UAV-Net: A Fast Aerial Vehicle Detector for Mobile Platforms Supplementary Material

Tobias Ringwald[†] Lars Sommer[‡*] Arne Schumann[*] Jürgen Beyerer[*‡] Rainer Stiefelhagen[†]

[†]CV:HCI
Karlsruhe Institute of Technology
Karlsruhe, Germany

[‡]Vision and Fusion Lab
Karlsruhe Institute of Technology
Karlsruhe, Germany

[*]Fraunhofer IOSB
Fraunhoferstraße 1
Karlsruhe, Germany

{firstname.lastname}@{kit.edu, iosb.fraunhofer.de}

1. Non-Maximum Suppression

Benchmarks in the main paper were conducted without timing the non-maximum suppression (NMS) stage in order to better judge the architectural changes of the networks. As NMS is crucial for a single stage detector like SSD, we also provide benchmarks with NMS for selected networks in Table 1 (DLR 3K) and 2 (VEDAI). A confidence threshold of 50% was used for the benchmarks.

2. Batch Size

As the aerial images had to be cut down into smaller tiles, it is often helpful to process them in batches. Table 3 shows VGG-SSD with 2 box sizes and UAV-Net$_{\varphi=0.50}$ in comparison. For UAV-Net$_{\varphi=0.50}$, a higher batch size can be used due to its low memory footprint. For VGG, batch sizes of 8 or higher could not be used due to memory exhaustion.

3. Qualitative Results

Due to space restrictions, the main paper only showed qualitative results for DLR 3K. We therefore also provide visualizations for VEDAI-1024 (see Figure 2), and UAVDT (see Figure 3). For DLR 3K, we provide further visualizations of crowded scenes in Figure 1 to emphasize the robustness of UAV-Net.

4. Modified Prediction Layers for UAVDT

For DLR 3K and VEDAI, 1×1 convolutions could replace the normal 3×3 convolution kernels as the receptive field is already noticeable larger than the size of present vehicles. For UAVDT, this assumption did not hold as the dataset contains a lot of different camera angles and ground sampling distances (GSD). Both DLR 3K and VEDAI were recorded with an approximately uniform GSD in bird's-eye view. Therefore, vehicles appeared with roughly the same size in all images. In the UAVDT dataset, some sequences were recorded in diagonal view and from different altitudes. Hence, objects can appear in different views which requires multiple default box sizes. We therefore conduct an ablation study to find the ideal filter and default box sizes for UAVDT. As base network we choose UAV-Net$_{\varphi=1.00}$ pretrained on DLR 3K (see main paper) and only adjust the filter size and number of default boxes in the prediction layers. Results are reported in Table 4. Note that the R-FCN reference from [1] was benchmarked on a different machine. The initial experiments for the filter size use the single box size and aspect ratios from the DLR 3K and VEDAI experiments. For additional default box sizes, we again use the clustering approach from [2] with an increasing amount of clusters (see Table 5). Default box aspect ratios were also clustered but set to the fixed set of two ratios $\{1.3, 2.2\}$ due to the lack of variation. Increasing the filter size also increases AP up to 5×5 kernels. Adding additional box sizes (denoted by c) also increases the AP. For $c = 5$, UAV-Net$_{\varphi=1.00}$ even surpasses the best reference model provided in [1]. For UAV-Net$_{\varphi=0.50}$ we also provide a model with 3×3 filters and $c = 4$, which provides a good trade-off between AP and inference speed.

5. UAV-Net Structure

Table 6 shows the network structure of UAV-Net for different φ values.

References

[1] D. Du, Y. Qi, H. Yu, Y. Yang, K. Duan, G. Li, W. Zhang, Q. Huang, and Q. Tian. The unmanned aerial vehicle benchmark: object detection and tracking. In *ECCV*, 2018. 1, 2

[2] C.-Y. Fu, W. Liu, A. Ranga, A. Tyagi, and A. C. Berg. DSSD: Deconvolutional single shot detector. *arXiv preprint arXiv:1701.06659*, 2017. 1

Network	AP (%)	Inference Speed (FPS)		
		Titan X	GTX 1060	Jetson TX2
VGG^{NMS}, 2 box sizes	**97.3**	24.2	9.9	1.2
UAV-Net$^{NMS}_{\varphi=1.00}$	97.2	179.2	82.1	15.0
UAV-Net$^{NMS}_{\varphi=0.50}$	97.1	237.6	112.4	20.9
UAV-Net$^{NMS}_{\varphi=0.15}$	91.3	**367.3**	**173.0**	**34.7**

Table 1: Influence of the NMS stage for selected networks on the DLR 3K dataset.

