Connecting Image Denoising and High-Level Vision Tasks via Deep Learning

Ding Liu, *Member, IEEE*, Bihan Wen, *Member, IEEE*, Jianbo Jiao, *Member, IEEE*, Xianming Liu,
Zhangyang Wang, *Member, IEEE*, and Thomas S. Huang, *Life Fellow, IEEE*

Abstract—Image denoising and high-level vision tasks are usually handled independently in the conventional practice of computer vision, and their connection is fragile. In this paper, we cope with the two jointly and explore the mutual influence between them with the focus on two questions, namely (1) how image denoising can help improving high-level vision tasks, and (2) how the semantic information from high-level vision tasks can be used to guide image denoising. First for image denoising we propose a convolutional neural network in which convolutions are conducted in various spatial resolutions via downsampling and upsampling operations in order to fuse and exploit contextual information on different scales. Second we propose a deep neural network solution that cascades two modules for image denoising and various high-level tasks, respectively, and use the joint loss for updating only the denoising network via back-propagation. We experimentally show that on one hand, the proposed denoiser has the generality to overcome the performance degradation of different high-level vision tasks. On the other hand, with the guidance of high-level vision information, the denoising network produces more visually appealing results. Extensive experiments demonstrate the benefit of exploiting image semantics simultaneously for image denoising and high-level vision tasks via deep learning. The code is available online: https://github.com/Ding-Liu/DeepDenoising

Index Terms—Deep learning, neural network, image denoising, high-level vision.

I. INTRODUCTION

CONVENTIONALLY, low-level image processing problems, such as image restoration and enhancement, and high-level vision tasks are handled separately by different approaches in computer vision. In this work, we connect them

Ding Liu is with Bytedance Inc., Mountain View, CA 61801 USA (e-mail: liudingdavy@gmail.com).

Bihan Wen is with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798 (e-mail: bihan.wen@ntu.edu.sg).

Jianbo Jiao is with the Department of Engineering Science, University of Oxford, Oxford OX1 2JD, U.K. (e-mail: jianbo.jiao@eng.ox.ac.uk).

Xianming Liu is with Horizon Robotics, Cupertino, CA 95014 USA (e-mail: xianming.liu@horizon.ai).

Zhangyang Wang is with the Department of Computer Science and Engineering, Texas A&M University, College Station, TX 77843 USA (e-mail: atlaswang@tamu.edu).

Thomas S. Huang is with the Beckman Institute, University of Illinois at Urbana–Champaign, Urbana, IL 61801 USA (e-mail: t-huang1@illinois.edu).

Digital Object Identifier 10.1109/TIP.2020.2964518

by showing the mutual influence between the two, i.e., visual perception and semantics, and propose a new perspective for solving both the low-level image processing and high-level computer vision problems in a single unified framework.

Image denoising, as one representative of low-level image processing problems, aims to recover the underlying image signal from its noisy measurement. Conventional image denoising approaches exploit either local or non-local image characteristics [1]–[6]. In recent years, we have witnessed the revival of deep learning in computer vision and deep neural networks have been developed for image denoising with notable performance improvement [7]–[11]. In this work, we propose a new convolutional neural network for image denoising. Inspired by U-Net [12], we conduct convolutions in different spatial scales via downsampling and upsampling operations before the resulting features are fused together, so that the kernels have a larger receptive field after all feature contraction. Downsampling operations also help to reduce the computation cost due to the reduction of feature map sizes. We conduct extensive experiments and show that our proposed denoising network achieves the state-of-the-art performance across benchmark datasets.

Many popular image denoising approaches only minimizes the mean square error (MSE) between the reconstructed and clean images whereas important image details are sometimes lost which results in image quality degradation; e.g., over-smoothing artifacts in some texture-rich regions are commonly observed in the denoised result from conventional methods. Meanwhile, the semantical information of the image is usually ignored during denoising. To overcome this drawback, we propose to cascade the network models for image denoising and a high-level vision task, respectively. During training, we jointly minimize the image reconstruction loss, the perceptual loss over the feature domain, and the high-level vision loss. With the guidance of image semantic information, the denoising network is able to further improve visual quality and generate more visually appealing results, which demonstrates the importance of semantic information for image denoising.

When high-level vision tasks are conducted on noisy data, image restoration is typically applied as an independent pre-processing step, which might be suboptimal for the ultimate goal [13]–[15]. Recent research reveals that neural networks trained for image classification can be easily fooled by small noise perturbation or other artificial patterns [16], [17]. However, to the best of our knowledge, the problem of how

low-level image processing could affect high-level semantical tasks is still not thoroughly studied [13], [14]. In this work, we observe an intriguing result that applying a conventional denoising method such as CBM3D [18] over noisy images may introduce unsatisfactory artifacts, which results in poor results of image classification and semantic segmentation. Therefore, we aim to produce an application-driven denoiser which is capable of simultaneously removing noise and preserving semantic-aware details for the high-level vision tasks. To this end, in our proposed cascaded architecture we keep the high-level vision network untouched but only use the gradients of the joint loss to update the denoising network, in order to make sure the denoising network has good generalization about other high-level vision tasks. This also ensures that the denoising module removes noises while preserving the important semantic details to produce correct high-level vision task results.

We systematically investigate the mutual influence between the low-level image denoising and high-level vision networks using our proposed architecture via numerous experiments. We show that the cascaded network trained with the joint loss not only boosts the perceptual quality of denoised images via image semantic guidance, but also substantially improves the accuracy of high-level vision tasks. Moreover, our proposed training strategy makes the trained denoising network generalizable to different high-level vision tasks. In other words, our denoising module trained for one high-level vision task can be directly plugged into other high-level tasks without fine-tuning either module, which facilitates the training effort when applied to various high-level vision tasks and keeps the high-level vision networks performing consistently for noisy and noiseless images.

In short, the main contributions of this paper are as follows:

- To the best of our knowledge, this is the first attempt to investigate the benefit of exploiting image semantics simultaneously for image denoising and high-level vision tasks under a unified deep learning framework.
- We demonstrate that high-level semantics can be used for image denoising to generate visually appealing results in a deep learning fashion.
- Our proposed training strategy enables the robustness of the trained denoising network to various high-level vision tasks so that it can directly work with various high-level vision networks without any fine-tuning.

This paper is built upon our previous work [19] with several notable improvements. First, compared with [19] we additionally incorporate the perceptual loss over the feature domain in our combined loss in order to capture more semantic information from multilevel feature domains. Second, we provide a detailed ablation study of the denoising network architecture, and we show an analysis of the gradient visualization of the cascade architecture, in order to further investigate how high-level image semantics guides low-level image denoising. Third, we conduct a subjective evaluation on several recent image denoising methods to thoroughly measure the visual quality of denoised results. Finally, we present a more comprehensive experiment section including comparisons of denoising results over more benchmark datasets,

and we extend our proposed denoising method to real noise removal.

The remainder of this paper is organized as follows. Section II reviews existing image denoising methods in the literature. Section III presents our image denoising network and the proposed cascaded architecture. Section IV provides the experimental results. Section V concludes this paper.

II. RELATED WORK

Denoising is the task of estimating the high-quality signal from its noisy measurements. Classical image denoising methods take advantage of local or non-local structures presented in the image explicitly. Natural images are well-known to be patch-wise sparse or compressible in transform domain, or over certain dictionary. Prior works [1], [20], [21] exploit such local structures and reduce noise by coefficient shrinkage for image restoration. The later approaches, including SSC [3], CSR [22], NCSR [4], GSR [23], WNNM [5], PCLR [24], PGPD [6], STROLLR [25], as well as BM3D [2] and its extension for color images CBM3D [18] – group similar patches within the image globally via block matching or clustering, and impose non-local structural priors on these groups, which usually lead to state-of-the-art image denoising performance.

More recently, the popular deep neural network techniques have been applied to low-level vision tasks. Specifically, a number of deep learning models have been developed for image denoising [7]–[9], [26]–[28], which can be classified into two categories: multilayer perception (MLP) based models and convolutional neural network (CNN) based models. Early MLP based models for image denoising include the stacked denoising autoencoder [7], which is an extension of the stacked autoencoder originally designed for unsupervised feature learning. A denoising autoencoder is proposed for both image denoising and blind image inpainting by Xie et al. [27]. Burger *et al.* [8] introduce a plain MLP and thoroughly compare its denoising performance with BM3D [2] in different experimental settings. CNNs are first utilized for image denoising by Jain and Seung [26]. Recent works [10], [29] attempt to unfold the iterative algorithms, and construct a cascaded convolutional filtering architecture for image denoising. Very deep networks are developed with skip connections for image restoration [9], [11], and achieve superior performance over other recent methods. Dilated convolutions are adopted to learn the residual image for denoising in [30]. A CNN is designed to implicitly learn the regularizer of the alternating minimization algorithm for regularization-based image restoration methods in [28]. The formatted residual information between the noiseless image and its denoised version is learned via a second network in [31]. The non-local self-similarity property of natural images are exploited with CNNs in [32], [33]. FFDNet is proposed in [34] to work on downsampled images in order to reduce the inference time. Non-local self-similarity of images is exploited and incorporated into a recurrent neural network in [35].

