to MobileNet V2 [2], ShuffleNet V2 [1] and QueryDet [4]. We select the feature maps generated from the layers (2, 4, 6) and (0, 1, 2) as the input of FPN for MobileNet V2 and ShuffleNet V2, respectively. Regarding QueryDet, we re-implement it by using the same setting for fair comparison. Particularly, we utilize the unified input size as 1,333×800 and omit the calculation of the P_2 layer.

B. Additional Ablation Studies

We show additional results by our approach using different residual structures, accelerating strategies, context clues and training epochs.

B.1. On Residual Structures

As shown in Fig. 2 of the main body, we adopt a residual structure to compensate the loss of contextual information due to sparse convolutions. We therefore compare the performance of the proposed CEASC network by using different residual structures, including: 1) "w.o. Res." without using the residual structure; 2) "Focal Res." using the raw input for skip connection, *i.e.* $\mathbf{F} := \mathbf{F} + \mathbf{X}$; and 3) "Ours" using the global contextual feature for skip connection, *i.e.* $\mathbf{F} := \mathbf{F} + \mathbf{G}$.

Method	mAP	AP_{50}	AP_{75}	GFLOPs	FPS
w.o. Res.	28.4	50.5	28.1	150.58	21.32
Focal Res.	28.1	49.8	27.4	153.52	19.98
Ours	**28.7**	**50.7**	**28.4**	**150.18**	**21.55**

Table A. Comparison in terms of mAP (%) and GFLOPs/FPS with different residual structures on VisDrone.

Method	mAP	AP_{50}	AP_{75}	GFLOPs	FPS
Baseline	28.4	50.0	27.8	524.95	13.46
FPN on P3+P4 only	28.1	49.9	27.5	494.42	15.63
DWS	24.5	43.2	23.9	157.74	20.84
Ours	**28.7**	**50.7**	**28.4**	**150.18**	**21.55**

Table B. Comparison in terms of mAP (%) and GFLOPs/FPS with different acceleration strategies on VisDrone.

Method	mAP	AP_{50}	AP_{75}	GFLOPs	FPS
Baseline	28.4	50.0	27.8	524.95	13.46
Interpolation [3]	26.9	48.5	26.1	212.15	13.80
Ours	**28.7**	**50.7**	**28.4**	**150.18**	**21.55**

Table C. Comparison in terms of mAP (%) and GFLOPs/FPS with different context clues on VisDrone.

As displayed in Table A, our approach reaches the best performance, highlighting its advantage in capturing global context.

† indicates equal contribution.
* refers to the corresponding author.

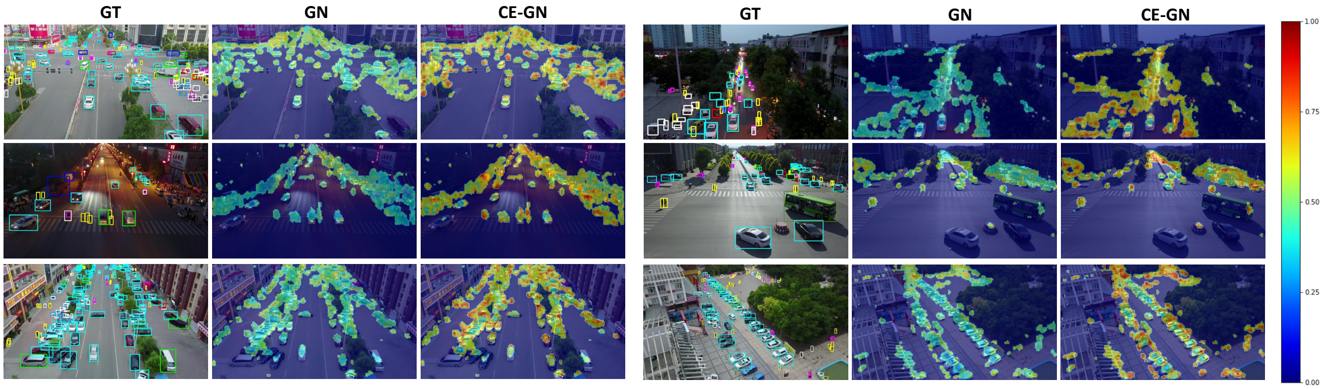

Figure A. Visualization of correlation between features generated by dense convolutions and sparse convolutions with distinct normalization schemes on VisDrone.

