

Do Noises Bother Human and Neural Networks In the Same Way? A Medical Image Analysis Perspective

Shao-Cheng Wen¹, Yu-Jen Chen¹, Zihao Liu², Wujie Wen³, Xiaowei Xu⁴,
Yiyu Shi⁵, Tsung-Yi Ho¹, Qianjun Jia⁴, Meiping Huang⁴, Jian Zhuang⁴

¹*Department of Computer Science, National Tsing Hua University, Hsinchu, Taiwan*

²*Department of Electrical and Computer Engineering, Florida International University, FL, USA*

³*Department of Electrical and Computer Engineering, Lehigh University, PA, USA*

⁴*Guangdong Provincial People's Hospital, Guangdong Academic of Medical Science, Guangzhou, China*

⁵*Department of Computer Science and Engineering, University of Notre Dame, IN, USA*

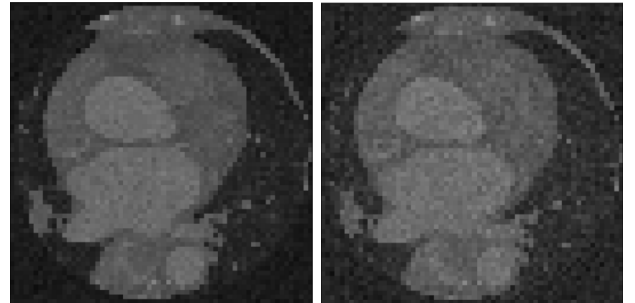
Abstract—Deep learning had already demonstrated its power in medical images, including denoising, classification, segmentation, etc. All these applications are proposed to automatically analyze medical images beforehand, which brings more information to radiologists during clinical assessment for accuracy improvement. Recently, many medical denoising methods had shown their significant artifact reduction result and noise removal both quantitatively and qualitatively. However, those existing methods are developed around human-vision, i.e., they are designed to minimize the noise effect that can be perceived by human eyes. In this paper, we introduce an application-guided denoising framework, which focuses on denoising for the following neural networks. In our experiments, we apply the proposed framework to different datasets, models, and use cases. Experimental results show that our proposed framework can achieve a better result than human-vision denoising network.

Index Terms—Denoising, Deep Learning

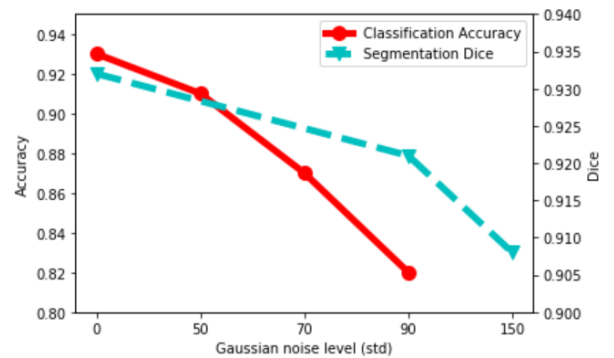
I. INTRODUCTION

THE prevalence of deep learning in medical image computing and analysis has greatly reduced the human effort and enhanced the efficiency of diagnosis and treatment [1], [2], [3], [4]. To achieve superb performance in such tasks, high-quality medical images are often indispensable for training and testing state-of-the-art deep learning models [5], [6]. Unfortunately, these raw images inevitably suffer from high-intensity noises (see Fig. 1 (a)) due to complex clinical scenarios [7], [8], [9], which significantly jeopardizing the capability of machine learning models on image segmentation and classification. As the example in Fig. 1 (b) shows, both image segmentation performance (Dice) and classification accuracy drop dramatically with the increase of noise level. Even neural network models are trained for better generalization by using images containing the same level of noise as that of testing ones, the performance can be decreased. In contrast, testing images in the medical domain usually can be much noisier than the training dataset. This further aggravates the accuracy problem for deep learning assisted medical imaging.

This research was approved by the Research Ethics Committee of Guangdong General Hospital, Guangdong Academy of Medical Science with the protocol No. 20140316.



(a) Noise-affected examples



(b) Accuracy/Dice vs Noise level

Fig. 1. (a) Demonstrates the noise-affected test images with $\mu = 0$, $\sigma = 90$ (left) and $\mu = 0$, $\sigma = 150$ (right), respectively. (b) segmentation and classification Dice/Accuracy w.r.t. Gaussian noise level. Note that we use the dirty Multi-Modality Whole Heart Segmentation (MM-WHS) dataset to train the segmentation model No-New-Net [10] and Classification Convolutional Neural Network (CCNN) [11]. Detailed experimental settings can be found in Section V.

To tackle this issue, image denoising is typically introduced as a pre-processing step before neural-network-based image classification or segmentation. Recently, most existing denoising methods [12], such as Residual Encoder-Decoder Convolution Neural Network (RED-CNN) [13] and Multi-Channel Denoising Convolution Neural Network (MCDnCNN) [14] utilized the power of deep neural network, which attempt to