Class information is exploited for improving image denoising in [36], where a two-stage strategy is adopted. Specifically,
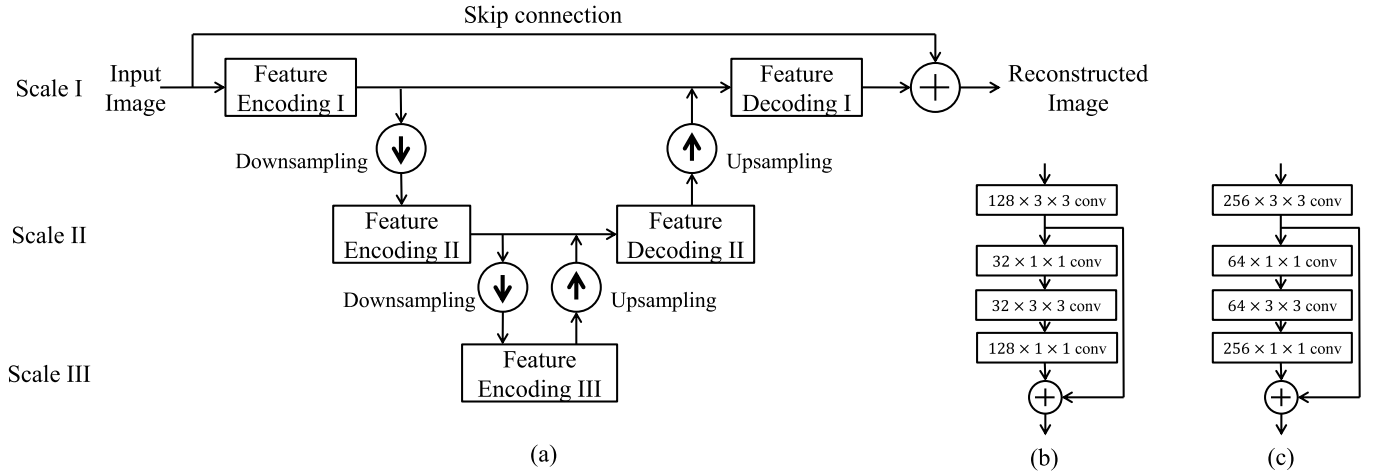
Fig. 1. (a) Overview of our proposed denoising network. (b) Architecture of the feature encoding module. (c) Architecture of the feature decoding module.

first the images are classified and then go through separate class-specific denoisers. As the classification and denoising stages are decoupled, this approach may lead to sub-optimal solution, comparing to our proposed end-to-end framework in which denoisers and classifiers are jointly trained.

III. METHOD

In this section, we first introduce the denoising network utilized in our framework, and then explain our framework that cascades this image denoising module and the module for high-level vision task with its training strategy and joint loss in detail.

A. Denoising Network

We propose a convolutional neural network for image denoising, which takes a noisy image as input and outputs the reconstructed image. This network conducts feature contraction and expansion through downsampling and upsampling operations, respectively. Each pair of downsampling and upsampling operations brings the feature representation into a new spatial scale, so that the whole network can process information on different scales.

Specifically, on each scale, the input is encoded after downsampling the features from the previous scale. After feature encoding and decoding possibly with features on the next scale, the output is upsampled and fused with the feature on the previous scale. Such pairs of downsampling and upsampling steps can be nested to build deeper networks with more spatial scales of feature representation, which generally leads to better restoration performance. Considering the trade-off between computation cost and restoration accuracy, we choose three scales for the denoising network in our experiments, while this framework can be easily extended for more scales. The ablation study of the denoising network architecture can be found in Section IV-A1.

These operations together are designed to learn the residual between the input and the target output and recover as many details as possible, so we use a long-distance skip connection to sum the output of these operations and the input image, in order to generate the reconstructed image. The overview is

in Figure 1 (a). Each module in this network will be elaborated on as follows.

1) Feature Encoding: We design one feature encoding module on each scale, which is one convolutional layer plus one residual block as in [37]. The architecture is displayed in Figure 1 (b). Note that each convolutional layer is immediately followed by spatial batch normalization and a ReLU neuron. From top to down, the four convolutional layers have 128, 32, 32 and 128 kernels in size of 3×3, 1×1, 3×3 and 1×1, respectively. The output of the first convolutional layer is passed through a skip connection for element-wise sum with the output of the last convolutional layer.

2) Feature Decoding: The feature decoding module is designed for fusing information from two adjacent scales. Two fusion schemes are tested: (1) concatenation of features on these two scales; (2) element-wise sum of them. Both schemes obtain similar denoising performance. Thus we choose the first scheme to accommodate feature representations of different channel numbers from two scales. We use a similar architecture as the feature encoding module except that the numbers of kernels in the four convolutional layers are 256, 64, 64 and 256. Its architecture is in Figure 1(c).

3) Feature Downsampling & Upsampling: Downsampling operations are adopted multiple times to progressively increase the receptive field of the following convolution kernels and to reduce the computation cost by decreasing the feature map size. The larger receptive field enables the kernels to incorporate larger spatial context for denoising. We use 2 as both the downsampling factor and the upsampling factor, and try two schemes for downsampling in the experiments: (1) max pooling with stride of 2; (2) conducting convolutions with stride of 2. Both schemes achieve similar denoising performance in practice, so we use the second scheme in the rest of the experiments for computation efficiency. Upsampling operations are implemented by deconvolution with 4×4 kernels, which aim to expand the feature map to the same spatial size as the previous scale.

Since all the operations in our proposed denoising network are spatially invariant, it has the merit of handling input images of arbitrary size.
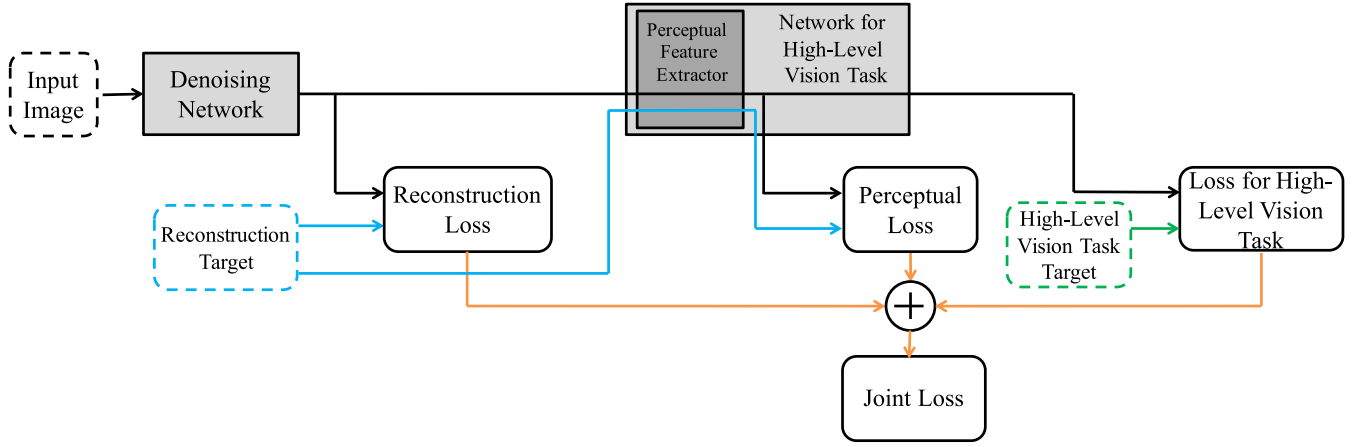
Fig. 2.   Overview of our proposed cascaded network.

B. When Image Denoising Meets High-Level Vision Tasks

We propose a robust deep architecture processing a noisy image input, via cascading a network for denoising and the other for high-level vision task, aiming to simultaneously:

1) reconstruct visually pleasing results guided by the high-level vision information, as the output of the denoising network;

2) attain sufficiently good accuracy across various high-level vision tasks, when trained for only one high-level vision task.

The overview of the proposed cascaded network is displayed in Figure 2. Specifically, given a noisy input image, the denoising network is first applied, and the denoised result is then fed into the following network for high-level vision task, which generates the high-level vision task output.

