UXDiff: Synthesis of X-Ray Image from Ultrasound Coronal Image of Spine with Diffusion Probabilistic Network

Yihao Zhou[1], Chonglin Wu[1], Xinyi Wang[1], and Yongping Zheng[1,2](✉)

[1] Department of Biomedical Engineering, The Hong Kong Polytechnic University, Hong Kong SAR, China
yongping.zheng@polyu.edu.hk

[2] Research Institute for Smart Ageing, The Hong Kong Polytechnic University, Hong Kong SAR, China

Abstract. X-ray radiography with measurement of the Cobb angle is the gold standard for scoliosis diagnosis. However, cumulative exposure to ionizing radiation risks the health of patients. As a radiation-free alternative, imaging of scoliosis using 3D ultrasound scanning has recently been developed for the assessment of spinal deformity. Although these coronal ultrasound images of the spine can provide angle measurement comparable to X-rays, not all spinal bone features are visible. Diffusion probabilistic models (DPMs) have recently emerged as high-fidelity image generation models in medical imaging. To enhance the visualization of bony structures in coronal ultrasound images, we proposed UX-Diffusion, the first diffusion-based model for translating ultrasound coronal images to X-ray-like images of the human spine. To mitigate the underestimation in angle measurement, we first explored using ultrasound curve angle (UCA) to approximate the distribution of X-ray under Cobb angle condition in the reverse process. We then presented an angle embedding transformer module, establishing the angular variability conditions in the sampling stage. The quantitative results on the ultrasound and X-ray pair dataset achieved the state-of-the-art performance of high-quality X-ray generation and showed superior results in comparison with other reported methods. This study demonstrated that the proposed UX-diffusion method has the potential to convert coronal ultrasound image of spine into X-ray image for better visualization.

Keywords: Diffusion model · Image-to-image translation · Image synthesis · Ultrasound imaging · Cobb angle · Scoliosis

1 Introduction

Adolescent idiopathic scoliosis (AIS), the most prevalent form of spinal deformity among children, and some patients tend to worsen over time, ultimately leading to surgical treatment if not being treated timely [1]. Determining the

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14228, pp. 3–12, 2023.
https://doi.org/10.1007/978-3-031-43996-4_1

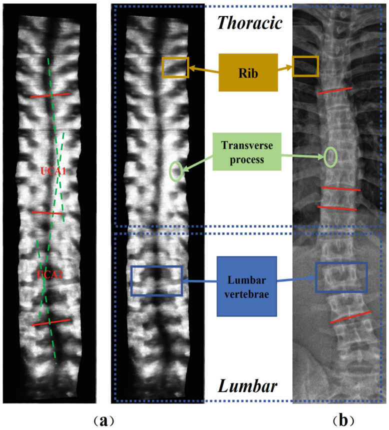


Fig. 1. Paired coronal ultrasound image with ultrasound curve angle (UCA) and X-ray image with Cobb angle. (**a**) Ultrasound images; (**b**) The corresponding X-ray image

observation interval is crucial for monitoring the likelihood of curve progression. Some scoliosis patients may progress within a short period. The gold standard for clinical diagnosis of scoliosis is the Cobb angle measured with X-ray imaging. However, the follow-up observation using X-ray requires an interval of 3–12 months because of the potential oncogenic effect of radiation exposure [2,3]. To reduce X-ray exposure and assess scoliosis frequently, coronal ultrasound imaging of spine formed by 3D ultrasound scanning has recently been developed and commercialized for scoliosis assessment [4]. This technique used a volume projection imaging (VPI) method to form the coronal ultrasound image, which contains the information of lateral curvature of spine (Fig. 1(**a**)) [5]. The ultrasound curve angle (UCA) can be obtained using the coronal ultrasound image of spine and has been demonstrated to be comparable with the radiographic Cobb angle [6]. However, clinicians hesitate to adopt this image modality since spinal images formed by VPI method are new to users, and the bone features look different from those in X-ray images (Fig. 1). If these bony features can be presented similarly to X-ray images, this radiation-free spine image will be more accepted. Moreover, such a conversion from ultrasound coronal images to X-ray-like images can not only help clinicians understand bone structure without any barrier, but also indirectly minimizes the patient's exposure to cumulative radiation.

