CROSS DOMAIN LOW-DOSE CT IMAGE DENOISING WITH SEMANTIC INFORMATION ALIGNMENT

Jiaxin Huang[1] *Kecheng Chen*[1] *Jiayu Sun*[2] *Xiaorong Pu*[3,1*] *Yazhou Ren*[1]

[1] School of Computer Science and Engineering,
University of Electronic Science and Technology of China, Chengdu, China
[2] West China Hospital, Sichuan University, Chengdu, China
[3] NHC Key Laboratory of Nuclear Technology Medical Transformation,
Mianyang Central Hospital, Mianyang, China

2022 IEEE International Conference on Image Processing (ICIP) | 978-1-6654-9620-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICIP46576.2022.9897265

ABSTRACT

Recently, cross domain adaptation has been applied into quite a few image restoration tasks. While promising performance has been achieved, the domain shift problem between the training set (a.k.a., source domain) and the testing set (a.k.a., target domain) in Low-dose Computed Tomography (LDCT) image denoising tasks is typically ignored by most existing methods. This is prone to the degradation of the denoising performance due to large discrepancy of feature distribution in each dataset from various vendors. Therefore, a simple yet effective LDCT denoising approach has been proposed in this paper to alleviate the domain shift between source and target domains through a novel semantic information alignment. Specifically, we first propose an adaptive version of random frequency mask (RFM) to extract the shared semantic information of cross domains. Then, we incorporate the mask into the existing denoiser to construct a semantic-information-guided objective. Experiments on synthetic and real datasets show our proposed method achieves impressive performance.

Index Terms— Medical image denoising, Low-dose CT, Domain adaptation, Deep learning

1. INTRODUCTION

Low-dose Computed Tomography (LDCT) scanning technology has been widely used as an important imaging modality in modern clinical diagnosis such as early screening of lung nodules and lung cancer, considering the potential radiation risk to the patient [1, 2]. However, reducing the radiation dose increases the noise in LDCT images, which could restrict the further improvement of diagnostic accuracy. Therefore, lots of researches have been proposed to remove the noise from LDCT images. The goal of their methods typically aims to learn a mapping from noisy to clean images on training data (namely source domain) firstly and desire to achieve impressive performance on testing data (namely target domain). However, one may suffer from the domain shift problem, which may be caused by the difference of the semantic information between source and target domains. For example, varying equipment parameters (e.g., the slice thickness) and individual properties may cause the different tissue percentage of obtained CT images between two domains. As a result, addressing the domain shift problem for LDCT denoising is of great interest.

Recently, domain adaptation [3] has attracted lots of attention due to the great progress achieved in style transfer [4], image translation [5] and image restoration tasks [6, 7, 8]. All of these methods are trying to alleviate the domain shift issue in their tasks, which could be divided into to two main streams. One is adversarial learning. Since the emergence of domain adversarial neural networks (DANN) [9], more and more frameworks have been proposed to apply adversarial training to align the source and target distributions on the feature-level [10] or pixel-level [11, 12]. The other focuses on learning an invariant representation from the source to the target domain [13]. For instance, building upon a domain adaptation formulation, Du et al. [7] manage to learn a discrete disentangling representation to align two domains, which is similar to the architecture proposed by Lee et al. [14]. However, they could not reconstruct LDCT images in high quality due to losing finer details or inconsistency backgrounds, which is determined by various features in frequency domain of the image.

In this paper, we tackle the domain adaptation problem for LDCT denoising tasks from a novel perspective, aiming at improving the adaptation ability of the models training from source domains to arbitrary target domains. Different from the methods mentioned above, we propose a play-and-plug medical image alignment model on semantic-wise. To be more specific, the model consists of an adaptive version of

*Corresponding author

Supported by NHC Key Laboratory of Nuclear Technology Medical Transformation (Mianyang Central Hospital) (No. 2021HYX017); Clinical Research Incubation Project, West China Hospital, Sichuan University (No. 2021HXFH004); Sichuan Science and Technology Program (Nos. 2022YFS0055, 2021YFS0172, and 2022YFS0047)

random frequency mask (RFM) and a discriminator. The former part is proposed to extract semantic information of LDCT images from two domains and the latter is used to align different features through adversarial learning. By incorporating the loss of the discriminator with the loss of denoising network and training them in an end-to-end manner, we can obtain better performance of solving both domain shift problem and LDCT image denoising task. The main contributions of the paper are summarized as:

• Propose an adaptive version of random frequency mask to extract semantic features of LDCT images from both source and target domains.

