SimROD: A Simple Baseline for Raw Object Detection with Global and Local Enhancements

Haiyang Xie[1,2], **Xi Shen**[2], **Shihua Huang**[2], **Qirui Wang**[1,2], **Zheng Wang**[1*]

[1]National Engineering Research Center for Multimedia Software, School of Computer Science, Wuhan University
[2]Intellindust AI Lab
{whuocean,wangzwhu}@whu.edu.cn, {shenxiluc,shihuahuang95}@gmail.com, wangqirui1937@163.com

Abstract

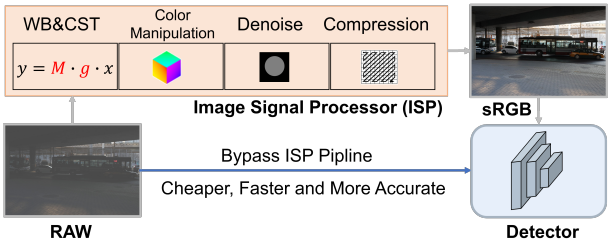
Most visual models are designed for sRGB images, yet RAW data offers significant advantages for object detection by preserving sensor information before ISP processing. This enables improved detection accuracy and more efficient hardware designs by bypassing the ISP. However, RAW object detection is challenging due to limited training data, unbalanced pixel distributions, and sensor noise. To address this, we propose SimROD, a lightweight and effective approach for RAW object detection. We introduce a Global Gamma Enhancement (GGE) module, which applies a learnable global gamma transformation with only four parameters, improving feature representation while keeping the model efficient. Additionally, we leverage the green channel's richer signal to enhance local details, aligning with the human eye's sensitivity and Bayer filter design. Extensive experiments on multiple RAW object detection datasets and detectors demonstrate that SimROD outperforms state-of-the-art methods like RAW-Adapter and DIAP while maintaining efficiency. Our work highlights the potential of RAW data for real-world object detection.

Project page — https://ocean146.github.io/SimROD2025/
Extended version — https://arxiv.org/abs/2503.07101

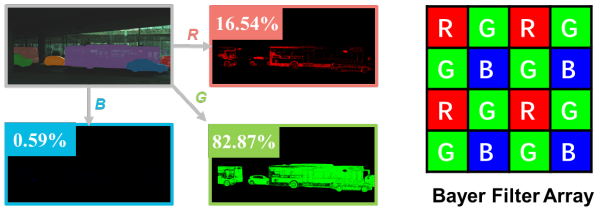
Introduction

Accurate object detection is crucial for autonomous driving, especially under challenging lighting and weather conditions. Traditional methods that rely on sRGB images often lose important details during processing. In contrast, RAW sensor data captures the unprocessed, richer signal from the sensor, preserving more details and a wider dynamic range (Xu et al. 2023; Cui and Harada 2024; Morawski et al. 2022; Chen, Tai, and Ma 2024). Moreover, as shown in Figure 1a, by directly using RAW data, there's no need for an ISP module, which can reduce system complexity, lower latency, and cut costs—key benefits for lightweight, real-time applications.

However, working with RAW data introduces several challenges, including limited training samples, unbalanced pixel distributions, and sensor noise. Current approaches typically rely on complex frameworks that integrate end-to-end optimization of ISP stages with object detection models. These methods explicitly designs learnable ISP stages to transform RAW data (Xu et al. 2023; Morawski et al. 2022; Mosleh et al. 2020; Yu et al. 2021). While these methods demonstrate promising results, they tend to be computationally expensive and introduce unnecessary design complexities. Additionally, modern cameras emphasize the green channel in their Bayer filter design (Zou, Yan, and Fu 2023; Bayer 1976) as the human eye is highly sensitive to green light in both bright and low-light conditions (Wald 1945). However, most existing methods treat RGB channels equally, overlooking the green channel's unique advantages



(a) Raw object detection enables bypassing the ISP.



(b) Green channels contain more informative signals.

Figure 1: *Top:* Advantages of RAW Data for Object Detection. Using RAW data eliminates the need for an ISP, reducing system complexity, latency, and cost—crucial for lightweight, real-time applications (Figure 1a). *Bottom:* Key Insights in SimROD. The green channel in RAW data carries more detailed information. The percentages indicate the proportion of color pixels with the highest intensity in the RGB channels—higher values mean richer details and lower noise in challenging lighting conditions (Figure 1b).

*Corresponding author

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

in RAW data.

In this work, we present SimROD, a simple yet effective approach that enhances RAW object detection performance while maintaining model simplicity. Our approach is based on two key insights: (1) learning an adapted global transformation might not be complicated but crucial for fine-grained tasks (Xu et al. 2023; Buckler, Jayasuriya, and Sampson 2017), and (2) the superior informativeness of the green channel (Figure 1b) in the RGGB Bayer pattern. By leveraging these insights, we introduce an efficient Global Gamma Enhancement (GGE) with only four learnable parameters, significantly reducing model complexity while achieving comparable performance to more complex methods. We also propose a Green-Guided Local Enhancement (GGLE) module that uses the green channel to refine local image details, further boosting detection accuracy.

