When Image Denoising Meets High-Level Vision Tasks: A Deep Learning Approach

Ding Liu[1], **Bihan Wen**[1], **Xianming Liu**[2], **Zhangyang Wang**[3], **Thomas S. Huang**[1] *
[1] University of Illinois at Urbana-Champaign, USA
[2] Facebook Inc.
[3] Texas A&M University, USA
{dingliu2,bwen3,t-huang1}@illinois.edu, xmliu@fb.com, atlaswang@tamu.edu

Abstract

Conventionally, image denoising and high-level vision tasks are handled separately in computer vision. In this paper, we cope with the two jointly and explore the mutual influence between them. First we propose a convolutional neural network for image denoising which achieves the state-of-the-art performance. 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 demonstrate 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 can generate more visually appealing results. To the best of our knowledge, this is the first work investigating the benefit of exploiting image semantics simultaneously for image denoising and high-level vision tasks via deep learning. The code is available online[1].

1 Introduction

A common approach in computer vision is to separate low-level vision problems, such as image restoration and enhancement, from high-level vision problems, and solve them independently. In this paper, we make their connection by showing the mutual influence between the two, i.e., visual perception and semantics, and propose a new perspective of solving both the low-level and high-level computer vision problems in a single unified framework, as shown in Fig. 1(a).

Image denoising, as one representative of low-level vision problems, is dedicated to recovering the underlying image signal from its noisy measurement. Classical image denoising methods take advantage of local or non-local structures presented in the image [Aharon *et al.*, 2006; Dabov *et al.*, 2007b; Mairal *et al.*, 2009; Dong *et al.*, 2013; Gu *et al.*, 2014; Xu *et al.*, 2015]. More recently, a number of deep learning models have been developed for image denoising which demonstrated superior performance [Vincent *et al.*, 2008; Burger *et al.*, 2012; Mao *et al.*, 2016; Chen and Pock, 2017; Zhang *et al.*, 2017a]. Inspired by U-Net [Ronneberger *et al.*, 2015], we propose a convolutional neural network for image denoising, which achieves the state-of-the-art performance.

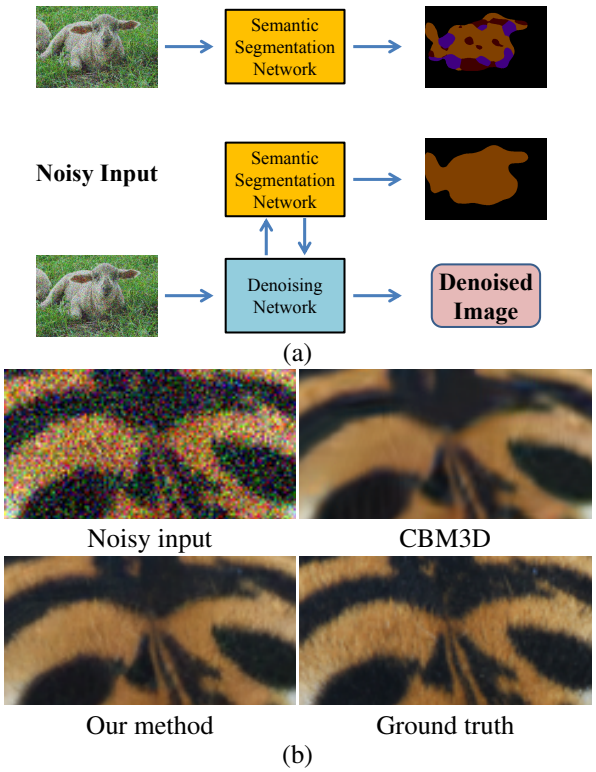


Figure 1: (a) Upper: conventional semantic segmentation pipeline; lower: our proposed framework for joint image denoising and semantic segmentation. (b) Zoom-in regions of a noisy input, its denoised estimates using CBM3D and our proposed method, as well as its ground truth.

