End-to-End Low-Light Enhancement for Object Detection with Learned Metadata from RAWs

Xuelin Shen[1]* **Haifeng Jiao**[1,3]* **Yitong Wang**[3] **Yulin He** [1] **Wenhan Yang**[2] †

[1]Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ)
[2]Peng Cheng Laboratory
[3]College of Computer Science and Software Engineering, Shenzhen University
shenxuelin@gml.ac.cn, jiaohaifeng@gml.ac.cn, wangyitong@gml.ac.cn
heyulin@gml.ac.cn, yangwh@pcl.ac.cn

Abstract

Although RAW images offer advantages over sRGB by avoiding ISP-induced distortion and preserving more information in low-light conditions, their widespread use is limited due to high storage costs, transmission burdens, and the need for significant architectural changes for downstream tasks. To address the issues, this paper explores a new raw-based machine vision paradigm, termed Compact RAW Metadata-guided Image Refinement (CRM-IR). In particular, we propose a Machine Vision-oriented Image Refinement (MV-IR) module that refines sRGB images to better suit machine vision preferences, guided by learned raw metadata. In detail, we propose a Cross-Modal Contextual Entropy (CMCE) network for raw metadata extraction and compression. It builds upon the latent representation and entropy modeling framework of learned image compression methods, and uniquely exploits the contextual correspondence between raw images and their sRGB counterparts to achieve more efficient and compact metadata representation. Additionally, we integrate priors derived from the ISP pipeline to simplify the refinement process, enabling a more efficient design. Such a design allows the CRM-IR to focus on extracting the most essential metadata from raw images to support downstream machine vision tasks, while remaining plug-and-play and fully compatible with existing imaging pipelines, without any changes to model architectures or ISP modules. We implement our CRM-IR scheme on various object detection networks, and extensive experiments under low-light conditions demonstrate that it can significantly improve performance with an additional bitrate cost of less than 10^{-3} bits per pixel. Code is available at `https://github.com/haifengjiao001/CRM-IR`.

1 Introduction

Raw images refer to unprocessed and uncompressed data captured directly from a camera's image sensor. Their retained sensor readings preserve linear scene radiance and full bit-depth precision. Raw images typically undergo in-camera Image Signal Processing (ISP) steps, including demosaicing, white balancing, gamma correction and compression, to remove perceptual redundancy and enhance visual appeal, ultimately producing the commonly seen sRGB images. However, as these ISP pipelines are primarily designed to satisfy human perceptual preferences, they often perform suboptimally in machine vision practice. Especially in low-light conditions, ISP pipelines apply nonlinear radiance amplification to enhance details and textures, inevitably amplifying the inherent noise introduced

*These authors contributed equally.
†Correspondence to: Wenhan Yang.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

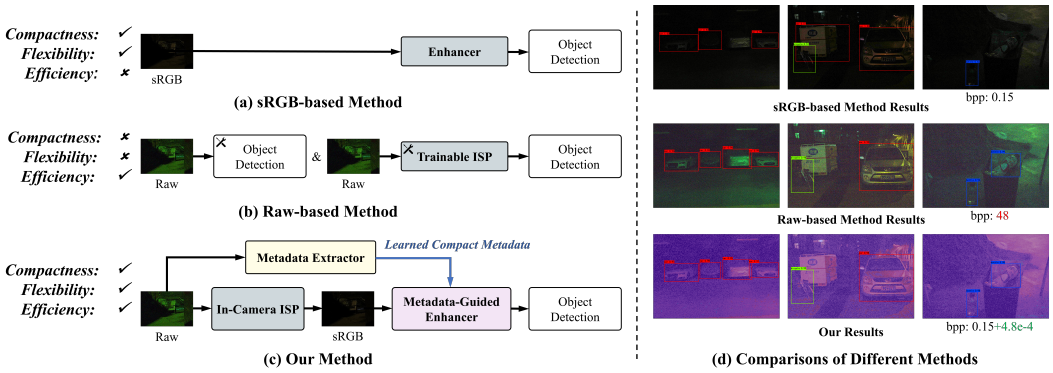

Figure 1: Comparison of existing low-light object detection methodologies: (a) sRGB-based methods, (b) raw-based methods and (c) the proposed CRM-IR. (d) visual results of representative approaches with respect to detection accuracy and transmission overhead, with the first and second rows reproduced from [16] and [17], respectively.

by physical factors during shooting, such as high ISO values, the use of flash and long exposure times [1, 2, 3, 4, 5]. Some studies apply low-light enhancement techniques to improve image quality under laser illumination, but these methods are constrained by the information loss during ISP processing [6, 7]. As a result, they can only adjust distributions and fail to fully recover many critical details. Moreover, these methods typically address human perceptual needs but do not sufficiently cater to the preferences or feature spaces required for machine vision, leaving this area of research still open.

Therefore, recent studies have revisited the full utilization of raw images, aiming to exploit their auxiliary and relatively clean visual information to improve machine vision applications [8, 9, 10]. Although the observed performance gains highlight the advantages of raw images over processed sRGB images in machine vision tasks, raw-based methods have shown progress but still face significant challenges that need to be addressed before they can be effectively applied in real applications. In particular, current approaches tend to take raw images as direct input [11, 12] and make substantial architectural changes to accommodate the different image format, whereas other studies [13, 14, 15] modify the in-camera ISP modules instead. These strategies pose flexibility issues as sRGB-based pipelines remain mainstream and perform well under normal lighting conditions. Moreover, the substantial storage and transmission resources requirements of raw images further hinder the practical adoption of raw-based solutions.