• Incorporate the mask into the existing denoiser to construct a semantic-information-guided objective via exploiting the properties of alignment module as the weightings of the loss function.

• Build a cross domain LDCT image denoising framework on the basis of the semantic features alignment.

2. PROPOSED METHOD

In this section, we focus on how to mask the LDCT image to extract semantic features for alignment and brief how to tune the overall network with the alignment model and the backbone via joint optimization. To our knowledge, this is the first attempt to solve the domain shift problem by aligning semantic features based on a frequency mask. The architecture of the proposed framework is illustrated in Fig. 1.

2.1. Preliminary

Given a source domain, denoted as S, there are n LDCT images with paired normal-dose CT (NDCT) references, denoted respectively as $\mathbf{X}^S = \{\mathbf{x}_1^S, \cdots, \mathbf{x}_n^S\}$ and $\mathbf{Y}^S = \{\mathbf{y}_1^S, \cdots, \mathbf{y}_n^S\}$. In target domain, denoted T, there are m LDCT images available, denoted as $\mathbf{X}^T = \{\mathbf{x}_1^T, \cdots, \mathbf{x}_m^T\}$.

2.2. Semantic Information Alignment Model

Our proposed model consists of the following steps: a) transform an input (a LDCT image from source or target domain) to the frequency domain, b) mask the frequency features with a sampled map based on latent distribution of LDCT images, c) convert the masked frequency representations back to the image that contains efficient semantic information, and d) align the shared features through adversarial learning.

Adaptive Version of Random Frequency Mask. Inspired by the random frequency mask module proposed by Yue et al. [16], we propose an adaptive version of random frequency mask with latent distribution to extract the semantic information shared by two domains. On the basis of discrete cosine transform (DCT) and inverse-DCT (I-DCT), we mitigate the perturbations encoded as high frequency components in LDCT image.

Without loss of generality, let a $H \times W \times C$ tensor denote the input of the module and $\mathbf{X} \in \mathbb{R}^{H \times W}$ denote one of the C channel slices in the input tensor. Discrete cosine transform (DCT) is adopted to compute the frequency representations of $\mathbf{X}$. The DCT result of $\mathbf{X}$ is represented as $\hat{\mathbf{X}} \in \mathbb{R}^{H \times W}$.

$$\hat{\mathbf{X}}(u,v) = c(u)c(v) \sum_{i=0}^{H-1} \sum_{j=0}^{W-1} \mathbf{Y}(i,j) \tag{1}$$

$$\mathbf{Y}(i,j) = \mathbf{X}(i,j) \frac{cos(i+0.5)\pi}{H \cdot u} \cdot \frac{cos(j+0.5)\pi}{W \cdot v}, \tag{2}$$

where $c(u)$ is a compensation coefficient, and the definition of $c(v)$ is the same as $c(u)$ [16].

Traditionally, the DCT maps of LDCT images (noisy images) have larger high-frequency components and smaller middle-frequency coefficients than NDCT images (clean ones) [17]. Especially, semantic information related to some original middle-frequency details is influenced by the noise to some degree [18]. Thus, we propose to reduce the perturbations by multiplying the DCT $\hat{\mathbf{X}}$ with a binary mask $\mathcal{M} \in \mathbb{R}^{H \times W}$:$\hat{\mathbf{X}}' = \hat{\mathbf{X}} \odot \mathcal{M}$, where $\odot$ is element-wise multiplication. Then, the masked DCT $\hat{\mathbf{X}}'$ is translated to the same image as the module input via inverse discrete cosine transform. The whole process of our proposed mask module is formulated as: $\mathbf{X}' = \mathcal{F}^{-1}(\mathcal{M} \odot \mathcal{F}(X))$, where $\mathcal{F}(\cdot)$ stands for DCT, and $\mathcal{F}^{-1}$ denotes the I-DCT. The module output $\mathbf{X}'$ has the same shape as the input $\hat{\mathbf{X}}$.