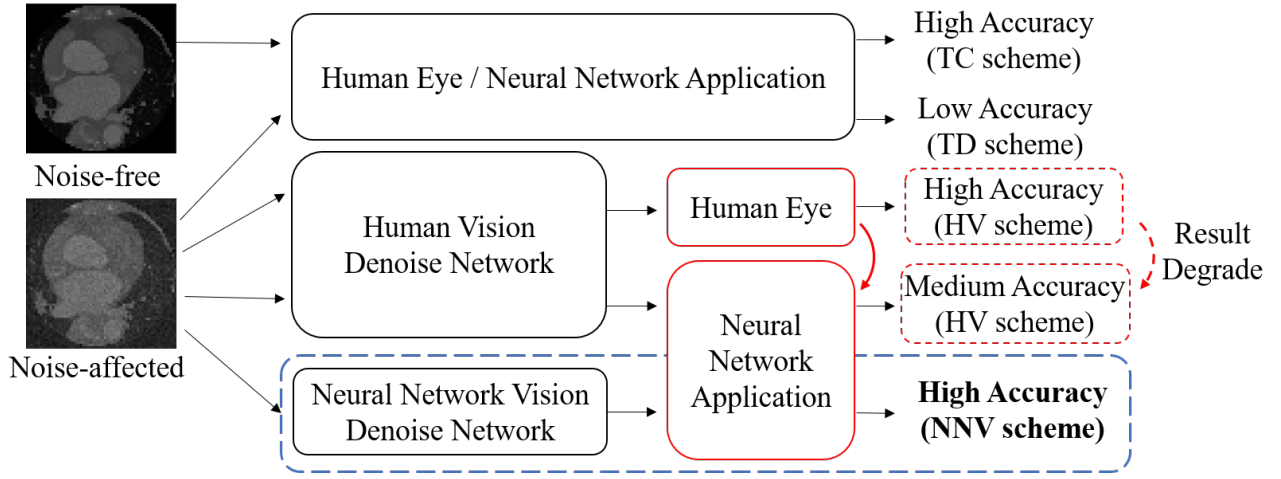


Fig. 2. Workflow comparison for input images with different noise level and different predict source (human and neural network application). The detail of all four schemes will be introduced in section V. We would like to point out that the result of the human-vision scheme may be degraded as the predict target changed (block marked with red). Note that the blue rectangle marked in the last row is the proposed workflow.

learn the distributions of the noise, so as to eliminate the noises in a more elaborate manner.

These denoising techniques could largely remove the noises based on the image visual quality measurements defined by human eyes, e.g., peak signal-to-noise ratio (PSNR) calculated by pixel-by-pixel difference between clean and its dirty version, and thereby enhance neural-network-based medical image segmentation or classification performance. We would like to argue that their advocated high denoising efficiency (dedicated to “human-vision”) may not be necessarily translated into impressive accuracy improvement for neural networks (or “neural-network-vision”). A clear workflows comparison for input images with different noise level and predict environment (human eye or neural network application) could be found in Fig. 2. From this figure, we would like to point out that the result of the human-vision scheme may be degraded as the predict environment changed.

In this paper, we propose to redefine the framework of medical image denoising by integrating the concept of “neural-network-vision”. Different from the human perceived visual distortion adopted by existing denoising solutions, the proposed framework evaluates the denoising efficiency directly through the perspective of neural network computation. As a result, such denoising can deal with the noises in a way that neural network favors [15], [16], so as to significantly boost the accuracy. We validate and compare our design with state-of-the-art denoising solutions, through comprehensive experiments on both image segmentation and classification tasks. Segmentation evaluation included two popular segmentation networks, No-New-Net [10] 2D and 3D version, under two different datasets, Multi-Modality Whole Heart Segmentation (MM-WHS) and Brain Tumor Segmentation challenge (BraTS). Classification evaluation was conducted through Classification Convolutional Neural Network (CCNN) [11] using Brain Tumor dataset. Examples demonstrated in Fig. 3 have shown the qualitative effect of our method on the segmentation results.

The main contributions of our work are as follows:

- A novel denoising framework guided by neural-network-vision is proposed.
- We proposed the very first application guided denoising network for image denoising by implementing the concept of “neural-network-vision”, and the denoising network can denoise images in a way desirable by any application network.
- Experimental results show that the proposed Neural-Network-Vision-based (NNV) image-denoising method outperforms any existing Human-Vision-based (HV) image-denoising methods in both segmentation and classification tasks.

II. MOTIVATION

In this section, we will demonstrate that human eyes and neural networks have very different understandings on image noises:

- 1) Keeping the noises that cannot be perceived by HV are unable to be effectively removed by current denoising methods due to HV-based quality judgement, may lead to considerable accuracy drop on deep-learning-based medical imaging.
- 2) Removing all noises that are obvious to human eyes via existing denoising methods according to human-visual rules, in contrast, may degrade the performance of deep-learning-based medical imaging.

The first argument has been proved by adversarial examples in deep learning security research—images containing very small and human-imperceptible noises can mislead the decision of deep learning models with high confidence [17].

The second argument can be roughly interpreted as follows: simply removing all noises perceivable by human eyes to make images completely noise-free, sometimes may result in inferior neural network decision making. Some noises could reinforce the information deemed to be important by neural networks for better image segmentation and recognition.

This can be clearly observed from Fig. 3, where Fig. 3 (b) is obtained by denoising dirty image (a) with RED-CNN

