A General Decoupled Learning Framework for Parameterized Image Operators

Qingnan Fan, Dongdong Chen, *Member, IEEE*, Lu Yuan, Gang Hua, *Fellow, IEEE*, Nenghai Yu and Baoquan Chen

Abstract—Many different deep networks have been used to approximate, accelerate or improve traditional image operators. Among these traditional operators, many contain parameters which need to be tweaked to obtain the satisfactory results, which we refer to as "parameterized image operators". However, most existing deep networks trained for these operators are only designed for one specific parameter configuration, which does not meet the needs of real scenarios that usually require flexible parameters settings. To overcome this limitation, we propose a new decoupled learning algorithm to learn from the operator parameters to dynamically adjust the weights of a deep network for image operators, denoted as the *base* network. The learned algorithm is formed as another network, namely the *weight learning* network, which can be end-to-end jointly trained with the *base* network. Experiments demonstrate that the proposed framework can be successfully applied to many traditional parameterized image operators. To accelerate the parameter tuning for practical scenarios, the proposed framework can be further extended to dynamically change the weights of only one single layer of the *base* network while sharing most computation cost. We demonstrate that this cheap parameter-tuning extension of the proposed decoupled learning framework even outperforms the state-of-the-art alternative approaches.

Index Terms—Image Processing and Computer Vision, Filtering, Restoration, Smoothing

1 Introduction

Image operators are fundamental building blocks for many computer vision tasks, such as image filtering and restoration. To obtain the desired results, many of these operators contain some parameters that need to be tweaked. We refer them as "parameterized image operators" in this paper. For example, parameters controlling the smoothness strength are widespread in most smoothing methods, and a parameter denoting the target upsampling scalar is always used in image super resolution.

Recently, many CNN based methods [1], [2], [3] have been proposed to approximate, accelerate or improve these parameterized image operators and achieved significant progress. However, we observe that the networks in these methods are often only trained for one specific parameter configuration, such as edge-preserving filtering [1] with a fixed smoothness strength, or super resolving low-quality images [2] with a particular downsampling scale. Many different models need to be retrained for different parameter settings, which is both storage-consuming and time-consuming. It also prohibits these deep learning solutions from being applicable and extendable to a much broader corpus of images.

In fact, given a specific network structure, when training separated networks for different parameter configurations $\overrightarrow{\gamma}_k$ as [1], [2], [3], the learned weights W_k are highly unconstrained and probably very different for each $\overrightarrow{\gamma}_k$. But can we find a common convolution weight space for different configurations by explicitly building their relationships? Namely, $W_k = h(\overrightarrow{\gamma}_k)$, where h can be a linear or non-linear function. In this way, we can adaptively change the weights of the single target network based on h in the runtime, thus enabling continuous parameter control.

To verify our hypothesis, we propose the first decoupled learning framework for parameterized image operators by decoupling the weights from the target network structure. Specifically, we employ a simple *weight learning* network $\mathcal{N}_{weight}$ as h to directly learn the convolution weights of one task-oriented *base* network $\mathcal{N}_{base}$. These two networks can be trained end-to-end. During the runtime, the *weight learning* network will dynamically update the weights of the *base* network according to different input parameters, thus making the *base* network generate different objective results. This should be a very useful feature in scenarios where users want to adjust and select the most visually pleasant results interactively.

We demonstrate the effectiveness of the proposed framework for many different types of applications, such as edge-preserving image filtering with different degrees of smoothness, image super resolution with different scales of blurring, and image denoising with different magnitudes of noise. We also demonstrate the extensibility of our proposed framework on multiple input parameters for a specific application, and combination of multiple different image processing tasks. Experimental results demonstrate that the proposed framework is able to achieve almost as good results as the one solely trained for a single parameter value.

- *Qingnan Fan is with Computer Science Department, Stanford University, Stanford, California 94305, US. E-mail: fqnchina@gmail.com*
- *Dongdong Chen and Nenghai Yu are with Department of Electronic Engineering and Information Science, University of Science and Technology of China, Hefei, Anhui 230026, China. E-mail: cd722522@mail.ustc.edu.cn, ynh@ustc.edu.cn*
- *Lu Yuan is with Microsoft Research, Redmond, Washington 98052, USA. E-mail: {luyuan}@microsoft.com*
- *Gang Hua is with Wormpex AI Research, Bellevue, Washington 98004, USA. E-mail: {ganghua}@gmail.com*
- *Baoquan Chen is with Peking University, Beijing 100871, China. E-mail: baoquan@pku.edu.cn*

Manuscript received September 27, 2018; revised April 17, 2019; accepted June 25, 2019. This work was supported in part by: National 973 Program (2015CB352501), NSFC-ISF (61561146397), the Natural Science Foundation of China under Grant U1636201 and 61629301. (Qingnan Fan and Dongdong Chen are co-first authors.)

Despite of its generality and flexibility, all the convolution weights of $\mathcal{N}_{base}$ need to be updated during the runtime. In the other words, the whole network needs to be re-evaluated when a new parameter is selected. This is very time-consuming and unacceptable in real user scenarios. To be adaptive for practical applications, the proposed framework can be further extended to dynamically change the weights of only one single layer while sharing most of the computation. With our default $\mathcal{N}_{base}$ network structure, experiments demonstrate this cheap parameter-tuning version outperforms the state-of-the-art methods [4], [5] by a large margin.

As an extra bonus, the proposed framework makes it easy to analyze the underlying working principle of the trained task-oriented network by visualizing different parameters. The knowledge gained from this analysis may inspire more promising research in this area. To sum up, the contributions of this paper lie in the following four aspects.

- We propose the first decoupled learning framework for parameterized image operators, where a *weight learning* network is learned to adaptively predict the weights for the task-oriented *base* network in the runtime.
- We show that the proposed framework can be learned to incorporate many different parameterized image operators and achieve very competitive performance with the one trained for a single specific parameter or operator.
- We extend our framework to enable cheap parameter tuning for real user scenarios, which outperforms many state-of-the-art methods by a large margin.
- We provide a unique perspective to understand the working principle of the trained task-oriented network with some valuable analysis and discussion, which may inspire more promising research in this area.

2 Related Work

In the past decades, many different image operators have been proposed for low level vision tasks. Previous work [6], [7], [8], [9] proposed different priors to smooth images while preserving salient structures. Some papers [10], [11] utilized the spatial relationship and redundancy to remove unpleasant noise in the image. Other papers [12], [13], [14] aimed to recover a high-resolution image from a low-resolution image. Among them, many operators are allowed to tune some built-in parameters to obtain different results, which is the focus of this paper.

With the development of deep learning techniques, many different neural networks are proposed to approximate, accelerate and improve these operators [1], [3], [15], [16], [17]. But their common limitation is that one model can only handle one specific parameter setting. To enable all other parameters, enormous different models need to be retrained, which is both storage-consuming and time-consuming. By contrast, our proposed framework allows us to input continuous parameters to dynamically adjust the weights of the task-oriented *base* network. Moreover, it can even be applied to multiple different parameterized operators with one single network.

Recently, Chen *et al.* [18] conducted a naive extension for parameterized image operators by concatenating the parameters as extra input channels to the network. Compared to their method, where both the network structure and weights maintain the same for different parameters, the weights of our *base* network are adaptively changed. Experimentally we find our framework outperforms their strategy by integrating multiple image operators. By decoupling the network structure and weights, our proposed framework also makes it easier to analyze the underlying working principle of the trained task-oriented network, rather than leaving it as a black box as in many previous works like [18].

To enable practical image processing on mobile devices, a simple scheme to accelerate an operator is to apply it at a low-resolution image then upsample the result by reintroducing high-resolution details, which is used in bilateral upsampling [19] and the fast guided filter [20]. To unify and generalize these two methods, [4] presents bilateral guided upsampling, which is further extended to [5] by incorporating such a technique into an end-to-end trained deep network. Though [4] and [5] are able to run very efficiently on even CPU devices, their resultant image quality is not good enough. By contrast, our cheap parameter-tuning extension directly runs the approximated operators on the original image resolution by sharing most of the computation costs for different operators, which not only enables real-time parameter adjustment on CPU devices but is also able to demonstrate its superior performance over these recent state-of-the-art approaches [4], [5] by a large margin.

Our method is also related to evolutionary computing and meta learning. Schmidhuber [21] suggested the concept of fast weights in which one network can produce context-dependent weight changes for a second network. Some other works [22], [23], [24] casted the design of an optimization algorithm as a learning problem, Recently, Ha *et al.* [25] proposed to use a static hypernetwork to generate weights for a convolutional neural network on MNIST and Cifar classification. They also leverage a dynamic hypernetwork to generate weights of recurrent networks for a variety of sequence modelling tasks. The purpose of their paper is to exploit weight sharing property across different convolution layers. But in our cases, we pay more attention to the common shared property among numerous input parameters and many different image operators.

3 Method

3.1 Problem Definition and Motivation

The input color image and the target parameterized image operators are denoted as $\mathcal{I}$ and $f(\overrightarrow{\gamma},\mathcal{I})$ respectively. $f(\overrightarrow{\gamma},\mathcal{I})$ transforms the content of $\mathcal{I}$ locally or globally without changing its dimension. $\overrightarrow{\gamma}$ denotes the parameters which determine the transform degree of f and may be a single value or a multi-value vector. For example, in L_0 smoothing[26], $\overrightarrow{\gamma}$ is the balance weight controlling the smoothness strength, while in RTV filter [8], it includes one more spatial gaussian variance. In most cases, f is a highly nonlinear process and solved by iterative optimization methods, which is very slow in runtime.

