DynamicISP: Dynamically Controlled Image Signal Processor for Image Recognition (Supplementary Material)

Masakazu Yoshimura Junji Otsuka Atsushi Irie Takeshi Ohashi
Sony Group Corporation
{masakazu.yoshimura, junji.otsuka, atsushi.irie, takeshi.a.ohashi}@sony.com

1. Details of the Implemented ISP

We implement the following ISP functions in a differentiable manner.

Auto Gain (AG)

Usual auto gains simply multiply some value, but we formulate as,

$$I_{AG}(X) = \begin{cases} \frac{1-p_h}{1-p_w} X \\ (if\, x < p_x(1-p_w),\, x \in X) \\ \frac{1-p_h}{1-p_w} X + \frac{p_h - p_w}{1-p_w} \\ (if\, p_x(1-p_w) + p_w < x, x \in X) \\ \frac{p_h}{p_w}(X - p_x(1-p_w)) + p_x(1-p_h) \\ (otherwise) \end{cases}, \tag{1}$$

where p_w, p_h, and p_x are parameters to be controlled, and x is the each pixel in the input image X. It intends to control how much and what range of domain should be emphasized with the third equation. The first and the second avoid overflow from $[0, 1]$ without clipping as shown in Fig. 1.

Denoiser (DN)

We utilize a simple Bilateral filter (BF) [13] as,

$$I_{DN}(X) = (1 - p_a) \cdot X + p_a \cdot BF(p_{\sigma s}, p_{\sigma i}; X), \tag{2}$$

where $p_{\sigma s}$ and $p_{\sigma i}$ are the parameters for the spatial and intensity variance and p_a is another parameter. We set the kernel size as five.

Sharpener (SN)

Simple Gaussian filter (GF) is used as,

$$I_{SN}(X) = (1 - p_a) \cdot X + p_a \cdot (X - GF(p_\sigma; X)). \tag{3}$$

The second term is the difference-of-Gaussians [10] whose kernel sizes are one and five;

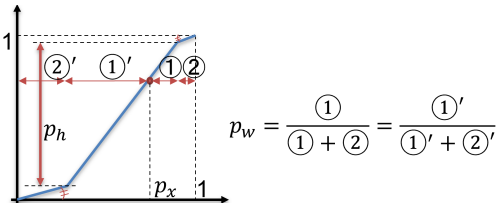


Figure 1. Auto gain function without clipping.

$$\begin{aligned} DoG(X) &= GF(p_{\sigma 1}, k_1; X) - GF(p_\sigma, k_2; X) \\ &= X - GF(p_\sigma, 5; X). \end{aligned} \tag{4}$$

Gamma (GM)

We follow the parameterization of gamma tone mapping in [11, 15] and implement it as differentiable;

