Attention-aware Learning for Hyperparameter Prediction in Image Processing Pipelines

Haina Qin[1,2], Longfei Han[3], Juan Wang[1], Congxuan Zhang[4], Yanwei Li[5], Bing Li[1,2(✉)], and Weiming Hu[1,2]

[1] NLPR, Institute of Automation, Chinese Academy of Sciences
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences
{qinhaina2020@,jun_wang@,bli@nlpr.,wmhu@nlpr.}ia.ac.cn
[3] Beijing Technology and Business University longfeihan@btbu.edu.cn
[4] Nanchang Hangkong University [5] Zeku Technology, Shanghai

Abstract. Between the imaging sensor and the image applications, the hardware image signal processing (ISP) pipelines reconstruct an RGB image from the sensor signal and feed it into downstream tasks. The processing blocks in ISPs depend on a set of tunable hyperparameters that have a complex interaction with the output. Manual setting by image experts is the traditional way of hyperparameter tuning, which is time-consuming and biased towards human perception. Recently, ISP has been optimized by the feedback of the downstream tasks based on different optimization algorithms. Unfortunately, these methods should keep parameters fixed during the inference stage for arbitrary input without considering that each image should have specific parameters based on its feature. To this end, we propose an attention-aware learning method that integrates the parameter prediction network into ISP tuning and utilizes the multi-attention mechanism to generate the attentive mapping between the input RAW image and the parameter space. The proposed method integrates downstream tasks end-to-end, predicting specific parameters for each image. We validate the proposed method on object detection, image segmentation, and human viewing tasks.

1 Introduction

Hardware ISPs are low-level image processing pipelines that convert RAW sensor data into images suitable for human viewing and downstream tasks. Hardware ISPs introduce several processing blocks that are less programmable and operate efficiently at real-time applications [43, 8]. ISPs are widely used in a variety of devices like cameras [3], smartphones, self-driving vehicles, and surveillance [30].

Existing processing blocks in hardware ISPs are configurable and sensitive with a set of user-tunable hyperparameters. These hyperparameters affect not only the output images but also the downstream tasks [31, 13]. It is important and still challenging to find optimal ISP hyperparameters for different specific tasks. In general, the industries rely on image experts to manually tune the parameters on a small typical dataset [1]. This artificial process is time-consuming

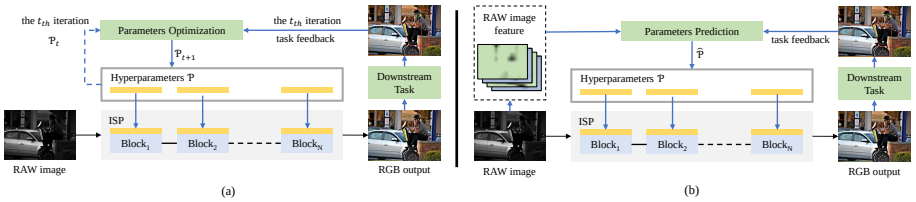

Fig. 1. Illustrations of different ISP tuning methods. (a) Previous methods leverage a mapping between the parameter space and high-level evaluation metrics based on differentiable approximations or hardware ISP. (b) Our proposed framework firstly constructs an attention-aware prediction network between RAW sensor data and parameter space, and then follows the previous work to create the mapping function between the parameter space and high-level evaluation metrics.

and biased toward human perception, and especially hard to subjectively find task-specific optimal hyperparameters for various downstream tasks, such as object detection and image segmentation [38, 39].

Hence, the potential of automated loss-based ISP hyperparameter optimization [35, 40] comes into sight. A full grid search is not an alternative way due to the large parameter space. Instead, several recent works reproduce the entire ISP transformation with software approximations [20, 17], then implement derivative-free methods [27] or gradient methods based on differentiable approximation [36, 15, 33]. These methods leverage a relationship between high-level evaluation metrics and the parameter space but ignore the mapping from the input RAW images to the parameter space. In addition, some methods try to directly optimize the hardware ISP with evolution strategy[11] in an end-to-end way [26]. They chose reasonably reduced search spaces and let the evolutionary algorithms do the exploration. However, during the inference stage, these methods should set fixed hyperparameters tuned in the training stage, which leads to being eclectic, not discriminative for a wide variety of input images.

In this work, we tune the ISP based on the statistical relationship among the input image, the parameter space, and the high-level evaluation metrics of downstream tasks. The optimal ISP settings should have high relevance with raw sensor data (low-level pixel-wise information [23], such as texture, exposure) and the scenario of downstream tasks (high-level semantic information, such as object location and object categories [9]). So we decouple the ISP pipelines into two modules: attention-aware prediction network and differentiable proxy network. The first module aims to construct a mapping function from RAW sensor data to parameters space, and the second one tries to reveal the relationship between the parameter space and high-level evaluation metrics, as shown in Figure 1.

To construct a mapping function from RAW sensor data to parameters space, we propose an attention-aware prediction network for ISP hyperparameter prediction, which is able to learn the natural information from the RAW low-level feature through Parameter Prediction Path. Further on, it is important to attach weights [24] to specific locations, which makes different hyperparameters distinc-

tively contribute to the input image. Hence, we introduce a multi-scale attention mechanism, named Multi-Attention Path, into the parameter prediction path, which aims to better represent the discrimination of different processing blocks. Multi-attention path enables the network exchange and aggregates the multi-scale information, and highlights the activation map for each hyperparameter of ISP by adaptively selecting multi-scale features. Meanwhile, to reveal the relationship between the parameter space and high-level evaluation metrics, we use a differentiable proxy network to mimic hardware ISP and use the output of the proxy network as the input for downstream tasks.

In our proposed framework, two differentiable networks, attention-aware prediction network and differentiable proxy network based on fully convolutional network(FCN) [36], are constructed and optimized end-to-end by using feedback from the high-level downstream task, and the predicted parameters of ISP are task-specific. We validate the proposed method in a variety of applications, including object detection, image segmentation, and human viewing. For these applications, we demonstrate that our method has better results compared with manual tuning methods and existing numerical optimization methods.

