Joint Spine Segmentation and Noise Removal from Ultrasound Volume Projection Images with Selective Feature Sharing

Zixun Huang, Rui Zhao, Frank H.F. Leung, Sunetra Banerjee, Timothy Tin-Yan Lee, De Yang, Daniel P.K. Lun, Kin-Man Lam, Yong-Ping Zheng, Sai Ho Ling

***Abstract*— Volume Projection Imaging from ultrasound data is a promising technique to visualize spine features and diagnose Adolescent Idiopathic Scoliosis. In this paper, we present a novel multi-task framework to reduce the scan noise in volume projection images and to segment different spine features simultaneously, which provides an appealing alternative for intelligent scoliosis assessment in clinical applications. Our proposed framework consists of two streams: i) A noise removal stream based on generative adversarial networks, which aims to achieve effective scan noise removal in a weakly-supervised manner, i.e., without paired noisy-clean samples for learning; ii) A spine segmentation stream, which aims to predict accurate bone masks. To establish the interaction between these two tasks, we propose a selective feature-sharing strategy to transfer only the beneficial features, while filtering out the useless or harmful information. We evaluate our proposed framework on both scan noise removal and spine segmentation tasks. The experimental results demonstrate that our proposed method achieves promising performance on both tasks, which provides an appealing approach to facilitating clinical diagnosis.**

***Index Terms*— Ultrasound volume projection imaging, intelligent scoliosis diagnosis, weakly-supervised scan noise removal, multi-task spine segmentation.**

I. Introduction

ADOLESCENT Idiopathic Scoliosis (AIS) is a serious deformity of the spinal cord, which develops over time and occurs in 2-4% of teenagers [1]. Currently, the Cobb angle based on radiography is the gold standard for the diagnosis of scoliosis [2]. However, exposure to X-rays is harmful to the human body, and an AIS patient needs to receive at least 25 X-rays during the whole treatment. 3D ultrasound imaging, as a radiation-free, inexpensive, and real-time imaging technique, has shown its remarkable reliability and feasibility in the screening and assessment of AIS in the literature [3]–[5].

In clinical AIS diagnosis with 3D ultrasound imaging, experts need to observe hundreds of images in a sequence of the whole spine region. This process is tedious and time-consuming. For faster diagnosis and better visualization of the spine structure, Volume Projection Imaging (VPI) was proposed to project the voxels of 3D ultrasound volume data onto a sequence of 2D spine coronal-plane images [7]. However, owing to the low quality (caused by speckle noise and contrast)

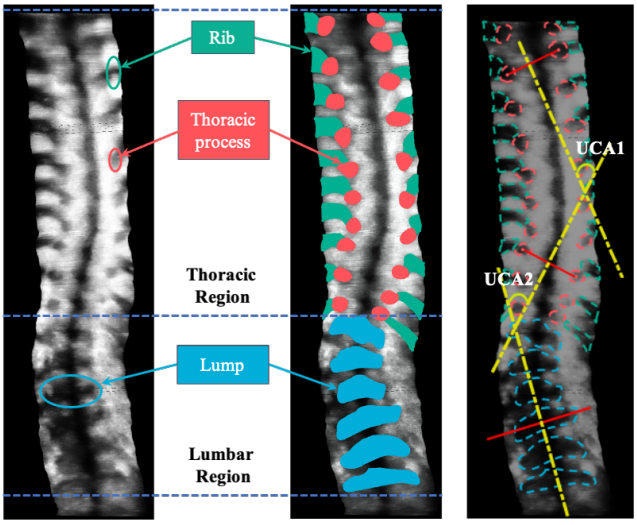


Fig. 1: An illustration of automatic scoliosis diagnosis from ultrasound volume projection images. The image on the left refers to the obtained ultrasound VPI image from a patient with scoliosis. The image in the middle presents different bone features in the VPI image. In the thoracic area, ribs and thoracic processes are annotated in green and red, respectively. In the lumbar region, the lumps, which are formed by the combined shadows of the partial bilateral inferior articular processes, laminae, and the superior articular processes of the inferior vertebrae, are labeled in blue. Following the ultrasound curve angle measurement (UCA) [6], the most tilted lump and paired transverse processes are utilized to calculate the spine deformity.

This study was substantially supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. R5017-18 & No. B-Q86J).

Zixun Huang, Rui Zhao, Frank H.F. Leung, Daniel P.K. Lun, and Kin-Man Lam are with the Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, Hong Kong. ({zixun.huang, rick10.zhao}@connect.polyu.hk, {frank-h-f.leung, enpklun, enkmlam}@polyu.edu.hk)

Sunetra Banerjee, Sai Ho Ling are with the School of Biomedical Engineering, University of Technology Sydney, Australia. (sunetra.banerjee@student.uts.edu.au, steve.ling@uts.edu.au)

Timothy Tin-Yan Lee, De Yang, Yong-Ping Zheng are with the Department of Biomedical Engineering, The Hong Kong Polytechnic University, Hong Kong. ({timothy.lee, de.de.yang}@connect.polyu.hk, yongping.zheng@polyu.edu.hk)

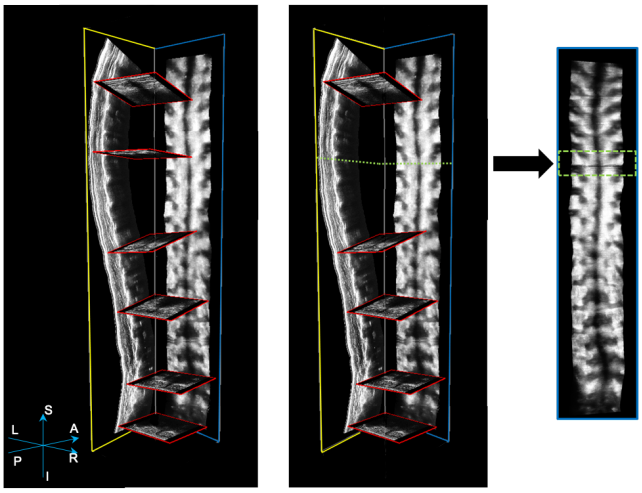


Fig. 2: An illustration of scan noise. (a) A typical 2D ultrasound sequence obtained from an AIS subject through the manual scan. Multiple B-mode images (illustrated as red frames within the volume) are captured and added together to form the 3D volume, which is later used to project the coronal (illustrated as the blue frame) and sagittal (illustrated as the yellow frame) images. (b) In case there is a missing B-mode image for various reasons (e.g., the 5th red frame from the bottom), the noise will be created (green dotted lines). (c) The resultant noise (indicated inside the green box) in the coronal projected image.

TABLE I: FLOPS and the computational time of applying **one** filter to process 3D and 2D data.

Image	Data size	Filter size	FLOPS	Computational Time
2D	(2048, 512)	3×3	1.049×10^7	0.014s
3D	(2048, 512, 512)	$3 \times 3 \times 3$	1.503×10^{10}	6.92s

of the ultrasound images and the acoustic shadow caused by high acoustic impedance of the bones [8], examinations require a rich sonographic experience from experts. The subjective factors behind personal experience are inevitable in manual scoliosis diagnosis. Therefore, current clinical workflow can greatly benefit from an automatic method for spine deformity measurement [7], [9]. As a pre-analyzing step for intelligent scoliosis diagnosis, spine segmentation functions to analyze and locate different bone features, which provides the basis for automatic spine deformity measurement. The idea is illustrated in Fig. 1.

On acquiring an ultrasound sequence, the operator can move the probe on the back smoothly and steadily with a relatively consistent speed in most cases. The probe touches the skin and scans from bottom to top along the spine, with ultrasound gel as a lubricant between the probe and the skin surface. However, owing to various reasons, including the resistance and fluidity of the skin, the resistance of the probe at certain parts of the spine, such as protruded spinous processes, will suddenly change. Therefore, the speed of the probe will increase sharply, resulting in fewer ultrasound frames in some specific areas. Consequently, the reconstructed 2D coronal

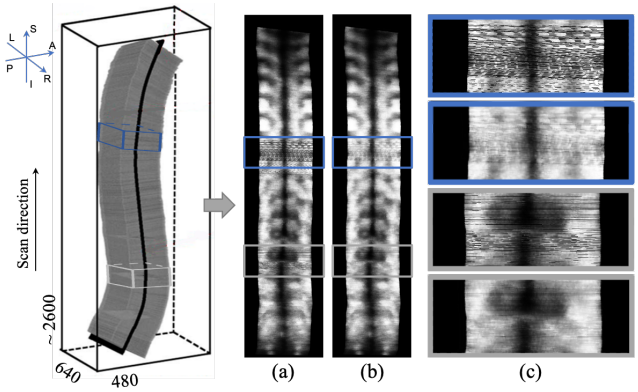


Fig. 3: An example of scan noise removal based on our proposed method for an ultrasound VPI image. The VPI image is generated by non-planar volume rendering from 3D ultrasound volume data [7]. (a) The original VPI image with severe scan noise, (b) The recovered image produced by our proposed method, (c) The scan noise and the recovered details highlighted in the blue and gray boxes.

images will suffer from strong scan noise, as shown in Fig. 3(a). Fig. 2 illustrates one of the possible formations of the noise. The scan noise exacerbates the difficulty in scoliosis assessment. To address this issue, a potential solution is to increase the sampling frequency in the acquisition. However, a high sampling frequency means a higher requirement for scanners, and increases the cost in clinical applications. From another perspective, densifying the sparse ultrasound volume data provides an algorithm-based approach to recovering the signal. However, the high computational complexity makes it impractical in real-time diagnosis. As a more practical solution, noise removal in 2D VPI images is of great interest thanks to its flexibility and efficiency. To show the computational benefit of the noise removal approach over the dense 3D reconstruction approach, we analyze the computational complexity in terms of the floating-point operations per second (FLOPS) and the computational time of each convolutional layer. As tabulated in Table I, FLOPS for processing volume data with 3D filters is about $1,400$ times larger than that with 2D filters on image data. Similarly, the computational cost on 3D data is also around 500 times larger than that on 2D data. With the rapid development of deep learning techniques, Deep Convolutional Neural Networks (DCNNs) have shown their great superiority in medical image restoration [10]–[12]. Despite their high effectiveness, the restoration based on DCNNs generally requires a large amount of paired data for learning. However, scan noise in VPI images is sensitive to various factors, such as ultrasound operators, probing, VPI settings, etc. This phenomenon makes the degradation model of scan noise unpredictable, which precludes us from synthesizing paired noisy and noise-free images for training. In practice, the acquired VPI images cannot produce the paired noise-free reference of each observation. These limitations trigger us to investigate a weakly supervised denoiser especially for scan noise. To this end, Generative Adversarial Networks (GANs)

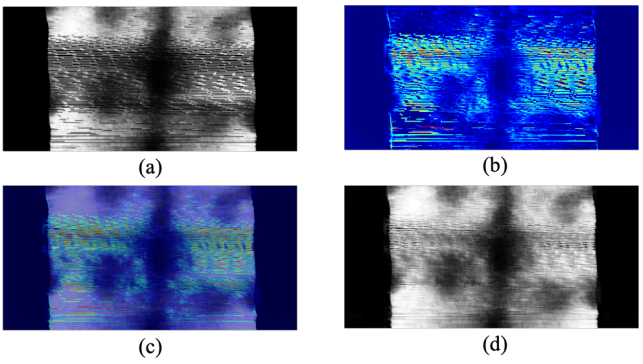


Fig. 4: The visualization of the feature map extracted from the second-last convolutional layer of the noise removal network. (a) The original noisy VPI image, (b) the extracted feature map, (c) overlap of the input image and the feature map, (d) the denoised VPI image.

provide an alternative to learning an image restoration model without paired-data supervision [13].