While popular image denoising algorithms reconstruct images by minimizing the mean square error (MSE), important image details are usually lost which leads to image quality degradation, e.g., over-smoothing artifacts in some texture-rich regions are commonly observed in the denoised output

*Ding Liu and Thomas S. Huang's research work was supported by the U.S. Army Research Office under Grant W911NF-15-1-0317.
[1]https://github.com/Ding-Liu/DeepDenoising

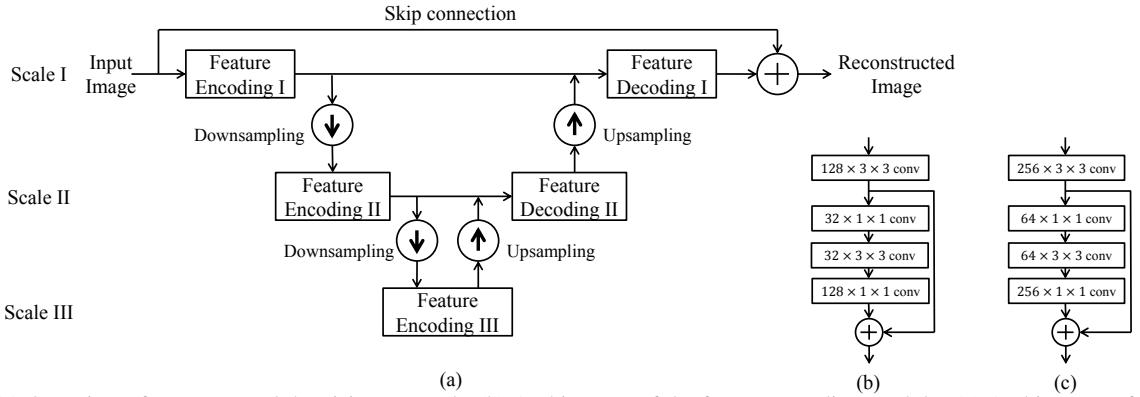


Figure 2: (a) Overview of our proposed denoising network. (b) Architecture of the feature encoding module. (c) Architecture of the feature decoding module.

from conventional methods, as shown in Fig. 1(b). To this end, we propose a cascade architecture connecting image denoising to a high-level vision network. We jointly minimize the image reconstruction loss 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 outputs, which demonstrates the importance of semantic information for image denoising.

When high-level vision tasks are conducted on noisy data, an independent image restoration step is typically applied as preprocessing, which is suboptimal for the ultimate goal [Wang *et al.*, 2016; Wu *et al.*, 2017; Liu *et al.*, 2017]. Recent research reveals that neural networks trained for image classification can be easily fooled by small noise perturbation or other artificial patterns [Szegedy *et al.*, 2013; Nguyen *et al.*, 2015]. Therefore, an application-driven denoiser should be capable of simultaneously removing noise and preserving semantic-aware details for the high-level vision tasks. Under the proposed architecture, we systematically investigate the mutual influence between the low-level and high-level vision networks. We show that the cascaded network trained with the joint loss not only boosts the denoising network performance 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 robust enough 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 finetuning either module, which facilitates the training effort when applied to various high-level tasks.

2 Method

We first introduce the denoising network utilized in our framework, and then explain the relationship between the image denoising module and the module for high-level vision tasks in detail.

2.1 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 tradeoff 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.

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 Fig. 2 (a). Each module in this network will be elaborated as follows.

Feature Encoding: We design one feature encoding module on each scale. which is one convolutional layer plus one residual block as in [He *et al.*, 2016]. The architecture is displayed in Fig. 2 (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 convolutioal layer is passed through a skip connection for element-wise sum with the output of the last convolutional layer.

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 of them obtain similar denosing 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 number of kernels in the four convolutional layers are 256, 64, 64 and 256. Its architecture is in Fig. 2(c).

