Learned Reverse ISP with Soft Supervision

Beiji Zou[1,2,✉] and Yue Zhang[1,2]

[1] School of Computer Science and Engineering, Central South University, Changsha 410083, China
[2] Hunan Engineering Research Center of Machine Vision and Intelligent Medicine, Changsha 410083, China
{bjzou,yuezhang}@csu.edu.cn

Abstract. RAW image serves as the foundation for camera imaging, which resides at the very beginning of the pipeline that generates sRGB images. Unfortunately, owing to special considerations, the information-rich RAW images are forfeited by default in most existing applications. To regain the RAW image, some works attempt to restore RAW images from RGB images. They focus on designing handcrafted model-based methods or complicated networks, however, ignoring the special property of RAW image, *i.e.*, high dynamic range. To make up for this deficiency, we introduce a novel soft supervision, derived from the high dynamic range. Specifically, we propose to soften the original ground-truth as a multivariate Gaussian distribution so that networks could learn much more information. Then, we introduce a soft supervision driven network (SSDNet), based on convolution and transformer, for effectively restoring RAW images from RGB images. Quantitative and qualitative results show the promising restoration performance of RGB-to-RAW. In particular, our method achieved fifth place in the S7 track of AIM Reversed ISP Challenge. The source code will be available at `https://github.com/yuezhang98/Learned-Reverse-ISP-with-Soft-Supervision`.

Keywords: Reversed ISP · Soft Supervision · Convolution · Transformer

1 Introduction

RAW image is single-channel raw data obtained from CMOS, which is linearly related to the scene irradiance. It is worth noting that the RAW image usually has a high dynamic range of at least 2^{10}, which provides more information and is helpful for image processes. However, due to various limitations, such as storage space, most devices have to give up saving RAW images and keep the sRGB image instead processed by the image signal processor (ISP). As a consequence, this complicates some important operations, such as image denoising, image enhancement, HDR and super-resolution [1,17,28,30,36,46]. For example, for denoising in the RAW domain, the statistical characteristics of the noise can be obtained by calibration, resulting in high performance noise reduction and $\frac{1}{3}$ less computation. Meanwhile, in the RGB domain, multiple nonlinear operations

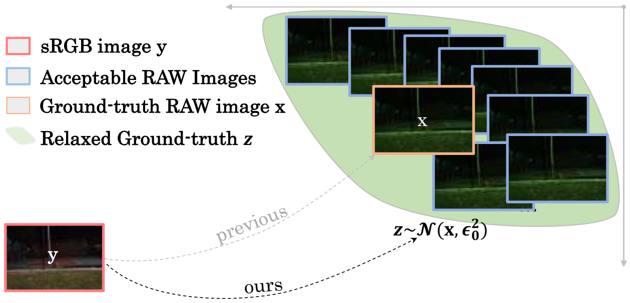

Fig. 1. The proposed soft supervision for learned reversed ISP. Previous methods usually enforce the network to learn the given ground-truth, while our method enables a wide range of acceptable supervision sampled from the relaxed multivariate distribution.

inside the ISP make the noise complex and more difficult to remove. Therefore, in this work, we focus on RGB-to-RAW mapping and enable it to be beneficial for image denoising.

Attracted by the beneficial attributes of RAW images, some works carry out the research on RGB to RAW images. Pioneeringly, Nguyen et al. [31,32] propose a fast breadth-first-search octree algorithm for finding the necessary control points to provide a mapping between the RGB and RAW colour spaces. Then, they demonstrate better practicality by using the recovered RAW for white balance correction and image deblurring. However, such simplified mapping ignores the complex processing flow inside the ISP, such as demosaicing, tone mapping, etc. To bridge this gap, Brooks et al. [5] propose a generic camera ISP process with five typical stages, where each step is approximated by an invertible function. This approach is fully interpretable and makes the ISP flow design more flexible. They apply the proposed algorithm in image denoising. Although it provides a better idea involving ISP, the process of ISP in practice is often more sophisticated and the limited parameters of the interface open to the public also restrict related research.

Recently, complementary learning-based approaches [34,51,48,13] have been proposed to alleviate this challenge. CycleISP [51] considers *cycle consistency* to learn the forward (RAW to RGB) and reverse (RGB to RAW) directions of ISP. They employ two different networks, which are trained end-to-end, to effectively assist the denoising task using converted RAW images. In the meantime, InvISP [48] enables RAW to RGB and RGB to RAW mapping by building on a single normalizing-flow-based invertible neural network [25]. However, both of them require a large amount of training data, which is a challenge because such datasets are difficult to collect. To address this problem, Conde et al. [13] propose a hybrid approach based on dictionary learning to achieve effective RGB to RAW, which maintains the advantages of both model-based and end-to-end learnable approaches. Unfortunately, these methods only focus on model design, in terms of either restoration performance or light-weight model, while discounting the characteristics of RAW image, namely the high dynamic range.

In this work, we introduce a novel soft supervision driven network (SSDNet) for RGB-to-RAW mapping. Specifically, we utilise the idea of partitioning to explore the RGB to RAW task systematically from the data and model perspectives respectively. Firstly, we pick up the missing piece of data trait, namely the high dynamic range, and propose to relax the given supervision. Although the high dynamic range of RAW allows for a more detailed representation of the data, it may increase the difficulty of network learning. To alleviate this problem, we propose to relax the original supervision to a multivariate Gaussian distribution, as shown in Figure 1. This brings two benefits: 1) the learning goal of the network is softened, simplifying the learning process; 2) the network could capture pixels around the given supervision during the training process, subtly performing data augmentation. We propose new loss functions based on the soft supervision that enable the network to further cope with the learning of multivariate Gaussian distributions. Secondly, we construct an encoder-decoder structure that incorporates convolution and Transformer [44]. As is well known, the early prevalent convolutional networks have a superior ability to perceive local structure. Besides, according to previous works [27,43,35], large receptive fields are favourable for image reconstruction. Inspired by the recent Transformer architecture [19,50], we made an organised combination of convolution and Transformer blocks, allowing the network to capture not only local details, but also the overall image pattern. Equipped with these two components, our SSDNet achieves state-of-the-art RGB-to-RAW mapping in Adobe FiveK dataset [48]. In particular, our method achieved fifth place in the S7 track of AIM Reversed ISP Challenge [14]. In addition, we have also demonstrated the effectiveness of the proposed method on image denoising.