In this paper, we conduct an in-depth study on GAN-based scan noise removal, and propose a novel weakly supervised denoiser that does not require the noisy-clean pairs for training. We further present a dual-adversarial learning strategy, which serves as an online augmentation approach to enhance the generalization ability of the learned denoiser. An example of the restored image patches based on our proposed method is presented in Fig. 3. It is clear that our proposed denoiser can effectively recover the noisy patterns from the original image, and provide better visual quality for medical experts to analyze the spinal deformity.

The integration of scan-noise removal and spine segmentation affects the efficiency and flexibility of a diagnosis system in clinical applications. Previous studies generally follow a cascaded fashion [14], [15]. They first perform image restoration to enhance the image quality (remove the noise), and then segment the objects of interest based on the recovered images. In spite of the simplicity, those frameworks suffer from low efficiency because the segmentation module has to wait for the results from the restoration modules. Moreover, the segmentation is performed on the reconstructed high-quality images, while the intermediate features are ignored. To better elaborate the intermediate features, we visualize the feature of the second last convolutional layer of the noise removal network as shown in Fig. 4. According to the features shown in Fig. 4(b) and (c), it is obvious that the intermediate features fill the noisy patterns and serve as a compact representation of the denoised output. In this paper, we argue that those intermediate features for reconstructing the high-quality images are important in enhancing the segmentation accuracy, because they synthesize the lost information in the low-quality image. Therefore, we propose to integrate our denoising network and the employed segmentation network in a multi-task manner, which accomplishes the restoration task and the segmentation task simultaneously for better performance and higher inference efficiency. Furthermore, to support the multi-task learning, we adopt a novel selective feature-sharing strategy, which facilitates the framework to select only meaningful features from the auxiliary tasks, and filter out the useless or harmful information. By these means, the proposed joint learning of scan noise removal and spine segmentation can mutually benefit each other.

This paper is an extension of our early work in [16]. We improve our previous method by designing a selective feature-sharing multi-task framework to replace the original cascaded learning strategy. The main contributions of this paper are summarized as follows:

- We first review the noise removal network proposed in our previous work in more detail. Specifically, we further visualize the extracted noisy patches from the training data, and present their distribution in terms of the degree of degradation. Based on the distribution, we perform a deeper analysis to show the feasibility of the adopted automatic annotation strategy. We also conduct more experiments to validate and evaluate the proposed noise removal network. We present the subjective results based on Mean Opinion Scores (MOS), as well as the objective results based on Scan Noise Removal Rate (SNRR) to verify the effectiveness of the proposed restoration method.
- We establish a multi-task framework with a novel selective feature-sharing strategy for learning the scan noise removal and spine segmentation jointly. To the best of our knowledge, it is the first attempt to perform selective feature sharing on the restoration and the segmentation of VPI images in medical image analysis.
- We conduct extensive experiments to demonstrate that the proposed weakly supervised denoiser is beneficial to both the visual quality and segmentation accuracy of the spine bone features. The adopted selective feature-sharing strategy can also enhance the inference efficiency and the performance.

The remainder of this paper is organized as follows. In Sec. II, we review some works related to our proposed methods. In Sec. III, we explicitly present the details of the proposed framework, including the preparation of the unpaired training samples for noise removal, the dual-adversarial strategy for enhancing the learning of the GAN-based network, and the selective feature-sharing mechanism for the multi-task framework's establishment. In Sec. IV, we introduce the experiment settings, and analyze the experimental results. Finally, we conclude this paper in Sec. V.

II. Related Works

A. Ultrasound Volume Projection Imaging

The construction of 3D data from a sequence of 2D ultrasound slices helps to reveal pathology that is obscured in 2D observations [17]. Recently, volume projection imaging (VPI) was proposed to visualize the spinal anatomy in 2D coronal-plane images from 3D ultrasound volume data [7]. The working pipeline of VPI is illustrated in Fig.5. Having acquired an ultrasound sequence of 2D slices, the squared distance weighted (SDW) interpolation [18] is utilized to reconstruct the 3D volume. Then the voxels of reconstructed

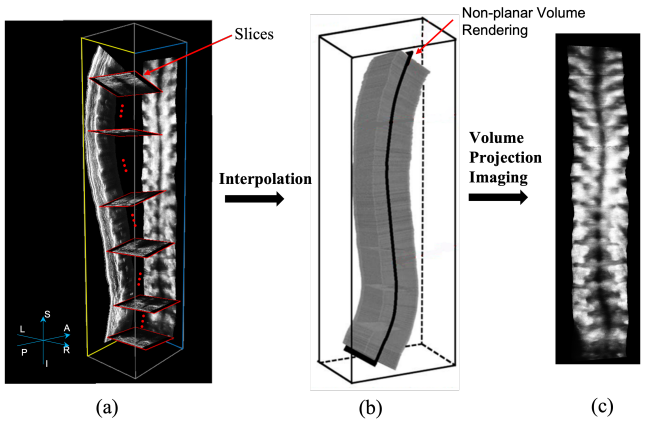


Fig. 5: An illustration of volume projection imaging. (a) The original 2D ultrasound sequence, (b) The reconstructed 3D volume, (c) The extracted VPI image.

3D volume are projected onto a 2D coronal plane by non-planar volume rendering. More details can be found in [7]. In this study, we process the 2D coronal-plane images from VPI to recover the corrupted details that resulted from excessive tilt and also to extract segmentation masks of different bone features. Our proposed algorithm thus improves the robustness and reliability of the scan system, so that, even in some of the worst cases, it can still generate satisfactory visual and segmentation results.

B. Scoliosis Diagnosis with Ultrasound

Previous studies related to ultrasound-based scoliosis measurement can be found. Berton et al. [19] developed a multiple feature extraction framework, and introduced a linear discriminant analysis (LDA) classifier for the segmentation of spinous processes. However, it was difficult for their performance to meet the requirements in clinical tasks. It is because experts should consider not only the spinous processes, but also the transverse processes and the laminae [20]. Recently, Ungi et al. [21] utilized convolutional neural networks (CNNs) to automatically segment the bone features from 2D ultrasound images in an end-to-end manner. However, since the segmentation is performed on sparse 2D images, the predicted segments are of low accuracy. Alternatively, volume projection imaging (VPI), as a 3D volume compression technique [7], provides a promising approach to visualize the whole spine anatomy based on the intensity of the voxels in ultrasound volumetric data. Owing to the superimposition of acoustic shadows on the superficial bone surfaces [8], the spinous processes are visible in VPI images. Chen et al. [4] proposed to manually measure spine deformity based on the middle dark spine profile in VPI images (VPI-SP), which was the first attempt to diagnose scoliosis using the VPI technique. To reduce the inter-observer and intra-observer variations caused by manual measurements in [4], Zhou et al. [22] proposed an automatic framework to model the middle spine curve in VPI images using a 6-th order polynomial. However, owing to the rotation of individual vertebrae in the axial and coronal planes, those VPI-SP based methods tend to underestimate the deformity in scoliosis assessment. To more accurately estimate the spine deformity, a more reliable approach is to compute the spine deformity using the paired thoracic processes and lumbar vertebrae, as shown in Fig. 1. The studies on spine segmentation from VPI images have founded an important basis for intelligent scoliosis diagnosis. Different related results have been obtained. Huang et al. [23] proposed an efficient regularization-based algorithm to address the occlusion issue in VPI images for enhanced spine segmentation. Zhao et al. [24] proposed to introduce the structure supervision to the representation learning in a self-attention manner for more effective spine segmentation. Lyu et al. [25] presented a dual-task framework with boundary detection as an auxiliary task to regularize spine segmentation. Banerjee et al. [26] proposed a lightweight UNet to perform effective spine segmentation with a low computational burden. In our previous study, we proposed a generative adversarial network with dual adversarial learning (DAGAN) to perform noise removal, which is cascaded with a segmentation network for enhanced spine segmentation [16]. However, DAGAN suffers from low efficiency because the segmentation model has to wait for the denoised image produced by the noise removal network. More importantly, the two tasks in DAGAN lack communication with each other, which restricts the knowledge interactions during the framework learning. These phenomena motivate us to investigate a more efficient learning approach in this paper that can perform effective information interactions in a multi-task fashion.

This work extends our previous work on DAGAN in three aspects. First, we perform a more comprehensive discussion on the proposed automatic annotation strategy and the dual-adversarial learning strategy in the noise removal task, through the analysis on the noisy patch distribution. Second, to promote the feature interaction of joint learning, we propose a multi-task framework with a selective feature-sharing strategy to transfer only the beneficial features from the auxiliary tasks. Third, our early work of DAGAN evaluates noise removal only via visualization. To qualitatively and quantitatively evaluate our denoising performance, we conduct a user study and propose an objective metric, i.e., Scan Noise Removal Rate, to show the superiority of our noise removal network.

C. Weakly Supervised Image Restoration

Similar to ultrasound scan noise, Computerized Tomography (CT) and Electron Tomography (ET) reconstructions usually encounter significant amounts of noise caused by low dose, limited, and fragmented data. Total variation (TV) regularization, which minimizes the total variance of an image while preserving the content of the original image, has been used extensively in the noise removal of CT and ET reconstruction [27]–[30]. Mahmood [27] adopted a total variance regularization strategy directly in sinogram for tomographic reconstruction. Traditional TV methods tend to over-smooth the image. Adaptive graph-based total variance (AGTV) [28] was proposed to preserve texture details and reduce artifacts caused by over-smoothing. However, AGTV does not consider the localization information of the graph patch. As an improved version of AGTV, non-local patch graph total