Our goal is to implement parameterized operator f with a base convolution network $\mathcal{N}_{base}$. In previous methods like [3], [17], given a specific network structure of $\mathcal{N}_{base}$, separated networks are trained for different parameter configurations $\overrightarrow{\gamma}_k$. In this way, the learned weights $\overrightarrow{W}_k$ of these separated networks are highly unconstrained and probably very different. But intuitively, for one specific image operator, the weights $\overrightarrow{W}_k$ of different $\overrightarrow{\gamma}_k$ might be related. So retraining separated models is too redundant. Motivated by this, we try to find a common weight space for different $\overrightarrow{\gamma}_k$ by adding a mapping constraint: $\overrightarrow{W}_k = h(\overrightarrow{\gamma}_k)$, where h can be a linear or non-linear function.

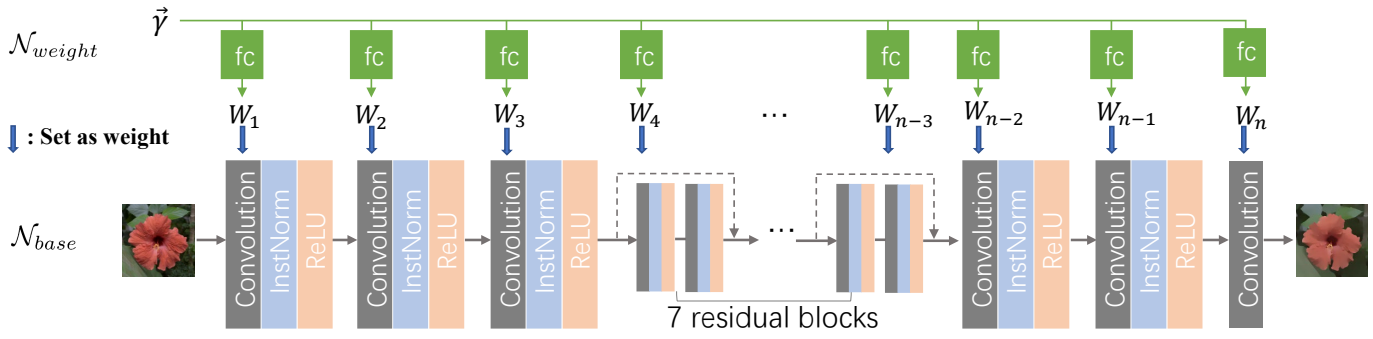


Fig. 1. Our system consists of two networks: the above *weight learning* network $\mathcal{N}_{weight}$ is designed to learn the convolution weights for the bottom *base* network $\mathcal{N}_{base}$. Given a parameterized image operator constraint by $\vec{\gamma}$, these two networks are jointly trained, and $\mathcal{N}_{weight}$ will dynamically update the weights of $\mathcal{N}_{base}$ for different $\vec{\gamma}$ in the inference stage.

In this paper, we directly learn h with another *weight learning* network $\mathcal{N}_{weight}$ rather than design it by handcraft. Assuming $\mathcal{N}_{base}$ is a fully convolutional network containing a total of n convolution layers, we denote their weights as $\vec{W}_k = (W_1, W_2, ..., W_n)$ respectively, then

$$(W_1, W_2, ..., W_n) = \mathcal{N}_{weight}(\vec{\gamma}) \tag{1}$$

where the input of $\mathcal{N}_{weight}$ is $\vec{\gamma}$ and the outputs are these weight matrices. In the training stage, $\mathcal{N}_{base}$ and $\mathcal{N}_{weight}$ can be jointly trained. In the inference stage, given different input parameter $\vec{\gamma}$, $\mathcal{N}_{weight}$ will adaptively change the weights of the target base network $\mathcal{N}_{base}$, thus enabling continuous parameter control.

Besides the original input image $\mathcal{I}$, the computed edge maps are shown to be a very important input signal for the target *base* network in [1]. Therefore, we also pre-calculate the edge map E of $\mathcal{I}$ and concatenate it to the original image as an extra input channel:

$$\begin{aligned} E_{x,y} = \frac{1}{4}\sum_c (&|\mathcal{I}_{x,y,c} - \mathcal{I}_{x-1,y,c}| + |\mathcal{I}_{x,y,c} - \mathcal{I}_{x+1,y,c}| \\ &+|\mathcal{I}_{x,y,c} - \mathcal{I}_{x,y-1,c}| + |\mathcal{I}_{x,y,c} - \mathcal{I}_{x,y+1,c}|) \end{aligned} \tag{2}$$

where x, y are the pixel coordinates and c refers to the color channels.

To jointly train $\mathcal{N}_{base}$ and $\mathcal{N}_{weight}$, we simply use pixel-wise L2 loss in the RGB color space as [18] by default:

$$\mathcal{L} = \|\mathcal{N}_{base}(\mathcal{N}_{weight}(\vec{\gamma}), \mathcal{I}, E) - f(\vec{\gamma}, \mathcal{I})\|^2 \tag{3}$$

3.2 Network Structure

As shown in Fig. 1, our *base* network $\mathcal{N}_{base}$ follows a similar network structure as [1]. We employ 20 convolutional layers with the same 3×3 kernel size, among which the intermediate 14 layers are formed as residual blocks. Except the last convolution layer, all the former convolutional layers are followed by an instance normalization [27] layer and a ReLU layer. To enlarge the receptive field of $\mathcal{N}_{base}$, the third convolution layer downsamples the dimension of feature maps by 1/2 using stride 2, and the third-to-last deconvolution layer (kernel size of 4×4) upsamples the downsampled feature maps to the original resolution symmetrically. In this way, the receptive field is effectively enlarged without losing too much image detail, and meanwhile the computation cost of intermediate layers is reduced. To further increase the receptive field, we also adopt the dilated convolution [28] as [18], more detailed network structure can be found in the supplementary material.

In this paper, the *weight learning* network $\mathcal{N}_{weight}$ simply consists of 20 fully connected (fc) layers by default. The i_{th} fc layer is responsible to learn the weights W_i for the i_{th} convolutional layer, which can be written as following:

$$W_i = A_i\vec{\gamma} + B_i, \qquad \forall i \in \{1, 2, ..., 20\} \tag{4}$$

Where A_i, B_i are the weight and bias of the i_{th} fc layer. Assuming the parameter $\vec{\gamma}$ has a dimension of m and W_i has a dimension of n_{wi}. The dimension of A_i and B_i would be $n_{wi} \times m$ and n_{wi} respectively.

Note in this paper, we don't intend to design an optimal network structure neither for the *base* network $\mathcal{N}_{base}$ nor the *weight learning* network $\mathcal{N}_{weight}$. On the contrary, we care more about whether it is feasible to learn the relationship between the weights of $\mathcal{N}_{base}$ and different parameter configurations $\vec{\gamma}$ even by such a simple *weight learning* network $\mathcal{N}_{weight}$.

3.3 Adaption to Cheap Parameter-Tuning

In the above default setting, the weights of all the convolution layers in $\mathcal{N}_{base}$ are learned by the *weight learning* network $\mathcal{N}_{weight}$ and would be dynamically changed with the input parameters. That is to say, if users want to do some parameter tuning for different visual effects, the whole network needs to be re-evaluated. However in order to obtain the best image quality, most current methods like [1], [17], [18], including our default design, often requires to run through a complex deep networks and is very expensive for computational cost.

To tackle the efficiency issue, we further extend our framework from learning all the convolution weights to learning the weights of only one single layer. Since all the following layers behind this adjustable layer need to be re-evaluated when fed with varying input features, it is better to put this layer as deeper as possible in the *base* network. In this way, the computation of all the preceding layers can be shared by different parameters, and only the layers after this single adjustable layer need to be re-evaluated. This is of great practicability for many real scenarios.

Specifically, that is to say, only the weights W_i of the ith layer is the function of input parameter $\vec{\gamma}$ as Equation (4) while all the remaining weights $W_k (k \neq i)$ are shared. During the runtime for parameter tuning, we only need to re-run the layers $k (k \geqslant i)$. Note that in our default network structure, W_i can be either the weights

of the convolutional layer or the scale and shift paramters in the following normalization layer. In this paper, by default, we choose the last instance normalization layer as the target layer and learn its scale and shift parameters, after which only one convolution layer needs to be run.

Though it seems more difficult for the network to adapt its behavior just with this single deep layer, we demonstrate its generality and effectiveness for different operators in the experiment section. Such a network design is able to outperform the state-of-the-art competitors [4], [5] by a large margin. More comprehensive ablation study about the learned layer type (convolution or instance normalization) and position k is conducted in the analysis section.

4 Experiments on the Proposed Framework

To demonstrate the ability of our proposed framework in incorporating parameterised image operators while maintaining their accuracy, we evaluate the proposed decoupled learning framework with different training configurations as shown from subsection 4.3 to 4.4. We leverage two representative types of image processing tasks: image filtering and image restoration. Within each of them, more than four popular operators are selected for detailed experiments. Below, we briefly introduce all the operators and their implementation details as follows.

4.1 Choice of Image Operators

Image Filtering: here we employ six popular image filters, denoted as L_0 [6], WLS [29], RTV [8], RGF [9], WMF [30] and LLF [31], which have been developed specifically for many different applications, such as edge-preserving image smoothing, texture removal, detail exaggeration, image abstraction, and image enhancement.

- L_0 **smooth** [6] - sharpening major image structures while eliminating a manageable degree of details by minimizing L_0 image gradients.
- **RTV** [8] - extracting structures from textures by optimizing the new inherent variation and relative total variation measures.
- **WLS** [29] - constructing edge-preserving multi-scale image decompositions for progressive coarsening of images.
- **RGF** [9] - removing textures by controlling detail smoothing under a scale measure.
- **WMF** [30] - an efficient 100+ times faster weighted median filter.
- **LLF** [31] - fast local Laplacian filters for tone mapping.

Image Restoration: The goal of image restoration is to recover a clear image from a corrupted image. In this paper we deal with four representative tasks in this venue: super resolution [32], denoising [33], deblocking [34] and derain [35], which have been studied with deep learning based approaches extensively. Except for the derain task, all the others are tested on various parameter settings that indicate the corruption level of the input image.