In conclusion, our main contributions are as follows:

- We present a new learning target for RGB-to-RAW mapping. The proposed soft supervision enables SSDNet to make effective use of data characteristics, resulting in better recovering performance.
- We present a novel network, SSDNet, that effectively fuses convolution and transformer, which is able to capture local and global feature for better restoration performance.
- Extensive experiments show that our method achieves state-of-the-art RAW restoration from RGB image. Our method also presents as a top solution in the novel AIM Reversed ISP Challenge[14]. Additionally, our methods could also benefit the downstream task, *e.g.*, image denoising.

2 Related work

2.1 Reversed ISP

ISP aims to convert the RAW data acquired from CMOS into natural RGB images. It involves a wide variety of designs and complex non-linear processes.

Therefore, RAW-to-RGB is an irreversible operation [48]. Furthermore, RAW images in practice applications are usually not preserved by default, due to their large amount of data. As technology evolves, realistic scenarios demand higher quality imaging via ISPs [16]. In particular, it is difficult to make considerable enhancements to an already converted RGB image. For this reason, some work has been initiated to investigate the mapping of RGB to RAW, with the expectation of enabling enhancement on RAW to acquire high quality images in the end. Nguyen et al. [31,32] proposed a fast breadth-first-search octree algorithm to provide a mapping between the RGB and RAW colour spaces. Then, Brooks et al. [5] complemented their research by taking ISP's internal modules into account. However, as ISP flows were inherently irreversible, this substitution of reversible functions for each of the ISP modules was inevitably introducing unreasonable errors. Fortunately, the rapid development of deep learning offers new ideas for this research [34,51,48,13]. For example, CycleISP [51] and InvISP [48] implemented RGB to RAW and RAW to RGB using deep networks. Yet, the desire for large amounts of training data limits their application. To address this problem, Conde et al. [13] proposed to use dictionary learning methods to replace important modules inside ISP, maintaining the advantages of both model-based and end-to-end learnable approaches. These studies of RGB to RAW have achieved promising results, in that they have neglected the special characteristics of RAW data, namely the high dynamic range. In this work, we shall pick up this piece and exploit it to enhance the performance of our proposed network.

2.2 Image Denoising

Image denoising is a fundamental and critical task, being required by various imaging systems as well as the pre-processing step for high-level vision algorithms. In recent years, deep learning-based image denoising has become popular in current research. The work on image denoising in growing numbers prefers deep CNNs and has achieved remarkable performance. DnCNN [53] utilized a convolutional neural network with 17 layers equipped with Batch Normalization and the concept of residual learning into image denoising. Its denoising performance outperforms all conventional algorithms, including BM3D [15]. This shows the immense potential of the CNN model in image denoising. Subsequently, more work focused on designing a wider and deeper network structure to achieve better denoising performance [54]. However, most of the previous work is designed for additive Gaussian white noise (AWGN), which is too simple and deviates from real scenes. To bridge this gap, Guo et al. [20] proposed CBDNet, which simulates various camera response functions. RIDNet [3] further introduced the feature attention mechanism to achieve impressive denoising performance on both synthetic and real-world datasets. Yue et al. [49] proposed Variational Denoising Network (VDN), exploiting the variational inference technique that enables VDN to simultaneously learn noise distributions. They try to enhance the performance of VDN by leveraging accurate noise information. In the account of the further prevalence of attention mechanisms, MIRNet [52] learned rich features through

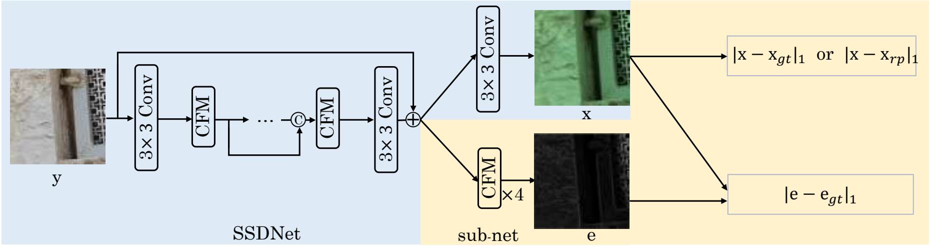

Fig. 2. The learning scheme for the proposed soft supervision. For an existed RGB-to-RAW network, we attach a sub-network in the end to learn the variance term in the proposed soft supervision. Practically, the sub-network is constructed by a four-CFM (Context Fusion Module), introduced in Section 3.2.

parallel multi-resolution convolutional flow as well as spatial attention and channel attention mechanisms. This network demonstrates the further potential of attentional mechanisms on image denoising. NBNet [12] further improved the denoising performance based on this prior as well.

In contrast to direct denoising in RGB images, we expect to achieve an earlier stage of noise reduction using RGB to RAW techniques. There are similar works to ours, CycleISP [51] and UPI [5]. But they ignore the attribute of RAW image, and this work is going to bridge that gap.

3 Proposed Method

3.1 Soft Supervision

We first exploit the characteristics of the training data, especially the high dynamic range. In contrast to previous methods [34,51,48,13] that struggle to allow the network to regress the established ground-truth, we propose to learn the soft supervision.