Through extensive experiments, we demonstrate that SimROD outperforms existing methods, such as RAW-Adapter (Cui and Harada 2024) and DIAP (Xu et al. 2023), on several standard RAW object detection benchmarks, including ROD (Xu et al. 2023), LOD (Hong et al. 2021), and Pascal-Raw (Omid-Zohoor, Ta, and Murmann 2014). For example, on the benchmark of Pascal-Raw (Omid-Zohoor, Ta, and Murmann 2014), following the setup of RAW-Adapter (Cui and Harada 2024), we achieve consistent performance improvement across different object detectors and different setups of Pascal-Raw (Omid-Zohoor, Ta, and Murmann 2014). Furthermore, we create a strong baseline for DIAP (Xu et al. 2023) by leveraging weights pre-trained on MS COCO (Lin et al. 2014), which raises its performance from 24.0% mAP to 30.7% mAP on the ROD dataset (Xu et al. 2023). SimROD achieves notable improvements even relative to this strong baseline.

To summarize, the main contributions of this work are as follows:

- We introduce SimROD, a simple yet effective approach for RAW object detection that combines global-to-local enhancements.
- Inspired by human visual system sensitivity and camera design, we confirm the informativeness of the green channel and develop a Green-Guided Local Enhancement module to refine local details and improve detection performance.
- Despite its simplicity, SimROD achieves state-of-the-art performance on ROD (Xu et al. 2023), LOD (Hong et al. 2021), and Pascal-Raw (Omid-Zohoor, Ta, and Murmann 2014), surpassing prior methods like RAW-Adapter (Cui and Harada 2024) and DIAP (Xu et al. 2023).

Motivation

The human visual system exhibits a pronounced sensitivity to green light wavelengths in both bright and dim light conditions, as evidenced by (Wald 1945). Due to this characteristic of human vision, cameras prioritize green channels in their Bayer filter design (Bayer 1976; Zou, Yan, and Fu 2023). Motivated by this biological and technical precedent, we explored the effectiveness of green channel for object detection by analyzing channel sensitivity and signal-to-noise ratio (SNR) for individual channels.

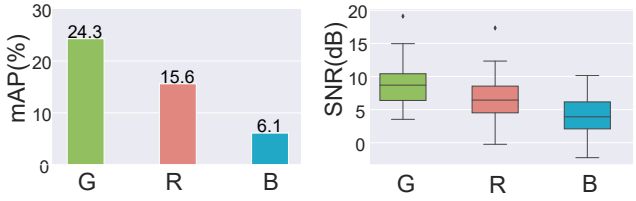


Figure 2: *Left:* We evaluate RAW object detection on the LOD dataset (Hong et al. 2021) using individual color channels—green (G), red (R), and blue (B)—with the state-of-the-art DIAP method (Xu et al. 2023). The results highlight the superior performance of G. *Right:* G has a significantly higher SNR than R and B, suggesting it may be more resistant to noise in extreme lighting conditions, potentially improving robustness.

- *Channel Sensitivity Analysis.* We utilized DIAP (Xu et al. 2023) on the LOD (Hong et al. 2021) dataset to independently assess the detection performance of each channel (green, red, and blue). As depicted in Figure 2 *left*, the green channel achieved the highest detection accuracy, surpassing the red and blue channels by substantial margins (approximately 10 and 20 AP points, respectively), underscoring its superior informativeness for object detection in RAW data.
- *Signal-to-Noise Ratio (SNR) Analysis.* As presented in Figure 2 *right*, the green channel consistently exhibits a higher SNR compared to the red and blue channels, suggesting it is less susceptible to noise, even under challenging lighting conditions. This robustness reinforces the efficacy of leveraging green channel guidance to improve object detection accuracy in extreme environments.

These findings underscore the green channel's potential to improve detection reliability, particularly in complex environments. Inspired by its demonstrated informativeness, we investigate a simple yet effective approach to fully leverage the green channel's strengths, thereby enhancing model performance.

Method

The proposed method's overall framework is illustrated in Figure 3. A RAW image is an unprocessed digital photograph that retains all the data from a camera's sensor, including the RGGB color pattern. Given a RAW image ($X_{\text{RAW}} \in \mathbb{R}^{2H\times 2W}$, we repack and convert it into a four-channel image $X_{\text{packed}} \in \mathbb{R}^{H\times W\times 4}$, where the last dimension represents the color channels in the RGGB pattern.

We first adjust the global pixel distribution of X_{packed} using the proposed *Global Gamma Enhancement* (GGE) module, which learns gamma transformation for each channel, resulting in $X_\gamma \in \mathbb{R}^{H\times W\times 4}$. This X_γ is then passed into the proposed *Green-Guided Local Enhancement* (GGLE) module for local region enhancement, producing an enhanced

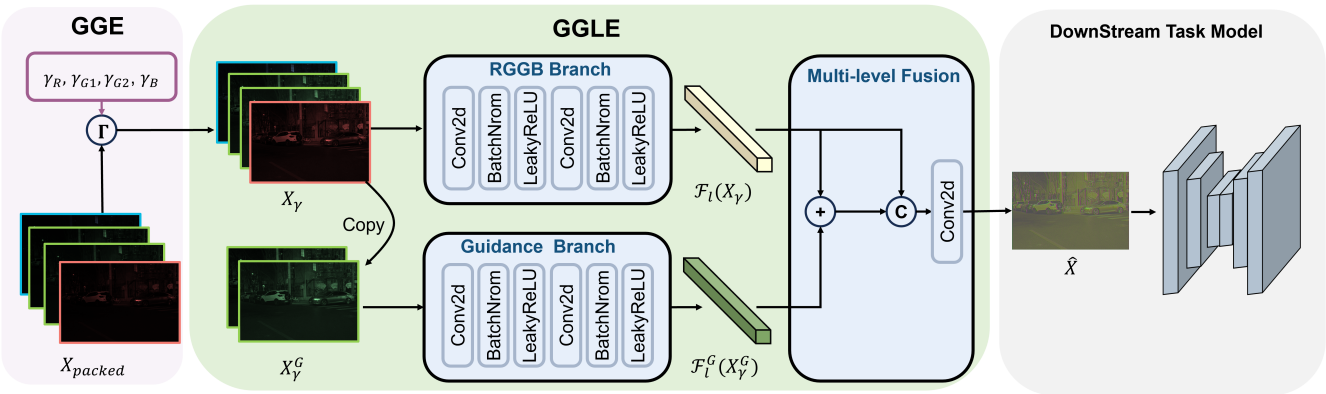


Figure 3: The overview of our proposed SimROD. Our SimROD takes a packed RAW image as input and first learns a global gamma transformation through the Global Gamma Enhancement (GGE) module. The transformed data is then processed by Green-Guided Local Enhancement (GGLE) to enhance local details.