- **super resolution** - increasing the resolution or enhancing the lost details from a low-resolution blurry image, which is controlled by a downsampling scale with bicubic interpolation.
- **denoising** - restore the clear image from a noisy image, which is composed of Gaussian white noise controlled by the Gaussian standard deviation.
- **deblocking** - recover the image details from a compressed JPEG image differentiated by the image quality factor.
- **derain** - removing rain streaks from a captured rainy image.

4.2 Implementation Details

Dataset: We take use of the 17k natural images in the PASCAL VOC dataset as the clear images to synthesize the ground truth training samples. The PASCAL VOC images are picked from Flicker, and consists of a wide range of viewing conditions. To evaluate our performance, 100 images from the dataset are randomly picked as the test data for the image filtering task. While for the restoration tasks, we take the well-known benchmark for each specific task for testing, which is specifically BSD100 (super resolution), BSD68 (denoise), LIVE1 (deblock), RAIN12 (derain). For the filtering task, we filter the natural images with the aforementioned algorithms to produce ground truth labels. As for the image restoration tasks, the clear natural image is taken as the target image while the synthesized corrupted image is used as input.

Parameter Sampling: To make our network able to handle continuous parameters, we generate training image pairs with a much broader scope of parameter values rather than a single one. We uniformly sample parameters in either the logarithm or the linear space depending on the specific application. If the upper bound of the parameter range is tens or even hundreds of times larger than the lower bound, the parameters are sampled in the logarithm space to balance their magnitudes, otherwise they are sampled in the linear space.

4.3 Results on the Single Parameterized Operator

Image Filtering. We first evaluate the performance of six image operators with various controllable parameters individually. We train one network for each parameter value (λ) in one operator, and also train a network jointly on continuous random values sampled from the operator's parameter range, which can be inferred from the λ column in Table 1. The performance of the two networks is evaluated on the test dataset with PSNR and SSIM error metrics. Since our goal is to measure the performance difference between these two strategies, we directly compute the absolute difference of their errors and demonstrate the results in Table 1.

As can be seen from the table, though our proposed framework trained with numerous parameter settings lags a little behind the one trained on a single parameter value, their difference is very small especially for the visually more important error metric SSIM. For each image operator, previous methods usually requires to train separate networks for each parameter value, while our proposed approach only trains one single network jointly. Moreover, these image operators are dedicated to different image processing applications, the proposed framework is still able to learn all of them well, which verifies the versatility and robustness of our strategy.

Some visual results of our proposed framework are shown in Figure 2. As can be seen, our single network trained on continuous random parameter values is capable of predicting high-quality smooth images of various strengths.

Image Restoration. We then evaluate the proposed framework on three popular image restoration tasks as shown in Table 2, which perform essentially different from image filtering. Our model learns to recover from different corrupted input images, instead

TABLE 1
Quantitative absolute difference between the network trained with a *single* parameter value and *numerous* random values for each image operator.

	L_0				WLS				RTV				RGF				WMF				LLF			
metric	λ	single	nume.	diff.	λ	single	nume.	diff.	λ	single	nume.	diff.	λ	single	nume.	diff.	λ	single	nume.	diff.	λ	single	nume.	diff.
PSNR	0.002	40.69	39.46	1.23	0.100	44.00	42.12	1.88	0.002	41.11	40.66	0.45	1.00	41.77	37.03	4.74	1.00	39.06	36.79	2.27	2	38.00	37.83	0.17
	0.004	38.96	38.72	0.24	0.215	43.14	42.64	0.50	0.004	40.91	41.10	0.19	3.25	38.36	38.27	0.09	3.25	39.78	38.76	1.02	3	34.64	35.71	1.07
	0.020	36.07	35.71	0.36	1.000	41.93	41.63	0.30	0.010	40.50	41.07	0.57	5.50	38.11	38.35	0.24	5.50	39.94	38.53	1.41	5	32.34	32.29	0.05
	0.093	33.08	31.92	1.16	4.641	39.42	39.64	0.22	0.022	41.07	40.77	0.30	7.75	37.65	37.99	0.34	7.75	40.06	39.20	0.86	7	30.11	29.91	0.20
	0.200	31.75	30.43	1.32	10.00	39.13	38.51	0.62	0.050	40.73	39.18	1.55	10.0	37.52	37.08	0.44	10.0	39.49	38.72	0.77	8	29.53	28.95	0.58
	ave.	36.11	35.25	**0.86**	ave.	41.52	40.91	**0.61**	ave.	40.86	40.55	**0.31**	ave.	38.68	37.74	**0.93**	ave.	39.66	38.40	**1.26**	ave.	32.93	32.94	**0.01**
SSIM	0.002	0.989	0.988	0.001	0.100	0.994	0.993	0.001	0.002	0.987	0.988	0.001	1.00	0.994	0.981	0.013	1.00	0.985	0.972	0.013	2	0.992	0.992	0
	0.004	0.986	0.987	0.001	0.215	0.993	0.993	0	0.004	0.989	0.990	0.001	3.25	0.986	0.986	0	3.25	0.985	0.979	0.006	3	0.988	0.990	0.02
	0.020	0.982	0.981	0.001	1.000	0.992	0.991	0.001	0.010	0.990	0.991	0.001	5.50	0.985	0.986	0.001	5.50	0.986	0.981	0.005	5	0.983	0.984	0.01
	0.093	0.977	0.973	0.004	4.641	0.987	0.989	0.002	0.022	0.992	0.992	0	7.75	0.984	0.985	0.001	7.75	0.986	0.985	0.001	7	0.977	0.977	0
	0.200	0.973	0.968	0.005	10.00	0.986	0.987	0.001	0.050	0.992	0.990	0.002	10.0	0.984	0.982	0.002	10.0	0.986	0.984	0.002	8	0.976	0.974	0.02
	ave.	0.981	0.979	**0.002**	ave.	0.990	0.990	**0**	ave.	0.990	0.990	**0**	ave.	0.986	0.984	**0.002**	ave.	0.985	0.980	**0.005**	ave.	0.983	0.983	**0**

TABLE 2
Quantitative absolute difference in PSNR (above) and SSIM (bottom) between the network trained on a *single* parameter value and *numerous* random values on the three image restoration tasks. Their parameters specifically mean downsampling scale (s), Gaussian standard deviation (σ) and JPEG quality (q).

Super Resolution				Denoising				Deblock			
s	single	nume.	diff.	σ	single	nume.	diff.	q	single	nume.	diff.
2	31.78	31.62	0.16	15	31.17	31.07	0.10	10	29.26	29.17	0.09
3	28.78	28.76	0.02	25	28.94	28.98	0.04	20	31.49	31.43	0.06
4	27.31	27.31	0	50	26.22	26.14	0.08				
ave.	29.29	29.23	**0.06**	ave.	28.77	28.73	**0.04**	ave.	30.37	30.30	**0.07**
2	0.894	0.892	0.002	15	0.881	0.883	0.002	10	0.817	0.817	0
3	0.798	0.796	0.002	25	0.821	0.822	0.001	20	0.881	0.882	0.001
4	0.728	0.726	0.002	50	0.722	0.718	0.004				
ave.	0.806	0.804	**0.002**	ave.	0.808	0.807	**0.001**	ave.	0.849	0.849	**0**

of learning to obtain different visual effects given the same input image as in the image filtering task.

As shown in Table 2, our results trained jointly on continuous random parameter values also show no big difference from the one trained solely on an individual parameter value, which further validate our algorithm in a broader image processing literature. Some corresponding visual results of our proposed framework are shown in Figure 3.

4.4 Results on Jointly Training Multiple Image Operators

Intuitively, another challenging case for our proposed framework is to incorporate multiple distinct image operators into a single learned neural network, which is much harder to be trained due to their different implementation details and purposes. To explore the potential of our proposed neural network, we experiment by jointly training over (*i*). 6 filtering based operators, (*ii*). 4 image restoration operators or (*iii*). all the 10 different operators altogether. To generate training images of each image operator, we sample random parameter values continuously within its parameter range. Since there is no parameter tuning for the derain task, we leverage its default parameter setting for training.

The input to the *weight learning* network now takes two parameters, one indicates the specific image operator while the other is the random parameter values assigned to the specified filter. These 10 image operators are denoted simply by 10 discrete values that range from 0.1 to 1.0 in the input parameter vector. Since the absolute parameter range may differ a lot from operator to operator, for example, [2,4] for super resolution and [0.002,0.2] for L_0 filter, we rescale the parameters in all the operators into the same numerical range to enable consistent back-propagated gradient magnitude.

As shown in Table 3, training on each individual image operator achieves the highest numerical score (#ope.=1), which is averaged over multiple different parameter settings just like in previous tables. While jointly training over either 6 image filters or 4 restoration tasks (#ope.=6/4), even for the case where all 10 image operators are jointly trained (#ope.=10), their average performance degrades but still achieves close results to the best score. It means with the same network structure, our framework is able to incorporate all these different image operators together into a single network without losing much accuracy.

4.5 Comparison with Baseline [18]

We compare our proposed framework with one naive approach that employs only the *base* network with additional input channels as [18], which indicates the parameter values and image operators separately. Each additional channel is occupied with a single value.

The results are shown in Table 3, which trains ten different image operators including both image filtering and restoration tasks together. We can see that the baseline from [18] lags behind us across all ten image operators. The potential reason for this phenomenon could be that it is more difficult to learn unified convolution weights to be suitable for tasks different in both goals and implementations. By contrast, the convolutional weights of our *base* network are adaptively changed for different tasks and input parameters. Theoretically speaking, our learned network should have the capability to express the base network with more numerous possibilities.