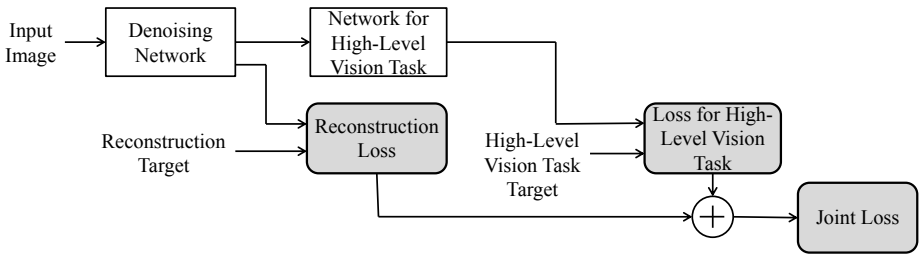


Figure 3: Overview of our proposed cascaded network.

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 the downsampling and upsampling factors, and try two schemes for downsampling in the experiments: (1) max pooling with stride of 2; (2) conducting convolutions with stride of 2. Both of them achieve similar denoisng performance in practice, so we use the second scheme in the rest 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.

2.2 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 denoisnig 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 Fig. 3. Specifically, given a noisy input image, the denosing 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.

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 [Johnson *et al.*, 2016]. The reason to adopt 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 other high-level tasks without finetuning 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.

Loss: The reconstruction loss of the denoising network is the mean squared error (MSE) between the denoising network output and the noiseless image. The losses of the classification network and the segmentation network 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 and the loss for high-level vision task, which can be represented as

$$L(F(x), y) = L_D(F_D(x), \tilde{x}) + \lambda L_H(F_H(F_D(x)), y), \quad (1)$$

where x is the noisy input image, $\tilde{x}$ is the noiseless image and y is the ground truth label of high-level vision task. F_D, F_H and F denote the denoising network, the network of high-level vision task and the whole cascaded network, respectively. L_D, L_H represent the losses of the denoising network and the high-level vision task network, respectively, while L is the joint loss, as illustrated in Fig. 3. λ is the weight for balancing the losses L_D and L_H.

3 Experiments

3.1 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 training set as in [Chen *et al.*, 2014]. The loss of training is equivalent to Eqn. 1 as $\lambda = 0$. 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.

We compare our denoisnig 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 widely used Kodak dataset[2], which consists of 24 color images. Table **??** shows the peak signal-to-noise ratio (PSNR) results for CBM3D [Dabov *et al.*, 2007a], TNRD [Chen and Pock, 2017], MCWNNM [Xu *et al.*, 2017], DnCNN [Zhang *et al.*, 2017a], and our proposed method. We do not list other methods [Burger *et al.*, 2012;