In previous studies, generative adversarial networks (GANs) have been widely used in medical image translation applications. Long et al. proposed an enhanced Cycle-GAN for integrated translation in ultrasound, which introduced a perceptual constraint to increasing the quality of synthetic ultrasound texture [7,8]. For scoliosis assessment, UXGAN has been added with an attention mechanism into the Cycle-GAN model to focus on the spine feature during modal transformation for translating coronal ultrasound image to X-ray image [9]. However, it worked well on mapping local texture but was not so good for the translation with more extensive geometric changes. Besides, it was only tested on patients with less than 20° of scoliosis, thus limiting its application for scoliosis assessment.

Diffusion and score-matching models, which have emerged as high-fidelity image generation models, have achieved impressive performance in the medical field [10,11]. Pinaya et al. adapted denoising diffusion probabilistic models(DDPM) for high-resolution 3D brain image generation [12]. Qing et al. investigated the performance of the DPM-based model with different sampling strategies for the conversion between CT and MRI [13]. Despite the achievements of all these previous works, there is no research on the diffusion model for converting coronal ultrasound images to X-ray-like images. So far, it is still a challenging topic because the difference in texture and shape between the two modalities are substantial.

In this study, based on the conditional diffusion model, we design an ultrasound-to-X-ray synthesis network, which incorporates an angle-embedding transformer module into the noise prediction model. We have found that the guidance on angle information can rectify for offsets in generated images with different amounts of the inclination of the vertebrae. To learn the posterior probability of X-ray in actual Cobb angle conditions, we present a conditional consistency function to utilize UCA as the prior knowledge to approximate the objective distribution. Our contributions are summarized as follows:

- We propose a novel method for the Us-to-X-ray translation using probabilistic denoising diffusion probabilistic model and a new angle embedding attention module, which takes UCA and Cobb angle as the prerequisite for image generation.
- Our attention module facilitates the model to close the objective X-ray distribution by incorporating the angle and source domain information. The conditional consistency function ensures that the angle guidance flexibly controls the curvature of the spine during the reverse process.
- correlation with authentic Cobb angle indicates its high potential in scoliosis evaluation.

2 Method

2.1 Conditional Diffusion Models for U2X Translation

Diffusion models are the latent variable model that attempts to learn the data distribution, followed by a Markovian process. The objective is to optimize the usual variational bound on negative log likelihood denoted as:

$$\mathbb{E}_{q(x_0)}[\log p_\theta(x_0)] \leq \mathbb{E}_{q(x_{0:T})}[\log p_\theta(x_{0:T}) - \log q(x_{1:T}|x_0)] \tag{1}$$

$p_\theta(x_{0:T})$ is the joint distribution called the reverse process. $q(x_{1:T}|x_0)$ is the diffusion process, progressively adding Gaussian noise to the previous state of the system. Given $x \sim p(x) \in \mathbb{R}^{W \times H}$ be an X-ray image, conditional diffusion models tend to learn the data distribution in condition y. In this work, we partition y

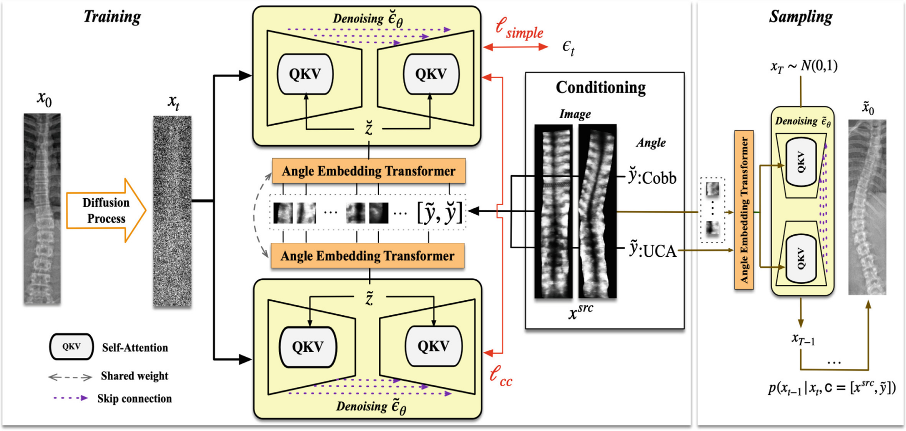


Fig. 2. *Training* The prediction models $\breve{\epsilon}_\theta$ and $\tilde{\epsilon}_\theta$ receive the corrupted image at the moment t of the diffusion process and output the noise at moment $t-1$ with an attention module for the introduction of the Cobb angle and UCA conditions, respectively; *Sampling* The embedded ultrasound images and its UCA are fed to the trained prediction model $\tilde{\epsilon}_\theta$ to generate X-ray-like images iteratively.

