RAW-Diffusion: RGB-Guided Diffusion Models for High-Fidelity RAW Image Generation (Supplementary Material)

Christoph Reinders[1†] Radu Berdan[2*] Beril Besbinar[2*] Junji Otsuka[3] Daisuke Iso[2]
[1]Leibniz University Hannover [2]Sony AI [3]Sony Group Corporation

In the supplementary material, the RAW-Diffusion hyper-parameters and the training details for object detection with Faster R-CNN and YOLOv8 are described. Furthermore, we provide additional quantitative and qualitative results for the RGB2RAW reconstruction and downstream object detection experiments.

1. Hyper-parameters

The hyper-parameters for the training, architecture, and diffusion process of RAW-Diffusion are detailed in Tab. 11. The number of base features is denoted by N_{Features} and the number of groups used in the group normalization is denoted by $N_{\text{Norm,Groups}}$. Further implementation details can be found in the published code: https://github.com/SonyResearch/RAW-Diffusion.

2. Object Detection Training Details

The experiments are performed using a Faster R-CNN [10] and YOLOv8 [5]. Faster R-CNN has a ResNet-50 backbone pretrained on ImageNet, and it is trained with RGB and RAW images normalized using the corresponding mean and standard deviation of the dataset. We apply random flip, random resize, and cropping as data augmentation, use an image size of 416×640, and train the network for 48 epochs. In particular, for YOLOv8, the images are normalized to $[0, 1]$, we use the same size of 416×640, and the model is finetuned from a COCO [7] pretrained checkpoint for 10 epochs using random flip augmentation. In addition, the RAW images are reduced to three channels by averaging the two green channels.

For the generation of the RAW datasets (Cityscapes-RAW and BDD100K-RAW) from large-scale RGB datasets, SRISP and RAW-Diffusion are trained on the images of the object detection datasets with a limited number of samples. When combining the original NOD images and the generated datasets, p_{gen} is set to 0.95.

† Work done during an internship at Sony AI. Corresponding author: reinders@tnt.uni-hannover.de
* These authors contributed equally to this work

3. Qualitative RGB2RAW Results

Additional qualitative RAW reconstruction results are shown in Fig. 6. The comparison includes all methods, i.e., U-Net [2], UPI [1], InvGrayscale [11], InvISP [12], InvISP$^+$ [12], ISPLess [3], ISPLess$^+$ [3], CycleR2R [6], RISPNet [4], Diffusion [8], SRISP [9], and RAW-Diffusion.

4. Detailed Object Detection Results

In Tab. 12 and Tab. 13, extended object detection results are presented using Faster R-CNN and YOLOv8, respectively. Performance metrics, including AP, AP_{50}, AP_{75}, AP_{S}, AP_{M}, and AP_{L}, are detailed. The experiments evaluate the adaptation performance on the target domain by training on a limited subset of 100 sample from the original NOD, denoted by RGB and RAW, respectively, and the combination with the generated datasets by SRISP and RAW-Diffusion, i.e., Cityscapes-RAW (SRISP), BDD100K-RAW (SRISP), Cityscapes-RAW (ours), and BDD100K-RAW (ours).

Additionally, the zero-shot performance is shown in Tab. 14 and Tab. 15 using Faster R-CNN and YOLOv8, respectively. The models are trained exclusively on the generated RAW datasets and evaluated on the test set of NOD.

5. Qualitative Object Detection Results

In Fig. 7, qualitative results from different object detection models using Faster R-CNN are presented. The models are trained on RGB images, RAW images, or a combination with the generated datasets. The experiment shows the advantages of RAW images compared to RGB images, especially in low-light scenarios. Furthermore, the integration of the generated RAW datasets, i.e., Cityscapes-RAW and BDD100K-RAW generated by SRISP and RAW-Diffusion, improves the precision.

6. Training on RGB datasets for object detection on RAW images

We analyze if object detection performance on RAW images improves by directly adding RGB samples to the original RAW training set instead of adding converted RAW images. The results of this experiment on NOD Nikon with Cityscapes and BDD100K, respectively, are shown in Tab. 16. Adding RGB training samples does not improve object detection performance as much as our generated RAW dataset. This highlights that the quality and distribution of the training dataset are crucial.

