RAWtoBit: A Fully End-to-end Camera ISP Network

Wooseok Jeong and Seung-Won Jung*

Department of Electrical Engineering, Korea University, Seoul, Korea
{561wesd, swjung83}@korea.ac.kr

Abstract. Image compression is an essential and last processing unit in the camera image signal processing (ISP) pipeline. While many studies have been made to replace the conventional ISP pipeline with a single end-to-end optimized deep learning model, image compression is barely considered as a part of the model. In this paper, we investigate the designing of a fully end-to-end optimized camera ISP incorporating image compression. To this end, we propose RAWtoBit network (RBN) that can effectively perform both tasks simultaneously. RBN is further improved with a novel knowledge distillation scheme by introducing two teacher networks specialized in each task. Extensive experiments demonstrate that our proposed method significantly outperforms alternative approaches in terms of rate-distortion trade-off.

Keywords: Camera network, knowledge distillation, image compression, image signal processing pipeline

1 Introduction

The image signal processing (ISP) pipeline is receiving increasing attention from the research community, as mobile devices are equipped with powerful hardware which can be utilized to process more sophisticated operations to boost performance [14]. A typical ISP pipeline includes several local and global operations, such as white balance, demosaicing, color correction, gamma correction, denoising, and tone mapping [22]. Since each of these operations is a research topic on its own, they are often separately optimized for a given ISP pipeline, which can be sub-optimal.

The deep learning-based approach has proven to be effective in various computer vision and image processing tasks, and consequently, many attempts have been made to replace conventional ISPs with convolutional neural networks (CNNs) [14,18,24,30]. While earlier learning-based works only dealt with the ISP components separately, such as demosaicing [32] and denoising [15,35,37], recent studies have paid attention to the design of a unified CNN that performs all ISP functionalities, which is referred to as an ISP-Net. For example, Schwartz et al. [24] proposed a two-stage ISP-Net for low-level and high-level operations and

* Corresponding author.

showed that sharing features between two stages leads to a better result. In [18], correlated ISP components are categorized into two groups and independently trained, followed by joint fine-tuning.

However, most previous ISP-Nets did not consider that sRGB images rendered from RAW are essentially followed by lossy compression, which may substantially alter the image quality. Although some studies [29, 30] have proposed to integrate JPEG simulation as a part of the model to take into account the compression artifacts, they are limited to the simulation, and the standard JPEG is still used to produce a bitstream.

The objective of image compression is to reduce bits required for storing and transmitting an image without largely affecting the perceived quality. Image compression is typically achieved by transforming the image, quantizing the transformed coefficients, and compressing the resultant representation using entropy coding [12]. In particular, the quantization introduces an inevitable error, where coarse quantization leads to bitrate reduction at the expense of distortion increase, giving rise to the rate-distortion trade-off. Under the principle of transform coding [12], many codecs have been developed to improve rate-distortion performance, including JPEG2000 [25] and versatile video coding (VVC) [4]. Most of the components in these existing codecs, however, are designed by human experts much like conventional ISP components, which promote researchers to design a CNN that performs image compression, which is referred to as a Comp-Net [1, 2, 8, 20]. Unlike conventional image compression techniques, Comp-Net is inherently differentiable and performs significantly better than the commonly used JPEG.

The advances in deep learning-based image processing and image compression motivate us to propose a fully end-to-end camera ISP network called RAWtoBit network (RBN). Our RBN takes RAW as an input as other ISP-Nets [18, 24, 29, 30] but outputs a bitstream, which can be decoded to reconstruct a high-quality sRGB image. To this end, we investigate two structures: cascaded and unified. Cascaded structure refers to a simple concatenation of ISP-Net and Comp-Net. However, the performance of Comp-Net can be upper-bounded by ISP-Net, resulting in sub-optimal rate-distortion performance. Unified structure refers to a single network that simultaneously performs the ISP operations and image compression. Although the unified structure can be easily implemented by training a Comp-Net with RAW-sRGB pairs with slight modification in network architecture, such a structure can also lead to sub-optimal rate-distortion performance since Comp-Net is not originally designed to perform complicated ISP operations. Observing that these two naïve approaches suffer from poor rate-distortion performance, we propose RBN to handle both tasks effectively. Furthermore, we present two teacher networks, namely the ISP teacher and the compression teacher, to guide RBN to reach a better rate-distortion trade-off. Experimental results demonstrate that our proposed RBN performs noticeably better than the alternative approaches. Our contribution can be summarized as follows:

- This work is the first attempt to integrate camera ISP and image compression in a single learning framework to the best of our knowledge. Unlike previous studies, our RBN takes RAW data as an input and produces a bitstream as an output.
- We propose a method that distills the knowledge from two teacher models, namely the ISP teacher and the compression teacher, to make RBN effectively performs both ISP and compression tasks.
- Extensive experimental results demonstrate that our RBN with knowledge distillation significantly improves rate-distortion performance over the cascaded or unified structure.