4.6 Generalization to Higher Dimensional Parameter Space

Except for experimenting on a single input parameter, we also demonstrate the experiments on inputting multiple types of pa-

TABLE 3
Numerical results of our proposed framework jointly trained over different number of image operators (#operators). "6/4" refers to the results jointly trained over either the front 6 filtering based approaches or the last 4 restoration tasks. "10" is the results of jointly training all 10 tasks. Our approach also achieves superior performance over one baseline [18] on all the 10 image operators.

metric	method	#ope.	L_0	WLS	RTV	RGF	WMF	LLF	SR	denoise	deblock	derain	average
PSNR	Ours	1	35.25	40.91	40.55	37.74	38.40	32.94	29.23	28.73	30.30	29.86	**34.40**
		6/4	33.27	37.39	37.00	35.41	36.06	30.08	28.89	28.67	30.10	30.32	**32.72**
		10	32.67	36.59	36.03	34.64	35.08	29.77	28.53	28.36	29.69	30.45	**32.18**
	[18]	10	31.01	34.85	34.20	33.10	33.61	28.58	28.21	28.05	29.48	29.12	**31.02**
SSIM	Ours	1	0.979	0.991	0.990	0.984	0.980	0.984	0.804	0.807	0.849	0.893	**0.926**
		6/4	0.969	0.980	0.979	0.974	0.967	0.976	0.797	0.800	0.842	0.893	**0.918**
		10	0.965	0.978	0.975	0.969	0.962	0.971	0.789	0.789	0.837	0.895	**0.913**
	[18]	10	0.949	0.969	0.964	0.953	0.947	0.961	0.779	0.777	0.834	0.863	**0.899**

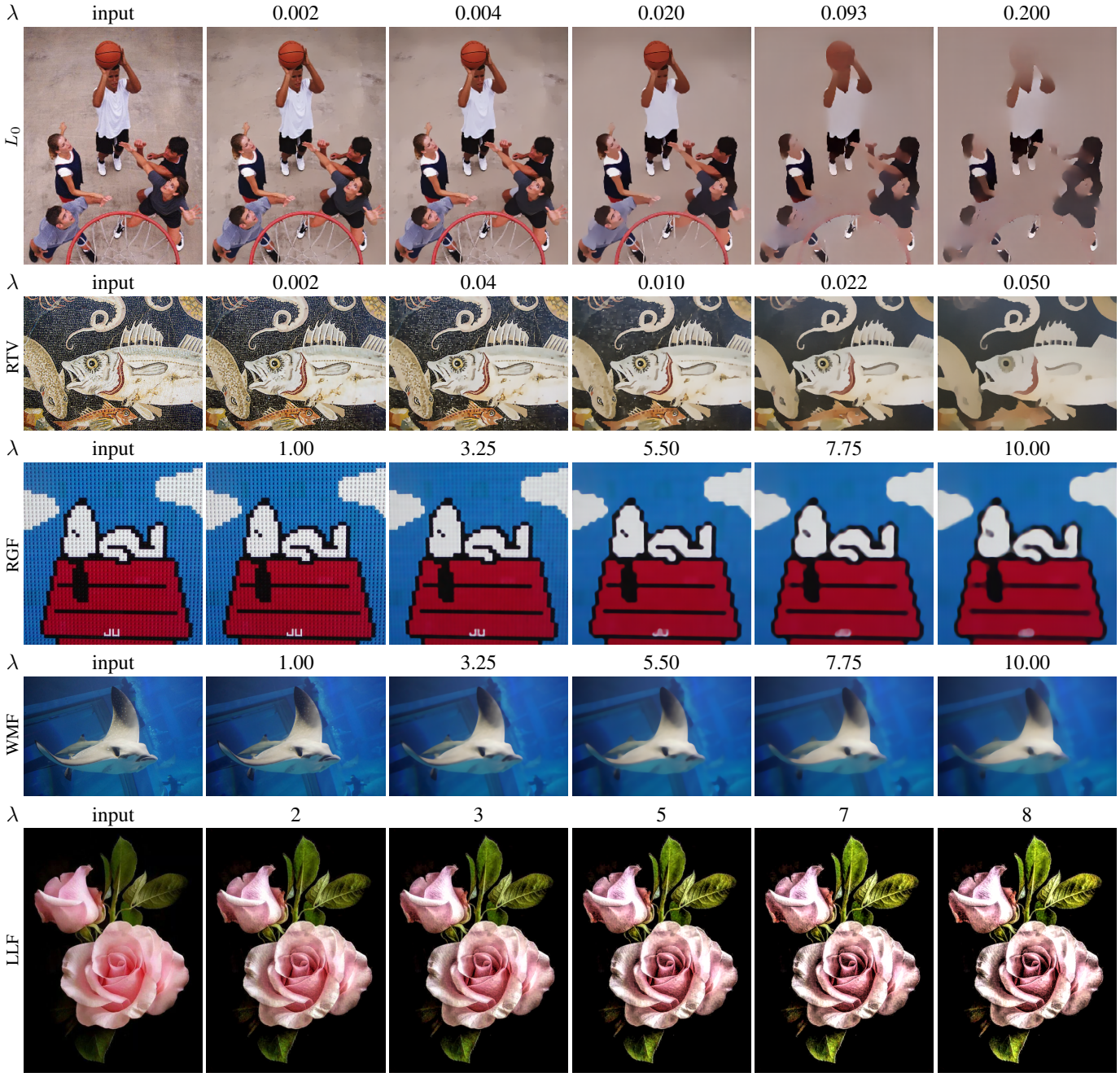


Fig. 2. Visual examples produced by our framework trained on continuous parameter settings of six image filters independently. Note all the visual effects for one filter are generated by a single network.

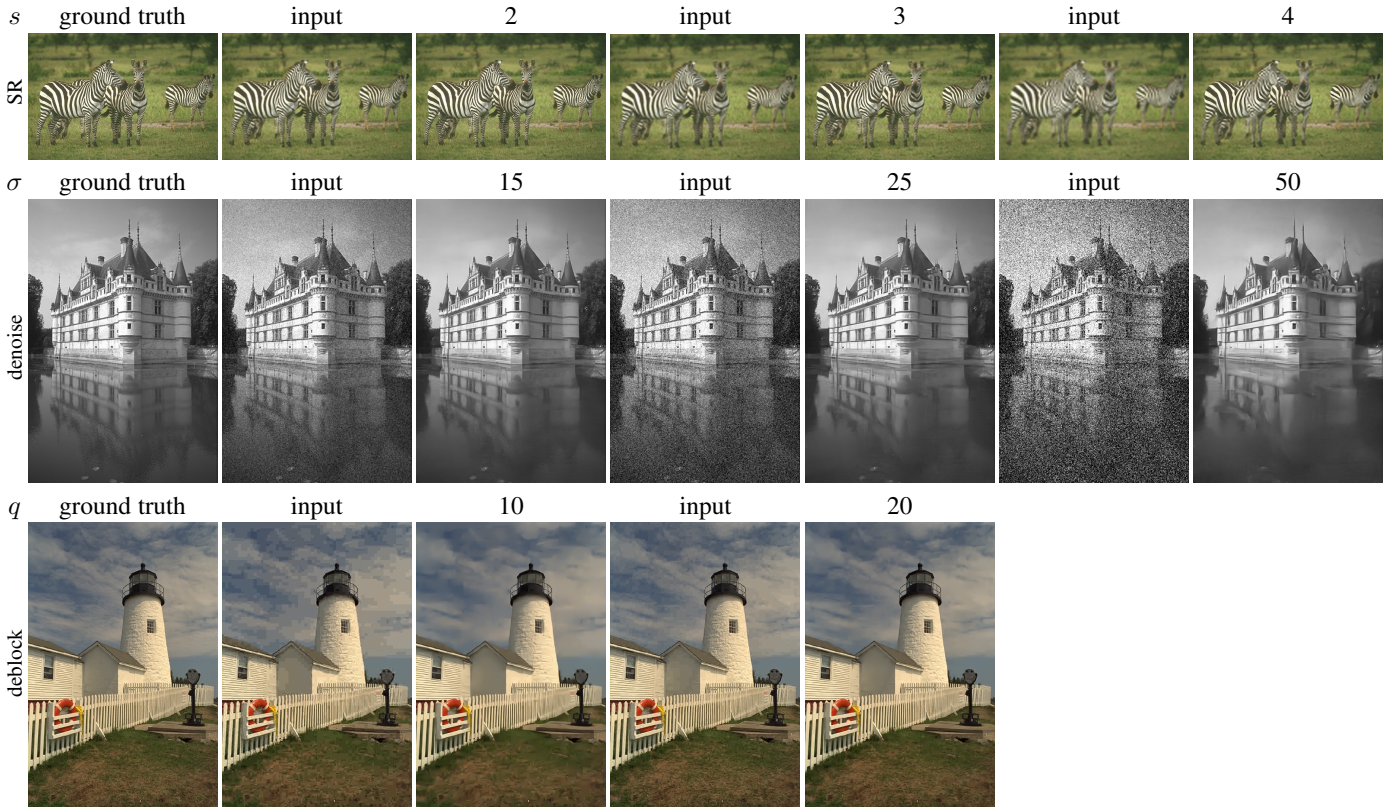


Fig. 3. Visual examples produced by our framework trained on continuous parameter settings of three image restoration tasks independently. Note all the clean images for one restoration task are generated by a single network.

TABLE 4
Quantitative evaluation (PSNR) on the higher dimensional parameter space of the RTV filter. "#dim." indicates the number of leveraged parameter dimensions, and "#ope." is the number of jointly trained image operators.

		λ (RTV)					
#ope.	#dim.	0.002	0.004	0.010	0.022	0.050	average
1	1	40.66	41.10	41.07	40.77	39.18	**40.55**
1	2	40.70	40.89	40.78	40.56	39.23	**40.43**
1	3	40.19	40.82	40.95	40.93	39.78	**40.53**
1	4	40.18	40.83	40.95	40.87	39.68	**40.50**
10	2	37.44	37.16	36.50	35.51	33.52	**36.03**

rameters that belongs to either one image operator or various operators.

In this section, we evaluate the performance on the famous texture removal tool RTV [8]. RTV has multiple parameters to adjust its visual effects, such as λ which balances between the data prior term and smoothness term in its energy function, σ which controls the spatial scale for computing the windowed variation and is even more effective in removing textures, ϵ_s that controls the sharpness of the final results and $maxIter$ that refers to the number of iterations in the optimization.