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

	$\sigma = 25$					$\sigma = 35$				$\sigma = 50$				
Image	**CBM3D**	**TNRD**	**MCWNNM**	**DnCNN**	**Proposed**	**CBM3D**	**MCWNNM**	**DnCNN**	**Proposed**	**CBM3D**	**TNRD**	**MCWNNM**	**DnCNN**	**Proposed**
01	29.13	27.21	28.66	29.75	**29.76**	27.31	26.93	28.10	**28.11**	25.86	24.46	25.28	26.52	**26.55**
02	32.44	31.44	31.92	32.97	**33.00**	31.07	30.62	31.65	**31.75**	29.84	29.12	29.27	30.44	**30.54**
03	34.54	32.73	34.05	34.97	**35.12**	32.62	32.27	33.37	**33.58**	31.34	29.95	30.52	31.76	**31.99**
04	32.67	31.16	32.42	32.94	**33.01**	31.02	30.92	31.51	**31.59**	29.92	28.65	29.37	30.12	**30.22**
05	29.73	27.81	29.37	30.53	**30.55**	27.61	27.53	28.66	**28.72**	25.92	24.37	25.60	26.77	**26.87**
06	30.59	28.52	30.18	31.05	**31.08**	28.78	28.44	29.37	**29.45**	27.34	25.62	26.70	27.74	**27.85**
07	33.66	31.90	33.36	34.42	**34.47**	31.64	31.53	32.60	**32.70**	29.99	28.24	29.51	30.67	**30.82**
08	29.88	27.38	29.39	30.30	**30.37**	27.82	27.67	28.53	**28.64**	26.23	23.93	25.86	26.65	**26.84**
09	34.06	32.21	33.42	34.59	**34.63**	32.28	31.76	33.06	**33.11**	30.86	28.78	30.00	31.42	**31.53**
10	33.82	31.91	33.23	34.33	**34.38**	31.97	31.51	32.74	**32.83**	30.48	28.78	29.63	31.03	**31.17**
11	31.25	29.51	30.62	31.82	**31.84**	29.53	29.04	30.23	**30.29**	28.00	26.75	27.41	28.67	**28.76**
12	33.76	32.17	33.02	34.12	**34.18**	32.24	31.52	32.73	**32.83**	30.98	29.70	30.00	31.32	**31.47**
13	27.64	25.52	27.19	**28.26**	28.24	25.70	25.40	26.46	**26.47**	24.03	22.54	23.70	24.73	**24.76**
14	30.03	28.50	29.67	30.79	**30.80**	28.24	28.05	29.17	**29.20**	26.74	25.67	26.43	27.57	**27.63**
15	33.08	31.62	32.69	33.32	**33.35**	31.47	31.15	31.89	**31.96**	30.32	29.07	29.59	30.50	**30.59**
16	32.33	30.36	31.79	32.69	**32.74**	30.64	30.15	31.16	**31.23**	29.36	27.82	28.53	29.68	**29.78**
17	32.93	31.20	32.39	**33.53**	33.50	30.64	30.75	31.96	**31.98**	29.36	28.07	28.98	30.33	**30.40**
18	29.83	28.00	29.46	30.40	**30.46**	28.00	27.70	28.72	**28.79**	26.41	25.06	25.94	27.03	**27.14**
19	31.78	30.01	31.29	**32.23**	32.30	30.19	29.86	30.80	**30.88**	29.06	27.30	28.44	29.34	**29.49**
20	33.45	32.00	32.78	34.15	**34.29**	31.84	31.32	32.73	**32.91**	30.51	29.24	29.79	31.28	**31.53**
21	30.99	29.09	30.55	31.61	**31.63**	29.17	28.86	29.94	**29.98**	27.61	26.09	27.13	28.27	**28.34**
22	30.93	29.60	30.48	**31.41**	31.38	29.36	28.93	29.94	**29.95**	28.09	27.14	27.47	28.54	**28.58**
23	34.79	33.68	34.45	35.36	**35.40**	33.09	32.79	33.86	**33.89**	31.75	30.53	30.96	32.18	**32.30**
24	30.09	28.17	29.93	**30.79**	30.77	28.19	28.17	28.98	**29.03**	26.62	24.92	26.37	27.18	**27.30**
Average	31.81	30.08	31.35	32.35	**32.39**	30.04	29.70	30.76	**30.83**	28.62	27.17	28.02	29.16	**29.27**

Table 1: Color image denoising results (PSNR) of different methods on Kodak dataset. The best result is shown in bold.

Zoran and Weiss, 2011; Gu *et al.*, 2014; Zhang *et al.*, 2017b] whose average performance is wore than DnCNN. The implementation codes used are from the authors' websites and the default parameter settings are adopted in our experiments.[3]

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 Kodak dataset.

3.2 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, 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 [Simonyan and Zisserman, 2014] is employed for image classification; we select *DeepLab-LargeFOV* in [Chen *et al.*, 2014] for semantic segmentation. We follow the preprocessing protocols (e.g. crop size, mean removal of each color channel) in [Simonyan and Zisserman, 2014] and [Chen *et al.*, 2014] 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 is empirically set as 0.25. 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 [Chen *et al.*, 2014], and test on its validation set. λ is empirically set as 0.5.

