

Supplementary Materials of An End-to-End Real-World Camera Imaging Pipeline

Anonymous Authors

1 ANALYSIS OF EXISTING DATASETS

The scarcity of datasets containing pairs of RAW and RGB images significantly constrains the development and evaluation of advanced image processing algorithms. At present, the principal datasets available are as follows:

The FiveK dataset[3], as outlined in Table 1, encompasses a collection of over 5,000 images sourced from five DSLR cameras. The shooting conditions of the FiveK data set are very ideal, with almost no noise. Its main purpose is to learn color correction tasks, so it is not suitable for training the entire ISPNet.

The SID (See In the Dark)[4] and ELD (Extreme Low-light Dataset) datasets are chiefly employed in research on noise reduction in RAW images under extremely low-light conditions. These datasets are limited by the absence of paired RGB images, which constrains their applicability to broader ISP tasks. Moreover, they exhibit a limited variety of imaging scenarios, which may introduce biases in algorithm development and assessment.

A case in point is the Zurich RAW to RGB dataset[5], which contains approximately 10,000 pairs of images, with RAW captures from Huawei P20 smartphones and corresponding RGB images acquired using Canon cameras. Due to the inherent challenges in aligning images from two distinct devices, this dataset employs the Scale-Invariant Feature Transform[8] (SIFT) method for image registration. Despite cropping the images to a standardized size of 448x448 pixels and excluding those with subpar alignment, significant discrepancies in image alignment persist. As depicted in Fig.1, careful examination of the designated areas—specifically within the red bounding boxes—reveals considerable misalignments between the paired images. This misalignment predisposes the training of computational models to pixel displacement artifacts, thereby impairing the sharpness and clarity of the generated images.

Previous RAW image processing methods usually use RawPy[1] to process RAW information and perform the ISP process. However, there are a series of problems in the RawPy processing pipeline, resulting in mediocre practicality. Specifically, images processed by RawPy exhibit deficiencies in several critical Image Signal Processing (ISP) stages. For instance, RawPy lacks capabilities in tonal mapping, which is pivotal for rendering images with high fidelity. This includes the absence of Global Tone Mapping (GTM) and Local Tone Mapping (LTM), as well as Adaptive Dynamic Range Compression (ADRC). These tone-mapping processes are essential for achieving a balanced exposure and dynamic range in the final image output, thereby affecting the overall quality and representativeness of the rendered RGB images. Furthermore, RawPy encounters issues with metadata loading, which can lead to inaccuracies in the processing pipeline. Metadata in image processing contains crucial information about the image settings and environment, such as exposure, color, and camera specifics. Errors in metadata loading can result in incorrect color rendition, exposure levels, and other image attributes that significantly impact the quality of the final

RGB output. Consequently, the limitations inherent in RawPy’s processing capabilities lead to RGB images that are of moderate quality and lack representativeness, which is a significant concern for applications requiring high-quality image renditions that closely mimic real-world ISP outcomes.

In summation, the extant datasets suffer from overly simplified modeling that fails to encapsulate the intrinsic characteristics of real-world camera imaging pipelines. This simplification leads to a practical gap between the dataset’s utility and the exigencies of real-world applications. Furthermore, the deficiency of extensive, high-caliber paired RAW and RGB image datasets severely impedes the training and validation of end-to-end imaging networks that are geared toward real-world deployment.

2 OUR DATASET DETAILS

To accumulate an extensive dataset to substantiate the research presented herein, we utilized a camera, the Canon M50. Employing the camera’s Auto Exposure (AE) professional mode while fixing the ISO sensitivity at 100, the camera automatically adjusted the exposure time and aperture to achieve optimal brightness levels for capturing images. This methodology facilitated the acquisition of RAW images alongside their corresponding RGB counterparts. Consequently, we amassed a large-scale, high-quality RAW-RGB paired dataset, which encompasses a broad spectrum of shooting scenarios, inclusive of varying lighting conditions and dynamic ranges. The dataset features a diverse array of scenes including, but not limited to, animals, landscapes, architecture, and portraiture.