image $\hat{X} \in \mathbb{R}^{H \times W \times 3}$. Finally, $\hat{X}$ is fed into a downstream task model.

This section is organized as follows: in Section , we provide a detailed description of the proposed the *Global Gamma Enhancement* (GGE) module; in Section , we detail the module *Green-Guided Local Enhancement* (GGLE); implementation details are provided in Section .

Global Gamma Enhancement

In visual perception tasks involving raw sensor data, pixel values are typically concentrated in the low range, which makes it challenging for deep neural networks to learn effectively and extract useful features (Xu et al. 2023). Therefore, dynamic range adjustment is an essential step in image signal processing pipelines to prepare raw data for object detection (Xu et al. 2023; Buckler, Jayasuriya, and Sampson 2017).

To address this issue, we propose a simple yet effective module named *Global Gamma Enhancement* (GGE). For a four-channel packed raw image $X_{\text{packed}} \in \mathbb{R}^{H \times W \times 4}$, with pixel values normalized to the range $[0, 1]$, we assign a learnable gamma parameter for each channel. For the *i*-th channel, the gamma transformation is defined as:

$$X_\gamma^i = \Gamma(X_{\text{packed}}^i; \gamma_i), \quad i \in R, G_1, G_2, B \tag{1}$$

where γ_i is simply a learnable parameter. Each channel is adjusted through a gamma transformation Γ scaled to the range $[0, 255]$, calculated as:

$$\Gamma(X_{\text{packed}}^i; \gamma_i) = 255 \cdot \left(X_{\text{packed}}^i\right)^{\gamma_i} \tag{2}$$

Discussion In contrast to the recent method introduced by (Xu et al. 2023), which predicts parameters for image-level RAW data adjustment, our proposed GGE contains only *four* parameters, thus the network only consists of a minimal number of parameters. This results in greater computational efficiency while achieving comparable or superior performance to (Xu et al. 2023) (Section). Notably, we observed that the gamma parameters predicted by the image-level adjustment (Xu et al. 2023) module remained largely unchanged, even when a completely random noise image was used as input.

Green-Guided Local Enhancement

The *Green-Guided Local Enhancement* (GGLE) module is designed to improve feature representations by exploiting the high-frequency details prevalent in the green channels of RAW data, formatted in the RGGB Bayer pattern. Specifically, GGLE processes these green channels independently alongside the full RGGB data, generating an optimized output tailored for downstream tasks such as object detection.

As illustrated in Figure 3, GGLE consists of two primary branches. The first branch, the RGGB branch, processes the complete RGGB data X_γ using a convolutional neural network, $\mathcal{F}_l$, which extracts spatial features from all channels to produce a feature map, $\mathcal{F}_l(X_\gamma)$, representing the full image context. The second branch, the Guidance branch, specifically targets the two green channels, $X_\gamma^{G_1}$ and $X_\gamma^{G_2}$, which are concatenated and processed through another convolutional network, $\mathcal{F}_l^G$, resulting in a green-focused feature map, $\mathcal{F}_l^G(X_\gamma^G)$, where $X_\gamma^G = [X_\gamma^{G_1}, X_\gamma^{G_2}]$.

The final output is generated by a multi-level fusion of $\mathcal{F}_l^G(X_\gamma^G)$ and $\mathcal{F}_l(X_\gamma)$, expressed as:

$$\hat{X} = \text{Conv}(\text{Concat}[\mathcal{F}_l(X_\gamma) + \mathcal{F}_l^G(X_\gamma^G), \mathcal{F}_l(X_\gamma)]) \tag{3}$$

Here, Conv represents a convolution, while Concat denotes feature concatenation. The resulting output, $\hat{X}$, is a three-channel representation that integrates structural details from the green channels across the RGB spectrum, enhancing performance in tasks that require high spatial resolution, such as object detection and segmentation.

Implementation Details

In the *Global Gamma Enhancement* (GGE) module, each γ_i is parameterized in a straightforward manner. For each γ_i, we define a learnable parameter $\alpha_i \in \mathbb{R}$, which is constrained to the range $(-1, 1)$ using the *tanh* activation function. This output is then linearly scaled to lie within a predefined range $(\gamma_{\text{min}}, \gamma_{\text{max}})$, where γ_{min} and γ_{max} are hyperpa-

rameters. Following the settings in (Xu et al. 2023), we set $\gamma_{\max} = 1/7.0$ and $\gamma_{\min} = 1/10.5$.

For the *Green-Guided Local Enhancement* (GGLE) module, both $\mathcal{F}_l$ and $\mathcal{F}_l^G$, used in the RGGB and Guidance branches respectively, employ a simple yet effective architecture consisting of convolutional layers, Batch Normalization (Ioffe and Szegedy 2015), and LeakyReLU (Nair and Hinton 2010) activation functions. Collectively, GGE and GGLE comprise a total of just 0.003 million parameters, rendering these modules lightweight compared to previous approaches. In contrast, prior methods often require hundreds of times more parameters, yet our approach achieves superior performance.