We train our network on the continuous parameter range of λ first, therefore the input parameter is a one-element vector and constrained in the one dimensional space. Its results are shown in Table 4 with #ope. and #dim. both equal to 1. Then we add one more parameter at a time to increase the input parameter dimension until all the aforementioned four parameters are included to form a four-element input vector, where #dim. equals to 4. Moreover, following the experimental setting in Table 3, we also train our network on 10 different image operators including the balance weight λ for RTV to form a two dimensional parameter space. We evaluate the performance on 5 discrete parameter values of λ that is commonly adopted for all the parameter dimensions.

We can see that when focusing on a single image operator (RTV) with up to even 4 parameters, the resultant performance of different parameter values are similar to each other in a reasonable range. But when we incorporate more image operators into joint training, the performance on RTV filter decreases significantly. A reasonable explanation for this phenomenon is that as the visual effects introduced by the additional input parameter are more different from the existing one, it becomes harder to incorporate the additional input parameter for joint training. Regarding the experiment in Table 4, the additional parameter for more image operators introduces more different visual effects compared to the four inherent parameters in the RTV filter, and hence becomes harder for joint training.

Note while incorporating more parameter dimensions into the training process, we don't modify the amount of training samples and iterations. Each training sample is generated by randomly sampling the leveraged parameters.

5 Experiments on the Cheap Parameter-Tuning Version

In this section, we evaluate the cheap parameter-tuning adaption of our proposed framework with many recent filtering and restoration approaches. Most compared approaches design some complex and advanced deep neural network to achieve the optimal performance for their specific dedicated tasks. In order to balance between the performance and computation costs, we slightly modify the

structure of $\mathcal{N}_{base}$ by utilizing the depth-wise convolution [36] and increasing the intermediate feature channels (64 to 128) in this section. We validate and study the effectiveness of the proposed framework, which not only runs efficiently for parameter tuning but also achieves state-of-the-art performance compared to many recent deep learning approaches.

5.1 Image Filtering

We compare our proposed framework with two recent state-of-the-art approaches [4], [5], which share the closest spirit as ours in reproducing the image filters as well as possible while running in real-time by varying between different image operators.

To conduct on a broader scope of image operators like the other two approaches, besides the aforementioned six image filters, we further add four operators with more different visual effects for a fair comparison as follows.

- **LLF remapping** [31] - fast local Laplacian filters for tone mapping by leveraging a remapping function.
- **WLS enhancement** [29] - detail exaggeration implemented via multi-scale image decompositions.
- **Stylization** [31] - transfer the look of one photographers masterpiece onto another photo via fast local Laplacian filters.
- **Abstraction** [46] - pencil drawing by combining the tone and stroke structures, which complement each other in generating visually constrained results.

Note the six image filters mentioned in Section 4 contains controllable parameters that indicate different visual effects, while each of the above four operators represents a unique fixed visual effect. All these ten image operators are fed through the networks for joint training. The input parameters are sampled similarly as in section 4.4.

The image operators deployed in [4], [5] are relatively diverse, which include both photographic effects [31] and image structure manipulations [26] that are also demonstrated in our paper. Even if our deployed image operators are not exactly the same as theirs [4], [5], we believe the idea proposed in their paper is general enough, and we apply them to the same operators in our paper (some included and some not included in their papers).

In Table 6, we demonstrate the numerical results of our proposed framework compared with the competitors BGU [4] and DBL [5]. For a fair comparison, the evaluation is conducted on one parameter setting (the default one) for each image operator following the other two approaches. As can be seen from the numerical errors (PSNR and SSIM), our results achieve significantly better results. The average PSNR over 10 image operators are about 5dB larger than the second best competitor.

Note both BGU and DBL learn to reproduce one specific operator with the default parameter setting, while *our approach learns all the 10 image operators with their full range of parameter values jointly within one single network*, which is much more difficult and challenging.

5.2 Image Restoration

Furthermore we compare our framework on the four aforementioned image restoration tasks: super resolution, denoising, deblocking and derain with many recent restoration methods as shown from Table 7 to Table 10. Except for the derain task, all the others are tested on various parameter settings that indicate the corruption level of the input image. Though these existing algorithms are dedicated on each specific field with specifically designed algorithms or network structures, our proposed framework is still able to achieve very competitive results on all these tasks. Need to note that all these tasks are jointly trained with one single network by our approach.

Given each new example for the restoration tasks, it needs to rerun the whole network from the beginning. Therefore it's not a very valid case to justify our cheap parameter-tuning framework. However, theoretically this experiment still demonstrates the extraordinary ability of deep networks to incorporate different restoration capabilities within a very limited parameterized operation (the last instance normalization layer).

TABLE 5
Running time evaluation (milliseconds) of our decouple learning framework and its cheap parameter tuning module on different image resolutions, along with the baseline [18]. It's evaluated on GPU devices by default without specifications.

Resolution	[18]	Ours	ParaTuning	ParaTuning (CPU)
VGA (640×480)	4.87	6.20	0.59	0.65
720p (1280×720)	5.02	6.45	0.72	0.85
1080p (1920×1080)	5.88	6.90	0.79	0.98

5.3 Running Time Comparison

In this section, we evaluate the running time of our proposed framework and the cheap parameter tuning module on different popular image resolutions. As shown in Table 5, given a new input, we need to run through the whole model that takes less than 10 milliseconds for a 1080p image, but while switching between different image operations it takes only less than 1 millisecond on either GPU or CPU device, which is almost ten times faster than running the full model. Our implementation leverages the multi-core CPU package (NNPACK) in the MXNet framework for acceleration, and our model runs on a 20-core CPU device.

We also evaluate the running time of the baseline [18]. Note their approach and ours share most of the *base* network structure, while their parameter tuning is implemented by adding additional input channels to the *base* network, which doesn't increase the computation cost much. On the other hand, our framework includes one more *weight learning* network and hence lags behind them on the running time. From this table, we can also see that the *weight learning* network takes around 1.02 milliseconds to process a 1080p image, which is very computationally efficient.

6 Understanding and Analysis

To better understand the *base* network $\mathcal{N}_{base}$ and the *weight learning* network $\mathcal{N}_{weight}$, we analyze some meaningful behavior behind our framework in this section.

6.1 Complexity of the Model Size

The saved weights of our proposed decouple learning framework come from two parts, the *weight learning* network and *base* network. Let's discuss our basic implementation where all the convolution weights are adaptively learned by the *weight learning* network with n input parameters. The *base* network contains 2,432 instance normalization weights and 696,256 convolution weights to be learned. While the convolution weights are adaptively predicted by the *weight learning* network, and hence are not required

TABLE 6
Quantitative comparison with state-of-the-art approaches in reproducing image operators.

metric	method	L_0	WLS	RTV	RGF	WMF	LLF	LLF remap	WLS enhance	Stylization	Abstraction	Average
PSNR	BGU	31.76	27.03	26.15	22.71	21.27	26.97	33.05	26.93	14.31	16.11	**24.62**
	DBL	28.67	30.63	28.52	27.11	26.88	25.13	29.34	28.29	25.08	19.61	**26.92**
	Ours	31.81	36.59	34.28	33.29	34.00	29.74	32.41	34.40	26.66	25.61	**31.87**
SSIM	BGU	0.912	0.915	0.848	0.776	0.765	0.936	0.978	0.931	0.673	0.427	**0.816**
	DBL	0.852	0.890	0.826	0.805	0.786	0.899	0.945	0.944	0.887	0.502	**0.833**
	Ours	0.946	0.971	0.948	0.945	0.940	0.967	0.969	0.986	0.927	0.835	**0.943**

TABLE 7
Quantitative results (PSNR/SSIM) of the JPEG deblocking task on the LIVE1 benchmark.

Quality	JPEG	ARCNN [34]	TNRD [37]	DnCNN [38]	MemNet [39]	Ours
10	27.77/0.7730	28.96/0.8076	29.15/0.8111	29.19/0.8123	29.45/0.8193	29.31/0.8170
20	30.07/0.8512	31.29/0.8733	31.46/0.8769	31.59/0.8802	31.83/0.8846	31.60/0.8816

TABLE 8
Quantitative results (PSNR/SSIM) of the image super resolution task on the BSD100 benchmark.

Scale	Bicubic	SRCNN [15]	VDSR [40]	DRCN [41]	DnCNN [38]	MemNet [39]	Ours
2	29.56/0.8431	31.36/0.8879	31.90/0.8960	31.85/0.8942	31.90/0.8961	32.08/0.8978	31.67/0.8934
3	27.21/0.7385	28.41/0.7863	28.82/0.7976	28.80/0.7963	28.85/0.7981	28.96/0.8001	28.76/0.7969
4	25.96/0.6675	26.90/0.7101	27.29/0.7251	27.23/0.7233	27.29/0.7253	27.40/0.7281	27.29/0.7257

TABLE 9
Quantitative results (PSNR) of the image denoising task on the BSD68 benchmark.

σ	BM3D [42]	WNNM [43]	TNRD [37]	DCDP [44]	Ours
15	31.07	31.37	31.42	31.63	31.48
25	28.57	28.83	28.92	29.15	29.11
50	25.62	25.87	25.97	26.19	26.22

TABLE 10
Quantitative results (PSNR) of the derain task on the RAIN12 benchmark.

DerainNet [45]	DnCNN [38]	Ours
28.94	30.90	30.08

to be saved. For each convolution layer in the *base* network, the *weight learning* network contains a single fully connected layer to predict its weights. Hence the learned weights for the fc layer are totally 2,088,768 (696,256×3), where n equals to 2 for the case of jointly training multiple image operators. Then the total weights required to save our decouple learning framework are 2,091,200 (2,088,768+2,432).