The dataset comprises a total of 4507 images with a resolution of 6000x4000 pixels (width x height). Out of these, 4057 images have been allocated to the training set, with the remaining 450 designated for the test set to evaluate the performance of the proposed framework. Selected samples from the dataset are exemplified in Fig.2, while comparisons between the collected RAW images and the corresponding RGB images—directly output by the Canon camera’s ISP—are provided in Fig.3. In addition, we show the comparison results of RGB rendered by Canon and RGB rendered by RawPy in Fig.4. Canon RGB can obtain higher-quality results.

Upon the completion of the image collection, a preprocessing routine was applied to the data. Utilizing the open-source RAW image processing library DCRAW, we extracted the raw data from the Canon CR3 format RAW images. Automatic White Balance (AWB) is a well-established direction in image processing research, and to obviate the influence of this module, we preadjusted the color channel balance of the RAW data according to the white balance coefficients found in the Canon RAW image metadata. The RAW images were ultimately stored in 16-bit PNG format, while the RGB images retained the high-quality output directly from the Canon’s default ISP.

Table 1: Existing RAW-RGB Datasets

Dataset	MIT-Adobe FiveK	SID	Zurich RAW to RGB
Num	5000	5094	48000
Resolution	2000×3008 - 4368×3912	1616×1080 - 5472×3648	448×448
Characteristics	Indoor and outdoor scenes	Monotonous scenes, focused on indoor and low-light	Severe misalignment in paired data
Source of RGB Images	Lacks paired RGB	Lacks paired RGB	RGB images from Canon cameras



RAW



RGB

Figure 1: Illustration of Misalignment in the ZRR Dataset. The branches in the red box are misaligned.

3 IMPLEMENTATION DETAILS

To validate the effectiveness of the proposed method, training was conducted on the training split of our RAW-RGB dataset, followed by both quantitative and qualitative evaluations on the test split. During training, data augmentation techniques, such as random cropping and flipping, were applied, with the images being further processed into 256×256 patches.

The optimization strategy is detailed below:

We employed the Rate-Distortion (R-D) loss function, extensively used in deep learning-based image compression, allowing for the training of models at different bitrates by setting specific distortion weight parameters λ , with values $[0.1, 0.025, 0.01, 0.0035]$ for this study. The Adam optimizer was used with beta parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$. Training started with an initial learning rate of 0.0001, and a multi-step learning rate adjustment strategy was implemented, reducing the learning rate by a factor of 10 at the 1,000,000th and 1,500,000th iterations, with a batch size set at 8.

4 ADVANTAGES OF AN END-TO-END METHOD

To prove that the superiority of our method lies not only in the designed module's ability to simulate the real imaging pipeline but also in the end-to-end holistic design. We add the **CADR** module based on LiteISPNet to learn coordinate-related distortion repair and coordinate-independent mapping compression module **CiMC** to obtain RRISPNet. We connect RRISPNet in series with the most advanced image compression methods VTM, TCM and MLIC to obtain a new cascade imaging framework. Compared with these three new cascade imaging frameworks, our RealCamNet still has significant advantages, with a BD-PSNR improvement of 0.75db, which can prove that the advantages of the proposed method are not limited to designing new modules to implement new functions, but also include end-to-end overall framework advantages. We demonstrate the advantages of the overall framework in Table 3.

