Toward RAW Object Detection: A New Benchmark and A New Model

Ruikang Xu[1]* Chang Chen[2]* Jingyang Peng[2]* Cheng Li[2]
Yibin Huang[2] Fenglong Song[2] Youliang Yan[2] Zhiwei Xiong[1]†

[1]University of Science and Technology of China [2]Huawei Noah's Ark Lab

xurk@mail.ustc.edu.cn, zwxiong@ustc.edu.cn,
{chenchang25, pengjingyang1, licheng89, huangyibin1, songfenglong, yanyouliang}@huawei.com

Abstract

In many computer vision applications (e.g., robotics and autonomous driving), high dynamic range (HDR) data is necessary for object detection algorithms to handle a variety of lighting conditions, such as strong glare. In this paper, we aim to achieve object detection on RAW sensor data, which naturally saves the HDR information from image sensors without extra equipment costs. We build a novel RAW sensor dataset, named ROD, for Deep Neural Networks (DNNs)-based object detection algorithms to be applied to HDR data. The ROD dataset contains a large amount of annotated instances of day and night driving scenes in 24-bit dynamic range. Based on the dataset, we first investigate the impact of dynamic range for DNNs-based detectors and demonstrate the importance of dynamic range adjustment for detection on RAW sensor data. Then, we propose a simple and effective adjustment method for object detection on HDR RAW sensor data, which is image adaptive and jointly optimized with the downstream detector in an end-to-end scheme. Extensive experiments demonstrate that the performance of detection on RAW sensor data is significantly superior to standard dynamic range (SDR) data in different situations. Moreover, we analyze the influence of texture information and pixel distribution of input data on the performance of the DNNs-based detector. Code and dataset will be available at https://gitee.com//mindspore/models/tree/master/research/cv/RAOD.

1. Introduction

Real-world scenes are complex and of high dynamic range (HDR), especially in extreme situations like the direct light of other vehicles. In many computer vision applications, such as autonomous driving and robotics, HDR data is important and necessary for making safety-critical decisions [34] since it extends the captured luminance. For instance, images may easily get over-exposed in brighter areas on standard dynamic range (SDR) images, but there may be important information in corresponding raw regions. To obtain the HDR data, recent works use additional cameras and even unconventional sensors, such as neuromorphic cameras and infrared cameras [13, 25], which inevitably brings extra costs. In this paper, we make the first effort to achieve object detection on RAW sensor data, which naturally stores HDR information without any additional burden.

RAW sensor data is generated from the image sensor, and is the input data of the image signal processor (ISP), rendering SDR images suitable for human perception and understanding. RAW sensor data is naturally HDR and save all information from image sensors. However, datasets from the RAW domain are difficult to collect, store and annotate. And, to the best of our knowledge, there is no large-scale HDR RAW sensor dataset available for object detection. Existing RAW sensor datasets are no more than 14-bit and not large enough for practical applications. For instance, PASCALRAW dataset [27] is of only 12-bit, which is not wide enough to handle complex lighting conditions. To fill this gap, we create a novel RAW sensor dataset for object detection on the driving scene, named as ROD, which consists of 25k annotated RAW sensor data in a 24-bit dynamic range on day and night scenarios.

On the other hand, most Deep Neural Networks (DNNs)-based object detection algorithms are designed for the common SDR data, which only records information in the 8-bit dynamic range. Hence, we first investigate the impact of dynamic range for DNNs-based detection algorithms and experimentally find that directly applying these DNNs-based detectors on HDR RAW sensor data results in significant performance degradation, and it gets worse when the dynamic range increases. Then, we analyze the key component of the ISP system and demonstrate the importance of dynamic range adjustment for RAW detection.

In this paper, we propose an adjustment method for the effective detection on RAW sensor data, which is jointly optimized with the downstream detection network in an end-to-end scheme. Note that our proposed method is trained together with the detector from scratch only using object an-

*Equal contribution. †Corresponding author. This work was done when Ruikang Xu was a research intern at Huawei Noah's Ark Lab.

notations as the supervision. To effectively exploit the HDR information from RAW sensor data, we devise an image-adaptive processing network to regulate RAW sensor data with learnable transformation functions. Specifically, we design two modules to adjust the dynamic range of RAW sensor data by image-level and pixel-level information. In addition, our proposed method is lightweight and computationally efficient.

Extensive experimental results on the proposed ROD dataset demonstrate that the performance of object detection on RAW sensor data is significantly superior to detection on SDR data in different scenarios. Our method also outperforms recent state-of-the-art neural ISP methods [23, 41]. Comprehensive ablation experiments show that our proposed method effectively improves the performance of DNNs-based object detection algorithms on HDR RAW sensor data. Furthermore, we analyze the influence of texture information and pixel distribution of the input data for the performance of the downstream detection network.

In summary, the main contributions are as follows:

- We build a novel RAW sensor dataset for object detection on HDR RAW sensor data, which contains 25k driving scenes on both day and night scenarios.
- We propose a simple and effective adjustment method for detection on HDR RAW sensor data, which is jointly optimized with the detector in an end-to-end manner.
- Extensive experiments demonstrate that object detection on HDR RAW sensor data significantly outperforms that on SDR data in different situations. It also shows that our method is effective and computationally efficient.

2. Related Work

2.1. HDR Imaging

HDR imaging is an important technique that greatly extends the dynamic range of exposure and accurately represents a wide range of illuminance, ranging from sunlight to shadows [5, 34, 35, 40]. Traditional methods take multiple SDR images under different exposures, then merge them with different weights to reproduce HDR data [6,18,26,37]. These methods may lead to ghosting in the generated HDR data due to misalignment caused by camera movement or changes in the scene [15, 38]. Recent HDR imaging methods use additional cameras and even unconventional sensors to combine a fusion camera system, such as neuromorphic cameras [4, 13, 33] and infrared cameras [19, 24, 25]. These methods add extra hardware costs and computational burdens. In addition, the camera sensors of the fusion system are not perfectly aligned, which distorts the reconstructed HDR data. In this work, we propose to achieve object detection on the RAW sensor data, which naturally stores the HDR information without any additional burden. Furthermore, we propose a new RAW sensor dataset with a SONY IMX490 sensor, which compacts sub-pixels with two exposure times each to generate the 24-bit RAW sensor data by a linear combination of four SDR RAW sensor data.