[3]For TNRD, we denoise each color channel using the grayscale image denoising implementation which is from the authors' website. TNRD for $\sigma = 35$ is not publicly available, so we do not include this case here.

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 [Huynh-Thu and Ghanbari, 2008]. 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.

Fig. 4 displays two image denoising examples from Kodak dataset. A visual comparison is illustrated for a zoom-in region: (II) and (III) are the denoising results using CBM3D [Dabov *et al.*, 2007a], and DnCNN [Zhang *et al.*, 2017a], respectively; (IV) is the proposed denoiser trained separately without the guidance of high-level vision information; (V) is the denoising result 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.

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

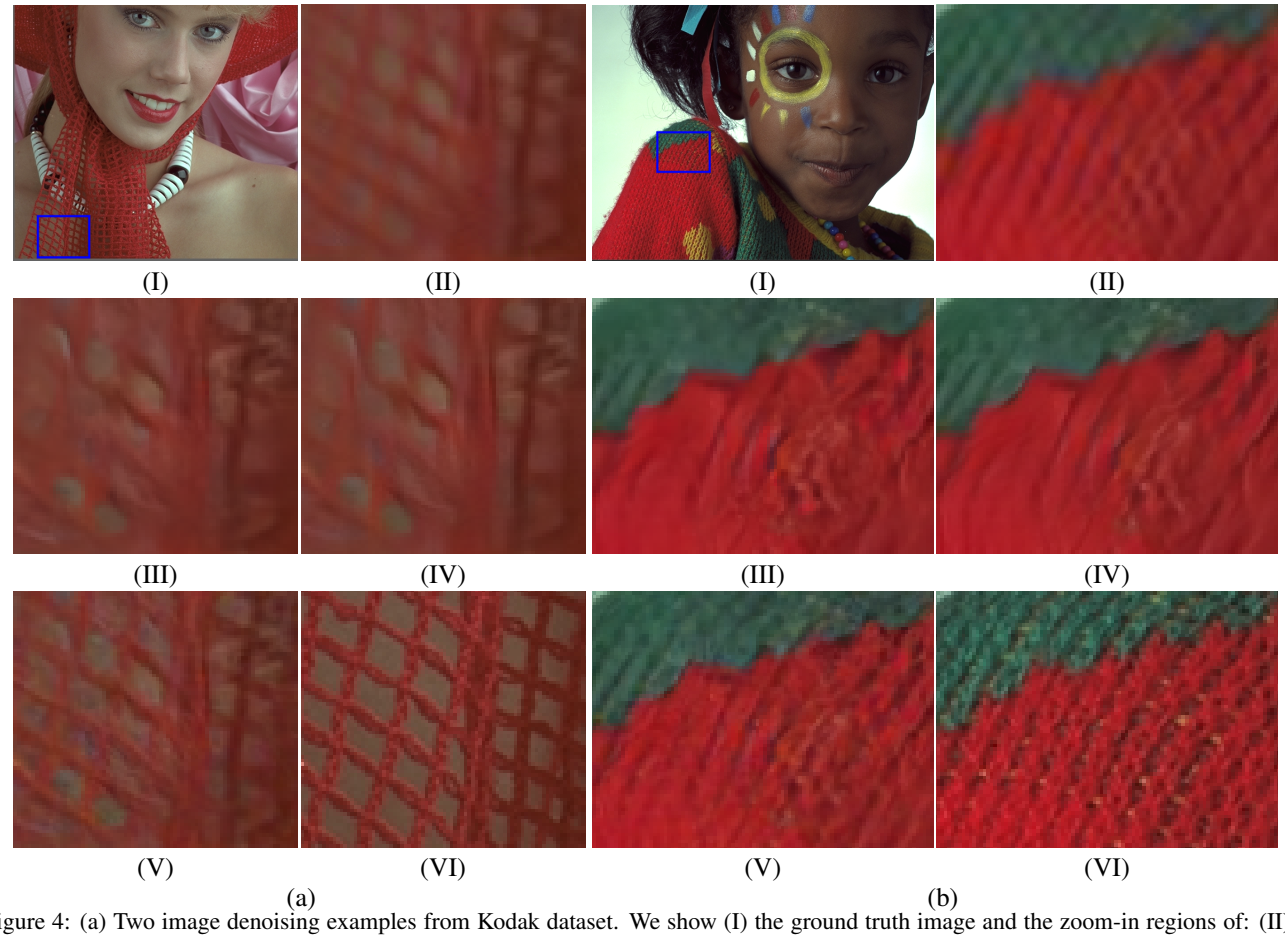