The contributions of this paper can be summarized as follows:

- We propose a novel framework for hyperparameter prediction in ISP that directly infers parameters based on RAW images while integrating downstream tasks in an end-to-end manner. For inference, our method can give distinctive results for each image suitable to the downstream task.
- We introduce a Multi-Attention structure for the parameter prediction network. It enables the network exchange and aggregate the multi-scale information, and highlight the activation map for hyperparameter of ISP.
- We validate the effectiveness of ISP hyperparameter prediction on 2D object detection, image segmentation, and human observation tasks. In these applications, our method approach outperforms existing numerical optimization methods and expert tuning.

Limitations: In the objective evaluation, such as object detection and image segmentation, the performance of the proposed method is compared with recent existing methods based on synthetic ISPs. However, the synthetic ISPs used have similar processing pipelines but not exactly the same. Further on, the subjective evaluation is compared with expert tuning methods, which bias towards human visual perception. Therefore, it is important to build a standard Synthetic ISP and standard subjective evaluation metrics. We believe that the release of relevant standard modules will be critical for future works on ISP tuning.

2 Related Work

There are several image processing components in the ISP pipelines. In traditional ISP, a specific algorithm is developed for each associated ISP component. Such a divide-and-conquer strategy decomposes the complex ISP design problem

into many sub-tasks. These sub-problems are always formed by tens to hundreds of handcrafted parameters and tuned towards to perception of imaging experts.

Recently, to tackle this challenging optimization problem, several automatic ISP tuning methods optimize the hyperparameters with downstream task feedback[5, 4, 44]. The impact of ISP hyperparameter on the performance of a downstream task is well explored in [2, 41, 7, 39]. Tseng *et al.* [36] optimizes ISP for object detection and classification using IoU. Mosleh *el al.* [26] utilizes object detection and object segmentation with mAP and PQ. Wu *et al.* [38] optimizes a simple ISP with an object detection task.

With the high-level feedback information, recent works always leverage a differentiable mapping between the parameter space and high-level evaluation metrics. Some existing optimization methods explore the optimal parameters from reasonably large reduced search spaces via an implicit end-to-end loss. For instance, Pfister *et al.* [29] proposes to optimize sparsity regularization for denoising. Nishimura *et al.* [27] optimizes software ISP model with a 0-th order Nelder-Mead method. However, it can only be used to optimize one ISP component at a time. Mosleh *et al.* [26, 32] directly optimizes hardware ISP by a novel CMA-ES strategy[12] with max-rank-based multi-objective scalarization and initial search space reduction. Robidoux *et al.* The other methods try to reproduce the entire ISP transformation with a CNN-based differentiable proxy [15, 10, 42]. For instance, Tseng *et al.* [36] trained an approximate CNN proxy model to mimic hardware ISP and optimized the differentiable CNN model with Stochastic Gradient Descent. Kim *et al.* [19] utilize the objective function of multi-output regression for modeling the relation between ISP parameter and IQM score. Onzon *et al.* [28] propose a neural network for exposure selection that is jointly end-to-end with an object detector and ISP pipeline. However, these methods should set fixed hyperparameters during the inference, which lack diversity and discrimination for various input RAW images.

Therefore, we believe that ISP tuning should require more knowledge to reveal the relationship among the raw image, parameter space, and the high-level evaluation metrics. The difference between our approach and the others is that we both construct an attention-aware prediction network and the differentiable proxy network. We not only train an ISP proxy to approximate the entire ISP as a RAW-to-output RGB image transfer function, but we also construct the parameter prediction network to explicitly joint optimized the trainable ISP with the downstream vision tasks.

3 Image Processing Pipelines

ISPs are low-level pipelines composed of many processing stages, generally converting RAW sensor pixels into human-viewing images. We briefly review the most common ISP modules and their associated parameters. The ISP whose parameters are optimized in this paper contains the following typical stages [6]:

(1) Optics and Sensor: The scene radiation is focused on the sensor through an assembly of lenses. The color filter array on the camera filters the light into

three sensor-specific RGB primaries, and the RAW pixel output of the sensor is linearly related to the irradiance falling on the sensor.

(2) Noise Reduction: Denoising is applied after the A/D conversion, which is done by blurring the image. Blurring will reduce noise but also remove details.

(3) DigitalGain and White Balance: After removing the black level bias and correcting defective pixels, the imaging sensor signal is amplified and digitized. The pixel values are color-corrected and gain-adjusted according to the white-balance matrices for common or automatically estimated illuminations.

(4) Demosaicking: Convert the color filter array over pixel sensors to RGB values for each pixel by performing interpolation.

(5) Color Space Transform: Map the white-balanced raw-RGB values to CIE XYZ using a 3x3 color space transform matrix, where CIE XYZ is a canonical color space definition.

(6) Sharpening: Compensate the outline of the image, enhance the edges and the part of the grayscale jump, make the image details enhanced.

(7) Color and Tone Correction: This is the stage to improve overall image appearance, including applying gamma curves and adjusting image contrast by histogram operations.

(8) Compression: Pixels values compressed to JPEG and storage.

Our ISP model f_{ISP} takes RAW Pixel values as input and models stages (2) to (8). This ISP converts the RAW image $\mathbf{I}$ into an RGB image $\mathbf{O}_{\text{ISP}}$.

$$\mathbf{O}_{\text{ISP}} = f_{\text{ISP}}(\mathbf{I}; \mathcal{P}),\ \mathcal{P} \in \mathbb{R}^{N}_{[0,1]} \tag{1}$$

where $\mathbf{I} \in \mathbb{R}^{W \times H}$, $\mathbf{O}_{\text{ISP}} \in \mathbb{R}^{W \times H \times 3}$. The conversion is modulated by the values of N continuous hyperparameters $\mathcal{P}$ with a range of values normalized to $[0, 1]$. For the discrete parameters in ISP, mapping them to continuous values within the range of values facilitates prediction [26].