2.2. Object Detection

Object detection aims at localizing a set of objects and recognizing their categories in an image, which is one of the most fundamental computer vision problems in the past few decades [7, 21, 36, 42]. Mainstream object detection pipelines can be roughly divided into two categories. One category is the one-stage proposal-free detector, which predicts detection results by regressing the coordinates of predefined anchors and simultaneously classifying categories by a single CNN, such as SSD [22] and RetinaNet [20]. Especially, YOLO series [8, 28–30] achieve promising performance on many benchmark datasets. Another category is the two-stage region proposal-based detector, which first extracts a set of regions of interest (RoIs) from input images, and then refines the location of each RoI and predicts its class labels, such as R-CNN [10], Faster R-CNN [31], DETR [3] and Sparse R-CNN [32] are powerful proposal-based detectors with high detection performance. However, the existing object detection methods are designed for the SDR data and cannot fit the HDR RAW sensor data, which results in significant performance degradation [43]. We propose a novel RAW sensor dataset and a method to achieve object detection on RAW sensor data. Specifically, we employ YOLO-X [8] and Sparse R-CNN [32] as the baseline detectors, which are the state-of-the-art methods of the above two representatives respectively.

3. RAW Object Detection Dataset

For object detection algorithms in many practical applications, such as autonomous driving, the HDR data is essential to handle complex real-world scenarios. RAW sensor data naturally stores the HDR information without additional equipment cost. To the best of our knowledge, there is no large-scale HDR RAW sensor dataset available for object detection. To fill this gap, we create a novel RAW sensor dataset for object detection on the HDR driving scene, named as ROD. We believe that the ROD dataset can serve as a benchmark for future works targeting object detection in RAW domain.

3.1. Data Collection and Processing

The ROD dataset consists of 25 thousand annotated RAW sensor data in both day and night scenarios, which is collected by a Sony IMX490 imaging sensor (Bayer sensor, with a resolution of 2880 × 1856). To generate the HDR data, we combine the RAW sensor data acquired with different exposure times. Specifically, the sensor compacts sub-pixels with two exposure times each, thus fused 24-bit RAW sensor data are a linear combination of four 12-bit RAW sensor data. As shown in Figure 1, the dataset presents a variety of driving scenes in unconstrained envi-

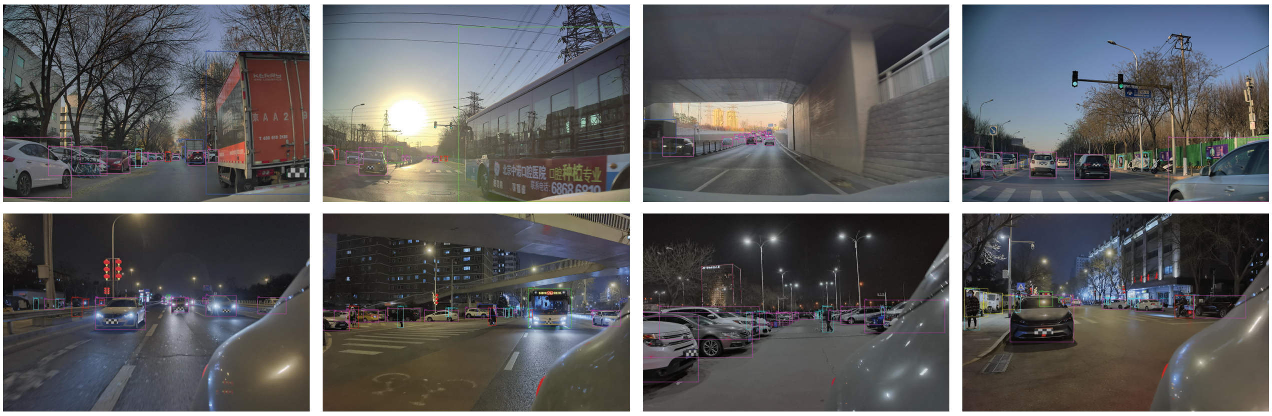


Figure 1. Example scenes in our ROD dataset. Top: Image instance captured in the day driving scenes. Bottom: Image instance captured in night driving scenes. We show the corresponding SDR data for better visualization.

Table 1. Comparison between the PASCALRAW, LOD, and our RAW sensor datasets.

Dataset	Sensor	Dynamic Range	Images	Category	Instance	Scenario
PASCALRAW [27]	Nikon D3200 DSLR	12-bit	4,259	3 classes	6,550	Day
LOD [39]	Canon EOS 5D Mark IV	14-bit	2,230	8 classes	9,726	Low-light
Ours	Sony IMX490	24-bit	25,207	6 classes	237,379	Day & Night

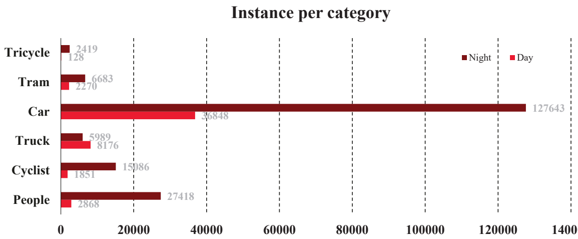


Figure 2. Number of instances per category for our ROD dataset.

ronments. We aim to build the HDR RAW sensor dataset for autonomous driving. Specifically, we annotate 237 thousand bounding boxes with 6 common class labels in the driving scene, which are car, pedestrian, cyclist, tram, tricycle, and track. The number of instances per category for all 6 categories collected is shown in Figure 2. The dataset will be made publicly available as a benchmark for future methods targeting object detection on the HDR RAW data.

