Inverting Image Signal Processing Pipeline with Diffusion Models

Xinman Liu, Xuanchi Ren, Ziyi Wu

1 Motivation

For professional users, RAW images are usually preferred over RGB images since they contain unprocessed scene irradiance. Such information is desirable for attaining more plausible visual effects and various image editing tasks. Recently, researchers point out that RAW images are also valuable for computer vision tasks, such as image super-resolution, image denoising, and reflection removal [1], [2]. However, since RAW images are memory-intensive, saving a pair of RAW and RGB images is not feasible. To enable users to get access to the RAW one, inverting the RGB images to RAW images becomes an important problem in computational photography.

Inverting the image signal processing (ISP) is a challenging problem since the ISP is a lossy pipeline, converting 12 or 14-bit RAW data to 8-bit RGB data. Especially, the over-exposed regions totally lose the corresponding data, resulting in a harder inversion. Moreover, there is also a lossy image compression process in the digital camera to save the RGB in the JPEG format. The current state-of-the-art method, Invertible ISP [3], utilizes the normalizing-flow-based models [4] with a differentiable JPEG simulator. However, due to the capacity limitation of flow-based models, their performance still has room for improvement, especially for the over-exposed areas.

The emergence of denoising diffusion probabilistic models (DDPM) [5] inspires us to investigate inverting the ISP process with the diffusion models. With their incredible generative power, in this work, we seek to develop a fully end-to-end framework to directly synthesize 14-bit RAW images from the RGB images.

2 Related Work

RAW Image Reconstruction. There is a line of work researching reconstructing RAW images from RGB ones. Afifi et al. [6] propose to model the RAW recovery process with camera-independent CIE-XYZ color space. Zamir et al. [2] propose to model the RGB-RAW-RGB pipeline in a cycle manner. Invertible ISP [3] proposes leveraging the invertible neural network to merge RAW-to-RGB and RGB-to-RAW mapping in a single model. Unlike previous work, we target to utilize the generative power of diffusion models to compensate for the gap between RAW data and RGB data, especially for the over-exposed part.

Diffusion Models. With the success of diffusion models in density estimation and sample quality, it is also introduced into the domains of images [7], video [8], and 3D [9] with an underlying neural backbone as a UNet [10]. The generative process of diffusion models is formulated as an iterative denoising procedure [5] with many variants. While researchers demonstrate that they can achieve amazing performance in these domains, there is no work combining RAW image reconstruction with diffusion models to fully dig into the potential of their generative power for the missing data.

3 Overview

Given an RGB image $X \in \mathbb{R}^{H\times W\times 3}$ as the condition, the learned diffusion model iteratively refines a noise $Y_T \sim \mathcal{N}(0, I)$ to Y_{T-1}, Y_{T-2}, ..., Y_0. And the final output of the diffusion model $Y_0 \in \mathbb{R}^{H\times W\times 3}$ is the demosaicked image. Then we apply the Bayer sampling function f_{Bayer} to attain the target RAW image $Z \in \mathbb{R}^{H\times W\times 1}$ by:

$$Z = f_{Bayer}(Y_0). \tag{1}$$

The architecture of our diffusion model is U-Net, and we follow [11] to train our model.

Moreover, due to the demand for high GPU resources of diffusion models, we are also interested in implementing a version of Latent Diffusion Models [10], which conducts the diffusion process in the latent space for the RAW image reconstruction. We plan to first get autoencoder models for RGB data and demosaicked images by fine-tuning the models provided by Stable Diffusion[1]. Then we can bridge the two latent spaces through the diffusion process.

4 Milestone

Our plan is as follows chronologically:

- We replicate the main baseline Invertible ISP [3] by run their training code by 11.20.
- We follow [11] to train a diffusion model to reconstruct RAW images from RGB images in terms of resolution 128×128 by 11.30.
- We are eager to extend our model to a latent diffusion version to enable higher-resolution synthesis, e.g., 256×256 by 12.05.

- *X. Liu (1004370637), X. Ren (1009173403), and Z. Wu (1007807526) are with University of Toronto.*
- *Authors are listed alphabetically.*

[1] https://github.com/CompVis/stable-diffusion

References

[1] X. Zhang, Q. Chen, R. Ng, and V. Koltun, “Zoom to learn, learn to zoom,” in *CVPR*, 2019.

[2] S. W. Zamir, A. Arora, S. H. Khan, M. Hayat, F. S. Khan, M. Yang, and L. Shao, “Cycleisp: Real image restoration via improved data synthesis,” in *CVPR*, 2020.

[3] Y. Xing, Z. Qian, and Q. Chen, “Invertible image signal processing,” in *CVPR*, 2021.

[4] L. Dinh, J. Sohl-Dickstein, and S. Bengio, “Density estimation using real NVP,” in *ICLR*, 2017.

[5] J. Ho, A. Jain, and P. Abbeel, “Denoising diffusion probabilistic models,” in *NeurIPS*, 2020.

[6] M. Afifi, A. Abdelhamed, A. Abuolaim, A. Punnappurath, and M. S. Brown, “CIE XYZ net: Unprocessing images for low-level computer vision tasks,” *T-PAMI*, 2022.

[7] P. Dhariwal and A. Nichol, “Diffusion models beat gans on image synthesis,” *NeurIPS*, vol. 34, pp. 8780–8794, 2021.

[8] J. Ho, T. Salimans, A. Gritsenko, W. Chan, M. Norouzi, and D. J. Fleet, “Video diffusion models,” *arXiv preprint arXiv:2204.03458*, 2022.

[9] B. Poole, A. Jain, J. T. Barron, and B. Mildenhall, “Dreamfusion: Text-to-3d using 2d diffusion,” *arXiv preprint arXiv:2209.14988*, 2022.

[10] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, “High-resolution image synthesis with latent diffusion models,” in *CVPR*, 2022.

[11] C. Saharia, J. Ho, W. Chan, T. Salimans, D. J. Fleet, and M. Norouzi, “Image super-resolution via iterative refinement,” *T-PAMI*, 2022.