Loss Function Our SimROD is an end-to-end framework, jointly optimizing the GGE and GGLE modules alongside the downstream model, thereby eliminating the need for additional loss functions tailored to these enhancement stages. Furthermore, we adopt the same loss functions as those employed in the original works (Zheng et al. 2021; Lin et al. 2017; Xie et al. 2021). For example, when using YoloX (Zheng et al. 2021) as the detector, optimization relies solely on the standard detection loss, which includes classification and regression components. The total loss function is defined as:

$$\mathcal{L}_{total} = \mathcal{L}_{cls} + \lambda \mathcal{L}_{reg} \tag{4}$$

where λ is the default balancing factor of the detector between the classification loss L_{cls} and the regression loss L_{reg} ($\lambda = 3$ in YoloX (Zheng et al. 2021)). This unified approach allows the enhancement modules to adapt naturally to the detection objectives, supporting end-to-end optimization of the entire framework.

Experiments

Datasets and Evaluation Metrics

We evaluate and compare our method against existing methods on four benchmark datasets: Pascal-Raw (Omid-Zohoor, Ta, and Murmann 2014), LOD (Hong et al. 2021), and ROD (Xu et al. 2023) for object detection, and ADE20K-Raw (Cui and Harada 2024) for semantic segmentation.

Pascal-Raw (Omid-Zohoor, Ta, and Murmann 2014). The Pascal-Raw dataset (Omid-Zohoor, Ta, and Murmann 2014) comprises 4,259 RAW images captured under standard lighting conditions with a Nikon D3200 DSLR camera, covering three object classes: person, car, and bicycle. The dataset is divided into 2,129 images for training and 2,130 images for testing. For our experiments, we utilize the preprocessed RAW data provided by RAW-Adapter (Cui and Harada 2024).

LOD (Hong et al. 2021). The LOD dataset (Hong et al. 2021) contains 2,230 RAW images captured with a Canon EOS 5D Mark IV camera in low-light conditions, covering eight object classes: bus, chair, TV monitor, bicycle, bottle, dining table, motorbike, and car. The dataset is split into 1,800 images for training and 430 images for testing. For all our experiments, we use the preprocessed RAW data provided by RAW-Adapter (Cui and Harada 2024) to ensure a fair comparison with RAW-Adapter.

ROD (Xu et al. 2023). The original ROD dataset (Xu et al. 2023) contains 25,207 RAW images, including 10k daytime scenes and 14k nighttime scenes, across six common object categories. Compared to other datasets, ROD is a larger-scale dataset with a more diverse range of scenes, focusing on urban driving scenarios. Due to limitations in dataset access, we were unable to adhere to the "standard" partitioning protocol defined by ROD (Xu et al. 2023). Despite repeated attempts to obtain the full dataset, only a subset of the training data has been publicly released. This subset consists of 16,089 RAW images, including 4,053 daytime scenes and 12,036 nighttime scenes, but it contains only five object classes instead of the six originally specified in (Xu et al. 2023). To ensure a fair evaluation, we randomly partitioned the publicly available subset into 80% for training and 20% for testing. This partition resulted in 3,245 daytime scenes and 9,626 nighttime scenes in the training set, with the remaining images reserved for testing. We strictly followed the official guidelines for data preprocessing and conducted multiple experimental trials to ensure the validity of the results. All references to the ROD dataset in this paper pertain specifically to this re-partitioned subset. We will make this subset publicly available to facilitate the re-implementation of our results.

ADE20K-Raw (Cui and Harada 2024). The ADE20K-RAW dataset (Cui and Harada 2024) is a RAW-format segmentation dataset derived from ADE20K (Zhou et al. 2017), consisting of 27,574 images synthesized using InvISP (Xing, Qian, and Chen 2021) by RAW-Adapter (Cui and Harada 2024). This dataset includes three versions—normal, dark, and over-exposure—that simulate different lighting conditions. The dataset follows the same training and testing splits defined by ADE20K (Zhou et al. 2017) for consistency.

Evaluation metrics. For object detection, we report standard average precision at an IoU threshold of 0.5 (AP_{50}) and the average precision across IoU thresholds from 0.5 to 0.95 (AP). For semantic segmentation, we use mean Intersection over Union (mIoU), which measures the average overlap between predicted and ground truth masks across all classes.

Training Details

All experiments in this section follow the settings of DIAP (Xu et al. 2023), unless stated otherwise.

For detection tasks, we conduct experiments using two object detectors: YoloX (Zheng et al. 2021) and RetinaNet (Lin et al. 2017), following the protocols of DIAP (Xu et al. 2023) and RAW-Adapter (Cui and Harada 2024). If not otherwise specified, all experiments were initialized with pre-trained weights.