2 Related Work

2.1 Camera ISP Network

An ISP-Net is trained to render sRGB images from RAW sensor data, either by explicitly supervising the subtasks of an ISP or by directly learning a RAW to sRGB mapping function, which in this case learns the necessary subtasks implicitly. Towards the latter approach, Schwartz et al. [24] presented an ISP-Net called DeepISP, where the two-stage network is employed for low-level and high-level operations, and demonstrated that the latter high-level enhancement task could be improved by sharing features from the former low-level task. Chen et al. [6] tackled the challenging problem of low light photography by learning the mapping from a RAW image captured in low light to its corresponding long-exposure sRGB image. Ignatov et al. [14] collected RAW and sRGB image pairs using a smartphone camera and a professional DSLR, respectively. By training a network with a hierarchical structure using such image pairs, their ISP-Net produced sRGB images with better visual quality than those rendered from the smartphone's built-in ISP. Zhang et al. [38] addressed the misalignment problem between RAW and sRGB image pairs [14] and proposed to warp the ground-truth sRGB images to the color corrected RAW images for supervision. However, none of the aforementioned ISP-Nets consider image compression, an essential component of practical camera ISPs.

Meanwhile, Xing et al. [30] proposed an invertible ISP-Net, where the forward pass produces sRGB images and the backward pass reconstructs RAW images. With the differentiable rounding function defined by the Fourier series, JPEG compression is simulated and integrated into the training procedure. Uhm et al. [29] cascaded an ISP-Net and a network that simulates JPEG compression to train the ISP-Net with lossy image compression in consideration. However, these ISP-Nets [29,30] are not trained in a fully end-to-end (i.e., RAW-to-bits) manner and rely on the standard JPEG compression at the inference stage, leading to sub-optimal rate-distortion performance.

2.2 Image Compression Network

Learned image compression has drawn attention from many researchers in recent years, and state-of-the-art methods [10] demonstrate improved rate-distortion

performance compared to hand-crafted codecs such as JPEG and JPEG2000. Learning-based methods share the same principle of transform coding as the conventional codecs; however, the nonlinear transformation is performed by a neural network instead of discrete cosine transform or wavelet transform, which are linear transformations, hence capable of learning a more complex representation. The most indispensable component of lossy image compression is undoubtedly the quantization, which is not differentiable, making the design of Comp-Net challenging. Many Comp-Nets thus have applied different strategies to deal with the quantization in the training process, e.g., by adding uniform noise [1] or replacing the derivative of the rounding function with identity function [26].

While early Comp-Nets [27, 28] employed recurrent neural networks, more recent Comp-Nets [1, 26] incorporated an autoencoder structure to minimize distortion and an entropy model estimation network to minimize bitrate. To this end, Ballé et al. [2] proposed to transmit side information to estimate the scale of the zero-mean Gaussian entropy model. Minnen et al. [20] further improved this idea by predicting the mean and scale of the Gaussian entropy model conditioned on hyperprior and decoded latent elements. Cheng et al. [8] achieved comparable performance to the latest VVC standard [4] by embedding attention modules in the network and using Gaussian mixtures as the entropy model. An interested reader can refer to the recent article [13] for a systematic literature review.

2.3 Knowledge Distillation

The main purpose of knowledge distillation (KD) is to transfer knowledge from a large teacher model to a more compact model without significantly compromising the performance. A comprehensive survey on KD [11] categorized existing schemes into three groups: response-based, feature-based, and relation-based KD. We focus on the feature-based KD, where the intermediate features are compared between teacher and student models as a means of knowledge transfer. Since Romero et al. [23] first proved the effectiveness of providing intermediate representation of a teacher model as a hint for training a student model, many approaches have been proposed to find a better form of KD. Zagoruyko et al. [33] proposed to transfer attention maps generated from the intermediate layers. Kim et al. [16] argued that the direct transfer of teacher's knowledge ignores the structural difference between the teacher and student models and introduced "factors", which are the features extracted and paraphrased from the teacher model, as a more effective form of knowledge. Chen et al. [7] developed a framework that can adaptively assign appropriate teacher layers to each student layer by attention allocation.

Although most KD schemes are developed for model compression, some studies have investigated KD from the models which perform related tasks. While not strictly following the general teacher-student principle in KD, Xu et al. [31] proposed to simultaneously perform depth estimation and scene parsing with guidance of intermediate auxiliary tasks such as surface normal estimation and semantic segmentation. Zhang et al. [36] introduced a logit and representation

graph for KD from multiple self-supervised auxiliary tasks. Similar to these approaches, we weakly disentangle the main objective into two tasks and perform KD using two teacher networks that better perform each task.

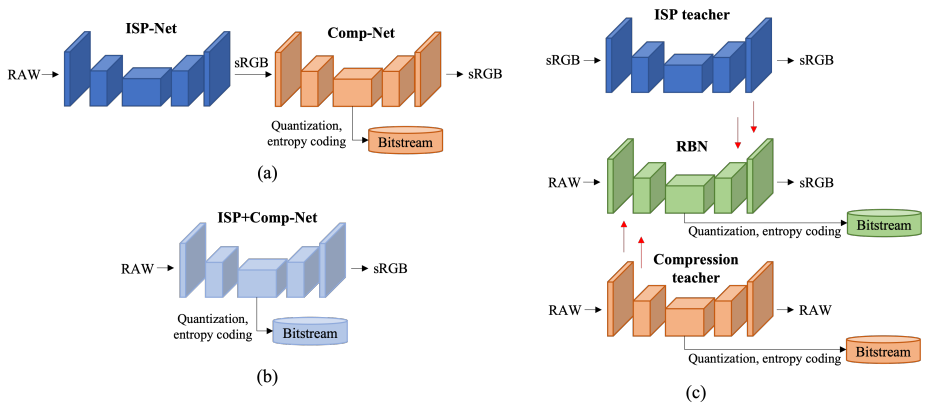

Fig. 1. Different configurations of end-to-end camera ISP networks: (a) Cascaded ISP-Net and Comp-Net, (b) unified ISP+Comp-Net, and (c) our proposed RBN with KD from two teachers. The red arrows represent the KD direction, and the context model for rate estimation is omitted for simplicity.