3.2. Comparison to Existing Datasets

The ROD dataset is composed of 24-bit RAW sensor data. It can be observed from Table 1 that our ROD dataset is larger than other datasets and has a higher dynamic range. Different from the PASCALRAW dataset [27] and the LOD dataset [39], our ROD dataset contains day and night scenarios. Although the LOD dataset has more categories, our ROD dataset has much more instances. The LOD dataset consists of the long-exposure RGB and short-exposure RAW pairs in daily scenes, which is not fit to understand the practical HDR driving scenes. The proposed ROD dataset consists of a large number of real-world driving scenes, which aims to facilitate object detection algorithms to be used in practical applications.

4. Analysis on RAW Detection

4.1. Impact of Dynamic Range for Object Detection

To handle a variety of lighting conditions, HDR data is necessary and important. But existing object detection algorithms are designed for 8-bit SDR data, which has a much lower dynamic range than our 24-bit HDR data. Hence, we try to investigate the impact of dynamic range on object detection at first. We train and test the YOLOX [8] with different parameters on RAW sensor data with different dynamic ranges and the corresponding SDR images, respectively. Specifically, the 10-bit dataset is captured by HUAWEI Mate20 cellphone, the 12-bit dataset is the PASCALRAW [27] dataset, and the 24-bit dataset is the day scenes of the proposed ROD dataset (10k images). The 10-bit dataset is only collected for ablation experiments, which consists of 8k images and 29k instances on the day scenario with the same categories as the proposed ROD dataset.

Experiment results are shown in Table 2. For the 10-bit and 12-bit datasets, the SDR data is generated by the ISPs of the corresponding imaging systems, respectively. For the 24-bit dataset, the SDR data is generated by the GEO GW5300 ISP, which is an advanced camera video processor system-on-chip designed for high-resolution sensor automotive applications. Its on-chip fusion algorithm combines up to four differently exposed images, achieving outstanding imaging quality. From the table, we can see that the performance of 10-bit RAW sensor data is close

Table 2. Impact of the dynamic range for object detection.

Date Type	10-bit dataset			12-bit dataset			24-bit dataset			Params
	AP	AP50	AP75	AP	AP50	AP75	AP	AP50	AP75	
SDR	43.8	65.6	47.4	67.3	93.6	78.4	52.1	74.6	56.8	0.90M
RAW	43.3	64.3	47.3	65.3	92.9	75.8	34.6	54.7	35.4	
SDR	48.1	69.4	53.2	70.9	94.9	84.0	63.3	88.4	69.6	2.27M
RAW	47.8	69.0	51.6	68.4	93.9	81.5	43.9	66.8	46.1	
SDR	51.8	73.2	56.5	72.8	95.5	86.2	69.7	91.3	76.7	8.92M
RAW	51.2	72.6	56.1	70.5	94.7	84.2	47.5	67.1	52.7	

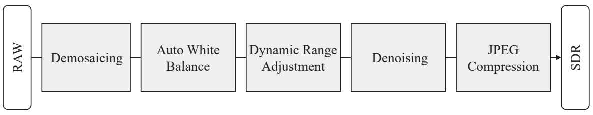


Figure 3. Key components of the software ISP pipeline.

Table 3. Ablation of the software ISP pipeline on 24-bit dataset.

Data Type	AP	AP50	AP75
RAW	32.3	53.7	32.9
RAW (DM+AWB)	34.6	54.7	35.4
RAW (DM+AWB+DRA)	52.1	74.4	56.9
RAW (ISP Pipeline)	53.3	76.8	58.6
RAW (DRA)	51.7	76.1	56.2
RAW (ISP Pipeline w/o DRA)	35.2	56.9	35.7

to the corresponding SDR data, but the performance of 12-bit RAW sensor data is lower than the corresponding SDR data. When the dynamic range increases to 24-bit, the performance of HDR RAW sensor data degrades significantly.

The experiment results demonstrate that DNNs-based object detection algorithms cannot handle the HDR data, and the performance degradation gets worse when dynamic range increases. Results also show that ISP system is important for DNNs-based object detection.

4.2. Ablation Study of ISP System

From the results of Table 2, we can see that the ISP system is beneficial to object detection, which transforms the HDR RAW sensor data into 8-bit SDR data. To investigate the key component of the ISP system for the DNNs-based detector, we perform the ablation of the ISP system on the proposed ROD dataset (24-bit).

We investigate the key components of the aforementioned ISP system, then simplify them to a multi-stage software ISP as shown in Figure 3. Definitions and descriptions of each state are as follows: 1) demosaicing (DM) is implemented by a convolution operation; 2) auto white balance (AWB) is a simple gray world algorithm; 3) dynamic range adjustment (DRA) is applied as a gamma-correction function; 4) denoising is a bilateral filter; 5) JEPG compression follows a standard JPEG algorithm. Results of ablation experiments are shown in Table 3. The experiment is conducted on YOLOX [8] with 0.90M parameters. We can see that all stages are useful, and dynamic range adjustment shows the most important impact. Especially, the performance of the RAW sensor data with only dynamic range adjustment is close to the full ISP.

In summary, we investigate the impact of dynamic range and perform the ablation of the ISP system. We find that the dynamic range adjustment is inevitable to object detection on the HDR RAW sensor data, since the higher the dynamic range, the more difficult it is to extract information by DNNs. Hence, we propose a method to adjust the dynamic range of RAW sensor data for object detection.

5. HDR RAW Detection Pipeline

5.1. Overview

The overall framework of our proposed method is shown in Figure 4. Our method is jointly optimized with the downstream detection network in an end-to-end scheme, which is trained together with the detector from scratch only using detection loss functions. Our method applies learnable transformation functions to effectively exploit the HDR information. Specifically, Our method respectively explores the image-level and pixel-level information to adjust the dynamic range of input data for object detection, which is light-weight and computationally efficient.

