Towards RAW Object Detection in Diverse Conditions

Supplementary Material

1. Statistics of AODRaw

More examples from AODRaw. We collect images under 9 conditions, as shown in Tab. 2 of the main paper. Tab. 1 shows **a specific example for each condition** for a better understanding. Furthermore, Fig. 1 shows **more examples of our AODRaw dataset** and the annotated bounding boxes, showing the diversity of the AODRaw.

Visualization of RAW images. To better show the domain gap between RAW and sRGB images, we show some RAW images and the corresponding sRGB images in Fig. 2. Here, the RAW images are visualized by demosaicing and normalizing them to the range of [0, 255], without any other preprocessing operations.

Annotated categories. Tab. 2 lists the annotated categories in AODRaw and the corresponding number of instances. We sort the categories according to the number of instances.

Categories per image. Fig. 3 shows the distribution of the number of categories in the images. For each condition, the maximum number of categories exceeds 10. More than half of the conditions have a maximum number that exceeds 15. Especially, the maximum number for outdoor daylight conditions is 19.

Instances per image. Fig. 4 shows the distribution of the number of instances in images. In all conditions, there are complex images containing hundreds of instances. The distribution of a single condition is close to the overall distribution, especially when the number of instances is less than 100. For cases exceeding 100, since there are fewer images in this range, there is some deviation between several conditions and the whole, *e.g.*, the condition of low-light and fog in outdoor scenes, as shown in Fig. 4h.

Bounding box size. Fig. 5 shows the distribution of the bounding box size for each condition, where small objects account for the majority in all conditions. Meanwhile, compared to the overall distribution of all conditions, the images in indoor scenes contain fewer small objects and more large objects, as shown in Fig. 5a and Fig. 5b. In contrast, the outdoor scenes have a similar distribution to the overall distribution with a large number of small objects.

2. Experiments Settings

Most hyperparameters follow the COCO [2] datasets in the mmdetection [1]. In data augmentations, the images are resized between 800 and 1024 along the shorter side, while the longer side is no larger than 2048. And the RandomFlip is used to augment images. For detectors using ConvNeXt [4], we set the layer-wise learning rate decay as 0.75 and the stochastic depth (DropPath) ratio as 0.4. For detectors using Swin-T [3], the stochastic depth ratio is 0.2. For RAW pre-training, we follow the official codes released by ConvNeXt [4] and Swin [3].

3. Distillation Implementation

Method. Besides the supervised classification loss function, we use logit-based and feature-based distillation for cross-domain distillation. For logit-based distillation, we denote z_s and z_t as the output of student and teacher, respectively. z_s and z_t have been normalized by the SoftMax function. Then, the logit-based loss is calculated as follows:

$$L_l = \mathrm{KLDivLoss}(y_s, y_t) = y_t \log \frac{y_t}{y_s}. \tag{1}$$

For feature-based distillation, we denote z_s and z_t as the global feature output by the student and teacher, respectively. For ConvNeXt, the features are acquired through 1) applying global average pooling to the output of the last block and 2) processing the feature using the last LayerNorm layer. The loss is calculated as follows:

$$L_f = \frac{1}{C} \sum_{i=0}^{C-1} |z_s^i - z_t^i|, \tag{2}$$

where C means the number of dimensions.

References

[1] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019. 1

[2] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 1

[3] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021. 1

[4] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *CVPR*, 2022. 1

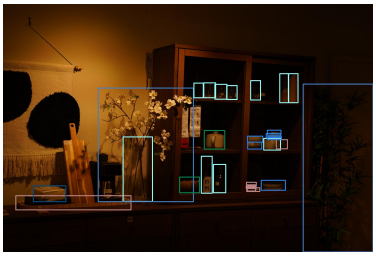
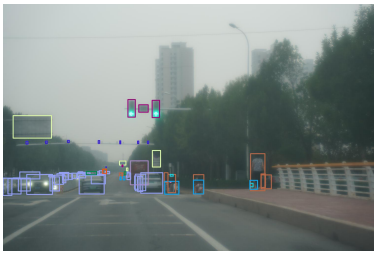
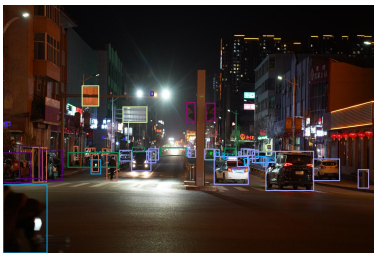
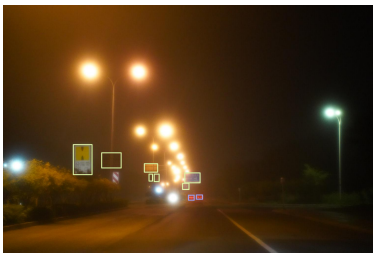
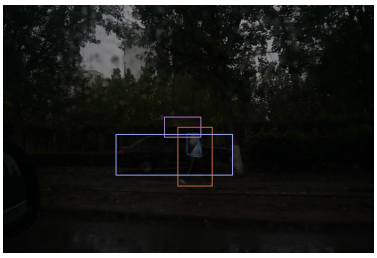
Indoor and Daylight	Indoor and Low-light	Outdoor, Daylight, and Clear Weather
		