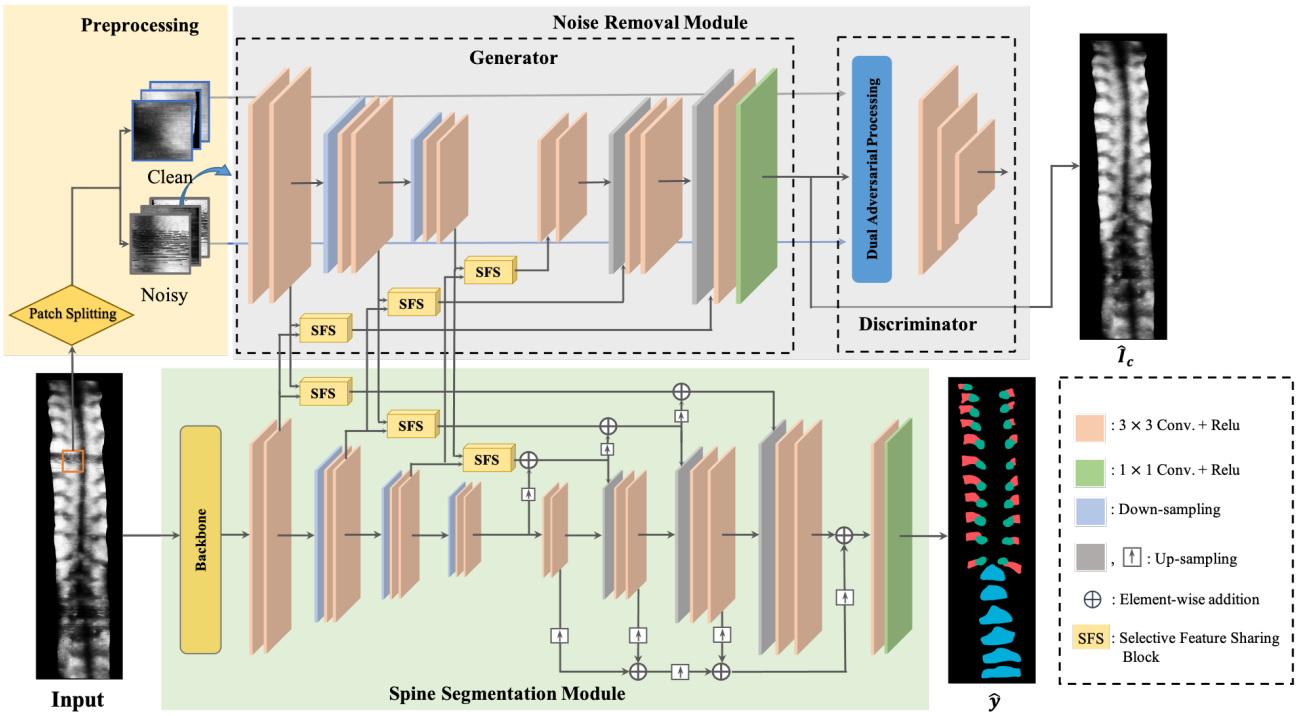


Fig. 6: An overview of the proposed framework, which consists of four main components, i.e, a preprocessing module, a noise removal module, a spine segmentation module, and selective feature-sharing blocks.

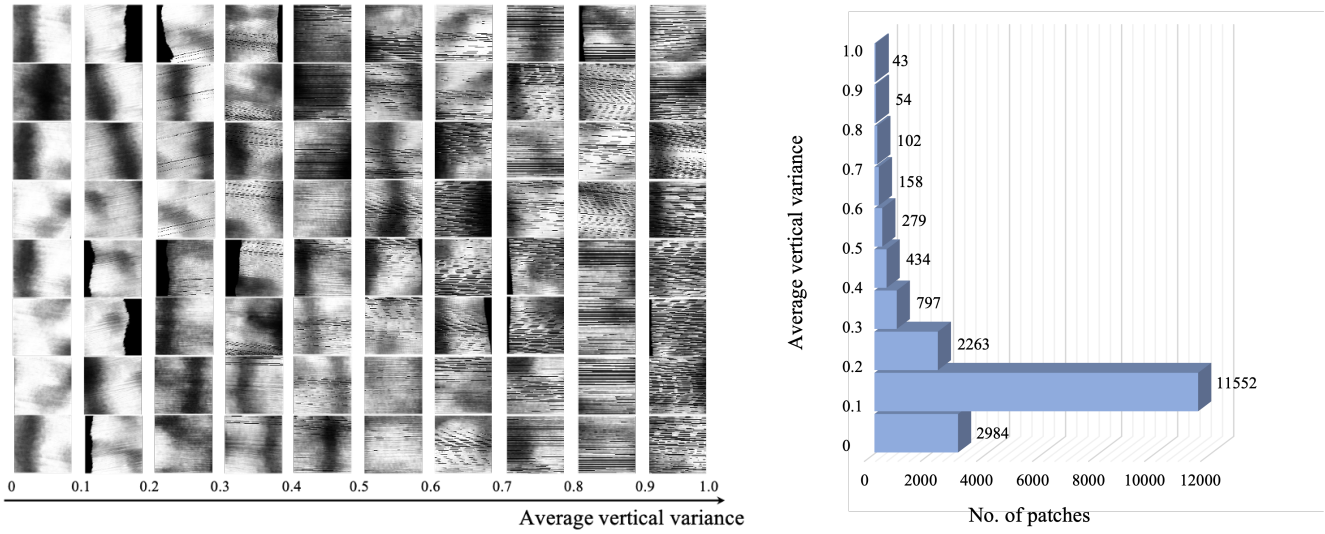
variation (NPGTV) [30] introduces the pixel coordinate as an ingredient to construct the K-nearest neighbor graph for effective denoising. We will compare the denoising results of various kinds of algorithms with those of our proposed algorithms in the experiment section.

Recently, generative adversarial networks (GANs) have been widely used in noise removal tasks and have achieved promising performance [31]–[35]. Matsui et al. [32] proposed a GAN-based model for single-image rain noise removal. Bobrow et al. [33] proposed a domain transformation model between coherent and incoherent illumination for laser speckle reduction. Yang et al. [35] proposed a perceptual loss function for training the GAN-based model to perform low-dose CT image denoising.

As introduced in Sec. I, it is difficult to acquire pairs of noisy-clean images from VPI for learning the noise removal task in a fully supervised manner. Thus, we propose a weakly-supervised denoising method in this paper. Recently, reference-free image denoising methods have been studied. Noise2Noise [36] was proposed to learn a deep denoiser without clean references. However, it requires extra knowledge of the noise generation model, which limits its application to ultrasound images. Noise2Void [37] and Noise2Self [38] are two self-supervision strategies for image restoration. However, they were shown to perform poorly on structured noise. On the other hand, generative adversarial networks (GANs) provide an alternative to solving the restoration problem in a weakly supervised manner. Hou et al. [39] adopted the cycle-adversarial strategy to reconstruct the image appearance for enhancing the segmentation accuracy of CT images. However, Liu et al. [40] showed that GAN-based methods would create some artifacts in the recovered images. To tackle this problem, they introduced a wavelet correction transfer network (WaveCT). However, the spectral-based supervision in [40] does not perform satisfactorily on structured noise removal. It can be seen that the aforementioned methods are not desirable for enhancing the quality of VPI images in clinical applications, which motivates us to explore a more effective denoiser for VPI images.

D. Multi-task Learning with Feature Selection

Instead of using the conventional cascaded approach to do the restoration and then the segmentation tasks, we adopt a multi-task approach to doing these two tasks in parallel for higher inference efficiency. The feature selection strategy plays the most important role in this multi-task learning, which has been widely investigated in natural language processing. Ruder et al. [41] proposed the Sluice Networks, in which a linear combination approach was designed to control the information flow between different tasks. Xiao et al. [42] took advantage of gated recurrent units and proposed a leaky unit with the property of remembering and forgetting knowledge. Zhao et al. [43] enhanced the leaky units in [42] and introduced the convolutional feature leaky units to perform feature selection between facial expression recognition and facial expression synthesis. In this paper, inspired by [42] and [43], we design a two-branch framework with selective feature sharing to bridge the helpful knowledge between the restoration task and the segmentation task.



(a) Visualization of the noisy patches based on the average vertical variance. (b) Distribution analysis on the noisy patches.

Fig. 7: The visualization and the distribution of the patches with different ranges of average vertical variance.

III. Methodology

In this section, we present the details of the proposed framework for joint scan noise removal and spine segmentation with selective feature sharing. We first overview the whole working pipeline of the framework, and then introduce the detailed design for each proposed component, including the generation of training samples, the dual-adversarial strategy, and the selective feature sharing mechanism.

A. Overview

The proposed selective feature-sharing multi-task framework is depicted in Fig. 6, which consists of four main components. i) A preprocessing module is employed to prepare unpaired samples for weakly supervised denoising learning. ii) A noise removal module is designed based on a generative adversarial network. It is a generator following a simple U-shape network architecture, which takes a noisy image as input, and estimates its corresponding noise-free image. iii) A spine segmentation module serves as a network for discriminating different spine features. iv) Selective feature-sharing (SFS) blocks are designed to connect the two modules (for noise removal and spine segmentation) with feature selection. It is worth noting that the preprocessing module and the discriminator only affect the training stage.

B. Training Patch Preparation for Noise Removal

The preprocessing step aims to prepare unpaired training samples for scan noise removal. We first extract the patches from the images by a sliding window of size $S \times S$ pixels, and split them into two categories, i.e., a noise-free group and a corrupted group. As shown in Fig. 3, the scan noise in the resulting VPI images follows a similar pattern. Owing to the probing characteristic, the projected images suffer from the corruption of horizontal dark lines. Based on this observation, we employ an edge detector to compute the average vertical variance $g(\boldsymbol{x})$ of the extracted patches $\boldsymbol{x}$, and then divide them into the positive and the negative groups by two predefined thresholds as follows:

$$f(\boldsymbol{x}) = \begin{cases} 0, & \text{if } g(\boldsymbol{x}) \geq \beta_n, \\ 1, & \text{if } g(\boldsymbol{x}) \leq \beta_p, \end{cases} \quad \text{with } g(\boldsymbol{x}) = \frac{1}{S^2} \sum_{i}^{S} \sum_{j}^{S} |h(\boldsymbol{x})_{i,j}|, \tag{1}$$

where $\boldsymbol{x}$ denotes an extracted image patch, β_n and β_p are the negative and positive thresholds for selecting the corrupted and clean patches respectively, $h(\cdot)$ represents the vertical edge detection function. By this means, we can synthesize a large amount of domain-transfer pairs for learning a denoiser in an unpaired manner.

To validate the adopted patch preparation strategy for training, we visualize the patch distribution in terms of different ranges of the average vertical variance, as shown in Fig. 7a. The average vertical variance of each patch is normalized by min-max feature scaling. As shown in Fig. 7a, the scan noise becomes stronger and stronger as the value of the average vertical variance rises. The patches in the range of $[0, 0.2]$ are basically noise-free, while the patches in the range of $(0.3, 1.0]$ contain severe scan noise. The patches in the range of $(0.2, 0.3]$ are noisy or noise-free. To avoid label confusion during training, we discard the patches in the range of $(0.2, 0.3]$. In other words, the negative and the positive thresholds, i.e., β_n and β_p, are set to 0.3 and 0.2, respectively, in this paper.