Given RAW sensor data X, we first down-sample it as X_{lr}, then feed it to the image-level adjustment module and the pixel-level adjustment module to learn the transformation functions for adjusting the image, denoted as

$$Y_I = g(X, \mathcal{F}_I(X_{lr}; \vartheta_I)). \tag{1}$$

$$Y_P = f(X, \mathcal{F}_P(X_{lr}; \vartheta_P)). \tag{2}$$

Here, Y_I and Y_P are outputs from the image-level and pixel-wise adjustment module, $g(X, \cdot)$ and $f(X, \cdot)$ stand for the image-level function and the pixel-level function, respectively. $\mathcal{F}_I(\cdot; \vartheta_I)$ and $\mathcal{F}_P(\cdot; \vartheta_P)$ stand for the image-level adjustment module and the pixel-level adjustment module. Then, the outputs of two functions are fused to generate the processed result for the downstream detector, denoted as

$$Y = \mathcal{F}_c((Y_I + Y_P)/2), \tag{3}$$

where $\mathcal{F}_c$ stands for the fusion convolution layers. Finally, the processed result is fed to the downstream detector for

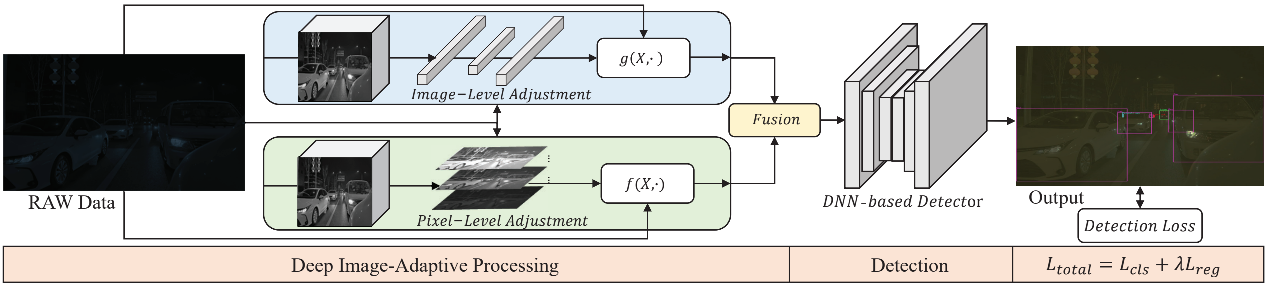


Figure 4. The pipeline of our proposed method. We feed the input RAW sensor data into the image-level adjustment module and the pixel-level adjustment module, respectively. The processed results of two modules are fused to generate the output image of our method, which is fed to the downstream detector. Our method is optimized with the detector in a end-to-end scheme using detection loss functions.

localization and classification. The information on local features and global characteristics guarantees the effectiveness of our method. And the operations performed in low-resolution space contributes to the overall efficiency.

5.2. Image-Level Adjustment

The pixels of the HDR RAW sensor data are mostly distributed in the low-value area, resulting in the information of the RAW data being hard to extract by DNNs. To adaptively adjust the distribution of pixels while preserving natural information at the same time, the image-level adjustment module explores the global information of the input image to regulate the RAW sensor data.

We first process the low-resolution image X_{lr} with a stack of standard strided convolutional layers to generate the features F^{l,n_g}, where n_g stands for the n-th convolution layers of the image-level adjustment module. The extracted features are fed to fully-connected layers for generating the hyperparameters γ of the image-level transformation function. To handle the intense dynamic adjustment, we define the transformation function as

$$Y_I = g(X, \mathcal{F}_I(X_{lr}; \vartheta_I)) = X^{\gamma}. \tag{4}$$

The output of the image-level adjustment module is fed to the downstream detection network. For each input data, the function can explore the global-wise feature to adaptively enhance the HDR information for effective detection.

5.3. Pixel-Level Adjustment

Standard tone modifications, such as exposure change and color curve adjustment, are commonly subtle and implemented by monotonic and image-level linear functions using global-wise information. In our work, we propose the pixel-level adjustment module, which utilizes the pixel-wise transformation function to explore the local-wise information for adjusting the RAW sensor data.

We process the low-resolution image X_{lr} with a stack of standard strided convolutional layers to generate the features F^{l,n_l}, where n_l stands for the n-th convolution layers of the pixel-level adjustment module. The extracted features F^{l,n_l} are used for generating the pixel-wise masks of the pixel-wise transformation function. The pixel-wise masks can be denoted as $m_i, i = 1, ..., K-1$, where K is the number of pieces. The pixel-wise transformation function is formulated as

$$Y_P = f(X, \mathcal{F}_P(X_{lr}; \vartheta_P)) = \sum\nolimits_{k=0}^{K-1} m_k \delta_k(X). \tag{5}$$

$$\delta_k(X) = \begin{cases} 0, & X \in \left[0, \frac{k}{K}\right) \\ X_{-}\frac{k}{K}, & X \in \left[\frac{k}{K}, \frac{k+1}{K}\right) \\ \frac{k}{K}. & X \in \left[\frac{k+1}{K}, 1\right] \end{cases} \tag{6}$$

Here $\delta_k(X)$ divides the input RAW sensor data into a stack of the piece based on the intensity of pixels. For each input data, the function can explore the local-wise feature to boost the texture information for detection.

5.4. Loss Function

Our proposed method is jointly optimized with the object detector in an end-to-end scheme without constraints on the processed image representation. In our experiments, we use YOLOX [8] and Sparse R-CNN [32] as the downstream detector. The loss is composed of classification and regression loss, which can be calculated as

$$L_{total} = L_{cls} + \lambda L_{reg}, \tag{7}$$

where λ is a balancing coefficient. Specifically, we use BCE Loss as the classification loss and intersection-over-union (IoU) loss as the regression loss.

6. Experiments

6.1. Dataset

Our experiments are conducted on the proposed ROD dataset. To eliminate the potential influence of domain shift, we divide the ROD dataset into two subsets based on the scenario. There are about 10k images for the day scenario and about 14k images for the night scenario. We randomly select the 9k annotated images from the day scenario and 13k annotated images of the night scenario as training datasets, respectively. The remaining data are used as corresponding test datasets, respectively.