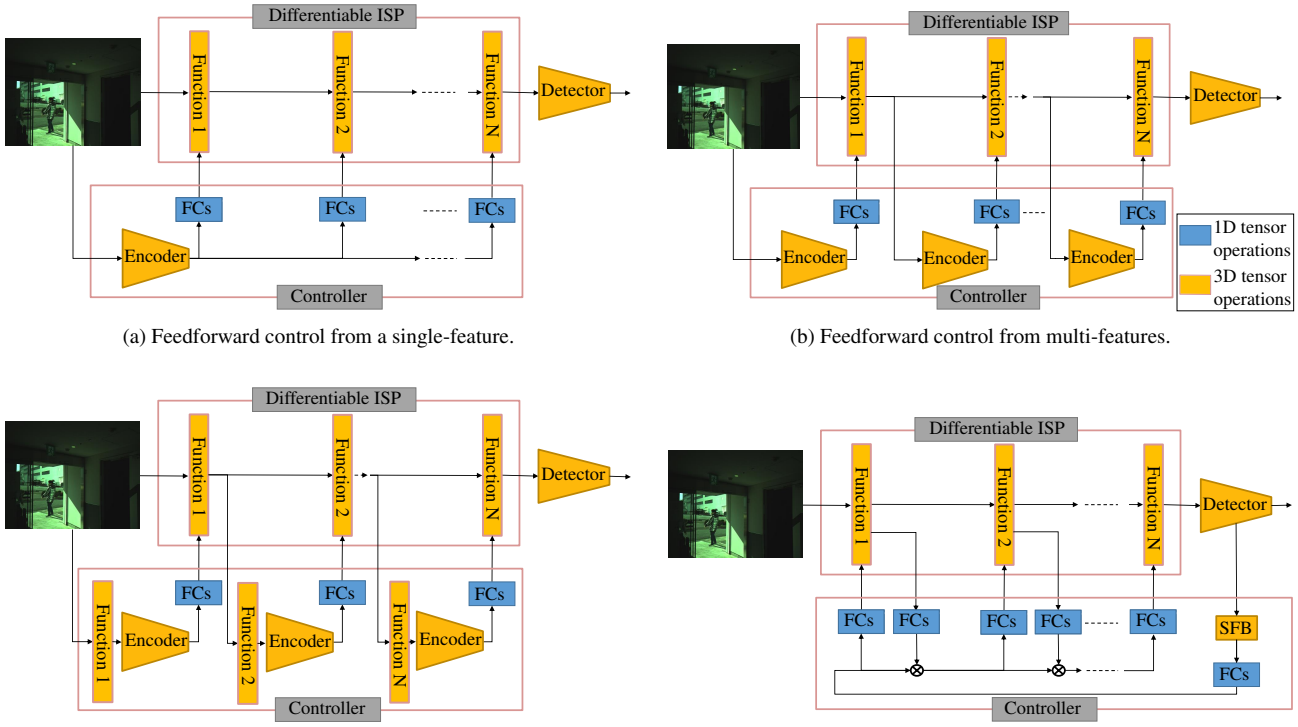
$$I_{GM}(X) = X^{\frac{1}{p_{\gamma 1}} \cdot \frac{1-(1-p_{\gamma 2})X^{\frac{1}{p_{\gamma 1}}}}{1-(1-p_{\gamma 2})p_k^{\frac{1}{p_{\gamma 1}}}}}. \tag{5}$$

Contrast Stretcher (CS)

We implement CS as a simple linear function of $I_{CS}(X) = q_b X + q_c$. Because DNN can process any range of value, we do not restrict the range.

Table 1. Control of Multi-layer ISPs on the human detection dataset. We add ISP layers in order of effect.

ISP components	w/o LU	w/ LU
GM	48.9	-
GM+CS	49.4	49.4
DN+GM+CS	49.2	49.4
DN+SN+GM+CS	48.0	**49.5**



(a) Feedforward control from a single-feature.

(b) Feedforward control from multi-features.

(c) Feedforward control by exemplifications when using several ISP parameter sets [4].

(d) Feedback control from a single-feature (ours).

Figure 2. Comparison with other possible controllers. The (a) is the most lightweight possible controller among feedforward control. The (b) is based on typical dynamic neural network architectures [14, 2, 3]. The (c) is a more advanced method of the (b) proposed for RAW image reconstruction [4]. The (d) is the proposed efficient feedback control.

Table 2. Comparison with other possible controllers on the human detection dataset. C is the computational cost of the ISP. ResNet18 or 4-layer CNN is used as the encoder for feedforward controls.