4 Method

4.1 Framework

For a RAW image $\mathbf{I}$ from an imaging sensor, our aim is to predict N parameters $\mathcal{P} = \{p_1, p_2, \ldots, p_{\text{N}}\}$ for the target ISP. The original image $\mathbf{I}$ generates an RGB output image $\mathbf{O}_{\text{ISP}} = f_{\text{ISP}}(\mathbf{I}; \mathcal{P})$ after the ISP processing under the parameter $\mathcal{P}$ setting, and $\mathbf{O}_{\text{ISP}}$ is used as the RGB image input for the downstream task. In this paper, we focus on downstream tasks that conform to human viewing preferences and visual analysis tasks.

Our approach is to learn an approximation function $\hat{\mathcal{P}} = f_{\text{pred}}(\mathbf{I}; \mathbf{W})$ that directly utilizes the raw sensor data to predict the parameter $\mathcal{P}$, where $\mathbf{W}$ denotes the trainable weights. Then we utilize the feedback from the downstream task, making $\mathbf{O}_{\text{ISP}} = f_{\text{ISP}}(\mathbf{I}; \hat{\mathcal{P}})$ favorable for it. The performance of f_{pred} is determined by the parameters $\mathbf{W}$ of the network, and then $\mathbf{W}$ is learned by minimizing the high-level loss function.

$$L_{\text{task}}(\mathbf{O}_{\text{ISP}}) = L_{\text{task}}(f_{\text{ISP}}(\mathbf{I}; \hat{\mathcal{P}})) = L_{\text{task}}(f_{\text{ISP}}(\mathbf{I}; f_{\text{pred}}(\mathbf{I}; \mathbf{W}))) \tag{2}$$

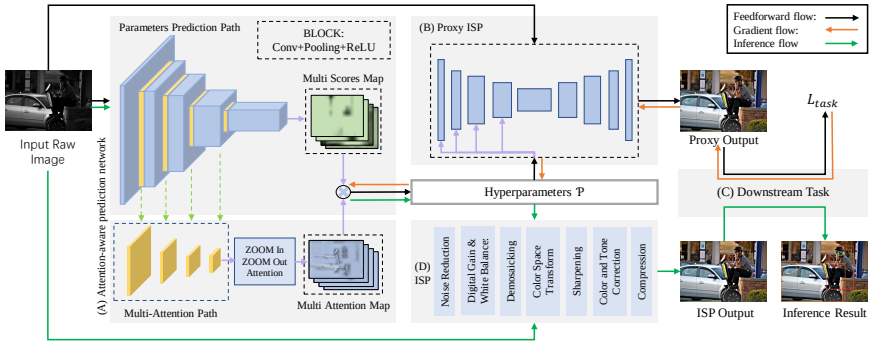

Fig. 2. Proposed attention-aware Learning framework and proxy architecture. (a) Attention-aware prediction network, each block consists of 3x3 conv, max pooling, and ReLU unit ; (b) Training ISP differentiable CNN proxy for mimic the hardware ISP; (c) Fixing learned parameters of the CNN-based ISP and the downstream task model, and optimizing input parameters itself given a high-level loss function. (d) Run-time execution using Hardware ISP architecture with predicted parameters.

where L_{task} is defined as the loss function for the downstream task, e.g., $\mathcal{L}_2$ loss with the reference image is used for conforming to human viewing preferences, and loss with a combination of classification and location regression is used for the object detection task. Since ISPs are non-differentiable black-box units, in order to utilize feedback from downstream tasks for parameter prediction network, a differentiable proxy network [36] based on FCN is built. The black-box ISPs are modeled as shown in Eq. 1. The proxy ISP $f_{\text{PROXY}}(\mathbf{I}, \mathcal{P}; \mathbf{W}_{\text{proxy}})$ consists of an fully connected network, taking $\mathbf{I}$ and hyperparameters $\mathcal{P}$ as inputs, and $\mathbf{W}_{\text{proxy}}$ as f_{PROXY} learnable CNN weights. The proxy should achieve such a goal: $\mathbf{O}_{\text{PROXY}} \approx \mathbf{O}_{\text{ISP}}$. The optimal weights $\mathcal{W}^*_{\text{proxy}}$ are optimized by minimizing the loss function and then froze for training the prediction network:

$$L_{\text{proxy}} = ||f_{\text{PROXY}}(\mathbf{I}, \mathcal{P}; \mathbf{W}_{\text{proxy}}) - f_{\text{ISP}}(\mathbf{I}; \mathcal{P})||. \tag{3}$$

After f_{pred} predicts the parameters $\hat{\mathcal{P}}$ of $\mathbf{I}$, $\hat{\mathcal{P}}$ with $\mathbf{I}$ are fed to the proxy ISP f_{PROXY} to produce an output image which is used as input for the downstream task. Since the proxy function f_{PROXY} is differentiable for $\mathbf{W}$, we can achieve an end-to-end learning process to jointly optimize the whole framework. The process of optimizing $\mathbf{W}$ on M images can be performed on the parameter prediction network f_{pred} and FCN-based proxy network using the supervised information from downstream tasks.

$$\mathbf{W}^*_{\text{task}} = \underset{\{\mathbf{W}\}}{argmin} \sum_{i=1}^{M} L_{\text{task}}(f_{\text{PROXY}}(\mathbf{I}_i, f_{\text{pred}}(\mathbf{I}_i; \mathbf{W}); \mathbf{W}^*_{\text{proxy}})). \tag{4}$$

In the inference stage, the parameter prediction network f_{pred} can estimate the optimal parameters $\mathcal{P}^*_{\text{task}}$ for the downstream task based on the raw data $\mathbf{I}$.

$$\mathcal{P}^*_{\text{task}} = f_{\text{pred}}(\mathbf{I}; \mathbf{W}^*_{\text{task}}). \tag{5}$$

Further on, aiming to discriminate the effects of the parameters on input images, we design a novel structure based on a multi-scale attention structure. Figure 2 illustrates our network structure.