This paper introduces the Compact RAW Metadata-guided Image Refinement (CRM-IR) paradigm to tackle challenges in low-light machine vision, characterized by strong *flexibility* and *compactness*. The approach extracts only essential raw image information, boosting downstream models, and integrates seamlessly into existing vision pipelines without altering current architectures or ISP procedures, as shown in Fig. 1. At its core, CRM-IR features a lightweight Image Refinement (MV-IR) module, which uses raw metadata as external conditions for pixel-level modifications to processed sRGB images *flexibly*. Through joint end-to-end training with metadata extraction and downstream networks, the MV-IR module flexibly adapts to machine vision needs. To ensure *compactness*, a raw-metadata encoder, based on Learned Image Compression (LIC), is introduced. This encoder extracts and compresses raw metadata efficiently, utilizing a cross-modal contextual entropy coding strategy that leverages semantic correspondence between the sRGB image and raw data for effective compression. Additionally, we present the Raw in Dark (RID) dataset, containing 500 annotated RAW sensor pairs from low-light daily scenes, further advancing RAW-based object detection. Extensive experiments demonstrate that CRM-IR achieves superior performance compared to existing methods, with minimal metadata transmission of less than 0.001 bits per pixel.

Our contributions are summarized as follows:

- We propose a novel raw-based machine vision paradigm that extracts only the most essential information from raw images to guide machine-oriented sRGB image refinement. This

design enables seamless integration as a plug-in within existing sRGB-based vision pipelines while maintaining minimal storage and transmission overhead.

- We introduce a novel raw-metadata encoder that effectively leverages the cross-modal contextual information between processed sRGB and raw images. This enables the proposed scheme to transmit only a minimal amount of raw metadata while significantly enhancing downstream performance.
- To advance RAW-based machine vision, we construct Raw in Dark (RID)—a diverse, large-scale dataset of 500 annotated RAW-sRGB pairs captured in real-world low-light scenarios across 8 object categories. RID fills key gaps in existing open-source datasets and serves as a strong benchmark for evaluating generalization in RAW-guided object detection through cross-dataset validation.

2 Related Work

2.1 Raw-based Machine Vision

The last decade has witnessed substantial progress in machine vision techniques, markedly improving scene-understanding accuracy [18, 19, 20, 21, 22] and inference speed [23, 24, 25, 26] and enabling widespread adoption in real-world applications. Nevertheless, robust performance under low-light conditions remains elusive. Dim illumination limits photon counts and produces a low signal-to-noise ratio [3, 4], while compensatory measures such as high-ISO amplification or extended exposure add sensor noise and motion blur [5]; together these factors shift the input distribution away from the priors learned during training and sharply reduce model accuracy. Early studies inserted a preprocessing stage to enhance semantic information before inference [27, 28], yet this approach did not achieve satisfactory performance and constrained the models' generalizability and scalability across real-world benchmarks and diverse tasks. Consequently, recent research has turned to raw images, seeking to harness their abundant unprocessed visual information to improve both image enhancement and downstream machine-vision performance. For instance, some studies [11, 12] adopt a straightforward strategy by training or fine-tuning downstream models directly on raw images, whereas others [13, 14, 15] focus on the ISP pipeline and employ differentiable image signal processors to generate sRGB images tailored for machine-vision tasks. Although these raw-based methods deliver substantial gains, they require full access to raw data, which hampers deployment in edge-to-cloud scenarios where transmitting full-resolution raw images is impractical.

2.2 Learned Image Compression

In recent years, Learned Image Compression (LIC) has progressed rapidly in step with deep-learning breakthroughs. In particular, Ballé *et al.* [29] spearheaded this shift by replacing handcrafted transforms, quantizers and entropy coders with a single, fully trainable pipeline. They later augmented their framework with a hyperprior that conditions each latent on auxiliary hyper-latents, markedly improving the rate-distortion trade-off over factorized-prior baselines [30]. Follow-up studies refined the entropy model by exploiting contextual cues: local part-of-image context [31], context spanning the entire image [32], checkerboard inference patterns [33], and channel-wise context [34, 35] have all been employed to boost accuracy or reduce computation. A recent exemplar, MLIC++ by Jiang *et al.* [36], fuses local, global and channel contexts in a multi-reference entropy model and already surpasses the latest coding standard, Versatile Video Coding (VVC) [37]. Meanwhile, other LIC works tend to enhance the analysis–synthesis backbone itself by adopting residual network [38], invertible network [39] and Swin-Transformer [40, 41], yielding richer latent representations and further gains in compression.

3 Method

3.1 Motivation and Overview

To address the limitations of existing raw-based methods in terms of flexibility and compactness, we propose a novel framework called Compact RAW Metadata-guided Image Refinement (CRM-IR). CRM-IR aims to fully leverage raw images to enhance downstream machine vision tasks while

maintaining a lightweight and pluggable structure compatible with existing vision pipelines. The overall framework is illustrated in Fig. 2, consisting of three key components.

1) Raw Metadata Extraction. Let x_r denote the input raw image and x_s represent its sRGB counterpart. At the imaging stage, we employ a raw metadata encoder $G(\cdot;\boldsymbol{\omega})$ parameterized by $\boldsymbol{\omega}$ that takes as input x_r while being conditioned on x_s, to identify and extract the raw metadata y,

$$y = G(x_s, x_r; \boldsymbol{\omega}). \tag{1}$$

2) Metadata Coding. Subsequently, a hyperprior-based entropy encoder $E(\cdot,\boldsymbol{\theta})$ parameterized by $\boldsymbol{\theta}$ is adopted to capture the statistical property of $\hat{y}$ under a multivariate Gaussian distribution, while simultaneously estimating and constraining its entropy, denoted as $E(y;\boldsymbol{\theta})$ for simplicity, which will be elaborated later.