Table 4. Quantitative comparison with YOLOX (0.90M) on the day and night scenarios of the ROD dataset in terms of AP, AR, AP50, and AP75. The best results are highlighted with bold fonts.

Method	Day				Night				Params (M)	Flops (G)
	AP	AR	AP50	AP75	AP	AR	AP50	AP75		
SDR	52.1	57.2	74.6	56.8	50.3	59.7	80.0	53.7	-	-
RAW	34.6	40.6	54.7	35.4	1.7	5.1	4.5	0.9	-	-
Gamma [12]	52.1	57.3	74.4	56.9	50.8	60.2	80.8	55.0	-	-
Mu-Log [2]	51.5	56.5	74.0	55.6	49.8	57.7	79.3	52.3	-	-
IA-Gamma [12]	53.5	59.7	78.2	57.1	51.8	59.8	81.4	55.8	0.02	0.97
IA-Mu-Log [2]	23.0	31.3	41.9	21.9	50.2	58.9	80.1	54.4	0.02	0.97
GTM [16]	45.1	51.4	68.4	47.6	1.7	4.2	4.1	1.1	0.02	0.97
GTM-DI [16]	45.7	51.9	69.8	48.6	4.8	11.0	10.5	3.7	0.02	0.97
MW-ISPNet [17]	43.4	50.4	67.2	45.8	33.6	44.6	59.1	33.3	9.14	1690.54
Lite-ISPNet [17]	46.8	52.4	68.9	47.3	37.5	47.1	61.9	36.6	5.94	2860.12
IA-ISPNet [23]	54.6	61.2	81.9	59.3	52.7	60.9	81.9	56.8	0.26	0.91
Ours	**58.7**	**63.9**	**85.3**	**61.3**	**54.2**	**61.7**	**83.0**	**58.2**	0.08	0.64

6.2. Experimental Setup and Implementation

For the real-time requirement of automatic driving scenarios, we adopt a YOLOX [8] model as the downstream detection network, which has only 0.90 (M) parameters and 2.27 (G) FLOPs. For training, we employ data augmentation strategies including random horizontal, flip, scale jitter of resizing, and Mosaic. During training and testing, we resize RAW sensor data to a size of 1280×1280. As for evaluation metrics, we adopt the average precision and the average recall over all IOU thresholds (AP and AR), AP at IOU thresholds 0.5 (AP50) and 0.75 (AP75).

We train all models for a total of 300 epochs with 5 epochs warmup on two subsets. We use stochastic gradient descent (SGD) for training. We use a learning rate of linear scaling [11] and the cosine learning rate schedule. The weight decay is 5×10^{-4} and the SGD momentum is 0.9. Our training follows the mini-batch strategy and the batch size is 32. During adjustment, we set the size of X_{lr} is 256×256, $n_g = 3$, and $n_l = 2$ for two modules. Our method is implemented by MindSpore [1].

6.3. Experimental Results

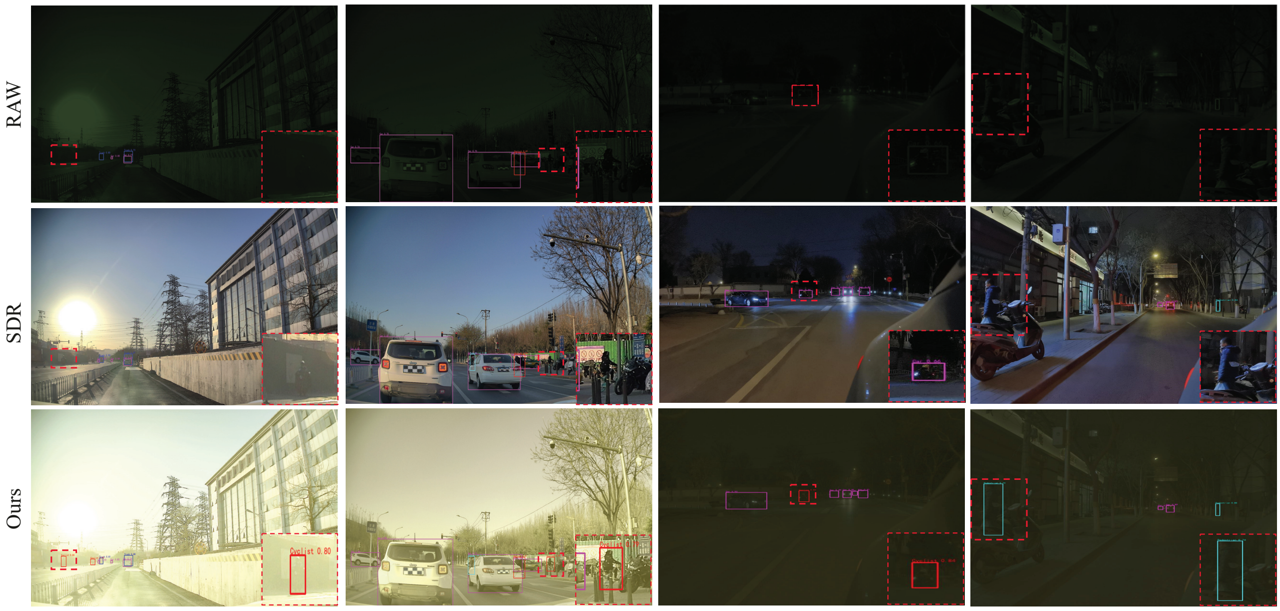
In order to verify the superiority of our method, we compare the proposed methods with typical state-of-the-art methods on the RAW object detection task, including traditional dynamic range compression algorithms (the Gamma correction algorithm (Gamma) [12] and the Mu-log correction algorithm (Mu-log) [2]), tone mapping methods (the image-adaptive Global-wise tone mapping (GTM) [16]), DNNs-based ISP methods (MW-ISPNet [17], Lite-ISPNet [41] and IA-ISPNet [23]). In addition, we modify the traditional dynamic range compression method to the image-adaptive Gamma correction algorithm (IA-Gamma) and the Mu-log correction algorithm (IA-Mu-Log). The image-adaptive strategy follows IA-ISPNet [23] for a fair comparison with our method. We also modify the GTM algorithm to dynamically learn the interval of piecewise linear functions for better dynamic range adjustment, named as GTM-DI. The SDR data is generated by the GEO GW5300 ISP. For a fair comparison, we train all methods with the downstream object detection network using detection loss functions in an end-to-end scheme.