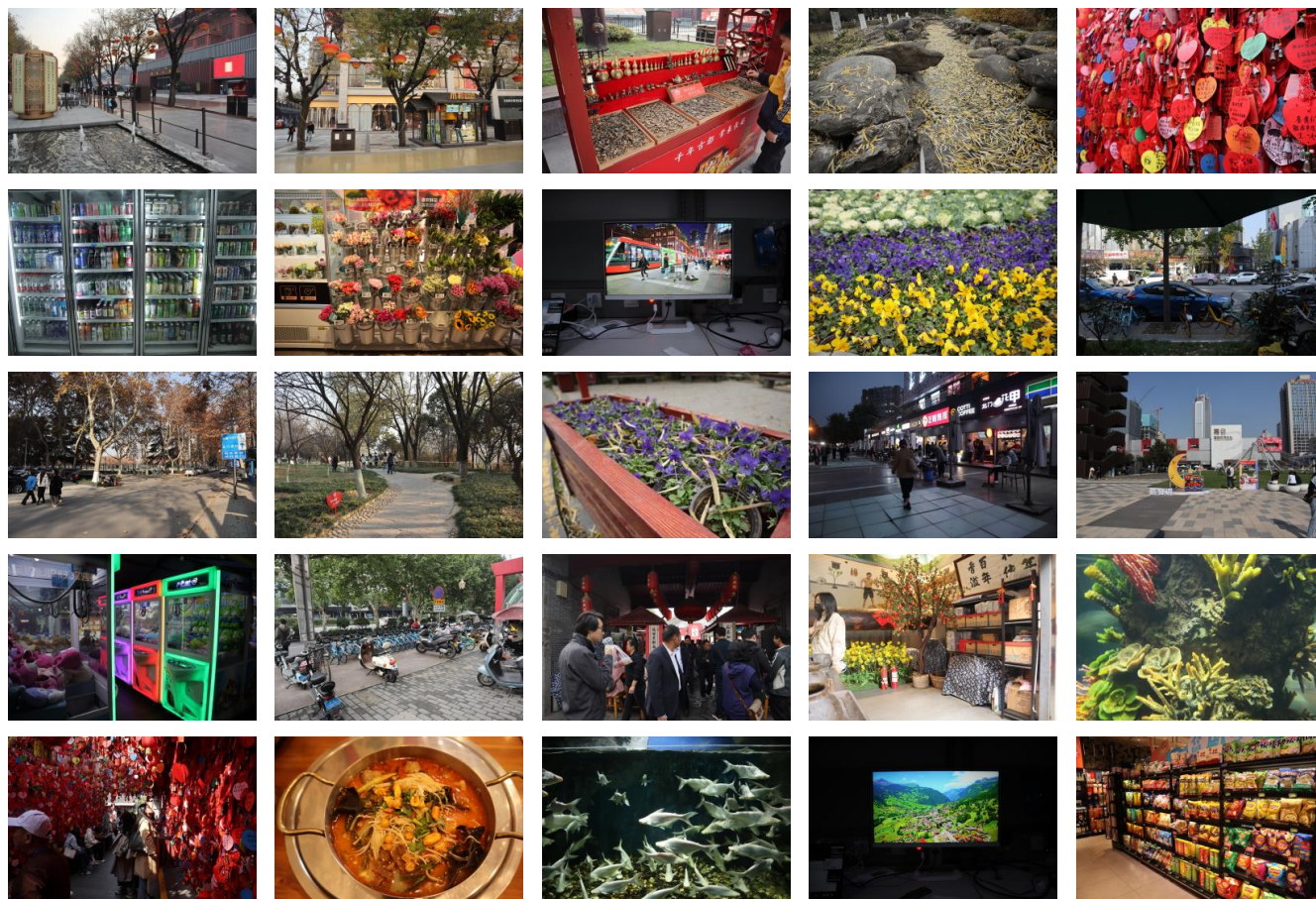


Figure 2: Sample images of the Our Dataset.



(a) 14bit RAW



(b) 8bit RGB

Figure 3: Canon RAW image and rendered RGB image.

5 DISADVANTAGES OF CASCADING SCHEME

The traditional cascaded framework of image processing is beset with several inefficiencies:



Figure 4: Canon rendered RGB vs RawPy rendered RGB image.

Table 2: Optimization Strategy Parameter Settings

Optimization Strategy		Parameter Settings
Loss Function		Rate-Distortion (R-D) Loss
Optimizer		Adam
Betas		[0.9, 0.999]
Learning Rate		0.0001
Iterations		2,000,000
Learning Rate Schedule	Reduce by 10x at 1,000,000 and 1,500,000 iterations	
Batch Size		8

- (1) **Independent demosaicing introduces redundant information through interpolation:** Initially, demosaicing interpolates the raw image data to synthesize high-resolution RGB images, which is followed by image compression. This interpolation introduces additional redundant information, detrimental to the efficiency of subsequent compression algorithms. Mathematically, the demosaic(upsampling) can be represented as a function f_{interp} that increases the resolution,

inadvertently adding noise and spurious data:

$$I_{\text{RGB}} = f_{\text{demosaic}}(I_{\text{raw}})$$

where I_{raw} and I_{RGB} represent the raw and the demosaic RGB images, respectively.

- (2) **Cumulative Error Propagation:** Modules in the cascade operate independently, precluding joint optimization. Each module, designed to perform a specific function, introduces errors, denoted by e_i , where i represents the stage in the

Table 3: Supplement rate-distortion results. We added the module proposed in this article based on LiteISP and cascaded it with the image compression method to form a new cascade framework. The RealCamNet still has significant advantages.