3) Image Refinement for Vision Tasks. At the application end, an image refinement model $M(\cdot;\boldsymbol{\phi})$ is incorporated, being responsible for adapting x_s to align with the requirements of downstream machine-vision tasks guided by y,

$$\widehat{x_s} = M(x_s, y; \boldsymbol{\phi}), \tag{2}$$

where $\widehat{x_s}$ is the refined sRGB image, $\boldsymbol{\phi}$ denotes the model parameter.

By feeding the $\widehat{x_s}$ to the downstream machine vision models $A(\cdot,\boldsymbol{\psi})$, the entire CRM-IR scheme is capable of end-to-end training under the following constraint,

$$\bar{\boldsymbol{\theta}}, \bar{\boldsymbol{\phi}}, \bar{\boldsymbol{\psi}}, \bar{\boldsymbol{\omega}} = \underset{(\phi,\psi,\theta,\omega)}{\arg\min} \sum_{(x_r,x_s,z)\in D} \lambda \cdot \mathcal{L}_A(z, A(M(x_s, G(x_s, x_r;\boldsymbol{\omega});\boldsymbol{\phi});\boldsymbol{\psi})) + E(y;\boldsymbol{\theta}), \tag{3}$$

where D denotes the training set, consisting of the raw image x_r, sRGB image x_s and annotation z for machine vision task. $\mathcal{L}_A$ measures the task performance, while λ is the Lagrange parameter to balance the bits cost and downstream task performance.

To meet the dual goals of *compactness* and *flexibility*, we incorporate the following designs:

- **Compactness**: We introduce a Cross-Modal Contextual Entropy Encoder (CMCE). The raw metadata must comprise only essential information from x_r while remaining independent of x_s with respect to machine vision requirements. CMCE employs the sRGB image x_s as a contextual prior during raw metadata compression. This approach effectively captures the inter-redundancy between the raw and the sRGB images.
- **Flexibility**: We design a lightweight Machine Vision-oriented Image Refinement (MV-IR) module, which modifies x_s at the pixel level under the guidance of metadata y. MV-IR introduces no changes to the image format, allowing seamless integration into existing pipelines without altering downstream architectures.

Detailed descriptions of the proposed CMCE and MV-IR are provided in the following subsections.

3.2 Cross-modal Contextual Entropy Encoder

In the field of LIC, context-based entropy modeling involves utilizing surrounding contexts, *i.e.*, information from already encoded latent elements, to dynamically estimate the probability distribution for encoding subsequent elements. This approach leverages complex correlations in the input images to improve compression efficiency. Motivated by this, we utilize the cross-modal dependency between the sRGB image x_s and raw image x_r, both available at the imaging end, to further enhance the compression process. In particular, the x_s would be first concatenated with the x_r while ensure that its latent elements are leveraged as contextual priors for estimating the probability distribution of the latent representation of x_r. After obtaining their joint latent representation y via Eqn. (1), a hyperencoder is introduced to extract side information $z = h_a(y)$, which captures the spatial dependencies among the elements of y.

In coding practice, y and z are typically passed through a uniform quantization process to obtain their integer forms $\hat{y}$ and $\hat{z}$ for compression purposes. During training, this quantization is approximated using uniform noise $\mathcal{U} \sim (-\frac{1}{2}, \frac{1}{2})$. As such, the actual rate estimation can be formulated as,

$$\mathcal{R}(\hat{y}) + \mathcal{R}(\hat{z}) = \mathbb{E}[-log_2(p_{\hat{y}|\hat{z}}(\hat{y}|\hat{z}))] + \mathbb{E}[-log_2(p_{\hat{z}}(\hat{z}))]. \tag{4}$$

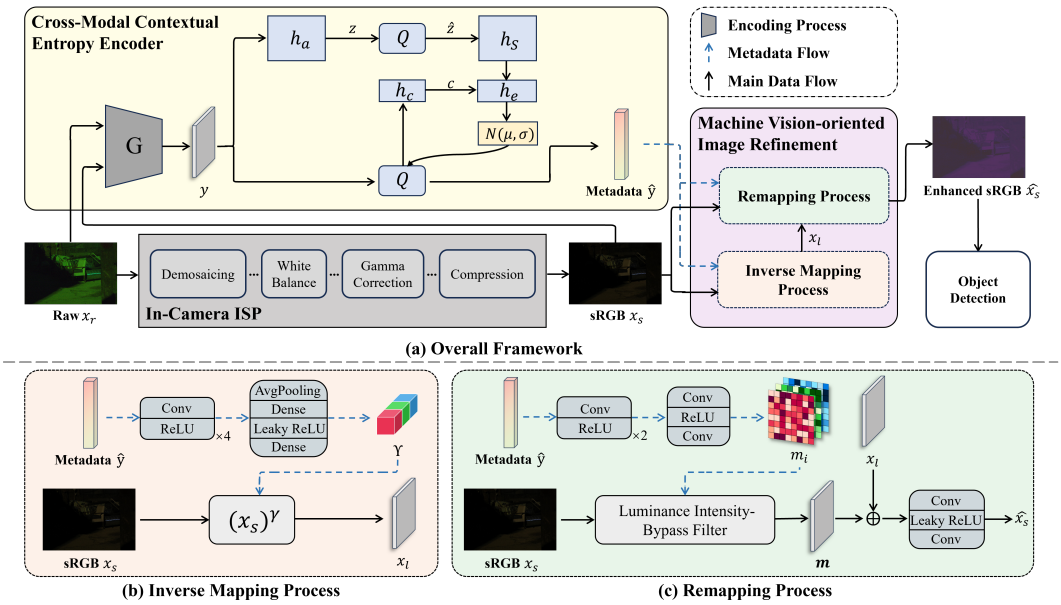

Figure 2: (a) Overall framework of the proposed CRM-IR framework. (b) and (c) details the inverse mapping and the remapping process of the proposed MV-IR module, respectively.