1) Training Strategy: First we initialize the network for high-level vision task from a network that is well-trained in the noiseless setting. We train the cascade of two networks in an end-to-end manner while fixing the weights in the network for high-level vision task. Only the weights in the denoising network are updated by the error back-propagated from the following network for high-level vision task, which is similar to minimizing the perceptual loss for image super-resolution [38]. The reason for adopting such a training strategy is to make the trained denoising network robust enough without losing the generality for various high-level vision tasks. More specifically, our denoising module trained for one high-level vision task can be directly plugged into networks for other high-level tasks without fine-tuning either the denoiser or the high-level network. Our approach not only facilitates the training effort when applying the denoiser to different high-level tasks while keeping the high-level vision network performing consistently for noisy and noise-free images, but also enables the denoising network to produce high-quality perceptual and semantically faithful results.

2) Loss: Recent works [39], [40] discover that multi-level feature statistics extracted from a trained CNN notably represent the visual semantics in various scales. We aim to use this property to enhance the quality of the denoised results from our framework. There are three loss components adopted in our framework, and we describe them sequentially as follows.

The reconstruction loss of the denoising network is the mean squared error (MSE) between the denoising network output and the noiseless image, which can be represented as

$$L_D(x, \tilde{x}) = \frac{1}{HW} \sum_{i=1}^{H} \sum_{j=1}^{W} (F_D(x)_{i,j} - \tilde{x}_{i,j})^2, \qquad (1)$$

where x is the noisy input image and $\tilde{x}$ is the noiseless image. F_D denotes the denoising network. x, $\tilde{x}$ and $F_D(x)$ are of size $H \times W$. $\tilde{x}_{i,j}$ indicates the pixel in the i-th column and the j-th row of $\tilde{x}$.

In addition to the reconstruction loss calculated on the image domain, we define the perceptual loss as the euclidean distance between the feature representations of a reconstructed image $F_D(x)$ and the reference image $\tilde{x}$ [38]. Here we choose the perceptual feature extractor $\phi(\cdot)$ to be the first several layers of the high-level vision network. Specifically, when we adopt VGG-based networks for image classification and semantic segmentation in our experiments, we use the feature $relu3\_3$ as the perceptual feature representation. The perceptual loss can be represented as

$$L_P(F_D(x), \tilde{x}) = \frac{1}{H_\phi W_\phi} \sum_{i=1}^{H_\phi} \sum_{j=1}^{W_\phi} (\phi(F_D(x))_{i,j} - \phi(\tilde{x})_{i,j})^2,$$
$$(2)$$

where the perceptual feature representations of $\phi(\tilde{x})$ and $\phi(F_D(x))$ are of size $H_\phi \times W_\phi$.

The losses of the classification task and the segmentation task both are the cross-entropy loss between the predicted label and the ground truth label. The joint loss is defined as the weighted sum of the reconstruction loss, the perceptual loss and the loss for high-level vision task, which can be represented as

$$L(x, \tilde{x}, y) = L_D(x, \tilde{x}) + \lambda_P L_P(F_D(x), \tilde{x}) + \lambda_H L_H(F_D(x), y),$$
$$(3)$$

where y is the ground truth label of high-level vision task. $L_H(F_D(x), y)$ represents the loss of the high-level vision network with $F_D(x)$ as the input, and L is the joint loss, as illustrated in Figure 2. λ_P and λ_H are the weights for balancing the losses L_D, L_P and L_H.

TABLE I

DIFFERENT NETWORK DESIGNS AND THEIR CORRESPONDING DENOISING PERFORMANCES. WE FINALLY
CHOOSE THE ARCHITECTURE IN BOLD FOR OUR DENOISING NETWORK

Scale No.	Skip connection	Encoding module	Decoding module	Parameters	PSNR
2	✓	(128, 32, 32, 128)	(128, 32, 32, 128)	384k	30.41
2	✗	(128, 32, 32, 128)	(256, 64, 64, 256)	637k	30.39
2	✓	(128, 32, 32, 128)	(256, 64, 64, 256)	637k	30.54
2	✓	(256, 64, 64, 256)	(256, 64, 64, 256)	1780k	30.70
3	✓	(128, 32, 32, 128)	(128, 32, 32, 128)	592k	30.64
3	✗	(128, 32, 32, 128)	(256, 64, 64, 256)	1,026k	30.52
3	**✓**	**(128, 32, 32, 128)**	**(256, 64, 64, 256)**	**1,026k**	**30.83**
3	✓	(256, 64, 64, 256)	(256, 64, 64, 256)	2,198k	30.87
4	✓	(128, 32, 32, 128)	(128, 32, 32, 128)	970k	30.76
4	✗	(128, 32, 32, 128)	(256, 64, 64, 256)	1,488k	30.67
4	✓	(128, 32, 32, 128)	(256, 64, 64, 256)	1,488k	30.91
4	✓	(256, 64, 64, 256)	(256, 64, 64, 256)	3,945k	30.96

IV. EXPERIMENTS

In this section, we first show the experiment results of our proposed denoising network, and then discuss its relation with high-level vision networks. Finally, we extend our denoising network to the case of real noise removal.

A. Image Denoising

Our proposed denoising network takes RGB images as input, and outputs the reconstructed images directly. We add independent and identically distributed Gaussian noise with zero mean to the original image as the noisy input image during training. We use the same training set as in [11], which is 432 color images from the Berkeley segmentation dataset [41]. The loss of separately training such a denoising network is equivalent to Equation 3 as $\lambda_P = \lambda_H = 0$. We use SGD with a batch size of 32, and the input patches are 48×48 pixels. The initial learning rate is set as 10^{-4} and is divided by 10 after every 500,000 iterations. The training is terminated after 1,500,000 iterations. We train a different denoising network for each noise level in our experiment.

1) Analysis of Network Architecture: To investigate the architecture of our proposed denoising network, we design a set of controlled experiments with different network structures, in order to analyze a number of design choices and examine their contributions to the overall denoising performance. Specifically, we change the number of spatial scales of feature maps, the design of the encoding and decoding modules, as well as the existence of the global skip connection. Recall that we use one convolutional layer plus one residual block from [37] as the encoding and decoding modules, and we test the case in which the four convolutional layers in each module have (128, 32, 32, 128) kernels or (256, 64, 64, 256) kernels. We use these models to test on the widely used Kodak dataset,[1] which consists of 24 color images with $\sigma = 35$. The average PSNR (in dB) and the number of model parameters are displayed in Table I. It is observed that the performance gets better when the computation cost becomes higher (i.e. more spatial scales and more model parameters). The global

skip connection improves the results. Considering the trade-off between the denoising performance and the computational cost, we adopt the encoding module of (128, 32, 32, 128) kernels and the decoding module of (256, 64, 64, 256) kernels with three spatial scales and the global skip connection in the rest of our experiments.

2) Comparisons With State-of-the-Art Methods: We compare our denoising network with several state-of-the-art color image denoising approaches on various noise levels: $\sigma = 25, 35$ and 50. We evaluate their denoising performance over the following two datasets: the Kodak dataset, and the CBSD68 dataset which consists of 68 color images as in [34]. Table II and III shows the peak signal-to-noise ratio (PSNR) results for CBM3D [18], MCWNNM [42], DnCNN [11], FFDNet [34] and our proposed method on these two datasets, respectively. It is noteworthy that we calculate PSNRs on the whole images of the Kodak dataset, rather than the cropped regions used in [34]. We do not list other methods [5], [8], [10], [20], [30] whose performance is worse than DnCNN or FFDNet. The implementation codes used are from the authors' websites[2] and the default parameter settings are adopted in our experiments.

It is clear that our proposed method outperforms all the competing approaches quantitatively across different noise levels. It achieves the highest PSNR in almost every image of the Kodak dataset, and obtains the highest average PSNR over the CBSD68 dataset.