Through our preliminary experiments, we find that the binary mask $\mathcal{M}$ determined by a Bernoulli distribution with a given probability [16] is not suitable when dealing with LDCT image, as the mask is simply determined by the distance from the lowest-frequency component corresponds to the frequency degree of a component. To efficiently extract the semantic information from LDCT images, learning the latent distributions of semantic-wise is essential. Therefore, we propose to use parameterized variational encoding network, first proposed by [19] to learn the latent distribution $P(\mathbf{z}|\mathbf{x})$ of LDCT image. Here, we denote $\mathbf{z}$ as the latent representation of $\mathbf{x}$. Considering that the distance from the lowest-frequency component corresponds to the frequency degree of a component, $P(\mathbf{z}|\mathbf{x})$ is used to decide whether to mask the current component or not. For each coefficient of position (u, v), we set its corresponding weight in the binary mask according to the latent distribution:

$$r_{(u,v)} = \frac{\sqrt{u^2 + v^2}}{r_{max}}, \tag{3}$$

where r_{max} equals to $\sqrt{(H-1)^2 + (W-1)^2}$ and denotes the maximum radius for a DCT map of size $H \times W$. To preserve the semantic information of the LDCT images, we keep such a DCT coefficient unchanged by setting $\mathcal{M}(u, v)$ as 1. Therefore, the mask $\mathcal{M}(u, v)$ is formally defined as:

$$\mathcal{M}(u,v) = \begin{cases} 1, & r_{(u,v)} > r_{max} \\ P(\mathbf{z}|\mathbf{x}), & 0 \le r_{(u,v)} \le r_{max} \end{cases} \tag{4}$$

After the process of masking, we align the semantic component extracted from both source and target domains, de-

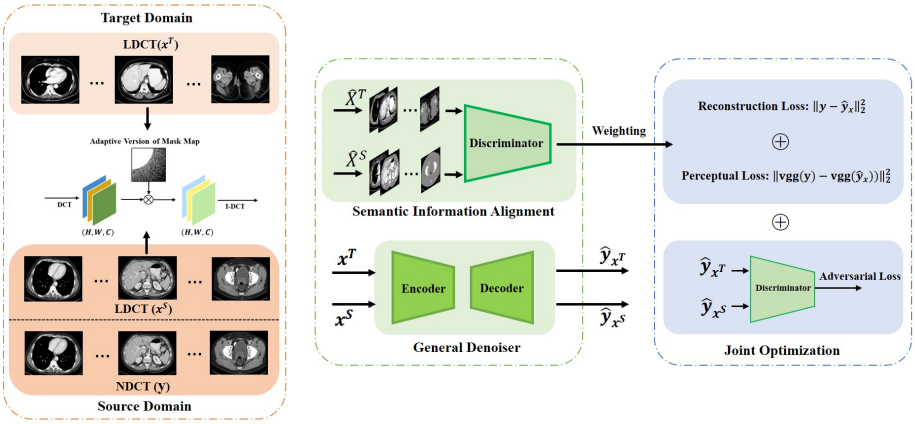

Fig. 1. The framework of our method for LDCT denoising. One can insert arbitrary LDCT denoisers into our proposed framework with a play-and-plug manner. A pretrained VGG-19 [15] network has been adopted to calculate the perpetual loss.

noted as $\mathbf{X}_n^{'}$ and $\mathbf{X}_m^{'}$ respectively, through adversarial learning. The alignment loss represents the similarity of the features shared by two different domains. In this paper, we argue that the lower similarity contributes to lower loss in the loss function of the denoiser. We denote the adversarial loss of semantic information alignment as $\mathcal{L}_w$:

$$\mathcal{L}_w = \mathbb{E}_{\hat{\mathbf{x}}^S \sim P(\hat{\mathbf{x}}^S)}[logD(\hat{\mathbf{x}}^T)] + \mathbb{E}_{\hat{\mathbf{x}}^T \sim P(\hat{\mathbf{x}}^T)}[log(1-D(\hat{\mathbf{x}}^T))] \quad (5)$$

2.3. Denoising Architecture

For the choice of denoiser, as previous studies, a convolution neural network (CNN) is also adopted in this paper, denoted as D. Thus, we adopt a conveying path-based convolutional encoder-decoder (CPCE) [20] as the backbone network by considering the performance and the convenience. The details of this backbone network can be found in [20].