3 Proposed Method

The proposed work is the first attempt to integrate ISP-Net and Comp-Net to the best of our knowledge. We thus first present two straightforward configurations of the ISP-Net and Comp-Net integration, namely cascaded structure (Section 3.1) and unified structure (Section 3.2). We then introduce our RBN, which is also based on the unified structure but specially designed and trained with our KD scheme (Section 3.3).

3.1 Cascaded Structure

A naïve approach to combine ISP and lossy compression is to cascade ISP-Net and Comp-Net, as shown in Fig. 1(a). An ISP-Net takes a RAW image $x_r \in \mathbb{R}^{4\times H/2\times W/2}$ as an input and produces an sRGB image $\hat{x}_s \in \mathbb{R}^{3\times H\times W}$, while Comp-Net takes an sRGB image as an input and generates a bitstream which can reconstruct an sRGB image. Both ISP-Net and Comp-Net are separately trained and cascaded. This configuration is not limited to specific ISP-Net and Comp-Net architectures, and in our study, we use LiteISPNet [38] and the context+hyperprior model [20] for ISP-Net and Comp-Net, respectively. In addition, as in [29], one can try fine-tuning ISP-Net in conjunction with Comp-Net to

take lossy compression into consideration. We also investigate the effectiveness of this fine-tuning in Section 4.

3.2 Unified Structure

Another way to achieve the same objective is to directly train Comp-Net with RAW-sRGB image pairs, as shown in Fig. 1(b). Note that conventional Comp-Nets input and output sRGB images, as shown in Fig. 1(a). However, our target network configuration requires the network to take a RAW image $x_r \in \mathbb{R}^{4\times H/2\times W/2}$ as an input and produce a bitstream which can reconstruct an sRGB image $\hat{x}_s \in \mathbb{R}^{3\times H\times W}$. Consequently, we modify the context+hyperprior model [20] to handle a four-channel input and add an additional inverse generalized divisive normalization (IGDN) and a transposed convolutional layer in the decoder to produce the sRGB image with the target size. The network is trained using the rate-distortion loss [20] while measuring the difference between ground-truth and decoded sRGB images.

Since conventional Comp-Nets, including the context+hyperprior model [20], are not designed to handle complicated ISP functions, it is expected that this unified model cannot perform both ISP and compression functionalities properly. More dedicated architecture design and training methodology are required to realize an effective end-to-end camera ISP, which is the motivation of the proposed RBN.

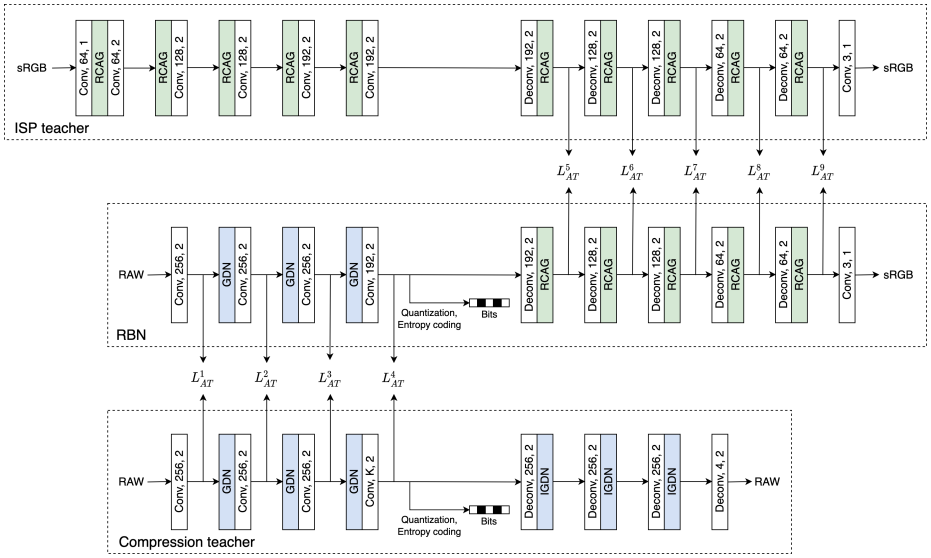

Fig. 2. Network architecture of RBN, consisting of (de)convolutional layers with the specified number of features and stride, RCAG [38], GDN, and IGDN. The applied context model for rate estimation [20] is omitted for simplicity.

3.3 RBN with KD

We now introduce our proposed RBN, which has a more appropriate architecture for the integration of ISP-Net and Comp-Net with a novel KD scheme, as depicted in Fig. 1(c). While the unified structure described in Section 3.2 does achieve the main objective of combining ISP-Net and Comp-Net into a single network, the performance is expected to be unsatisfactory in terms of the rate-distortion trade-off. This is because Comp-Net cannot fully handle the transformation of the RAW image into latent representation for compression and necessary ISP operations at the same time. Hence, we design RBN to be capable of performing both tasks with guidance from two teacher networks, namely the ISP teacher and the compression teacher.