For YoloX (Zheng et al. 2021), we use the official training strategies, including standard data augmentation techniques such as random horizontal flipping, scale jittering through resizing, and Mosaic augmentation (Bochkovskiy, Wang, and Liao 2020). Both training and testing data are resized to 640×640. The model is trained for 300 epochs, with five warmup epochs, using the SGD optimizer with a momentum of 0.9. We apply a cosine learning rate schedule and use a

Method	LOD		Pascal-Raw		ROD		Add.
	AP	AP_{50}	AP	AP_{50}	AP	AP_{50}	Params (M)
Demosacing	26.5	46.0	66.8	92.8	4.8	9.5	0.000
Gamma	25.7	44.2	69.0	94.5	7.6	14.3	0.000
IA	25.1	43.9	68.9	94.2	30.6	53.3	0.176
Raw-or-Cook†	18.0	36.3	61.6	91.2	-	-	0.000
GenISP†	20.5	39.8	60.6	89.5	-	-	0.220
RAW-Adapter†	26.4	45.1	67.5	93.7	N/A	N/A	0.460
DIAP	25.9	43.4	68.7	94.2	30.7	53.4	0.260
Our SimROD	**26.7(+0.8)**	**46.3(+2.9)**	**69.7(+1.0)**	**95.1(+1.1)**	**33.1(+2.4)**	**57.6(+4.2)**	0.003 (1%)

Table 1: Results with YoloX-Tiny (Zheng et al. 2021), following DIAP's benchmark (Xu et al. 2023). Performance metrics (AP and AP_{50}) for YoloX-Tiny across different methods including IA (Xu et al. 2023), Raw-or-cook (Ljungbergh et al. 2023), GenISP (Morawski et al. 2022), RAW-Adapter (Cui and Harada 2024) and DIAP (Xu et al. 2023). The best performance for each dataset is highlighted in bold. The table also includes the number of parameters (in millions). The Pascal-Raw (Omid-Zohoor, Ta, and Murmann 2014) in this table is the normal light version. † indicates reproduced results. N/A means the model did not converge.

batch size of 12. The training process of Pascal-Raw (Omid-Zohoor, Ta, and Murmann 2014) and LOD (Hong et al. 2021) takes 2 hours on three NVIDIA RTX 3090 GPUs. The training process of ROD (Xu et al. 2023) takes 10 hours on three NVIDIA RTX 4090 GPUs. For initialization, we use COCO (Lin et al. 2014) pre-trained weights, which improve the performance from 24.0% AP to 30.7% AP compared to the approach proposed in the original DIAP (Xu et al. 2023) and construct a strong baseline.

For the RetinaNet (Lin et al. 2017) detector, we use the MMDetection framework and its default data augmentation pipeline, which includes random cropping, random flipping, and multi-scale testing (Cui and Harada 2024). We set the learning rate for the proposed SimROD to 3e-3.

For the segmentation task, we followed the settings from RAW-Adapter (Cui and Harada 2024) and used Segformer (Xie et al. 2021) with the MIT-B5 (Xie et al. 2021) backbone. We trained the model on four NVIDIA RTX 4090 GPUs, with learning rates adjusted empirically for the proposed method: 8e-5 for the normal-light dataset of Pascal-Raw (Cui and Harada 2024), 7e-5 for the dark dataset of Pascal-Raw (Cui and Harada 2024), and 9e-5 for the over-exposure dataset (Cui and Harada 2024).

Comparison with State-of-the-Art Methods

RAW Object Detection on Pascal-Raw (Omid-Zohoor, Ta, and Murmann 2014), LOD (Hong et al. 2021), and ROD (Xu et al. 2023) Datasets. Table 1 and Table 2 demonstrate that our proposed SimROD consistently outperforms existing methods across all datasets while being highly efficient. To ensure a fair comparison with RAW-Adapter (Cui and Harada 2024) and DIAP (Xu et al. 2023), we strictly follow their official settings with YoloX-Tiny (Zheng et al. 2021) and RetinaNet-R50 (Lin et al. 2017). For YoloX-Tiny (Table 1), SimROD achieves notable AP gains of +2.4 on ROD, +0.8 on LOD, and +1.0 on Pascal-Raw, surpassing DIAP (Xu et al. 2023) with just 0.003M additional parameters. For RetinaNet-R50 (Table 2), SimROD reaches an impressive 63.4% AP_{50} on LOD, outperforming RAW-Adapter (Cui and Harada 2024) and DIAP (Xu et al. 2023), while requiring only 0.7% of RAW-Adapter's additional parameters. These results highlight SimROD's strong generalization, superior accuracy, and exceptional parameter efficiency, making it a highly effective solution for RAW object detection.

Semantic Segmentation on ADE20K-Raw (Zhou et al. 2017; Cui and Harada 2024) with RAW Data. To further validate SimROD's effectiveness, we evaluate it on semantic segmentation using ADE20K-Raw with Segformer (Xie et al. 2021), following RAW-Adapter (Cui and Harada 2024). As shown in Table 3, SimROD achieves the best performance across normal and low-light conditions, while achieving competitive performance on over-exposure conditions, demonstrating strong generalization. Despite adding only 0.003M parameters, it remains highly efficient while delivering superior accuracy. These results reinforce our object detection findings, proving SimROD's versatility in RAW data processing. Its potential extends beyond detection, making it applicable to broader vision tasks.

Analysis and Discussion

Ablation Study. Table 4 presents an ablation study on the LOD (Hong et al. 2021) and Pascal-Raw (Omid-Zohoor, Ta, and Murmann 2014) datasets using YoloX-Tiny (Zheng et al. 2021) to evaluate the impact of different enhancement modules. The best AP and AP_{50} results are achieved when all four components are used, indicating the combined enhancements' positive effect on model performance. Moreover, the green channel guidance outperforms the red and blue guidance, indicating the unique value of the green channel. Notably, using the red and blue channel guidance even leads to the performance drop on LOD (Hong et al. 2021) which is low-light, noisy, and more challenging. We also evaluate RGGB Table 4. We find that that *solely using the Green Channels (GG) alone outperforms the combination*

