Supplementary of Dark-ISP: Enhancing RAW Image Processing for Low-Light Object Detection

Jiasheng Guo[1,*] Xin Gao[1,*] Yuxiang Yan[1] Guanghao Li[1] Jian Pu[1,†]

[1] Institute of Science and Technology for Brain-inspired Intelligence, Fudan University

{guojs22, gaoxin23, yxyan22, ghli22}@m.fudan.edu.cn, {jianpu}@fudan.edu.cn

1. Glossary and Terminology

1.1. Image Signal Processing (ISP) Terms

The role of the ISP is to render the photon data collected by the sensor into the desired image as perfectly as possible. In this section, we outline the steps involved in processing the images recorded by the sensor in a digital camera to produce the final image:

(i) **Camera Sensor**: The camera sensor consists of a two-dimensional grid of photodiodes, where each photodiode is a semiconductor device that converts photons (light radiation) into charge, corresponding to a single pixel in the image. Color filters are placed over these photodiodes to produce color. This arrangement of color filter arrays (CFA) is typically named after Bryce Bayer [4].

(ii) **Pre-processing**: Initial operations applied to Bayer sensor data to correct inherent sensor limitations and prepare the image for further processing. This includes adjustments such as black level correction, which sets the baseline pixel value to zero, and lens shading correction, which compensates for vignetting effects caused by lens imperfections.

(iii) **Noise Reduction**: Techniques employed to minimize unwanted random variations in pixel intensity, known as noise, which can degrade image quality. Noise reduction is crucial for enhancing visual quality and is closely related to exposure time and camera ISO settings.

(iv) **Demosaicing**: The process of reconstructing a full-color image from incomplete color samples output by an image sensor overlaid with the Bayer pattern. This involves interpolating the missing color information to produce a complete RGB image.

(v) **White Balance**: A method used to adjust the colors in an image to match the perceived color of the scene, ensuring that objects that appear white in person are rendered white in the photo. This process compensates for the color temperature of the illumination source.

(vi) **Color Space Transformation**: The conversion of image data from one color space to another. In digital imaging, this often involves mapping white-balanced pixel data to an intermediate color space (e.g., CIEXYZ) and then to a display-referred color space (e.g., sRGB), typically using 3×3 transformation matrices specific to the camera.

(vii) **Color and Tone Correction**: Adjustments made to the color balance and tonal range of an image to achieve the desired visual appearance. These corrections are often implemented using 3D and 1D lookup tables (LUTs) and may include tone mapping to compress the dynamic range of the image.

(viii) **Sharpening**: Techniques applied to enhance the perceived sharpness of an image by emphasizing edges and fine details. Methods such as unsharp masking or deconvolution are commonly used to achieve this effect.

1.2. RAW Image Formats

- **Bayer RAW image**: A Bayer image is a format that captures color information from the sensor arrangement of color filter arrays.
- **RGB-RAW image**: A demosaiced Bayer image is converted to a standard color space and stored in 8-bit RGB format(.png, .JPG, .JEPG).

[*] These authors contributed equally to this work.

[†] Corresponding author.

Table A1. Detection performance comparison on the real-world LOD dataset for COCO metric.

Image Format	Method	ResNet50			ResNet18		
		mAP	mAP 50	mAP 75	mAP	mAP 50	mAP 75
RGB RAW-RGB	default ISP	52.2	77.9	58.1	45.9	73.9	49.5
	demosaic	53.4	79.5	59.7	48.3	76.9	53.2
	LIS [2]	52.0	78.4	57.5	46.4	74.0	49.5
	FeatEnHancer [5]	53.9	80.5	58.5	50.0	78.2	55.3
	RAW-Adapter [3]	53.3	79.8	58.8	47.5	75.7	52.5
Bayer	default ISP	58.4	83.7	65.9	53.2	80.5	58.9
	demosaic	57.2	82.9	64.4	51.5	79.2	57.0
	SID [1]	56.5	82.5	63.0	51.9	79.7	57.5
	LIS [2]	57.8	83.5	65.6	52.4	79.8	58.2
	FeatEnHancer [5]	58.6	84.3	65.7	53.4	81.1	59.9
	RAW-Adapter [3]	57.1	83.2	64.3	52.5	80.4	58.3
	Our Dark-ISP	**58.9**	**84.4**	**66.9**	**53.7**	**81.5**	**60.9**

Table A2. Ablation study for Linear component. Camera means the conventional White Balance and Color Space Transform operations in the ISP. Local indicates that we only allow the parameter matrix to perform Local Attention operations with the image. Global signifies that we only allow the parameter matrix to engage in Global Attention operations with the image. CCM refers to our approach of using only the 3×3 camera color correction matrix as the prediction target.