To obtain $p_{\hat{y}|\hat{z}}(\hat{y}|\hat{z})$, a multivariate Gaussian prior is introduced, parameterized by a hyperdecoder $h_s(\cdot)$,

$$p_{\hat{y}|\hat{z}}(\hat{y}|\hat{z}) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}), [\boldsymbol{\mu}, \boldsymbol{\sigma}] \leftarrow h_s[\hat{z}; \theta_s], \tag{5}$$

where the θ_s denotes the parameters of $h_s(\cdot)$. Within this process, an autoregressive context model h_c is employed to capture the contextual information within $\hat{y}$. It utilizes both the side information $\hat{z}$ and already processed and the already processed portion of $\hat{y}$ to predict the distribution of the remaining elements. Accordingly, the causal context c_i for a given latent element is obtained via $c_i = h_c(\hat{\mathbf{y}}_{<i}; \theta_c)$, where $\hat{\mathbf{y}}_{<i}$ denotes the causal context of y_i. The prediction of the Gaussian parameters can be roughly formulated as a function involving the hyper-decoder, context model and entropy parameter network,

$$\mu_i, \sigma_i = h_e(h_s(\hat{z}; \theta_s), c_i; \theta_e), \tag{6}$$

where $h_e(\cdot; \theta_e)$ denotes the entropy parameter network with parameter θ_e. Thus, we obtained

$$p_{\hat{y}|\hat{z}}(\hat{y}|\hat{z}) = \prod_i \left(\mathcal{N}(\mu_i, \sigma_i) * \mathcal{U}(-\frac{1}{2}, \frac{1}{2}) \right) (\hat{y}_i). \tag{7}$$

As for the entropy coding of the side information $\hat{z}$, *i.e.,* the estimation of $\mathbb{E}[-log_2(p_{\hat{z}}(\hat{z}))$ in Eqn. (4) since no hyperprior is available for this process, a simple factorized density model is employed. This model is directly adopted from existing LIC works and is not elaborated upon here.

3.3 Machine Vision-oriented Image Refinement based on Raw Metadata

As revealed by existing raw-based works [1], within the ISP process, white balancing and gamma correction are the main modules that negatively affect downstream machine vision tasks, as both involve nonlinear mapping of pixel luminance to a discrete domain aligned with human perceptual preferences. Therefore, the main idea of the proposed MV-IR is to first convert the sRGB image back to a linear space, guided by the raw metadata. Subsequently, an additional remapping process is introduced to enhance details and structures in low-light regions, specifically tailored to machine vision requirements.

In particular, during the inverse remapping process, to simplify the task, we introduce an additional prior based on gamma correction, *i.e.,* we aim to predict an image-wise gamma correction parameter that is also aligned with the ISP process. Thus, the inverse mapping process can be formulated by,

$$x_l = (x_s)^{\gamma}, \gamma = F_{im}(\hat{y}), \tag{8}$$

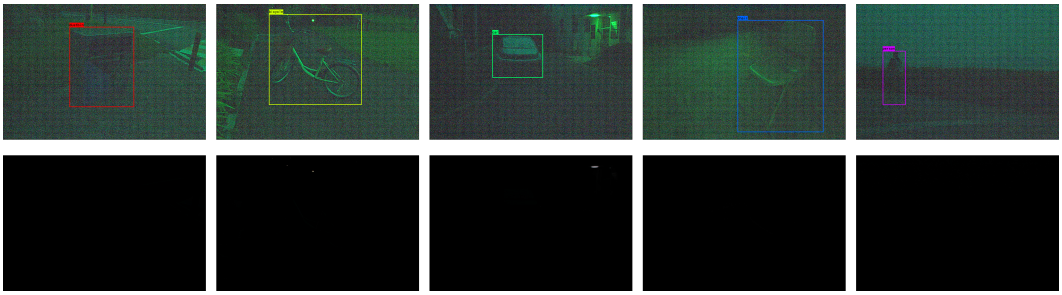

Figure 3: Samples from the RID dataset: The first row shows RAW images and the second row shows sRGB images processed by the camera's internal ISP.

where x_l denotes the linear representation of x_s, and $F_{im}(\cdot)$ is responsible for predicting the gamma parameter. This prediction is straightforward with the assistance of the raw metadata and can be efficiently implemented using only a few convolutional layers, as shown in Fig. 2 (b).

As for the remapping process illustrated in Fig. 2 (c), $F_{rm}(\cdot)$ is employed with the aim of learning a pixel-wise modulation map $\mathbf{m}$, conditional on x_s and $\hat{y}$. Considering that our method is specifically designed for low-light scenarios, we introduce a luminance region-wise learning strategy, *i.e.*, explicitly learning distinct modulation maps $[m_1, m_2]$ corresponding to low-light and normal-light regions.

$$m_i = F_{rm}(x_s, \hat{y}), i \in (1, 2), \tag{9}$$

where $F_{rm}(\cdot)$ is for the modulation map prediction. They would be subsequently merged to $\mathbf{m}$ via through luminance intensity-bypass filtering operations.

$$\mathbf{m} = \sum_{i=1}^{2} m_i \cdot \delta_i(x_s), \tag{10}$$

where $\delta_1(\cdot)$ and $\delta_2(\cdot)$ denote pixel-wise bypass filters that selectively pass the corresponding elements of the learned maps when the pixel intensities fall within the low-light and normal-light regions, respectively. Afterward, the modulation map $\mathbf{m}$ is applied to the linear representation x_l via summation and subsequently fed into lightweight fusion layers $F_f(\cdot)$ for further adjustment, producing the final output.