Given an sRGB image $\mathbf{y} \in \mathbb{R}^{\mathrm{W}\times\mathrm{H}\times 3}$, we apply SSDNet to get the reversed RAW image $\mathbf{x} \in \mathbb{R}^{\mathrm{W}\times\mathrm{H}}$. At the same time, we can also obtain the corresponding absolute error map $\mathbf{e}_{gt} = |\mathbf{x} - \mathbf{x}_{gt}|$, where $\mathbf{x}_{gt} \in \mathbb{R}^{\mathrm{W}\times\mathrm{H}}$ is the ground-truth. In fact, the error map contains additional information about the ground-truth, usually being neglected in previous methods. To further exploit this information, we relax the original ground-truth in conjunction with this error map into a multivariate Gaussian distribution, as soft supervision $\mathbf{x}_{sp}$, so that the network can fully capture the supervision information.

$$\mathbf{x}_{sp} \sim \mathcal{N}(\mathbf{x}_{gt},\ \mathbf{e}_{gt}) \tag{1}$$

Eq.1 contains all the supervisory information, even the recoverability of the network itself. To learn $\mathbf{x}_{sp}$, a simple approach is attaching a sub-network in the network, as shown in Figure 2, and then using L1 loss enable the whole network to learn the mean and variance, i.e. the given ground-truth $\mathbf{x}_{gt}$ and the error map $\mathbf{e}_{gt}$.

$$\arg\min\ \mathcal{L} = \|\mathbf{x} - \mathbf{x}_{gt}\|_1 + \|\mathbf{e} - \mathbf{e}_{gt}\|_1 \tag{2}$$

where $\mathbf{e}$ is the output of the sub-network. However, this approach is not really taking advantage of the learned distribution. For this reason, we propose a **testing-resampling** strategy to enhance the restoration results. Specifically, we sample 50 instances in $\mathcal{N}(\mathbf{x},\, \mathbf{e})$ and then average them to get the final results. Obviously, this approach brings additional computation costs.

In contrast, we provide another **training-resampling** strategy to eliminate the above problems. We resort to the reparameterization trick [24,22] to resample the supervision in the distribution as new supervision, $\mathbf{x}_{rp}$.

$$\mathbf{x}_{rp} = \mathbf{x}_{gt} + \mathbf{z} * \mathbf{e}_{gt},\ \ \mathbf{z} \sim \mathcal{N}(\mathbf{0},\ \mathbf{1}) \tag{3}$$

In this way, SSDNet would capture not a given supervision during training, but a large number of possible samples existing in the soft supervision $\mathbf{x}_{sp}$. Practically, we utilize the following loss function:

$$\arg\min\ \mathcal{L} = \|\mathbf{x} - \mathbf{x}_{rp}\|_1 + \alpha * \|\mathbf{e} - \mathbf{e}_{gt}\|_1 \tag{4}$$

where α is the hyper-parameter to balance learning targets. During training, we would attach a sub-network on the main network, the same as Figure 2. While testing, the sub-network could be simply deprecated for saving memory.

3.2 SSDNet

Throughout the existing RGB-to-RAW models, we note that the existing reversed ISP models are pure CNNs [34,51,48,13], which have a limited receptive field that suffers from the limited restoration performance. Recently, transformer-based image restoration networks [50,10] have also demonstrated impressive performance benefiting from their global receptive fields, although they are again limited by weak local modelling capabilities [9,19,21]. To address the shortcomings of these two architectures, we propose a hybrid module based on the large kernel convolution and self-attention, the context fusion module(CFM). Then, we take CFM as a basic block to build SSDNet. To illustrate it clearly, we first introduce the overall network and then elaborate on the proposed CFM.

Overall Architecture. The overview architecture is shown at the top of Figure 3. The SSDNet is based on a symmetrical UNet [38] architecture. SSDNet has three encoder stages and three corresponding decoder stages. Given an RGB image $\mathbf{y} \in \mathbb{R}^{\mathrm{W}\times\mathrm{H}\times 3}$, the network first applies a 3×3 convolutional layer to project the image into feature space. Then, SSDNet first encodes the projected features. Particularly, at the end of each encoder, the feature maps, $\mathbf{X} \in \mathbb{R}^{\mathrm{H}\times\mathrm{W}\times\mathrm{C}}$, are downsampled to $\frac{1}{2}\times$ scale, $\mathbf{X} \in \mathbb{R}^{\frac{1}{2}\mathrm{H}\times\frac{1}{2}\mathrm{W}\times 2\mathrm{C}}$, with a 3×3 convolutional layer and PixelShuffle [41]. After encoding, the final high-level features are decoded to the original feature space. More precisely, the feature maps are up-sampled to $2\times$ scale with a 3×3 convolutional layer and PixelUnShuffle operation before