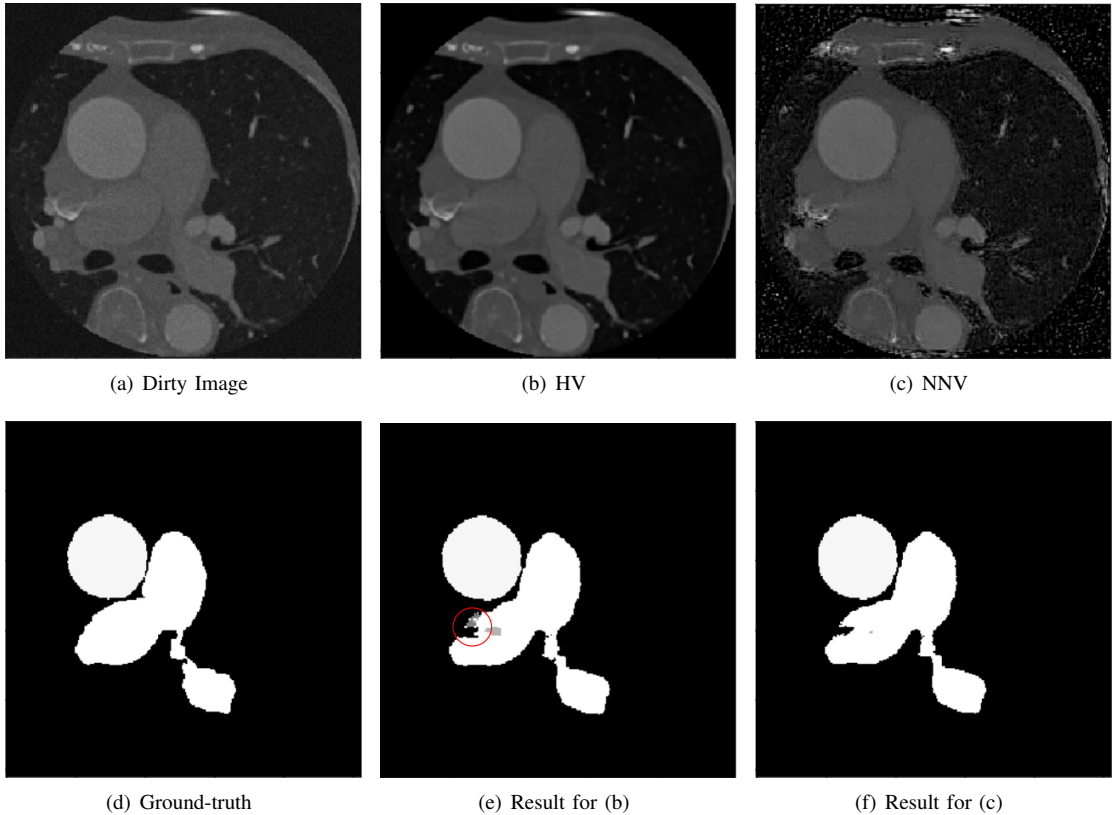


Fig. 3. Segmentation result comparison for (a) dirty image and the denoised image through (b) Human-Vision (HV) and (c) Neural-Network-Vision (NNV). The corresponding dice score are 1.0 and 0.704, and 0.849. As shown in (d), the ground-truth segment the image into two class, ascending aorta and the pulmonary artery. However, the region circled with red in (e) is the result of HV misclassified the pulmonary artery into the left atrium blood cavity, while (f) is correctly classified.

guided by an HV rule. Visually, Fig. 3 (b) has a much lower level of noise compared with Fig. 3 (c), which is denoised by an NNV manner tailored for deep learning by deliberately keeping some noises. Yet surprisingly, processing both denoised images with the same segmentation network No-New-Net, suggests that the noisier one, i.e., Fig. 3 (c) denoised by NNV, could achieve a much higher Dice score than the clean version, i.e., Fig. 3 (b) denoised by HV on image segmentation. The detailed segmentation result comparison is illustrated in Fig. 3 (e) and (f).

III. THE DIFFERENCE BETWEEN HUMAN-VISION AND NEURAL-NETWORK-VISION

Frequency results In order to know the difference between HV and NNV, we first transfer an image into the frequency domain by 8×8 2D Discrete Cosine Transform (DCT). In this way, the image is split into multiple small 8×8 frequency coefficients blocks. We then put the frequency coefficients belong to the same frequency components together to measure its distribution of this image (totally 64 frequency distributions). Since all the distributions obey normal distribution (i.e., mean is 0), thus the standard deviation (SD) indicates the energy in each frequency component of this image (i.e., large SD means more energy in this frequency component). Fig. 4 shows the heat map of SD at each frequency component, where (a) is

clean image, (b) indicates HV-based denoised image, and (c) represents the proposed NNV-based denoised image.

Obviously, the NNV-based denoised image (c) has more comprehensive information in high frequency domain compared with clean image (a).

This indicates how the segmentation network wants to change the denoised image, i.e., the additional information added in NNV-based denoised image is guided by segmentation network.

Frequency analysis Assume x_k is a single pixel of a raw image \mathbf{X} , and x_k can be represented by 8×8 DCT:

$$x_k = \sum_{i=0}^{i=7} \sum_{j=0}^{j=7} c_{(k,i,j)} \cdot b_{(i,j)} \quad (1)$$

where $c_{(k,i,j)}$ and $b_{(i,j)}$ are the DCT coefficient and corresponding basis function at 64 different frequencies, respectively.

Since the human visual system is less sensitive to high-frequency components, HV-based denoising is achieved by intentionally discarding the high-frequency parts $c_{(k,i,j)}$. On the contrary, Deep-Neural-Networks (DNN) examine the importance of the frequency information in a quite different way. The gradient of the DNN function F with respect to a basis function $b_{(i,j)}$ can be calculated as:

$$\frac{\partial F}{\partial b_{(i,j)}} = \frac{\partial F}{\partial x_k} \times \frac{\partial x_k}{\partial b_{i,j}} = \frac{\partial F}{\partial x_k} \times c_{(k,i,j)} \quad (2)$$

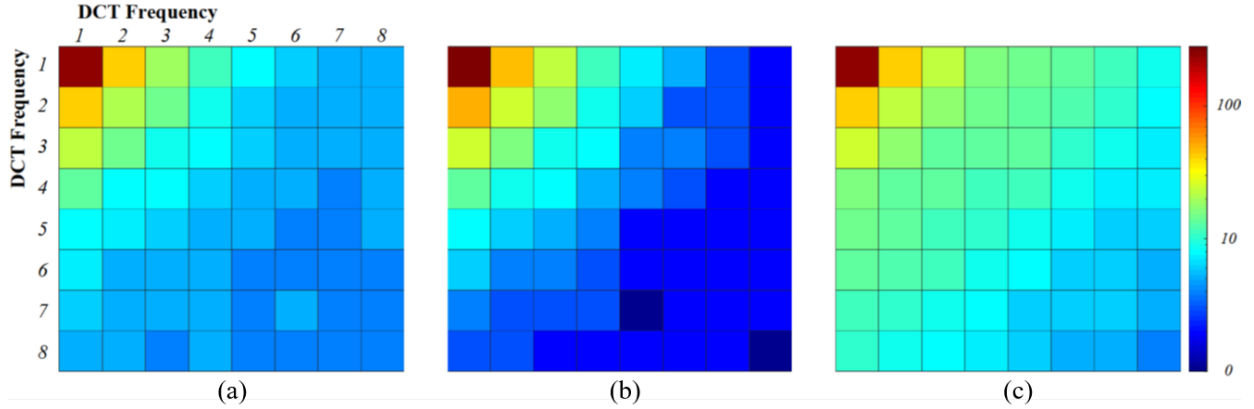


Fig. 4. DCT frequency distributions at different examples: (a) Test image; (b) HV-based denoising; (c) NNV-based denoising

Eq. 2 implies that the contribution of a frequency component $(b_{i,j})$ of a single pixel x_k to the DNN learning will be mainly determined by its associated DCT coefficient $(c_{(k,i,j)})$ and the importance of the pixel $(\frac{\partial F}{\partial x_k})$. Here $\frac{\partial F}{\partial x_k}$ is obtained after the DNN training, while $c_{(k,i,j)}$ will be distorted by filtering before training. If $c_{(k,i,j)} = 0$, the frequency feature $(b_{i,j})$, which may carry important details for DNN feature map extraction, cannot be learned by DNN for weights updating, causing a lower accuracy.