Outdoor, Daylight, and Fog Weather	**Outdoor, Daylight, and Rain Weather**	**Outdoor, Daylight, and Rain+Fog Weather**
		
Outdoor, Low-light, and Clear Weather	**Outdoor, Low-light, and Fog Weather**	**Outdoor, Low-light, and Rain Weather**
		

Table 1. The example of each condition in AODRaw.

Name	Instance	Name	Instance	Name	Instance	Name	Instance
car	42798	person	16864	traffic sign	13458	surveillance camera	9344
traffic light	8884	motorcycle	7389	truck	4271	chair	3548
bottle/cup	3168	bicycle	2134	garbage can	1993	traffic cone	1929
table	1480	tricycle	1268	helmet	1211	umbrella	1142
pillow	1055	lamp	1046	handbag/satchel	1001	bus	973
potted plant	910	hat	893	backpack	818	phone	685
plate	663	vase	639	monitor	586	bus stop sign	528
desk lamp	431	tent	406	sofa	373	clock	333
bowl	322	pen	228	crane	196	wine glass	191
bench	191	keyboard	175	mirror	170	bed	158
mouse	155	pot	151	earphone	151	fire hydrant	145
toilet paper	138	spoon	105	laptop	104	sink	96
watch	93	fire extinguisher	90	suitcase	85	train	72
dog	64	computer box	58	refrigerator	43	cans	42
vending machine	34	airplane	32	boat	29	cat	23
toilet	20	scissors	19				

Table 2. The annotated categories and the number of instances per category.