B. When Image Denoising Meets High-Level Vision Tasks

We choose two high-level vision tasks as representatives in our study: image classification and semantic segmentation [43], which have been dominated by deep network based models. We utilize two popular VGG-based deep networks in our system for each task, respectively. *VGG-16* in [44] is employed for image classification; we select *DeepLab-LargeFOV* in [45] for semantic segmentation. We follow the

---

[1] http://r0k.us/graphics/kodak/

[2] CBM3D: http://www.cs.tut.fi/ foi/GCF-BM3D
MCWNNM: https://github.com/csjunxu/MCWNNM-ICCV2017
DnCNN: https://github.com/cszn/DnCNN
FFDNet: https://github.com/cszn/FFDNet

TABLE II

COLOR IMAGE DENOISING RESULTS (PSNR) OF DIFFERENT METHODS ON THE KODAK DATASET. THE BEST RESULT IS SHOWN IN BOLD

Image	$\sigma = 25$					$\sigma = 35$					$\sigma = 50$				
	CBM3D	MCWNNM	DnCNN	FFDNet	Proposed	CBM3D	MCWNNM	DnCNN	FFDNet	Proposed	CBM3D	MCWNNM	DnCNN	FFDNet	Proposed
01	29.13	28.66	29.75	29.65	**29.76**	27.31	26.93	28.10	28.01	**28.11**	25.86	25.28	26.52	26.45	**26.55**
02	32.44	31.92	32.97	32.86	**33.00**	31.07	30.62	31.65	31.60	**31.75**	29.84	29.27	30.44	30.35	**30.54**
03	34.54	34.05	34.97	34.92	**35.12**	32.62	32.27	33.37	33.38	**33.58**	31.34	30.52	31.76	31.77	**31.99**
04	32.67	32.42	32.94	32.90	**33.01**	31.02	30.92	31.51	31.49	**31.59**	29.92	29.37	30.12	30.09	**30.22**
05	29.73	29.37	30.53	30.35	**30.55**	27.61	27.53	28.66	28.52	**28.72**	25.92	25.60	26.77	26.67	**26.87**
06	30.59	30.18	31.05	30.99	**31.08**	28.78	28.44	29.37	29.34	**29.45**	27.34	26.70	27.74	27.72	**27.85**
07	33.66	33.36	34.42	34.31	**34.47**	31.64	31.53	32.60	32.54	**32.70**	29.99	29.51	30.67	30.66	**30.82**
08	29.88	29.39	30.30	30.04	**30.37**	27.82	27.67	28.53	28.30	**28.64**	26.23	25.86	26.65	26.50	**26.84**
09	34.06	33.42	34.59	34.52	**34.63**	32.28	31.76	33.06	33.00	**33.11**	30.86	30.00	31.42	31.40	**31.53**
10	33.82	33.23	34.33	34.24	**34.38**	31.97	31.51	32.74	32.68	**32.83**	30.48	29.63	31.03	30.99	**31.17**
11	31.25	30.62	31.82	31.70	**31.84**	29.53	29.04	30.23	30.14	**30.29**	28.00	27.41	28.67	28.56	**28.76**
12	33.76	33.02	34.12	34.04	**34.18**	32.24	31.52	32.73	32.65	**32.83**	30.98	30.00	31.32	31.24	**31.47**
13	27.64	27.19	**28.26**	28.17	28.24	25.70	25.40	26.46	26.39	**26.47**	24.03	23.70	24.73	24.67	**24.76**
14	30.03	29.67	30.79	30.67	**30.80**	28.24	28.05	29.17	29.07	**29.20**	26.74	26.43	27.57	27.48	**27.63**
15	33.08	32.69	33.32	33.24	**33.35**	31.47	31.15	31.89	31.82	**31.96**	30.32	29.59	30.50	30.44	**30.59**
16	32.33	31.79	32.69	32.66	**32.74**	30.64	30.15	31.16	31.14	**31.23**	29.36	28.53	29.68	29.63	**29.78**
17	32.93	32.39	**33.53**	33.47	33.50	30.64	30.75	31.96	31.94	**31.98**	29.36	28.98	30.33	30.33	**30.40**
18	29.83	29.46	30.40	30.31	**30.46**	28.00	27.70	28.72	28.63	**28.79**	26.41	25.94	27.03	26.96	**27.14**
19	31.78	31.29	**32.23**	32.12	32.30	30.19	29.86	30.80	30.70	**30.88**	29.06	28.44	29.34	29.26	**29.49**
20	33.45	32.78	34.15	34.03	**34.29**	31.84	31.32	32.73	32.65	**32.91**	30.51	29.79	31.28	31.21	**31.53**
21	30.99	30.55	31.61	31.53	**31.63**	29.17	28.86	29.94	29.88	**29.98**	27.61	27.13	28.27	28.21	**28.34**
22	30.93	30.48	**31.41**	31.33	31.38	29.36	28.93	29.94	29.88	**29.95**	28.09	27.47	28.54	28.48	**28.58**
23	34.79	34.45	35.36	35.31	**35.40**	33.09	32.79	33.86	33.82	**33.89**	31.75	30.96	32.18	32.22	**32.30**
24	30.09	29.93	**30.79**	30.66	30.77	28.19	28.17	28.98	28.89	**29.03**	26.62	26.37	27.18	27.11	**27.30**
Average	31.81	31.35	32.35	32.25	**32.39**	30.04	29.70	30.76	30.69	**30.83**	28.62	28.02	29.16	29.10	**29.27**

TABLE III

COLOR IMAGE DENOISING RESULTS (PSNR) OF DIFFERENT METHODS ON THE CBSD68 DATASET. THE BEST RESULT IS SHOWN IN BOLD

	CBM3D	MCWNNM	DnCNN	FFDNet	Proposed
$\sigma = 25$	30.71	30.23	31.23	31.21	**31.32**
$\sigma = 35$	28.89	28.57	29.58	29.58	**29.73**
$\sigma = 50$	27.38	26.63	27.92	27.76	**28.09**

preprocessing protocols (e.g. crop size, mean removal of each color channel) in [44] and [45] accordingly while training and deploying them in our experiments.

As for the cascaded network for image classification and the corresponding experiments, we train our model on ILSVRC2012 training set, and evaluate the classification accuracy on ILSVRC2012 validation set. λ_P and λ_H are empirically set as 0.5 and 0.25, respectively. As for the cascaded network for image semantic segmentation and its corresponding experiments, we train our model on the augmented training set of Pascal VOC 2012 as in [45], and test on its validation set. λ_P and λ_H are empirically set as 0.5 and 0.5, respectively.

1) High-Level Vision Information Guided Image Denoising: The typical metric used for image denoising is PSNR, which has been shown to sometimes correlate poorly with human assessment of visual quality [46]. Since PSNR depends on the reconstruction error between the denoised output and the reference image, a model trained by minimizing MSE on the image domain should always outperform a model trained by minimizing our proposed joint loss (with the guidance of high-level vision semantics) in the metric of PSNR. Therefore, we emphasize that the goal of our following experiments is not to pursue the highest PSNR, but to demonstrate the qualitative difference between the model trained with our proposed joint loss and the model trained with MSE on the image domain.

In our cascade architecture, we study the mutual influence between image denoising and high-level vision tasks. Considering image classification as an example of high-level vision tasks, the loss used in our framework is composed of (1) the image reconstruction loss from the denoising network, (2) the perceptual loss from the VGG-based network, and (3) the classification loss from the VGG-based network. First we analyze the contribution of each loss component, by visualizing the gradient with respect to the input noisy image for applying only (1) the image reconstruction loss, (2) the perceptual loss, and (3) the classification loss, respectively. Several examples from ILSVRC2012 validation set are displayed in Figure 3.

Figure 3(c) shows that the energy is spread out across the visualization of gradient map with respect to the input noisy image for only the reconstruction loss, since the additive noise is distributed ubiquitously in the same level over the image, and the reconstruction loss is not spatially biased. In contrast, Figure 3(e) shows the visualization of gradient map with respect to the input noisy image for only the classification loss, where the energy is more concentrated on regions with class-specific details, which usually contain key features to recognize the class of images. These regions are generally the texture-rich regions that are prone to suffer from oversmoothing in conventional denoising approaches, such as the fur of fox and dog, as well as the feather of bird in Figure 3. Figure 3(d) is the visualization of gradient map with respect to the input noisy image for only the perceptual loss, and the energy tends to focus on a number of local features. In other words, with the class information as auxiliary supervision, our denoising network trained with the joint loss is able to denoise such class-specific regions differently from the rest of image, which demonstrates that these three types of losses work in a complementary manner for denoising. In addition, since the joint loss is the linear combination of these three losses, by adjusting the weights of the joint loss, one can control the visual quality of the denoising results which preserves different types/regions of image details.