Method	BD-Rate↓	BD-PSNR↑	BD-MSSSIM↑	BD-LPIPS↓	BD-DeltaE↓
RR-ISP(Ours)+VTM(H.266)	-22.099	2.1416	0.7971	-0.0093	-0.8042
RR-ISP(Ours)+TCM(CVPR'23)	-30.4073	2.2157	1.126	-0.005	-0.8113
RR-Codec(Ours)	-39.0842	2.9603	1.6392	-0.0162	-1.1709

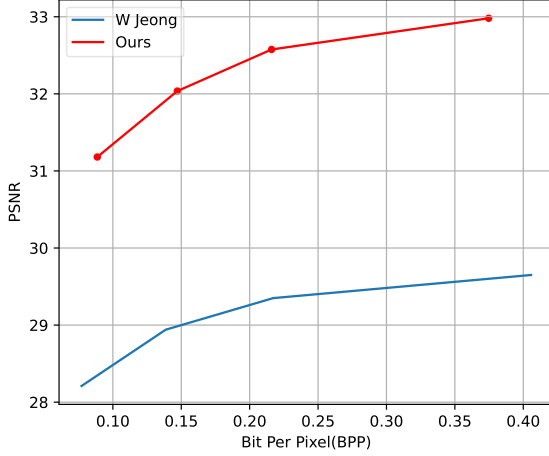


Figure 5: Comparison with the method proposed by W Jeong.

processing pipeline. These errors accumulate, impacting the final image quality:

$$I_{\text{final}} = \left(\prod_i f_i + e_i \right) (I_{\text{RGB}})$$

where I_{final} is the final processed image and f_i the operation of the i -th module.

- (3) **Computational Redundancy:** Each module typically consists of a neural network performing a task in three stages: encoding, operating in latent space, and decoding. This is represented as a combination of encoding E_i , Enhance operation O_i , and decoding D_i functions:

$$I_{\text{processed}} = E_1(D_1(E_2(D_2(\dots E_n(D_n(I_{\text{RGB}})) \dots)))$$

The multiple encodings and decodings to and from the latent space lead to computational redundancy, negatively impacting the efficiency of the pipeline.

These limitations highlight the need for an integrated approach that optimizes the image processing workflow as a unified task, thus mitigating noise amplification and computational overhead while enhancing image quality.

6 DETAIL OF COLOR PRIOR ENCODER CPE

We show the specific structures of the global color encoder and local color encoder in the CPE module in Figure 6. These two modules extract global color priors and local color priors respectively.

7 DETAIL OF CHANNEL-SPATIAL ATTENTION MODULE CSA

We show the detailed structure of the SWA module and CWRA module in Figure 7. These two modules perform channel attention and spatial attention enhancement respectively.

8 ANALYSIS OF RAWTOBIT

The RAWtoBit[6] method is designed to compress the RAW data into a bitstream, which is subsequently decoded and processed through the introduced RCAG[9] to perform the ISP process. The RAWtoBit framework intends to transport the RAW data to the decoder before executing the RAW-to-RGB conversion process. Compared with RGB images, RAW images have higher bit depth and wider color range, so it is not necessary to retain all the information in RAW.

Furthermore, the RAWtoBit scheme does not account for the complexity of the real-world imaging pipeline, manifesting a lack of modeling for various essential processes such as spatially variant distortions, tone mapping, and other device-specific characteristics.

In conclusion, our proposed method addresses these deficits by embodying a more comprehensive and accurate representation of the real-world end-to-end imaging pipeline. By integrating the necessary modeling of spatial distortions and tone mapping, we substantially enhance the fidelity of the processed images. We report experimental results of our method with previous RAWtoBit on real-world imaging tasks in Fig. 5, which further demonstrates the superior performance of the proposed end-to-end scheme.

9 SUBJECTIVE QUALITY COMPARISON

To verify the subjective quality of the proposed method, we use MUSIQ[7] and CLIPQA[2] to evaluate the results obtained by different methods and compare the performance of different methods through the Rate-Distortion curve. Detailed results are shown in Fig. 8.

10 MORE QUALITATIVE RESULTS

We show more subjective results in Fig. 9 and Fig. 10.