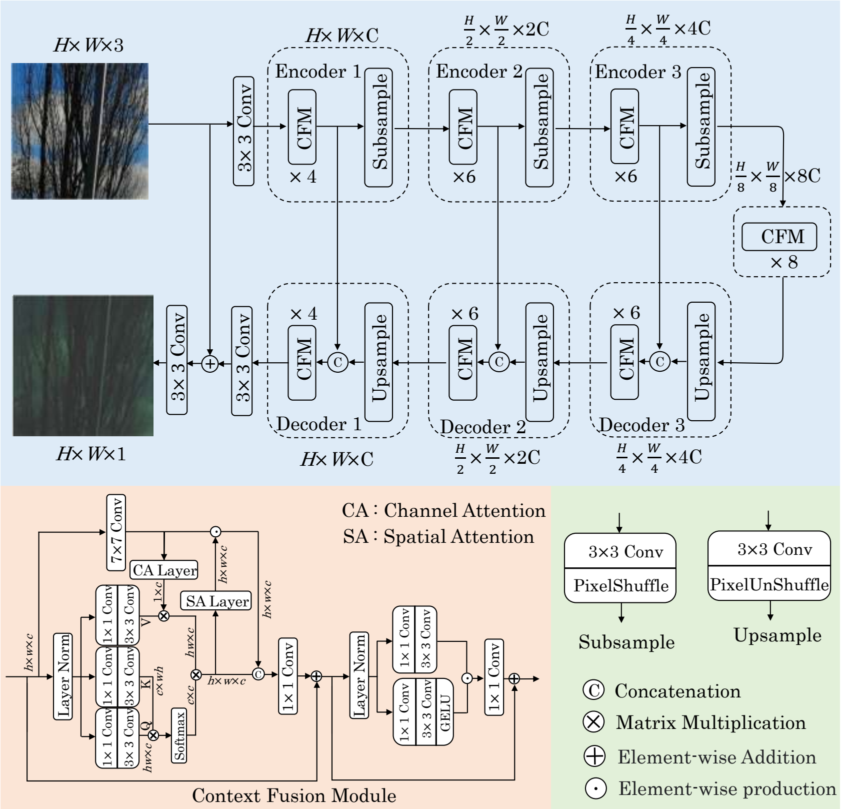

Fig. 3. The architecture of the proposed SSDNet. The proposed SSDNet has a multi-scale hierarchical design and it is made up of the Context fusion module (CFM). The CFM consists of an efficient mixture of the convolution and Transformer blocks as well as the simplified Gated-Dconv Feed-Foreard Network (GDFN) [50].

each decoder stage. This layer reduces the feature channels by half, identical to the inverse operation in the encoder stage. After that, skip-connection passes the low-level feature maps from the corresponding encoder stage. The up-sampled feature and the encoder's feature jointly compose the input to the matching decoder. Then, the recovered feature is re-projected to the RGB image space, $\mathbf{r} \in \mathbb{R}^{\mathrm{W}\times\mathrm{H}\times 3}$, by another 3×3 convolutional layer. An RGB residual map is obtained by $\mathbf{y}_r = \mathbf{y} + \mathbf{r}$. At last, a 3×3 convolutional layer is applied to compact the RGB residual map $\mathbf{y}_r$ and output a single-channel RAW image, $\mathbf{x} \in \mathbb{R}^{\mathrm{W}\times\mathrm{H}\times 1}$. Its basic building blocks of the encoder and decoder follow the same Context Fusion Module (CFM).

Context Fusion Module. As shown in the left bottom of Figure 3, CFM has two stages, respectively the convolution-transformer mixed module and the feed forward network. For the first part, we follow Restormer and employ Multi-Dconv Head Transposed Attention (MDTA) to reduce the computational burden of self-attention. The MDTA performs self-attention in the channel dimension, and this coarse-grained technique may lead to poor feature extraction. Thus, they add depth-wise convolution after the attention layer to alleviate this problem. Furthermore, we propose to enhance MDTA by extracting the input features in parallel using a large kernel (7 × 7) convolution and then fuse the output of MDTA and large kernel convolution. This simple operation can effectively enhance the full CFM.

Specifically, for a layer normalized tensor $\mathbf{Y} \in \mathbb{R}^{H\times W\times C}$, MDTA first produces *query* ($\mathbf{Q}$), *key* ($\mathbf{K}$) and *value* ($\mathbf{V}$). Then, the query and key would be transposed such that their dot-product interaction could generate a smaller attention map of size $\mathbb{R}^{C\times C}$. At the same time, we enhance the input future with another accompanied 7 × 7 convolution. In particular, we bring the local modelling capability of self-attention based on the dynamic weights of the convolution operation in the channel dimension [23]; and then integrate the global modelling capability for the convolution operation by generating global dynamic weights on the space based on the self-attention mechanism. The general process of enhanced MDTA are as follows:

$$\begin{aligned}
&\hat{\mathbf{Y}} = f(\text{Attention}\,(\mathbf{Q},\mathbf{K},\mathbf{V}_c),\ \mathbf{Y}_s) + \mathbf{Y},\\
&\text{Attention}\,(\mathbf{Q},\mathbf{K},\mathbf{V}_c) = \mathbf{V}\cdot\text{Softmax}(\mathbf{K}\cdot\mathbf{Q}),\ \mathbf{V}_c = w_c \times \mathbf{V}\\
&\mathbf{Y}_s = w_s \times \text{Conv}(\mathbf{Y})
\end{aligned} \tag{5}$$

where $\mathbf{Y}$ and $\hat{\mathbf{Y}}$ are the input and output feature maps; f is the concatenate fusion operation, followed a 1 × 1 convolution to compress the channel; $\mathbf{Q} \in \mathbb{R}^{HW\times C}$, $\mathbf{K} \in \mathbb{R}^{C\times HW}$ and $\mathbf{V} \in \mathbb{R}^{HW\times C}$ projections are reshaped from the original size $\mathbb{R}^{H\times W\times C}$. $\mathbf{V}_c$ and $\mathbf{Y}_s$ are respectively enhanced by the dynamic channel weight w_c and the dynamic spatial weight w_s.

For the second part, feed-forward network, we simplify the original Gated-Dconv Feed-Forward Network (GDFN) [50]. Specifically, we take compressed depth-wise convolutions to reduce the computational cost. Overall, the computational process are as follows:

$$\begin{aligned}
\widetilde{\mathbf{Y}} &= \text{Gating}\left(\hat{\mathbf{Y}}\right) + \hat{\mathbf{Y}},\\
\text{Gating}(\hat{\mathbf{Y}}) &= \text{GELU}(W_c(\text{LN}(\hat{\mathbf{Y}})))\odot W_c(\text{LN}(\hat{\mathbf{Y}})),
\end{aligned} \tag{6}$$

where $\odot$ is element-wise multiplication, W_c represents the compressed convolution, and LN denotes the layer normalization [4]. Equipped with EMDTA (enhanced EMDA) and SGDFN (simplified GDFN), CFM could effectively capture the fine-grained details. Especially, for soft supervision learning, we apply a sub-network with four-CFMs attached on SSDNet, as shown in Figure 2.