Network	AP (%)	Inference Speed (FPS)		
		Titan X	GTX 1060	Jetson TX2
VGG^{NMS}	**96.4**	16.5	5.6	0.7
UAV-Net$^{NMS}_{\varphi=1.00}$	95.7	120.6	47.3	9.6
UAV-Net$^{NMS}_{\varphi=0.50}$	95.2	163.1	69.2	13.5
UAV-Net$^{NMS}_{\varphi=0.15}$	93.5	**243.6**	**118.0**	**21.8**

Table 2: Influence of the NMS stage for selected networks on the VEDAI dataset.

Batch Size	VGG^{NMS}	UAV-Net$^{NMS}_{\varphi=0.50}$
1	0.7	13.5
2	0.7	14.6
4	0.7	15.1
8	–	15.3
16	–	**15.5**
32	–	**15.5**

Table 3: Jetson TX2 inference speed with different batch sizes for VGG (2 box sizes) and UAV-Net$_{\varphi=0.50}$ (on VEDAI, both including NMS).

Network	AP (%)	Inference Speed (FPS)		
		Titan X	GTX 1060	Jetson TX2
R-FCN [1]	34.35	4.7	–	–
UAV-Net$^{1\times1,c=1}_{\varphi=1.00}$	26.21	**214.0**	**98.8**	**18.3**
UAV-Net$^{3\times3,c=1}_{\varphi=1.00}$	27.52	207.4	89.1	17.2
UAV-Net$^{5\times5,c=1}_{\varphi=1.00}$	28.28	169.9	68.9	14.2
UAV-Net$^{7\times7,c=1}_{\varphi=1.00}$	27.91	117.8	45.9	8.5
UAV-Net$^{9\times9,c=1}_{\varphi=1.00}$	27.78	90.2	32.6	6.3
UAV-Net$^{11\times11,c=1}_{\varphi=1.00}$	26.47	69.7	26.2	4.8
UAV-Net$^{5\times5,c=2}_{\varphi=1.00}$	29.83	133.5	54.7	10.9
UAV-Net$^{5\times5,c=3}_{\varphi=1.00}$	32.63	113.4	47.4	9.4
UAV-Net$^{5\times5,c=4}_{\varphi=1.00}$	33.70	91.2	38.4	7.2
UAV-Net$^{5\times5,c=5}_{\varphi=1.00}$	**34.52**	80.1	34.7	6.6
UAV-Net$^{1\times1,c=5}_{\varphi=1.00}$	30.08	100.4	46.5	8.7
UAV-Net$^{3\times3,c=4}_{\varphi=1.00}$	32.76	112.2	51.5	9.0
UAV-Net$^{3\times3,c=5}_{\varphi=1.00}$	33.48	96.4	42.4	7.8
UAV-Net$^{3\times3,c=4}_{\varphi=0.50}$	31.82	132.5	69.2	11.4

Table 4: Effect of different filter sizes and default box sizes in the prediction layers of UAV-Net for UAVDT. c is the number of clusters from Table 5.

Clusters	Default box sizes (in px)
1	$\{31.4\}$ (from DLR 3K)
2	$\{30.6, 82.9\}$
3	$\{26.2, 48.4, 130.3\}$
4	$\{21.3, 38.4, 80.5, 256.1\}$
5	$\{19.2, 34.9, 58.5, 112.8, 291.6\}$

Table 5: Default box sizes for different cluster sizes.

Layer	Filter Size	φ	Filter Count
conv1	3×3	1.00	64
		0.50	40
		0.15	8
fire2/s3×3	3×3	1.00	16
		0.50	12
		0.15	4
fire2/e3×3	3×3	1.00	64
		0.50	24
		0.15	4
fire2/e1×1	1×1	1.00	64
		0.50	24
		0.15	8
fire3/s1×1	1×1	1.00	16
		0.50	16
		0.15	12
fire3/e3×3	3×3	1.00	64
		0.50	56
		0.15	20
fire3/e1×1	1×1	1.00	64
		0.50	24
		0.15	4
fire4/s3×3	3×3	1.00	32
		0.50	32
		0.15	24
fire4/e3×3	3×3	1.00	128
		0.50	116
		0.15	36
fire4/e1×1	1×1	1.00	128
		0.50	4
		0.15	4
fire5/s1×1	1×1	1.00	32
		0.50	28
		0.15	8
fire5/e3×3	3×3	1.00	128
		0.50	84
		0.15	4
fire5/e1×1	1×1	1.00	128
		0.50	4
		0.15	4
mbox_loc	1×1	*	20
mbox_conf	1×1	*	10

Table 6: UAV-Net structure for the DLR 3K and VEDAI experiments.