REFERENCES

- [1] [n. d.]. RawPy: Python library for raw image processing. <https://pypi.org/project/rawpy/>. Accessed: yyyy-mm-dd.
- [2] Firstname Author and Secondname Another. 2021. CLIPQA: No-reference Image Quality Assessment via CLIP. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [3] Vladimir Bychkovsky, Sylvain Paris, Eric Chan, and Frédo Durand. 2011. The MIT-Adobe FiveK Dataset for Raw Image Processing. In *Computational Photography (ICCP), 2011 IEEE International Conference on*. IEEE.

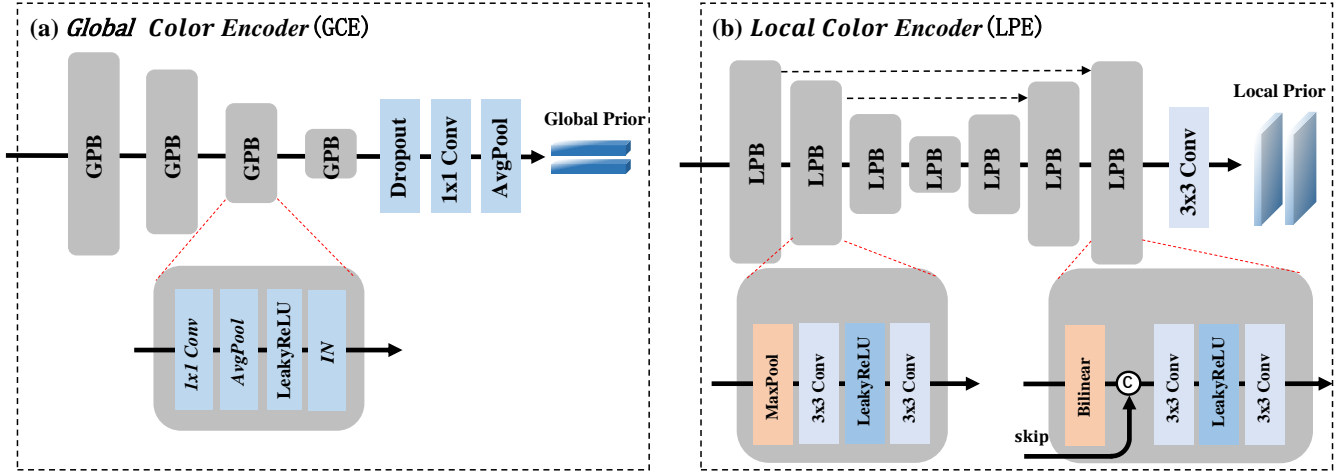


Figure 6: The structure of Global Color Encoder and Local Color Encoder.