ISP		GFLOPS	#params [M]	latency [ms]	controller [ms]	AP [%]
GM	(a) (encoder = 4-layer CNN)	14.03+C	**14.33**	11.2	3.3	47.5
	+ our "RO+"	14.03+C	**14.33**	11.5	3.6	48.8
	(a) (encoder = ResNet18)	20.99+C	25.77	13.2	5.3	46.1
	+ our "RO+"	20.99+C	25.77	13.6	5.7	**49.4**
	(d) ours	**13.65+C**	15.66	**8.3**	**0.4**	49.1
DN+SN+GM+CS	(a) (encoder = ResNet18)	20.99+C	25.77	31.1♠	5.3	48.3
	+ our "RO+" and "LU"	20.99+C	25.87	31.6♠	5.8	48.5
	(b) (encoder = 4-layer CNN)	15.23+C	**14.72**	29.1♠	3.3	47.7
	+ our "RO+"	15.23+C	**14.72**	29.5♠	3.7	49.0
	(c) (encoder = 4-layer CNN)	15.93+6C	14.73	101.1♠	9.2 ♣	49.4
	(d) ours	**13.65+C**	15.76	**27.7**♠	**1.9**	**49.6**

♠: Our implementations of DN and SN have a lot of duplicate calculations and are not optimized.

♣: Because of the above, the controller's latency is measured in case all four layers are GM to make a fair comparison.

2. Additional Evaluations

2.1. Evaluation on Human Detection

Multi-Layer Control

A more detailed ablation study is performed for multi-layer control. Here, we add ISP layers in order of effect one by one. As the Table 1 shows, the proposed method without LU struggles to control multi-layer ISPs. The proposed LU successfully disentangles the difficulty of multi-layer control and boosts the accuracy from the setting of only containing GM tone mapping.

Table 3. Detailed evaluation on LODDataset [7] trained with simulated RAW-like data converted from COCO Dataset [9].

	mAP@0.5:0.95 per exposure ratio to default						
	1/10	1/20	1/30	1/40	1/50	1/100	Ave.
as is	23.3	16.7	14.4	114.1	13.4	3.6	14.3
SID [1]	25.8	20.0	16.4	15.1	13.2	6.7	16.2
Zero DCE [5]	32.5	25.3	23.4	21.5	17.8	8.9	21.6
REDI [8]	33.6	30.2	26.1	24.6	23.4	14.1	25.4
H. Yang et. al. [7]	38.5	31.7	29.3	27.8	27.1	18.1	28.8
diff. tuning (GM) [15]	42.8	34.9	39.5	38.9	29.4	12.6	33.0
NeuralAE (GM)[12]	42.3	35.3	38.5	40.6	29.9	15.6	33.7
NeuralAE (GM+CS)[12]	37.0	31.5	33.4	35.1	28.5	18.3	30.6
ours (GM)	45.2	**41.1**	51.1	**49.1**	**41.4**	33.9	43.6
ours (AG+GM+CS)	**45.8**	**41.1**	**52.1**	48.6	41.2	**34.9**	**44.0**

Comparison with Other Possible Controllers

In this section, the proposed controller is compared with other possible controllers, especially feedforward controllers because most of the dynamic neural networks methods [14, 2, 3] have successfully controlled DNN parameters based on feedforward controls. Fig. 2(a) controls all functions based on a single-feature. It is different from the typical controllers for dynamic neural networks [14, 2, 3] but the most lightweight possible feedforward controller. Fig. 2(b) is based on typical dynamic neural network architectures that control each layer with an output of the previous layer. The problem in applying it to the ISP control is that the output of the previous layer is just an image, so it is necessary to create features from scratch using an encoder. Fig. 2(c) is a more advanced method of Fig. 2(b) proposed for RAW image reconstruction [4]. Several processed images with different parameter sets are input to the encoder as exemplifications, and the output parameters from the encoder are determined as the weighted average of the parameter sets. Note that the inverse pipeline is not implemented because it is trained only with detection loss in our problem setup. The number of exemplifications is set as five. We use two types of networks with different computational costs as encoders in Fig. 2(a), (b), and (c): ResNet18 [6] or 4-layer light-weight CNNs with ReLU activations, whose kernel sizes, strides, and output channel sizes are (3, 3, 3, 3), (2, 2, 2, 2), and (16, 32, 64, 128). Fig. 2(d) is the proposed feedback control from a feature.