Figure 4 displays two image denoising examples from the Kodak dataset. A visual comparison is illustrated for a zoom-in region: (II) and (III) are the denoising results using CBM3D [18], and DnCNN [11], respectively; (IV) is the proposed denoiser trained separately without the guidance of high-level vision information; (V) is the denoising result
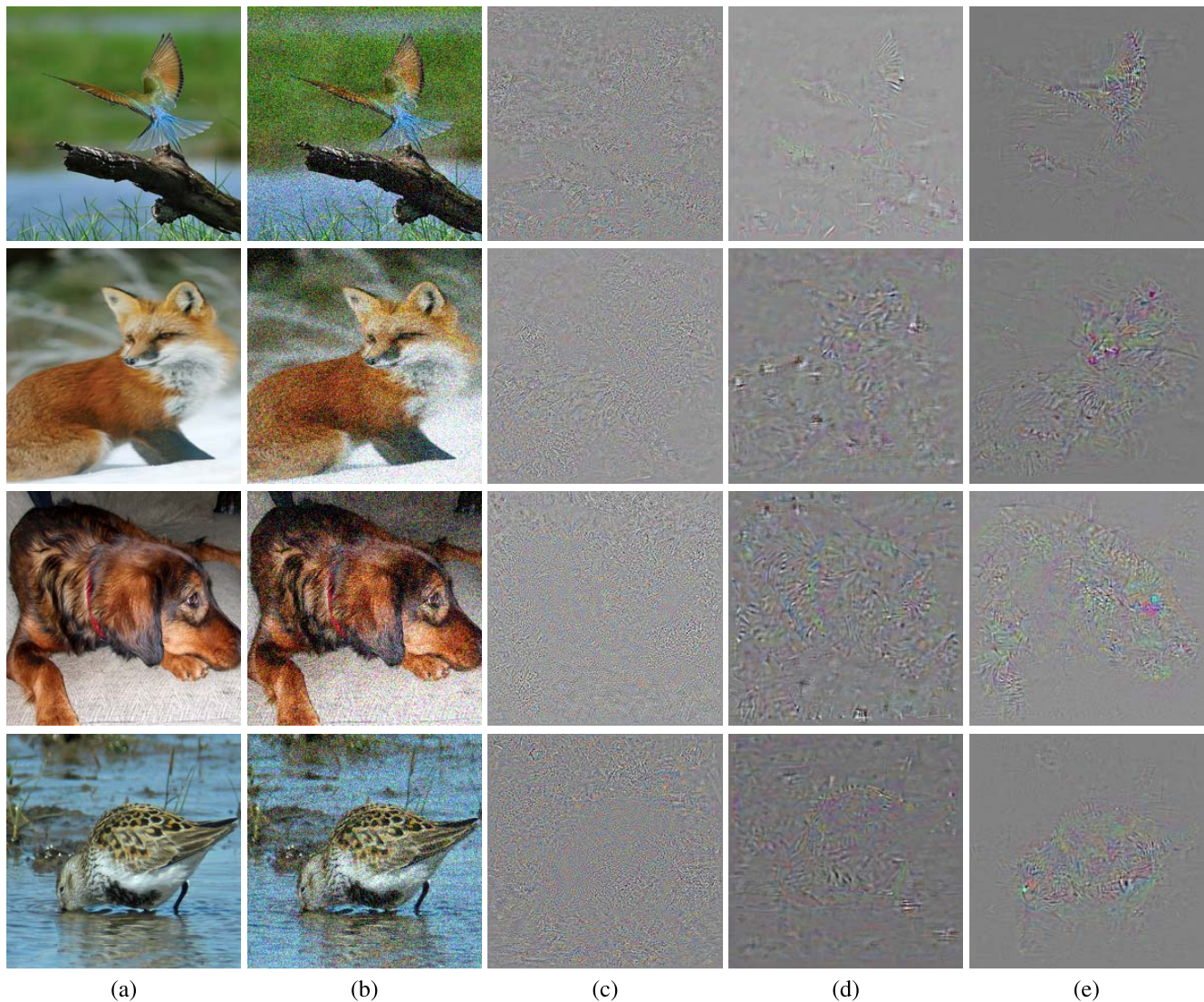
Fig. 3. From left to right: (a) Ground truth images in ILSVRC2012 validation set. (b) Noisy images. (c) The visualization of gradient with respect to the input noisy image for only the reconstruction loss. (d) The visualization of gradient with respect to the input noisy image for only the perceptual loss. (e) The visualization of gradient with respect to the input noisy image for only the classification loss.

using the proposed denoising network trained jointly with a segmentation network. We can find that the results using CBM3D, DnCNN and our separately trained denoiser generate oversmoothing regions, while the jointly trained denoising network is able to reconstruct the denoised image which preserves more details and textures with better visual quality.

In order to more thoroughly quantify the perceptual quality of the denoised results, we perform a mean opinion score (MOS) test. We compare the following four denoising methods: CBM3D, DnCNN, our proposed denoising network without high-level guidance and with high-level guidance. In our test, for each noisy image the denoised results from these four methods are shown simultaneously. We use scores from 1 to 4 to indicate from low image quality to high image quality, and ask the raters to rank the denoised results by assigning corresponding scores. 15 raters are invited to participate our test. In each test, 5 images from the Kodak dataset and 5 images from the CBSD68 dataset are randomly

selected and the denoised results are from the noise level of 50. The MOS result of our perceptual evaluation is shown in Figure 5. We find that our proposed denoising method with high-level guidance achieves the highest score, showing high-level semantics is able to improve the visual perception in our proposed framework.

2) Generality of the Denoiser for High-Level Vision Tasks: We now investigate how the image denoising can enhance the high-level vision applications, including image classification and semantic segmentation, over the ILSVRC2012 and Pascal VOC 2012 datasets, respectively. The noisy images ($\sigma = 15, 30, 45, 60$) are denoised and then fed into the VGG-based networks for high-level vision tasks. To evaluate how different denoising schemes contribute to the performance of high-level vision tasks, we experiment with the following cases:

- Noisy images are directly fed into the high-level vision network, termed as *VGG*. This approach serves as the baseline.
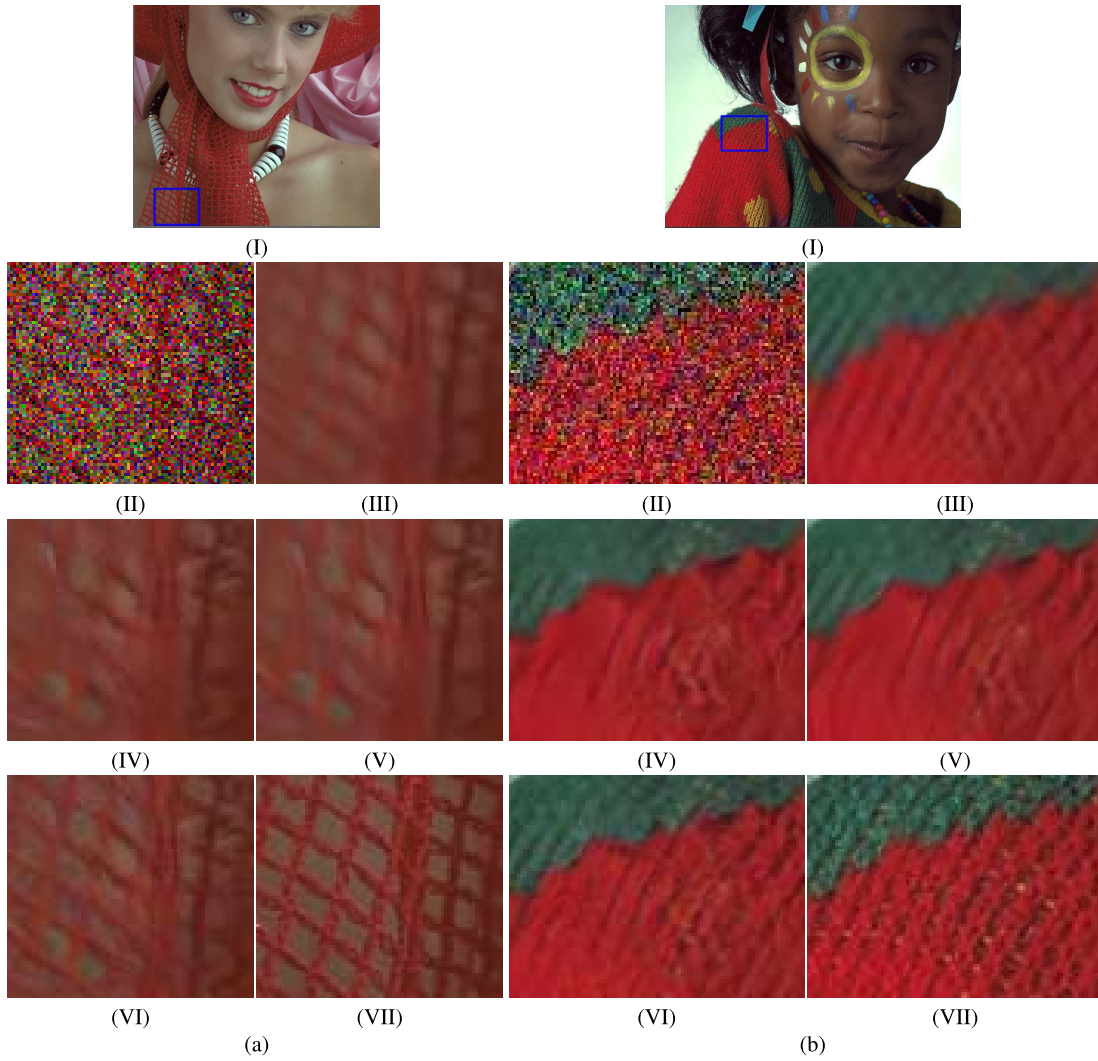
Fig. 4. Two image denoising examples from the Kodak dataset with noise level of 50 are displayed in (a) and (b). We show (I) the ground truth image and the zoom-in regions of: (II) the noisy image; (III) the denoised image by CBM3D; (IV) the denoised image by DnCNN; the denoising result of our proposed model (V) without the guidance of high-level vision information; (VI) with the guidance of high-level vision information and (VII) the ground truth.
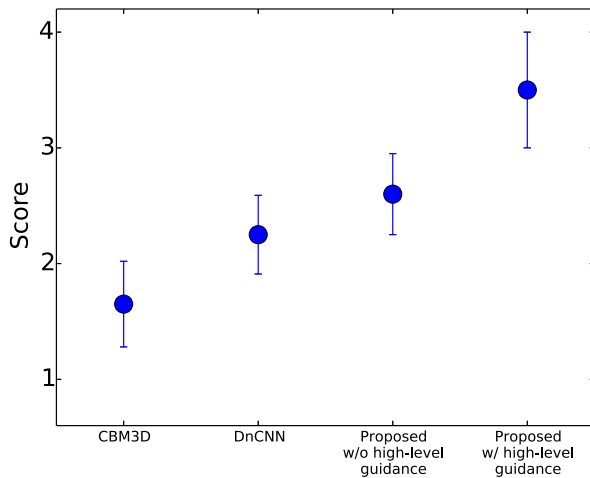