Fig. 2 illustrates the detailed network architecture of RBN. We design RBN to have a heterogeneous encoder and decoder. The encoder of RBN follows the general structure in image compression with a series of strided convolution and generalized divisive normalization (GDN) [20]. Specifically, we modify the encoder architecture of one of the representative Comp-Nets [20] that sets the number of channels and kernel size as 192 and 5, respectively, to compress a three-channel sRGB image. Since our RBN takes a RAW image that is packed into four channels, the receptive field can grow uncontrollably quickly if a large kernel size is used. Hence, we use the kernel size of 3 while increasing the number of channels to 256. The last convolutional layer produces latent representation with 192 channels. The decoder architecture is modified from LiteISPNet [38], which is one of the state-of-the-art ISP-Nets. In particular, we replace inverse wavelet transform with transposed convolution and use two residual channel attention blocks in the residual channel attention group (RCAG). Note that no skip connection exists between the encoder and decoder since the decoded latent vector alone should be capable of reconstructing an sRGB image. To perform entropy coding and entropy model estimation, we leverage the context+hyperprior model [20], which estimates the mean and scale of the Gaussian entropy model using the spatially adjacent decoded latent elements and hyperprior.

Although RBN can be trained in an end-to-end manner, it may still suffer from sub-optimal rate-distortion performance since joint learning of compression and ISP is challenging. To overcome this issue, on the one hand, we guide the encoder of RBN to focus more on image compression using the compression teacher. As shown in Fig. 2, the compression teacher network takes a RAW image $x_r \in \mathbb{R}^{4\times H/2\times W/2}$ and produces a bitstream that can reconstruct a RAW image $\hat{x}_r \in \mathbb{R}^{4\times H/2\times W/2}$. Because the encoder of the compression teacher is trained to find compact representation for efficient compression, we consider that the knowledge from the encoder of the compression teacher can be distilled to the encoder of RBN. On the other hand, we guide the decoder of RBN to focus more on reconstructing the sRGB image from the latent representation by using the ISP teacher. As shown in Fig. 2, the ISP teacher is designed as an sRGB autoencoder such that its decoder can best perform the sRGB image reconstruction from low-dimensional latent representation. Consequently, we consider that the knowledge from the decoder of the ISP teacher can be distilled to the decoder of

RBN. Note that except for the last convolutional layer of the compression teacher network, the encoder and decoder pairs between the two teacher networks and RBN have identical structures to facilitate KD.

To perform KD, we adopt the attention transfer [33], where the spatial attention maps evaluated from the intermediate layers of the teacher and student networks are compared. In the original work [33], the attention map is defined as the sum of absolute values along the channel dimension of the output of the intermediate layer. We empirically found it to be ineffective to take absolute values, and thus we define the attention map as the direct sum along the channel dimension:

$$M_j = \sum_{i=1}^{C_j} A_j\left(i,:,:\right), \tag{1}$$

where $A_j \in \mathbb{R}^{C_j \times H_j \times W_j}$ is the output of the j-th intermediate layer, $M_j \in \mathbb{R}^{H_j \times W_j}$ is the attention map of the j-th intermediate layer, C_j, H_j, and W_j are the corresponding channel dimension, height, and width, respectively. This modification is necessary because we apply the attention transfer between the outputs of the convolutional layers and not the ReLU activation, thus taking absolute values can lead to the loss of directional information embedded in the output tensor. The attention loss for KD, L_{AT}, is defined as the mean squared error between the normalized attention maps of the teacher and student networks:

$$\begin{aligned} L_{AT} &= \sum_{j=1}^{n_p} \alpha_j L_{AT}^j, \\ L_{AT}^j &= \frac{1}{N_j} \left\| \frac{M_j^S}{\left\|M_j^S\right\|_2} - \frac{M_j^T}{\left\|M_j^T\right\|_2} \right\|_2^2, \end{aligned} \tag{2}$$

where $\|\cdot\|_2$ measures the L2-norm, $M_j^S(M_j^T)$ is the j-th attention map of the student (teacher) network, n_p is the number of pairs of the attention maps, N_j is the number of elements in the j-th attention map, and α_j is the weight for the j-th loss term. Inspired by [21], we set α_j to make the attention loss relatively higher than the rate-distortion loss during the early training phase, and decay it as the training progresses. In this way, our RBN can initially focus on KD and progressively switch to the main objective of the rate-distortion optimization. To this end, α_j is chosen as:

$$\alpha_j = \alpha_0 \cdot \gamma^{k^2}, \tag{3}$$

where α_0 is the initial value, and γ is the decay factor. α_0 is set to 10^6 for KD between the two encoders and 10^5 for KD between the two decoders, while γ is set to 0.99999 for both cases. In other words, α_j slowly decreases as training epoch k increases. The final loss function is defined as follows:

$$L_{total} = L_R + \lambda L_D + L_{AT}, \tag{4}$$

where L_R and L_D are the rate and distortion loss terms defined in [20], respectively, and the trade-off parameter λ determines the rate-distortion trade-off.

4 Experimental Results

4.1 Implementation Details

Dataset and Settings To train and test the models, we collected 487 RAW images taken with Nikon D700 from the MIT-Adobe FiveK dataset [5]. As in [29, 30, 34], the ground-truth sRGB images were rendered using the LibRaw library [19] and split into a ratio of 80:5:15 for training, validation, and testing. The training image patches were randomly extracted from the RAW images with the dimension of $4 \times 128 \times 128$ by four-channel packing (RGGB), and their corresponding patches were extracted from the sRGB images with the dimension of $3 \times 256 \times 256$. Data augmentation, including random flip and rotation, was applied during the training. Our experiments were conducted using PyTorch with Adam optimizer [17] on Intel i9-10980XE and NVIDIA RTX 3090.