2.4. Joint Optimization

In this paper, we creatively define adversarial loss of the alignment model as a weight to determine the combination of reconstruction loss and perceptual loss, which are used to induce weighting factors of the objective via an adversarial manner. A joint optimization loss function is constructed for denoising purpose as following:

$$\mathcal{L}_{all} = \mathcal{L}_{adv} + \lambda_0 \mathcal{L}_{rec} + \lambda_1 \mathcal{L}_{per} \quad (6)$$

Reconstruction Loss. We apply a general reconstruction loss $\mathcal{L}_{rec}$ to facilitate the training:

$$\mathcal{L}_{rec} = \|\mathbf{y} - \hat{\mathbf{y}}_{\mathbf{x}^S}\|_2^2 \quad (7)$$

Perceptual Loss. Inspired by perception loss [15], the feature from the deeper layers of the pretrained model contain semantic meanings only, which are noiseless or with little noise. In this paper, the loss $\mathcal{L}_{per}$ could be formulated as

$$\mathcal{L}_{per} = \|vgg(\mathbf{y}) - vgg(\hat{\mathbf{y}}_{\mathbf{x}^S})\|_2^2, \quad (8)$$

we use VGG-19 pretrained network on ImageNet.

Adversarial Loss. We impose domain adversarial loss $\mathcal{L}_{adv}$:

$$\mathcal{L}_{adv} = \mathbb{E}_{\mathbf{y} \sim P(\mathbf{y})}[logD(\mathbf{y})] + \mathbb{E}_{\mathbf{y} \sim P(\hat{\mathbf{y}_z})}[log(1 - D(\hat{\mathbf{y}_z}))], \quad (9)$$

where $\hat{\mathbf{y}_z}$ and $\mathbf{y}$ attempt to discriminate the realness of generated images from target domain. We define adversarial loss as a weight to determine the combination of reconstruction loss and perceptual loss.

Table 1. Quantitative analysis of denoising performance.

	LDCT	CPCE	CycleGAN	DRIT++	Ours
			PSNR(mean±var)		
$\mathcal{S}\rightarrow\mathcal{R}$	34.21±0.81	38.58±0.73	39.89±0.79	40.54±0.67	**42.27±0.63**
$\mathcal{R}\rightarrow\mathcal{S}$	34.21±0.81	36.31±0.89	37.11±0.77	39.42±0.72	**40.92±0.58**
			SSIM(mean±var)		
$\mathcal{S}\rightarrow\mathcal{R}$	0.93±0.79	0.95±0.73	0.97±0.68	0.98±0.67	**0.99±0.51**
$\mathcal{R}\rightarrow\mathcal{S}$	0.93±0.79	0.94±0.78	0.95±0.77	0.95±0.62	**0.98±0.53**

3. EXPERIMENTAL RESULTS

3.1. Model Structure and Details

Baseline Methods. In this paper, we evaluate our proposed method based on both synthetic and real clinical datasets. Existing deep learning-based LDCT denoising methods typically perform training on datasets. To follow this protocol, the modularized CPCE has been adopted and trained on the NIH-AAPM-Mayo dataset [20] and natural images with simulated Gaussian noise, respectively. To be fair, we utilize open-source pretraining models of CycleGAN [6] given by authors for comparison. In addition, considering Lee et al. [14] also proposed a method (named DRIT++) for image restoration task, we select it as a representative domain adaptation method to compare the performance on LDCT denoising tasks. In all experiments, we randomly crop patches

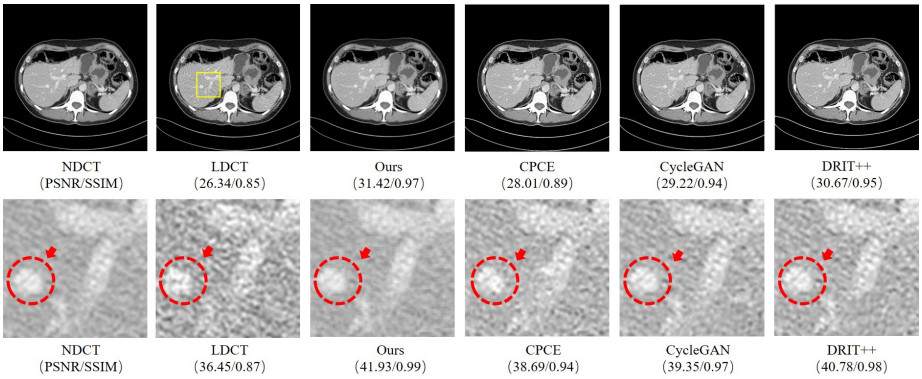

Fig. 2. A patient example from the L209 case in the LDCT-and-Projection-data dataset [21]. We compare different methods under the circumstance of training on synthetic dataset and testing on real dataset. The display window is [40, 400]HU. The yellow rectangle denotes ROI area.