As analyzed above, most saved weights come from the *weight learning* network, which is equivalent to n+1 times the convolution weights in the *base* network. Regardless of the detailed *base* network design, such a conclusion still holds. While jointly training multiple image operators, the total saved weights of our framework are almost 3 times the weights of the independent *base* network, however it implements theoretically numerous image operations defined by the continuous input parameter range. In our PyTorch implementation, the saved model for jointly training 10 image operators with two input parameters only occupies 5.33 MB, which is sufficient for embedded systems.

While adapted for cheap parameter tuning, only the weights in a single instance normalization layer are learned by the *weight learning* network. The total saved weights of our framework are even closer to the one of an independent *base* network, therefore it becomes more efficient to jointly train multiple image operators. As the *base* network is bigger in this case, it takes 7.99 MB to save the whole model.

6.2 The Effective Receptive Field

In neuroscience, the receptive field is the particular region of the sensory space in which a stimulus will modify the firing of one specific neuron. The large receptive field is also known to be important for modern convolutional networks. Different strategies are proposed to increase the receptive field, such as deeper network structure or dilated convolution. Though the theoretical receptive field of one network may be very large, the real effective receptive field may vary with different learning targets. So how is the effective receptive field of $\mathcal{N}_{base}$ changed with different parameters $\overrightarrow{\gamma}$ and $\mathcal{I}$? Here we use L_0 smoothing [26] as the default example operator.

In Fig. 4, we study the effective receptive field of a non-edge point, a moderate edge point, and a strong edge point with different smoothing parameters λ respectively. To obtain the effective receptive field for a specific spatial point p, we first feed the input image into the network to get the smoothing result, then propagate the gradients back to the input while masking out the gradient of all points except p. Only the points whose gradient value is large than $0.025 * grad_{max}$ ($grad_{max}$ is the maximum gradient value of input gradient) are considered within the receptive field and marked as green in Fig. 4. From Fig. 4, we observe three important phenomena: 1) For a non-edge point,

Fig. 4. Effective receptive field of L_0 smoothing for different spatial positions and parameter λ. The top to bottom indicate the effective receptive field of a non-edge point, a moderate edge point, and a strong edge point.

the larger the smoothing parameter λ is, the larger the effective field is, and most effective points fall within the object boundary. 2) For a moderate edge point, its receptive field stays small until a relatively large smoothing parameter is used. 3) For a strong edge point, the effective receptive field is always small for all the different smoothing parameters. It means, on one hand, the *weight learning* network $\mathcal{N}_{weight}$ can dynamically change the receptive field of $\mathcal{N}_{base}$ based on different smoothing parameters. On the other hand, the *base* network $\mathcal{N}_{base}$ itself can also adaptively change its receptive field for different spatial points.

6.3 Investigation into the Learned Weight

In this subsection, we discuss about the relationship between different learned convolutional kernels, and alternatives of the learned parameter position and type to better understand our proposed framework.

6.3.1 Difference of the learned convolution kernels between jointly trained network and solely trained network

To analyse the difference of the convolution weights between networks jointly trained on numerous random parameter values ("nume.") and a single parameter value ("fixed"), we compute their correlation coefficient, individual mean and variance for each layer as shown in Table 11.

As can be seen, the correlation coefficient is almost 0 everywhere, which means there is no linear relationship between the two groups of convolution kernels. The absolute mean and variance of jointly trained network is also significantly smaller than that of the solely trained network. Therefore, in each way, their learned convolution weights are very different from each other, even if the learned smoothing effect is almost the same (PSNR/SSIM: 35.51dB/0.983 (nume.) *vs.* 35.83dB/0.982 (fixed)).

This simple experiment further verifies the huge solution space in the form of learned convolution kernels. Two exactly same results may be represented by very different convolution weights. The linear transformation in our proposed weight learning network actually connects all the different image operators and constrains their learned convolution weights in a limited high dimensional space.

6.3.2 Study of the learned parameter position and type

For most image operators presented in the paper, there is only one parameter needed. It might be too complex to use a *weight learning* network that only takes a scalar to manipulate all the convolution layers in the *base* network. To conduct a comprehensive study of the effects of where and what the learned controllable parameters are, we start by three experimental settings, (i). learned instance normalization parameters at different network depth. (ii). learned different types of convolution parameters at a fixed network depth. (iii). learned parameters in either all the convolutional or instance normalization layers.

We compare the performance of different network settings in Table 12. These experiments are made by reproducing the 10 image filters with the base network described in Section 3.2.

(i). At first, we adjust the position of learned controllable parameters with four discrete network depth. It ranges from the first (1) instance normalization layer to the last (19) layer (the 20th convolutional layer is not followed by instance normalization). From the table, we can see that as the learned parameters become deeper in the network, the performance degrades. It's a reasonable phenomenon, since the deeper the parameters are, the smaller the network capacity to differentiate various image operators is, and thus it's harder to incorporate all the operators in the network.

We can also observe that the average error (PSNR) over all the 10 image operators only decreases by about 2.5dB from the first layer to the last layer, however it significantly improves the running speed when varying between different image operators in the network.

(ii). To explore the influence of other types of parameters in the network to separate these image operators, we replace the controllable parameters in the instance normalization layer with the ones in the convolutional layer.

Each convolutional layer contains a $M \times N \times k \times k$ weight tensor and a M-length bias vector. M is the number of output

TABLE 11
Comparison between the statistics of convolution kernels learned with random parameter values and a single parameter value. The numbers are generated based on the WLS filter while λ equals to 10.

layer index	1	2	3	4	5	6	7	8	9	10
correlation	0.005	-0.001	0.008	0.004	-0.005	0.007	0.008	0.010	-0.002	-0.012
mean (nume.)	0.219	0.066	0.004	-0.069	-0.227	-0.183	-0.214	-0.131	-0.152	-0.141
mean (single)	0.542	0.801	0.604	-0.571	-2.090	-0.685	-0.520	-1.638	-1.604	-1.326
variance (nume.)	15.514	12.166	15.511	18.231	17.676	20.408	16.578	19.420	16.447	18.292
variance (single)	542.28	373.70	490.53	419.55	482.71	559.00	532.14	505.67	471.50	437.69
layer index	11	12	13	14	15	16	17	18	19	20
correlation	0.011	0.184	-0.014	0.001	0.012	0.024	0.011	-0.001	-0.071	-0.017
mean (nume.)	-0.296	-0.168	-0.084	-0.310	-0.137	-0.201	0.003	-0.239	-0.407	0.634
mean (single)	-1.117	-1.134	-2.417	-1.437	-1.043	-1.977	-0.596	1.898	-3.203	3.358
variance (nume.)	17.704	20.141	18.767	21.080	28.461	23.795	19.475	14.592	13.703	4.220
variance (single)	549.51	500.61	585.75	507.40	792.44	588.06	581.66	501.08	487.62	140.33

TABLE 12
Numerical analysis of learned parameter type and position. "norm" and "conv" indicates the learned parameter type in either the instance normalization layers or convolution layers. The number in the brackets behind the parameter type refer to the position where the learned layer is located, "1" for the 1st layer and "all" for all the layers.

metric	method	L_0	WLS	RTV	RGF	WMF	LLF	LLF remap	WLS enhance	Stylization	Abstraction	Average
PSNR	norm(1)	31.06	34.86	33.75	32.87	33.49	29.22	32.21	32.59	27.26	24.81	**31.21**
	norm(7)	30.86	34.75	33.54	32.97	33.39	29.24	31.99	32.79	27.07	24.81	**31.14**
	norm(14)	30.64	34.31	33.06	32.42	33.21	29.15	31.80	32.53	27.05	24.83	**30.90**
	norm(19)	28.53	31.22	29.46	29.72	30.69	27.03	30.36	29.78	25.57	24.38	**28.67**
	conv(19)	30.57	34.17	32.45	31.84	32.96	28.80	32.07	31.95	27.46	26.30	**30.85**
	conv channel1(19)	29.54	32.80	31.14	30.94	31.97	28.09	31.53	31.39	27.07	25.89	**30.03**
	conv channel2(19)	24.87	26.63	24.86	26.17	27.48	23.26	24.86	25.29	21.82	24.45	**24.96**
	norm(all)	31.71	35.51	34.31	33.10	33.96	29.56	32.45	33.11	27.32	25.21	**31.62**
	conv(all)	31.64	35.02	33.67	32.81	33.97	29.59	32.52	32.69	28.12	26.35	**31.63**
SSIM	norm(1)	0.949	0.971	0.959	0.954	0.948	0.966	0.981	0.982	0.923	0.819	**0.945**
	norm(7)	0.945	0.969	0.958	0.955	0.948	0.967	0.980	0.982	0.917	0.812	**0.943**
	norm(14)	0.935	0.963	0.942	0.947	0.945	0.966	0.979	0.980	0.917	0.816	**0.939**
	norm(19)	0.871	0.890	0.809	0.881	0.876	0.947	0.970	0.965	0.889	0.796	**0.889**
	conv(19)	0.931	0.948	0.913	0.932	0.932	0.965	0.979	0.989	0.911	0.826	**0.932**
	conv channel1(19)	0.905	0.927	0.862	0.917	0.921	0.957	0.976	0.974	0.906	0.822	**0.916**
	conv channel2(19)	0.741	0.727	0.629	0.700	0.737	0.844	0.899	0.914	0.848	0.741	**0.778**
	norm(all)	0.954	0.972	0.960	0.956	0.953	0.970	0.982	0.984	0.924	0.823	**0.947**
	conv(all)	0.954	0.969	0.952	0.953	0.954	0.969	0.982	0.982	0.925	0.833	**0.947**

feature channels, N is the number of input feature channels, k is the kernel size. M equals to N for the intermediate convolutional layers in our baseline network. Since the convolutional layer is followed by an instance normalization layer, the added bias in the convolutional layer will be balanced out by the following normalization operation and becomes useless. Therefore we experiment only by varying the parameters in the convolutional weights.

We study with three types of learned convolutional parameters, which are separately: "conv", all the weight kernels ($M\times N\times k\times k$); "conv channel1", one slice of weight kernels ($M\times 1\times k\times k$); "conv channel2", another slice of weight kernels ($1\times N\times k\times k$). The learned convolutional layer shares the same network depth as the last instance normalization layer.