4 Experimental Results

In this section, we perform extensive experiments to quantitatively and qualitatively verify the effectiveness of our method to perform RAW image reconstruction. Two objective metrics were used in the quantitative evaluation, including

Peak Signal to Noise Ratio (PSNR) and Structural Similarity (SSIM) [45]. PSNR computes the peak signal power to reconstruction error power ratio, while SSIM measures the structural similarity between the reconstructed image and the supervised image. Without loss of generality, we employ RGGB in the Bayer CFA pattern for all our experiments. Certainly, our algorithm can be easily extended to other Bayer patterns, such as RGBG. In addition, we have also conducted experiments on a downstream task, RAW image denoising, to verify the usefulness of the learned inverse ISP model beyond RAW image reconstruction.

4.1 Experimental setup

Datasets

AIM Reversed ISP-S7 dataset [14]. AIM Reversed ISP Challenge aims at creating a solution for recovering the camera's RAW with the corresponding RGB images processed by the in-camera ISP. The competition expects that such a solution should generate reasonable RAW images and by doing so, other downstream tasks, such as denoising, super-resolution or colour invariance, can benefit from the generation of such synthetic data. To this end, the competition organisers have collected a number of pair pairs (4320 for training, 480 for testing), taken from a Samsung S7 phone [40]. Especially, the RAW images were captured by IMX260 sensor with GRBG Bayer pattern.
MIT-Adobe FiveK dataset. We utilise the training-test data from [48], which were collected from the MIT-Adobe FiveK dataset [8] for the Canon EOS 5D subset (777 image pairs) and the Nikon D700 subset (590 image paris). This dataset randomly divides the two sets of data (Canon, Nikon) into training and test groups in a ratio of 85:15, respectively. Following [48,13], we use the LibRaw library to render ground-truth sRGB images from the RAW images.
SIDD dataset [1]. Real noise in real scenes is more challenging to deal with, therefore this dataset provides realistic noisy images, including both RAW sensor data and sRGB data. These images were captured by five smartphone cameras, under different lighting conditions, poses and ISO levels. The collected images have more noise compared to DSLR images and are due to the small aperture and sensor limitations. This dataset is available in 320 pairs of ultra-high resolution images (e.g. 5328 x 3000) for training and 1280 pairs of images for validation. This work takes SIDD [1] to investigate the application of learned RAW image reconstruction.

Implementation details

We train our SSDNet in an end-to-end manner and do not perform any pre-training process. In the training stage, we use AdamW [37] optimizer with momentum terms (0.9, 0.999). All the models are trained with 276,000 iterations. The initial learning rate is 3×10^{-4}, and it remains constant for the first 92,000 iterations and decreases to 1×10^{-6} for the next 184,000 iterations with the cosine

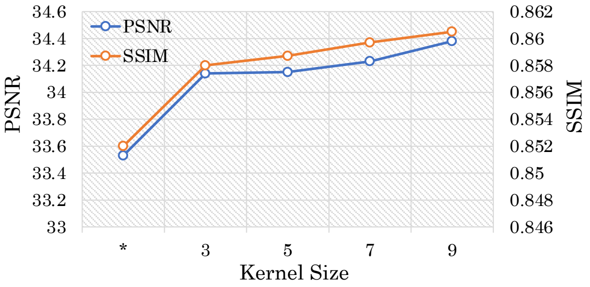

Fig. 4. Ablation study on the convolutional branch in CFM, especially the influence of its kernel size. $'*'$ means CFM does not incorporate the convolutional branch and the two interactions.

annealing strategy. We apply the random rotation and flipping for data augmentation. We train the network using a progressive patch size growing strategy [50]. Specifically, the network is initially trained with 128x128 image patches. During the iterations, the patch size is increased in steps of 64 to $(192, 256, 320, 384)$ in $(92K, 156K, 204K, 240K)$ iterations. In addition, to save memory, the batch size is reduced from the initial 64 to $(32, 16, 8, 8)$ in line with the increasing patch size. Especially, for **Training-resampling** strategy, we empirically set $\alpha = 0.001$. Our experiments are conducted on 8× 3090 Ti GPUs, typically taking about 4 days to complete the training.

4.2 Ablation Studies

Convolutional branch in CFM. Here, we explore the influence of the convolutional branch in CFM, including its kernel size. As shown in Figure 4, the convolutional branch brings at least 0.5dB enhancement. Furthermore, our model yields better performance as the convolutional kernel size increases[3]. We consider that the larger the convolution kernel, the larger the effective receptive field [29], which results in more reliable dynamic weights. To balance computational cost and performance, we finally choose the 7×7 kernel size.

Interactions in CFM. We verify the importance of the interaction operations, respectively the channel attention and spatial attention for Self-attention and Convolutional branches. As shown in Table 1, both of these interactions are effective in improving the restoration performance. Top performance is achieved in combination, with a 0.7dB improvement over the baseline, which demonstrates the importance of interactions. Especially, the spatial interaction shows greater importance than channel interaction, which means introducing local information modelling capability to the self-attention module is much easier and more effective than introducing global modelling capability to the convolution. We suppose

[3] Due to the limitation of computational resources, we have applied the convolution operation with maximum 9×9 kernel size.

Table 1. Ablation study on the interactions in CFM with the kernel size 7 in the convolutional branch. CA means channel attention, SA means spatial attention. 0^{st} model is the baseline without convolutional branch and any interaction.

Interactions	0^{st}	1^{st}	2^{nd}	3^{rd}	4^{th}
CA?	-	✗	✓	✗	✓
SA?	-	✗	✗	✓	✓
PSNR	33.53	33.89	34.03	34.18	**34.23**
SSIM	0.8520	0.8576	0.8580	0.8594	**0.8596**

Table 2. Ablation study on the Soft Supervision. $'*'$ means training and testing models in a traditional fashion. 50 or 100 in **Test-resampling** indicates that 50 or 100 instances are sampled to average.

Strategy	*	Test-resampling(50)	Test-resampling(100)	Training-resampling
PSNR/SSIM	34.23/0.8596	34.94/0.8651	**34.95/0.8651**	34.45/0.8624

that the self-attention module has a larger potential for adaptation than convolution.