into the set of Cobb angle $\breve{y}$ and paired ultrasound image $x^{src} \sim p'(x) \in \mathbb{R}^{W \times H}$. Then the training objective can be reparameterized as the prediction of mean of noising data distribution with y at all timestep t:

$$\mu(x_t, t, y) = \frac{1}{\sqrt{a_t}}(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon(x_t, t, y)) \tag{2}$$

$$\ell_{simple}(\breve{\epsilon}_\theta) = \sum_{t=1}^{T} \mathbb{E}_{x_0 \sim q(x_0), \epsilon_t \sim \mathcal{N}(0,I)}[\left\| \breve{\epsilon}_\theta^{(t)}(\sqrt{a_t}x_0 + \sqrt{1-\alpha_t}\epsilon_t, y) - \epsilon_t \right\|_2^2] \tag{3}$$

2.2 Attention Module with Angle Embedding

Theoretically, if the denoising $\breve{\epsilon}_\theta$ is correct, then as $T \rightarrow \infty$, we can obtain X-ray images that the sample paths are distributed as $p_\theta(x_{t-1}|x_t, c = [\breve{y}, x^{src}])$. However, acquiring Cobb angles from real X-ray images is not feasible during the sampling stage. Accordingly, we propose an auxiliary model $\tilde{\epsilon}_\theta$, taking estimated UCA as the prior knowledge, to approximate the objective distribution. Specifically, let $\tilde{y}$ be the UCA of ultrasound images (Fig. 2). This transformer-based module establishes the relationship between $\tilde{y}$ and $\breve{y}$. We run a linear projection of dimension $\mathbb{R}^{dim}$ over the two scalars of angle, and a learnable 1D position embedding is employed to them, representing their location information. Then we reshape them to the same dimension as image tokens. A shared weight attention module compute the self-attention after taking all image tokens, and the projection of $\tilde{y}$ or $\breve{y}$ as input.

Algorithm 1. UXDiffusion Training

```
**Require**: $\mathbf{c}$: Condition information
**Optimization parameters**: $\breve{\theta}$, $\tilde{\theta}$, Angle-Embedding module $\theta_{\mathcal{A}}$
repeat
    $x_0 \sim q(x_0)$
    $x^{src}, \breve{y}, \tilde{y} \Longleftarrow$ Paired source image, Cobb angle, UCA retrieval
    $t \sim Uniform(\{1, ..., T\})$
    $x_t \Longleftarrow$ Corrupt target image $x_0$
    $\epsilon \sim \mathcal{N}(0, I)$
    Taking gradient descent step on
        $\nabla_{\breve{\theta}, \theta_{\mathcal{A}}} \left\| \epsilon - \breve{\epsilon}_\theta^{(t)}(x_t, \mathbf{c} = [x^{src}, \breve{y}]) \right\|^2$ // Eq. 3
    $\breve{\theta}$ Freeze
    Taking gradient descent step on
        $\nabla_{\tilde{\theta}, \theta_{\mathcal{A}}} \left\| \breve{\epsilon}_\theta^{(t)}(x_t, \mathbf{c} = [x^{src}, \breve{y}]) - \tilde{\epsilon}_\theta^{(t)}(x_t, \mathbf{c} = [x^{src}, \tilde{y}]) \right\|^2$ // Eq. 5
until converged
```

$$z = \begin{cases} \mathcal{A}_\theta([E(\tilde{y}); E(-1); x_{src}^1 E; \cdots x_{src}^N E;] + E_{pos}) & \text{if angle is UCA} \\ \mathcal{A}_\theta([E(-1); E(\breve{y}); x_{src}^1 E; \cdots x_{src}^N E;] + E_{pos}) & \text{else} \end{cases} \tag{4}$$

where $\mathcal{A}$ is the operation of a standard transformer encoder [14]. N is the patch number of the source image. Since the transformer can record the indexes of each token, we set the value of the patch of $\breve{y}$ to -1 when predicting the $\tilde{\epsilon}(c = [x^{src}, \tilde{y}, \breve{y} = \phi])$, and vice versa. The dimension of output z matches the noise prediction model's input for the self-attention mechanism. Thus, the sequence of the patch of the angle and image can be entered point-wise into the denoising model.