As shown in Fig 4 (a), clean image has comprehensive information in all frequency domains, however (b) HV-based method discard the high frequency information which will make DNN hard to learn these features. The NNV-based method can add more features in all frequency components to make the DNN easier to learning or training.

IV. NEURAL-NETWORK-VISION-BASED DENOISING

In this section, we would like to discuss about the concept for the proposed framework. First, we would like to define the denoising network as function $f : R^{m \times n} \rightarrow R^{m \times n}$ and the segmentation neural network as operation $g : R^{m \times n} \rightarrow R^k$, where m and n are the width and the height of the input image, respectively and k is the number of categories for each pixel.

For segmentation, the objective function of the training procedure which obtains the weights θ_g of model g can be written as:

$$\min_{\theta_g} \mathcal{L}_g(g(x; \theta_g), y) \quad (3)$$

where x and y are the input image and the ground-truth of the model, respectively. \mathcal{L}_g denotes the loss function which is minimized by optimizer.

On the other hand, the HV-based denoising function f is trained individually as:

$$\min_{\theta_f} \mathcal{L}_f(f(x; \theta_f), y) \quad (4)$$

where the weights θ_f is optimized by minimizing loss function \mathcal{L}_f .

The proposed NNV denoising framework is proved to have at least same power with application model itself. The weights θ of the NNV-based denoising network inside proposed framework f is minimized through loss function \mathcal{L}_g as:

$$\min_{\theta_f} \mathcal{L}_g(g(f(x; \theta_f); \theta_g), y) \quad (5)$$

While θ_f is given in worst case, the denoising model f could only learn to make $f(x; \theta_f) = x$, which means simply output the input. However, as substitute $f(x; \theta_f) = x$ into the framework $g(f(x; \theta_f))$, it would just make the framework as $g(x; \theta_g)$, which is equivalent with doing segmentation itself. Though this statement, we believe denoising network is required in our experiments.

V. EXPERIMENTS

A. Datasets

We use two segmentation datasets and one classification dataset to evaluate the proposed framework. Start with the first segmentation dataset, Multi-Modality Whole Heart Segmentation (MM-WHS) [18] dataset was acquired at Shanghai Shuguang Hospital, China, using routine cardiac CT angiography. All the image cover the whole heart from the upper abdominal to the aortic arch. The slices were acquired in the axial view. This dataset aims to accurately segment all the substructures of the whole heart into seven categories and background, as eight classes. In our experiment, we considered the dataset as clean images even though some of them still remain noises. We then synthesized the dirty dataset by superposing the noise to the clean image, which follows normal Gaussian distribution and Poisson distribution. Moreover, 2,557 images from 19 series are used as the training set, and the test set contains 363 images from 1 scanning series.

Second, we examined the experiments with Brain Tumor segmentation challenge (BraTS) [19] segmentation dataset. This dataset includes 210 High Grade Glioma (HGG) cases, which consist of a T1 weighted, a post-contrast T1-weighted, a T2-weighted, and a Fluid-Attenuated Inversion Recovery (FLAIR) MRI for each patient. We chose post-contrast T1-weighted images as our input in the experiment. Each tumor is segmented into edema, necrosis, and non-enhancing tumor, and active/enhancing tumor, which is in 5 categories (background + 4 classes). We split the dataset into the training set and test set with 1,125 images and 129 images, respectively.

For classification dataset, we utilized a brain tumor public dataset [20]. The objective of this dataset is to correctly classify the input into one of the three grades (Grade II, Grade III, and Grade IV). This dataset contains 233 patients with a total of 3,064 brain images with meningioma, glioma, and pituitary tumors, which is in three grades. We split them into training set with 2,500 images and test set with 300 images, not including in the training set. The images are T1-weighted contrast enhanced MRI images of axial (transverse plane), coronal (frontal plane), or sagittal (lateral plane) planes.

B. Experiment Schemes

In this paper, we compared our Neural-Network-Vision (NNV) based denoising scheme with three other schemes, segmentation or classification network Trained with Clean images (TC) and Trained with Dirty images (TD) which has the same noise level as test images, and the Human-Vision (HV) based denoising, respectively. To train all the four schemes, noises were added to datasets with different noise levels. For TC and TD schemes, they were trained using clean images and dirty images, respectively. For HV scheme, the denoising network was independently trained using paired clean images and dirty images and optimized using pixel-wise loss function, such as Mean Squared Error (MSE). Finally, the proposed framework, NNV scheme was trained with dirty images and optimized by the loss function which considered the difference between the output of the following neural networks and its ground-truth, such as cross-entropy loss.

C. Experiment Setup

In our paper, three application models were used, including a 3D segmentation model, a 2D segmentation model, and a 2D classification model. For 3D scenario, every experiment scheme involved required input volumes which were scaled into $64 \times 64 \times 64$. As for the experiments based on the 2D model, the input images were scaled to the size of 256×256 . While training, each scheme required different training epoch. We followed the default settings mentioned in the referred paper. For HV and MV schemes, the number of epoch was set to 300. Xavier uniform initializer [21] was used for all kernel in every convolution and deconvolution layer. The batch size was set to 1. Adam optimization [22] was used to minimize the loss function with learning rate at 0.00001. Moreover, all experiments done in our paper were implemented in Python3 with TensorFlow 1.14 over NVIDIA GeForce RTX 2080 Ti GPU.

D. Referenced Application Neural Network Model

In this section, we will briefly introduce the model we applied as the vehicle.

For denoising network, RED-CNN [13] and MCDnCNN [14] were used in baseline framework and the proposed framework. RED-CNN, known as Residual Encoder-Decoder Convolutional Neural Network, contains five convolution and deconvolution layers. With connection between encoder and decoder block, residual information could be carried to the

latter layer. In Chen et al. [13], it is proposed to reconstruct a denoised image from low-dose CT image. Multi-Channel Denoising Convolutional Neural Network (MCDnCNN) [14] is a modified version from DnCNN [23], which considers the neighboring slices for a better result. It is basically formed by eight convolution layers with batch normalization and a final convolution layer. Proposed by Jiang et al., it aims to denoise noises inside MRI images.