Soft Supervision. Soft supervision is another critical element in delivering the effectiveness of SSDNet. This section focuses on verifying its validity, involving two sampling methods, namely Testing-resampling and Training-resampling. As shown in Table 2, our soft supervision brings at least 0.22dB gain for RAW restoration. The Testing-resampling strategy performs better than the Training-resampling strategy. We assume that this gap is derived from the robustness of an ensemble-like **Test-resampling** strategy (it averages 50 outputs as the final restored image). Since this strategy imposes an additional computational burden that is not as simple and elegant as the end-to-end **Training-resampling** strategy. Therefore, we adopt the training-resampling strategy by default in subsequent experiments.

4.3 RAW Image Reconstruction

In this section, we show the experimental results on RAW image reconstruction, especially the performance on AIM Reversed ISP Challenge [14] and a general benchmark, MIT-fiveK [48,8].

AIM Reversed ISP Challenge
We submitted a result obtained via the proposed method to the AIM Reversed ISP challenge [14]. In order to exploit the potential performance of our method to the maximum, we simultaneously employed self-ensemble strategy [26] during the testing phase. As shown in Table 3, our final submitted model achieved

Table 3. Quantitative RAW Reconstruction results at the final test sets of AIM Challenge on Reversed ISP - Track1 S7 [14]. All our results are boosted by self-ensemble strategy [26]. Best results are in **bold**. Ours are in blue.

Method	Test1		Test2		Params.	Runtime	GPU
	PSNR	SSIM	PSNR	SSIM	(M)	(ms)	
NOAHTCV	**31.86**	**0.83**	**32.69**	**0.88**	5.6	25	V100
MiAlgo	31.39	0.82	28.56	0.85	4.5	18	2080
CASIA LCVG	30.19	0.81	31.47	0.86	464	31	A100
HIT-IIL	29.12	0.80	29.98	0.87	116	19818	V100
CS^{2}U (Ours)	**29.13**	**0.79**	**29.95**	**0.84**	105	1300	3090
SenseBrains	28.36	0.80	30.08	0.86	69	50	V100
PixelJump	28.15	0.80	n/a	n/a	6.64	40	3090
HiImage	27.96	0.79	n/a	n/a	11	200	3090
0noise	27.67	0.79	29.81	0.87	86	6	2080
OzU VGL	27.89	0.79	28.83	0.83	2.8	10	3090
CVIP	27.85	0.80	29.50	0.86	0.17	19	Q6000

Table 4. Quantitative RAW Reconstruction evaluation among our model and other baselines on MIT-fiveK [8,48]. Bese results are in **bold**.

Method	Nikon/PSNR	Canon/PSNR
UPI [5]	29.30	-
CycleISP [51]	29.40	31.71
InvGrayscale [47]	33.34	34.21
U-Net	38.24	41.52
Invertible-ISP (w/o JPEG) [48]	43.29	45.72
Invertible-ISP (with JPEG Fourier) [48]	44.42	46.78
Learnable Dictionaries [13]	43.62	50.08
Ours	**47.84**	**53.60**

35.31dB PSNR on the validation set. And our method achieves the fifth place on the test set with 31.02dB PSNR.

We show qualitative results in Figure 5, comparing with the results of other 10 participating teams in the competition. Our model achieves promising RAW recovery results, and compared to the top ones, the differences are rarely noticeable. Similarly, our algorithms enable effective restoration for challenging regions, especially the rich textures, as shown in the second line of Figure 5.

General RAW Reconstruction Benchmark

We compared our SSDNet with existing inverse ISP methods, including UPI [5], CycleISP [51], InvGrayscale [47], Invertible-ISP (w/o JPEG) [48] and Learnable Dictionaries [13]. We used the same training dataset as these methods, with PSNR and SSIM scores reported by [13]. Quantitative comparisons with existing inverse ISP methods are shown in Table 4. Our SSDNet has achieved comparable performance with previous state-of-the-art methods, even far outperforming previous state-of-the-art methods by 4.2 dB on the Nikon dataset.

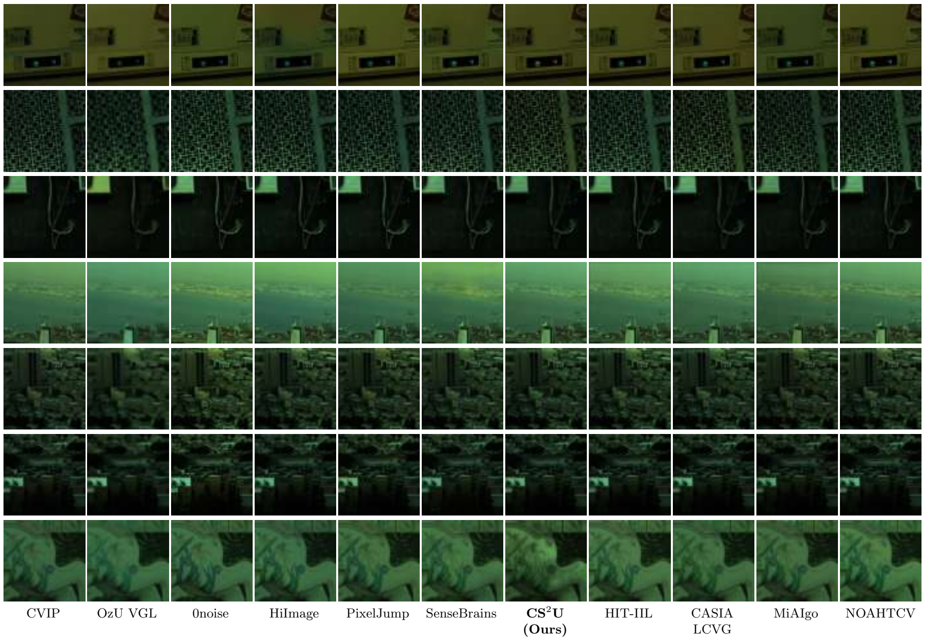

Fig. 5. RAW image restoration for the AIM 2022 Challenge [14]. Individual modules effectively enhance the restoration results. All visualisations are rendered by a simple ISP model (`https://github.com/mv-lab/AISP`).