Figure 4: (a) Two image denoising examples from Kodak dataset. We show (I) the ground truth image and the zoom-in regions of: (II) the denoised image by CBM3D; (III) the denoised image by DnCNN; the denoising result of our proposed model (IV) without the guidance of high-level vision information; (V) with the guidance of high-level vision information and (VI) the ground truth.

		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	67.0	66.4
	Top-5	68.9	84.8	84.9	87.6	87.2
σ=45	Top-1	24.3	55.2	54.6	63.0	62.0
	Top-5	46.1	79.4	78.8	84.6	84.0
σ=60	Top-1	11.4	50.0	50.1	59.2	57.0
	Top-5	26.3	74.2	74.5	81.8	80.2

Table 2: Classification accuracy after denoising noisy image input, averaged over ILSVRC2012 validation dataset. Red is the best and blue is the second best results.

cases:

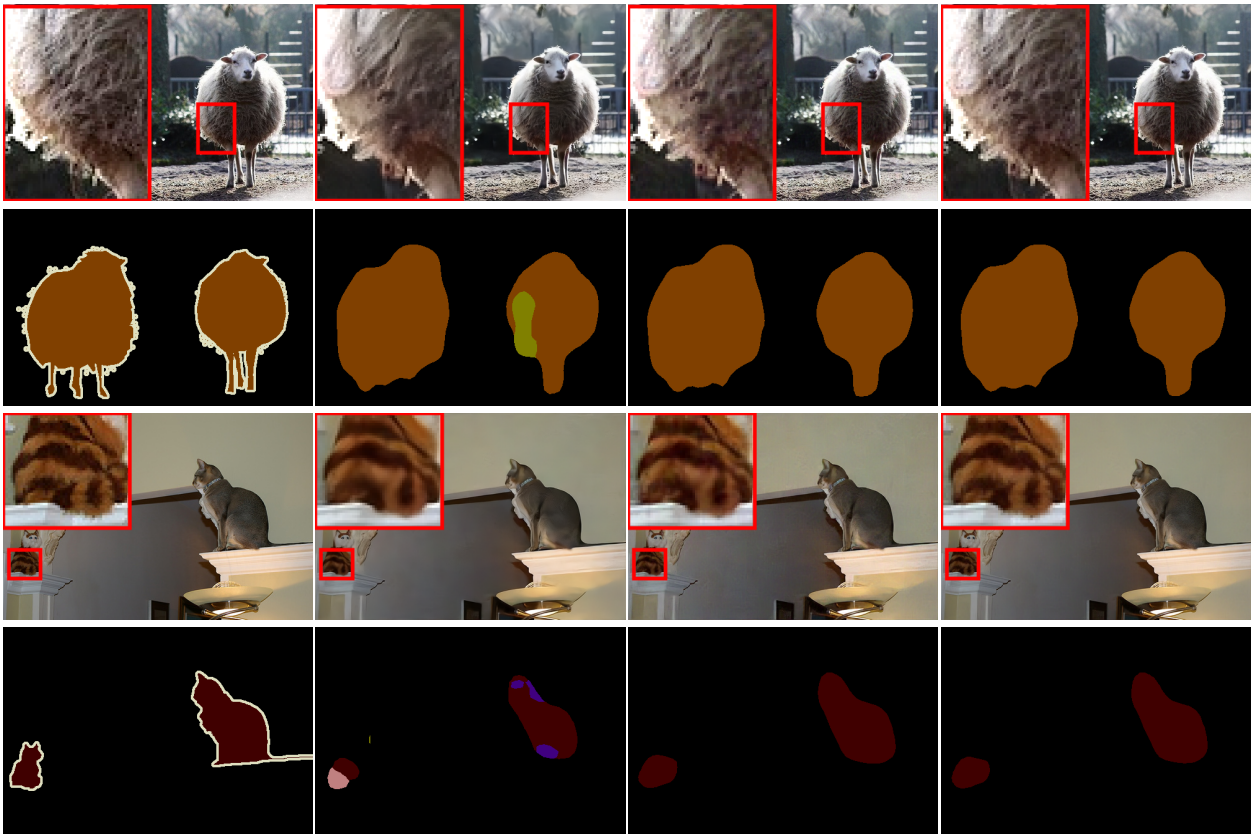
- noisy images are directly fed into the high-level vision network, termed as *VGG*. This approach serves as the baseline;
- noisy images are first denoised by CBM3D, and then fed into the high-level vision network, termed as *CBM3D+VGG*;
- noisy images are denoised via the separately trained denoising 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)*.