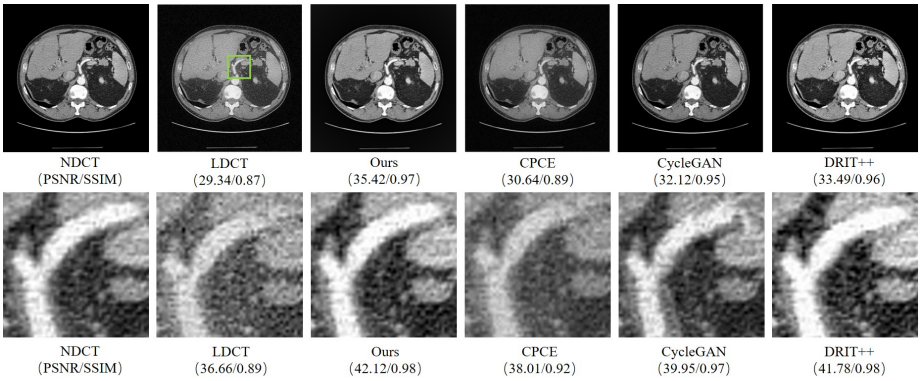

Fig. 3. A patient example in the LIDC-IDRI dataset [22]. We compare different methods under the circumstance of training on real dataset and testing on synthetic dataset. The display window is [40, 400]HU. The green rectangle denotes ROI area.

with batch size of 16 for training. Hyper-parameters are set to λ_0=1-λ_1=0.67.

Datasets. A real clinical dataset authorized by Mayo Clinic is used to evaluate our proposed method, which contains 5936 images in 512 $\times$ 512 resolution from 10 subjects. We randomly select 4000 images as training set, the remaining is as testing set. A synthetic dataset has been constructed by adding random noise to the NDCT images in LIDC-IDRI dataset [22], an extra normal-dose CT dataset. In the experiments, we construct the synthetic datasets by introducing a noise extracting method proposed by Chen et al. [22], which randomly adds the noise extracted from LDCT images to the NDCT images.

3.2. Results

In the experiments, the real clinical dataset and the synthetic dataset are denoted as $\mathcal{R}$ and $\mathcal{S}$ respectively. The L209 case in LDCT-and-Projection-data dataset [21] is used to compare the denoising performances of different methods from $\mathcal{S}$ to $\mathcal{R}$ and the case selected from LIDC-IDRI dataset shows the results from $\mathcal{R}$ to $\mathcal{S}$. As shown in Fig. 2 and Fig. 3, we can observe that our proposed method achieves relatively better noise suppression on both domains compared with other baseline methods. To further evaluate the performance of the details, zoomed region of interest (ROI) of Fig. 2 and Fig. 3 are also given. Table 1 shows the values (mean±var) of quantitative analysis of denoising performance. Peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) are used for evaluation. For PSNR and SSIM, the higher the better. We observe that our method achieves the best performance.

4. CONCLUSION

In this paper, we propose a play-and-plug medical image alignment model on semantic-wise to solve the domain shift problem. Aided by effective semantic alignment, our cross domain method could reconstruct LDCT images with finer details and better visual perception. Experiments on both synthetic and real LDCT image denoising show our method achieves better performance than baseline methods.

5. REFERENCES

[1] Kecheng Chen, Xiaorong Pu, Yazhou Ren, Hang Qiu, Haoliang Li, and Jiayu Sun, "Low-dose ct image blind denoising with graph convolutional networks," in *ICONIP*, 2020, pp. 423–435.

[2] Kecheng Chen, Kun Long, Yazhou Ren, Jiayu Sun, and Xiaorong Pu, "Lesion-inspired denoising network: Connecting medical image denoising and lesion detection," in *ACM MM*, 2021, p. 3283–3292.