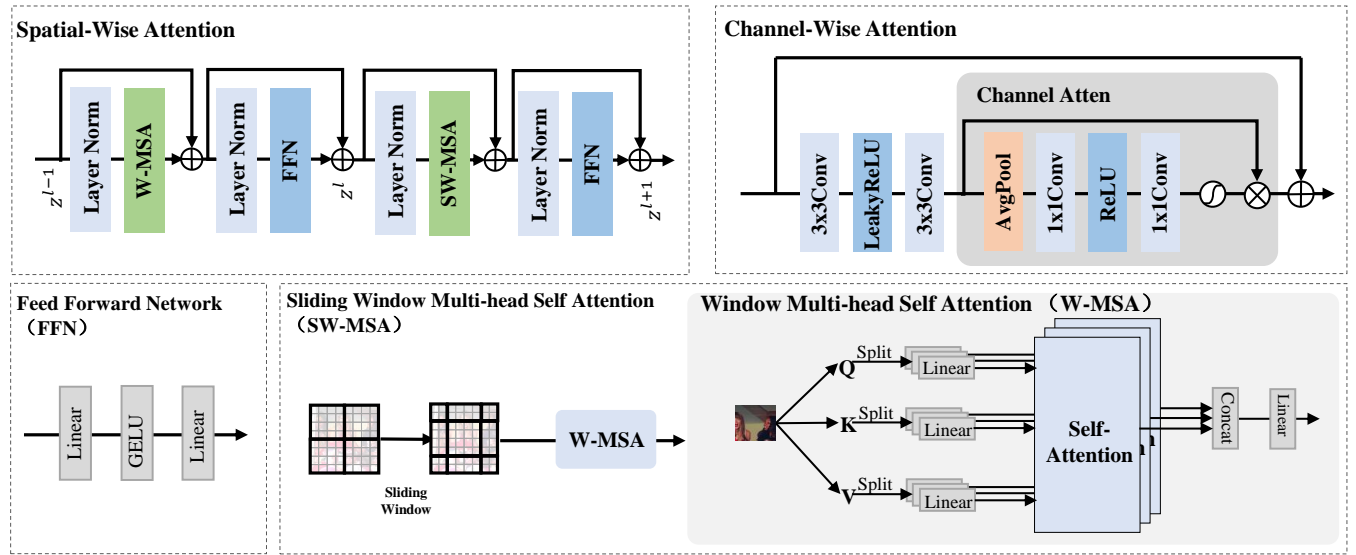


Figure 7: The structure of SWA and CWRA in CSA module.

- [4] Chen Chen, Qifeng Chen, Jia Xu, and Vladlen Koltun. 2018. Learning to See in the Dark. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
- [5] Andrey Ignatov, Radu Timofte, Zhilu Zhang, Ming Liu, Haolin Wang, Wangmeng Zuo, Jiawei Zhang, Ruimao Zhang, Zhanglin Peng, Sijie Ren, et al. 2020. Aim 2020 challenge on learned image signal processing pipeline. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part III* 16. Springer, 152–170.
- [6] Woosok Jeong and Seung-Won Jung. 2022. RAWtoBit: A Fully End-to-end Camera ISP Network. In *European Conference on Computer Vision*. Springer, 497–513.
- [7] Y. Ke, J. Li, J. Yang, and X. Wang. 2021. MUSIQ: Multi-scale Image Quality Transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.
- [8] David G Lowe. 2004. Distinctive image features from scale-invariant keypoints. *International journal of computer vision* 60 (2004), 91–110.
- [9] Zhilu Zhang, Haolin Wang, Ming Liu, Ruohao Wang, Jiawei Zhang, and Wangmeng Zuo. 2021. Learning raw-to-srgb mappings with inaccurately aligned supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4348–4358.

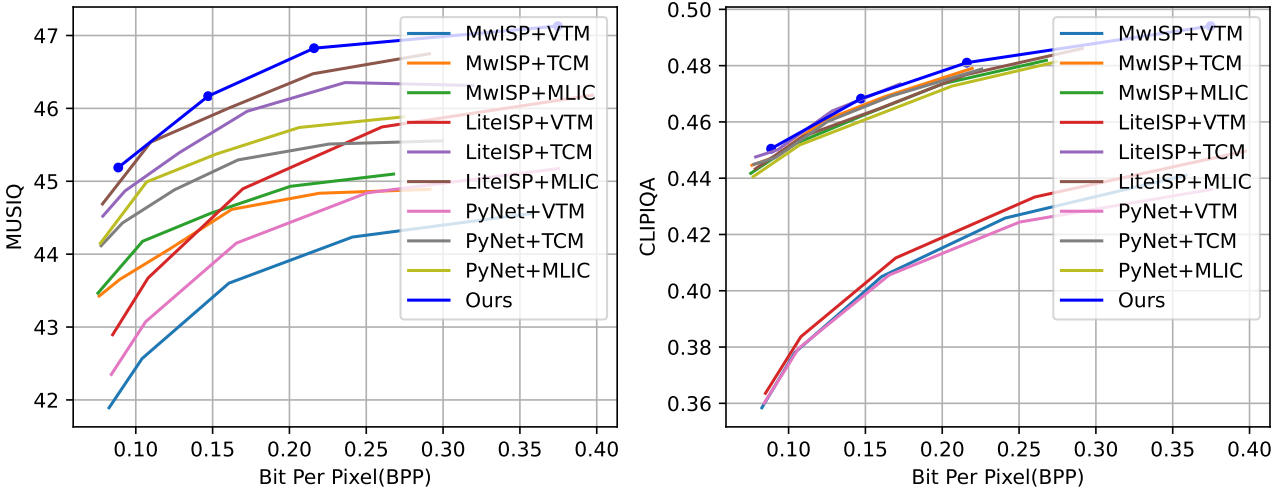


Figure 8: Subjective Quality Rate-Distortion Curve.