Method	mAP	Param(MB)
Camera	66.4	-
Local	67.0	0.168
Global	68.6	0.177
Our Dark-ISP(CCM 3×3)	69.2	0.343
Our Dark-ISP(CCM 3×4)	**70.4**	0.345

- **sRGB image**: A standard RGB (Red, Green, Blue) color space created cooperatively by HP and Microsoft for use on monitors, printers, and the internet. It defines a specific range of colors that can be displayed, ensuring consistency across different devices.

2. Supplementary Experimental Results

This section provides additional experimental details and results not presented in the main paper, including performance evaluations and ablation studies.

2.1. Detection Performance on LOD Dataset

Tab. A1 presents the detection performance results on the LOD dataset using COCO metrics. As shown, our method outperforms competing approaches in both mAP, mAP_{50}, and mAP_{75}, across both ResNet50 and ResNet18 backbones.

2.2. Ablation Study: Linear Component

The Linear Component aims to learn a linear transformation that incorporates both global and local information from the image. We examine the impact of the two types of information on detection results, as shown in Tab. A2. The White Balance and Color Space Transformation operations from traditional ISP are used as baseline for comparison. Both types of information contribute to the learning of image color transformations. Compared to the local information (67.0 mAP), the global information (68.6 mAP) shows a more significant improvement. This may be because the linear operations in traditional ISP are inherently global operations, making the global information more compatible. Nevertheless, the information gains from the two are not coupled; thus, integrating them together leads to an additional performance improvement. Additionally, we also try using only the CCM matrix as the prediction target. Its size is 3x3, requiring the prediction of only 9 parameters, which makes its flexibility lower compared to our 3x4 joint matrix (12 parameters), resulting in a performance of 69.2 mAP.

2.3. Ablation Study: Self-Boost Regularization Starting Epoch

We explore its impact on model convergence and performance. Self-Boost Regularization embodies the idea that 'there is a hierarchy in learning and specialization in skills.' It allows advanced deep nonlinear features U to enhance the learning of weakly gradient-propagating shallow linear features I', provided that U possesses sufficient knowledge. In practice, Self-Boost Regularization is activated after a few epochs of training to prevent U from misleading I' before it has adequately learned. Therefore, we conducted an ablation study on the activation epoch regarding this point, and the results are shown in Fig. A1. It can be observed that at the beginning of training, due to U not being fully converged and the randomness of the weights being high, the learning of I' was misled, resulting in a performance drop compared to the baseline without this loss (**68.7 mAP**). As the activation epoch is pushed further back, U becomes increasingly stable and starts to correctly guide I', leading to a gradual performance increase. After activation at the 10th epoch, it reaches its peak at **70.4 mAP** .

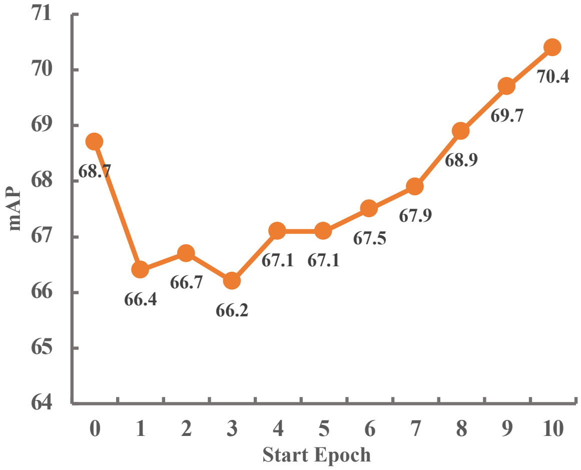


Figure A1. Ablation Study on the Starting Epoch of Self-Boost Regularization.

3. Nonlinear Function Design and Analysis

3.1. Impact of Tone Mapping on Image Representation

In our study, we demonstrated the impact of tone mapping functions with different shapes on images. As shown in Fig. A2, the concave function stretches the pixel values in the low-brightness regions while compressing the high-brightness areas, resulting in a brighter overall image. In contrast, the convex function has the opposite effect, making the originally dark regions even darker. The S-curve function further accentuates the shadow effects of objects, we believe it can also be beneficial for object detection tasks.