[3] Sinno Jialin Pan and Qiang Yang, "A survey on transfer learning," *TKDE*, vol. 22, no. 10, pp. 1345–1359, 2009.

[4] Tero Karras, Samuli Laine, and Timo Aila, "A style-based generator architecture for generative adversarial networks," in *CVPR*, 2019, pp. 4401–4410.

[5] Taesung Park, Alexei A Efros, Richard Zhang, and Jun-Yan Zhu, "Contrastive learning for unpaired image-to-image translation," in *ECCV*, 2020, pp. 319–345.

[6] Daniel Zoran and Yair Weiss, "From learning models of natural image patches to whole image restoration," in *ICCV*, 2011, pp. 479–486.

[7] Wenchao Du, Hu Chen, and Hongyu Yang, "Learning invariant representation for unsupervised image restoration," in *CVPR*, 2020, pp. 14483–14492.

[8] Yuanjie Shao, Lerenhan Li, Wenqi Ren, Changxin Gao, and Nong Sang, "Domain adaptation for image dehazing," in *CVPR*, 2020, pp. 2808–2817.

[9] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky, "Domain-adversarial training of neural networks," *JMLR*, vol. 17, no. 1, pp. 2096–2030, 2016.

[10] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell, "Adversarial discriminative domain adaptation," in *CVPR*, 2017, pp. 7167–7176.

[11] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker, "Learning to adapt structured output space for semantic segmentation," in *CVPR*, 2018, pp. 7472–7481.

[12] Yun-Chun Chen, Yen-Yu Lin, Ming-Hsuan Yang, and Jia-Bin Huang, "Crdoco: Pixel-level domain transfer with cross-domain consistency," in *CVPR*, 2019, pp. 1791–1800.

[13] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang, "Diverse image-to-image translation via disentangled representations," in *ECCV*, 2018, pp. 35–51.

[14] Hsin-Ying Lee, Hung-Yu Tseng, Qi Mao, Jia-Bin Huang, Yu-Ding Lu, Maneesh Singh, and Ming-Hsuan Yang, "Drit++: Diverse image-to-image translation via disentangled representations," *IJCV*, vol. 128, no. 10, pp. 2402–2417, 2020.

[15] Justin Johnson, Alexandre Alahi, and Li Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *ECCV*, 2016, pp. 694–711.

[16] Jiutao Yue, Haofeng Li, Pengxu Wei, Guanbin Li, and Liang Lin, "Robust real-world image super-resolution against adversarial attacks," in *ACM MM*, 2021, pp. 5148–5157.

[17] Zheng Zhang, Zhihui Lai, Zi Huang, Wai Keung Wong, Guo-Sen Xie, Li Liu, and Ling Shao, "Scalable supervised asymmetric hashing with semantic and latent factor embedding," *TIP*, vol. 28, no. 10, pp. 4803–4818, 2019.

[18] Zheng Zhang, Luyao Liu, Yadan Luo, Zi Huang, Fumin Shen, Heng Tao Shen, and Guangming Lu, "Inductive structure consistent hashing via flexible semantic calibration," *TNNLS*, vol. 32, no. 10, pp. 4514–4528, 2020.

[19] Diederik P Kingma and Max Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[20] Hongming Shan, Yi Zhang, Qingsong Yang, Uwe Kruger, Mannudeep K Kalra, Ling Sun, Wenxiang Cong, and Ge Wang, "3-d convolutional encoder-decoder network for low-dose ct via transfer learning from a 2-d trained network," *TMI*, vol. 37, no. 6, pp. 1522–1534, 2018.

[21] Kenneth Clark, Bruce Vendt, Kirk Smith, John Freymann, Justin Kirby, Paul Koppel, Stephen Moore, Stanley Phillips, David Maffitt, Michael Pringle, et al., "The cancer imaging archive (tcia): maintaining and operating a public information repository," *Journal of digital imaging*, vol. 26, no. 6, pp. 1045–1057, 2013.

[22] Kecheng Chen, Jiaxin Huang, Jiayu Sun, Yazhou Ren, and Xiaorong Pu, "Task-driven deep learning for ldct image denoising," in *ISICDM*, 2020, pp. 35–39.