$$\widehat{x_s} = F_f(\mathbf{m} + x_l). \tag{11}$$

3.4 Raw in Dark(RID) Dataset

This subsection introduces the RID dataset to enable cross-dataset generalization validation, given the scarcity of open-source real–world benchmarks. Compared with the widely employed raw-based object detection dataset LOD [42], RID broadens the range of scenes and object categories and places particular emphasis on images captured under extremely low-light conditions. As for the source image collection, a Canon EOS80D was used to capture 500 paired samples, each consisting of a RAW image and its sRGB-JPEG counterpart. We selected scenes that span a range of everyday environments, including dimly lit underground parking garages, nighttime roadsides and poorly illuminated indoor rooms, all characterized by deep shadows and severely limited visibility. We selected ISO settings of 50 and 100 and exposure times of 1/125s and 1/250s to replicate the photon-starved conditions typical of extremely dark surveillance scenarios.

Regarding the annotations, professional annotators are employed to create accurate instance-level labels for widely studied classes, *i.e.*, car, bicycle, chair, to support cross-task generalization validation, and further categories including person, zebra crossing, emergency exit sign, fire extinguisher and dustbin are included to extend the dataset's coverage. Representative examples are exhibited in Fig. 3. It is worth noting that RID is an ongoing project. We plan to further expand the dataset by employing various shooting devices, including different cameras and smartphones, with the goal of investigating the influence of diverse ISP procedures. Additionally, we will include more annotations for other critical machine vision tasks, such as semantic segmentation and instance segmentation, aiming to support the broader machine vision research community.

Table 1: Quantitative comparison results on the YOLOv3 backbone. The bold and underlined font indicate the best and second-best results, respectively. For reference, JPEG-format sRGB images have an average bpp of 0.15, whereas raw images require 48 bpp.

Method	In-Dataset Val. (LOD)				Cross-Dataset Val. (RID)			
	mAP	mAP50	mAP75	bpp	mAP	mAP50	mAP75	bpp
YOLOv3	40.00	67.67	43.44	-	<u>45.67</u>	65.80	<u>63.75</u>	-
Zero-DCE	41.19	67.79	45.79	-	45.26	64.38	62.76	-
KinD	41.22	67.03	44.60	-	44.37	63.76	58.55	-
YOLA	41.00	68.49	<u>45.87</u>	-	38.29	60.22	46.29	-
RAOD	<u>41.60</u>	**70.49**	42.29	-	39.74	**68.49**	41.36	-
Ours	**42.14**	<u>69.11</u>	**46.02**	4.88e-4	**51.72**	<u>65.81</u>	**63.85**	4.86e-4

4 Experiments

At the experimental stage, we implemented our CRM-IR scheme on a set of object detection models and conducted extensive comparisons with other low-light object detection paradigms, including both sRGB-based and raw-based methods, to demonstrate the superiority of our approach. Moreover, comprehensive ablation studies were performed to validate the effectiveness of the proposed CMCE and MV-IR modules.

4.1 Experimental Settings

Datasets. Low-light Object-Detection (LOD) dataset [42] is employed as our benchmark, which contains 2230 paired RAW and sRGB-JPEG format image pairs collected by a Canon EOS 5D Mark IV camera, covering both low-light and normal daylight scenes, where only the low-light parts are selected in our experiments. In particular, 1,784 training pairs and 446 test pairs are selected for model training and evaluation, respectively. These images contain a total of 9,726 labeled instances spanning 8 common object classes: car, motorbike, bicycle, chair, dining table, bottle, TV monitor and bus. In addition, our proposed RID dataset is employed to perform cross-dataset testing, aiming to evaluate the generalization capability of the employed methods.

Baselines. Three milestone object detection models are employed as baseline models, including YOLOv3 [43], Faster R-CNN [44] and CenterNet [21]. Moreover, four state-of-the-art low-light object detection schemes are evaluated, including three sRGB-based methods: Zero-DCE [16], KinD [45] and YOLA [46], all of which follow a similar pipeline that first enhances low-light images before feeding them into downstream object detection models. Moreover, one raw-based method RAOD [17] is also adopted. It follows a similar enhancement-based pipeline by leveraging raw images to guide the enhancement process, but it overlooks the transmission cost of raw data. For a fair comparison, all competing schemes are implemented on the same baseline models as our method and trained from scratch using the same dataset.

Evaluation Criterion. We adopt the commonly used mean average precision regarding mAP, mAP50 and mAP75 to measure the downstream task performance. Moreover, to demonstrate the compactness of our method, we report the bpp as an indicator of storage or transmission resource requirements across different methods. For the employed methods, we report the bpp of sRGB images and raw images for sRGB-based and raw-based approaches, respectively. In contrast, for our method, we additionally report the bpp of the raw metadata alongside the sRGB image.

Implementation Details. During training, data augmentation strategies are employed, including random horizontal flips and random scale jitter during resizing. All models were trained for 300 epochs using the Adam optimizer [47]. A linear scaling learning rate with a cosine decay schedule was employed, starting from an initial learning rate of 5e-4. The weight decay was set to 0, momentum was 0.9 and the training batch size was set to 8. During both training and testing, all input images were resized to 512×512. All experiments were conducted using PyTorch on an NVIDIA RTX 6000 Ada Generation GPU with 48 GB of memory.

Table 2: Quantitative comparison results on the Faster R-CNN backbone. The bold and underlined font indicate the best and second-best results, respectively.