Quantitative Evaluation The quantitative comparison results on the day and night scenarios are shown in Table 4. We can see that the DNNs-based detector is ineffective on night scenario RAW sensor data, which shows that the information of HDR RAW sensor data is difficult to be extracted by DNNs. The performance of our method surpasses SDR data with improvements of 6.6% and 3.9% on the day and night scenarios, respectively. The comparison results demonstrate that the detection on RAW sensor data is significantly superior to the detection on SDR data. In addition, our method effectively boosts the performance of the DNNs-based detector on RAW sensor data with only 0.08 (M) parameters and 0.64 (G) FLOPs. Results from MW-ISPNet and Lite-ISPNet show that simply increasing the model capacity does not necessarily lead to performance gains, which in turn shows the superiority of our method.

Qualitative Evaluation In Figure 5, we show qualitative results of original RAW data, SDR data, and our method by visualizing detection results with confidence scores over 0.4 in the day and night scenarios of the ROD dataset. For the day scenario, we can see that detection on the HDR RAW data with our method can effectively deal with the strong glare of sunlight and the severe lighting variance, but detection on the SDR data fails in these challenging cases. For the night scenario, we can see that detection on the HDR RAW data with our method can effectively handle the low-light condition and accurately recognize objects. In summary, the results of the qualitative evaluation demonstrate that detection on the HDR RAW data can handle a variety of lighting conditions to make safety-critical decisions.

6.4. Ablation Studies

Model Generalization To better validate the effectiveness of object detection on the RAW sensor data, we per-



(a) Detection on the day scenario

(b) Detection on the night scenario

Figure 5. Visual examples of object detection. (a) and (b) are detection results on the day and night scenarios of the ROD dataset, respectively. From top to bottom are the results of RAW data, SDR data, and our method, respectively. Our method significantly outperforms the SDR data. Please zoom in for confidence scores and class predictions. More visual results are in the supplementary document.

Table 5. Quantitative comparison with Sparse R-CNN (104.54M) on the day scenario of the ROD dataset.

Method	AP	AR	AP50	AP75
SDR	73.5	80.8	91.8	84.0
RAW	66.3	73.6	88.1	78.9
Gamma [12]	73.7	82.0	92.2	83.1
Mu-Log [2]	72.7	81.4	91.0	84.0
IA-Gamma [12]	75.1	82.4	92.4	85.7
IA-Mu-Log [2]	74.2	80.2	91.2	84.6
GTM [16]	71.4	78.5	89.4	82.2
GTM-DI [16]	72.6	79.2	89.6	82.5
MW-ISPNet [17]	71.6	77.8	91.4	84.2
MW-ISPNet [17]	72.7	79.2	91.9	85.2
IA-ISPNet [23]	75.6	81.2	91.6	85.1
Ours	**77.4**	**83.6**	**93.2**	**87.3**

Table 6. Quantitative comparison with YOLOX (8.92M) on the day scenario of the ROD dataset.

Method	AP	AR	AP50	AP75
SDR	69.3	72.4	91.3	76.7
RAW	47.5	52.2	67.1	52.7
Gamma [12]	71.2	74.7	94.2	82.4
Mu-Log [2]	69.1	72.8	93.9	78.1
IA-Gamma [12]	72.4	75.6	94.4	82.3
IA-Mu-Log [2]	42.7	64.6	46.9	48.0
GTM [16]	66.0	70.3	88.9	76
GTM-DI [16]	66.4	71.9	90.3	72.7
MW-ISPNet [17]	51.3	66.4	83.3	71.2
Lite-ISPNet [17]	54.6	68.8	85.3	77.2
IA-ISPNet [23]	73.1	76.7	94.5	83.1
Ours	**75.5**	**78.6**	**94.9**	**83.9**

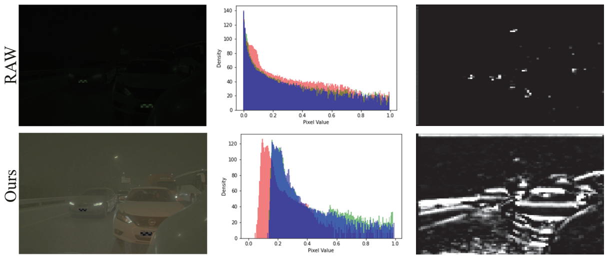
form a comparison experiment with the proposal-based detector on the day scenario of the ROD dataset. We employ Sparse R-CNN as the downstream detector and the results of experiments are shown in Tabel 5. From the table, we can see that our method significantly outperforms the SDR data on the proposal-based detector, which outperforms the SDR data and IA-ISPNet by 3.9% and 1.8%, respectively. The results demonstrate that detection on the RAW sensor data is effective and outperforms the SDR data both with the proposal-free and the proposal-based detector.

Model Size To evaluate the impact of the model size for detection on the RAW sensor data, we employ the YOLOX with different parameters on the day scenario of the ROD dataset. We increase the number of channels of all detector convolutions by a factor of 4, increasing the number of model parameters to 8.92(M). The experiment results are shown in Table 6. We can see that the proposed model outperforms the SDR data by 6.2%. The results demonstrate that detection on the RAW sensor data outperforms the SDR data with a large model size.