Fig. 5. Mean opinion scores of the subjective evaluation for four denoising methods. The longer bar indicates the larger variance of the scores.

- Noisy images are first denoised by CBM3D, and then fed into the high-level vision network, termed as *CBM3D+VGG*.

- Noisy images are denoised via the separately trained denoising network, and then fed into the high-level vision network, termed as *Separate+VGG*.
- Our proposed approach: noisy images are processed by the cascade of these two networks, which is trained using the joint loss, termed as *Joint Training*.
- A denoising network is trained with the classification network in our proposed approach, but then is connected to the segmentation network and evaluated for the task of semantic segmentation, or vice versa. This is to validate the generality of our denoiser for various high-level tasks, termed as *Joint Training (Cross-Task)*.

Note that the weights in the high-level vision network are initialized from a well-trained network under the noiseless setting and not updated during training in our experiments.

Table IV and Table V list the performance of high-level vision tasks, i.e., top-1 and top5 accuracy for classification and mean intersection-over-union (IoU) without conditional random field (CRF) postprocessing for semantic segmentation. We notice that the baseline VGG approach obtains much lower accuracy than all the other cases, which shows the necessity of
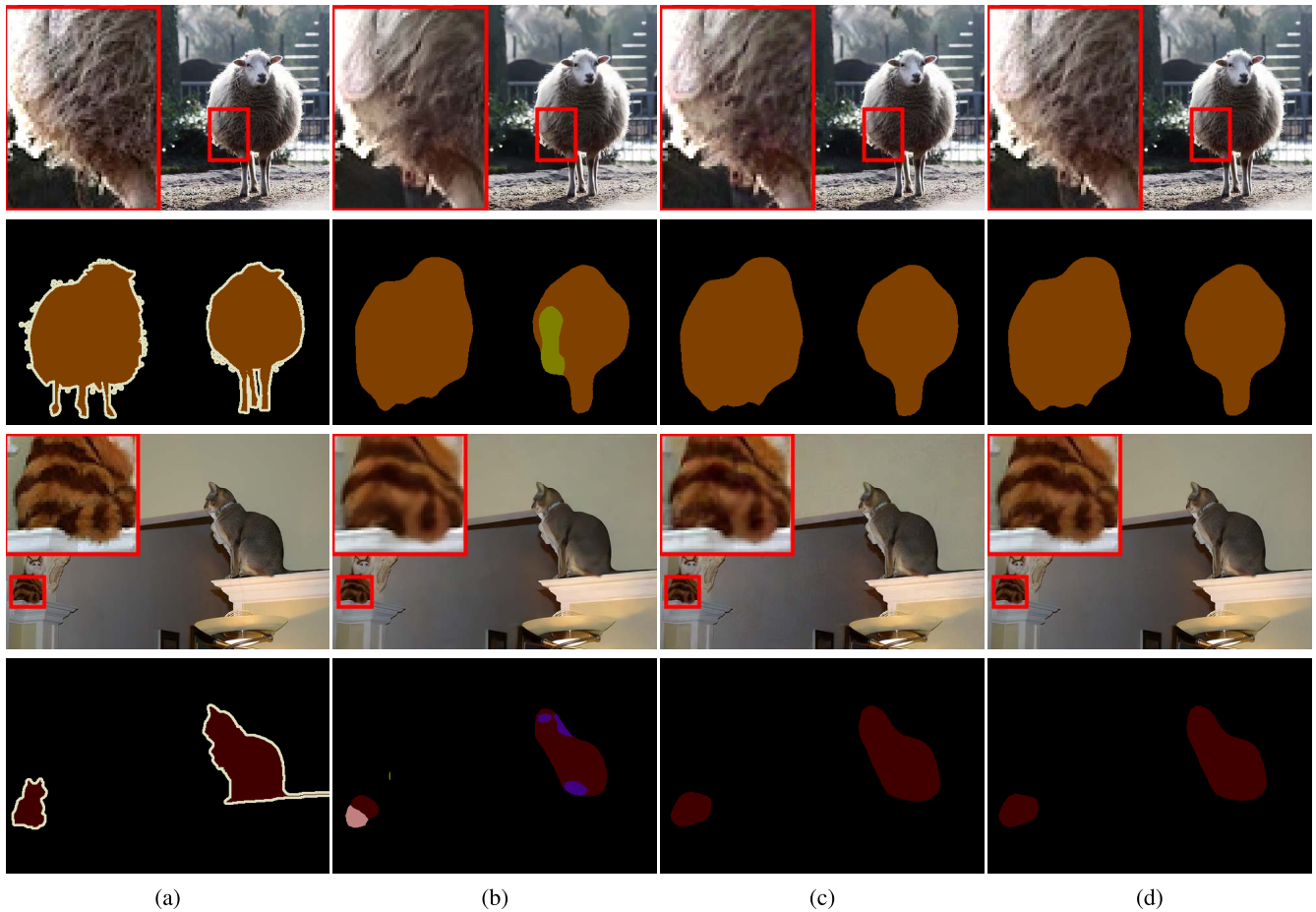
(a) (b) (c) (d)

Fig. 6. Two semantic segmentation examples from Pascal VOC 2012 validation set. From left to right: (a) the ground truth image, the denoised image using (b) the separately trained denoiser, (c) the denoiser trained with the reconstruction and segmentation joint loss, and (d) the denoiser trained with the classification network and evaluated for semantic segmentation. Their corresponding segmentation label maps are shown below. The zoom-in region which generates inaccurate segmentation in (b) is displayed in the red box.

TABLE IV

CLASSIFICATION ACCURACY AFTER DENOISING NOISY IMAGE INPUT, AVERAGED OVER ILSVRC2012 VALIDATION DATASET. RED IS THE BEST AND BLUE IS THE SECOND BEST RESULTS

		VGG	CBM3D + VGG	Separate + VGG	Joint Training	Joint Training (Cross-Task)
σ=15	Top-1	62.4	68.2	68.3	69.9	69.8
	Top-5	84.2	88.8	88.7	89.5	89.4
σ=30	Top-1	44.4	62.3	62.7	67.1	66.4
	Top-5	68.9	84.8	84.9	87.6	87.2
σ=45	Top-1	24.3	55.2	54.6	63.2	62.1
	Top-5	46.1	79.4	78.8	84.7	84.0
σ=60	Top-1	11.4	50.0	50.1	59.4	57.2
	Top-5	26.3	74.2	74.5	81.9	80.3

TABLE V

SEGMENTATION RESULTS (mIOU) AFTER DENOISING NOISY IMAGE INPUT, AVERAGED OVER PASCAL VOC 2012 VALIDATION DATASET. RED IS THE BEST AND BLUE IS THE SECOND BEST RESULTS

	VGG	CBM3D + VGG	Separate + VGG	Joint Training	Joint Training (Cross-Task)
σ=15	56.78	59.58	58.70	60.47	60.41
σ=30	43.43	55.29	54.13	57.88	56.30
σ=45	27.99	50.69	49.51	54.84	54.03
σ=60	14.94	46.56	46.59	52.05	51.84

image denoising as a preprocessing step for high-level vision tasks on noisy data. When we only apply denoising without considering high-level semantics (e.g., in CBM3D+VGG and Separate+VGG), it also fails to achieve high accuracy due to the artifacts introduced by the denoisers. The proposed Joint Training approach achieves sufficiently high accuracy across various noise levels.