4.2 Attention-aware Parameter Prediction Network

Parameter Prediction Path The semantic information contained in the RAW Image largely determines the parameter fetching, so we design a parameter prediction network that generates a prediction path with multiple encoders to implement the encoding from the RAW Image to the target parameters, which is illustrated in Figure 2. For the input Bayer array, we split the input data into three channels with RGB values respectively. Then, the formatted input is fed into a fully convolutional neural network. Followed by several convolutional blocks, the spatial resolution of the feature map is smaller than the size of the input RAW Image, and the number of channels is the number N of parameters $\mathcal{P}$. The output feature maps are finally passed to a weighted pooling layer for local-to-global aggregation:

$$p_i = \sum_{j=1}^{m} c_i(R_j) g_i(R_j),\ i = 1, \ldots, N \tag{6}$$

where $\mathcal{R} = \{R_1, R_2, \ldots, R_m\}$ is a set of overlapping perceptual field regions of the original image $\mathbf{I}$, m is the number of R. Meanwhile, different local regions will generate different local parameter prediction $g_i(R_j)$ because they have different semantic information. The $c_i(R_j)$ represents the degree of attention of the parameter p_i for the local region R_j, which will be expressed in detail in Multi-Attention Path.

For predicting the value $g_i(R_j)$ of the parameter p_i on the local region R, we utilize five encoders for extracting the higher-level features from the image. In Parameter Prediction Path, different local regions in the image are progressively abstracted into high-level representations. Each encoder contains a 3x3 convolution followed by a ReLU unit and a 2x2 maximum pooling operation for span-2 downsampling. During each encoder, the number of feature channels is doubled. Subsequent convolution kernels with 1x1xN are used to downsample the feature maps while generating parameter estimation maps for N channels.

Multi-Attention Path We believe that the parameter prediction network should highly focus on the semantically informative parts while ignoring the semantically ambiguous regions on the input image. For this purpose, we added the attention path to the network, shown in Eq. 6. For the specific local region R, the value of the function $c_i(R)$ will reflect the effectiveness of the parameter p_i to the image region R. If the semantic information on R is informative for the setting of the parameter p_i, the value of $c_i(R)$ will be large, which will make the prediction p_i on R, $g_i(R)$ have a greater influence.

Since different parameters p of ISP processing blocks have different functions, the attentive regions are not the same for different p. Here, we combine the multi-scale features from the parameter prediction path to generate Multi-level attention, and Figure 2 shows the structure. Specifically, we add one channel to each

encoder. The feature maps on these five channels are named Multi-Level features $\{\mathbf{F}_1,\mathbf{F}_2,\mathbf{F}_3,\mathbf{F}_4,\mathbf{F}_5\}$, which contains different levels of semantic information of the input image. Then, we introduce a zoom-in-zoom-out attention module, which utilizes interpolation (UpSampling) and max-pooling (DownSampling) operations to keep the same size as $\mathbf{F}_3$, and concatenates the feature maps separately. Finally, the multi-scale features are downsampled twice, and undergo convolution of 3x3 and 1x1xN to generate the Multi-Attention Path. The size is the same as the size of the parameter prediction path outputs, and the number of channels for both maps is the number of parameters N. The i_{th} channel are corresponding to the parameter p_i.

The prediction path $G(R)$ generates the following set of parameter predictions, which correspond to local regions R for different parameters p:

$$G(R) = \{g_1(R), g_2(R), \dots g_N(R)\} \tag{7}$$

At the same time, the multi-attention path $C(R)$ can predict the attention of the local region R_j for different parameters p:

$$C(R) = \{c_1(R), c_2(R), \dots c_N(R)\} \tag{8}$$

Finally, as in the Eq. 6, the results of the parameter prediction path and multi-attention path are integrated to generate the global prediction results for all parameters. In this process, each parameter can automatically select and fuse the attention at different levels to obtain targeted regions.

5 Experiments

5.1 Settings

We validate the proposed ISP hyperparameter prediction method on the following downstream tasks and datasets:

(1) 2D object detection using [31] on MS COCO [22]. Using a synthetic (simulated) ISP processing simulated RAW as the upstream module for the task. The processing blocks of the ISP are described in Sec.3 with 20 ISP hyperparameters.

(2) Instance segmentation using [13] on COCO, which has the same ISP settings as the 2D object detection task.

(3) Perceptual image quality for human viewing. The dataset was collected by SONY IMX766 CMOS sensor. The reference images are obtained by Qualcomm Spectra 580 ISP processing, with 32 expert-tuned ISP hyperparameters. The synthetic ISP processing 4096x3072 RAW data from the sensor is the upstream module for the task.

In the training stage, the RAW image is the input to the model and generates the corresponding predicted hyperparameters. The predicted parameters and the RAW image are used as input to the proxy ISP, which later outputs the RGB image. The proxy ISP has fixed network weights as described in Sec.4.1. We train the parameter prediction network from scratch using the loss of the