2.3 Consistent Conditional Loss

Training UXDiffusion. As described in Algorithm 1, the input to the denoising model $\tilde{\epsilon_\theta}$ and $\breve{\epsilon_\theta}$ are the corrupt image, source image and the list of angle. Rather than learning in the direction of the gradient of the Cobb angle, the auxiliary denoising model instead predicts the score estimates of the posterior probability in UCA conditions for approximating the X-ray distribution under the actual Cobb angle condition. The bound can be denoted as $KL(p_\theta(x_{t-1}|x_t, \tilde{y}, x^{src})||p_\theta(x_{t-1}|x_t, \breve{y}, x^{src}))$, where KL denotes Kullback-Leibler divergence. The reverse process mean comes from an estimate $x_\theta(\tilde{y}) \approx x_\theta(\breve{y})$ plugged into $q(x_{t-1}|x_t, y)$. Note that the parameterization of the mean of distribution can be simplified to the noise prediction [15]. The posterior probability of X-ray in the condition of Cobb angle can be calculated by minimizing the conditional consistency loss defined as:

$$\ell_{cc} = \sum_{t=1}^{T} \mathbb{E}[\left\| \epsilon_\theta^{(t)}(x_t, x^{src}, \breve{y}) - \epsilon_\theta^{(t)}(x_t, x^{src}, \tilde{y}) \right\|_2^2] \tag{5}$$

Sampling with UCA. We follow the sampling scheme in [16] to speed up the sampling process. The embedded ultrasound image and corresponding UCA are fed into the trained noise estimation network $\tilde{\epsilon}$ to sample the X-ray-like image iteratively.

3 Experiments

3.1 Dataset

The dataset consists of 150 paired coronal ultrasound and X-ray images. Each patient took X-ray radiography and ultrasound scanning at the same day. Patients with BMI indices greater than $25.0\,\mathrm{kg/m^2}$ and patients with scoliosis angles exceeding 60° were excluded from this study, as the 7.5 MHz ultrasound transducer could not penetrate well for those fatty body and the spine would deform and rotate too much with large Cobb angle, thus affecting ultrasound image quality. The Cobb angle was acquired from an expert with 15 years of experience on scoliosis radiographs, while the UCA was acquired by two raters with at least 5 years of evaluating scoliosis using ultrasound. We manually aligned and cropped the paired images to the same reference space. The criterion for registration was to align the spatial positions of the transverse and spinous processes and the ribs. We resized them into 256 × 512, and grouped the data by 90, 30 and 30 for training, validation and test, respectively.

3.2 Implementation Details

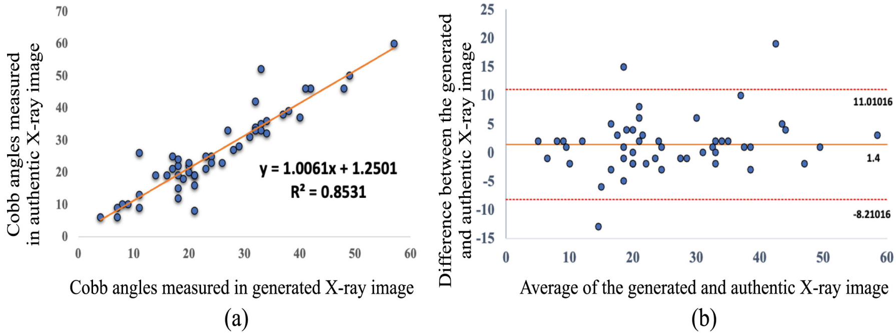
We followed the same architecture of DDPM, using a U-Net network with attention blocks to predict ϵ. For the angle embedding attention module, we transformed the token in ViT [14] for classification into the tokens for the angle list. The transformer encoder blocks were used to compute the self-attention on the embedding angle and image tokens. We set $T = 2000$ and forward variances from 10^{-4} to 0.02 linearly. We apply the exponential moving average strategy to update the model weights, the decay rate is 0.999. All the experiments were conducted on a 48GB NVIDIA RTX A6000 GPU.

3.3 Synthesis Performance

In this section, three different types of generated models were chosen for comparison: 1) GAN-based model for paired image-to-image translation [17], because of the paired data we use; 2) GAN-based model for unpaired image-to-image translation [9], the first model to synthesize X-ray image from coronal ultrasound image, is based on CycleGAN; 3) Classifier-free conditional diffusion-based model [18], which proposed a conditional diffusion model toward medical image segmentation (Table 1). We could transform the model into our U2X task by replacing the segmentation image with the ultrasound image. The visualization comparison is presented in Fig. 4, and the quantitative comparison is demonstrated. Our proposed model has a higher quality in the generation of vertebral

Table 1. Quantitative comparison. The average SSIM and PSNR scores of X-ray-like images using different methods.