C. Noise Removal with Dual Adversarial Learning

The noise removal module aims to restore clean images from the noisy inputs, which contains two main components, i.e., the generator for processing the images and the discriminator for the images synthesized from the generator against the images from the real training dataset. Given an unpaired input

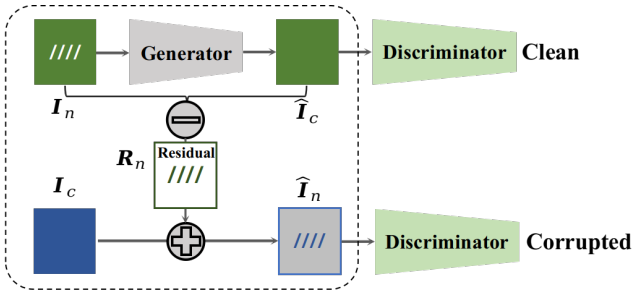


Fig. 8: An illustration of the proposed dual adversarial learning for on-the-fly augmentation in generative adversarial networks. ($\boldsymbol{I}_n$ and $\boldsymbol{I}_c$ are the unpaired noisy and clean images, $\hat{\boldsymbol{I}}_c$ denotes the estimated noiseless image, $\hat{\boldsymbol{I}}_n$ represents the synthesized corrupted sample.)

couple $\{\boldsymbol{I}_n, \boldsymbol{I}_c\}$, where $\boldsymbol{I}_n$ and $\boldsymbol{I}_c$ denote the unpaired noisy and clean images respectively, a generator $\mathcal{G}$ is employed to obtain the noiseless estimation, denoted as $\hat{\boldsymbol{I}}_c = \mathcal{G}(\boldsymbol{I}_n)$. The discriminator $\mathcal{D}_{\text{clean}}$ is employed to discriminate the noiseless estimation $\hat{\boldsymbol{I}}_c$ and the real clean image $\boldsymbol{I}_c$. Thus, the learning objective is formulated as follows:

$$\begin{aligned}\mathcal{L}_{\text{adv}} =& \mathbb{E}_{\boldsymbol{I}_c}[\log(\mathcal{D}_{\text{clean}}(\boldsymbol{I}_c))] + \mathbb{E}_{\hat{\boldsymbol{I}}_c}[\log(1-\mathcal{D}_{\text{clean}}(\hat{\boldsymbol{I}}_c))], \\ & \text{with } \hat{\boldsymbol{I}}_c = \mathcal{G}(\boldsymbol{I}_n),\end{aligned} \tag{2}$$

where $\mathcal{L}_{\text{adv}}$ denotes the adversarial loss to be minimized, and $\mathbb{E}$ refers to the expectation. As shown in Fig. 7b, the samples in noisy and noise-free groups are highly imbalanced. Therefore, we use a dual adversarial strategy to serve as an on-the-fly augmentation to enhance the adversarial training. As shown in Fig. 8, consider the estimated noiseless image $\hat{\boldsymbol{I}}_c$, we compute the residual between the input noisy image and the output estimation as $\boldsymbol{R}_n = \boldsymbol{I}_n - \hat{\boldsymbol{I}}_c$, which represents the noisy patterns in the input image. Then, we add the noisy residual back to another clean image $\boldsymbol{I}_c$, denoted as $\hat{\boldsymbol{I}}_n = \boldsymbol{I}_c + \boldsymbol{R}_n$, which synthesizes a new corrupted sample $\hat{\boldsymbol{I}}_n$. Therefore, the generator aims to mislead another discriminator $\mathcal{D}_{\text{noisy}}$ for classifying the synthetic noisy and real noisy images as follows:

$$\begin{aligned}\mathcal{L}_{\text{dua}} =& \mathbb{E}_{\boldsymbol{I}_n}[\log(\mathcal{D}_{\text{noisy}}(\boldsymbol{I}_n))] + \mathbb{E}_{\hat{\boldsymbol{I}}_n}[\log(1-\mathcal{D}_{\text{noisy}}(\hat{\boldsymbol{I}}_n))], \\ & \text{with } \hat{\boldsymbol{I}}_n = \boldsymbol{I}_n - \mathcal{G}(\boldsymbol{I}_n) + \boldsymbol{I}_c,\end{aligned} \tag{3}$$

where $\mathcal{L}_{\text{dua}}$ denotes the dual adversarial loss. Therefore, the overall learning objective $\mathcal{L}_{\text{denoise}}$ for the noise removal branch is defined as follows:

$$\mathcal{L}_{\text{denoise}} = \mathcal{L}_{\text{adv}} + \mathcal{L}_{\text{dua}} + \alpha||\boldsymbol{I}_n - \hat{\boldsymbol{I}}_c||_2^2, \tag{4}$$

where $||\cdot||_2^2$ refers to the $L2$ distance, which is employed to regularize the generator to preserve the image content, and α denotes a hyperparameter controlling the trade-off between noise removal and content preservation.

D. Spine Segmentation

The spine segmentation branch aims to perform semantic segmentation to effectively separate different bone features. Since the bone features are of different scales, we adopt a commonly used segmentation head, i.e., Feature Pyramid

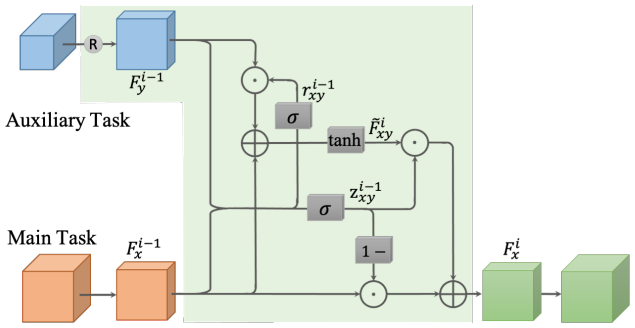


Fig. 9: An illustration of the proposed selective feature-sharing (SFS) blocks in our framework. ($\boldsymbol{F}_x^{i-1}$ and $\boldsymbol{F}_y^{i-1}$ are the features in the main task and the auxiliary task respectively, $\boldsymbol{r}_{xy}^{i-1}$ denotes the leaky gate that selects the information from the auxiliary task, $\boldsymbol{z}_{xy}^{i-1}$ denotes the memory gate that determines the features from the main task, $\widetilde{\boldsymbol{F}}_{xy}^{i}$ denotes the fused features from the two tasks.)

Network (FPN) [44], to do semantic segmentation. FPN fuses the features from different scales of the input image, which benefits the segmentation of small foreground objects, such as thoracic vertebrae. Given an input image $\boldsymbol{I} \in \mathbb{R}^{1\times H\times W}$, where H and W denote its height and width respectively, the FPN backbone $\mathcal{F}$, together with the segmentation head, is employed to predict the foreground segment mask, denoted as $\hat{\boldsymbol{y}} = \mathcal{F}(\boldsymbol{I})$. $\hat{\boldsymbol{y}} \in \mathbb{R}^{N\times H\times W}$ is the output logit tensor whose number of channels N is equal to the number of classes. Here, we consider three foreground objects, i.e., rib, thoracic process, and lumbar, together with one background, making $N = 4$. To supervise the learning of spine segmentation, we adopt the segmentation loss $\mathcal{L}_{\text{segment}}$ based on the pixel-wise cross entropy (ℓ_{CE}) loss as follows:

$$\begin{aligned}\mathcal{L}_{\text{segment}} &= \frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W}\ell_{\text{CE}}(\hat{\boldsymbol{y}}_{ij}, \boldsymbol{y}_{ij}), \\ \ell_{\text{CE}} &= -\log\left(\frac{\exp(\hat{y}[class])}{\sum_k \exp(\hat{y}[k])}\right),\end{aligned} \tag{5}$$

where $\boldsymbol{y}$ denotes the ground-truth segment masks, with a onehot logit vector at each pixel position. $\hat{y}[k]$ and $\hat{y}[class]$ refer to the k-th and the ground-truth elements of the predicted logit vector $\hat{\boldsymbol{y}}$, respectively.

The overall learning objective $\mathcal{L}$ for training the whole framework is defined as:

$$\mathcal{L} = \mathcal{L}_{\text{segment}} + \lambda\mathcal{L}_{\text{denoise}}, \tag{6}$$

where λ denotes the hyperparameter controlling the trade-off between the two different tasks.

E. Selective Feature-Sharing (SFS) Blocks

Different from the cascaded structure in previous studies of joint noise removal and segmentation, we propose to adopt a parallel learning strategy with selective feature sharing to improve model efficiency. Inspired by the gate recurrent units for controlling the information flow in handling sequential data [45], we design feature-sharing blocks to control the

information flow between different tasks. The structure of the feature-sharing block is illustrated in Fig. 9. We consider to transfer features from the auxiliary task y to the main task x at the i-th sharing step. A spatial rescaling operator is first employed to normalize the spatial size of the feature in the auxiliary task $\boldsymbol{F}_y^{i-1}$ to be the same as the features in the main task $\boldsymbol{F}_x^{i-1}$. After that, we define a leaky gate $\boldsymbol{r}_{xy}^{i-1}$ to select the information in the auxiliary task, which is beneficial to the main task, as follows:

$$\boldsymbol{r}_{xy}^{i-1} = \sigma(\boldsymbol{W}_{\boldsymbol{r}}^{i-1} * [\boldsymbol{F}_x^{i-1}, \boldsymbol{F}_y^{i-1}]), \tag{7}$$

where $\sigma(\cdot)$ is the sigmoid function, $\boldsymbol{W}_{\boldsymbol{r}}^{i-1}$ denotes the learnable convolutional kernels to generate the leaky gate scores, and $[\cdot,\cdot]$ refers to channel concatenation. Having obtained the leaky gate $\boldsymbol{r}_{xy}^{i-1}$, we employ it to fuse the features from the two tasks as follows:

$$\widetilde{\boldsymbol{F}}_{xy}^{i} = \tanh(\boldsymbol{W}^{i-1} * (\boldsymbol{r}_{xy}^{i-1} \odot \boldsymbol{F}_y^{i-1}) + \boldsymbol{U}^{i-1} * \boldsymbol{F}_x^{i-1}), \tag{8}$$

where $\boldsymbol{W}^{i-1}$ and $\boldsymbol{U}^{i-1}$ denote the learnable convolutional kernels, $\odot$ is the element-wise multiplication operator. Moreover, a memory gate $\boldsymbol{z}_{xy}^{i-1}$ is further designed to determine the features that should be memorized from the $(i-1)$-th step to the i-th step in the main task x, which is formulated as follows.

$$\boldsymbol{z}_{xy}^{i-1} = \sigma(\boldsymbol{W}_{\boldsymbol{z}}^{i-1} * [\boldsymbol{F}_x^{i-1}, \boldsymbol{F}_y^{i-1}]), \tag{9}$$

where $\boldsymbol{W}_{\boldsymbol{z}}^{i-1}$ denotes the learnable convolutional kernels to generate the memory gate scores. Finally, we construct the output features for the main task by aggregating the information controlled by the leaky and memory gates as follows:

$$\boldsymbol{F}_x^{i} = (1 - \boldsymbol{z}_{xy}^{i-1}) \odot \boldsymbol{F}_x^{i-1} + \boldsymbol{z}_{xy}^{i-1} \odot \widetilde{\boldsymbol{F}}_{xy}^{i}. \tag{10}$$

By this means, we effectively build the connections between the main task and the auxiliary task with meaningful feature selection. It is worth noting that, as shown in Fig. 6, the feature-sharing blocks are placed in pairs in the noise removal module and the spine segmentation module. Thus, the information flow is bidirectional between these two tasks.

F. Training Scheme

We train the whole framework in an end-to-end manner under the learning objective defined in Eq. (6) to jointly learn the two tasks and optimize those feature-sharing blocks. Our proposed framework consists of two main modules, i.e., a noise removal module and a spine segmentation module. From the perspective of task complexity, the segmentation task is more complicated than the noise removal task in learning, since the segmentation task is a pixel-wise multi-class classification task while the noise removal task is a binary classification task. In addition, when we independently solve the two problems, it takes about 8 hours to train up a segmentation network, while training a denoising network only requires about 4 hours. Therefore, in each iteration, we empirically train the noise removal module once and the spine segmentation module twice. When training the noise removal module, we update the discriminator for three optimization steps through experimental findings, and then update the generator for one step. It is worth mentioning that, whichever module is trained, the feature-sharing blocks are optimized to select beneficial features from the representations in the two tasks. This training scheme is summarized in a PyTorch-like pseudo code in the Appendix.

IV. Experiments

In this section, we introduce the experiments for evaluating the performance of the proposed multi-task framework. We first describe the experimental settings, including the preparation of the collected data samples and the implementation of the proposed framework. Secondly, we present and analyze the results, and compare them with those from other state-of-the-art methods. Finally, we show the ablation studies for validating the proposed designs in our framework.