3.2. Design of Non-Convex Polynomial Basis Functions

Based on the above physical properties, we designed the basis functions in the nonlinear component. They satisfy the following properties: First, each function must pass through the points (0,0) and (1,1), representing the minimum and maximum brightness values of each pixel. Secondly, they must be non-convex functions on the interval [0,1] for dim image tones in low light environment. Therefore we selected polynomial functions of orders one to eight that satisfy the above two properties as bases to learn the ideal nonlinear mapping that best fits the detection task. The specific expressions of each basis function are shown in Fig. A3(a). We introduced skip connections to prevent the vanishing gradient problem caused by an increase in the number of network layers. To be compatible with this operation, each basis function needs to subtract x from itself, and the resulting shape of each function are shown in Fig. A3(b).

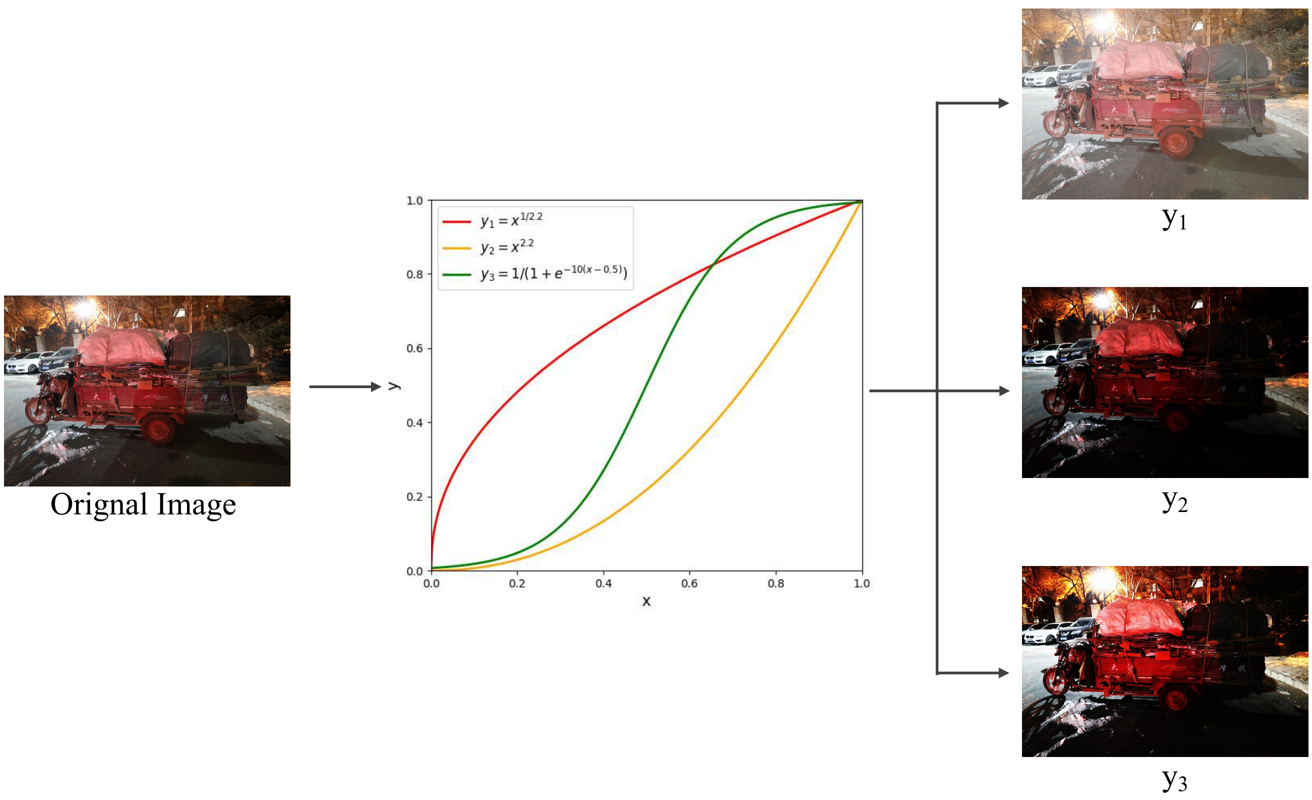


Figure A2. The impact of tone mapping functions with different shapes on images.