Cascaded Structure We trained a LiteISPNet with the L2 loss. A batch size of 16 was used in this experiment. The model was trained for 24k iterations with the learning rate of 5×10^{-5} and fine-tuned for 2.4k iterations with the learning rate of 5×10^{-6}. For the context+hyperprior model, we leveraged the pre-trained model provided by [3]. We also experimented with the fine-tuning of the LiteISPNet in conjunction with the context-hyperprior model. In this case, after the pre-training of the LiteISPNet for 24k iterations, fine-tuning was performed in the cascaded configuration while fixing the parameters of the context-hyperprior model.

Unified Structure The modified context-hyperprior model as described in Section 3.2 was trained for 1.6M iterations. We used a batch size of 8 in this experiment. The learning rate was set to 5×10^{-5} and reduced to 5×10^{-6} after 1.5M iterations. The trade-off parameter λ was chosen from the set $\{0.0035, 0.0067, 0.013, 0.025, 0.0483, 0.0932, 0.18\}$, resulting seven different models with different rate-distortion performance.

RBN with KD To train the teacher and student networks, we set the initial learning rate to 5×10^{-5} and then decayed it to 5×10^{-6}. Specifically, the ISP teacher was trained for 580k iterations, where the learning rate was decayed after 480k iterations. The compression teacher network was trained for 2M iterations, where the learning rate was decayed after 1.5M iterations. We trained separate compression teacher models for each rate-distortion trade-off by choosing λ from the set $\{0.0035, 0.0067, 0.013, 0.0483, 0.0932, 0.18, 0.36\}$ to support seven models with different rate-distortion performance. The channel dimension of the last convolutional layer of the encoder of the compression teacher, i.e., K in Fig. 2, was set to 192 for the three lower bitrate settings ($\lambda = 0.0035, 0.0067, 0.013$) and 320 for the four higher bitrate settings. Our proposed RBN was trained for 1M iterations, where the learning rate was decayed after 900k iterations. We also trained separate RBNs with the same λ values used in the corresponding compression teachers. We used a batch size of 8 in this experiment.

4.2 Performance Comparisons

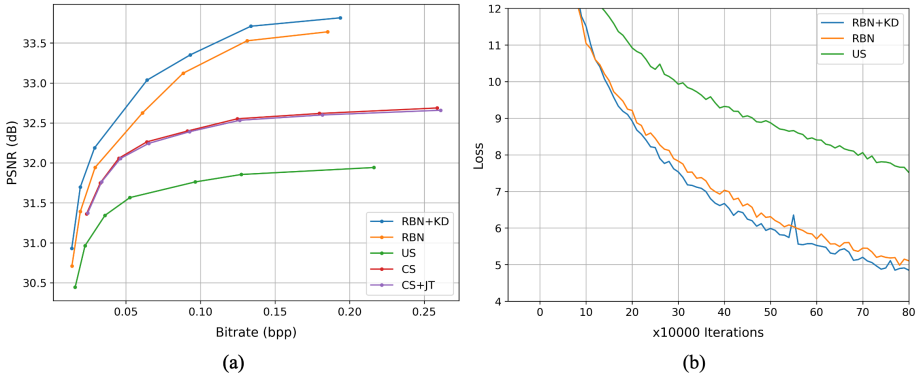

Fig. 3. Experimental results: (a) Rate-distortion performance comparisons and (b) training loss for the models trained with $\lambda = 0.18$.

Quantitative results The rate-distortion performance of the proposed method is shown in Fig. 3(a), where the average PSNR between the ground-truth and reconstructed sRGB images and average bits per pixel (bpp) were measured from 73 test images. Note that we evaluated bpp in terms of the resolution of the reconstructed sRGB image and not the four-channel packed RAW image. Compared to the baseline approaches such as the cascaded structure (CS) and unified structure (US), RBN outperformed rate-distortion performance, and RBN with KD (RBN+KD) further improved performance. We notice that the performance gap becomes more significant in the high bitrate region, which is important since high-bitrate compression is usually used in practical camera ISPs. Note that the average PSNR between the ground-truth and reconstructed sRGB images obtained by the LiteISPNet used in the cascaded structure is 32.82 dB, which means that the cascaded structure cannot exceed this value regardless of compression rates. Hence the overall rate-distortion performance of the cascaded structure is bounded by the performance of the ISP-Net. As for the joint fine-tuning of the ISP-Net in the cascaded structure (CS+JT), we did not find it effective. We suspect that, unlike the JPEG compression artifacts, which have some form of regularity in that they appear in every 8×8 block boundary, the effect of Comp-Net is more subtle and complex, making it difficult to be effectively captured by the joint fine-tuning of the network. The unified structure performed the worst among the compared methods. This result is expected as the network architecture aims to extract and normalize the features for compression and thus has difficulty performing local and global operations of the ISP. We have also experimented cascading other ISP-Nets with Comp-Net, as well as cascading Comp-Net with ISP-Net, where the result can be found in the supplementary material.

Fig. 3(b) shows training loss plots for the unified structure, RBN, and RBN with KD. As suggested in [21], the initial training phase is critical for the network to form a necessary connection to optimize the loss function. By enforcing the network to focus on KD in the initial training stage, the loss of RBN with KD reduced slower than that of RBN in early iterations. However, RBN with KD eventually reached a better optimization point, leading to better rate-distortion performance. It is also clear that the unified structure poorly optimized the rate-distortion loss.