Method	In-Dataset Val. (LOD)				Cross-Dataset Val. (RID)			
	mAP	mAP50	mAP75	bpp	mAP	mAP50	mAP75	bpp
Faster R-CNN	39.96	67.12	41.74	-	37.23	63.79	39.30	-
Zero-DCE	40.50	67.94	42.56	-	37.73	64.57	40.07	-
KinD	39.83	66.86	42.23	-	37.11	63.54	39.76	-
YOLA	40.31	68.22	42.31	-	37.55	64.83	39.84	-
RAOD	<u>41.32</u>	**69.79**	<u>43.14</u>	-	<u>38.50</u>	**66.33**	<u>40.62</u>	-
Ours	**41.75**	<u>68.49</u>	**43.94**	3.26e-4	**38.89**	<u>65.09</u>	**41.37**	3.23e-4

Table 3: Quantitative comparison results on the CenterNet backbone. The bold and underlined font indicate the best and second-best results, respectively.

Method	In-Dataset Val. (LOD)				Cross-Dataset Val. (RID)			
	mAP	mAP50	mAP75	bpp	mAP	mAP50	mAP75	bpp
CenterNet	40.41	65.36	42.60	-	40.33	62.63	44.67	-
Zero-DCE	41.44	65.85	**44.90**	-	40.30	62.64	44.53	-
KinD	40.10	62.49	44.23	-	40.28	62.48	44.52	-
YOLA	41.74	66.18	<u>44.55</u>	-	42.04	66.48	<u>44.85</u>	-
RAOD	**42.89**	**69.25**	43.94	-	**43.45**	**70.28**	43.22	-
Ours	<u>42.05</u>	<u>68.09</u>	44.52	1.38e-3	<u>42.31</u>	<u>68.28</u>	**44.91**	1.39e-3

4.2 Experimental Results

The quantitative evaluation results regarding the baseline models of YOLOv3, Faster R-CNN and CenterNet are established in Table 1, Table 2, Table 3, respectively. First, the effectiveness of the proposed method can be easily observed comparing with the baseline models, as leading to an average mAP improvement of 2.14%, 1.79% and 1.64% regarding YOLOv3, Faster R-CNN and Centernet, respectively. Considering bitrate performance, our method introduces only 4.88e-4, 3.26e-4 and 1.38e-3 bpp of raw-metadata overhead for the three baseline models, respectively. Relative to the original sRGB bitrate of 0.15 bpp, this increase is under 0.5 percent and therefore negligible in practice, thanks to the CMCE module's ability to capture and exploit the contextual correspondence between each raw image and its sRGB counterpart.

Compared with state-of-the-art sRGB-based methods Zero-DCE, KinD and YOLA, our scheme achieves an average mAP improvement of about 0.94%, 1.60% and 0.96%, confirming that leveraging raw-image information effectively overcomes the limitations of conventional ISP processing. Compared with the raw-based method RAOD, our approach achieves comparable performance while transmitting only a compact metadata stream instead of the entire raw image. RAOD benefits from full access to the 48 bpp raw data, yet this requirement imposes a prohibitive storage and transmission load, making the scheme impractical for many real-world deployments. Cross-dataset evaluations confirm the strong generalization capability of our method, as it maintains nearly the same level of object-detection accuracy in unseen scenarios.

Fig. 4 provides a set of intuitive comparisons. Things have to be mentioned that, distinct visualization strategies were adopted for a clear representation. For the Ground Truth, detection bounding boxes are visualized directly on the original low-light input. For the other approaches that follow the "*enhancement-then-detection*" pipeline, detection results are shown on their corresponding enhanced images. Fig. 4 demonstrates that low light sRGB images contain considerable environmental noise that the enhancement stage cannot fully remove, causing frequent mis-detections. By contrast, incorporating raw information reduces noise in the enhance version and substantially improves downstream detection accuracy.

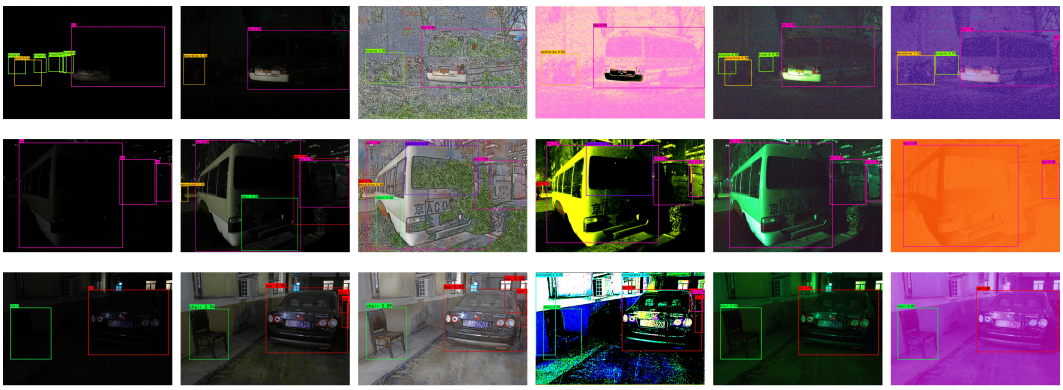

Figure 4: Visualization of object-detection results: the first, second and third rows correspond to methods based on YOLOv3, Faster R-CNN and CenterNet, respectively.

4.3 Ablation Studies

To evaluate the distinct contributions of our proposed modules and the utility of leveraging raw image data, we performed a detailed ablation study on the LOD datasets. We compare our full model against several ablated variants:

w/o **CMCE:** In this configuration, the metadata extraction and compression stages are disabled, so no raw metadata are passed to the MV-IR module. The MV-IR remains active but is driven by a non-informative placeholder implemented as a zero tensor.

w/o **MV-IR:** We retain the CMCE module but employ the enhancement module with simple concatenation, where the extracted metadata and the sRGB image features are combined via concatenation.