As can be seen from the table, learning all the weight kernels in the convolutional layer (conv) achieves the highest performance, which is very reasonable. While learning all the convolutional kernels that are only related to one single input channel (conv channel1), the performance degrades by about 0.8dB. Learning all the parameters that are related to one single output channel (conv channel2) obtains the worst performance. Note even if "conv channel1" and "conv channel2" share the same number of controllable parameters, all the learned parameters in "conv channel2" only result in adjustment of a single channel of output feature map, which influence the following layers much less than "conv channel1" and hence achieve worse results.

Interestingly, by learning the convolutional parameters (conv and conv channel1) achieves better results than learning instance normalization parameters at the same network depth. This may be because the convolutional layer is ahead of the instance normalization, and the learned convolutional parameters (even $M\times k\times k$ for "conv channel1", k equals to 3 in our case) is much more than the normalization's ($M\times 2$).

(iii). To further explore the difference between learning convolution and normalization layers, we try to learn the parameters in either all the convolutional layers or all the instance normalization layers in the network, which should theoretically achieves the best performance of incorporating all the image operators.

As shown in the "conv(all)" and "norm(all)" rows, surprisingly they achieve almost the same performance on the either PSNR or SSIM error metric. This means it already touches the top

TABLE 13
Quantitative evaluation of a few variants of the proposed network trained on the WLS filter [29]. We experiment separately by training only the base network on a fixed single parameter value ("single (base)"), extending the weight learning network to two or more fully connected layers trained on numerous random parameter values with ("nume. (fc2R)") and without ("nume. (fc2)") ReLU layers inbetween.

λ	single		single (base)		nume.		nume. (fc2R)		nume. (fc2)	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
0.100	44.00	0.994	44.16	0.994	42.12	0.993	42.02	0.992	42.36	0.993
0.215	43.14	0.993	43.08	0.993	42.64	0.993	42.23	0.993	42.43	0.993
1.000	41.93	0.992	40.61	0.991	41.63	0.991	40.25	0.991	40.28	0.991
4.641	39.42	0.987	38.01	0.988	39.64	0.989	37.31	0.988	37.35	0.988
10.00	39.13	0.986	36.83	0.986	38.51	0.987	36.00	0.985	35.94	0.986
average	41.52	0.990	40.54	0.990	40.91	0.990	39.56	0.990	39.67	0.990

TABLE 14
Quantitative evaluation of our proposed framework trained only with a set of fixed parameter values ("various (fixed)") on the L_0 smoother [6]. The parameters used for training our framework are taken from the 5 non-boldface parameters between [0.002, 0.2] in the table. The extra 4 parameters with boldface are only used in the test stage. The absolute difference between the network trained on a single parameter ("single") and various fixed parameters ("various (fixed)") is displayed in the bottom.

		0.0020	**0.0025**	**0.0033**	0.0043	0.0200	0.0928	**0.1200**	**0.1600**	0.2000	average
single	PSNR	40.69	40.19	39.77	38.96	36.07	33.08	31.78	31.13	31.75	35.94
	SSIM	0.989	0.987	0.986	0.986	0.982	0.977	0.973	0.972	0.973	0.981
various (fixed)	PSNR	39.61	39.33	38.95	38.51	35.37	31.80	31.40	30.69	30.54	35.13
	SSIM	0.988	0.987	0.986	0.986	0.979	0.972	0.972	0.971	0.970	0.979
difference	PSNR	1.08	**0.86**	**0.82**	0.45	0.7	1.28	**0.38**	**0.44**	1.21	0.81
	SSIM	0.001	**0**	**0**	0	0.003	0.005	**0.001**	**0.001**	0.003	0.002

performance limit of incorporating the 10 image operators with this neural network.

Note learning all the normalization parameters (norm(all)) only obtains a little improvement (0.41dB) than learning the one in the first layer (norm(1)), from which we can see it's more important which position the controllable parameters are instead of how many layers the parameters are located in.

6.4 Interpretation of the Weight Learning Network

To help understand the connection between the *base* network $\mathcal{N}_{base}$ and the *weight learning* network $\mathcal{N}_{weight}$, we decompose the parameter vector $\overrightarrow{\gamma}$ and the weight matrix A_i into independent elements $\gamma_1, ..., \gamma_m$ and $A_{i0}, ..., A_{im}$ respectively, then:

$$(A_i \overrightarrow{\gamma} + B_i) \otimes x = \sum_{k=1}^{m} \gamma_k A_{ik} \otimes x + B_i \otimes x \qquad (5)$$

where $\otimes$ denotes convolution operation, and m is the dimension of $\overrightarrow{\gamma}$. In other words, the one convolution layer, whose weights are learned with one single fc layer, is exactly equivalent to a multi-path convolution block as shown in Figure 5. Learning the weight and bias of the single fc layer is equivalent to learning the common basic convolution kernels $B_i, A_{i1}, A_{i2}, ..., A_{im}$ in the convolution block.

6.5 Analysis of More Variants of Our Proposed Network

In this subsection, we experiment with a few variants of our network to justify its effectiveness.

6.5.1 Training the base network only

Since training a fully convolutional network alone has been employed frequently by previous image processing papers [17], [1], [18], which presents a strong baseline, we experiment with this alternative ("nume. (base) ") which only leverages the *base* network by training it on one specific parameter configuration. As show in Table 13, our proposed framework ("single") achieves even better performance than the baseline under the PSNR error metric.

6.5.2 Training with more fc layers

We also try a deeper *weight learning* network with more fully connected layers. Here we simply add one more fully connected layer to the default *weight learning* network, and demonstrate the performance of its two variants with ("nume. (fc2R)") and without ("nume. (fc2)") ReLU between these two layers respectively. As shown in Table 13, they all achieve comparable performance with that of single layer ("nume.") used in our paper. The potential reason for this phenomenon is that this one-layer *weight learning* network is sufficient for adaptively learning various parameter settings, while adding more weights/complexity to the network does not contribute to the performance much.

6.6 Interpolation Ability of the Proposed Framework on Unseen Input Parameters

Since the *weight learning* network contains a single fully connected layer with no non-linear activation layers, the predicted convolution weights should be a linear transformation of the input parameters. Given such a fact, any convolutional kernels of a specific parameter should be the linear interpolation of the other two. Hence we are curious about the interpolation ability of the proposed framework.

To verify such a property, we train the network using only a few fixed parameter values, which corresponds to the 5 non-boldface parameter values in Table 14. But in the test stage, we

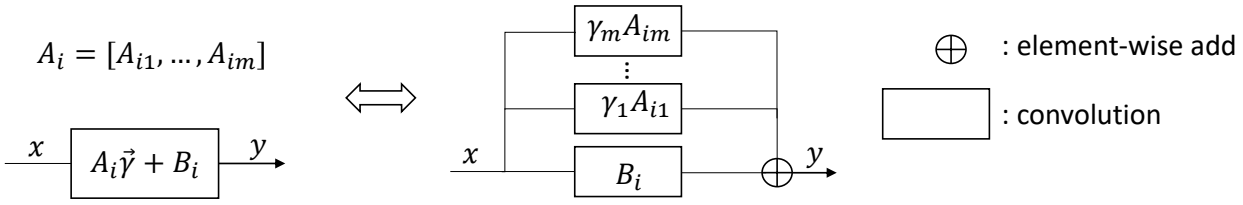


Fig. 5. Equivalent analysis of the connection between the *base* network $\mathcal{N}_{base}$ and the *weight learning* network $\mathcal{N}_{weight}$. One convolution layer whose weights are learnt by the fc layer is exactly equivalent to a multi-path convolution blocks.

use another 4 parameter values (boldface in Table 14) that have not been seen by the network in the training stage but are between the lower and upper bound of its parameter range. As shown in Table 14, we can see that the network performs similarly to the one trained with only one parameter value, but more importantly for the interpolated boldface parameter values that the network does not recognize, it also surprisingly achieves very comparable results.

This means that a few parameter values are already sufficient for learning a good linear transformation in the *weight learning* network from input parameters to convolution weights. However, as in a real scenario, the number of such fixed training parameters is usually difficult to decide, due to many different parameter ranges of image operators. As a result, we choose to sample random parameter values instead of only a few of them for training in our paper.

7 Conclusion

In this paper, we propose the first decouple learning framework for parameterized image operators, where the weights of the task-oriented *base* network $\mathcal{N}_{base}$ are decoupled from the network structure and directly learned by another *weight learning* network $\mathcal{N}_{weight}$. These two networks can be easily end-to-end trained, and $\mathcal{N}_{weight}$ dynamically adjusts the weights of $\mathcal{N}_{base}$ for different parameters $\vec{\gamma}$ during the runtime. We show that the proposed framework can be applied to different parameterized image operators, such as image smoothing, denoising and super resolution, while obtaining comparable performance as the network trained for one specific parameter configuration. It also has the potential to jointly learn multiple different parameterized image operators within one single network. For real user scenarios, we further extend our framework to enable cheap parameter tuning, which obtains superior performance over state-of-the-art methods by a large margin. To better understand the working principle, we also provide some valuable analysis and discussions, which may inspire more promising research in this direction. More theoretical analysis is worthy of further exploration in the future.

References

[1] Q. Fan, J. Yang, G. Hua, B. Chen, and D. Wipf, "A generic deep architecture for single image reflection removal and image smoothing," in *Proceedings of the 16th International Conference on Computer Vision (ICCV)*, 2017, pp. 3238–3247.

[2] J. Kim, J. Kwon Lee, and K. Mu Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1646–1654.

[3] L. Xu, J. Ren, Q. Yan, R. Liao, and J. Jia, "Deep edge-aware filters," in *International Conference on Machine Learning*, 2015, pp. 1669–1678.

[4] J. Chen, A. Adams, N. Wadhwa, and S. W. Hasinoff, "Bilateral guided upsampling," *ACM Transactions on Graphics (TOG)*, vol. 35, no. 6, p. 203, 2016.