This clearly demonstrates the effectiveness of the proposed SSDNet.

Discussions
Although we have achieved promising results for RGB to RAW, the network structure used nevertheless exhibited a large number of parameters, as shown in the right side of Table 3. We believe that combining the proposed learning paradigm with the model-based method [13] would alleviate this problem and even further improve their performance, which would also be an interesting piece of future work.

4.4 Application to Image Denoising

In this section, we leverage our learned RAW reconstruction model to RAW denoising. We follow CycleISP [51], firstly generating synthetic data using the inverse ISP model and then using them to train the denoiser. Specifically, we synthesize 100K noisy-clean image pairs, and then we modify the SSDNet, by only changing the input channel to 4, to be a RAW denoiser. We report our denoising results in Table 5 and compare them with the state-of-the-art RAW denoising methods available. Note CycleISP [51] was trained on 1 million images, while our model only trained with 10% data. Still, Our method has achieved comparable

Table 5. RAW denoising results on the SIDD dataset [1]. Best results are in bold.

Method	RAW		sRGB	
	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
EPLL [55]	40.73	0.935	25.19	0.842
GLIDE [42]	41.87	0.949	25.98	0.816
TNRD [11]	42.77	0.945	26.99	0.744
MLP [7]	43.17	0.965	27.52	0.788
KSVD [2]	43.26	0.969	27.41	0.832
NLM [6]	44.06	0.971	29.39	0.846
WNNM [18]	44.85	0.975	29.54	0.888
BM3D [15]	45.52	0.980	30.95	0.863
FoE [39]	45.78	0.966	35.99	0.90
DnCNN [53]	47.37	0.976	38.08	0.935
N3Net [33]	47.56	0.976	38.32	0.938
UPI [5]	48.89	0.982	**40.17**	**0.962**
CycleISP [51]	52.41	0.990	39.47	0.918
Learnable Dictionaries [13]	**52.48**	0.990	-	-
Ours	51.73	**0.993**	39.33	0.955

results with the latest methods [51,13], which demonstrates the application of our inverse ISP model to the downstream task, RAW image denoising.

5 Conclusions

In this work, we propose a new learning target and an effective network for data-driven RGB to RAW learning. We first exploit the unique property of RAW image, i.e. high dynamic range, by suggesting relaxing the supervision to a multivariate Gaussian distribution in order to learn images that are reasonable for a given supervision. Then, we propose the encoder-decoder architecture SS-DNet, which is inspired by the Transformer architecture and is straightforward and effective. Combined with the above two components, our method achieves an effective RGB to RAW mapping. Experimental results show that our algorithm achieves promising results on both synthetic and real datasets available. In particular, our method achieved the fifth place in the S7 track of AIM Reversed ISP Challenge. At last, we demonstrate the application of RGB to RAW on a denoising task, implying that this research can be a valid tool for RAW image denoising.

Acknowledgment

This work was supported by the National Key R&D Program of China under Grant 2018AAA0102100.

References

1. Abdelhamed, A., Lin, S., Brown, M.S.: A high-quality denoising dataset for smartphone cameras. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1692–1700 (2018)
2. Aharon, M., Elad, M., Bruckstein, A.: K-svd: An algorithm for designing overcomplete dictionaries for sparse representation. IEEE Transactions on signal processing **54**(11), 4311–4322 (2006)
3. Anwar, S., Barnes, N.: Real image denoising with feature attention. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3155–3164 (2019)
4. Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization. arXiv:1607.06450 (2016)
5. Brooks, T., Mildenhall, B., Xue, T., Chen, J., Sharlet, D., Barron, J.T.: Unprocessing images for learned raw denoising. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11036–11045 (2019)
6. Buades, A., Coll, B., Morel, J.M.: A non-local algorithm for image denoising. In: 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05). vol. 2, pp. 60–65. IEEE (2005)
7. Burger, H.C., Schuler, C.J., Harmeling, S.: Image denoising: Can plain neural networks compete with bm3d? In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1256–1272 (2012)
8. Bychkovsky, V., Paris, S., Chan, E., Durand, F.: Learning photographic global tonal adjustment with a database of input/output image pairs. In: CVPR 2011. pp. 97–104. IEEE (2011)
9. Chen, Q., Wu, Q., Wang, J., Hu, Q., Hu, T., Ding, E., Cheng, J., Wang, J.: Mixformer: Mixing features across windows and dimensions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5249–5259 (2022)
10. Chen, X., Wang, X., Zhou, J., Dong, C.: Activating more pixels in image super-resolution transformer. arXiv preprint arXiv:2205.04437 (2022)
11. Chen, Y., Yu, W., Pock, T.: On learning optimized reaction diffusion processes for effective image restoration. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5261–5269 (2015)
12. Cheng, S., Wang, Y., Huang, H., Liu, D., Fan, H., Liu, S.: Nbnet: Noise basis learning for image denoising with subspace projection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4896–4906 (2021)
13. Conde, M.V., McDonagh, S., Maggioni, M., Leonardis, A., Pérez-Pellitero, E.: Model-based image signal processors via learnable dictionaries. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 481–489 (2022)
14. Conde, M.V., Timofte, R., et al.: Reversed image signal processing and raw reconstruction. aim 2022 challenge report. In: Proceedings of the European Conference on Computer Vision Workshops (ECCVW) (2022)
15. Dabov, K., Foi, A., Katkovnik, V., Egiazarian, K.: Image denoising by sparse 3-d transform-domain collaborative filtering. IEEE Transactions on image processing **16**(8), 2080–2095 (2007)
16. Delbracio, M., Kelly, D., Brown, M.S., Milanfar, P.: Mobile computational photography: A tour. arXiv preprint arXiv:2102.09000 (2021)
17. Gharbi, M., Chaurasia, G., Paris, S., Durand, F.: Deep joint demosaicking and denoising. ACM Transactions on Graphics **35**(6) (2016)