	VGG	CBM3D + VGG	Separate + VGG	Joint Training	Joint Training (Cross-Task)
σ=15	56.78	59.58	58.70	60.46	60.41
σ=30	43.43	55.29	54.13	57.86	56.29
σ=45	27.99	50.69	49.51	54.83	54.01
σ=60	14.94	46.56	46.59	52.02	51.82

Table 3: 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.

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 **??** and Table **??** 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



(a) (b) (c) (d)

Figure 5: 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.

much lower accuracy than all the other cases, which shows the necessity of 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 **??**, 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 **??**. These two experiments show 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.

Fig. 5 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 and the joint loss (cross-task), while the zoom-in region of its denoised image for inaccurate segmentation in Fig. 5 (b) contains oversmoothing artifacts. On the contrary, both the Joint Training and Joint Training (Cross-Task) approaches achieve finer segmentation result and produce more visually pleasing denoised outputs simultaneously.

4 Conclusion

Exploring the connection between low-level vision and high-level semantic tasks is of great practical value in various applications of computer vision. In this paper, we tackle this challenge in a simple yet efficient way by allowing the high-level semantic information flowing back to the low-level vision part, which achieves superior performance in both image denoising and various high-level vision tasks. In our method, the denoiser trained for one high-level task has the generality to other high-level vision tasks. Overall, it provides a feasible and robust solution in a deep learning fashion to real world problems, which can be used to handle other corruptions [Liu *et al.*, 2016; Li *et al.*, 2017].

References

[Aharon *et al.*, 2006] Michal Aharon, Michael Elad, and Alfred Bruckstein. K-SVD : An algorithm for designing overcomplete dictionaries for sparse representation. *TSP*, 54(11):4311–4322, 2006.

[Burger *et al.*, 2012] Harold C Burger, Christian J Schuler, and Stefan Harmeling. Image denoising: Can plain neural networks compete with bm3d? In *CVPR*, pages 2392–2399. IEEE, 2012.

[Chen and Pock, 2017] Yunjin Chen and Thomas Pock. Trainable nonlinear reaction diffusion: A flexible framework for fast and effective image restoration. *TPAMI*, 39(6):1256–1272, 2017.

[Chen *et al.*, 2014] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062*, 2014.

[Dabov *et al.*, 2007a] Kostadin Dabov, Alessandro Foi, Vladimir Katkovnik, and Karen Egiazarian. Color image denoising via sparse 3d collaborative filtering with grouping constraint in luminance-chrominance space. In *ICIP*, volume 1, pages I–313. IEEE, 2007.