Table 11. Hyper-parameters of RAW-Diffusion.

	Hyper-parameter	Value
Training	Training Steps	70k
	Optimizer	AdamW, β=[0.9, 0.999]
	Weight Decay	0.0
	Learning Rate	0.0001, linearly decreasing to zero
	Batch Size	4
Architecture	$N_{\text{ResBlocks}}$	2
	N_{Features}	32
	Feature Expansion	$(1, 1, 2, 2, 4, 4)$
	$N_{\text{Norm,Groups}}$	8
	Attention Block Resolutions	$16 \times 16, 8 \times 8$
	$N_{\text{GM,ResBlocks}}$	4
	$C_{\text{GM,Features}}$	64
Diffusion	Schedule	Linear, $\beta_1 = 0.0001$ to $\beta_T = 0.02$
	Steps	1000

Table 12. Object detection results using Faster R-CNN that is trained using 100 NOD training samples (RGB and RAW). Additionally, Cityscapes-RAW and BDD100K-RAW generated by SRISP and RAW-Diffusion are integrated. The best result is shown in bold, and the second best underlined.

	NOD Nikon					
Training Dataset	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
RGB	19.1±0.2	36.8±0.4	17.7±0.6	2.0±0.1	16.8±0.4	44.6±0.3
RAW	18.2±0.2	35.3±0.2	17.1±0.2	1.6±0.0	15.9±0.2	43.1±0.3
RAW + Cityscapes-RAW (SRISP)	23.0±0.6	43.9±0.4	21.9±1.5	2.5±0.1	21.4±0.6	49.8±1.4
RAW + BDD100K-RAW (SRISP)	24.2±0.3	45.7±0.4	22.6±0.7	2.9±0.3	21.7±0.3	51.8±1.2
RAW + Cityscapes-RAW (ours)	<u>24.7</u>±0.3	<u>46.3</u>±0.6	<u>23.9</u>±0.2	<u>3.7</u>±0.3	<u>22.8</u>±0.6	<u>52.1</u>±0.9
RAW + BDD100K-RAW (ours)	**26.5**±0.3	**49.3**±0.5	**25.3**±0.5	**4.4**±0.2	**23.9**±0.4	**54.6**±0.6
	NOD Sony					
Training Dataset	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
RGB	19.2±0.5	38.2±0.8	17.6±0.5	1.1±0.2	18.4±0.9	38.2±0.6
RAW	18.0±0.1	35.8±0.3	16.3±0.4	1.1±0.3	16.6±0.1	36.8±0.3
RAW + Cityscapes-RAW (SRISP)	23.1±0.4	46.9±0.4	20.0±0.1	2.0±0.2	21.3±1.0	44.4±0.1
RAW + BDD100K-RAW (SRISP)	26.0±0.2	50.5±0.6	24.2±0.5	3.1±0.5	24.4±0.3	46.8±0.6
RAW + Cityscapes-RAW (ours)	<u>26.2</u>±0.3	<u>50.6</u>±0.6	<u>24.9</u>±0.4	<u>3.4</u>±0.5	<u>25.4</u>±0.6	<u>46.9</u>±0.7
RAW + BDD100K-RAW (ours)	**28.6**±0.1	**55.2**±0.4	**26.4**±0.3	**4.2**±0.1	**27.7**±0.4	**49.6**±0.3

Table 13. Object detection results using YOLOv8 evaluating the performance on 100 NOD training samples (RGB and RAW) and the integration of Cityscapes-RAW and BDD100K-RAW generated by SRISP and RAW-Diffusion.