18. Gu, S., Zhang, L., Zuo, W., Feng, X.: Weighted nuclear norm minimization with application to image denoising. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2862–2869 (2014)
19. Guo, J., Han, K., Wu, H., Tang, Y., Chen, X., Wang, Y., Xu, C.: Cmt: Convolutional neural networks meet vision transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12175–12185 (2022)
20. Guo, S., Yan, Z., Zhang, K., Zuo, W., Zhang, L.: Toward convolutional blind denoising of real photographs. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1712–1722 (2019)
21. Han, K., Wang, Y., Chen, H., Chen, X., Guo, J., Liu, Z., Tang, Y., Xiao, A., Xu, C., Xu, Y., et al.: A survey on vision transformer. IEEE transactions on pattern analysis and machine intelligence (2022)
22. He, X., Cheng, J.: Revisiting l1 loss in super-resolution: A probabilistic view and beyond. arXiv preprint arXiv:2201.10084 (2022)
23. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7132–7141 (2018)
24. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
25. Kingma, D.P., Dhariwal, P.: Glow: Generative flow with invertible 1x1 convolutions. Advances in neural information processing systems **31** (2018)
26. Lim, B., Son, S., Kim, H., Nah, S., Mu Lee, K.: Enhanced deep residual networks for single image super-resolution. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 136–144 (2017)
27. Liu, P., Zhang, H., Zhang, K., Lin, L., Zuo, W.: Multi-level wavelet-cnn for image restoration. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 773–782 (2018)
28. Liu, Y.L., Lai, W.S., Chen, Y.S., Kao, Y.L., Yang, M.H., Chuang, Y.Y., Huang, J.B.: Single-image hdr reconstruction by learning to reverse the camera pipeline. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)
29. Luo, W., Li, Y., Urtasun, R., Zemel, R.: Understanding the effective receptive field in deep convolutional neural networks. Advances in neural information processing systems **29** (2016)
30. Maini, R., Aggarwal, H.: A comprehensive review of image enhancement techniques. arXiv preprint arXiv:1003.4053 (2010)
31. Nguyen, R.M., Brown, M.S.: Raw image reconstruction using a self-contained srgb-jpeg image with only 64 kb overhead. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1655–1663 (2016)
32. Nguyen, R.M., Brown, M.S.: Raw image reconstruction using a self-contained srgb–jpeg image with small memory overhead. International journal of computer vision **126**(6), 637–650 (2018)
33. Plötz, T., Roth, S.: Neural nearest neighbors networks. Advances in Neural information processing systems **31** (2018)
34. Punnappurath, A., Brown, M.S.: Learning raw image reconstruction-aware deep image compressors. IEEE transactions on pattern analysis and machine intelligence **42**(4), 1013–1019 (2019)
35. Purohit, K., Rajagopalan, A.: Region-adaptive dense network for efficient motion deblurring. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 11882–11889 (2020)

36. Qian, G., Gu, J., Ren, J.S., Dong, C., Zhao, F., Lin, J.: Trinity of pixel enhancement: a joint solution for demosaicking, denoising and super-resolution. arXiv preprint arXiv:1905.02538 (2019)
37. Reddi, S.J., Kale, S., Kumar, S.: On the convergence of adam and beyond. arXiv preprint arXiv:1904.09237 (2019)
38. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
39. Roth, S., Black, M.J.: Fields of experts. International Journal of Computer Vision **82**(2), 205–229 (2009)
40. Schwartz, E., Giryes, R., Bronstein, A.M.: Deepisp: Toward learning an end-to-end image processing pipeline. IEEE Transactions on Image Processing **28**(2), 912–923 (2018)
41. Shi, W., Caballero, J., Huszár, F., Totz, J., Aitken, A.P., Bishop, R., Rueckert, D., Wang, Z.: Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1874–1883 (2016)
42. Talebi, H., Milanfar, P.: Global image denoising. IEEE Transactions on Image Processing **23**(2), 755–768 (2013)
43. Tian, C., Fei, L., Zheng, W., Xu, Y., Zuo, W., Lin, C.W.: Deep learning on image denoising: An overview. Neural Networks (2020)
44. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
45. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004)
46. Wronski, B., Garcia-Dorado, I., Ernst, M., Kelly, D., Krainin, M., Liang, C.K., Levoy, M., Milanfar, P.: Handheld multi-frame super-resolution. ACM Transactions on Graphics (TOG) **38**(4), 1–18 (2019)
47. Xia, M., Liu, X., Wong, T.T.: Invertible grayscale. ACM Transactions on Graphics (TOG) **37**(6), 1–10 (2018)
48. Xing, Y., Qian, Z., Chen, Q.: Invertible image signal processing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6287–6296 (2021)
49. Yue, Z., Yong, H., Zhao, Q., Zhang, L., Meng, D.: Variational denoising network: Toward blind noise modeling and removal. arXiv preprint arXiv:1908.11314 (2019)
50. Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.H.: Restormer: Efficient transformer for high-resolution image restoration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5728–5739 (2022)
51. Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.H., Shao, L.: Cycleisp: Real image restoration via improved data synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2696–2705 (2020)
52. Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.H., Shao, L.: Learning enriched features for real image restoration and enhancement. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16. pp. 492–511. Springer (2020)

53. Zhang, K., Zuo, W., Chen, Y., Meng, D., Zhang, L.: Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. IEEE transactions on image processing **26**(7), 3142–3155 (2017)
54. Zhang, K., Zuo, W., Zhang, L.: Ffdnet: Toward a fast and flexible solution for cnn-based image denoising. IEEE Transactions on Image Processing **27**(9), 4608–4622 (2018)
55. Zoran, D., Weiss, Y.: From learning models of natural image patches to whole image restoration. In: 2011 international conference on computer vision. pp. 479–486. IEEE (2011)