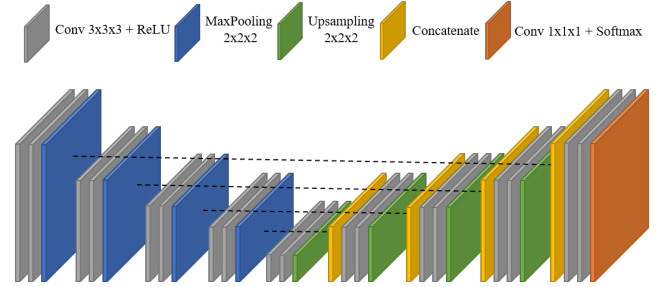


Fig. 5. The network structure of No-New-Net [10], which is a famous U-Net based architecture.

For segmentation, since U-Net based architecture is well-known for image segmentation, No-New-Net [10] 3D and 2D version were selected and implemented. Originally, No-New-Net was examined by using a brain tumor segmentation dataset, and results in rank two of brain tumor segmentation challenge (BraTS) 2018. Thus, we designed a similar model by replacing all 3D layers into 2D version.

At last, we implemented the Classification Convolutional Neural Network (CCNN), which follows the model mentioned in Sultan et al. [11]. The model is built based on basic convolutional neural network (CNN) with four max pooling layers. It is proposed to classify different grades of brain tumor through MRI image.

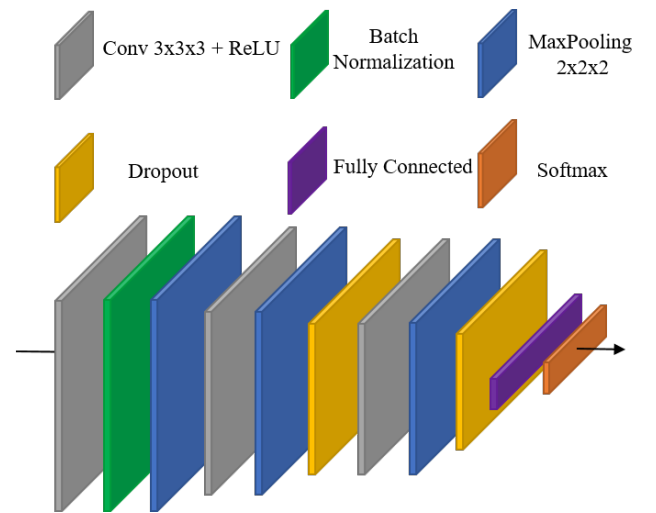


Fig. 6. The network structure of Classification Convolutional Neural Network (CCNN) [11].

E. Evaluation Metrics

For segmentation results, we follow existing works [10], [24], [25] applying Dice score and Hausdorff Distance d_H for evaluation. Dice score is influenced more with the overlap percentage between prediction and ground-truth. Besides, Hausdorff Distance calculates the largest distance between prediction and ground-truth boundary, which is influenced by the boundary distance.

The two metrics could be formulated as:

$$Dice = \frac{2|A \cap B|}{|A| + |B|} \quad (6)$$

$$d_H = \max \left\{ \sup_{a \in A} \inf_{b \in B} d(a, b), \sup_{b \in B} \inf_{a \in A} d(a, b) \right\} \quad (7)$$

where A and B denote two sets as ground-truth and segmentation prediction, respectively. In Eq. 7, sup and inf are supremum and infimum of sets, respectively. $d(\cdot)$ could be defined as any distance calculation function. In our experiment, we applied Euclidean distance in our experiment. Moreover, most existing works [26] show that in medical imaging, a Dice improvement over 0.01 is already significant when comparing with the same neural networks with different settings.

For classification, top-k accuracy is the most common metric for evaluation. In our paper, since only a few categories were desired to be classified, top-1 accuracy was applied and reported in Section VI.

VI. RESULTS

In this section, we will discuss the experimental results which are completed using two datasets, two noise types, and two denoising networks on segmentation and classification for the four experiment schemes, as TC, TD, HV, and NNV schemes mentioned in Section V. In order to show the flexibility of the proposed framework, we test all four trained schemes with different noise level included in the training set.

A. Results Analysis for Segmentation

We first evaluated how NNV can improve the segmentation accuracy over HV and segmentation network itself by using No-New-Net [10] 2D segmentation network. To show the feasibility, two datasets were applied to the experiment.

Table I reports the mean and the standard deviation (SD) of the test result using MM-WHS and BrasTS datasets, with two different noise added, Gaussian noise and Poisson noise for four experiment schemes. To show the flexibility of the proposed framework, two denoise methods, RED-CNN and MCDnCNN were implemented to both HV and NNV schemes. For a fair comparison, both schemes in each denoising method were trained with the same hyperparameters and settings.

We start our discussion on No-New-Net 2D segmentation. For MM-WHS dataset, first of all, TD scheme achieved 0.134 better Dice than TC scheme on average. This is as expected since the network could learn the feature extracted from noisier images. Secondary, since denoising method was included, we believe that HV denoise network is still somehow effective.

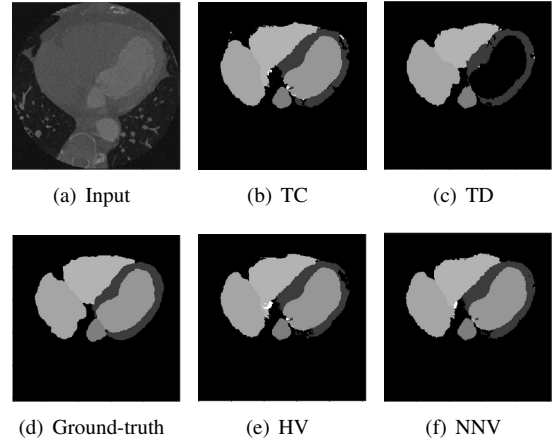


Fig. 7. Segmentation comparison for (a) input image from MM-WHS dataset. (b-c) are the segmentation results of TC and TD schemes, which denoising network is not included. (d) is the ground-truth and (e-f) are the segmentation results of HV and NNV schemes using RED-CNN denoising network. Note that Gaussian white noise $\sigma = 70$ is added to simulate dirty images in this case.