The results are shown in Table 2. In the case of the single-layer ISP setting, the feedforward control exceeds the accuracy of the proposed method by using a large encoder (ResNet18). However, in the experimental setting with the 4-layer CNN, where the computational cost is still higher than the proposed method, the accuracy is inferior. This result indicates that the feedback control is more efficient. By adding more convolutions to the "Semantic Feature Branch" (SFB), the proposed feedback control might improve the accuracy. In our setting, SFB contains only one convolution layer. In addition, the proposed RO+ for controlling a difficult function is found to be effective even for the feedforward controls.

In the case of the multi-layer ISP setting, the proposed method outperforms feedforward controls with lower computational cost. Although the (c) architecture is accurate, it takes a high computational cost because it needs multiple computations of ISPs and encoders. Limitted to feedforward controls, a comparison of (a) and (b) shows that it is more efficient to encode the previous layer's image with multiple small encoders than with one large encoder. On the other hand, our feedback control achieves higher accuracy despite the fact that the control is based on a single shallow layer of feature (the output of the first stage of the detector's ResNet18 backbone). This should be because the following two factors outweigh the difficulty of controlling from a single encoder. One factor should be the advantage that the controller is able to extract what is captured by the detector directly. The other factor should be the effectiveness of the proposed training method for feedback control (PI).

Lastly, the proposed method is lightweight because it does not require image encoders and performs almost only 1D tensor operations.

2.2. Evaluation on Low-Light Recognition

Training with Simulated RAW Images

A more detailed comparison than Table 7 of the main paper is shown in Table 3. It is broken down by the level of under-exposure. Our method obtains the highest accuracy among all levels of under-exposure. The dynamic ISP control is able to convert a broad luminance distribution environment to a preferable distribution for the detector. The visualized comparison is in Fig. 3 and Fig. 4.

Table 4. Evaluation on LODDataset [7] trained with real dark RAW data in LODDataset.

	ISP	mAP@0.5:0.95
H. Yang et. al. [7]	-	44.7
NeuralAE [12]	GM	45.0
NeuralAE [12]	GM+CS	45.5
ours	GM	45.4
ours	AG+GM+CS	**46.2**

Training with Real Dark RAW Images

We also evaluate the case of real RAW images used for training. The real RAW images are randomly split into training data of 1830 images and test data of 400, the same with [7]. The result is shown in Table 4. Our method is confirmed effective for small amounts of real RAW training data.