	NOD Nikon					
Training Dataset	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
RGB	22.7±1.7	32.9±10	22.5±1.8	1.7±0.3	19.6±2.1	55.9±4.1
RAW	25.6±0.2	45.1±0.1	25.6±0.3	2.3±0.1	21.6±0.3	62.4±0.1
RAW + Cityscapes-RAW (SRISP)	29.1±0.1	48.6±0.2	29.4±0.1	3.1±0.1	25.3±0.1	66.0±0.4
RAW + BDD100K-RAW (SRISP)	<u>32.0</u>±0.3	<u>53.1</u>±0.5	<u>32.1</u>±0.4	<u>4.2</u>±0.2	<u>28.6</u>±0.2	<u>68.7</u>±0.5
RAW + Cityscapes-RAW (ours)	29.9±0.3	50.2±0.5	30.7±0.7	3.3±0.1	26.6±0.3	66.4±0.3
RAW + BDD100K-RAW (ours)	**32.6**±0.1	**54.3**±0.2	**32.6**±0.7	**4.9**±0.1	**29.0**±0.3	**69.2**±0.2
	NOD Sony					
Training Dataset	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
RGB	18.4±0.3	33.8±0.2	18.4±0.4	1.9±0.0	15.5±2.0	47.3±8.1
RAW	27.6±0.4	49.3±0.2	27.0±0.9	1.6±0.2	23.0±0.1	57.9±0.5
RAW + Cityscapes-RAW (SRISP)	29.5±0.4	51.7±0.6	29.1±0.7	2.9±0.3	25.2±0.2	58.0±1.6
RAW + BDD100K-RAW (SRISP)	<u>32.0</u>±0.7	<u>55.3</u>±0.8	<u>31.6</u>±0.8	<u>3.2</u>±0.0	<u>29.5</u>±0.5	60.0±1.2
RAW + Cityscapes-RAW (ours)	31.0±0.3	54.8±0.5	31.0±0.5	<u>3.2</u>±0.1	27.0±0.3	<u>60.6</u>±0.4
RAW + BDD100K-RAW (ours)	**33.6**±0.1	**58.3**±0.2	**33.8**±0.0	**4.2**±0.0	**31.0**±0.1	**61.9**±0.1