A. Dataset

In our experiments, we acquired 2D VPI images using the Scolioscan system (Model SCN801, Telefield Medical Imaging Ltd, Hong Kong). The frame rate of the system is 60 fps. The ultrasound probe scans from bottom to top along the spine with an average scanning speed of 2.0 cm/s. Our experimental procedures involving human subjects were approved by the Institutional Review Board [1]. The subjects gave informed consent to their inclusion in this study as required, and the work adheres to the Declaration of Helsinki. We collected 109 images from 109 patients (82 females and 27 males) with an average age of 15.6 ± 2.7 years. The mean of the body mass index (BMI) of the subjects was $18.3 \pm 2.1\ kg/m^2$. All the patients had no neuromuscular or congenital problems and received standard diagnosis and treatment. The patients suffer from different degrees of spine deformity, which presents further challenges to the generalization of our proposed framework. Having obtained the sequence of 2D ultrasound slices, the system employs the squared distance weighted (SDW) interpolation to reconstruct the 3D volume data as shown in Fig. 5. Based on different imaging depths, nine 2D coronal images can be obtained from the 3D ultrasound voxels. Among those nine observations, only the image with the best quality was manually selected by the experts for our experiments. The bone features are annotated by three ultrasound experts, one has more than 2-year experience while the other two have 5-year experience. These images were of different sizes, but basically have a resolution of about $2,600 \times 640$ pixels with a spacing of $0.15mm \times 0.15mm$. To spatially normalize the collected images, we resize them into 2048×512 pixels for both training and validation. In the training stage, we further extracted image patches of size 128×128 pixels as the training samples. Random rescalings with the ratio uniformly sampled from the range $(0.5, 2.0)$, random crop, and random flip over were performed as data augmentation. To perform a general test on the proposed framework, we employed the 5-fold subject-independent cross-validation. The whole dataset was uniformly divided into 5 folds. In each round, we trained

[1]This study was performed in line with the principles of the Declaration of Helsinki. Approval was granted by the Ethics Committee of The Hong Kong Polytechnic University (06 Sep 2018/HSEARS20180906005)

the network on 4 folds and validated the remaining 1 fold. We repeated the procedure 5 times to validate the framework with the whole dataset, and the final results were obtained by averaging the accuracy over the 5 rounds.

B. Implementation Details

We implemented our framework with PyTorch and MMSegmentation[2]. ResNet101[3] [46] was adopted as the backbone network for feature extraction in the spine segmentation module. The Feature Pyramid Network[4] (FPN) [44] was employed as the segmentation head to fuse multi-scale features for segmentation. For the noise removal module, we established a simple U-Net architecture, as shown in Fig. 6, as the generator. A four-layer plain convolutional network, followed by two fully connected layers, was built as the discriminator. All the convolutional kernels in the framework were of size 3×3, except for those in the residual connections, where 1×1 convolutional filters were used. We utilized ReLU as the activation function following each convolutional layer.

There are algorithms proposed to stabilize the training of GANs, such as spectral normalization [47], orthogonal regularization [48], and adaptive discriminator augmentation [49]. In our noise removal task, the noisy and clean patches are all extracted from ultrasound VPI images, which indicates they follow a similar distribution. To reduce the computational complexity in deployment, we adopted a simple baseline of GAN for our denoising task.

In the training stage, we employed Adam to optimize the noise removal network and the spine segmentation network jointly. The hyperparameter in the objective function, i.e., α in Eq. (4) and λ in Eq. (6) were empirically set to 10^{-4} and 1.0 by grid search. We trained both networks for 8×10^4 iterations with the learning rate gradually decreasing from 10^{-3} to 10^{-5} based on the cosine annealing schedule [50] [5]. The weight decay was set to 5×10^{-4} for regularization. The mini-batch size for both tasks was fixed at 8.

The training was performed on two Nvidia GeForce RTX 3090 GPUs, and it took about 12 hours to finish a round of learning. Our implementation and some samples from the collected dataset are available at `https://github.com/jacksonhzx95/Joint_segmentation_denoise_for_scoliosis.git`.

C. Metrics

To evaluate the performance of our proposed framework on spine segmentation, we adopt the widely used metrics of Dice similarity score (Dice), precision, and average Hausdorff distance (AHD), which are formulated as follows.

$$\text{Dice} = \frac{2TP}{2TP + FP + FN}, \tag{11}$$

$$\text{Precision} = \frac{TP}{TP + FP}, \tag{12}$$

$$\text{AHD}(\mathcal{A}, \mathcal{B}) = max(d(\mathcal{A}, \mathcal{B}), d(\mathcal{B}, \mathcal{A})), \tag{13}$$

where TP, TN, FP, and FN refer to true positive, true negative, false positive, and false negative points, respectively. $\mathcal{A}$ and $\mathcal{B}$ denotes the ground-truth and predicted segmentation masks, respectively. $d(\mathcal{A}, \mathcal{B})$ is the directed average Hausdorff distance given by

$$d(\mathcal{A}, \mathcal{B}) = \frac{1}{V} \sum_{a \in \mathcal{A}} \min_{b \in \mathcal{B}} \|a - b\|, \tag{14}$$

where V is the total number of the bone feature pixels from $\mathcal{A}$.

As introduced in Sec. I, it is difficult to obtain the paired samples of noisy and noise-free images, and thus the conventional image quality measurements, e.g., Peak Signal-to-Noise Ratio (PSNR) and Structure Similarity Index (SSIM), are not available in this study. Instead, we designed a scan noise removal rate (SNRR) metric to quantitatively evaluate the noise removal performance, which is defined as follows:

$$\text{SNRR} = \frac{|N - N'|}{N}, \tag{15}$$

where N, N' denotes the number of the noisy patches detected by a vertical edge detector from the original and denoised images, respectively. The details of the noisy patch detector are summarized in Algorithm 2 in the Appendix.

D. Results on Spine Segmentation

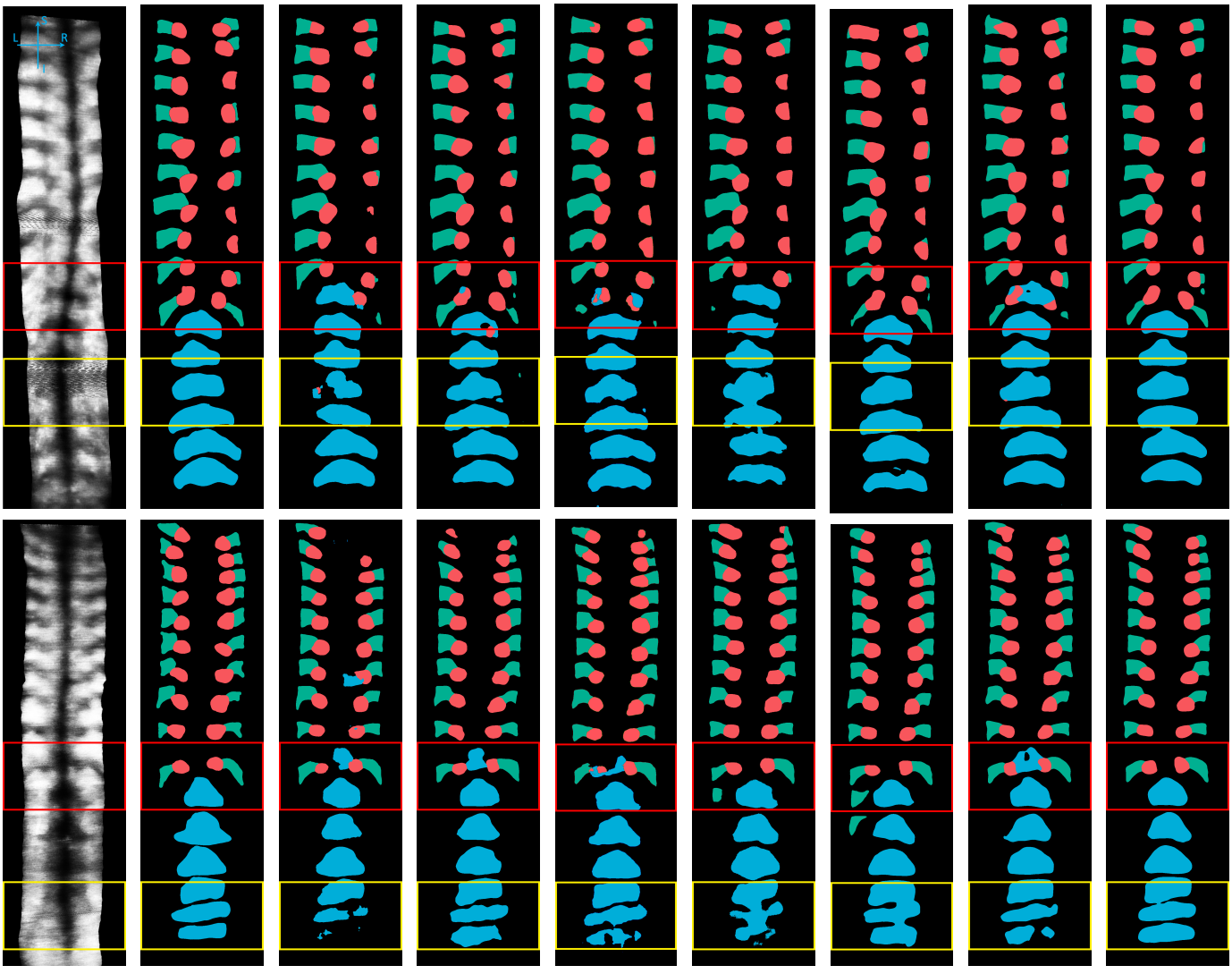
We first evaluate our proposed framework on spine segmentation by comparing it with other state-of-the-art segmentation methods under the same settings, including the benchmark methods of UNet [51], FPN [44] and HRNet [52], the state-of-the-art algorithms of nnUNet [54] and UNet++ [53] for medical image segmentation, the multi-task algorithms of MASSL [55] and DCR [25], and the methods of DAGAN [16] and SEAM [24] especially designed for ultrasound VPI images. It is worth noting that our previous work DAGAN [16] also aims to recover those scan noises in VPI images. However, it performs restoration and segmentation independently. All the comparing methods are established with the source codes provided by their original authors and are applied to our spine segmentation task, except for DCR [25] and MASSL [55]. We re-implemented DCR [25] and MASSL [55] following the descriptions in their papers.

The comparison results are tabulated in Table II. It can be seen that our proposed feature-selective joint-learning framework outperforms the three baseline methods, i.e., UNet [51], FPN [44] and HRNet [52], by a large margin on all the evaluation metrics. This shows the effectiveness of the proposed joint learning scheme for the spine segmentation task. Comparing with the state-of-the-art methods designed for medical image segmentation, i.e., nnUNet [54] and UNet++ [53], we observe an improvement of over 1% on the average Dice score and 3% on the average precision, respectively. We consider the reason to be that the strong scan noise in the VPI images corrupts the discriminative patterns of spine

[2]https://github.com/open-mmlab/mmsegmentation
[3]https://github.com/pytorch/vision/blob/master/torchvision/models/resnet.py
[4]https://github.com/jwyang/fpn.pytorch
[5]https://github.com/pytorch/pytorch/blob/master/torch/optim/lr_scheduler.py

TABLE II: Quantitative segmentation results in terms of the Dice score (Dice (%)), Precision (Pre. (%)), and Average Hausdoff Distance (AHD (mm)) based on different methods (Mean ± Standard Deviation (SD)).