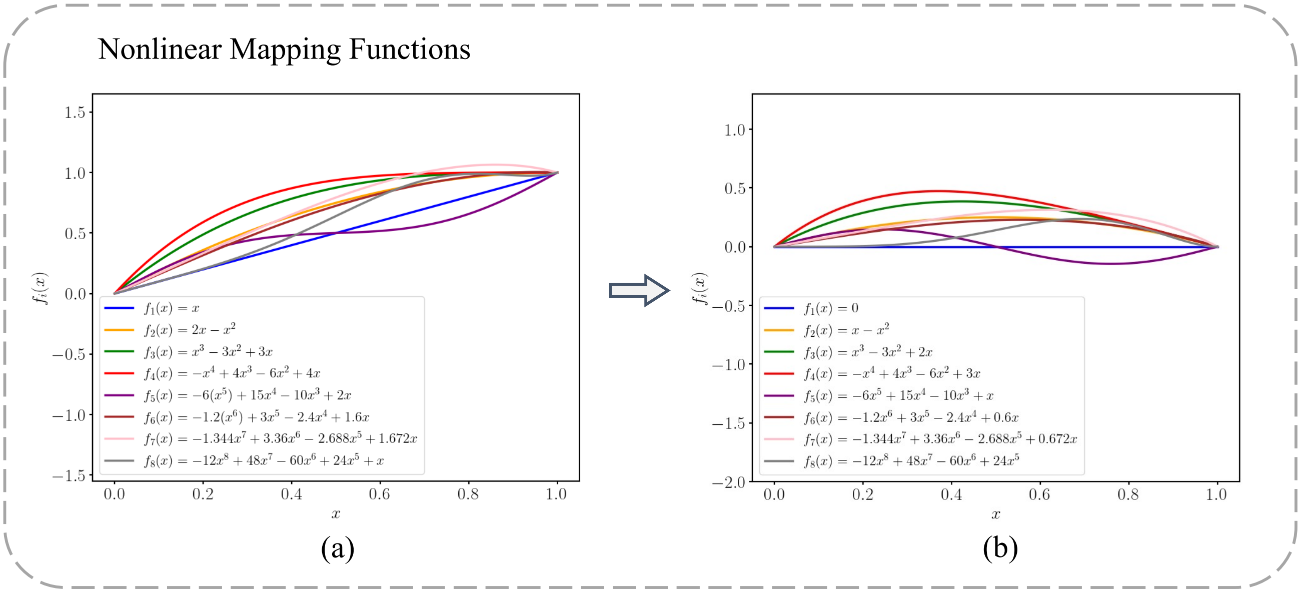


Figure A3. Specific functions of nonlinear mapping bases. (a) Predefined first to eighth order polynomial basis functions. (b) Subtracting x to align with the skip connection and maintain curve shape.

4. Discussion About Self-Boost Regularization

4.1. The validity of the $\tilde{P}$ definition

To verify the essence of the Self-Boost regularization derived from the defined equation

$$\tilde{P} := U \cdot I^T \cdot (I \cdot I^T)^{-1}, \tag{A1}$$

which ensures that $P' \cdot I$ approaches U sufficiently, we calculated the average values of $|U - P' \cdot I|$ and the norm of $\frac{\partial}{\partial P'}|U - P' \cdot I|^2$ for each epoch after activating this mechanism from the 10th epoch during training. As shown in Fig. A4, both values gradually converge to zero, demonstrating the effectiveness of this definition.

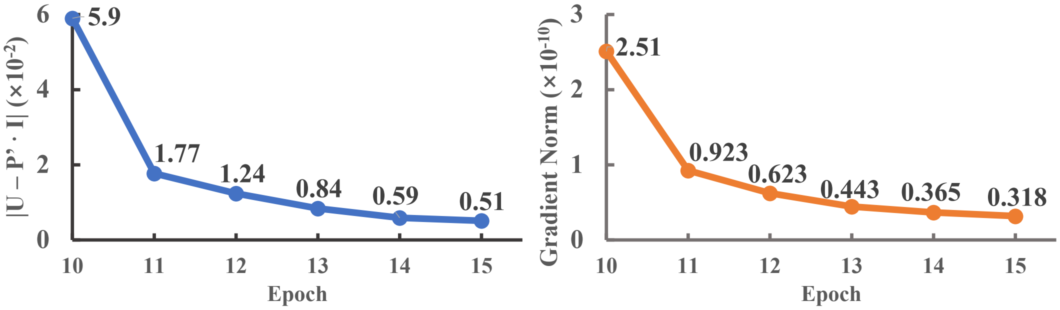


Figure A4. The trend of $|U - P'I|$ and the norm of the gradient.