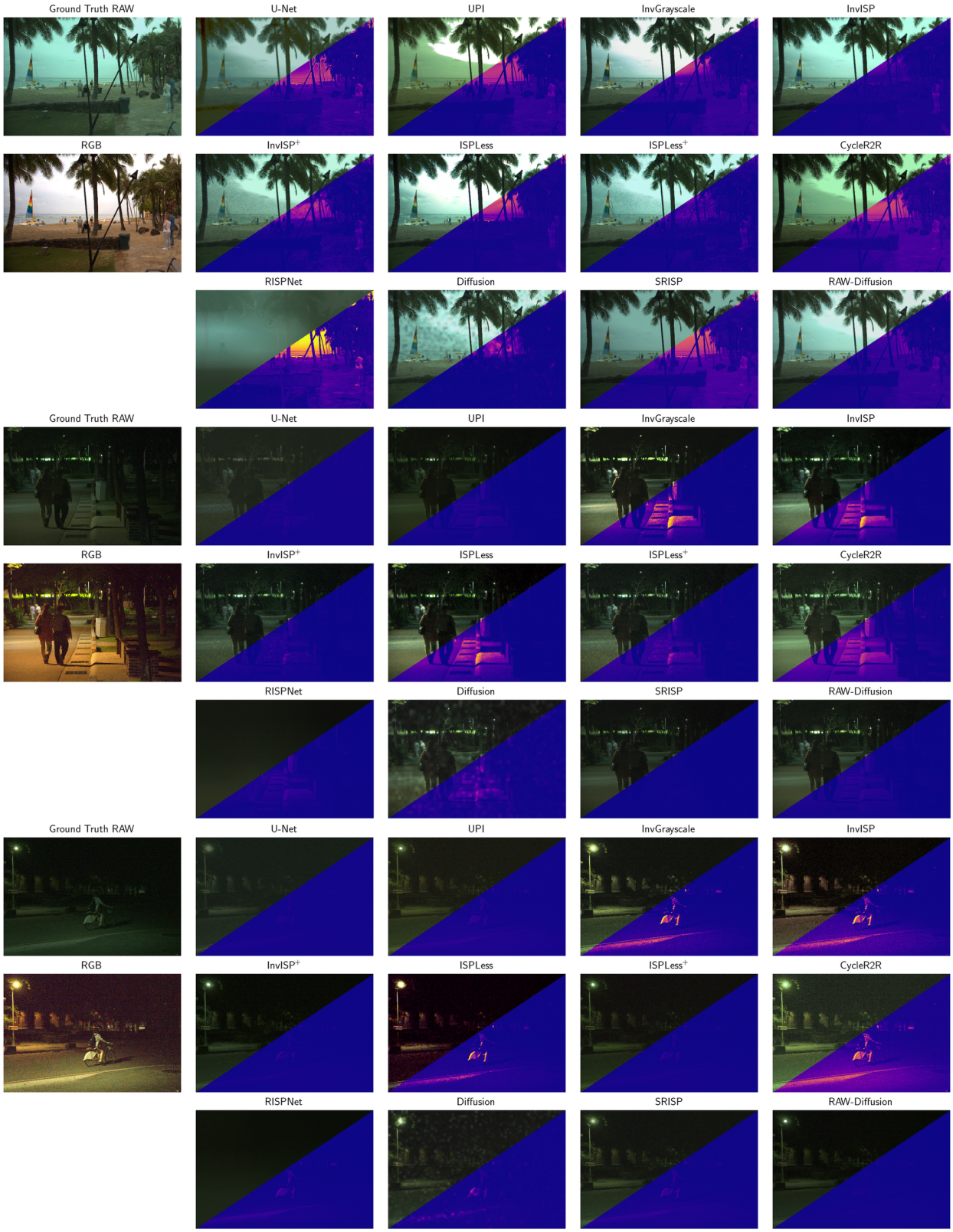

Figure 6. Qualitative results on FiveK Canon (top), NOD Nikon (middle), and NOD Sony (bottom). The reconstructed RAW image and the error map are presented for each method. The RAW images are shown with a gamma correction of $1/2.2$ for visualization.

Table 14. Zero-shot object detection results using Faster R-CNN. The models are trained exclusively on the generated datasets and evaluated on the NOD test set.

Training Dataset	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
	NOD Nikon					
Cityscapes-RAW (SRISP)	7.7±1.0	16.8±2.7	6.4±0.6	0.4±0.2	9.5±0.7	23.8±1.4
BDD100K-RAW (SRISP)	18.4±0.3	35.9±0.3	15.8±0.5	2.5±0.2	16.7±0.6	38.1±0.2
Cityscapes-RAW (ours)	12.0±0.7	23.4±1.6	11.4±0.5	1.5±0.4	11.2±0.7	26.8±1.7
BDD100K-RAW (ours)	**22.0**±0.1	**43.1**±0.3	**18.8**±0.5	**3.5**±0.2	**20.8**±0.2	**43.4**±0.7
	NOD Sony					
Cityscapes-RAW (SRISP)	3.8±1.1	9.3±2.8	2.6±0.5	0.2±0.1	4.7±0.6	14.1±1.7
BDD100K-RAW (SRISP)	16.9±0.6	34.2±0.6	14.5±0.8	.8±0.3	15.6±0.6	30.2±0.9
Cityscapes-RAW (ours)	13.7±1.3	29.4±2.3	11.7±0.8	1.9±0.1	13.5±1.5	26.5±2.5
BDD100K-RAW (ours)	**21.6**±0.1	**43.7**±0.4	**18.7**±0.2	**3.3**±0.2	**20.9**±0.2	**37.2**±0.4

Table 15. Zero-shot object detection results using YOLOv8. The models are trained exclusively on the generated datasets and evaluated on the NOD test set.