Figure 1: Qualitative results for UAV-Net$_{\varphi=0.50}$ (left column) and UAV-Net$_{\varphi=0.15}$ (right column) on DLR 3K. A confidence threshold of 50% was used to generate the detections (green boxes = true positives, yellow boxes = false negatives, red boxes = false positives). Note that the shown images are not full tiles from the dataset but instead cropped for visualization purposes.

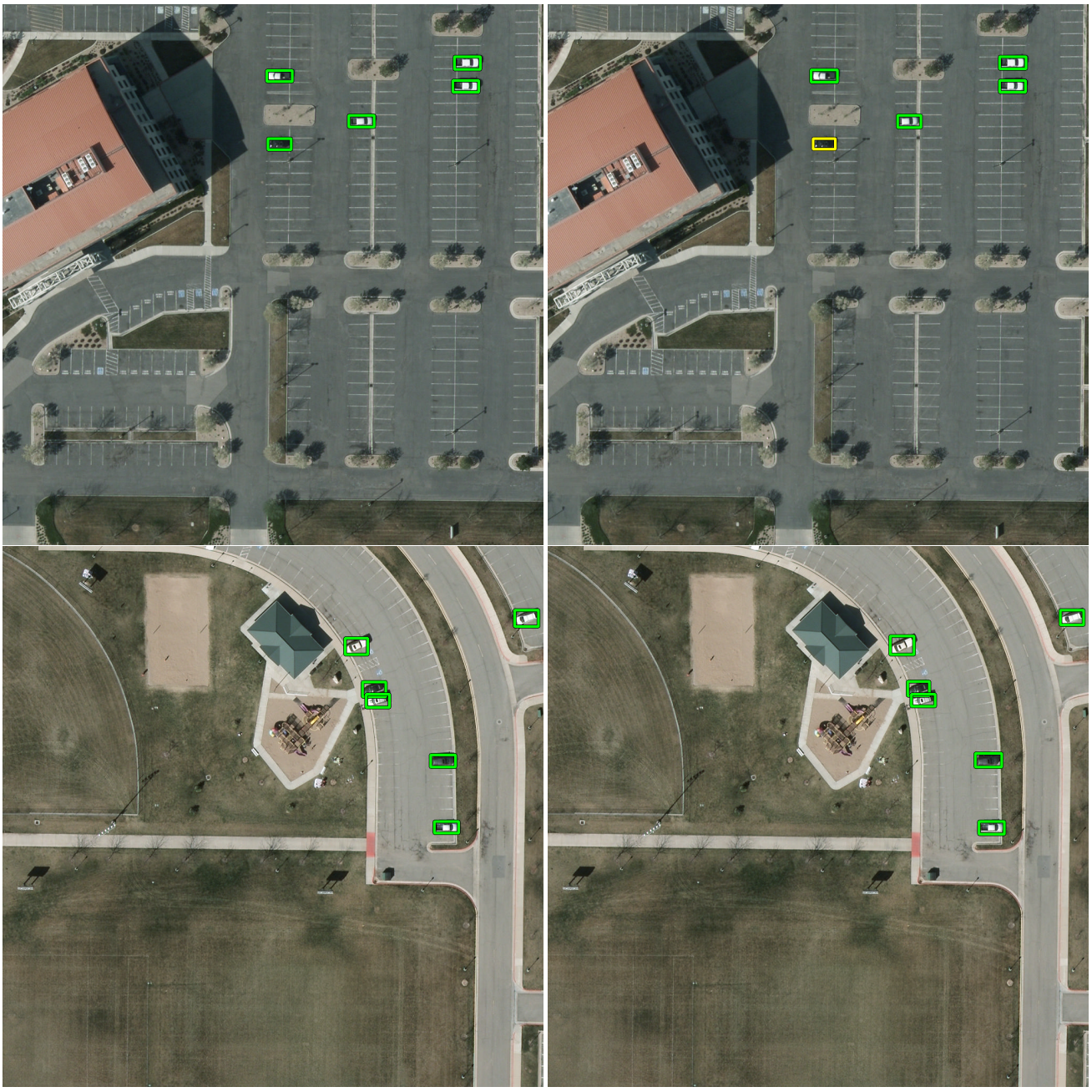

Figure 2: Qualitative results for UAV-Net$_{\varphi=0.50}$ (left column) and UAV-Net$_{\varphi=0.15}$ (right column) on VEDAI-1024. A confidence threshold of 50% was used to generate the detections (green boxes = true positives, yellow boxes = false negatives, red boxes = false positives).

Figure 3: Qualitative results for UAV-Net$^{5\times5,c=5}_{\varphi=1.00}$ (left column) and UAV-Net$^{3\times3,c=4}_{\varphi=0.50}$ (right column) on UAVDT. A confidence threshold of 50% was used to generate the detections (white boxes). Best viewed in the digital version.