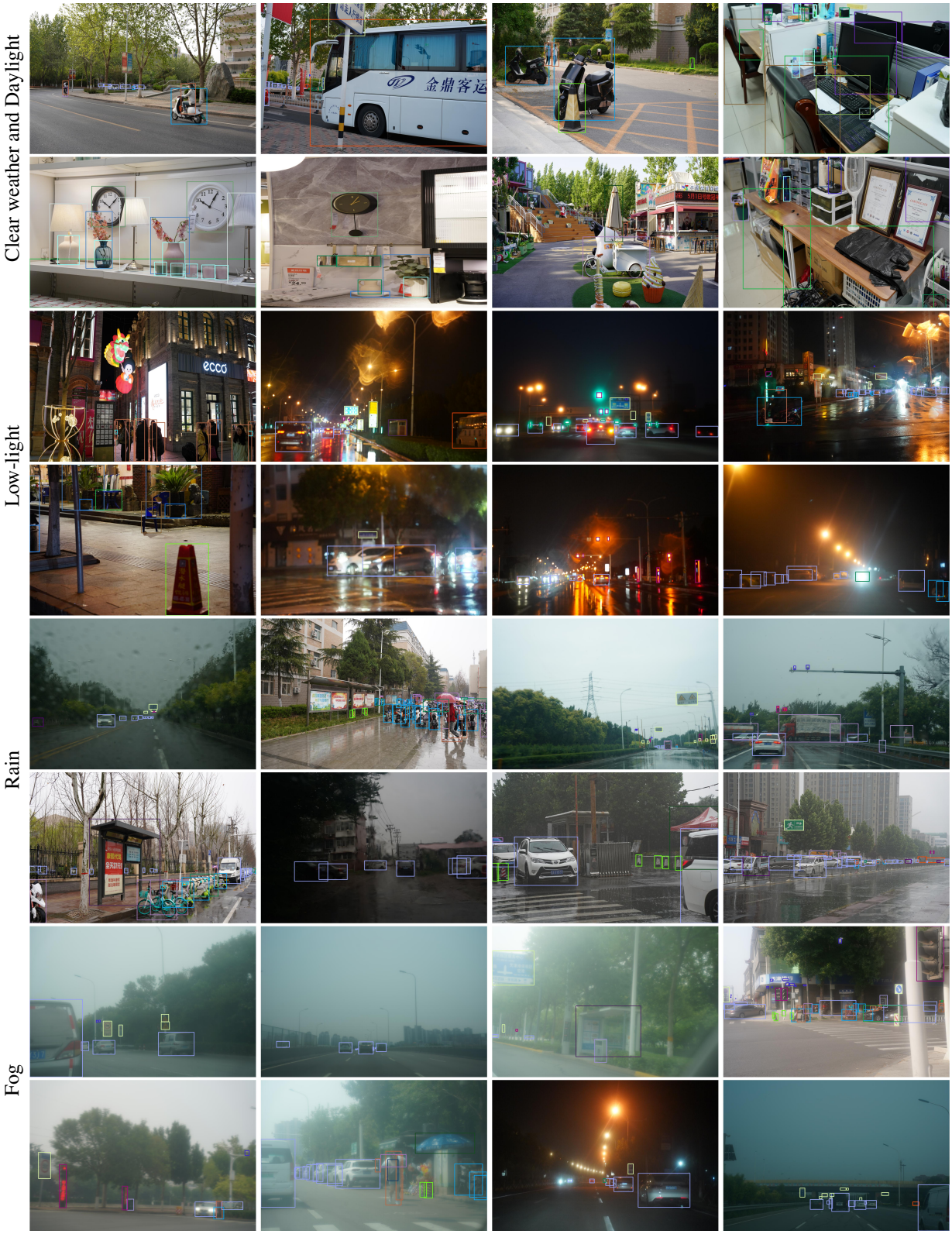
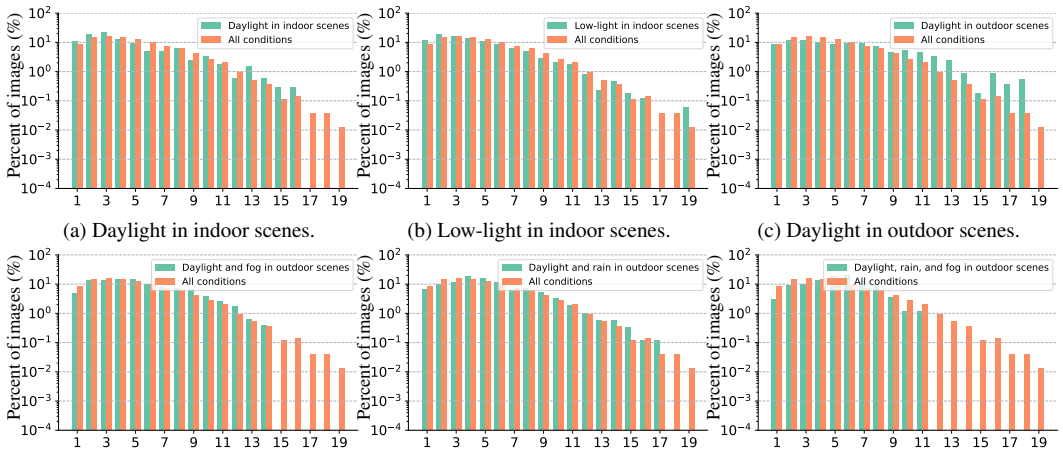


Figure 1. Example of the images in the AODRaw.

Figure 2. Example of the RAW images in the AODRaw.

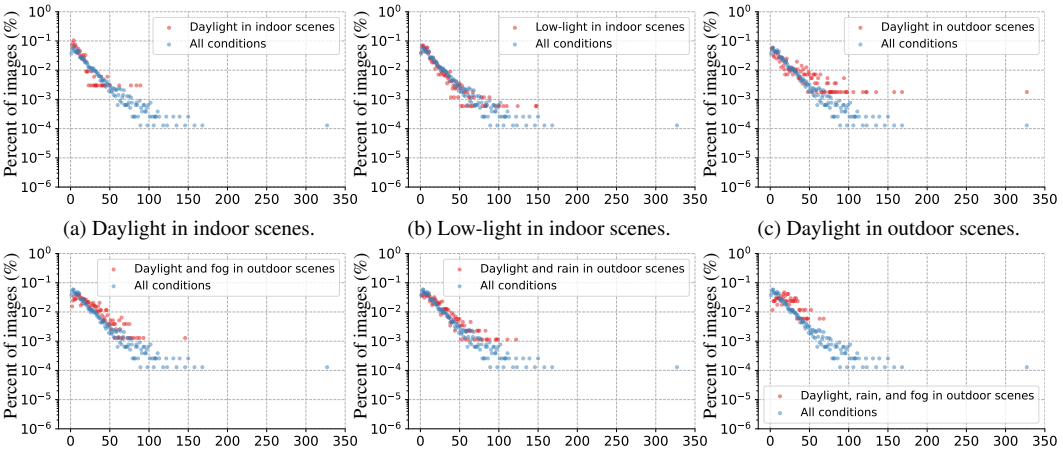


(a) Daylight in indoor scenes. (b) Low-light in indoor scenes. (c) Daylight in outdoor scenes.