Method	SSIM	PSNR (dB)
Pix2pix [17]	0.673 ± 0.05	18.44 ± 3.27
UXGAN [9]	0.715 ± 0.04	20.86 ± 2.85
MedSegDiff [18]	0.729 ± 0.04	21.02 ± 1.77
UXDiff (Ours)	0.740 ± 0.04	21.55 ± 1.69



(a)

(b)

Fig. 3. (**a**) Correlation between Cobb angles measured with the synthesized image and ground-truth X-ray images. The coefficient of determination $R^2 = 0.8531$; (**b**) Bland-Altman plot of Cobb angles for the synthesized image and ground-truth X-ray image.

contour edges compared to the baselines. Also, the synthesis images have the same spine curvature orientation as ground-true images, with high fidelity of structural information. Then, we measured the structural similarity (SSIM) and peak signal-to-noise ratio (PSNR) to evaluate the performance of the synthesized images. The results show that our model outperforms paired and unpaired GAN-based methods by 2.5–6.7 points on SSIM and is also better than the reference diffusion-based method. Our model achieves the highest value of PSNR along with the lowest standard deviation, showing the stability of the model. We believe that the diffusion-based baseline only considers the source images and disregards the scoliosis offset of the generated X-ray images, which can be addressed using angle-based guidance. Experimental results demonstrate that the performance of predicting the conditional distribution with angle guidance is superior to simply using the source image.

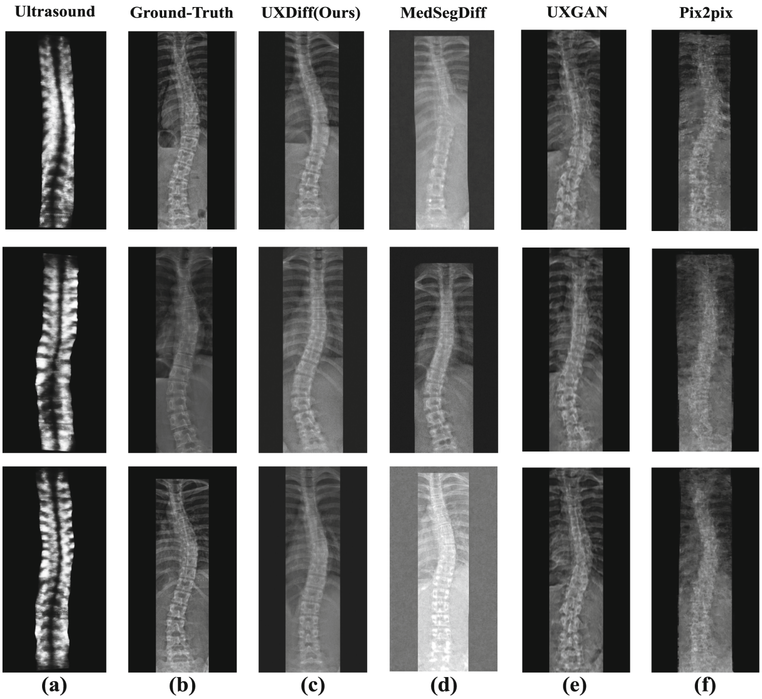


Fig. 4. Qualitative comparison of Us-to-X-ray image translation. The first two columns show the ultrasound (**a**) and ground-truth X-ray images (**b**), respectively. Our model generate a more realistic X-ray-like image that reveals the anatomical structure of spine curvature, with higher visualization quality of vertebrae than baselines (**c**–**f**).

3.4 Comparison with Cobb Angle

Since the Cobb angle is widely used as the gold standard for scoliosis assessment, our objective is to synthesize an X-ray-like image that can be applied to the measurement of the Cobb angle. As depicted in Fig. 3, we measure the Cobb angle difference between the synthesized image and the original X-ray image. Then we use linear regression to study the correlation between the Cobb angles measured using the generated and original images and the Bland-Altman plot to analyze the agreement between the two Cobb angles. The result demonstrates that our model can generate images maintaining a high consistency in restoring the overall curvature of the bones. The coefficient of determination is $R^2 = 0.8531 (p < 0.001)$. The slope of the regression line is 45.17° for all parameters, which is close to the ideal value of 45°. Bland-Altman plots demonstrate the measured angle value difference between the generated and GT X-rays. The mean difference is 1.4° for all parameters indicating that it is comparable with the ground-truth Cobb angle. The experiments demonstrate that the proposed model for calculating the Cobb angle is practical and can be used to evaluate scoliosis.