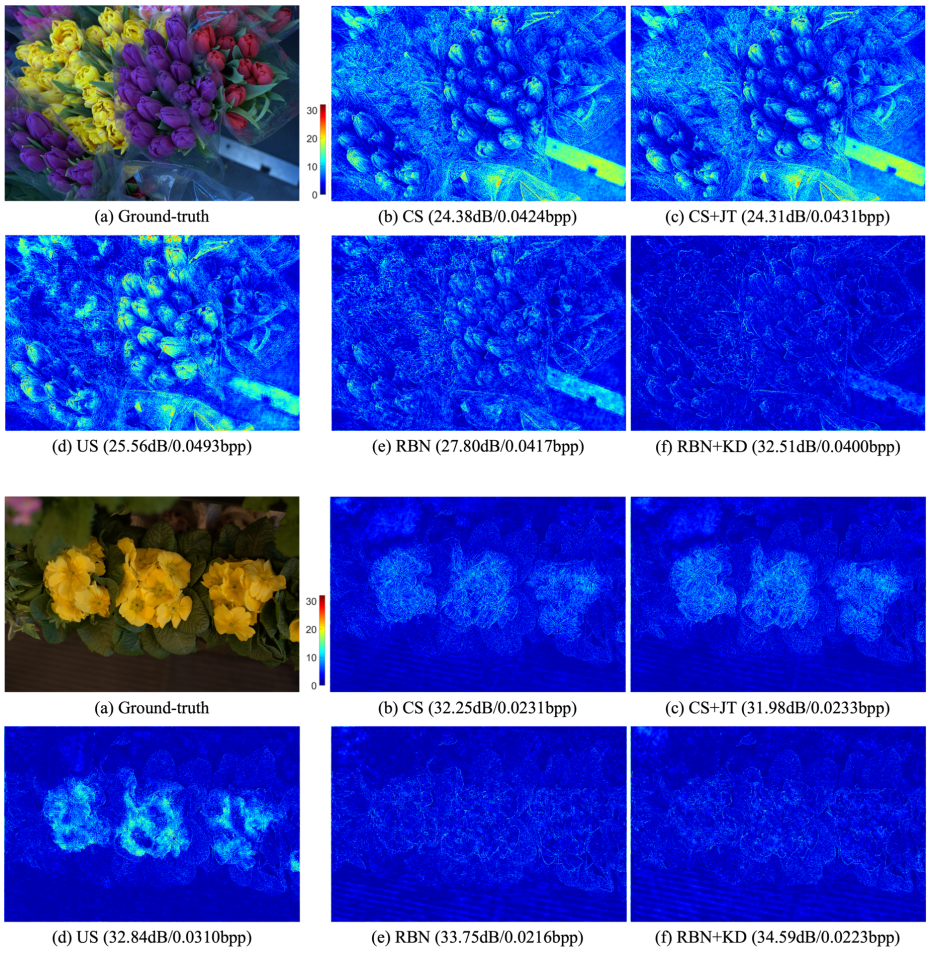

Fig. 4. Visual comparisons: (a) Ground-truth sRGB image and the error maps between the ground-truth and reconstructed sRGB images for (b) cascaded structure, (c) cascaded structure with joint fine-tuning, (d) unified structure, (e) RBN, and (f) RBN+KD.

Qualitative results The visual comparisons of the proposed methods are shown in Fig. 4. These results were obtained from the models trained with $\lambda = 0.013$ for the unified structure and RBN and $\lambda = 0.0035$ for the cascaded structure. For each method, we visualize the error map between the ground-truth and resultant sRGB images for comparison. It can be seen that the images obtained from the cascaded and unified structures exhibit significant errors. These two structures suffer from not only reproducing image textures such as flower petals but also rendering global color tones, demonstrating that these naïve approaches are insufficient to handle compression and necessary ISP operations at the same time. Meanwhile, the proposed RBN renders sRGB images with less distortion, especially in the texture-rich region. RBN with KD further reduces distortion, yielding the highest image quality at similar compression rates. Additional visual comparison in Fig. 5 clearly shows that RBN with KD renders color more accurately.

(a) Ground-truth | (b) CS (32.66dB/0.0674bpp) | (c) CS+JT (32.59dB/0.0679bpp) | (d) US (33.55dB/0.0844bpp) | (e) RBN (35.17dB/0.0578bpp) | (f) RBN+KD (37.84dB/0.0617bpp)

Fig. 5. Visual comparisons: (a) Ground-truth sRGB image and reconstructed sRGB images obtained by (b) cascaded structure, (c) cascaded structure with joint fine-tuning, (d) unified structure, (e) RBN, and (f) RBN+KD.

4.3 Ablation Studies

We conducted ablation studies on the KD scheme used in the proposed method to verify its effectiveness. Fig. 6(a) shows rate-distortion performance for the RBNs trained with guidance from either the ISP teacher or the compression teacher only, as well as the RBNs with full KD and without any KD. The hyperparameters α_0 and γ were kept the same for these four compared models. Unexpectedly, removing either one of the teacher networks resulted in inferior performance even compared with the RBN without KD. We conjecture that the hyperparameter settings of $\alpha_0 = 10^6$ for the encoder, $\alpha_0 = 10^5$ for the decoder, and $\gamma = 0.99999$ were empirically chosen for the KD with two teachers; thus, simply removing one of the teacher networks might not result in a better rate-distortion performance. Next, to verify our claim of the ineffectiveness of deriving the attention map by taking the absolute value of the tensor, we report

the result for the model trained with such method in Fig. 6(b). It is clear that the method mentioned above actually hinders the performance of RBN.