(d) Daylight and fog in outdoor scenes. (e) Daylight and rain in outdoor scenes. (f) Daylight, rain, and fog in outdoor scenes.

(g) Low-light in outdoor scenes. (h) Low-light and fog in outdoor scenes. (i) Low-light and rain in outdoor scenes.

Figure 3. Distribution of the number of categories in images of each condition. The horizontal axis represents the number of categories.

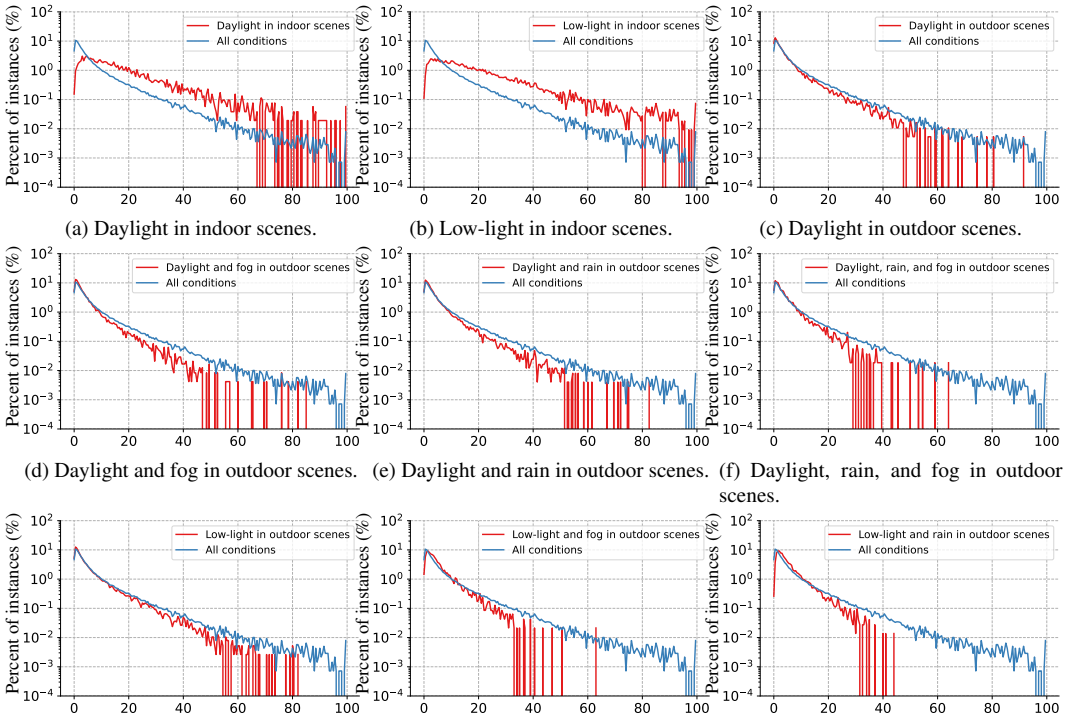


(a) Daylight in indoor scenes. (b) Low-light in indoor scenes. (c) Daylight in outdoor scenes.

(d) Daylight and fog in outdoor scenes. (e) Daylight and rain in outdoor scenes. (f) Daylight, rain, and fog in outdoor scenes.

(g) Low-light in outdoor scenes. (h) Low-light and fog in outdoor scenes. (i) Low-light and rain in outdoor scenes.

Figure 4. Distribution of the number of instances in images of each condition. The horizontal axis represents the number of instances.



(a) Daylight in indoor scenes. (b) Low-light in indoor scenes. (c) Daylight in outdoor scenes.

(d) Daylight and fog in outdoor scenes. (e) Daylight and rain in outdoor scenes. (f) Daylight, rain, and fog in outdoor scenes.

(g) Low-light in outdoor scenes. (h) Low-light and fog in outdoor scenes. (i) Low-light and rain in outdoor scenes.

Figure 5. Relative bounding box size $\sqrt{\frac{\text{box area}}{\text{image area}}}$ of each condition. The horizontal axis represents the relative bounding box size.