4 Conclusion

This paper developed a synthesized model for translating coronal ultrasound images to X-ray-like images using a probabilistic diffusion network. Our purpose is to use a single network to parameterize the X-ray distribution, and the generated images can be applied to the Cobb angle measurement. We achieved this by introducing the angular information corresponding to ultrasound and X-ray image to the model for noise prediction. An attention module was proposed to guide the model for generating high-quality images based on embedded image and angle information. Furthermore, to overcome the unavailability of the Cobb angle in the sampling process, we presented a conditional consistency function to train the model to learn the gradient according to the UCA for approximating the X-ray distribution in the condition of Cobb angle. Experiments on paired ultrasound and X-ray coronal images demonstrated that our diffusion-based method advanced the state-of-the-art significantly. In summary, this new model has great potential to facilitate 3D ultrasound imaging to be used for scoliosis assessment with accurate Cobb angle measurement and X-ray-like images obtained without any radiation.

Acknowledgement. This study was support by The Research Grant Council of Hong Kong (R5017-18).

References

1. Reamy, B.V., Slakey, J.B.: Adolescent idiopathic scoliosis: review and current concepts. Am. Family Phys. **64**(1), 111–117 (2001)
2. Yamamoto, Y., et al.: How do we follow-up patients with adolescent idiopathic scoliosis? Recommendations based on a multicenter study on the distal radius and ulna classification. Eur. Spine J. **29**, 2064–2074 (2020)
3. Knott, P., et al.: SOSORT 2012 consensus paper: reducing X-ray exposure in pediatric patients with scoliosis. Scoliosis **9**(1), 4 (2014)
4. Zheng, Y.-P., et al.: A reliability and validity study for Scolioscan: a radiation-free scoliosis assessment system using 3D ultrasound imaging. Scoliosis Spinal Disord. **11**, 1–15 (2016)
5. Cheung, C.-W.J., Zhou, G.-Q., Law, S.-Y., Mak, T.-M., Lai, K.-L., Zheng, Y.-P.: Ultrasound volume projection imaging for assessment of scoliosis. IEEE Trans. Med. Imaging **34**(8), 1760–1768 (2015)
6. Lee, T.T.-Y., Lai, K.K.-L., Cheng, J.C.-Y., Castelein, R.M., Lam, T.-P., Zheng, Y.-P.: 3D ultrasound imaging provides reliable angle measurement with validity comparable to x-ray in patients with adolescent idiopathic scoliosis. J. Orthop. Transl. **29**, 51–59 (2021)
7. Teng, L., Fu, Z., Yao, Y.: Interactive translation in echocardiography training system with enhanced cycle-GAN. IEEE Access **8**, 106147–106156 (2020)
8. Zhu, J.-Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2223–2232 (2017)

9. Jiang, W., Yu, C., Chen, X., Zheng, Y., Bai, C.: Ultrasound to X-ray synthesis generative attentional network (UXGAN) for adolescent idiopathic scoliosis. Ultrasonics **126**, 106819 (2022)
10. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Adv. Neural. Inf. Process. Syst. **33**, 6840–6851 (2020)
11. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. arXiv preprint arXiv:2011.13456 (2020)
12. Pinaya, W.H.L., et al.: Brain imaging generation with latent diffusion models. In: Mukhopadhyay, A., Oksuz, I., Engelhardt, S., Zhu, D., Yuan, Y. (eds.) DGM4MICCAI 2022. LNCS, vol. 13609, pp. 117–126. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-18576-2_12
13. Lyu, Q., Wang, G.: Conversion between CT and MRI images using diffusion and score-matching models. arXiv preprint arXiv:2209.12104 (2022)
14. Dosovitskiy, A., et al.: An image is worth 16×16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
15. Nichol, A.Q., Dhariwal, P.: Improved denoising diffusion probabilistic models. In: International Conference on Machine Learning, pp. 8162–8171. PMLR (2021)
16. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
17. Isola, P., Zhu, J.-Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1125–1134 (2017)
18. Wu, J., Fang, H., Zhang, Y., Yang, Y., Xu, Y.: MedSegDiff: medical image segmentation with diffusion probabilistic model. arXiv preprint arXiv:2211.00611 (2022)