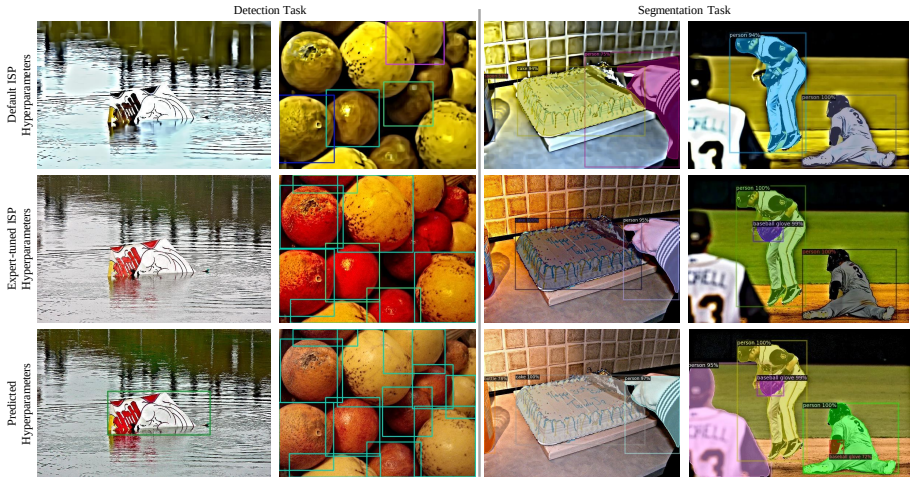

Fig. 3. Image understanding evaluation: Object Detection on COCO (left), Image Segmentation on COCO (right). Default ISP hyperparameters (top), expert-tuned hyperparameters (middle), and our method (bottom). Our prediction achieve better performance for downstream tasks.

downstream task and the RMSprop optimizer. For the object detection and image segmentation tasks, we use the loss with a combination of classification and location regression; for the human viewing optimization task, we use the $\mathcal{L}_2$ loss between the RGB output and the reference image. The learning rate is initially set to 10^{-4} and reduced to 10^{-6} after 200 epochs. The training was performed for 400 epochs. More details of the training procedure are described in the supplementary material.

In the evaluation stage, we use the RAW image from the test set as model input to predict the hyperparameters for each RAW image. The RAW image and hyperparameter pairs are processed into the ISP to obtain the RGB output. For the object detection and image segmentation tasks, the RGB outputs are used as input, and the evaluation metric uses mean average precision(mAP) [22]. The PSNR and SSIM between the RGB outputs and the corresponding reference images are used as evaluation metrics for the human viewing optimization task.

5.2 ISP Hyperparameter Prediction for Object Detection

In this task, we use an existing sRGB dataset and a synthetic ISP to evaluate our hyperparametric prediction method. Since the input of the ISP is RAW data, we processed sRGB images with RAW data simulation [18]. For the task of object detection using [31] on the MS COCO dataset [22], we trained and tested our hyperparameters prediction results. The results in Table 1 demonstrate that our method has a greater improvement in the default parameters than the recent methods and has the best final results. While the images produced by our expert-tuned ISP are more consistent with human perception (see Figure 3), the images

Table 1. Synthetic ISP optimization for Object Detection on COCO. It is important to point out that our ISP has similar processing pipelines compared with the ISPs used in other methods. So we borrow the results from [26].

	ISP Model	$mAP_{0.5}$	$mAP_{0.75}$	$mAP_{0.5:0.95}$
Default Parameters	Synthetic ISP [26]	0.15	-	-
blockwise-tuned for Object Detection		0.20	-	-
Expert-tuned for Image Quality		0.35	-	-
Hardware-tuned for Object Detection		0.39	-	-
Default Parameters	Synthetic ISP Sec. 3	0.34	0.22	0.21
Expert-tuned for Image Quality		0.56	0.40	0.37
Predicted Parameters (Ours)		**0.61**	**0.44**	**0.41**

produced by our predicted parameters result in better performance for object detection tasks. These images have emphasized texture details and color features that are more in line with the preferences of the object detection compared to the expert-tuned images.

5.3 ISP Hyperparameter Prediction for Image Segmentation

Table 2. Synthetic ISP optimization for Image Segmentation on COCO. It is important to point out that our ISP has similar processing pipelines compared with the ISPs used in other methods. So we borrow the results from [26].

	ISP Model	$mAP_{0.5}$	$mAP_{0.75}$	$mAP_{0.5:0.95}$
Default Parameters	Synthetic ISP [26]	0.12	-	-
Expert-tuned for Image Quality		0.26	-	-
Hardware-tuned for Segmentation		0.32	-	-
Default Parameters	Synthetic ISP Sec. 3	0.22	0.13	0.12
Expert-tuned for Image Quality		0.46	0.28	0.27
Predicted Parameters (Ours)		**0.52**	**0.33**	**0.31**

For the image segmentation task, we trained the prediction network end-to-end with [13] a downstream task and validated the results on synthetic ISP and simulated RAW COCO datasets. The results in Table 2 are shown that our method has more significant improvement than other methods, and better final results, especially better than the default parameters(baseline), Fig. 3 demonstrates an example of our method with default parameters and expert-tuned parameters for instance segmentation. It can be seen that our predicted parameters can adjust the texture and color of the image to match the preferences of the image segmentation task. The parameters predicted by our model for each

image achieve an improvement of 0.3 $\text{mAP}_{0.5}$ compared to the default parameters. Also, compared to the imaging experts tuned parameters for the task, the parameters predicted by our model can achieve a $\text{mAP}_{0.5}$ improvement of 0.1.

5.4 ISP Hyperparameter Prediction for Human Viewing

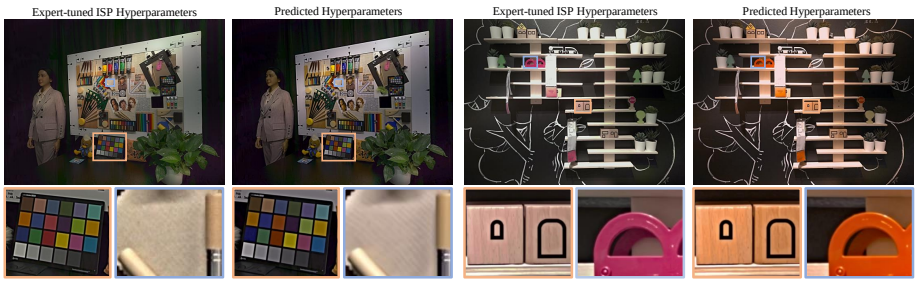

Fig. 4. Comparison of Expert-tuned hyperparameters and our method. These images are processed by a synthetic ISP described in Sec. 3 to match the image quality of the reference images generated by the Qualcomm Spectra580 ISP.