Method	LOD		Pascal-Raw		Add.
	AP	AP_{50}	AP	AP_{50}	Params (M)
Demosacing	58.5	58.5	89.2	89.2	0.000
Default ISP	58.4	58.4	89.6	89.6	0.000
Karaimer et Brown	54.4	54.4	89.4	89.4	-
Lite-ISP	-	-	88.5	88.5	-
InvISP	56.9	56.9	87.6	87.6	-
Dirty-Pixel	61.6	61.6	89.7	89.7	4.28
DIAP†	59.1	59.1	89.5	89.5	0.260
RAW-Adapter	62.1	62.1	89.7	89.7	0.46
Our SimROD	**63.4(+1.3)**	**63.4(+1.3)**	**90.1(+0.4)**	**90.1(+0.4)**	0.003 (0.7%)

Table 2: Results with RetinaNet-R50 (Lin et al. 2017), following RAW-Adapter's benchmark (Cui and Harada 2024). Performance metrics (AP and AP_{50}) for RetinaNet-R50 across different methods including Demosaicing (Cui and Harada 2024), Default ISP (Omid-Zohoor, Ta, and Murmann 2014), Karaimer et Brown (Karaimer and Brown 2016), Lite-ISP (Zhang et al. 2021), InvISP (Xing, Qian, and Chen 2021) and Dirty-Pixel (Diamond et al. 2021). The best performance for each dataset is highlighted in bold. The table also includes the number of parameters (in millions). The Pascal-Raw (Omid-Zohoor, Ta, and Murmann 2014) in this table is the normal light version. † indicates results reproduced using the official code. RetinaNet-R50 shows similar AP/AP_{50} due to strong performance and low small-object ratios in LOD/Pascal-Raw (5%/3% vs. 42% in COCO).

Method	Normal	Over-exp	Dark	Add. Params. (M)
Demosacing	47.47	45.69	37.55	0.00
Lite-ISP	43.22	42.01	5.52	0.00
InvISP	47.82	44.30	4.03	0.00
SID	-	-	37.06	0.00
DNF	-	-	35.88	0.00
Dirty-Pixel	47.86	46.50	38.02	4.28
RAW-Adapter	47.95	**46.62**	38.75	0.30
Our SimROD	**48.42**	45.96	**38.85**	0.003

Table 3: Segmentation Performance on ADE20K-Raw (Zhou et al. 2017; Cui and Harada 2024). Comparison of semantic segmentation (mIoU) results under normal, overexposure (*Over-exp*), and low-light (*Dark*) conditions using Segformer (Xie et al. 2021). Bold numbers indicate the best results for each condition.

Pretrained Weight	GGE	GGLE (RGGB)	GGLE (Guidance)	LOD AP_{50}	Pascal-Raw AP_{50}
○	○	○	○	27.7	85.0
●	○	○	○	44.6	93.0
●	●	○	○	45.0	94.3
●	●	●	○	45.1	94.7
●	●	●	R	44.9	94.5
●	●	●	B	44.2	94.2
●	●	●	RB	44.5	94.3
●	●	●	RGGB	46.2	94.5
●	●	●	GG	**46.3**	**95.1**

Table 4: Ablation Study on the Impact of Different Components over Pascal-Raw (Omid-Zohoor, Ta, and Murmann 2014) and LOD (Hong et al. 2021) datasets with YoloX-Tiny (Zheng et al. 2021). The term 'Guided Channel' refers to the use of a specific channel for local enhancement in GGLE, while GGLE without a guided channel indicates a configuration with only the RGGB branch. Bold numbers indicate the best performance.

of all channels (RGGB). This is attributed to the significant noise in R and B of RGGB, which affects the learning of the model. The inferior performance of RB compared to R alone further corroborates this, suggesting that combining all channels as guidance does not yield better results.

GGE v.s. DIAP (Xu et al. 2023) We compares DIAP (Xu et al. 2023) and our GGE across ROD (Xu et al. 2023), Pascal-Raw (Omid-Zohoor, Ta, and Murmann 2014), and LOD (Hong et al. 2021) datasets in Table 5. Our GGE achieves comparable or slightly improved performance in certain metrics while reducing parameters and GFLOPs, reflecting a more efficient model design. Note that GGE only contains four learnable parameters.

Sampling Frequency of the Green Channel. We investigate how the green channel's sampling frequency affects RAW object detection performance (Table **??**). In our experiments, we feed the green channel into DIAP (Xu et al. 2023) and compare two strategies: the default 1× sampling based on the RGGB pattern versus a reduced 0.5× sampling (using only one green value). The results show that lowering the green channel frequency leads to a significant performance drop, especially on the low-light LOD dataset. This confirms that the denser spatial information provided by the green channel—due to its higher frequency in the Bayer pattern—is crucial for robust performance.

The Value of γ in Training . As shown in Fig. 4, the γ values slightly increase during training leading to the result-

Dataset	Metric	DIAP	Our GGE
ROD	AP	30.7	**30.9**
	AP_{50}	53.4	**53.5**
Pascal-Raw	AP	**68.7**	68.4
	AP_{50}	94.2	**94.3**
LOD	AP	**25.9**	25.5
	AP_{50}	43.4	**46.2**
	Params (M)	0.260	0.000
	GFLOPs	1.504	0.000

Table 5: Comparison of DIAP (Xu et al. 2023) and our GGE on ROD (Xu et al. 2023), Pascal-Raw (Omid-Zohoor, Ta, and Murmann 2014), and LOD (Hong et al. 2021) datasets with YoloX-Tiny (Zheng et al. 2021).Results are shown in AP and AP_{50} for each dataset, with additional parameters and GFLOPs also provided.