Dynamic Range We perform a comparison experiment using YOLOX with different parameters on different dynamic range datasets. The results of the experiments are shown in Table 7. From the table, we can see that our method significantly boosts the performance of RAW sensor data, which outperforms the SDR data on different dynamic range datasets. Specifically, the proposed method outperforms the SDR data by 1.0 percent and 1.4 percent on

Table 7. Quantitative comparison with YOLOX (0.90M) on different dynamic range datasets.

Method	10-bit Dataset		12-bit Dataset		24-bit dataset		Params
	AP	AP50	AP	AP50	AP	AP50	
SDR	43.8	65.6	67.3	93.6	52.1	74.6	0.90(M)
RAW	43.3	64.3	65.3	92.9	34.6	54.7	
Ours	44.8	66.2	68.7	94.2	58.7	85.3	
SDR	48.1	69.4	70.9	94.9	63.3	88.4	2.27(M)
RAW	47.8	69.0	68.4	93.9	43.9	66.8	
Ours	50.6	71.2	70.2	94.9	67.8	92.2	
SDR	51.8	73.2	72.8	95.5	69.7	91.3	8.92(M)
RAW	51.2	72.6	70.5	94.7	47.5	67.1	
Ours	54.1	74.9	72.5	95.2	75.5	94.9	



(a) Data (b) Distribution (c) Feature visualization

Figure 6. Pixel distribution and feature visualization of RAW sensor data and the processed result of our method.

the 10-bit dataset and 12-bit dataset with YOLOX (0.90M), respectively. The results demonstrate that our method can effectively improve the performance of object detection on the RAW sensor data with different dynamic ranges.

6.5. Analysis

We experimentally find that directly applying the RAW sensor data to the DNNs-based object detection methods results in a significant performance drop in different scenarios and it gets even worse when the dynamic range increases. We hypothesize the reason is that RAW sensor data raises the difficulty of feature extraction in DNNs. Since the pixels of RAW sensor data are distributed in the low-value area resulting in a lack of texture information, making it difficult for DNNS to recognize and understanding [9]. As shown in the top row of Figure 6, there is very little information of the features from RAW sensor data. What's more, considering the case of imaging a strong glare in an extremely dark scene, which means several close-to-one values inside a nearly zero-value background. If directly processing these raw data, those large values will dominate the gradient descent process, and spread out when it goes deeper. Whereas, our method balances well between those ones and zeros, which are both meaningful for downstream detectors.

To analyze the impact of the texture information of RAW sensor data on the performance of DNNs-based detection methods, we employ the entropy of the gray-level co-occurrence matrix (GLCM) [14] as the metric to evaluate the necessity of dynamic range adjustment methods. We employ the YOLOX (0.90M) as the detection network for analysis. As shown in Table 8 we can see that the dynamic range adjustment method is effective to boost texture information, and the performance of detection is positively associated with the entropy of GLCM. And Figure 6 shows that our proposed method significantly boosts the information of the features extracted by the DNN. Experiment results demonstrate that the pixel distribution and texture information of RAW sensor data is important factors for detection.

Table 8. Impact of texture information on the performance of detection with YOLOX on the day scenario of the ROD dataset.

Method	Skew	Entropy of GLCM	AP
RAW	8.1742	11.1691	34.6
GTM-DI [16]	2.3311	20.9431	45.1
Gamma [12]	0.8873	24.0634	52.1
IA-Gamma [12]	0.6098	24.1645	53.5
Ours	0.9719	24.5954	58.7

7. Conclusion and Discussion

In this paper, we propose to achieve end-to-end object detection on RAW sensor data, which naturally stores the HDR information without extra equipment cost. For DNN-based detection methods to extract and explore the HDR information of RAW sensor data, we build a novel RAW sensor dataset, named ROD, which consists of 25k annotated RAW sensor data in a 24-bit dynamic range in day and night driving scenarios. Based on the ROD dataset, we investigate the impact of dynamic range on object detection and propose a method for effective detection on RAW sensor data. Specifically, we devise an image-adaptive network to regulate RAW sensor data with learnable transformation functions, which adjusts the dynamic range by image and pixel-level information. Extensive experiments on the ROD dataset demonstrate that the performance of detection on RAW sensor data is significantly superior to detection on SDR data in different situations.

Despite object detection on RAW sensor data with our proposed method can effectively handle a variety of light conditions and significantly outperforms SDR data, it still needs extra processing before DNNs-based detectors. It is worth noting that, for a large proposal-based detector, the performance drop caused by RAW sensor data is smaller than with an efficient proposal-free detector. Hence, we believe that well-designed detector networks can directly handle the HDR RAW sensor data without any processing, which can further exploit the information of RAW sensor data and improve the efficiency and effectiveness of detection. This is considered as our future work.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grants 62131003 and 62021001. We gratefully acknowledge the support of MindSpore used for this research.

References

[1] https://www.mindspore.cn. 6

[2] Helder Araujo and Jorge M Dias. An introduction to the log-polar mapping [image sampling]. In *Proceedings II Workshop on Cybernetic Vision*, pages 139–144. IEEE, 1996. 6, 7

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 2

[4] Zehao Chen, Qian Zheng, Peisong Niu, Huajin Tang, and Gang Pan. Indoor lighting estimation using an event camera. In *CVPR*, 2021. 2

[5] Zhen Cheng, Tao Wang, Yong Li, Fenglong Song, Chang Chen, and Zhiwei Xiong. Towards real-world hdrtv reconstruction: A data synthesis-based approach. In *ECCV*, 2022. 2

[6] Paul E Debevec and Jitendra Malik. Recovering high dynamic range radiance maps from photographs. In *ACM SIGGRAPH*. 2008. 2

[7] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010. 2

[8] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021. 2, 3, 4, 5, 6

[9] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *ICLR*, 2018. 8

[10] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014. 2

[11] Priya Goyal, Piotr Dollar, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017. 6

[12] Hongwei Guo, Haitao He, and Mingyi Chen. Gamma correction for digital fringe projection profilometry. *Applied optics*, 43(14):2906–2914, 2004. 6, 7, 8