Unlike visual analysis tasks, subjective image quality is an attribute that describes a preference for a particular image rendering [25]. This particular image rendering should consider the visibility of the distortions in an image, such as colour shifts, blurriness, noise, and blockiness [34]. Our goal is to predict the hyperparameters on the synthetic ISP corresponding to the RAW image such that the distance between the output of the synthetic ISP and the reference image is minimized.

In this task, we train the hyperparameter prediction network using the $\mathcal{L}_2$ distance between the RGB output and the reference image as the loss function described in the previous section. The proposed method predicts the hyperparameters of the synthetic ISP corresponding to the RAW image. We collected a new dataset for training and testing the performance of the proposed method on this task. To sufficiently validate the model, this dataset contains images from 108 different scenes. The RAW image is acquired by the IMX766 sensor, and the corresponding reference image is generated by the Spectra 580 ISP, where the hyperparameters of the ISP are set manually by our imaging experts. We compared with the expert-tuned synthetic ISP output and the default hyperparameter settings, and the results are shown in Figure 5 (a). Compared with the expert-tuned parameters, we have better PSNR and SSIM results between the reference image and the RGB output produced by the predicted parameters. This indicates that our predicted parameters are more consistent with human preferences. Figure 4 shows our results on the human viewing task. It can be seen that our predicted parameters produce images with clearer texture details, better noise control, and human-adapted color features than the expert-tuned.

	PSNR	SSIM
Default $\hat{\mathcal{P}}$	10.97	0.39
Expert-tuned	17.11	0.57
Predicted $\hat{\mathcal{P}}$	**19.96**	**0.73**

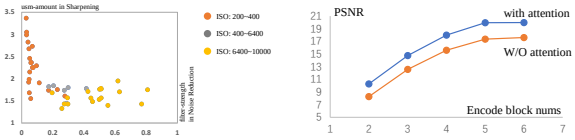

Fig. 5. (a)Left: ISP tuning for Perceptual image quality using expert-tuned and the proposed methods. (b)Middle: The scatter plot of the predicted hyperparameters on different ISO ranges. (c)Right: The PSNR perfomance versus the number of encoding blocks on the human viewing dataset.

5.5 Ablation Study

Parameter Prediction Path In general, the industries rely on image experts to manually tune the parameters on a small typical dataset. The automatic ISP methods set fixed hyperparameters tuned with downstream tasks in an end-to-end way. In contrast, the proposed method can predict specific parameters for each input image. We verify that this prediction method for RAW images has better results than fixed parameters in this ablation study.

To demonstrate the diversity of the predicted parameters across images, we chose several images based on different ISO ranges and plot scatter plots with two parameters in the Noise Reduction module and the Sharpening module, as shown in Fig. 5 (b). It shows that our method can self-adaptively optimize the parameters on different lighting conditions. For noisy images in low light conditions (high ISO), the predicted parameter is preferred to increase the noise reduction level (filter strength) on the BM3D module. The results in Fig. 5 (a) show that distinctive prediction parameters provide better performance than fixed expert-tuned parameters. We also evaluate the performance and efficiency of our network architecture design. We test the network performance by changing the number of encoding blocks on the human viewing dataset. On this basis we further validate our attention module, shown in Fig. 5 (c). Meanwhile, compared with the previous methods [35, 26] which required hundreds of loops in the ISP tuning process, our method is more efficient, a 4096x3072 RAW image with a 2x down-sample can be processed in 0.5s by our method (on an NVIDIA RTX3090) and 129.39 GFLOPS.

Also, we selected 300 representative images from the COCO [22] training set and predict parameters by the proposed method. The mode of the discrete predicted parameters and the mean of the continuous parameters as the fixed hyperparameters, and fed into the synthetic ISP with the corresponding raw image to generate sRGB output for detection and segmentation tasks. Table 3 demonstrate the value of diverse parameters for different images; averaging and fixing the diverse parameters is a compromise process. It is important and challenging to find optimal ISP hyperparameters for specific tasks.

Multi-Attention Path In Multi-Attention Path, the scoring map for each parameter is multiplied by its corresponding attention map. The prediction results reflect the distinctive contributions of image patches to multi-target parameters. Figure 6 shows the attention maps created by our feature attention-based

Table 3. Ablation study of evaluate the visual analysis results based on fixed parameters and diverse predicted parameters. And effectiveness analysis on proposed multi-attention path. The first column is the results for object detection, and the second is the results for image segmentation.

	Detection			Segmentation		
	mAP 0.5	mAP 0.75	mAP 0.5:0.95	mAP 0.5	mAP 0.75	mAP 0.5:0.95
Predicted $\hat{\mathcal{P}}$ (Fixed)	0.57	0.40	0.38	0.48	0.29	0.27
Predicted $\hat{\mathcal{P}}$ (W/O Attention)	0.57	0.41	0.38	0.47	0.30	0.28
Predicted $\hat{\mathcal{P}}$ (Single Attention)	0.58	0.43	0.40	0.50	0.31	0.30
Predicted $\hat{\mathcal{P}}$ (Self-attention [37])	0.57	0.42	0.38	0.48	0.30	0.28
Diverse Predicted $\hat{\mathcal{P}}$ (Ours)	**0.61**	**0.44**	**0.41**	**0.52**	**0.33**	**0.31**

network. The Sharpening module (second row, prediction $\hat{\mathcal{P}}_1$) does not significantly affect texture-flat areas so that the attention map activates the features in texture-rich areas in the image. The Noise Reduction module (third row, prediction $\hat{\mathcal{P}}_2$) is prone to remove noise in the background areas so that the background with details in the image are activated via the attention map. For $\hat{\mathcal{P}}_N$ (bottom row) in the WB module, a wider range of color features are activated as reference areas for color prediction.