Methods	Lump			Thoracic			Rib			*Ave.*			FPS
	Dice	Pre.	AHD	Dice	Pre.	AHD	Dice	Pre.	AHD	Dice	Pre.	AHD	
UNet [51]	80.26±6.06	78.40±8.47	9.92±7.54	73.64±5.88	79.10±7.15	2.55±1.61	76.02±6.10	79.13±7.98	2.93±2.61	76.64±6.01	78.88±7.87	5.14±3.92	4.0
FPN [44]	83.31±5.70	81.77±7.73	5.21±5.37	74.84±5.45	79.37±6.61	2.08±1.64	76.86±5.91	80.06±7.14	2.35±1.65	78.34±5.69	80.40±7.16	3.21±2.89	**4.2**
HRNet [52]	83.03±5.43	85.63±5.88	1.97±1.61	76.00±4.90	75.41±7.68	1.63±0.78	77.27±5.14	78.05±6.47	1.65±0.84	78.77±5.16	79.69±6.67	1.75±1.07	2.4
UNet++ [53]	82.28±6.89	83.82±7.96	3.16±3.66	73.46±5.72	77.81±7.18	3.19±2.69	76.84±5.74	76.28±8.22	2.79±2.13	77.52±6.12	79.30±7.79	3.05±2.82	3.2
nnUNet [54]	83.11±5.41	81.59±7.62	2.11±**0.79**	77.06±4.60	77.90±7.05	1.69±1.34	78.32±**4.71**	80.46±7.45	1.77±1.29	79.47±**4.91**	79.98±7.37	1.86±1.14	0.3
DAGAN [16]	83.92±5.42	83.03±6.51	4.54±4.8	76.01±4.74	77.45±6.41	2.11±1.41	78.06±5.27	81.50±6.19	2.37±2.14	79.33±5.14	80.66±6.37	3.01±2.78	1.7
SEAM [24]	84.40±6.21	85.05±7.04	2.95±4.38	76.36±4.82	76.50±6.85	1.90±1.10	77.79±5.64	80.11±7.05	2.07±1.49	79.52±5.56	80.55±6.98	2.31±2.32	3.8
MASSL [55]	82.11±5.76	82.57±5.83	5.03±5.21	76.03±4.78	77.71±6.38	2.39±2.05	77.56±5.98	82.36±5.83	2.25±1.30	78.57±5.51	80.88±6.30	3.22±2.85	1.5
DCR [25]	83.01±5.34	78.94±7.53	2.34±1.48	76.56±4.60	75.62±6.79	1.53±**0.42**	78.03±4.83	79.60±6.17	1.60±0.72	79.20±4.92	78.06±6.83	1.81±0.87	1.2
Ours (UNet)	82.80±5.23	84.40±6.54	2.15±1.39	76.31±4.56	78.39±**6.26**	1.56±0.44	78.08±5.41	79.37±7.15	1.86±1.03	79.06±5.00	80.72±6.65	1.86±0.95	1.9
Ours (HRNet)	**85.60**±5.01	85.29±6.29	2.34±2.58	77.27±4.71	77.10±7.15	1.65±0.79	77.88±5.10	81.81±6.52	**1.54**±0.80	80.25±4.94	81.40±6.65	1.84±1.39	1.2
∼ w/o NR	83.31±5.70	81.77±7.73	5.21±5.37	74.84±5.45	79.37±6.61	2.08±1.64	76.86±5.91	80.06±7.14	2.35±1.65	78.34±5.69	80.40±7.16	3.21±2.89	**4.2**
∼ w/o DA	85.09±5.02	84.47±6.71	2.49±2.20	77.13±4.88	77.78±7.06	1.92±1.23	77.49±5.49	81.39±6.71	2.24±1.75	79.90±5.13	81.21±6.83	2.22±1.72	2.3
∼ w/ CC	84.44±**4.88**	84.13±**6.37**	3.12±3.19	76.15±5.41	**79.52**±6.91	2.16±1.82	75.56±6.03	82.18±6.41	2.38±1.36	78.71±5.44	81.95±6.57	2.55±2.12	2.6
Ours	85.57±4.98	**86.47**±6.55	**1.69**±1.05	**78.00**±**4.25**	79.12±**6.26**	**1.37**±0.47	**78.50**±5.51	**83.17**±**5.62**	**1.54**±**0.67**	**80.69**±**4.91**	**82.92**±**6.14**	**1.53**±**0.73**	2.3



(a) Input (b) Manual (c) UNet (d) FPN (e) DCR (f) nnUNet (g) SEAM (h) DAGAN (i) Ours

Fig. 10: A visualization of the segmentation results based on different methods. The red regions denote the segmented thoracic processes, the green regions denote the segmented ribs, and the blue ones denote the segmented lumps in the lumbar region. Highlighted in the red boxes are the challenging areas around the boundary of the thoracic and lumbar regions. Highlighted in the yellow boxes are areas with the most notable differences.

bones, which restricts those methods from fully investigating the discriminative features for segmentation. Our proposed framework also surpasses those previous studies especially designed for spine segmentation, i.e., DAGAN [16] and SEAM

TABLE III: The p-value of the proposed method vs. three candidate algorithms respectively (nnUnet [54], DAGAN [16], FPN [44]).

Methods	Dice	Precision	AHD
FPN [44]	0.0001	0.0001	0.0001
nnUNet [54]	0.0001	0.0001	0.0002
DAGAN [16]	0.0001	0.0001	0.0001

[24], by approximately 1.1% and 1.4% respectively on the average Dice score, which shows its superiority in performing spine segmentation from ultrasound VPI images.

To further demonstrate the advantages of the proposed method, we visualize two samples from the testing split of the dataset to illustrate the segmentation improvement of our proposed method. The results are presented in Fig. 10. It can be seen that the baseline methods, i.e., UNet [51] and FPN [44], are significantly affected by the corrupted patterns from the scan noise, and predict unsatisfactory results in the third lumbar vertebra. Another challenging area is around the boundary of the thoracic and lumbar regions, as shown in the red boxes of Fig. 10. Apparently, the multi-task algorithm, i.e., DCR [25], the state-of-the-art medical image segmentation method, i.e., nnUNet [54], as well as the spine segmentation method, i.e., DAGAN [16], tend to predict a false alarm of lumbar vertebra at this place. SEAM [24] can produce comparable results with our proposed method to eliminate this false alarm, because SEAM [24] further considers the structure supervision in learning. Our proposed method, benefiting from the proposed selective feature-sharing joint-learning strategy, can accurately predict the spine features without structure supervision. Moreover, as the structure knowledge is inaccurate at the boundary of the images, SEAM tends to make incorrect predictions at the top and bottom regions, while our proposed method can predict more appealing segmentation masks in all regions of the image.

To statistically evaluate the significance of the obtained comparison results, we perform the paired sample t-test [56] on our dataset. Specifically, we compare the adopted evaluation metrics, i.e., Dice, Precision, and AHD, of our method and the benchmark methods, including FPN, DA-GAN, nnUNet. The obtained p-values with a threshold of 0.01 for validating the statistical significance are tabulated in Table III. The results indicate the differences in performance are statistically significant, which means our proposed method improves the spine segmentation performance over those benchmark methods.

We also compare the runtime results based on different methods as shown in the last column of Table II. Frame per second (FPS) is adopted to measure the speed for inference. Our proposed method is more efficient than the cascaded method, i.e., DAGAN [16], for jointly solving the two problems. It is also faster than the state-of-the-art medical image segmentation method, i.e., nnUNet [54]. Although its speed is lower than the baseline methods, we consider the FPS of 2.3 is sufficient for real-time clinical applications.

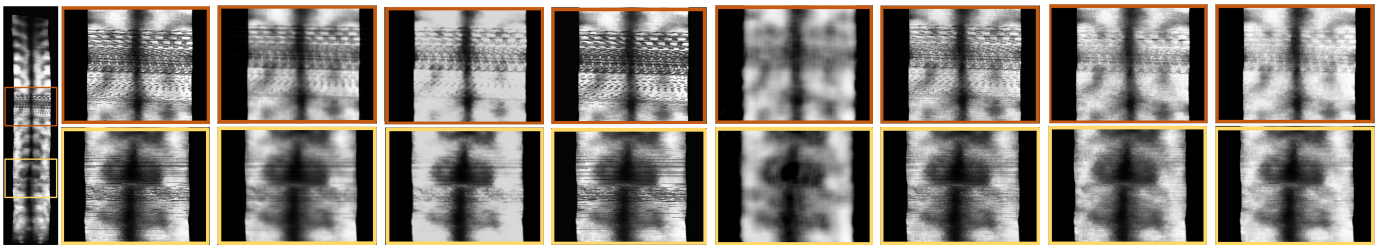
TABLE IV: Qualitative scores for user study and quantitative results for scan noise removal rate (SNRR (%)). A higher Mean Opinion Score (MOS) indicates better visual quality, and a higher SNRR indicates better noise removal performance.

Methods	Org.	N2V	DIP	CYC	GSD	NPGTV	DAGAN	Ours
MOS	5.04	4.925	1.72	5.25	3.63	2.93	6.02	**6.56**
SNRR	-	48.02	**97.75**	70.22	92.45	85.24	80.48	**97.61**

E. Results on Scan Noise Removal

To validate the effectiveness of our proposed method on scan noise removal, we performed user studies and scan noise removal rate based on the comparison with other weakly supervised denoisers that do not require the paired clean-noisy images for learning, such as Noise2Void (N2V) [37], Deep Image Prior (DIP) [57], Cycle-GAN (CYC) [58], and the total variance based denoising algorithms such as Graph-based Sinogram Denoising (GSD) [27], and non-local patch graph total variation (NPGTV) [30], as well as our previous work DAGAN [16]. In the user study, we randomly selected 15 testing images, and invited 8 ultrasound experts to assess the restored image quality. The restoration results from different methods were presented in random order. For each testing sample, the experts were asked to give a score based on its visual quality, with 6 and 1 indicating the highest and lowest quality respectively. We averaged the scores for the same method over the selected 15 images. A higher value of mean opinion score (MOS) implies better perception, and is more desirable to facilitate the scoliosis diagnosis. Before doing the user study, all the experts are asked to do a validation test (8 testing samples) to show whether the MOS of the participants in the user study is consistent with the objective metrics (e.g., SSIM and PSNR). The results and analysis of the validation test are included in the Appendix.