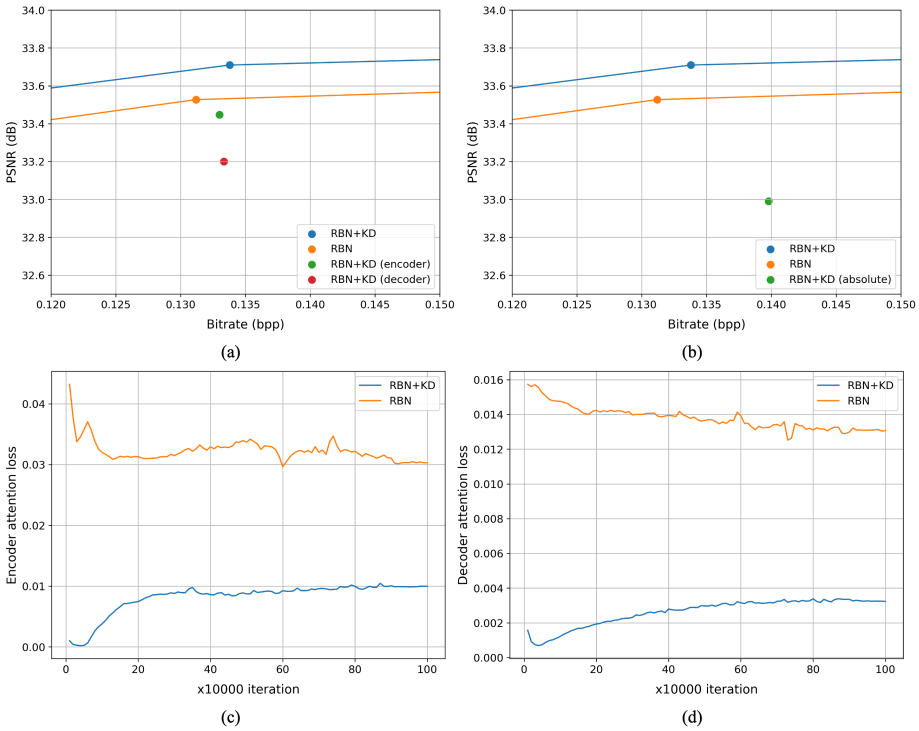

Fig. 6. Ablation studies for models trained with $\lambda = 0.18$: (a) RBNs with different KD strategies and (b) RBNs with different attention maps for KD, and attention loss plots of RBNs with and without KD: (c) encoder and (d) decoder.

Figs. 6(c) and (d) show the encoder and decoder attention loss plots for the RBNs trained with and without KD, respectively. Although the RBN without KD was not trained to optimize these attention loss terms, these losses decreased to some extent as the training progressed, which suggests that the attention losses are related to the main objective of the rate-distortion trade-off. In the RBN with KD, the attention losses decreased at the early iterations but then increased since the weight for the attention loss is decreased by (3) as training proceeded. In other words, the optimization with the attention loss in the RBN with KD contributed to finding good initial conditions for the model to reach a better rate-distortion trade-off. Also, note that KD does not incur additional computational cost in the inference stage. Comparison on the computational cost of the experimented models can be found in the supplementary material.

Fig. 7. Samples of the sRGB images rendered from RAW captured with Nikon D7000. The RBN+KD model used for testing was trained with RAW from Nikon D700. Bilinear demosaicing is applied to the RAW for visualization.

4.4 Generalization Test

Since different camera sensors have different characteristics, a dedicated RBN is required for each specific camera model in principle. Here, to test the generalization ability of the RBN with KD trained on the Nikon D700 subset from the MIT-Adobe FiveK dataset [5], we applied the trained model to the test images captured by a different camera sensor. Fig. 7 shows some test results on the Nikon D7000 subset from the RAISE dataset [9]. It can be seen that our results still exhibit high contrast and well-adjusted color with rich textures, indicating that sensor-specific training is desired but not mandatory, at least for the sensors from the same manufacturer.

5 Conclusions

In this paper, we presented the first approach to integrate the ISP and image compression into a single framework. To this end, we designed a network called RBN to perform both tasks simultaneously and effectively. Compared to the naïve baselines of the cascaded and unified structures, RBN exhibits significantly better rate-distortion performance. In addition, we further improved RBN by introducing KD from the two teacher networks specialized in each task. Experimental results demonstrated that RBN with KD shows a noticeable performance increase over the alternative approaches. We hope our work inspires further research on the fully end-to-end camera ISP network.

Acknowledgements
This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea Government (MSIT) (No. 2022R1A2C2002810).