4.2. Numerical Stability of the Self-Boost Regularization

In Section 3.3 of the main text, we use the inverse of $(I \cdot I^T)^{-1}$ in Eq. (9) and Eq. (10) to compute the self-supervised signal for regularizing the learnable camera parameters P'. However, when performing matrix inversion operations, it is essential to consider the condition number of the matrix $I \cdot I^T$ to ensure numerical stability. This is because, when the condition number is large (e.g., approaching or exceeding 10^5 or higher), the matrix may become ill-conditioned, meaning it is numerically close to being singular or degenerate. In such cases, the inversion or pseudoinversion process can become unstable, leading to inaccurate results and potentially compromising our method.

From a theoretical perspective, $I \cdot I^T$ is almost always invertible, as guaranteed by Sard's theorem. Consequently, $I \cdot I^T$ is almost always positive definite, and its condition number is given by the ratio of the largest to the smallest eigenvalue:

$$\kappa(A) = \frac{\lambda_{\max}}{\lambda_{\min}},$$

where $\lambda_{\max}$ and $\lambda_{\min}$ denote the maximum and minimum eigenvalues of A, respectively. For non-invertible cases, we resort to using `torch.linalg.pinv`, which internally computes the pseudoinverse via Singular Value Decomposition (SVD). This is a highly robust method, as SVD performs a numerically stable decomposition of the matrix. For matrices with large condition numbers, SVD reduces numerical errors by truncating small singular values, thus stabilizing the pseudoinverse computation. This truncation effectively discards certain regularization directions, as it introduces numerical instability by neglecting smaller components. However, this does not significantly impact the final regularization of the camera parameters, as we are only selecting the more reliable directions for regularization.

Additionally, we evaluated the condition number of $I \cdot I^T$ for the images used in experiments. As shown in Fig. A5, the overall condition numbers of the matrices are not large, indicating that the inversion operations performed during our approach are numerically stable and do not pose significant risks for instability.

5. Implementation Details

5.1. Architectural Details of Global-Local Attention

Fig. A6 provides a detailed visualization of the global-local attention mechanism used in our model. The figure illustrates how attention is applied locally and globally to enhance feature learning.

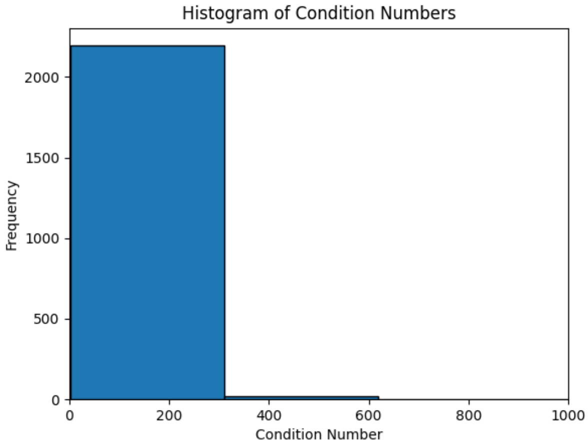


Figure A5. Histogram of the Condition Number of $I \cdot I^T$.