Thus, for both denoising methods, RED-CNN and MCDnCNN, HV scheme improved the Dice and the Hausdorff over both schemes without denoising network by 0.103 and 0.107, 0.77 and 0.85, respectively. Finally, NNV scheme achieved another improvement over HV scheme by 0.019 higher Dice and 0.054 lower Hausdorff distance for RED-CNN and 0.018 and 0.069 for MCDnCNN. Fig. 7 shows the input, ground-truth, and the segmentation results of four schemes.

As for the comparison with Poisson noise added in MM-WHS dataset, we can notice that TD scheme achieves higher Dice than TC scheme with 0.062 improvement. Compared with the TD scheme, the proposed NNV scheme achieves the optimal performance in both metrics which yields an improvement of 0.126 higher Dice and 0.895 lower Hausdorff distance for RED-CNN and 0.117 and 0.835 for MCDnCNN.

We also applied the same experiment to BraTS dataset. The statistic result of the experiment is also reported in Table I. We can notice that the improvement trend of all the schemes on MM-WHS dataset is the same as that on BraTS dataset. As for Gaussian noise experiment, for those schemes without denoising, TD scheme outperforms TC scheme with Dice 0.016 on average. However, as the denoising network is included, HV scheme beats TD scheme with 0.08 and 0.204 on Dice score and Hausdorff distance, respectively. Finally, NNV scheme successfully results in the highest Dice 0.585 and the lowest Hausdorff distance 2.340 on average among all four schemes. Poisson noise experiment also has a similar improvement trend compared with Gaussian noise experiment. TD scheme again improves the Dice score with 0.063 than TC scheme. HV scheme beat TD scheme with slightly improvement. And finally the proposed NNV scheme lead the optimal score in both Dice score and Hausdorff distance with 0.01 and 0.067, respectively for RED-CNN and 0.004 and 0.045, respectively for MCDnCNN.

We further applied the experiment to No-New-Net 3D version, the comparison are reported in Table II. In this experiment, Gaussian white noise with $\sigma = 90$ was superposed

TABLE I

STATISTIC RESULT COMPARISON (MEAN \pm SD) FOR NO-NEW-NET 2D (256 \times 256) MODEL USING MM-WHS AND BRATS DATASETS. RED-CNN AND MCDnCNN DENOISING NETWORKS WERE USED. GAUSSIAN WHITE NOISE $\sigma = 70$ WAS ADDED TO SIMULATE DIRTY IMAGES.

Schemes		Gaussian Noise				Poisson Noise			
		MM-WHS dataset		BraTS dataset		MM-WHS dataset		BraTS dataset	
		Dice	Hausdorff	Dice	Hausdorff	Dice	Hausdorff	Dice	Hausdorff
w/o Denoise	TC	0.542 \pm 0.251	2.723 \pm 0.976	0.481 \pm 0.184	3.047 \pm 0.837	0.641 \pm 0.217	2.655 \pm 1.725	0.525 \pm 0.186	2.502 \pm 1.055
	TD	0.676 \pm 0.257	2.815 \pm 2.116	0.497 \pm 0.203	2.570 \pm 0.760	0.703 \pm 0.160	2.685 \pm 1.538	0.588 \pm 0.168	2.334 \pm 0.732
RED-CNN	HV	0.779 \pm 0.177	1.953 \pm 1.373	0.577 \pm 0.169	2.366 \pm 0.727	0.779 \pm 0.182	1.994 \pm 1.398	0.590 \pm 0.169	2.320 \pm 0.721
	NNV	0.798\pm0.167	1.899\pm1.333	0.585\pm0.164	2.340\pm0.781	0.829\pm0.150	1.790\pm1.227	0.602\pm0.163	2.253\pm0.740
MCDnCNN	HV	0.783 \pm 0.176	1.965 \pm 1.392	0.575 \pm 0.172	2.416 \pm 0.745	0.757 \pm 0.190	2.067 \pm 1.419	0.596 \pm 0.161	2.326 \pm 0.699
	NNV	0.802\pm0.171	1.896\pm1.340	0.585\pm0.163	2.341\pm0.774	0.820\pm0.155	1.850\pm1.293	0.600\pm0.169	2.281\pm0.770

TABLE II

STATISTIC RESULT COMPARISON (MEAN \pm SD) FOR NO-NEW-NET 3D (64 \times 64 \times 64) MODEL USING MM-WHS DATASET. BOTH TRAINING AND TEST SET CONTAINS GAUSSIAN WHITE NOISE WITH $\sigma = 90$.

Schemes		Dice	Sensitivity	Specificity
w/o Denoise	TC	0.817 \pm 0.089	0.817 \pm 0.093	0.997 \pm 0.001
	TD	0.827 \pm 0.065	0.807 \pm 0.083	0.998 \pm 0.000
RED-CNN	HV	0.830 \pm 0.060	0.820 \pm 0.065	0.998 \pm 0.000
	NNV	0.840\pm0.053	0.825\pm0.061	0.998 \pm 0.000
MCDnCNN	HV	0.837 \pm 0.058	0.825 \pm 0.064	0.998 \pm 0.000
	NNV	0.841\pm0.054	0.826\pm0.062	0.998 \pm 0.000

to the MM-WHS dataset. From the table, similar to 2D mmwhs segmentation experiment, TD scheme again achieves 0.01 higher Dice over the TC scheme. With denoising network included, HV and MV schemes outperform TD scheme up to 0.014 Dice and 0.019 sensitivity on average. Furthermore, compared with HV scheme, MV scheme achieves slightly higher Dice and sensitivity at 0.01 and 0.005, respectively for RED-CNN, and 0.004 and 0.001, respectively for MCDnCNN. However, since the size of input volume is only 64 \times 64 \times 64, all the four schemes had originally achieved over 0.8 Dice score. Thus, the improvement, though smaller than that in the No-New-Net 2D segmentation presented in the paper, is still significant. Moreover, it can be clearly seen that all the experiments result in high specificity, which means that most true negative cases can be correctly segmented.