References

1. Ballé, J., Laparra, V., Simoncelli, E.P.: End-to-end optimized image compression. In: Proceedings of the International Conference on Learning Representations (2017)
2. Ballé, J., Minnen, D., Singh, S., Hwang, S.J., Johnston, N.: Variational image compression with a scale hyperprior. In: Proceedings of the International Conference on Learning Representations (2018)
3. Bégaint, J., Racapé, F., Feltman, S., Pushparaja, A.: CompressAI: a PyTorch library and evaluation platform for end-to-end compression research. arXiv preprint arXiv:2011.03029 (2020)
4. Bross, B., Wang, Y.K., Ye, Y., Liu, S., Chen, J., Sullivan, G.J., Ohm, J.R.: 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)
5. Bychkovsky, V., Paris, S., Chan, E., Durand, F.: Learning photographic global tonal adjustment with a database of input/output image pairs. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 97–104. IEEE (2011)
6. Chen, C., Chen, Q., Xu, J., Koltun, V.: Learning to see in the dark. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3291–3300 (2018)
7. Chen, D., Mei, J.P., Zhang, Y., Wang, C., Wang, Z., Feng, Y., Chen, C.: Cross-layer distillation with semantic calibration. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 7028–7036 (2021)
8. Cheng, Z., Sun, H., Takeuchi, M., Katto, J.: Learned image compression with discretized Gaussian mixture likelihoods and attention modules. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7939–7948 (2020)
9. Dang-Nguyen, D.T., Pasquini, C., Conotter, V., Boato, G.: Raise: A raw images dataset for digital image forensics. In: Proceedings of the ACM Multimedia Systems Conference. pp. 219–224 (2015)
10. Gao, G., You, P., Pan, R., Han, S., Zhang, Y., Dai, Y., Lee, H.: Neural image compression via attentional multi-scale back projection and frequency decomposition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14677–14686 (2021)
11. Gou, J., Yu, B., Maybank, S.J., Tao, D.: Knowledge distillation: A survey. International Journal of Computer Vision **129**(6), 1789–1819 (2021)
12. Goyal, V.K.: Theoretical foundations of transform coding. IEEE Signal Processing Magazine **18**(5), 9–21 (2001)
13. Hu, Y., Yang, W., Ma, Z., Liu, J.: Learning end-to-end lossy image compression: A benchmark. IEEE Transactions on Pattern Analysis and Machine Intelligence (2022)
14. Ignatov, A., Van Gool, L., Timofte, R.: Replacing mobile camera ISP with a single deep learning model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 536–537 (2020)
15. Kim, D.W., Ryun Chung, J., Jung, S.W.: GRDN: Grouped residual dense network for real image denoising and GAN-based real-world noise modeling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 2086–2094 (2019)

16. Kim, J., Park, S., Kwak, N.: Paraphrasing complex network: Network compression via factor transfer. In: Proceedings of the Advances in Neural Information Processing Systems. vol. 31 (2018)
17. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
18. Liang, Z., Cai, J., Cao, Z., Zhang, L.: CameraNet: A two-stage framework for effective camera ISP learning. IEEE Transactions on Image Processing **30**, 2248–2262 (2021)
19. LibRaw LLC: Libraw (2019), https://www.libraw.org/
20. Minnen, D., Ballé, J., Toderici, G.D.: Joint autoregressive and hierarchical priors for learned image compression. In: Proceedings of the Advances in Neural Information Processing Systems. vol. 31 (2018)
21. Passalis, N., Tzelepi, M., Tefas, A.: Heterogeneous knowledge distillation using information flow modeling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2339–2348 (2020)
22. Ramanath, R., Snyder, W.E., Yoo, Y., Drew, M.S.: Color image processing pipeline. IEEE Signal Processing Magazine **22**(1), 34–43 (2005)
23. Romero, A., Ballas, N., Kahou, S.E., Chassang, A., Gatta, C., Bengio, Y.: FitNets: Hints for thin deep nets. In: Proceedings of the International Conference on Learning Representations (2015)
24. 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)
25. Taubman, D., Marcellin, M.: JPEG2000 image compression fundamentals, standards and practice: Image compression fundamentals, standards and practice, vol. 642. Springer Science & Business Media (2012)
26. Theis, L., Shi, W., Cunningham, A., Huszár, F.: Lossy image compression with compressive autoencoders. In: Proceedings of the International Conference on Learning Representations (2017)
27. Toderici, G., O'Malley, S.M., Hwang, S.J., Vincent, D., Minnen, D., Baluja, S., Covell, M., Sukthankar, R.: Variable rate image compression with recurrent neural networks. In: Proceedings of the International Conference on Learning Representations (2016)
28. Toderici, G., Vincent, D., Johnston, N., Jin Hwang, S., Minnen, D., Shor, J., Covell, M.: Full resolution image compression with recurrent neural networks. In: Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition. pp. 5306–5314 (2017)
29. Uhm, K.H., Choi, K., Jung, S.W., Ko, S.J.: Image compression-aware deep camera ISP network. IEEE Access **9**, 137824–137832 (2021)
30. 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)
31. Xu, D., Ouyang, W., Wang, X., Sebe, N.: PAD-Net: Multi-tasks guided prediction-and-distillation network for simultaneous depth estimation and scene parsing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 675–684 (2018)
32. Ye, W., Ma, K.K.: Color image demosaicing using iterative residual interpolation. IEEE Transactions on Image Processing **24**(12), 5879–5891 (2015)
33. Zagoruyko, S., Komodakis, N.: Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In: Proceedings of the International Conference on Learning Representations (2017)

34. 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)
35. Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.H., Shao, L.: Multi-stage progressive image restoration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14821–14831 (2021)
36. Zhang, C., Peng, Y.: Better and faster: Knowledge transfer from multiple self-supervised learning tasks via graph distillation for video classification. In: Proceedings of the International Joint Conference on Artificial Intelligence. pp. 1135–1141 (2018)
37. 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)
38. Zhang, Z., Wang, H., Liu, M., Wang, R., Zhang, J., Zuo, W.: Learning RAW-to-sRGB mappings with inaccurately aligned supervision. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4348–4358 (2021)