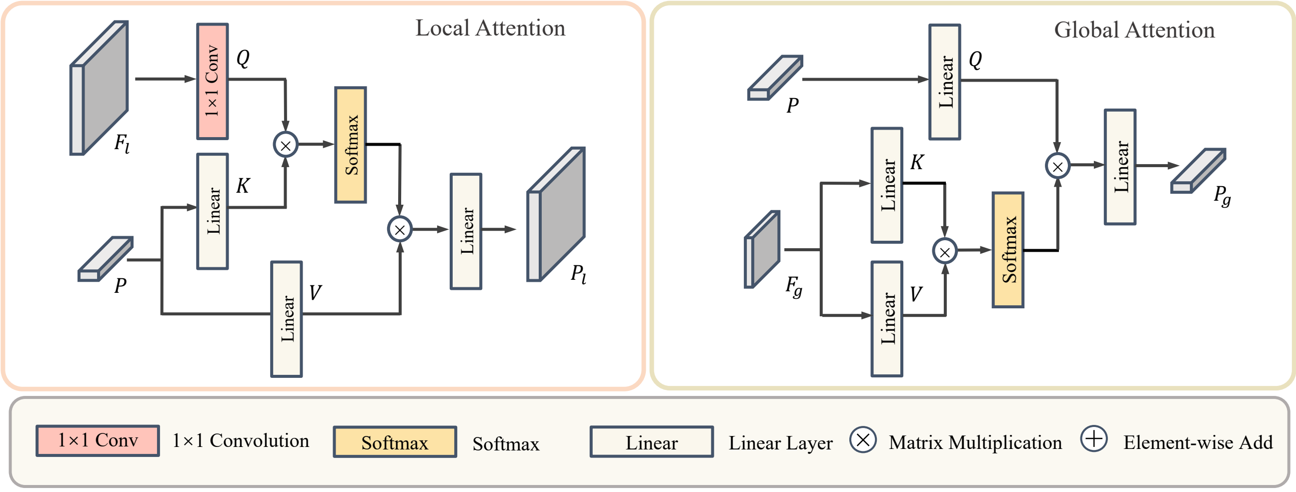


Figure A6. The Local Attention and Global Attention

5.2. Implementation of Comparative Methods

This section details the implementation of baseline methods compared in the main paper, including FeatEnHancer and RAW-Adapter [3]. For a fair comparison, the implementation of all methods is consistent with the open-source libraries in the **github**. In the original method, LIS and RAW-Adapter [3] use RAW-RGB data, which represents the average of the two green channels of the Bayer image after compressing the bit depth and storing it in the sRGB color space. FeatEnHancer [5] takes dark RGB as input directly, and SID [1] takes Bayer image as input directly. In our experiments, we also implemented the aforementioned comparative methods on data generated directly from Bayer images. For LIS [2] and RAW-Adapter [3], we similarly averaged the two green channels as in LIS [2], but without bit depth compression, directly passing the float-type tensor to the network. For FeatEnHancer [5], we processed the Bayer image using the same workflow as the default ISP used in the LED [6] open-source library, obtaining a float-type representation of the RGB image before inputting it to the network.

6. Efficiency Analysis

We conducted an efficiency analysis of each model on the LOD dataset, while also introducing two additional methods for comparison: ROAD[8] and IA-ISP[7]. Both methods utilize the Bayer format as input for end-to-end training of the down-stream task. As shown in A3, Dark-ISP maintains high efficiency and low memory usage, with inference time comparable to other methods while achieving superior performance. Each module in our method effectively leverages rich prior knowledge, resulting in a lightweight ISP module that strikes an optimal balance between performance and efficiency, making it highly suitable for real-world applications.

Table A3. Efficiency analysis for each method. The parameters exclusively count the modules apart from the downstream task detector.

Methods	mAP	Params(MB)	Inference Time (ms)	GFLOPS
SID[1]	64.7	29.60	3.48	97.91
LIS[2]	67.9	3.30	3.24	51.95
RAOD[8]	66.0	0.28	3.28	51.50
IA-ISP[7]	67.0	0.63	3.56	51.75
RAW-Adapter[3]	66.2	2.19	3.51	52.23
FeatEnHancer[5]	67.0	0.53	3.95	78.58
Our Dark-ISP	**70.4**	0.49	3.42	83.32

7. Additional Visual Comparisons

This section includes additional qualitative comparisons of detection results on the NOD dataset. Fig. A7 and Fig. A8 demonstrate the qualitative performance comparison of Dark-ISP with other state-of-the-art methods (FeatEnHancer [5], RAW-Adapter [3]). Our method consistently detects more objects with higher accuracy in low-light conditions, avoiding false positives and missed detections.