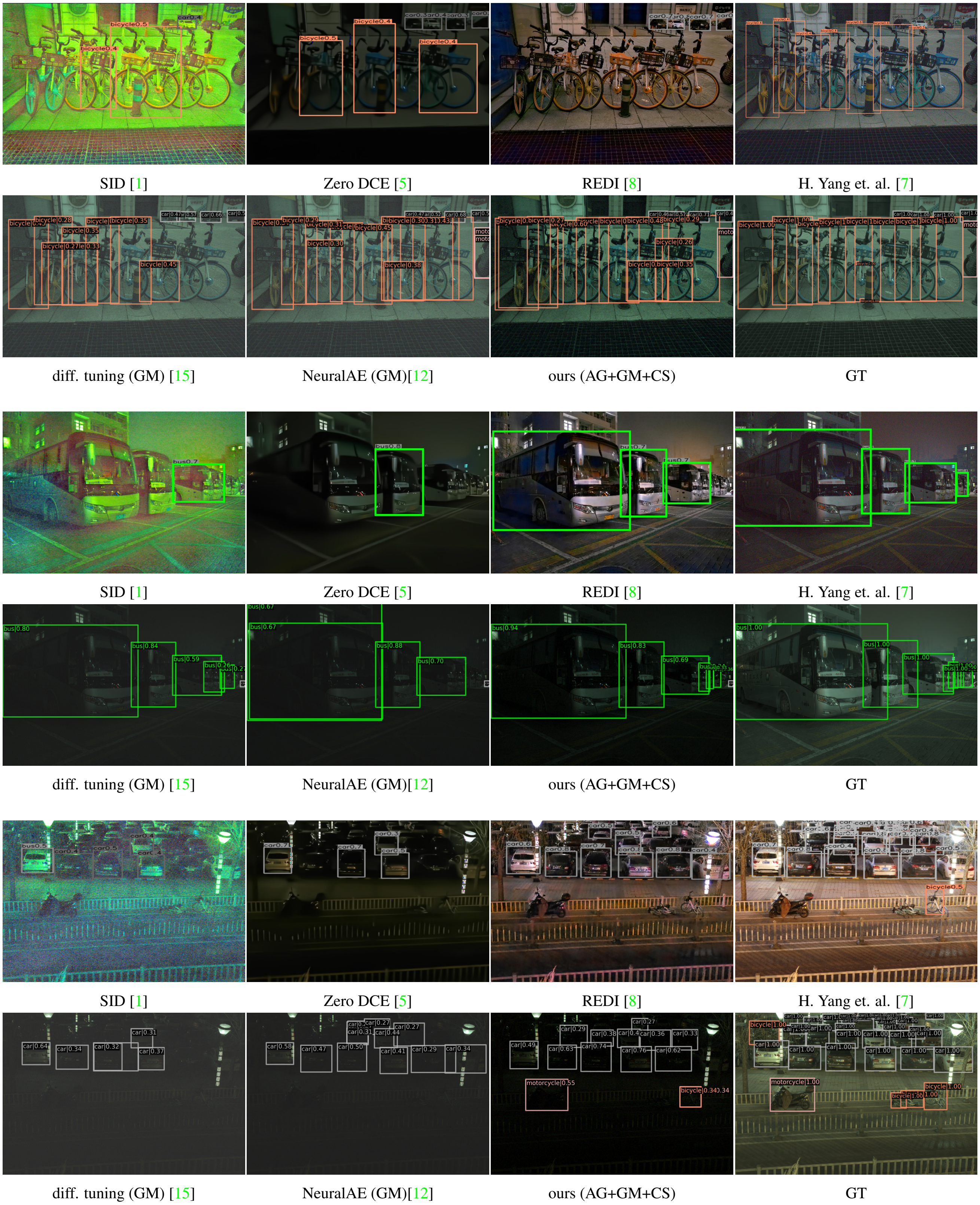
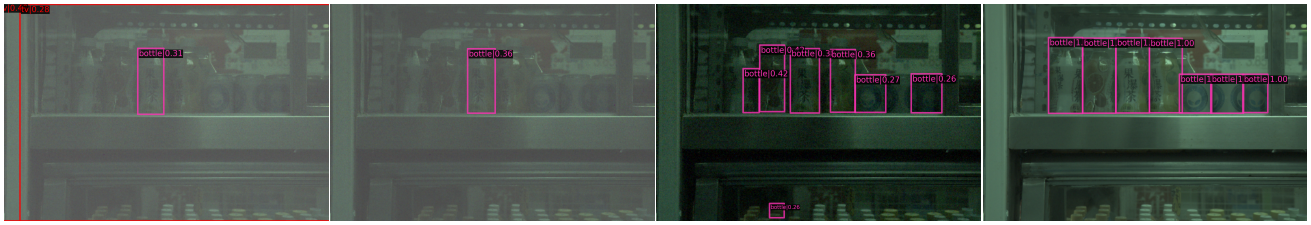
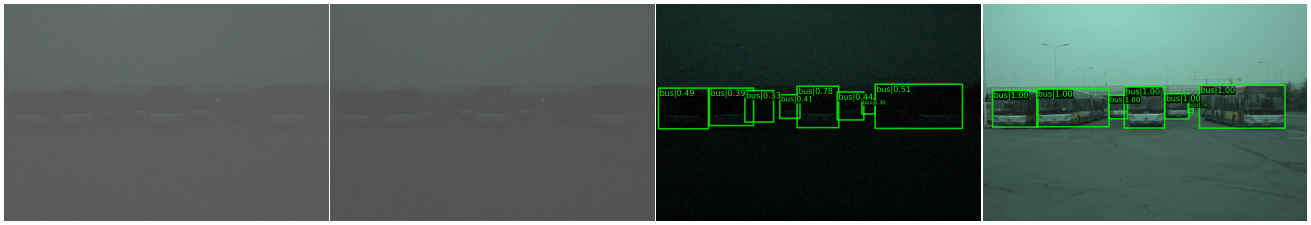