As for the case of Joint Training (Cross-Task), first we train the denoising network jointly with the segmentation network and then connect this denoiser to the classification network. As shown in Table IV, its accuracy remarkably outperforms the cascade of a separately trained denoising network and

a classification network (i.e., Separate+VGG), and is comparable to our proposed model dedicatedly trained for classification (Joint Training). In addition, we use the denoising network jointly trained with the classification network to connect the segmentation network. Its mean IoU is much better than Separate+VGG in Table V. These two experiments show that the high-level semantics of different tasks are universal in terms of low-level vision tasks, which is in line with intuition, and the denoiser trained in our method has the generality for various high-level tasks.

Figure 6 displays two visual examples of how the data-driven denoising can enhance the semantic segmentation performance. It is observed that the segmentation result of the denoised image from the separately trained denoising network has lower accuracy compared to those using the joint loss
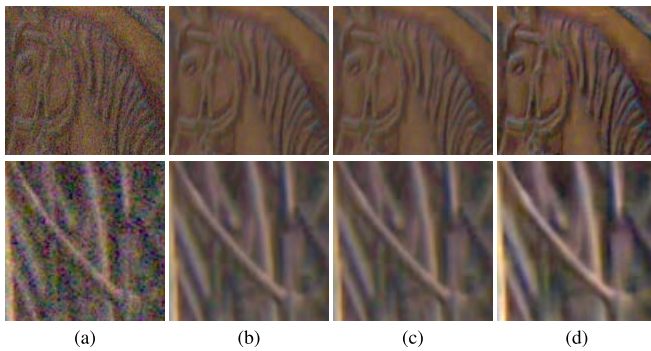
Fig. 7. Two denoising examples on real noisy images from DND. (a) The crop region of original images. (b) The denoising results from CBM3D. (c) The denoising results from DnCNN. (d) The denoising results from our method, i.e., the denoiser trained with the joint loss.

and the joint loss (cross-task), while the zoom-in region of its denoised image for inaccurate segmentation in Figure 6(b) contains oversmoothing artifacts. In contrast, both the Joint Training and Joint Training (Cross-Task) approaches achieve finer segmentation result and produce more visually pleasing denoised outputs simultaneously.

C. Extension to Real Noise Removal

Synthetic Gaussian noise removal has been extensively studied in previous works [1], [2], [8], [21]. However, it is unclear if denoisers can be generalized from removing synthetic noise to handling realistic noise, which is studied in [47]. Here, we further test our method on real noisy images in the Darmstadt Noise Dataset (DND) [48], where the images are taken with higher ISO, and the corruption can be more complicated than i.i.d. Gaussian noise [49]. Denoising examples using CBM3D, DnCNN and our proposed method are shown in Figure 7. It can be seen that CBM3D and DnCNN are prone to generate oversmoothing regions, while our proposed method preserves more local edges and high-frequency components, which leads to denoised images with better visual quality. This shows that our denoising method is robust to the noise statistical distribution, and thus demonstrates good generalization to realistic corruption.

V. Conclusion

Investigating the relation between low-level image processing and high-level semantic tasks has great practical value in various applications of computer vision. In this paper, we handle these two components in a simple yet efficient way by allowing the high-level semantic information to flow back to the low-level image processing part, which achieves superior performance in both image denoising and various high-level vision tasks. In addition, the denoiser trained for one high-level vision task in this manner has the robustness to other high-level vision tasks. Overall, it provides a feasible and robust solution in a deep learning fashion to real world problems, For future work, we will explore to embed the high-level semantics to more low-level vision tasks, e.g., super-resolution [50], [51], as well as to bring more types of high-level vision tasks [52]–[54] into consideration.

References

[1] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Process.*, vol. 54, no. 11, pp. 4311–4322, Nov. 2006.

[2] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, "Image denoising by sparse 3-D transform-domain collaborative filtering," *IEEE Trans. Image Process.*, vol. 16, no. 8, pp. 2080–2095, Aug. 2007.

[3] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, "Non-local sparse models for image restoration," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep. 2009, pp. 2272–2279.

[4] W. Dong, L. Zhang, G. Shi, and X. Li, "Nonlocally centralized sparse representation for image restoration," *IEEE Trans. Image Process.*, vol. 22, no. 4, pp. 1620–1630, Apr. 2013.

[5] S. Gu, L. Zhang, W. Zuo, and X. Feng, "Weighted nuclear norm minimization with application to image denoising," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 2862–2869.

[6] J. Xu, L. Zhang, W. Zuo, D. Zhang, and X. Feng, "Patch group based nonlocal self-similarity prior learning for image denoising," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 244–252.

[7] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proc. 25th Int. Conf. Mach. Learn. (ICML)*. New York, NY, USA: ACM, 2008, pp. 1096–1103.

[8] H. C. Burger, C. J. Schuler, and S. Harmeling, "Image denoising: Can plain neural networks compete with BM3D?" in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2392–2399.

[9] X. Mao, C. Shen, and Y.-B. Yang, "Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections," in *Proc. NIPS*, 2016, pp. 2802–2810.

[10] Y. Chen and T. Pock, "Trainable nonlinear reaction diffusion: A flexible framework for fast and effective image restoration," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1256–1272, Jun. 2017.

[11] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a Gaussian Denoiser: Residual learning of deep CNN for image denoising," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3142–3155, Jul. 2017.

[12] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. MICCAI*. Cham, Switzerland: Springer, 2015, pp. 234–241.

[13] Z. Wang, S. Chang, Y. Yang, D. Liu, and T. S. Huang, "Studying very low resolution recognition using deep networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4792–4800.

[14] J. Wu, R. Timofte, Z. Huang, and L. Van Gool, "On the relation between color image denoising and classification," Apr. 2017, *arXiv:1704.01372*. [Online]. Available: https://arxiv.org/abs/1704.01372

[15] D. Liu, B. Cheng, Z. Wang, H. Zhang, and T. S. Huang, "Enhance visual recognition under adverse conditions via deep networks," *IEEE Trans. Image Process.*, vol. 28, no. 9, pp. 4401–4412, Sep. 2019.

[16] C. Szegedy *et al.*, "Intriguing properties of neural networks," Dec. 2013, *arXiv:1312.6199*. [Online]. Available: https://arxiv.org/abs/1312.6199

[17] A. Nguyen, J. Yosinski, and J. Clune, "Deep neural networks are easily fooled: High confidence predictions for unrecognizable images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 427–436.

[18] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, "Color image denoising via sparse 3D collaborative filtering with grouping constraint in luminance-chrominance space," in *Proc. ICIP*, vol. 1, Sep./Oct. 2007, pp. I-313–I-316.

[19] D. Liu, B. Wen, X. Liu, Z. Wang, and T. Huang, "When image denoising meets high-level vision tasks: A deep learning approach," in *Proc. 27th Int. Joint Conf. Artif. Intell. (IJCAI)*, Jul. 2018.

[20] D. Zoran and Y. Weiss, "From learning models of natural image patches to whole image restoration," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 479–486.

[21] B. Wen, S. Ravishankar, and Y. Bresler, "Structured overcomplete sparsifying transform learning with convergence guarantees and applications," *Int. J. Comput. Vis.*, vol. 114, nos. 2–3, pp. 137–167, Sep. 2015.

[22] W. Dong, X. Li, L. Zhang, and G. Shi, "Sparsity-based image denoising via dictionary learning and structural clustering," in *Proc. CVPR*, Jun. 2011.

[23] J. Zhang, D. Zhao, and W. Gao, "Group-based sparse representation for image restoration," *IEEE Trans. Image Process.*, vol. 23, no. 8, pp. 3336–3351, Aug. 2014.

[24] F. Chen, L. Zhang, and H. Yu, "External patch prior guided internal clustering for image denoising," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 603–611.

[25] B. Wen, Y. Li, and Y. Bresler, "When sparsity meets low-rankness: Transform learning with non-local low-rank constraint for image restoration," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017.

[26] V. Jain and S. Seung, "Natural image denoising with convolutional networks," in *Proc. NIPS*, 2009, pp. 769–776.

[27] J. Xie, L. Xu, and E. Chen, "Image denoising and inpainting with deep neural networks," in *Proc. NIPS*, 2012, pp. 341–349.

[28] Y. Kim, H. Jung, D. Min, and K. Sohn, "Deeply aggregated alternating minimization for image restoration," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017.

[29] U. Schmidt and S. Roth, "Shrinkage fields for effective image restoration," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 2774–2781.

[30] K. Zhang, W. Zuo, S. Gu, and L. Zhang, "Learning deep CNN Denoiser prior for image restoration," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017.