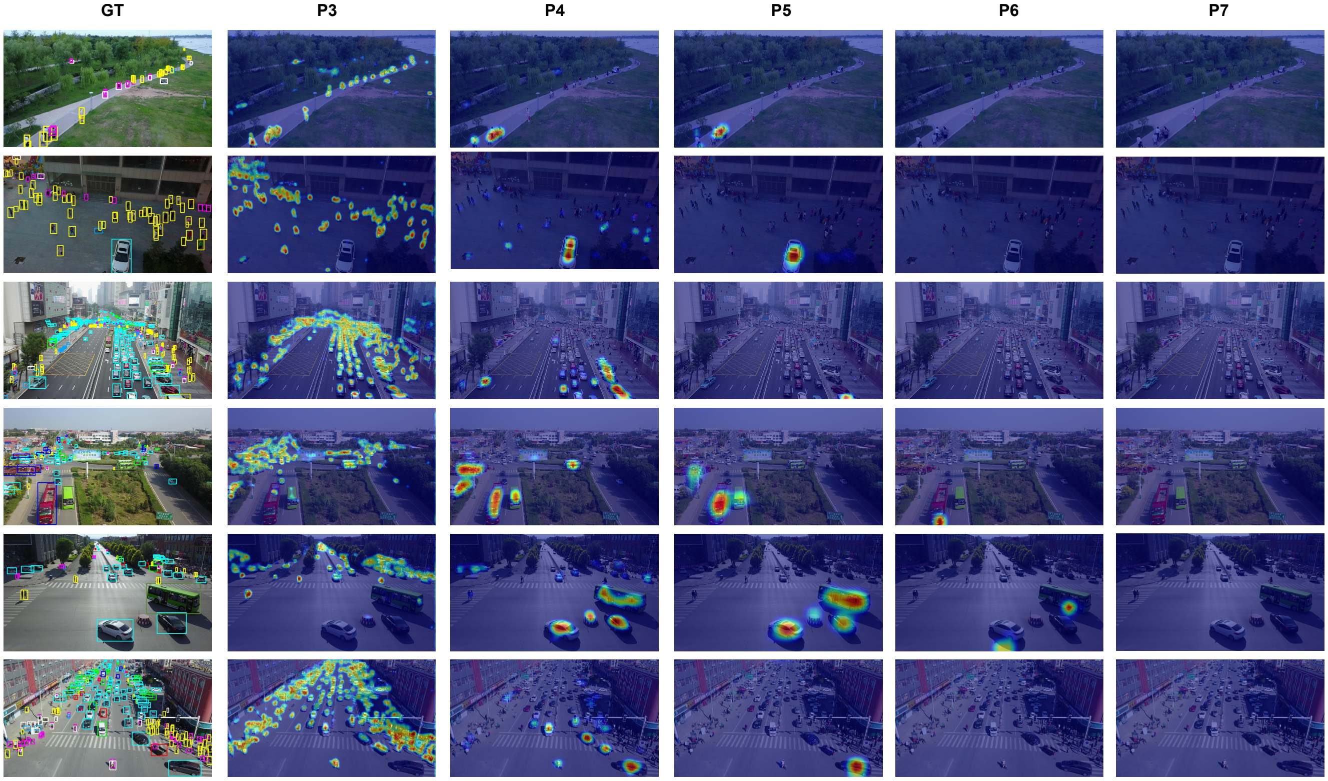

Figure B. Visualization of dynamic masks estimated by AMM at different layers (from 'P3' to 'P7') in FPN of GFL V1 on VisDrone. Highlighted areas are activated for computation.

B.2. On Acceleration Strategies

In Table 8 of the main body, we compare our approach with the state-of-the-art lightweight models for drone-based object detection. Here, we carry out the ablation study on our approach using distinct acceleration strategies including: 1) "FPN on P3+P4 only" that only adopts the P3 and P4 layers for FPN, since the P5-P7 layers are unlikely to be activated for sparse convolutions as observed in Fig. B; and 2) "DWS" that utilizes the Depth-Wise Separable (DWS) convolution as in MobileNet, instead of the normal 3 × 3 convolution in the "Baseline", *i.e.* the original GFL V1 detector.

The results are summarized in Table B, indicating that our approach outperforms the counterparts both in accuracy and efficiency, due to the sparse convolutions optimized by context-enhancement and adaptive multi-layer masking.

Epoch	Method	mAP	AP_{50}	AP_{75}	AR_1	AR_{10}	AR_{100}	AR_{500}	GFLOPs	FPS
12	Baseline	27.8	49.2	27.3	0.63	6.27	34.7	44.2	524.95	13.48
	Ours (CEASC)	27.8	**49.3**	**27.4**	**0.67**	**6.47**	**34.8**	**44.4**	**151.93**	**21.52**
15	Baseline	28.4	50.0	27.8	0.62	6.36	35.6	44.9	524.95	13.46
	Ours (CEASC)	**28.7**	**50.7**	**28.4**	**0.65**	**6.56**	35.6	**45.0**	**150.18**	**21.55**
24	Baseline	28.9	50.9	28.4	**0.72**	6.53	35.7	45.2	524.95	13.41
	Ours (CEASC)	**29.1**	**51.3**	**28.7**	0.70	**6.90**	**36.0**	**45.4**	**151.42**	**21.49**

Table D. Comparison in terms of AP/AR (%) and GFLOPs/FPS with the GFL V1 base detector using different training epochs on VisDrone.