Training Dataset	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
	NOD Nikon					
Cityscapes-RAW (SRISP)	19.1±0.8	33.5±1.2	19.0±0.8	1.7±0.2	17.2±0.4	48.9±2.0
BDD100K-RAW (SRISP)	25.2±1.2	43.9±2.6	24.3±1.1	3.1±0.2	23.9±1.3	53.8±2.8
Cityscapes-RAW (ours)	25.8±0.7	44.1±0.8	25.7±1.0	2.7±0.2	23.1±0.4	58.8±1.9
BDD100K-RAW (ours)	**28.8**±0.6	**49.8**±0.9	**27.6**±0.9	**4.3**±0.4	**26.8**±0.1	**59.9**±2.1
	NOD Sony					
Cityscapes-RAW (SRISP)	17.4±1.0	34.0±1.3	16.2±1.6	1.8±0.1	15.8±0.6	39.7±2.3
BDD100K-RAW (SRISP)	24.0±1.4	43.6±2.7	23.1±1.3	2.2±0.2	22.8±1.6	45.5±2.2
Cityscapes-RAW (ours)	26.1±0.5	48.8±0.1	25.0±1.2	2.6±0.1	23.4±0.3	53.6±1.7
BDD100K-RAW (ours)	**29.2**±0.6	**53.6**±1.0	**28.1**±1.1	**4.0**±0.3	**27.3**±0.4	**54.4**±1.9

Table 16. Analysis of integrating the original RGB dataset and our generated RAW dataset. The Average Precision (AP) is shown for each experiment.

Training Dataset	Faster R-CNN	YOLOv8
RAW	18.2±0.2	25.8±0.5
RAW + Cityscapes-RGB	23.0±0.2	26.1±0.3
RAW + BDD100K-RGB	24.5±0.3	27.3±0.3
RAW + Cityscapes-RAW (ours)	24.7±0.3	29.9±0.3
RAW + BDD100K-RAW (ours)	**26.5**±0.3	**32.6**±0.1