To verify the effectiveness of multi-attention, we replace the Multi-attention Path with Single-Attention Path (the same attention map is used for all parameters predictions). We also test the results by removing the attention path and only using Parameters Prediction Path for parameter prediction. It can be seen from Table 3 that the methods with attention path have better performance than those without attention module. The results indicate that it is a benefit for learning the mapping between the raw input image and the parameter space by highlighting high-value image regions. Meanwhile, the performance of uniform single-attention for multi parameters is not well performed than Multi-Attention scheme, which indicates that the effectiveness of the proposed multi-attention path. In addition, in contrast to existing common designs, the attention structure is specifically designed for ISP parameters prediction tasks, highlighting different image details for various parameters predictions. To compare the multi-attention structure with other common designs, we replace the multi-attention structure by combining Parameters Prediction Path with Self-attention [37]. Experimental results show that the multi-attention structure that can combine features at different scales and generate attention maps is more suitable for ISP hyperparameters prediction tasks and better performance.

6 Conclusions and future work

In this paper, we propose a novel ISP tuning framework named attention-aware learning for hyperparameter prediction, compared with the existing CNN methods to simulate commercial ISPs, our method try to mimic the expert-tuning

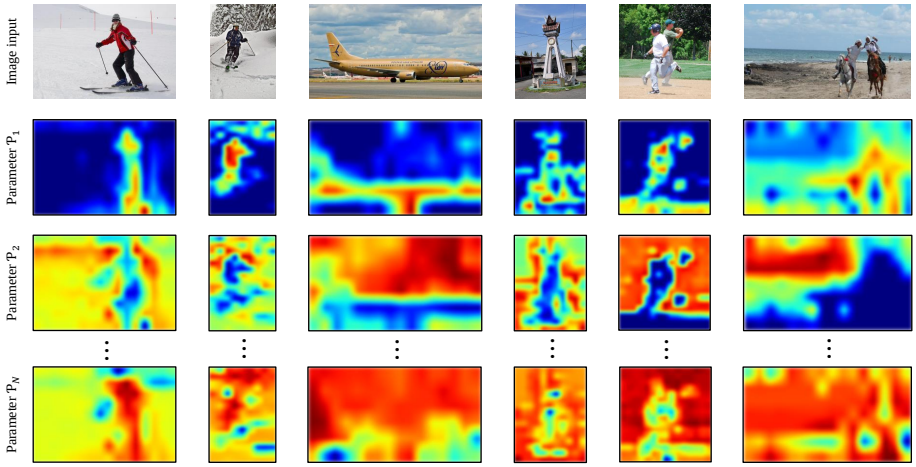

Fig. 6. Examples of attention map outputs by our network. It shows three attention maps corresponding to different parameters for each image. $\hat{\mathcal{P}}_1$ is in the sharpening module, $\hat{\mathcal{P}}_2$ is in the noise reduction module, and $\hat{\mathcal{P}}_N$ is in the WB module.

procedure via predicting the parameters for all ISP modules. It simultaneously constructs the mapping functions between RAW input & parameter space and the parameter space & high-level metrics. Considering that different hyperparameters make a distinctive contribution to the input image, we design a novel multi-attention structure for jointly predicting the hyperparamter. Experimental results demonstrate that the proposed method can improve the performance of models.

The presented method needs to predict the hyperparameters directly while the ISP processes the image and maintains the efficiency of ISP. The efficiency of the proposed method is a crucial issue for deployment to ISPs for application. The next step is to make the model lightweight through algorithms compression and acceleration, including pruning policy [14, 21] and quantization policy [16], can be an avenue of future research. Running the method efficiently on ISPs is one of our future works.

Acknowledgement

This work was supported by the National Key Research and Development Program of China (Grant No. 2020AAA0105802), the Natural Science Foundation of China (Grant No. 62036011,62192782, 61721004,62122086, 61906192, U1936204), the Key Research Program of Frontier Sciences, CAS, Grant No. QYZDJ-SSW-JSC040, Beijing Natural Science Foundation (No. 4222003).

References

1. Ieee standard for camera phone image quality. IEEE Std 1858-2016 (Incorporating IEEE Std 1858-2016/Cor 1-2017) pp. 1–146 (2017). https://doi.org/10.1109/IEEESTD.2017.7921676
2. Bardenet, R., Brendel, M., Kégl, B., Sebag, M.: Collaborative hyperparameter tuning. In: International conference on machine learning. pp. 199–207 (2013)
3. van Beek, P., Wu, C.T.R., Chaudhury, B., Gardos, T.R.: Boosting computer vision performance by enhancing camera isp. Electronic Imaging **2021**(17), 174–1 (2021)
4. Bergstra, J., Bardenet, R., Bengio, Y., Kégl, B.: Algorithms for hyper-parameter optimization. Advances in neural information processing systems **24** (2011)
5. Bergstra, J., Yamins, D., Cox, D.: Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. In: International conference on machine learning. pp. 115–123 (2013)
6. Brown, M.S., Kim, S.: Understanding the in-camera image processing pipeline for computer vision. In: IEEE International Conference on Computer Vision. vol. 3 (2019)
7. Buckler, M., Jayasuriya, S., Sampson, A.: Reconfiguring the imaging pipeline for computer vision. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 975–984 (2017)
8. Cao, Y., Wu, X., Qi, S., Liu, X., Wu, Z., Zuo, W.: Pseudo-isp: Learning pseudo in-camera signal processing pipeline from a color image denoiser. arXiv preprint arXiv:2103.10234 (2021)
9. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European Conference on Computer Vision. pp. 213–229 (2020)
10. Chen, C., Chen, Q., Xu, J., Koltun, V.: Learning to see in the dark. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3291–3300 (2018)
11. Cheung, E.C., Wong, J., Chan, J., Pan, J.: Optimization-based automatic parameter tuning for stereo vision. In: 2015 IEEE International Conference on Automation Science and Engineering (CASE). pp. 855–861 (2015)
12. Hansen, N., Müller, S.D., Koumoutsakos, P.: Reducing the time complexity of the derandomized evolution strategy with covariance matrix adaptation (cma-es). Evolutionary computation **11**(1), 1–18 (2003)
13. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 2961–2969 (2017)
14. He, Y., Zhang, X., Sun, J.: Channel pruning for accelerating very deep neural networks. In: Proceedings of the IEEE international conference on computer vision. pp. 1389–1397 (2017)
15. 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)
16. Jacob, B., Kligys, S., Chen, B., Zhu, M., Tang, M., Howard, A., Adam, H., Kalenichenko, D.: Quantization and training of neural networks for efficient integer-arithmetic-only inference. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2704–2713 (2018)
17. Karaimer, H.C., Brown, M.S.: A software platform for manipulating the camera imaging pipeline. In: European Conference on Computer Vision. pp. 429–444. Springer (2016)