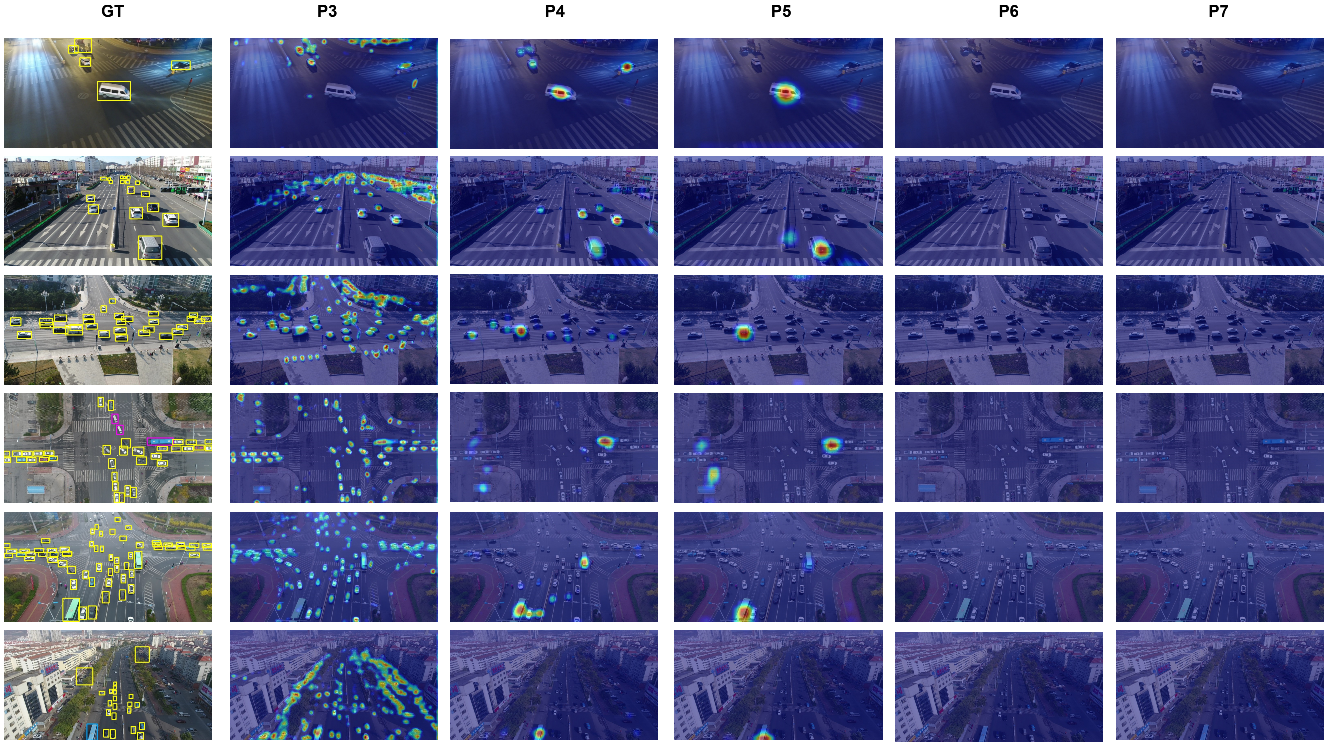

Figure C. Visualization of dynamic masks estimated by AMM at different layers (from 'P3' to 'P7') in FPN of GFL V1 on UAVDT. Highlighted areas are activated for computation.

B.3. On Context Clues

In Sec. 3.1.2 of the main body, we mention an interpolation method to generate ignored pixels from focal areas [3], and an ablation study is conducted for comparison. The results in Table C reveal that interpolation incurs a drop on the accuracy but consumes more computations.

B.4. On Training Epochs

In the literature, some studies train their models for varying epochs (*e.g.* 12 or 24). We thus provide more results by using such numbers of training epochs in addition to 15 adopted in this work. As displayed in Table D, our approach consistently boosts the performance by a large margin with different training epochs. When more training epochs are used, our approach reaches a higher accuracy, where 15 is a good trade-off.

B.5. More Visualized Results

More results are visualized in Fig. A as supplements to Fig. 3 of the main body. The features normalized by CE-GN have higher correlation with dense convolutions than GN, indicating that CE-GN enhances focal features with the assistance of global context.

In Fig. B and Fig. C, we visualize more results as supplements to Fig. 4 of the main body. As illustrated, the mask generated by our approach well covers foreground areas, indicating that sparse convolutions spend most computations on foreground, thus promoting the efficiency without sacrificing much precision.

References

[1] Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *ECCV*, 2018. 1

[2] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *CVPR*, 2018. 1

[3] Zhenda Xie, Zheng Zhang, Xizhou Zhu, Gao Huang, and Stephen Lin. Spatially adaptive inference with stochastic feature sampling and interpolation. In *ECCV*, 2020. 1, 3

[4] Chenhongyi Yang, Zehao Huang, and Naiyan Wang. Querydet: Cascaded sparse query for accelerating high-resolution small object detection. In *CVPR*, 2022. 1