G Sampling Frequency	LOD		Pascal-Raw	
	mAP	AP_{50}	mAP	AP_{50}
0.5× (reduced)	22.0	40.3	67.5	94.3
1× (default)	24.3	42.0	67.9	94.2

Table 6: Effect of Green Channel Sampling Frequency. Reducing the green channel sampling frequency to 0.5× leads to a significant performance drop, especially on the low-light LOD dataset, highlighting the importance of dense green channel information.

ing pixel values are slightly reduced. This is reasonable, as the initial value of 0.114 is lower than the commonly used default of 1/2.2 in standard gamma correction.

Conclusion

In this work, we presented SimROD, a simple yet effective approach for enhancing object detection performance on RAW data. SimROD introduces a streamlined solution featuring the Global Gamma Enhancement (GGE) with only four learnable parameters, achieving competitive performance while maintaining low model complexity. Furthermore, our exploration revealed that the green channel holds more informative signals, leading to the development of the Green-Guided Local Enhancement (GGLE) module, which enhances local image details effectively. Extensive experiments across multiple RAW object detection datasets and detectors, as well as the RAW segmentation dataset, demonstrate the effectiveness of our SimROD.

Related Work

Object Detection on RAW Data Object detection has been an active area of research in computer vision (Everingham et al. 2010; Lin et al. 2014; Wu, Sahoo, and Hoi 2020; Tian et al. 2025; Ma et al. 2024; Wang et al. 2025b,a). RAW data serves as the input to the image signal processor (ISP) and has attracted significant attention due to its unique

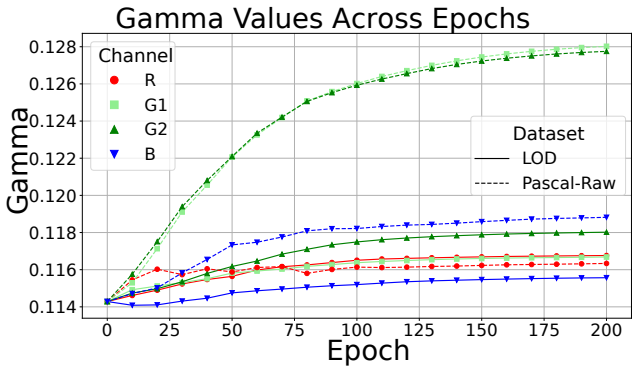


Figure 4: γ across epochs on LOD and Pascal-Raw.

value.(Xu et al. 2023; Karaimer and Brown 2016; Nishimura et al. 2018; Ramanath et al. 2005). Considerable efforts have been made to utilize the RAW data to improve detection performance and robustness. (Hong et al. 2021) designed an auxiliary task for image reconstruction on synthetic datasets to improve detection performance. (Onzon, Mannan, and Heide 2021) proposed estimating camera exposure parameters from previous frames to achieve well-exposed images. (Morawski et al. 2022) designed an ISP pipeline called GenISP, which applies learnable white balance and color correction transformations to RAW data. (Yoshimura et al. 2023a) decomposed control into the entire dataset, along with fine-tuning for individual images. They introduced a latent update style controller to manage the stages of a differentiable ISP. (Xu et al. 2023) developed an image-level adjustment module and a pixel-level adjustment module to learn transformations. (Chen, Tai, and Ma 2024) proposed an activation function to extract features from RAW data effectively. (Cui and Harada 2024) combined an input-level adapter with a model-level adapter to enhance downstream task performance. Additionally, (Yoshimura et al. 2023b) proposed using noise-accounted raw augmentation to enhance recognition performance. Despite these advancements, most of these approaches often rely on simulating expert-tuning ISP stages, resulting in redundant parameters and increased computational complexity, which can adversely affect both accuracy and efficiency.

Image Enhancement based on RAW Data Many works have focused on enhancing images from RAW data. (Brooks et al. 2019) introduced a method to transform sRGB images into synthetic RAW data and trained a neural network model for denoising. (Zou, Yan, and Fu 2023) suggested separating challenging and easier regions and using dual intensity and global spatial guidance to reconstruct images from RAW images. (Guo, Liang, and Zhang 2021) proposed leveraging the green channel as a prior to jointly perform demosaicking and denoising on images, demonstrating the unique value of the green channel. However, these methods depend on paired datasets, which, even when artificially synthesized, often result in suboptimal data that may not be directly applicable to object detection tasks.

Acknowledgements

We thank Caizhi Zhu, Yinqiang Zheng, Xuanlong Yu, Zechao Hu, Hao Li, Zhengwei Yang, Song Ouyang, Jingyu Xu, Likai Tian, and Runqi Wang for their valuable support. This work was funded by the National Natural Science Foundation of China (Grant No. 62571379). The numerical calculations in this paper have been done on the supercomputing system in the Supercomputing Center of Wuhan University.

References

Bayer, B. E. 1976. Color imaging array. U.S. Patent.

Bochkovskiy, A.; Wang, C.-Y.; and Liao, H.-Y. M. 2020. Yolov4: Optimal speed and accuracy of object detection. *arXiv*.

Brooks, T.; Mildenhall, B.; Xue, T.; Chen, J.; Sharlet, D.; and Barron, J. T. 2019. Unprocessing images for learned raw denoising. In *CVPR*.

Buckler, M.; Jayasuriya, S.; and Sampson, A. 2017. Reconfiguring the imaging pipeline for computer vision. In *ICCV*.

Chen, H.; Tai, H.-S.; and Ma, K. 2024. Guiding a Harsh-Environments Robust Detector via RAW Data Characteristic Mining. In *AAAI*.