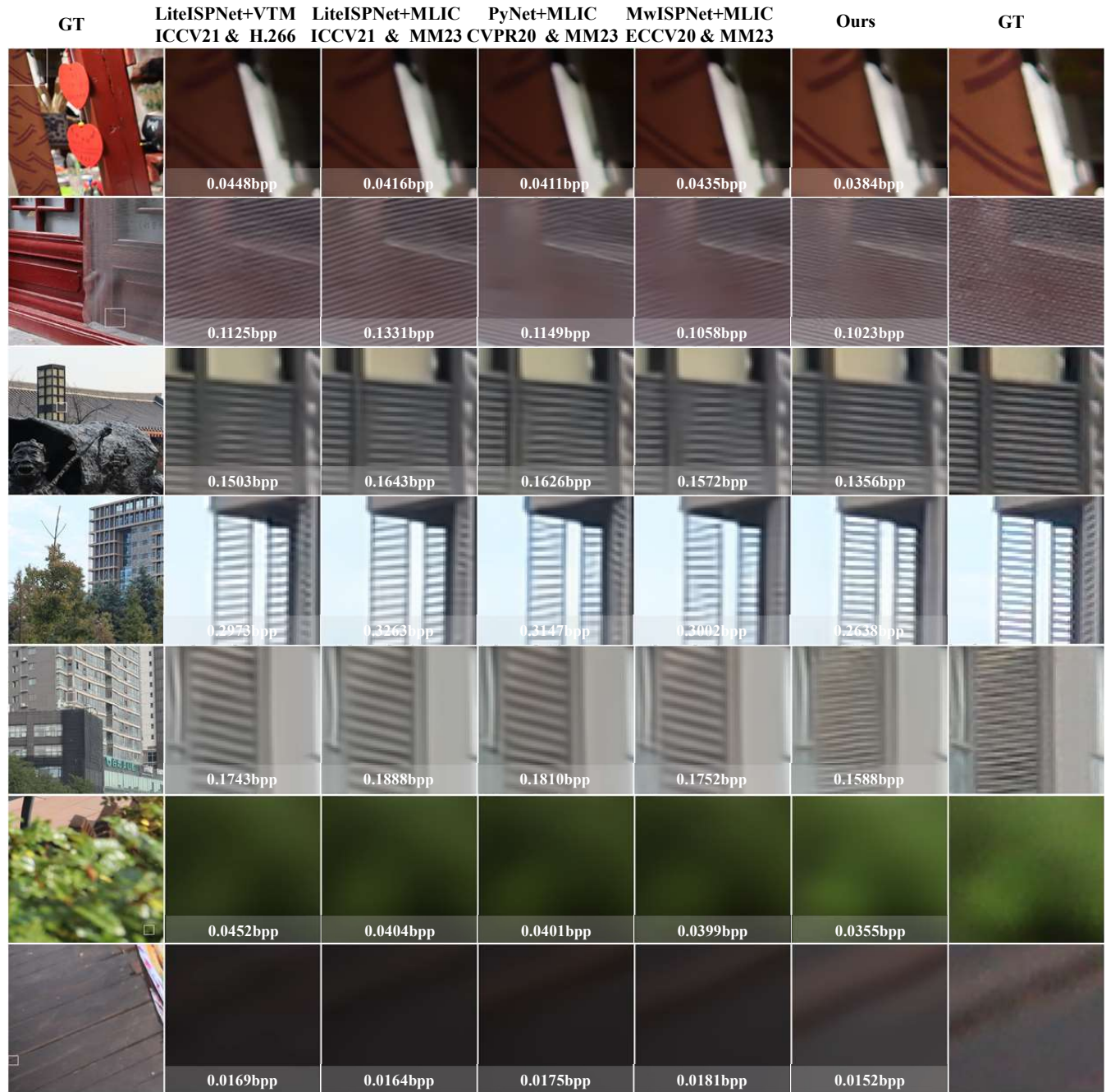


Figure 9: Qualitative Results (1).