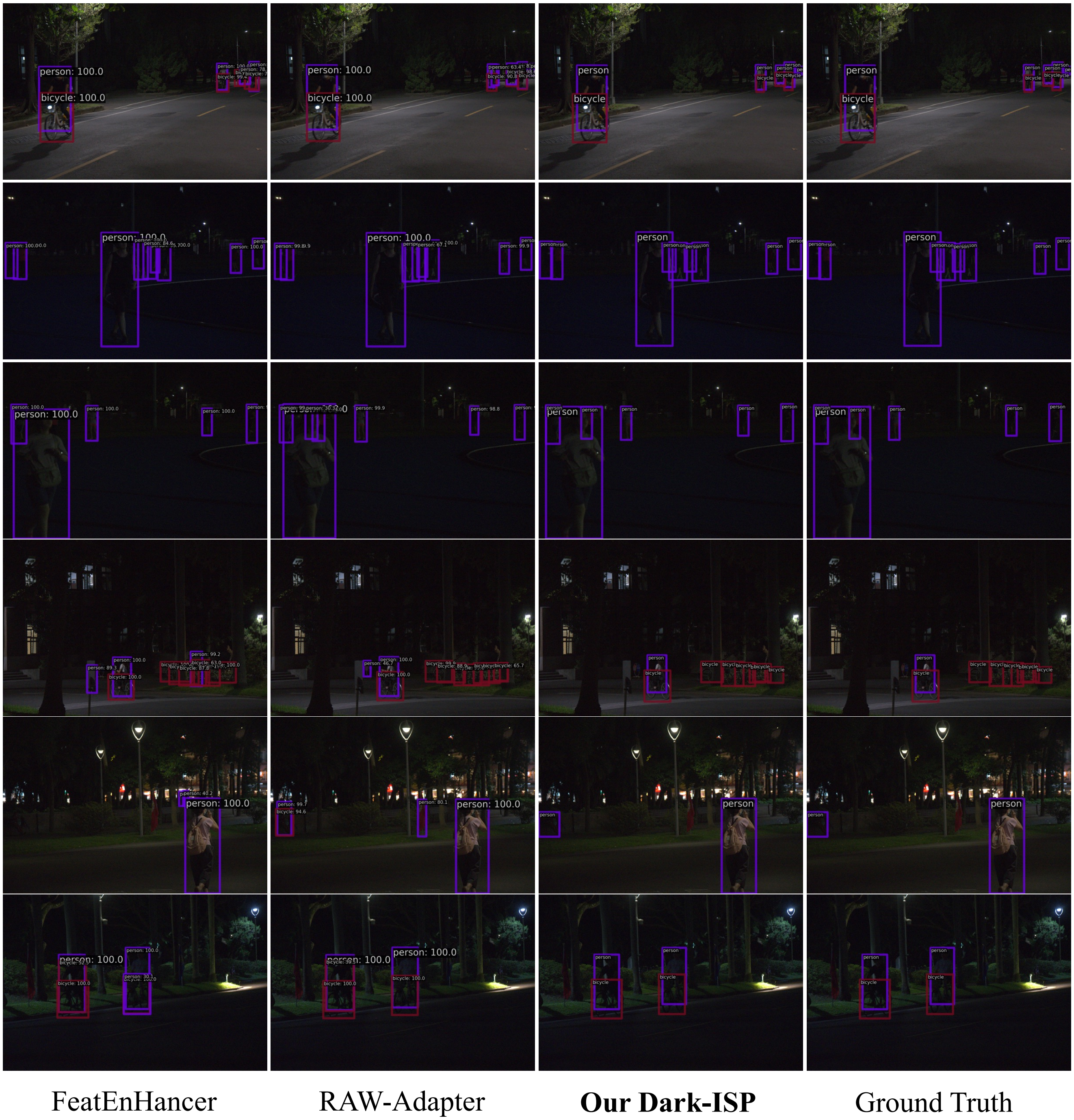
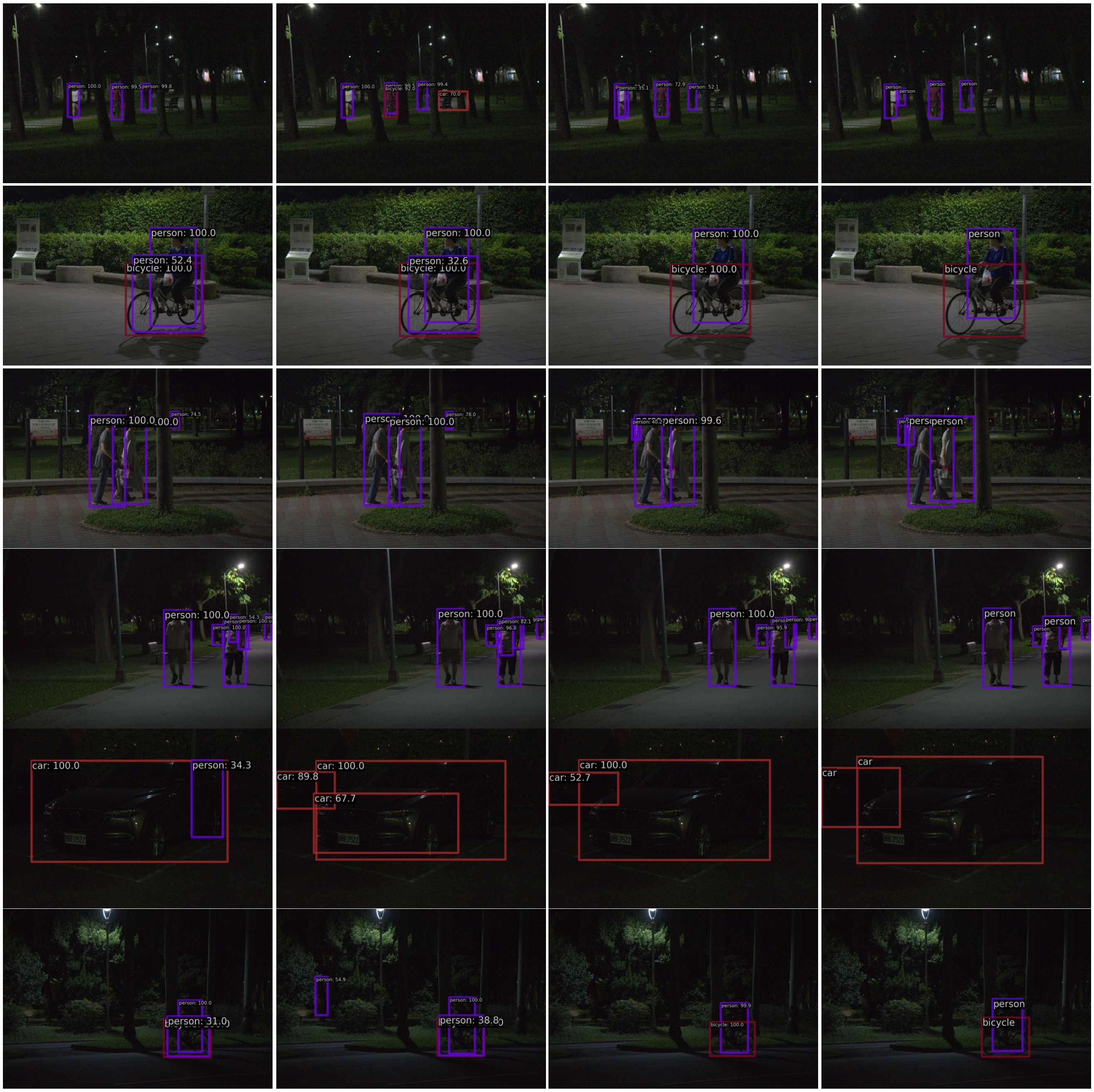