[31] J. Jiao, W.-C. Tu, S. He, and R. W. H. Lau, "FormResNet: Formatted residual learning for image restoration," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 1034–1042.

[32] S. Lefkimmiatis, "Non-local color image denoising with convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3587–3596.

[33] S. Lefkimmiatis, "Universal denoising networks : A novel CNN architecture for image denoising," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3204–3213.

[34] K. Zhang, W. Zuo, and L. Zhang, "FFDNet: Toward a fast and flexible solution for CNN-based image denoising," *IEEE Trans. Image Process.*, vol. 27, no. 9, pp. 4608–4622, Sep. 2018.

[35] D. Liu, B. Wen, Y. Fan, C. C. Loy, and T. S. Huang, "Non-local recurrent network for image restoration," in *Proc. NIPS*, 2018, pp. 1673–1682.

[36] T. Remez, O. Litany, R. Giryes, and A. M. Bronstein, "Class-aware fully convolutional Gaussian and Poisson denoising," *IEEE Trans. Image Process.*, vol. 27, no. 11, pp. 5707–5722, Nov. 2018.

[37] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[38] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. ECCV*. Cham, Switzerland: Springer, 2016, pp. 694–711.

[39] L. Gatys, A. S. Ecker, and M. Bethge, "Texture synthesis using convolutional neural networks," in *Proc. NIPS*, 2015, pp. 262–270.

[40] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2414–2423.

[41] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *Proc. 8th IEEE Int. Conf. Comput. Vis. (ICCV)*, vol. 2, Nov. 2002, pp. 416–423.

[42] J. Xu, L. Zhang, D. Zhang, and X. Feng, "Multi-channel weighted nuclear norm minimization for real color image denoising," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, vol. 2, Oct. 2017, pp. 1096–1104.

[43] P. Wang *et al.*, "Understanding convolution for semantic segmentation," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2018, pp. 1451–1460.

[44] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," Sep. 2014, *arXiv:1409.1556*. [Online]. Available: https://arxiv.org/abs/1409.1556

[45] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected CRFs," Dec. 2014, *arXiv:1412.7062*. [Online]. Available: https://arxiv.org/abs/1412.7062

[46] Q. Huynh-Thu and M. Ghanbari, "Scope of validity of PSNR in image/video quality assessment," *Electron. Lett.*, vol. 44, no. 13, p. 800, 2008.

[47] S. Guo, Z. Yan, K. Zhang, W. Zuo, and L. Zhang, "Toward convolutional blind denoising of real photographs," in *Proc. CVPR*, 2019, pp. 1712–1722.

[48] T. Plotz and S. Roth, "Benchmarking denoising algorithms with real photographs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1586–1595.

[49] C. Liu, R. Szeliski, S. Bing Kang, C. Zitnick, and W. Freeman, "Automatic estimation and removal of noise from a single image," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 2, pp. 299–314, Feb. 2008.

[50] D. Liu, Z. Wang, B. Wen, J. Yang, W. Han, and T. S. Huang, "Robust single image super-resolution via deep networks with sparse prior," *IEEE Trans. Image Process.*, vol. 25, no. 7, pp. 3194–3207, Jul. 2016.

[51] D. Liu *et al.*, "Learning temporal dynamics for video super-resolution: A deep learning approach," *IEEE Trans. Image Process.*, vol. 27, no. 7, pp. 3432–3445, Jul. 2018.

[52] C. Yan *et al.*, "A fast Uyghur text detector for complex background images," *IEEE Trans. Multimedia*, vol. 20, no. 12, pp. 3389–3398, Dec. 2018.

[53] C. Yan, L. Li, C. Zhang, B. Liu, Y. Zhang, and Q. Dai, "Cross-modality bridging and knowledge transferring for image understanding," *IEEE Trans. Multimedia*, vol. 21, no. 10, pp. 2675–2685, Oct. 2019.

[54] D. Liu, Z. Zhao, X. Wang, Y. Hu, L. Zhang, and T. Huang, "Improving 3D human pose estimation via 3D part affinity fields," in *Proc. WACV*, 2019, pp. 1004–1013.
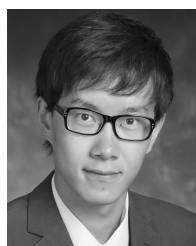
Ding Liu (Member, IEEE) received the Ph.D. degree from the University of Illinois at Urbana–Champaign, USA, in 2018. He is currently a Research Scientist with Bytedance Inc., Mountain View, CA, USA. His research experience encompasses using deep learning to solve low-level vision problems, including image restoration and image enhancement. He has research interests in the broad area of computer vision, image processing, and deep learning.
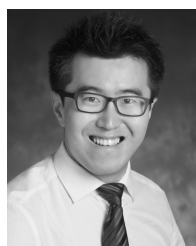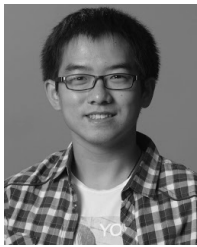
Bihan Wen (Member, IEEE) received the B.Eng. degree in electrical and electronic engineering from Nanyang Technological University, Singapore, in 2012, and the M.S. and Ph.D. degrees in electrical and computer engineering from the University of Illinois at Urbana–Champaign, USA, in 2015 and 2018, respectively. He is currently a Nanyang Assistant Professor with the School of Electrical and Electronic Engineering, Nanyang Technological University. His research interests span areas of machine learning, computational imaging, computer vision, image and video processing, and big data applications. He was a recipient of the 2016 Yee Fellowship and the 2012 Professional Engineers Board Gold Medal.
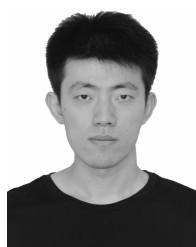
Jianbo Jiao (Member, IEEE) received the Ph.D. degree in computer science from the City University of Hong Kong in 2018. He was a Visiting Scholar with the Beckman Institute, University of Illinois at Urbana–Champaign from 2017 to 2018. He is currently a Post-Doctoral Researcher with the Department of Engineering Science, University of Oxford. His research interests include computer vision and machine learning.

Xianming Liu received the B.S. and M.S. degrees from the Harbin Institute of Technology in 2008 and 2010, respectively, and the Ph.D. degree from the Electrical and Computer Engineering Department, University of Illinois at Urbana–Champaign, under the supervision of Prof. T. S. Huang. He is currently a Senior Research Scientist with the Applied AI Lab, Horizon Robotics. His research interests include machine learning, computer vision, spatial computing, and embedded machine learning platforms. He was a recipient of the Kodak Fellowship 2015.

Zhangyang Wang (Member, IEEE) received the B.E. degree from the University of Science and Technology of China in 2012, and the Ph.D. degree from the Electrical and Computer Engineering Department, University of Illinois at Urbana–Champaign (UIUC) in 2016, under the supervision of Prof. T. S. Huang. He was a former Research Intern with the U.S. Army Research Lab in 2013, Adobe Research in 2014, and Microsoft Research in 2015. He is currently an Assistant Professor of computer science and engineering with Texas A&M University. He has coauthored more than 70 articles and has published several books and chapters. His research has been addressing machine learning, computer vision and multimedia signal processing problems, and their interdisciplinary applications using advanced feature learning and optimization techniques. He has been granted three patents and has received around 20 research awards and scholarships.

Thomas S. Huang (Life Fellow, IEEE) received the B.S. degree in electrical engineering from National Taiwan University, Taipei, Taiwan, and the M.S. and Sc.D. degrees in electrical engineering from the Massachusetts Institute of Technology (MIT), Cambridge. He was on the Faculty of the Department of Electrical Engineering, MIT, from 1963 to 1973, and on the Faculty of the School of Electrical Engineering and the Director of the Laboratory for Information and Signal Processing, Purdue University from 1973 to 1980. In 1980, he joined the University of Illinois at Urbana–Champaign, where he is currently a William L. Everitt Distinguished Professor of electrical and computer engineering, and a Research Professor with the Coordinated Science Laboratory, and with the Beckman Institute for Advanced Science. He is the Technology and Co-Chair of the Institute's major research theme Human–Computer Intelligent Interaction. His professional interests lie in the broad area of information technology, especially the transmission and processing of multidimensional signals. He has published 21 books and over 600 articles in network theory, digital filtering, image processing, and computer vision. He is a member of the National Academy of Engineering and the Academia Sinica, China, a Foreign Member of the Chinese Academies of Engineering and Sciences, and a fellow of the International Association of Pattern Recognition, and the Optical Society of America. He received the Honda Lifetime Achievement Award, the IEEE Jack Kilby Signal Processing Medal, and the King-Sun Fu Prize of the International Association for Pattern Recognition.