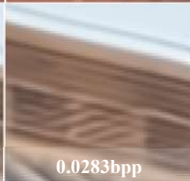
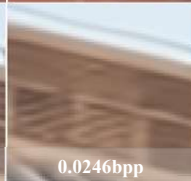

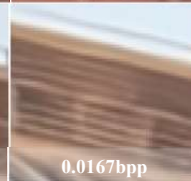


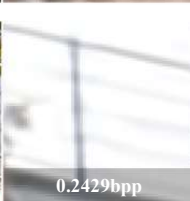
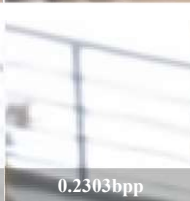
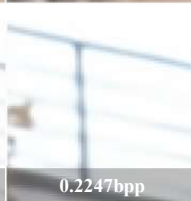
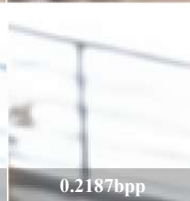



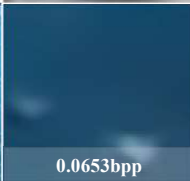
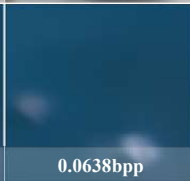
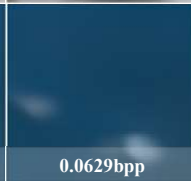
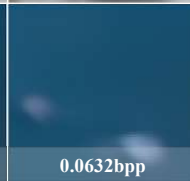
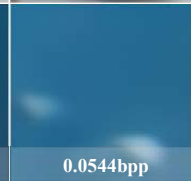


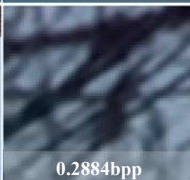
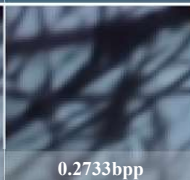
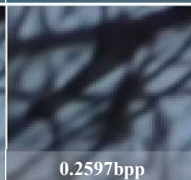
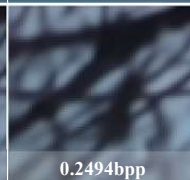
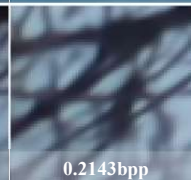


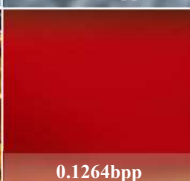
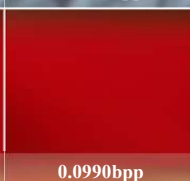
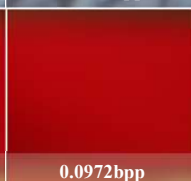
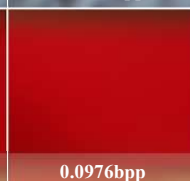
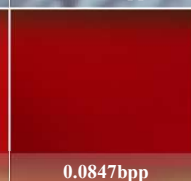


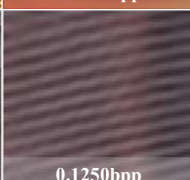
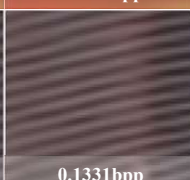

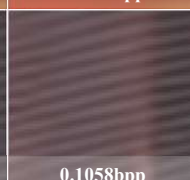
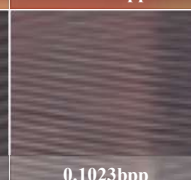

GT	LiteISPNet+VTM ICCV21 & H.266	LiteISPNet+MLIC ICCV21 & MM23	PyNet+MLIC CVPR20 & MM23	MwISPNet+MLIC ECCV20 & MM23	Ours	GT
	 0.1646bpp	 0.1613bpp	 0.1557bpp	 0.1501bpp	 0.1403bpp	
	 0.0172bpp	 0.0283bpp	 0.0246bpp	 0.0280bpp	 0.0167bpp	
	 0.2429bpp	 0.2303bpp	 0.2247bpp	 0.2187bpp	 0.1901bpp	
	 0.0653bpp	 0.0638bpp	 0.0629bpp	 0.0632bpp	 0.0544bpp	
	 0.2884bpp	 0.2733bpp	 0.2597bpp	 0.2494bpp	 0.2143bpp	
	 0.1264bpp	 0.0990bpp	 0.0972bpp	 0.0976bpp	 0.0847bpp	
	 0.1250bpp	 0.1331bpp	 0.1149bpp	 0.1058bpp	 0.1023bpp	

Figure 10: Qualitative Results (2).