Cui, Z.; and Harada, T. 2024. RAW-Adapter: Adapting Pre-trained Visual Model to Camera RAW Images. In *ECCV*.

Diamond, S.; Sitzmann, V.; Julca-Aguilar, F.; Boyd, S.; Wetzstein, G.; and Heide, F. 2021. Dirty pixels: Towards end-to-end image processing and perception. *TOG*.

Everingham, M.; Van Gool, L.; Williams, C. K.; Winn, J.; and Zisserman, A. 2010. The pascal visual object classes (voc) challenge. *IJCV*.

Guo, S.; Liang, Z.; and Zhang, L. 2021. Joint Denoising and Demosaicking With Green Channel Prior for Real-World Burst Images. *TIP*.

Hong, Y.; Wei, K.; Chen, L.; and Fu, Y. 2021. Crafting Object Detection in Very Low Light. In *BMVC*.

Ioffe, S.; and Szegedy, C. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv*.

Karaimer, H. C.; and Brown, M. S. 2016. A software platform for manipulating the camera imaging pipeline. In *ECCV*.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017. Focal loss for dense object detection. In *ICCV*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *ECCV*.

Ljungbergh, W.; Johnander, J.; Petersson, C.; and Felsberg, M. 2023. Raw or cooked? object detection on raw images. In *Scandinavian Conference on Image Analysis*.

Ma, C.; Liu, Y.-L.; Wang, Z.; Liu, W.; Liu, X.; and Wang, Z. 2024. HumanNeRF-SE: A Simple yet Effective Approach to Animate HumanNeRF with Diverse Poses. In *CVPR*.

Morawski, I.; Chen, Y.-A.; Lin, Y.-S.; Dangi, S.; He, K.; and Hsu, W. H. 2022. Genisp: Neural isp for low-light machine cognition. In *CVPR*.

Mosleh, A.; Sharma, A.; Onzon, E.; Mannan, F.; Robidoux, N.; and Heide, F. 2020. Hardware-in-the-loop end-to-end optimization of camera image processing pipelines. In *CVPR*.

Nair, V.; and Hinton, G. E. 2010. Rectified linear units improve restricted boltzmann machines. In *ICML*.

Nishimura, J.; Gerasimow, T.; Sushma, R.; Sutic, A.; Wu, C.-T.; and Michael, G. 2018. Automatic ISP Image Quality Tuning Using Nonlinear Optimization. In *ICIP*.

Omid-Zohoor, A.; Ta, D.; and Murmann, B. 2014. PASCALRAW: raw image database for object detection. *Stanford Digital Repository*.

Onzon, E.; Mannan, F.; and Heide, F. 2021. Neural auto-exposure for high-dynamic range object detection. In *CVPR*.

Ramanath, R.; Snyder, W.; Yoo, Y.; and Drew, M. 2005. Color image processing pipeline. *SPS*.

Tian, L.; Zhao, J.; Hu, Z.; Yang, Z.; Li, H.; Jin, L.; Wang, Z.; and Li, X. 2025. CCIN: Compositional Conflict Identification and Neutralization for Composed Image Retrieval. In *CVPR*.

Wald, G. 1945. Human vision and the spectrum. *Science*.

Wang, R.; Ma, C.; Li, G.; Xu, H.; Li, Y.; and Wang, Z. 2025a. You Think, You ACT: The New Task of Arbitrary Text to Motion Generation. In *ICCV*.

Wang, R.; Ma, C.; Zhao, J.; Xu, H.; Sun, D.; Chen, H.; Xiong, L.; Wang, Z.; and Li, X. 2025b. Leader is Guided: Interactive Motion Generation via Lead-Follow Paradigm and Trajectory Guidance. In *MM*.

Wu, X.; Sahoo, D.; and Hoi, S. C. 2020. Recent advances in deep learning for object detection. *Neurocomputing*.

Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J. M.; and Luo, P. 2021. SegFormer: Simple and efficient design for semantic segmentation with transformers. *NeurIPS*.

Xing, Y.; Qian, Z.; and Chen, Q. 2021. Invertible image signal processing. In *CVPR*.

Xu, R.; Chen, C.; Peng, J.; Li, C.; Huang, Y.; Song, F.; Yan, Y.; and Xiong, Z. 2023. Toward raw object detection: A new benchmark and a new model. In *CVPR*.

Yoshimura, M.; Otsuka, J.; Irie, A.; and Ohashi, T. 2023a. Dynamicisp: dynamically controlled image signal processor for image recognition. In *ICCV*.

Yoshimura, M.; Otsuka, J.; Irie, A.; and Ohashi, T. 2023b. Rawgment: Noise-accounted raw augmentation enables recognition in a wide variety of environments. In *CVPR*.

Yu, K.; Li, Z.; Peng, Y.; Loy, C. C.; and Gu, J. 2021. Reconfigisp: Reconfigurable camera image processing pipeline. In *ICCV*.

Zhang, Z.; Wang, H.; Liu, M.; Wang, R.; Zhang, J.; and Zuo, W. 2021. Learning raw-to-srgb mappings with inaccurately aligned supervision. In *ICCV*.

Zheng, G.; Songtao, L.; Feng, W.; Zeming, L.; Jian, S.; et al. 2021. YOLOX: Exceeding YOLO series in 2021. *arXiv*.

Zhou, B.; Zhao, H.; Puig, X.; Fidler, S.; Barriuso, A.; and Torralba, A. 2017. Scene parsing through ade20k dataset. In *CVPR*.

Zou, Y.; Yan, C.; and Fu, Y. 2023. Rawhdr: High dynamic range image reconstruction from a single raw image. In *ICCV*.