18. Kim, S.J., Lin, H.T., Lu, Z., Süsstrunk, S., Lin, S., Brown, M.S.: A new in-camera imaging model for color computer vision and its application. IEEE Transactions on Pattern Analysis and Machine Intelligence **34**(12), 2289–2302 (2012)
19. Kim, Y., Lee, J., Kim, S.S., Yang, C., Kim, T., Yim, J.: Dnn-based isp parameter inference algorithm for automatic image quality optimization. Electronic Imaging **2020**(9), 315–1 (2020)
20. 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)
21. Lin, J., Rao, Y., Lu, J., Zhou, J.: Runtime neural pruning. Advances in neural information processing systems **30** (2017)
22. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755 (2014)
23. Liu, D., Wen, B., Jiao, J., Liu, X., Wang, Z., Huang, T.S.: Connecting image denoising and high-level vision tasks via deep learning. IEEE Transactions on Image Processing **29**, 3695–3706 (2020)
24. Majumdar, P., Singh, R., Vatsa, M.: Attention aware debiasing for unbiased model prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4133–4141 (2021)
25. Mantiuk, R.K., Tomaszewska, A., Mantiuk, R.: Comparison of four subjective methods for image quality assessment. In: Computer graphics forum. vol. 31, pp. 2478–2491. Wiley Online Library (2012)
26. Mosleh, A., Sharma, A., Onzon, E., Mannan, F., Robidoux, N., Heide, F.: Hardware-in-the-loop end-to-end optimization of camera image processing pipelines. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7529–7538 (2020)
27. Nishimura, J., Gerasimow, T., Sushma, R., Sutic, A., Wu, C.T., Michael, G.: Automatic isp image quality tuning using nonlinear optimization. In: 2018 25th IEEE International Conference on Image Processing. pp. 2471–2475. IEEE (2018)
28. Onzon, E., Mannan, F., Heide, F.: Neural auto-exposure for high-dynamic range object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7710–7720 (2021)
29. Pfister, L., Bresler, Y.: Learning filter bank sparsifying transforms. IEEE Transactions on Signal Processing **67**(2), 504–519 (2018)
30. Phan, B., Mannan, F., Heide, F.: Adversarial imaging pipelines. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16051–16061 (2021)
31. Redmon, J., Farhadi, A.: Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767 (2018)
32. Robidoux, N., Capel, L.E.G., Seo, D.e., Sharma, A., Ariza, F., Heide, F.: End-to-end high dynamic range camera pipeline optimization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6297–6307 (2021)
33. 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)
34. Thung, K.H., Raveendran, P.: A survey of image quality measures. In: IEEE international conference for technical postgraduates. pp. 1–4 (2009)

35. Tseng, E., Mosleh, A., Mannan, F., St-Arnaud, K., Sharma, A., Peng, Y., Braun, A., Nowrouzezahrai, D., Lalonde, J.F., Heide, F.: Differentiable compound optics and processing pipeline optimization for end-to-end camera design. ACM Transactions on Graphics **40**(2), 1–19 (2021)
36. Tseng, E., Yu, F., Yang, Y., Mannan, F., Arnaud, K.S., Nowrouzezahrai, D., Lalonde, J.F., Heide, F.: Hyperparameter optimization in black-box image processing using differentiable proxies. ACM Transactions on Graphics **38**(4), 27–1 (2019)
37. 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)
38. Wu, C.T., Isikdogan, L.F., Rao, S., Nayak, B., Gerasimow, T., Sutic, A., Ainkedem, L., Michael, G.: Visionisp: Repurposing the image signal processor for computer vision applications. In: IEEE International Conference on Image Processing. pp. 4624–4628. IEEE (2019)
39. Yahiaoui, L., Hughes, C., Horgan, J., Deegan, B., Denny, P., Yogamani, S.: Optimization of isp parameters for object detection algorithms. Electronic Imaging **2019**(15), 44–1 (2019)
40. Yang, C., Kim, J., Lee, J., Kim, Y., Kim, S.S., Kim, T., Yim, J.: Effective isp tuning framework based on user preference feedback. Electronic Imaging **2020**(9), 316–1 (2020)
41. Yogatama, D., Mann, G.: Efficient transfer learning method for automatic hyperparameter tuning. In: Artificial intelligence and statistics. pp. 1077–1085. PMLR (2014)
42. Yu, K., Li, Z., Peng, Y., Loy, C.C., Gu, J.: Reconfigisp: Reconfigurable camera image processing pipeline. arXiv preprint arXiv:2109.04760 (2021)
43. 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)
44. Zoph, B., Vasudevan, V., Shlens, J., Le, Q.V.: Learning transferable architectures for scalable image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8697–8710 (2018)