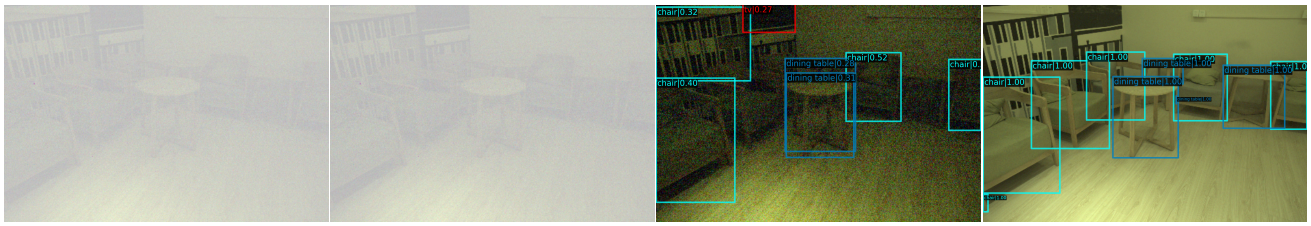
Figure 3. The visualization result on LODDataset [7] trained with simulated RAW-like data converted from COCO Dataset [9]. The results of SID, Zero DCE, REDI, and H. Yang et. al. are from the [7]'s paper.



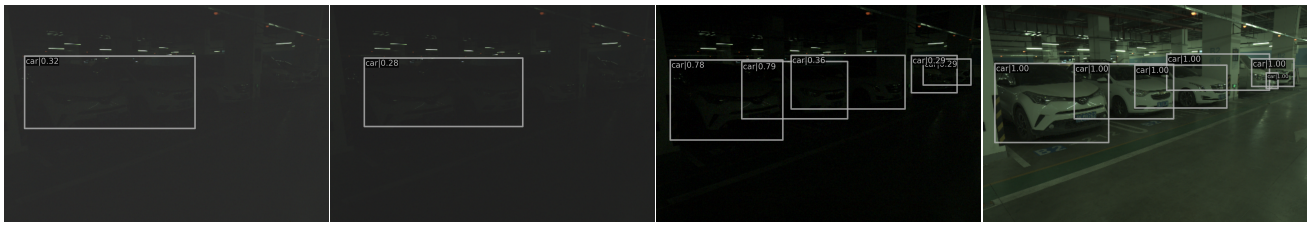
diff. tuning (GM) [15] NeuralAE (GM)[12] ours (AG+GM+CS) GT



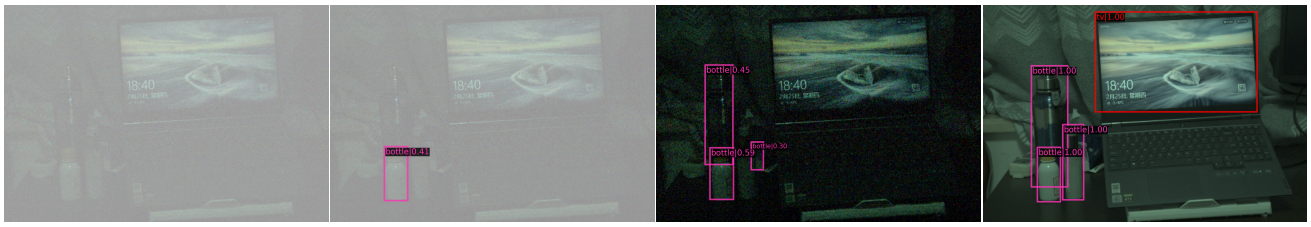
diff. tuning (GM) [15] NeuralAE (GM)[12] ours (AG+GM+CS) GT