The results are listed in Table IV. We adopt the Mean Opinion Score (MOS) and scan noise removal rate (SNRR) to measure the quality of the images. The MOS denotes the average opinion scores from all the medical experts based on the 15 testing samples. In this user study, we also included the original images to indicate the baseline performance. It is worth noting that the original images are reconstructed from an adaptive interpolation method, i.e., the squared distance weighted (SDW) interpolation [18], which can be regarded as an interpolation baseline of 3D reconstruction methods. It can be seen from the table that DIP obtains the lowest MOS (1.72) while it gets the highest SNRR (97.75%). This implies that DIP [57] deteriorates the original image content. Although the TV-based algorithms, i.e., GSD [27] and NPGTV [30], achieve comparable performance with ours on SNRR, their MOS are even lower than that of the original image. This shows that the TV-based algorithms can remove noise but fail to preserve the content of the original image. It is seen that our proposed method produces the highest MOS of 6.56, and the second-highest SNRR of 97.61%, which implies our proposed model can more effectively restore the image content to facilitate the scoliosis assessment. More importantly, our proposed method also surpasses DAGAN [16] by about 0.5 on MOS and around 17% on SNRR, respectively. This

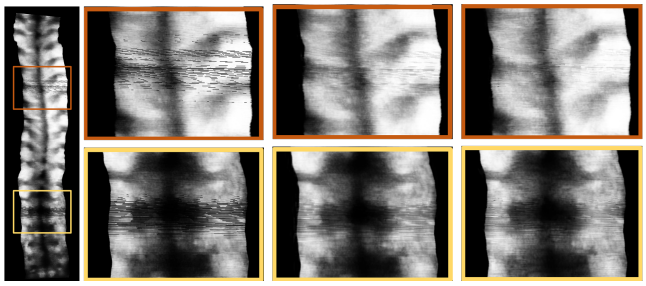


(a) Original (b) GSD (c) NPGTV (d) N2V (e) DIP (f) CYC (g) DAGAN (h) Ours

Fig. 11: Visualization of the noise removal results based on different methods.

TABLE V: Quantitative segmentation results of Folder 1 of dataset in terms of the Dice score (Dice (%)), Precision (Pre. (%)), and Average Hausdoff Distance (AHD (mm)) based on different training schemes. (Training scheme: segmentation training vs. noise removal training)

Training Scheme	Dice	Pre.	AHD
1: 3	74.42	82.43	3.65
1: 2	78.04	82.76	2.50
1: 1	80.85	83.14	2.34
2: 1	**81.27**	83.52	**1.45**
3: 1	80.21	**83.86**	2.17



(a) Original (b) ∼w/o DA (c) Full

Fig. 12: Visualization of the noise removal results for the ablation study, using the proposed full framework and the framework without dual-adversarial (DA) learning.

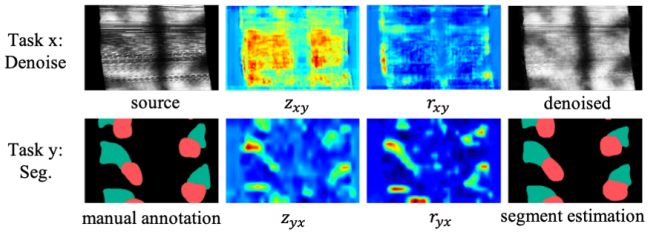


Fig. 13: Visualization of the learned gate values on one testing sample. The visualization is achieved based on the method in Grad-CAM [59].

demonstrates that the proposed joint-learning strategy, based on feature selection, can promote the scan noise removal task by transferring the beneficial supervision from segmentation.

To better show the visual quality of the restored images from different methods, we present two examples in Fig. 11, and compare our proposed methods with the other weakly supervised denoisers. It can be observed that the self-supervised methods N2V [37] and DIP [57] fail to reduce the structured noise. CYC [58] can reduce the scan noise, but the result is not satisfactory enough. GSD [27] and NPGTV [30] can remove scan noise but at the cost of smoothing the content of the original image. DAGAN [16] produces a comparable visual quality of the proposed method. When focusing on the details of the recovered area, our proposed method can recover the noisy area more effectively. Overall, our proposed method achieves the most appealing visual results.

F. Ablation Study

To validate the effect of different training schemes in the proposed framework, we tried different training schemes and tests in Folder 1 of our dataset, as shown in Table V. It is seen that the best experimental results can be obtained by training the denoising branch once and the segmentation branch twice in an iteration.

To show the effectiveness of the different designs in the proposed framework, we performed ablation studies. Specifically, we iteratively eliminate the noise removal module and the dual-adversarial strategy in the framework, denoted as "∼ w/o NR" and "∼ w/o DA" respectively, and compare the results with those of the proposed full network. It is worth noting that the model without the noise removal module becomes a single-task network for spine segmentation, and its performance should be the same as the FPN baseline. We also replace the proposed selective feature-sharing (SFS) blocks with a simple channel concatenation module, denoted as "∼ w/ CC", to validate the effect of SFS. Moreover, to verify the benefits from the auxiliary noise removal task and the flexibility of the adopted segmentation backbones, we replaced the original segmentation network, i.e., FPN, with UNet and HRNet. The results on spine segmentation are summarized in Table II. It can be seen that our joint learning framework is beneficial to the segmentation performance, as the Dice score improves about 2.3%, 2.4% and 1.5% as compared to the FPN, UNet and HRNet baseline, respectively. In addition, different feature-sharing strategies also affect the segmentation performance. We can see that the results from the model using channel concatenation for feature sharing drop by about 2% as compared to the full framework. This demonstrates the advantage of the proposed selective sharing blocks for

multi-task learning. We also investigate the proposed dual-adversarial (DA) learning. Since our proposed method is a multi-task framework, the performance of one task will also affect the performance of the other task. It is obvious that when eliminating the DA strategy, the segmentation results decrease about 0.7%. Although DA is targeted to improve the GAN-based noise removal task, it is also important to the segmentation in the proposed feature-sharing multi-task framework. We visualize a sample trained without DA in Fig. 12 to show the effect of DA on noise removal.

To clearly elaborate the ability of the proposed selective feature sharing module, we visualize the gate value obtained from a testing sample by Grad-CAM [59]. The results are shown in Fig. 13. In terms of the leaky gate r, the feature selection in the noise removal task is constructed by focusing on the noisy area that needs to be recovered. In contrast, the segmentation task pays more attention to the most informative regions associated with the spine features, aiming to learn the relationship between the source and manual annotation. Similarly, the memory gate z aims to selectively transfer the features from the previous layers. Thus, the segmentation task still focuses on the expressive spine feature region, while the noise removal task recognizes the noise feature required to be recovered. In conclusion, the gates in the selective feature sharing module can be considered as a task-based attention strategy in multi-task learning.

V. Conclusion

In this paper, we have proposed a multi-task framework with selective feature-sharing mechanism to handle both the scan noise removal and spine segmentation tasks simultaneously. Our proposed framework employs a two-stream network with the proposed feature-sharing blocks to selectively transfer features between these two tasks, which enables the framework to deliver only the beneficial features, while filtering out the useless or harmful information in task interactions. To overcome the difficulty in acquiring paired clean-noisy samples, we present a weakly supervised denoiser based on generative adversarial networks, which learns effective scan noise removal with unpaired training samples. To enhance the denoising performance, we introduce a dual-adversarial strategy to augment the sample pairs in an on-the-fly manner. We evaluate the performance of our framework in terms of both visual quality and segmentation accuracy on clinical VPI images. The results demonstrate that our method shows appealing performance on both tasks, which makes it a potential approach in clinical diagnosis.

VI. Appendix

Validation test for user study

Before performing the user study in our experiments for evaluating the visual quality of different restored images, all the experts were asked to do a validation test (8 testing samples) to show whether the MOS of the participants in the user study is consistent with the objective metrics (e.g., SSIM and PSNR). For the validation test, we employed 4 traditional image processing algorithms to the original VPI images, such as 4×4 downsampling (4DS), 8×8 downsampling (8DS), adding Gaussian noise with σ of 0.2 (GN), and average blur filtering (AB) with the filter size of 10×10. The resulting images from different algorithms (including original images) were presented in random order. According to the results of the validation test (shown in Table VI), the MOS from the participants in the user study matches well to the objective metrics, which implies that a higher MOS can generally indicate better visual quality.

Metric	Original	4DS	8DS	GN	AB
PSNR	∞	25.78	24.35	15.74	25.40
SSIM	1	0.80	0.74	0.10	0.75
MOS	4.03	3.06	2.27	2.68	2.97

TABLE VI: Quantitative results in terms of the PSNR (dB), SSIM and qualitative results in terms of Mean Opinion Score (MOS) based on different image processing algorithms.

Algorithm 1 PyTorch-like Pseudo Code of Our Training Scheme in One Iteration

```
# (I_n, I_c, y): a mini-batch of training samples
    consisting of noisy images, unpaired clean images,
    and the ground-truth segmentation masks of I_n.
# (true, fake): tensors consisting of one and zero
    values for supervising GAN-based learning.
# f: our proposed framework for joint learning noise
    removal and spine segmentation.
# (f_d_c, f_d_n): discriminators for GAN-based noise
    removal learning.

for (I_n, I_c, y) in loader:
   # Training noise removal branch
   for _ in range(2):
      # Training discriminator in noise removal branch
      I_c_hat = f(I_n)
      I_n_hat = I_c + I_n - I_c_hat
      I_c_hat.detach()
      I_n_hat.detach()
      for _ in range(3):
         for i in (c & n):
            pre_t = f_d_i(I_i)
            pre_f = f_d_i(I_i_hat)
            error = (BCEWithLogitsLoss(pre_t, true) +
                BCEWithLogitsLoss(pre_f, fake)) / 2
            error.backward()
            f_d_i.update()
      # Training generator in noise removal branch
      I_c_hat = f(I_n)
      I_n_hat = I_c + I_n - I_c_hat
      pre_c = f_d_c(I_c_hat)
      pre_n = f_d_n(I_n_hat)
      loss = BCEWithLogitsLoss(pre_c, true) +
          BCEWithLogitsLoss(pre_n, true) + alpha *
          MSELoss(I_n, I_c_hat)
      loss.backward()
      f.update()
   # Training segmentation branch
   y_hat = f(I_n)
   loss = CrossEntropyLoss(y_hat, y)
   loss.backward()
   f.update()
```

Algorithm 2 Pseudo code of noisy patch detector for quantitative measurement on noise removal task.

Initialization: The number of noisy patches: $N = 0$; The sliding window size: $s \times s$; The stride of the slide window: p; The threshold of the noisy patch: T; The position of the slide window: vertical v_s, horizontal h_s;

Output: The number of noisy patches: N

```
for all testing images img_i do
    get the width w_i and height h_i of img_i
    for v_s = 0 to (h_i − s) with stride s do
        for h_s = 0 to (w_i − s) with stride s do
            crop sub-image img_sub from img_i:
            img_sub = img_i[v_s : v_s + s][h_s : h_s + s]
            calculate the average vertical energy e_ave:
            e_ave = g(img_sub)                          ▷ (Eq. 1)
            if e_ave > T then
                N = N + 1
            end if
        end for
    end for
end for
```

REFERENCES

[1] L. Ramirez, N. G. Durdle, and V. J. Raso et al., "A support vector machines classifier to assess the severity of idiopathic scoliosis from surface topography," *IEEE Transactions on Information Technology in Biomedicine*, vol. 10, no. 1, pp. 84–91, 2006.

[2] J. COBB, "Outline for the study of scoliosis," *Instr Course Lect AAOS*, vol. 5, pp. 261–275, 1948.

[3] Y. Wong, K. K. Lai, and Y. P. Zheng et al., "Is radiation-free ultrasound accurate for quantitative assessment of spinal deformity in idiopathic scoliosis (is): A detailed analysis with eos radiography on 952 patients," *Ultrasound in Medicine & Biology*, vol. 45, no. 11, pp. 2866–2877, 2019.

[4] W. Chen, E. H. Lou, and P. Q. Zhang et al., "Reliability of assessing the coronal curvature of children with scoliosis by using ultrasound images," *Journal of Children's Orthopaedics*, vol. 7, no. 6, pp. 521–529, 2013.