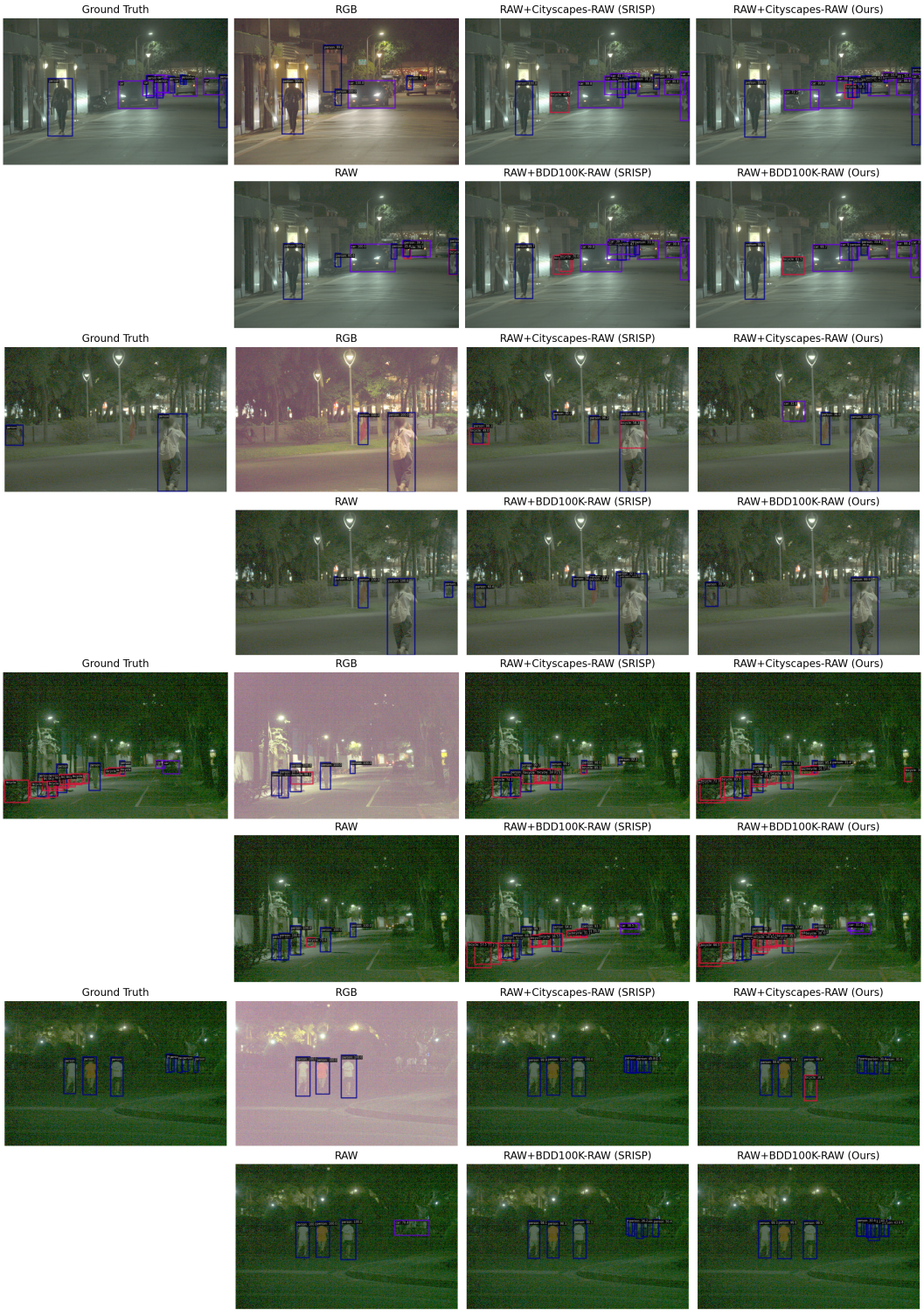

Figure 7. Qualitative object detection results from various models on the test set of NOD Nikon (first and second row) and Sony (third and forth row).

References

[1] Tim Brooks, Ben Mildenhall, Tianfan Xue, Jiawen Chen, Dillon Sharlet, and Jonathan T. Barron. Unprocessing Images for Learned Raw Denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11036–11045, 2019.

[2] Marcos V Conde, Radu Timofte, Yibin Huang, Jingyang Peng, Chang Chen, Cheng Li, Eduardo Pérez-Pellitero, Fenglong Song, Furui Bai, Shuai Liu, and others. Reversed image signal processing and RAW reconstruction. AIM 2022 challenge report. In *Proceedings of the European Conference on Computer Vision Workshops (ECCVW)*, pages 3–26. Springer, 2022.

[3] Gourav Datta, Zeyu Liu, Zihan Yin, Linyu Sun, Akhilesh R. Jaiswal, and Peter A. Beerel. Enabling ISPless Low-Power Computer Vision. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2429–2438. IEEE, 2023.

[4] Xiaoyi Dong, Yu Zhu, Chenghua Li, Peisong Wang, and Jian Cheng. RISPNet: A network for Reversed image signal processing. In *Computer Vision – ECCV 2022 Workshops*, pages 445–457, 2023.

[5] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics YOLOv8. 2023.

[6] Zhihao Li, Ming Lu, Xu Zhang, Xin Feng, M. Salman Asif, and Zhan Ma. Efficient visual computing with camera RAW snapshots. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–18, 2024.

[7] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*, 2014.

[8] Xinman Liu, Xuanchi Ren, and Ziyi Wu. Inverting Image Signal Processing Pipeline with Diffusion Models. Technical report, University of Toronto, 2022.

[9] Junji Otsuka, Masakazu Yoshimura, and Takeshi Ohashi. Self-Supervised Reversed Image Signal Processing via Reference-Guided Dynamic Parameter Selection. *arXiv preprint arXiv:2303.13916*, 2023.

[10] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

[11] Menghan Xia, Xueting Liu, and Tien-Tsin Wong. Invertible grayscale. *ACM Transactions on Graphics (SIGGRAPH Asia 2018 issue)*, 37(6):246:1–246:10, Nov. 2018.

[12] Yazhou Xing, Zian Qian, and Qifeng Chen. Invertible Image Signal Processing. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6283–6292, 2021.