[Dabov *et al.*, 2007b] Kostadin Dabov, Alessandro Foi, Vladimir Katkovnik, and Karen Egiazarian. Image denoising by sparse 3-d transform-domain collaborative filtering. *TIP*, 16(8):2080–2095, 2007.

[Dong *et al.*, 2013] Weisheng Dong, Lei Zhang, Guangming Shi, and Xin Li. Nonlocally centralized sparse representation for image restoration. *TIP*, 22(4):1620–1630, 2013.

[Gu *et al.*, 2014] Shuhang Gu, Lei Zhang, Wangmeng Zuo, and Xiangchu Feng. Weighted nuclear norm minimization with application to image denoising. In *CVPR*, pages 2862–2869, 2014.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[Huynh-Thu and Ghanbari, 2008] Quan Huynh-Thu and Mohammed Ghanbari. Scope of validity of psnr in image/video quality assessment. *Electronics letters*, 44(13):800–801, 2008.

[Johnson *et al.*, 2016] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, pages 694–711. Springer, 2016.

[Li *et al.*, 2017] Boyi Li, Xiulian Peng, Zhangyang Wang, Jizheng Xu, and Dan Feng. Aod-net: All-in-one dehazing network. In *ICCV*, 2017.

[Liu *et al.*, 2016] Ding Liu, Zhaowen Wang, Bihan Wen, Jianchao Yang, Wei Han, and Thomas S Huang. Robust single image super-resolution via deep networks with sparse prior. *IEEE TIP*, 25(7):3194–3207, 2016.

[Liu *et al.*, 2017] Ding Liu, Bowen Cheng, Zhangyang Wang, Haichao Zhang, and Thomas S Huang. Enhance visual recognition under adverse conditions via deep networks. *arXiv preprint arXiv:1712.07732*, 2017.

[Mairal *et al.*, 2009] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman. Non-local sparse models for image restoration. In *ICCV*, 2009.

[Mao *et al.*, 2016] Xiaojiao Mao, Chunhua Shen, and Yu-Bin Yang. Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections. In *NIPS*, pages 2802–2810. 2016.

[Nguyen *et al.*, 2015] Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *CVPR*, pages 427–436, 2015.

[Ronneberger *et al.*, 2015] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241. Springer, 2015.

[Simonyan and Zisserman, 2014] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[Szegedy *et al.*, 2013] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[Vincent *et al.*, 2008] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *ICML*, pages 1096–1103. ACM, 2008.

[Wang *et al.*, 2016] Zhangyang Wang, Shiyu Chang, Yingzhen Yang, Ding Liu, and Thomas S Huang. Studying very low resolution recognition using deep networks. In *CVPR*, pages 4792–4800, 2016.

[Wu *et al.*, 2017] Jiqing Wu, Radu Timofte, Zhiwu Huang, and Luc Van Gool. On the relation between color image denoising and classification. *arXiv preprint arXiv:1704.01372*, 2017.

[Xu *et al.*, 2015] Jun Xu, Lei Zhang, Wangmeng Zuo, David Zhang, and Xiangchu Feng. Patch group based nonlocal self-similarity prior learning for image denoising. In *CVPR*, pages 244–252, 2015.

[Xu *et al.*, 2017] Jun Xu, Lei Zhang, David Zhang, and Xiangchu Feng. Multi-channel weighted nuclear norm minimization for real color image denoising. 2017.

[Zhang *et al.*, 2017a] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *TIP*, 2017.

[Zhang *et al.*, 2017b] Kai Zhang, Wangmeng Zuo, Shuhang Gu, and Lei Zhang. Learning deep cnn denoiser prior for image restoration. In *CVPR*, July 2017.

[Zoran and Weiss, 2011] Daniel Zoran and Yair Weiss. From learning models of natural image patches to whole image restoration. In *ICCV*, pages 479–486. IEEE, 2011.