B. Results Analysis for Classification

TABLE III

CLASSIFICATION ACCURACY COMPARISON FOR CLASSIFICATION CONVOLUTIONAL NEURAL NETWORK (CCNN), USING BRAIN TUMOR DATASET WITH THREE DIFFERENT NOISE LEVELS $\sigma = 50, 70, 90$ GAUSSIAN WHITE NOISE ADDED. NOTE THAT TRAIN SET AND TEST SET CONTAIN SAME NOISE LEVEL IN THIS EXPERIMENT.

Cases	w/o Denoise		RED-CNN	
	TC	TD	HV	NNV
Gaussian white noise $\sigma = 50$	0.370	0.936	0.946	0.960
Gaussian white noise $\sigma = 70$	0.350	0.923	0.933	0.940
Gaussian white noise $\sigma = 90$	0.343	0.890	0.926	0.936

In this section, we also extend the evaluation to classification using CCNN proposed by Saltan et al. [11]. Similar to previous section, we explored how the NNV-based denoising framework works among different noise levels. Thus, three datasets were synthesized with $\sigma = 50, 70, 90$ Gaussian white noise superposed to Brain Tumor dataset. Note that both training set and test set contain the same noise level.

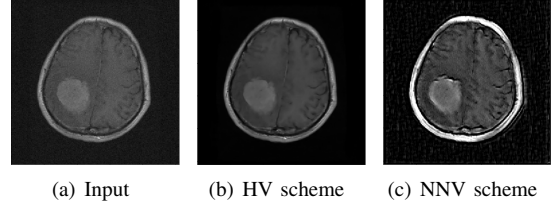


Fig. 8. Classification example for the (a) dirty input and the corresponded denoised image using (b) HV and (c) NNV experiment scheme from brain tumor dataset. The HV scheme misclassified the glioma into Grade III, which NNV scheme correctly classified into the Grade II.

From Table III, we can again observe that in both TC and TD schemes, the higher level the noise is, the lower accuracy does the classification network results in. Obviously, TD scheme outperformed TC scheme in the experiment, which had up to 2.6 \times higher accuracy. However, when it comes to comparison between HV and NNV schemes, all numbers outperformed schemes without denoising network. That is, denoising network is required for testing on dirty images. Based on the observation between HV and NNV scheme, NNV scheme successfully improved the accuracy up to 0.140 in all three cases.

Furthermore, we also show a visualize example for classification in Fig. 8. In this case, glioma in Fig. 8 (b) HV scheme was misclassified into Grade III. However, Fig. 8 (c) NNV scheme successfully matched the correct class as Grade II. Since the confidence for Grade III and Grade II in this case is relatively close, as we sharpen the contour of the glioma become sharper in Fig 8 (c), the classification model could classify the grade correctly.

C. Results for Different Noise Levels in Training Set and Test Set

In this section, we test all the four trained schemes with different noise levels from the training set to show the effectiveness of our framework.

Firstly, we trained the four schemes for No-New-Net 2D segmentation network using MM-WHS and BraTS datasets with $\sigma = 70$ Gaussian white noise added. Table IV shows the test results for test set with $\sigma = 50$ and $\sigma = 150$ Gaussian white noise superposed for MM-WHS and BraTS datasets, respectively. For both datasets, our expectation still holds. The two schemes with denoising network included outperformed the other two without denoising up to 0.037 higher Dice and

TABLE IV

RESULTS (MEAN \pm SD) FOR DIFFERENT NOISE LEVEL OCCURRED IN TRAINING SET AND TEST SET. NOTE THAT THE EXPERIMENT WAS TRAINED USING NO-NEW-NET 2D SEGMENTATION ON MM-WHS AND BRA-TS DATASET, RESPECTIVELY WITH $\sigma = 70$ GAUSSIAN WHITE NOISE ADDED.

Dataset	Noise Level		w/o Denoise		RED-CNN	
			TC	TD	HV	NNV
MM-WHS dataset	$\sigma = 50$	Dice	0.574 \pm 0.247	0.684 \pm 0.259	0.717 \pm 0.191	0.721\pm0.192
		Hausdorff	2.583 \pm 1.005	2.789 \pm 2.127	2.168 \pm 1.456	2.132\pm1.398
	$\sigma = 150$	Dice	0.396 \pm 0.229	0.605\pm0.209	0.455 \pm 0.243	0.488 \pm 0.231
		Hausdorff	3.411 \pm 0.966	3.150 \pm 1.950	3.076 \pm 0.998	3.011\pm0.989
BraTS dataset	$\sigma = 50$	Dice	0.529 \pm 0.192	0.521 \pm 0.205	0.565 \pm 0.174	0.573\pm0.177
		Hausdorff	2.605 \pm 0.653	2.461 \pm 0.804	2.420 \pm 0.683	2.352\pm0.808
	$\sigma = 150$	Dice	0.437 \pm 0.178	0.504 \pm 0.180	0.580 \pm 0.160	0.584\pm0.170
		Hausdorff	3.406 \pm 0.920	2.492 \pm 0.749	2.405 \pm 0.707	2.333\pm0.778

TABLE V

RESULTS FOR DIFFERENT NOISE LEVEL OCCURRED IN TRAINING SET AND TEST SET. NOTE THAT ALL EXPERIMENT IS TRAINED FOR CLASSIFICATION USING BRAIN TUMOR DATASET WITH $\mu = 0$, $\sigma = 70$ GAUSSIAN NOISE ADDED.

Cases	w/o Denoise		RED-CNN	
	TC	TD	HV	NNV
Gaussian white noise $\sigma = 50$	0.370	0.860	0.953	0.973
Gaussian white noise $\sigma = 90$	0.343	0.906	0.920	0.923

0.657 Hausdorff distance for MM-WHS dataset. For BraTS dataset, similar result obtained. NNV scheme led the best Dice 0.584 and Hausdorff distance 2.333 among all four experiment schemes.

Second, we brought our experiment to classification. In Table V, we applied the four schemes, which were trained with images containing Gaussian white noise with $\sigma = 70$ to test set with Gaussian white noise $\sigma = 50$ and $\sigma = 90$, respectively. Obviously, all the four schemes performed as our expectation. NNV scheme outperformed the other three schemes in all the three cases, where HV, TD, and TC schemes had 0.02, 0.11, and 0.6 higher accuracy, respectively for Gaussian white noise with $\sigma = 50$ and 0.003, 0.017, and 0.58, respectively for Gaussian white noise with $\sigma = 90$.