[13] Jin Han, Chu Zhou, Peiqi Duan, Yehui Tang, Chang Xu, Chao Xu, Tiejun Huang, and Boxin Shi. Neuromorphic camera guided high dynamic range imaging. In *CVPR*, 2020. 1, 2

[14] Robert M Haralick, Karthikeyan Shanmugam, and Its' Hak Dinstein. Textural features for image classification. *IEEE Transactions on Systems, Man, and Cybernetics*, (6):610–621, 1973. 8

[15] Jun Hu, Orazio Gallo, Kari Pulli, and Xiaobai Sun. Hdr deghosting: How to deal with saturation? In *CVPR*, 2013. 2

[16] Yuanming Hu, Hao He, Chenxi Xu, Baoyuan Wang, and Stephen Lin. Exposure: A white-box photo post-processing framework. *ACM Transactions on Graphics*, 37(2):1–17, 2018. 6, 7, 8

[17] Andrey Ignatov, Radu Timofte, Zhilu Zhang, Ming Liu, Haolin Wang, Wangmeng Zuo, Jiawei Zhang, Ruimao Zhang, Zhanglin Peng, Sijie Ren, et al. Aim 2020 challenge on learned image signal processing pipeline. In *ECCVW*, 2020. 6, 7

[18] Nima Khademi Kalantari, Ravi Ramamoorthi, et al. Deep high dynamic range imaging of dynamic scenes. *ACM Transactions on Graphics*, 36(4):144–1, 2017. 2

[19] Hui Li and Xiao-Jun Wu. Densefuse: A fusion approach to infrared and visible images. *IEEE Transactions on Image Processing*, 28(5):2614–2623, 2019. 2

[20] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, 2017. 2

[21] 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. 2

[22] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *ECCV*, 2016. 2

[23] Wenyu Liu, Gaofeng Ren, Runsheng Yu, Shi Guo, Jianke Zhu, and Lei Zhang. Image-adaptive yolo for object detection in adverse weather conditions. *arXiv preprint arXiv:2112.08088*, 2021. 2, 6, 7

[24] Jiayi Ma, Pengwei Liang, Wei Yu, Chen Chen, Xiaojie Guo, Jia Wu, and Junjun Jiang. Infrared and visible image fusion via detail preserving adversarial learning. *Information Fusion*, 54:85–98, 2020. 2

[25] Jiayi Ma, Yong Ma, and Chang Li. Infrared and visible image fusion methods and applications: A survey. *Information Fusion*, 45:153–178, 2019. 1, 2

[26] Kede Ma, Zhengfang Duanmu, Hanwei Zhu, Yuming Fang, and Zhou Wang. Deep guided learning for fast multi-exposure image fusion. *IEEE Transactions on Image Processing*, 29:2808–2819, 2019. 2

[27] A Omid-Zohoor, D Ta, and B Murmann. Pascalraw: raw image database for object detection, 2014. 1, 3

[28] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *CVPR*, 2016. 2

[29] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *CVPR*, 2017. 2

[30] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 2

[31] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *NeurIPS*, 2015. 2

[32] Peize Sun, Rufeng Zhang, Yi Jiang, Tao Kong, Chenfeng Xu, Wei Zhan, Masayoshi Tomizuka, Lei Li, Zehuan Yuan, Changhu Wang, et al. Sparse r-cnn: End-to-end object detection with learnable proposals. In *CVPR*, 2021. 2, 5

[33] Bishan Wang, Jingwei He, Lei Yu, Gui-Song Xia, and Wen Yang. Event enhanced high-quality image recovery. In *ECCV*, 2020. 2

[34] Lin Wang and Kuk-Jin Yoon. Deep learning for hdr imaging: State-of-the-art and future trends. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 1, 2

[35] Tao Wang, Yong Li, Jingyang Peng, Yipeng Ma, Xian Wang, Fenglong Song, and Youliang Yan. Real-time image enhancer via learnable spatial-aware 3d lookup tables. In *ICCV*, 2021. 2

[36] Xiongwei Wu, Doyen Sahoo, and Steven CH Hoi. Recent advances in deep learning for object detection. *Neurocomputing*, 396:39–64, 2020. 2

[37] Han Xu, Jiayi Ma, and Xiao-Ping Zhang. Mef-gan: Multi-exposure image fusion via generative adversarial networks. *IEEE Transactions on Image Processing*, 29:7203–7216, 2020. 2

[38] Qingsen Yan, Dong Gong, Javen Qinfeng Shi, Anton van den Hengel, Chunhua Shen, Ian Reid, and Yanning Zhang. Dual-attention-guided network for ghost-free high dynamic range imaging. *International Journal of Computer Vision*, 130(1):76–94, 2022. 2

[39] Hong Yang, Wei Kaixuan, Chen Linwei, and Fu Ying. Crafting object detection in very low light. In *BMVC*, 2021. 3

[40] Mingde Yao, Dongliang He, Xin Li, Zhihong Pan, and Zhiwei Xiong. Bidirectional translation between uhd-hdr and hd-sdr videos. *IEEE Transactions on Multimedia*, 2023. 2

[41] Zhilu Zhang, Haolin Wang, Ming Liu, Ruohao Wang, Jiawei Zhang, and Wangmeng Zuo. Learning raw-to-srgb mappings with inaccurately aligned supervision. In *ICCV*, 2021. 2, 6

[42] Zhong-Qiu Zhao, Peng Zheng, Shou-tao Xu, and Xindong Wu. Object detection with deep learning: A review. *IEEE Transactions on Neural Networks and Learning Dystems*, 30(11):3212–3232, 2019. 2

[43] Wei Zhou, Xiangyu Zhang, Hongyu Wang, Shenghua Gao, and Xin Lou. Raw bayer pattern image synthesis for computer vision-oriented image signal processing pipeline design. *arXiv e-prints*, pages arXiv–2110, 2021. 2