Figure A7. Visual comparisons on the NOD dataset.



FeatEnHancer RAW-Adapter **Our dark-ISP** Ground Truth

Figure A8. Visual comparisons on the NOD dataset.

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. 2, 6, 7

[2] Linwei Chen, Ying Fu, Kaixuan Wei, Dezhi Zheng, and Felix Heide. Instance segmentation in the dark. *International Journal of Computer Vision*, 131(8):2198–2218, 2023. 2, 6, 7

[3] Ziteng Cui and Tatsuya Harada. Raw-adapter: Adapting pre-trained visual model to camera raw images. In *European Conference on Computer Vision*, pages 37–56. Springer, 2024. 2, 6, 7

[4] Mauricio Delbracio, Damien Kelly, Michael S Brown, and Peyman Milanfar. Mobile computational photography: A tour. *Annual review of vision science*, 7(1):571–604, 2021. 1

[5] Khurram Azeem Hashmi, Goutham Kallempudi, Didier Stricker, and Muhammad Zeshan Afzal. Featenhancer: Enhancing hierarchical features for object detection and beyond under low-light vision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6725–6735, 2023. 2, 6, 7

[6] Xin Jin, Jia-Wen Xiao, Ling-Hao Han, Chunle Guo, Ruixun Zhang, Xialei Liu, and Chongyi Li. Lighting every darkness in two pairs: A calibration-free pipeline for raw denoising. In *ICCV*, 2023. 6

[7] Wenyu Liu, Gaofeng Ren, Runsheng Yu, Shi Guo, Jianke Zhu, and Lei Zhang. Image-adaptive yolo for object detection in adverse weather conditions. In *AAAI*, number 2, 2022. 6, 7

[8] Ruikang Xu, Chang Chen, Jingyang Peng, Cheng Li, Yibin Huang, Fenglong Song, Youliang Yan, and Zhiwei Xiong. Toward raw object detection: A new benchmark and a new model. In *CVPR*, 2023. 6, 7