VII. CONCLUSION

In this paper, due to the observation that neural network applications focus on different sight from human eyes, we introduced a neural-network-vision-based denoising framework. Unlike previous human-vision-based denoising methods, our framework could perform a better result for neural network application. By evaluating the experiment through different networks, noise types, and datasets on segmentation and classification, experimental results have shown the effectiveness and the feasibility of the proposed framework.

REFERENCES

- [1] X. Xu, Y. Ding, S. X. Hu, M. Niemier, J. Cong, Y. Hu, and Y. Shi, "Scaling for edge inference of deep neural networks," *Nature Electronics*, vol. 1, no. 4, pp. 216–222, 2018.
- [2] Y. Ding, J. Liu, X. Xu, M. Huang, J. Zhuang, J. Xiong, and Y. Shi, "Uncertainty-aware training of neural networks for selective medical image segmentation," in *Medical Imaging with Deep Learning*, 2020.
- [3] X. Xu, Q. Lu, L. Yang, S. Hu, D. Chen, Y. Hu, and Y. Shi, "Quantization of fully convolutional networks for accurate biomedical image segmentation," in *CVPR*, 2018, pp. 8300–8308.
- [4] T. Wang, X. Xu, J. Xiong, Q. Jia, H. Yuan, M. Huang, J. Zhuang, and Y. Shi, "Ica-unet: Ica inspired statistical unet for real-time 3d cardiac cine mri segmentation," in *MICCAI*, 2020, pp. 447–457.
- [5] X. Xu, T. Wang, Y. Shi, H. Yuan, Q. Jia, M. Huang, and J. Zhuang, "Whole heart and great vessel segmentation in congenital heart disease using deep neural networks and graph matching," in *MICCAI*, 2019, pp. 477–485.
- [6] X. Xu, T. Wang, J. Zhuang, H. Yuan, M. Huang, J. Cen, Q. Jia, Y. Dong, and Y. Shi, "Imagechd: A 3d computed tomography image dataset for classification of congenital heart disease," in *MICCAI*, 2020, pp. 77–87.
- [7] F. E. Boas, and D. Fleischmann, "Ct artifacts: causes and reduction techniques," *Imaging Med*, vol. 4, no. 2, pp. 229–240, 2012.
- [8] K. Krupa, and M. Bekiesińska-Figatowska, "Artifacts in magnetic resonance imaging," *Polish journal of radiology*, vol. 80, p. 93, 2015.
- [9] J. Liu, Y. Ding, J. Xiong, Q. Jia, M. Huang, J. Zhuang, B. Xie, C.-C. Liu, and Y. Shi, "Multi-cycle-consistent adversarial networks for ct image denoising," in *ISBI*, 2020, pp. 614–618.
- [10] F. Isensee, P. Kickingeder, W. Wick, M. Bendszus, and K. H. Maier-Hein, "No new-net," in *International MICCAI Brainlesion Workshop*, 2018, pp. 234–244.
- [11] H. H. Sultan, N. M. Salem, and W. Al-Atabany, "Multi-classification of brain tumor images using deep neural network," *IEEE Access*, vol. 7, pp. 69215–69225, 2019.
- [12] Y.-J. Chen, Y.-J. Chang, S.-C. Wen, Y. Shi, X. Xu, T.-Y. Ho, Q. Jia, M. Huang, and J. Zhuang, "Zero-Shot Medical Image Artifact Reduction," in *ISBI*, 2020, pp. 862–866.
- [13] H. Chen, Y. Zhang, M. K. Kalra, F. Lin, Y. Chen, P. Liao, J. Zhou, and G. Wang, "Low-dose ct with a residual encoder-decoder convolutional neural network," *IEEE transactions on medical imaging*, vol. 36, no. 12, pp. 2524–2535, 2017.
- [14] D. Jiang, W. Dou, L. Vosters, X. Xu, Y. Sun, and T. Tan, "Denoising of 3d magnetic resonance images with multi-channel residual learning of convolutional neural network," *Japanese journal of radiology*, vol. 36, no. 9, pp. 566–574, 2018.
- [15] D. Liu, B. Wen, J. Jiao, X. Liu, Z. Wang and T. S. Huang, "Connecting Image Denoising and High-Level Vision Tasks via Deep Learning," *IEEE Transactions on Image Processing*, vol. 29, pp. 3695–3706, 2020.
- [16] Z. Fan, L. Sun, X. Ding, Y. Huang, C. Cai, and J. Paisley, "A Segmentation-Aware Deep Fusion Network for Compressed Sensing MRI," in *ECCV*, 2018, pp. 55–70.
- [17] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," arXiv preprint arXiv:1312.6199, 2013.
- [18] X. Zhuang, "Challenges and methodologies of fully automatic whole heart segmentation: a review," *Journal of healthcare engineering*, vol. 4, no. 3, pp. 371–407, 2013.
- [19] B.H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest, et al., "The multimodal brain tumor image segmentation benchmark (brats)," *IEEE transactions on medical imaging*, vol. 34, no. 10, pp. 1993–2024, 2014.
- [20] J. Cheng, "brain tumor dataset," 2016.
- [21] X. Glorot, and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, 2010, pp. 249–256.
- [22] D. P. Kingma, and J. Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980, 2014.
- [23] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising," *IEEE Transactions on Image Processing*, vol. 26, no. 7, pp. 3142–3155, 2017.

- [24] A. Mortazi, J. Burt, and U. Bagci, "Multi-planar deep segmentation networks for cardiac substructures from mri and ct," in *International Workshop on Statistical Atlases and Computational Models of the Heart*, 2017, pp. 199—206.
- [25] A. Zhao, G. Balakrishnan, F. Durand, J. V. Guttag, and A. V. Dalca, "Data augmentation using learned transformations for one-shot medical image segmentation," in *CVPR*, 2019, pp. 8543—8553.
- [26] D. Karimi, and S. E. Salcudean, "Reducing the hausdorff distance in medical image segmentation with convolutional neural networks," arXiv preprint arXiv:1904.10030, 2019.