[5] M. Young, D. L. Hill, and R. Zheng et al., "Reliability and accuracy of ultrasound measurements with and without the aid of previous radiographs in adolescent idiopathic scoliosis (ais)," *European Spine Journal*, vol. 24, no. 7, pp. 1427–1433, 2015.

[6] T. T. Lee, K. K. Lai, and J. C. Cheng et al., "3d ultrasound imaging provides reliable angle measurement with validity comparable to x-ray in patients with adolescent idiopathic scoliosis," *Journal of Orthopaedic Translation*, vol. 29, pp. 51–59, 2021.

[7] C. W. J. Cheung, G. Q. Zhou, and S. Y. Law et al., "Ultrasound volume projection imaging for assessment of scoliosis," *IEEE Transactions on Medical Imaging*, vol. 34, no. 8, pp. 1760–1768, 2015.

[8] P. U. Pandey, N. Quader, and P. Guy et al., "Ultrasound bone segmentation: A scoping review of techniques and validation practices," *Ultrasound in Medicine & Biology*, vol. 46, no. 4, pp. 921–935, 2020.

[9] R. C. Brink, S. P. Wijdicks, and I. N. Tromp et al., "A reliability and validity study for different coronal angles using ultrasound imaging in adolescent idiopathic scoliosis," *The Spine Journal*, vol. 18, no. 6, pp. 979–985, 2018.

[10] D. Mahapatra, B. Bozorgtabar, and R. Garnavi, "Image super-resolution using progressive generative adversarial networks for medical image analysis," *Computerized Medical Imaging and Graphics*, vol. 71, pp. 30–39, 2019.

[11] H. Zhao, Z. Ke, and N. Chen et al., "A new deep learning method for image deblurring in optical microscopic systems," *Journal of biophotonics*, vol. 13, no. 3, pp. e201960147, 2020.

[12] S. V. M. Sagheer and S. N. George, "A review on medical image denoising algorithms," *Biomedical Signal Processing and Control*, vol. 61, pp. 102036, 2020.

[13] H. S. Park, J. Baek, and S. K. You et al., "Unpaired image denoising using a generative adversarial network in x-ray ct," *IEEE Access*, vol. 7, pp. 110414–110425, 2019.

[14] S. Ishii, S. Lee, and H. Urakubo et al., "Generative and discriminative model-based approaches to microscopic image restoration and segmentation," *Microscopy*, vol. 69, no. 2, pp. 79–91, 2020.

[15] D. Auroux, "From restoration by topological gradient to medical image segmentation via an asymptotic expansion," *Mathematical and Computer Modelling*, vol. 49, no. 11-12, pp. 2191–2205, 2009.

[16] Z. Huang, R. Zhao, and F. H. F. Leung et al., "Da-gan: Learning structured noise removal in ultrasound volume projection imaging for enhanced spine segmentation," in *IEEE International Symposium on Biomedical Imaging*. IEEE, 2021, pp. 770–774.

[17] B. Lichtenbelt, R. Crane, and S. Naqvi, *Introduction to volume rendering*, Prentice-Hall, Inc., 1998.

[18] Q. H. Huang, Y. P. Zheng, and M. H. Lu et al., "Development of a portable 3d ultrasound imaging system for musculoskeletal tissues," *Ultrasonics*, vol. 43, no. 3, pp. 153–163, 2005.

[19] F. Berton, F. Cheriet, and M. C. Miron et al., "Segmentation of the spinous process and its acoustic shadow in vertebral ultrasound images," *Computers in Biology and Medicine*, vol. 72, pp. 201–211, 2016.

[20] W. Chen, L. H. Le, and E. H. Lou, "Ultrasound imaging of spinal vertebrae to study scoliosis," *Open Journal of Acoustics*, vol. 2, no. 3, pp. 95–103, 2012.

[21] T. Ungi, H. Greer, and K. Sunderland et al., "Automatic spine ultrasound segmentation for scoliosis visualization and measurement," *IEEE Transactions on Biomedical Engineering*, 2020.

[22] G. Q. Zhou, W. W. Jiang, and K. L. Lai et al., "Automatic measurement of spine curvature on 3-d ultrasound volume projection image with phase features," *IEEE Transactions on Medical Imaging*, vol. 36, no. 6, pp. 1250–1262, 2017.

[23] Z. Huang, L. W. Wang, and F. H. F. Leung et al., "Bone feature segmentation in ultrasound spine image with robustness to speckle and regular occlusion noise," in *IEEE International Conference on Systems, Man, and Cybernetics*. IEEE, 2020, pp. 1566–1571.

[24] R. Zhao, Z. Huang, and T. Liu et al., "Structure-enhanced attentive learning for spine segmentation from ultrasound volume projection images," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2021, pp. 1195–1199.

[25] J. Lyu, X. Bi, and S. Banerjee et al., "Dual-task ultrasound spine transverse vertebrae segmentation network with contour regularization," *Computerized Medical Imaging and Graphics*, vol. 89, pp. 101896, 2021.

[26] S. Banerjee, J. Lyu, and Z. Huang et al., "Light-convolution dense selection u-net (lds u-net) for ultrasound lateral bony feature segmentation," *Applied Sciences*, vol. 11, no. 21, pp. 10180, 2021.

[27] F. Mahmood, N. Shahid, and P. Vandergheynst et al., "Graph-based sinogram denoising for tomographic reconstructions," in *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2016, pp. 3961–3664.

[28] F. Mahmood, N. Shahid, and U. Skoglund et al., "Adaptive graph-based total variation for tomographic reconstructions," *IEEE Signal Processing Letters*, vol. 25, no. 5, pp. 700–704, 2018.

[29] W. Wei, B. Zhou, and D. Połap al., "A regional adaptive variational pde model for computed tomography image reconstruction," *Pattern Recognition*, vol. 92, pp. 64–81, 2019.

[30] Y. Zhang, J. Wu, and Y. Kong et al., "Image denoising via a non-local patch graph total variation," *PloS one*, vol. 14, no. 12, pp. e0226067, 2019.

[31] A. Guo, L. Fang, and M. Qi et al., "Unsupervised denoising of optical coherence tomography images with nonlocal-generative adversarial network," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–12, 2020.

[32] T. Matsui and M. Ikehara, "Gan-based rain noise removal from single-image considering rain composite models," in *European Signal Processing Conference*. IEEE, 2021, pp. 665–669.

[33] T. L. Bobrow, F. Mahmood, and M. Inserni et al., "Deeplsr: a deep learning approach for laser speckle reduction," *Biomedical optics express*, vol. 10, no. 6, pp. 2869–2882, 2019.

[34] H. Shan, Y. Zhang, and Q. Yang et al., "3-d convolutional encoder-decoder network for low-dose ct via transfer learning from a 2-d trained network," *IEEE Transactions on Medical Imaging*, vol. 37, no. 6, pp. 1522–1534, 2018.

[35] Q. Yang, P. Yan, and Y. Zhang et al., "Low-dose ct image denoising using a generative adversarial network with wasserstein distance and perceptual loss," *IEEE Transactions on Medical Imaging*, vol. 37, no. 6, pp. 1348–1357, 2018.

[36] J. Lehtinen, J. Munkberg, and J. Hasselgren et al., "Noise2noise: Learning image restoration without clean data," in *International Conference on Machine Learning*. PMLR, 2018, pp. 2965–2974.

[37] K. Alexander, B. Tim-Oliver, and J. Florian, "Noise2void-learning denoising from single noisy images," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2019, pp. 2129–2137.
[38] J. Batson and L. Royer, "Noise2self: Blind denoising by self-supervision," in *International Conference on Machine Learning*. PMLR, 2019, pp. 524–533.
[39] Y. Huo, Z. Xu, and S. Bao et al., "Adversarial synthesis learning enables segmentation without target modality ground truth," in *IEEE International Symposium on Biomedical Imaging*. IEEE, 2018, pp. 1217–1220.
[40] Z. Liu, X. Yang, and R. Gao et al., "Remove appearance shift for ultrasound image segmentation via fast and universal style transfer," in *IEEE International Symposium on Biomedical Imaging*. IEEE, 2020, pp. 1824–1828.
[41] S. Ruder, J. Bingel, and I. Augenstein et al., "Latent multi-task architecture learning," in *AAAI Conference on Artificial Intelligence*, 2019, vol. 33, pp. 4822–4829.
[42] L. Xiao, H. Zhang, and W. Chen et al., "Learning what to share: Leaky multi-task network for text classification," in *International Conference on Computational Linguistics*, 2018, pp. 2055–2065.
[43] R. Zhao, T. Liu, and J. Xiao et al., "Deep multi-task learning for facial expression recognition and synthesis based on selective feature sharing," in *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2021, pp. 4412–4419.
[44] T. Y. Lin, P. Dollár, and R. Girshick et al., "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*. IEEE, 2017, pp. 2117–2125.
[45] K. Cho, B. Van Merriënboer, and C. Gulcehre et al., "Learning phrase representations using rnn encoder-decoder for statistical machine translation," in *Conference on Empirical Methods in Natural Language Processing*, 2014, p. 1724–1734.
[46] K. He, X. Zhang, and S. Ren et al., "Deep residual learning for image recognition," in *IEEE/CVF International Conference on Computer Vision and Pattern Recognition*. IEEE, 2016, pp. 770–778.
[47] T. Miyato, T. Kataoka, and M. Koyama et al., "Spectral normalization for generative adversarial networks," in *International Conference on Learning Representations*, 2018.
[48] A. Brock, J. Donahue, and K. Simonyan, "Large scale gan training for high fidelity natural image synthesis," in *International Conference on Learning Representations*, 2018.
[49] T. Karras, M. Aittala, and J. Hellsten et al., "Training generative adversarial networks with limited data," in *Conference on Neural Information Processing Systems*, 2020.
[50] I. Loshchilov and F. Hutter, "Sgdr: Stochastic gradient descent with warm restarts," in *International Conference on Learning Representations*, 2017.
[51] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2015, pp. 234–241.
[52] J. Wang, K. Sun, and T. Cheng et al., "Deep high-resolution representation learning for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.
[53] Z. Zhou, M. M. R. Siddiquee, and N. Tajbakhsh et al., "Unet++: Redesigning skip connections to exploit multiscale features in image segmentation," *IEEE Transactions on Medical Imaging*, vol. 39, no. 6, pp. 1856–1867, 2019.
[54] F. Isensee, P. F. Jaeger, and S. A. Kohl et al., "nnu-net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature Methods*, vol. 18, no. 2, pp. 203–211, 2021.
[55] S. Chen, G. Bortsova, and A. G. U. Juárez et al., "Multi-task attention-based semi-supervised learning for medical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 457–465.
[56] H.A. David and J.L. Gunnink, "The paired t test under artificial pairing," *The American Statistician*, vol. 51, no. 1, pp. 9–12, 1997.
[57] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Deep image prior," in *IEEE/CVF International Conference on Computer Vision and Pattern Recognition*, 2018, pp. 9446–9454.
[58] J. Y. Zhu, T. Park, and P. Isola et al., "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *IEEE/CVF International Conference on Computer Vision*. IEEE, 2017, pp. 2223–2232.
[59] R. R. Selvaraju, M. Cogswell, and A. Das et al., "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *IEEE/CVF International Conference on Computer Vision*. IEEE, 2017, pp. 618–626.