[5] M. Gharbi, J. Chen, J. T. Barron, S. W. Hasinoff, and F. Durand, "Deep bilateral learning for real-time image enhancement," *ACM Transactions on Graphics (TOG)*, vol. 36, no. 4, p. 118, 2017.

[6] L. Xu, C. Lu, Y. Xu, and J. Jia, "Image smoothing via l 0 gradient minimization," in *ACM Transactions on Graphics (TOG)*, vol. 30, no. 6. ACM, 2011, p. 174.

[7] L. Karacan, E. Erdem, and A. Erdem, "Structure-preserving image smoothing via region covariances," *ACM Transactions on Graphics (TOG)*, vol. 32, no. 6, p. 176, 2013.

[8] L. Xu, Q. Yan, Y. Xia, and J. Jia, "Structure extraction from texture via relative total variation," *ACM Transactions on Graphics (TOG)*, vol. 31, no. 6, p. 139, 2012.

[9] Q. Zhang, X. Shen, L. Xu, and J. Jia, "Rolling guidance filter," in *European conference on computer vision*. Springer, 2014, pp. 815–830.

[10] A. Buades, B. Coll, and J.-M. Morel, "A non-local algorithm for image denoising," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 2. IEEE, 2005, pp. 60–65.

[11] M. Elad and M. Aharon, "Image denoising via sparse and redundant representations over learned dictionaries," *IEEE Transactions on Image processing*, vol. 15, no. 12, pp. 3736–3745, 2006.

[12] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE transactions on image processing*, vol. 19, no. 11, pp. 2861–2873, 2010.

[13] J. Sun, Z. Xu, and H.-Y. Shum, "Image super-resolution using gradient profile prior," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008, pp. 1–8.

[14] M. E. Tipping and C. M. Bishop, "Bayesian image super-resolution," in *Advances in neural information processing systems*, 2003, pp. 1303–1310.

[15] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 2, pp. 295–307, 2016.

[16] V. Jain and S. Seung, "Natural image denoising with convolutional networks," in *Advances in Neural Information Processing Systems*, 2009, pp. 769–776.

[17] S. Liu, J. Pan, and M.-H. Yang, "Learning recursive filters for low-level vision via a hybrid neural network," in *European Conference on Computer Vision*. Springer, 2016, pp. 560–576.

[18] Q. Chen, J. Xu, and V. Koltun, "Fast image processing with fully-convolutional networks," in *IEEE International Conference on Computer Vision*, vol. 9, 2017.

[19] J. Kopf, M. F. Cohen, D. Lischinski, and M. Uyttendaele, "Joint bilateral upsampling," *ACM Transactions on Graphics (ToG)*, vol. 26, no. 3, p. 96, 2007.

[20] K. He and J. Sun, "Fast guided filter. arxiv 2015," *arXiv preprint arXiv:1505.0099*.

[21] J. Schmidhuber, "Learning to control fast-weight memories: An alternative to dynamic recurrent networks," *Neural Computation*, vol. 4, no. 1, pp. 131–139, 1992.

[22] M. Andrychowicz, M. Denil, S. Gomez, M. W. Hoffman, D. Pfau, T. Schaul, B. Shillingford, and N. De Freitas, "Learning to learn by gradient descent by gradient descent," in *Advances in Neural Information Processing Systems*, 2016, pp. 3981–3989.

[23] O. Wichrowska, N. Maheswaranathan, M. W. Hoffman, S. G. Colmenarejo, M. Denil, N. de Freitas, and J. Sohl-Dickstein, "Learned optimizers that scale and generalize," in *International Conference on Machine Learning*, 2017.

[24] Y. Chen, M. W. Hoffman, S. G. Colmenarejo, M. Denil, T. P. Lillicrap, and N. de Freitas, "Learning to learn for global optimization of black box functions," in *International Conference on Machine Learning*, 2017.

[25] D. Ha, A. Dai, and Q. V. Le, "Hypernetworks," *ICLR*, 2018.

[26] L. Xu, C. Lu, Y. Xu, and J. Jia, "Image smoothing via l0 gradient minimization," *ACM Transactions on Graphics (SIGGRAPH Asia)*, 2011.

[27] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis," in *Proc. CVPR*, 2017.

[28] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," in *ICLR*, 2016.

[29] Z. Farbman, R. Fattal, D. Lischinski, and R. Szeliski, "Edge-preserving decompositions for multi-scale tone and detail manipulation," in *ACM Transactions on Graphics (TOG)*, vol. 27, no. 3. ACM, 2008, p. 67.

[30] Q. Zhang, L. Xu, and J. Jia, "100+ times faster weighted median filter (wmf)," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 2830–2837.

[31] M. Aubry, S. Paris, S. W. Hasinoff, J. Kautz, and F. Durand, "Fast local laplacian filters: Theory and applications," *ACM Transactions on Graphics (TOG)*, vol. 33, no. 5, p. 167, 2014.

[32] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *European conference on computer vision*. Springer, 2014, pp. 184–199.

[33] X. Mao, C. Shen, and Y.-B. Yang, "Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections," in *Advances in neural information processing systems*, 2016, pp. 2802–2810.

[34] C. Dong, Y. Deng, C. Change Loy, and X. Tang, "Compression artifacts reduction by a deep convolutional network," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 576–584.

[35] X. Fu, J. Huang, D. Zeng, Y. Huang, X. Ding, and J. Paisley, "Removing rain from single images via a deep detail network," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1715–1723.

[36] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," *CVPR*, 2017.

[37] Y. Chen and T. Pock, "Trainable nonlinear reaction diffusion: A flexible framework for fast and effective image restoration," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 6, pp. 1256–1272, 2017.

[38] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising," *IEEE Transactions on Image Processing*, vol. 26, no. 7, pp. 3142–3155, 2017.

[39] Y. Tai, J. Yang, X. Liu, and C. Xu, "Memnet: A persistent memory network for image restoration," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4539–4547.

[40] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR Oral)*, June 2016.

[41] J. Kim, J. Kwon Lee, and K. Mu Lee, "Deeply-recursive convolutional network for image super-resolution," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1637–1645.

[42] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, "Image denoising by sparse 3-d transform-domain collaborative filtering," *IEEE Transactions on image processing*, vol. 16, no. 8, pp. 2080–2095, 2007.

[43] S. Gu, L. Zhang, W. Zuo, and X. Feng, "Weighted nuclear norm minimization with application to image denoising," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 2862–2869.

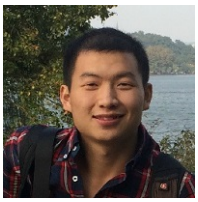
[44] K. Zhang, W. Zuo, S. Gu, and L. Zhang, "Learning deep cnn denoiser prior for image restoration," in *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, 2017.

[45] X. Fu, J. Huang, X. Ding, Y. Liao, and J. Paisley, "Clearing the skies: A deep network architecture for single-image rain removal," *IEEE Transactions on Image Processing*, vol. 26, no. 6, pp. 2944–2956, 2017.

[46] C. Lu, L. Xu, and J. Jia, "Combining sketch and tone for pencil drawing production," in *Proceedings of the Symposium on Non-Photorealistic Animation and Rendering*. Eurographics Association, 2012, pp. 65–73.

[47] J. Koutnik, F. Gomez, and J. Schmidhuber, "Evolving neural networks in compressed weight space," in *Proceedings of the 12th annual conference on Genetic and evolutionary computation*. ACM, 2010, pp. 619–626.

Qingnan Fan is a Postdoctoral Scholar in the Computer Science Department of Stanford University. He received his PhD degree from Shandong University in 2019. His research interests mainly include image/video processing and 3D vision.



Dongdong Chen is a Ph.D. student from the University of Science and Technology of China. He is also a joint PhD between his university and Microsoft Asia. His research interests mainly include style transfer, image generation, low-level image processing and object detection.

Lu Yuan received his PhD degree from the Department of Computer Science and Engineering at the Hong Kong University of Science and Technology in 2009. Now he is a Senior Research Manager in Microsoft Redmond. His research interests include computer vision, applied machine learning and computational photography.

Gang Hua is the Vice President and Chief Scientist of Wormpex AI Research. He was an Associate Professor of Computer Science in Stevens Institute of Technology between 2011 and 2015, while holding an Academic Advisor position at IBM T. J. Watson Research Center. He has published more than 130 peer reviewed papers in top conferences such as CVPR/ICCV/ECCV, and top journals such as T-PAMI and IJCV. To date He holds 18 issued U.S Patents and also has 14 more U.S. Patents Pending. He is an IAPR Fellow, an ACM Distinguished Scientist, and a Senior Member of IEEE. His research focuses on artificial intelligence, computer vision, pattern recognition, machine learning, and robotics, with primary applications in the cloud and mobile intelligence domain.

Nenghai Yu is a full Professor at University of Science and Technology of China. He is also the director of Information Processing Center of USTC,deputy director of academic committee of School of Information Science and Technology. He was a visiting scholar in Institute of Production Technology, Faculty of Engineering, University of Tokyo, in 1999 and did cooperative research as the senior visiting scholar in Dept. of Electrical Engineering, Columbia University, from Apr. to Oct. 2008. His research focuses on image processing and video analysis, multimedia communication, media content security, Internet information retrieval, data mining and content filtering, network communication and security.

Baoquan Chen is a Professor of Peking University, where he is the Executive Director of the Center on Frontiers of Computing Studies. Prior to the current post, he was the Dean of School of Computer Science and Technology at Shandong University, and the founding director of the Visual Computing Research Center, Shenzhen Institute of Advanced Technology (SIAT), Chinese Academy of Sciences (2008-2013), and a faculty member at Computer Science and Engineering at the University of Minnesota at Twin Cities (2000-2008). His research interests generally lie in computer graphics, visualization, and human-computer interaction, focusing specifically on large-scale city modeling, simulation and visualization.