diff. tuning (GM) [15] NeuralAE (GM)[12] ours (AG+GM+CS) GT



diff. tuning (GM) [15] NeuralAE (GM)[12] ours (AG+GM+CS) GT



diff. tuning (GM) [15] NeuralAE (GM)[12] ours (AG+GM+CS) GT

Figure 4. The visualization result on LODDataset [7] trained with simulated RAW-like data converted from COCO Dataset [9]. More challenging images than images in Fig. 3 are collected.

References

[1] Chen Chen, Qifeng Chen, Jia Xu, and Vladlen Koltun. Learning to see in the dark. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3291–3300, 2018. 3, 2.2

[2] Yinpeng Chen, Xiyang Dai, Mengchen Liu, Dongdong Chen, Lu Yuan, and Zicheng Liu. Dynamic convolution: Attention over convolution kernels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11030–11039, 2020. 2, 2.1

[3] Yinpeng Chen, Xiyang Dai, Mengchen Liu, Dongdong Chen, Lu Yuan, and Zicheng Liu. Dynamic relu. In *European Conference on Computer Vision*, pages 351–367. Springer, 2020. 2, 2.1

[4] Marcos V Conde, Steven McDonagh, Matteo Maggioni, Ales Leonardis, and Eduardo Pérez-Pellitero. Model-based image signal processors via learnable dictionaries. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 481–489, 2022. 1, 2, 2.1

[5] Chunle Guo, Chongyi Li, Jichang Guo, Chen Change Loy, Junhui Hou, Sam Kwong, and Runmin Cong. Zero-reference deep curve estimation for low-light image enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1780–1789, 2020. 3, 2.2

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2.1

[7] Yang Hong, Kaixuan Wei, Linwei Chen, and Ying Fu. Crafting object detection in very low light. In *Proceedings of the British Machine Vision Virtual Conference*, 2021. 3, 4, 2.2, 2.2, 3, 4

[8] Mohit Lamba and Kaushik Mitra. Restoring extremely dark images in real time. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3487–3497, 2021. 3, 2.2

[9] 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 *European conference on computer vision*, pages 740–755. Springer, 2014. 3, 3, 4

[10] David Marr and Ellen Hildreth. Theory of edge detection. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 207(1167):187–217, 1980. 1

[11] Ali Mosleh, Avinash Sharma, Emmanuel Onzon, Fahim Mannan, Nicolas Robidoux, and Felix Heide. Hardware-in-the-loop end-to-end optimization of camera image processing pipelines. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7529–7538, 2020. 1

[12] Emmanuel Onzon, Fahim Mannan, and Felix Heide. Neural auto-exposure for high-dynamic range object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7710–7720, 2021. 3, 4, 2.2, 2.2

[13] Carlo Tomasi and Roberto Manduchi. Bilateral filtering for gray and color images. In *Sixth international conference on computer vision (IEEE Cat. No. 98CH36271)*, pages 839–846. IEEE, 1998. 1

[14] Brandon Yang, Gabriel Bender, Quoc V Le, and Jiquan Ngiam. Condconv: Conditionally parameterized convolutions for efficient inference. *Advances in Neural Information Processing Systems*, 32, 2019. 2, 2.1

[15] Masakazu Yoshimura, Junji Otsuka, Atsushi Irie, and Takeshi Ohashi. Rawgment: Noise-accounted raw augmentation enables recognition in a wide variety of environments. *arXiv preprint arXiv:2210.16046*, 2022. 1, 3, 2.2, 2.2