The ablation results are presented in Table 4, where the effectiveness of each specially designed module is evident. As shown, excluding the raw metadata yields an average mAP50 decrease of about 0.58%, 2.99% and 2.24% regarding YOLOv3, Faster R-CNN and CenterNet, respectively. Meanwhile, omitting the MV-IR module results in an overall 1.91%, 3.26% and 1.77% decrease in YOLOv3, Faster R-CNN and CenterNet, respectively.

Table 4: Ablation Study Results of CMCE and MV-IR modules

Detector	CMCE	MV-IR	mAP	mAP50	mAP75
YOLOV3	✗	✗	40.00	67.67	43.44
	✗	✓	41.06	68.53	44.78
	✓	✗	41.19	67.20	45.21
	✓	✓	42.14	69.11	46.02
Faster R-CNN	✗	✗	39.96	67.12	41.74
	✗	✓	39.65	65.85	42.40
	✓	✗	38.84	65.23	41.13
	✓	✓	41.75	68.49	43.94
CenterNet	✗	✗	40.41	65.36	42.60
	✗	✓	41.16	65.85	43.67
	✓	✗	41.64	66.32	44.25
	✓	✓	42.05	68.09	44.52

5 Conclusion

To conclude, this paper explores a new raw-based object-detection paradigm called Raw Metadata-guided Image Refinement (CRM-IR). Compared with existing raw-based approaches,

CRM-IR is characterized by its flexibility and compactness: it integrates seamlessly into current machine vision pipelines while explicitly accounting for the transmission and storage overhead of raw information. For flexibility, we introduce a Machine Vision-oriented Image Refinement (MV-IR) module that adjusts sRGB images to better match machine vision preferences, functioning as a standalone preprocessing step without altering network architectures or in-camera ISP modules. CRM-IR further leverages raw metadata collected at the imaging stage; by extracting only the essential raw information and using the paired sRGB image as contextual prior, it delivers significant gains in downstream object-detection accuracy with negligible additional bitrate.

Acknowledgments

This work was in part by the Interdisciplinary Frontier Research Project of PCL under Grant 2025QYB013, in part by the Major Key Project of PCL (PCL2025A03), in part by Guangdong Basic and Applied Basic Research Foundation under Grant 2024A1515010454, in part by the Open Research Fund from Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ) under Grant No. GML-KF-24-27, in part by the Natural Science Foundation of Guangdong Province under Grant 2023A1515011667.

References

[1] Haofeng Huang, Wenhan Yang, Yueyu Hu, Jiaying Liu, and Ling-Yu Duan. Towards low light enhancement with raw images. *IEEE Transactions on Image Processing*, 31:1391–1405, 2022.

[2] Yufei Wang, Yi Yu, Wenhan Yang, Lanqing Guo, Lap-Pui Chau, Alex C Kot, and Bihan Wen. Beyond learned metadata-based raw image reconstruction. *International Journal of Computer Vision*, 132(12):5514–5533, 2024.

[3] Kaixuan Wei, Ying Fu, Yinqiang Zheng, and Jiaolong Yang. Physics-based noise modeling for extreme low-light photography. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):8520–8537, 2021.

[4] Wei Wang, Xin Chen, Cheng Yang, Xiang Li, Xuemei Hu, and Tao Yue. Enhancing low light videos by exploring high sensitivity camera noise. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4111–4119, 2019.

[5] Tamer Rabie. Adaptive hybrid mean and median filtering of high-iso long-exposure sensor noise for digital photography. *Journal of Electronic Imaging*, 13(2):264–277, 2004.

[6] 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.

[7] Lingyu Zhu, Wenhan Yang, Baoliang Chen, Hanwei Zhu, Xiandong Meng, and Shiqi Wang. Temporally consistent enhancement of low-light videos via spatial-temporal compatible learning. *International Journal of Computer Vision*, 132(10):4703–4723, 2024.

[8] Barak Battash, Haim Barad, Hanlin Tang, and Amit Bleiweiss. Mimic the raw domain: Accelerating action recognition in the compressed domain. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 684–685, 2020.

[9] Yukihiro Sasagawa and Hajime Nagahara. Yolo in the dark-domain adaptation method for merging multiple models. In *Proceedings of the European Conference on Computer Vision*, pages 345–359, 2020.

[10] Yufan Liu, Jiajiong Cao, Weiming Bai, Bing Li, and Weiming Hu. Learning from the raw domain: Cross modality distillation for compressed video action recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1–5, 2023.

[11] Xiangyu Zhang, Ling Zhang, and Xin Lou. A raw image-based end-to-end object detection accelerator using hog features. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 69(1):322–333, 2022.

[12] Zhong-Yu Li, Xin Jin, Bo-Yuan Sun, Chun-Le Guo, and Ming-Ming Cheng. Towards raw object detection in diverse conditions. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8859–8868, 2025.

[13] Ziteng Cui and Tatsuya Harada. Raw-adapter: Adapting pre-trained visual model to camera raw images. In *Proceedings of the European Conference on Computer Vision*, pages 37–56, 2024.

[14] Haina Qin, Longfei Han, Juan Wang, Congxuan Zhang, Yanwei Li, Bing Li, and Weiming Hu. Attention-aware learning for hyperparameter prediction in image processing pipelines. In *Proceedings of the European Conference on Computer Vision*, pages 271–287, 2022.

[15] Yujin Wang, Tianyi Xu, Zhang Fan, Tianfan Xue, and Jinwei Gu. Adaptiveisp: Learning an adaptive image signal processor for object detection. *Advances in Neural Information Processing Systems*, 37:112598–112623, 2024.

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

[17] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13384–13393, 2023.

[18] 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 Systems*, 30(11): 3212–3232, 2019.

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

[20] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *Proceedings of the European Conference on Computer Vision*, pages 21–37. Springer, 2016.

[21] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. Centernet: Keypoint triplets for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6569–6578, 2019.

[22] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF international Conference on Computer Vision*, pages 9627–9636, 2019.

[23] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 580–587, 2014.

[24] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14454–14463, 2021.

[25] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 779–788, 2016.

[26] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021.

[27] Xiaojie Guo, Yu Li, and Haibin Ling. Lime: Low-light image enhancement via illumination map estimation. *IEEE Transactions on Image Processing*, 26(2):982–993, 2017.

[28] Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Deep retinex decomposition for low-light enhancement. In *Proceedings of the British Machine Vision Conference*, 2018.

[29] Johannes Ballé, Valero Laparra, and Eero P Simoncelli. End-to-end optimized image compression. In *Proceedings of the International Conference on Learning Representations*, 2017.

[30] Johannes Ballé, David Minnen, Saurabh Singh, Sung Jin Hwang, and Nick Johnston. Variational image compression with a scale hyperprior. In *Proceedings of the International Conference on Learning Representations*, 2018.

[31] David Minnen, Johannes Ballé, and George D Toderici. Joint autoregressive and hierarchical priors for learned image compression. *Advances in Neural Information Processing Systems*, 31, 2018.

[32] Zongyu Guo, Zhizheng Zhang, Runsen Feng, and Zhibo Chen. Causal contextual prediction for learned image compression. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(4):2329–2341, 2021.

[33] Dailan He, Yaoyan Zheng, Baocheng Sun, Yan Wang, and Hongwei Qin. Checkerboard context model for efficient learned image compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14771–14780, 2021.

[34] David Minnen and Saurabh Singh. Channel-wise autoregressive entropy models for learned image compression. In *Proceedings of the IEEE International Conference on Image Processing*, pages 3339–3343, 2020.

[35] Dailan He, Ziming Yang, Weikun Peng, Rui Ma, Hongwei Qin, and Yan Wang. Elic: Efficient learned image compression with unevenly grouped space-channel contextual adaptive coding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5718–5727, 2022.

[36] Wei Jiang, Jiayu Yang, Yongqi Zhai, Peirong Ning, Feng Gao, and Ronggang Wang. Mlic: Multi-reference entropy model for learned image compression. In *Proceedings of the ACM International Conference on Multimedia*, pages 7618–7627, 2023.

[37] Benjamin Bross, Ye-Kui Wang, Yan Ye, Shan Liu, Jianle Chen, Gary J Sullivan, and Jens-Rainer Ohm. Overview of the versatile video coding (VVC) standard and its applications. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(10):3736–3764, 2021.

[38] F Chen, Y Xu, and L Wang. Two-stage octave residual network for end-to-end image compression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 3922–3929, 2022.

[39] Y Xie, K Cheng, and Q Chen. Enhanced invertible encoding for learned image compression. In *Proceedings of the ACM International Conference on Multimedia*, pages 162–170, 2021.

[40] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international Conference on Computer Vision*, pages 10012–10022, 2021.

[41] Ming Lu, Peiyao Guo, Huiqing Shi, Chuntong Cao, and Zhan Ma. Transformer-based image compression. In *Proceedings of the Data Compression Conference*, pages 469–469, 2022.

[42] Yang Hong, Kaixuan Wei, Linwei Chen, and Ying Fu. Crafting object detection in very low light. In *British Machine Vision Conference*, volume 1, page 3, 2021.

[43] Joseph Redmon and Ali Farhadi. YOLOv3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[44] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems*, 28, 2015.

[45] Yonghua Zhang, Jiawan Zhang, and Xiaojie Guo. Kindling the darkness: A practical low-light image enhancer. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 1632–1640, 2019.

[46] Mingbo Hong, Shen Cheng, Haibin Huang, Haoqiang Fan, and Shuaicheng Liu. You only look around: Learning illumination-invariant feature for low-light object detection. *Advances in Neural Information Processing Systems*, 37:87136–87158, 2024.

[47] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations*, 2015.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: In the abstract and introduction, we delineate our research scope, address the limitations of RAW image utilization, and highlight our proposed methodology and key contributions.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [No]

   Justification: Potential limitations of our method may include inference latency. This was not detailed in the main body of the paper due to page limits.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: This paper does not introduce novel theoretical contributions. The equations used in the paper are intended to explain the working principles of each module.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We provide a comprehensive description of our experimental setup in Section 4, which covers datasets, baselines, evaluation metrics, and pertinent experimental details. To further support reproducibility, our code will be released as open source.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Data and code for our proposed method are publicly available. Please refer to the Abstract for links to the code repository and dataset.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: In Section 4, we have provided our experimental settings, such as datasets and optimizers. We will subsequently organize and open-source our code, making detailed specifics/information available.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Owing to constraints on time and page length, the current presentation of our experimental results does not include an analysis of error or other statistical significance measures.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: In section 4, we specified in our implementation details that an NVIDIA RTX 6000 Ada Generation GPU with 48 GB of memory was used for training.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in this paper fully complies with the NeurIPS Code of Ethics (https://neurips.cc/public/EthicsGuidelines).

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

Justification: We introduce a novel RAW-based machine vision paradigm that offers a new approach for industrial applications, which we consider a positive societal contribution. Furthermore, this work is not expected to have any adverse societal impacts.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The paper poses no such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All assets (code, data, models) employed in this paper have been appropriately credited/cited.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.

- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Our work introduces new methods and datasets, while have alredy been released.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.