PLOS ONE

RESEARCH ARTICLE

# Screening of adolescent idiopathic scoliosis using generative adversarial network (GAN) inversion method in chest radiographs

**Jun Soo Lee**[1☯], **Keewon Shin**[2☯], **Seung Min Ryu**[2,3☯], **Seong Gyu Jegal**[2], **Woojin Lee**[4], **Min A. Yoon**[5], **Gil-Sun Hong**[5], **Sanghyun Paik**[4], **Namkug Kim**[6*]

**1** Department of Industrial Engineering, Seoul National University, Seoul, Korea, **2** Department of Biomedical Engineering, Asan Medical Institute of Convergence Science and Technology, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Republic of Korea, **3** Department of Orthopedic Surgery, University of Ulsan College of Medicine, Asan Medical Center, Seoul, Korea, **4** Department of Radiology, Hanyang University Hospital, Seoul, Korea, **5** Department of Radiology, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Korea, **6** Department of Convergence Medicine, Asan Medical Institute of Convergence Science and Technology, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Republic of Korea

☯ These authors contributed equally to this work.
* namkugkim@gmail.com





**Citation:** Lee JS, Shin K, Ryu SM, Jegal SG, Lee W, Yoon MA, et al. (2023) Screening of adolescent idiopathic scoliosis using generative adversarial network (GAN) inversion method in chest radiographs. PLoS ONE 18(5): e0285489. https://doi.org/10.1371/journal.pone.0285489

**Editor:** Muhammad Fazal Ijaz, Sejong University, REPUBLIC OF KOREA

**Received:** January 27, 2023

**Accepted:** April 25, 2023

**Published:** May 22, 2023



**Data Availability Statement:** All relevant data are within the paper and its Supporting information files.

**Funding:** SMR This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2021R1A6A3A01088445). NO - Include this sentence at the end of your statement: The funders had no role in study design, data collection and

## Abstract

### Objective

Conventional computer-aided diagnosis using convolutional neural networks (CNN) has limitations in detecting sensitive changes and determining accurate decision boundaries in spectral and structural diseases such as scoliosis. We devised a new method to detect and diagnose adolescent idiopathic scoliosis in chest X-rays (CXRs) employing the latent space's discriminative ability in the generative adversarial network (GAN) and a simple multi-layer perceptron (MLP) to screen adolescent idiopathic scoliosis CXRs.

### Materials and methods

Our model was trained and validated in a two-step manner. First, we trained a GAN using CXRs with various scoliosis severities and utilized the trained network as a feature extractor using the GAN inversion method. Second, we classified each vector from the latent space using a simple MLP.

### Results

The 2-layer MLP exhibited the best classification in the ablation study. With this model, the area under the receiver operating characteristic (AUROC) curves were 0.850 in the internal and 0.847 in the external datasets. Furthermore, when the sensitivity was fixed at 0.9, the model's specificity was 0.697 in the internal and 0.646 in the external datasets.

analysis, decision to publish, or preparation of the manuscript.

**Competing interests:** The authors have declared that no competing interests exist.

## Conclusion

We developed a classifier for Adolescent idiopathic scoliosis (AIS) through generative representation learning. Our model shows good AUROC under screening chest radiographs in both the internal and external datasets. Our model has learned the spectral severity of AIS, enabling it to generate normal images even when trained solely on scoliosis radiographs.

## Introduction

Adolescent idiopathic scoliosis (AIS) is the most common spinal deformity [1] and is defined as a 10˚ or more spinal curvature of unknown etiology in persons 10 to 18 years old [2]. AIS has an overall 0.47% to 5.2% prevalence [3]; thus, screening in school-aged adolescents is imperative because early detection can reduce the need for surgery through non-surgical management such as bracing [4]. However, Cobb's angle measurements conveyed a 4˚ to 8˚ intra- and inter-observer variability [5, 6], with one study revealing a potential maximum inter-observer measurement error up to 11.8˚ [7]. In addition, Cobb's angle manual measurement is labor-intensive and time-consuming [8]. Therefore, it would be conducive to authentic clinical practice if Cobb's angle was only measured in suspected scoliosis patients, which can be completed during health checkups.

Thus, many studies actively research AIS diagnosis using deep learning models with convolutional neural networks (CNN). For example, one study developed an automated scoliosis screening algorithm using a deep learning model for naked back photos [2]. Furthermore, as several segmentation studies incorporate deep learning [9], some authors have developed an automated Cobb's angle measurement algorithm using the detection or segmentation provided by another deep learning model [10–12]. However, supervised learning will inevitably function poorly with external dataset images. Moreover, CNN is affected by image textures [13], so it may not be preferable for diagnosing spectrum disorders that affect the global spinal structure, such as scoliosis. Therefore, a method that can withstand continuous or spectral progression would complement the clinical field's discriminative diagnosis tendency, which inevitably binarizes patients with borderline symptoms into the 'normal' category.

In this regard, we noticed an alternative machine learning field trend: finding a lower dimensional data representation with preferable properties such as discernibility through data learning distribution. Specifically, we developed a deep learning AIS diagnosis model with classification features that are latent vectors acquired from query images extracted using GAN inversion. In this feature-extracting process, we incorporated GAN trained from an imbalanced dataset that does not include standard counterpart samples to maximize the differentiating ability. Therefore, we conducted an ablation study to find and evaluate a model that exhibits optimal performance and diagnostic power. The primary purposes of the study are:

- Empirically proving GAN's ability to generate normal images through partial normality, provided in training, set with symptoms.
- Developing a novel method for differentiating given data using GAN as a feature extractor.

Our method is expected to detect spectral disease progressions with high sensitivity.

## Literature review

Representation learning is a machine learning method allowing a system to learn feature identification from a substantial amount of unlabeled data [14–16]. While most works learn data

distribution representation using a discriminative objective, some studies have utilized a generative learning approach that generates or models pixels in the input space [17–19]. Multiple investigations have attempted to interpret the relationship between data and its latent representations in generative models. Arora et al. [20] proposed that using a generative adversarial network (GAN) as a feature learner is a practical approach due to its low support in GANs-learned distribution; they also provided a theoretical background for this approach. Salimans et al. [21] proposed semi-supervised learning through discriminator modification to classify K-classes and a virtual class for detecting fake samples. On the other hand, Srinivasu et al. [22] injected information from the generative model in tackling discriminative tasks by using auto-generated images reconstructed using variational autoencoders [18] as additional data for model training. He et al. proposed a novel unsupervised learning approach using masked auto-encoders to reconstruct inputs from partially masked versions of themselves [23]. This allows the model to learn suitable representations for downstream tasks, particularly in computer vision applications. While their approach is purely unsupervised in the upstream phase, our proposed method leverages semi-supervised learning on a dataset with abnormal labels, making it somewhat similar to He et al.'s approach.

These studies displayed the data distribution and learned latent space correlation, thus suggesting a similar correlation in a specific data sample-latent vector pair. However, the above-mentioned methods are limited as they require extra architectural modules or modifications. Thus, methods that manipulate the latent vector were introduced. StyleGAN [24] introduced intermediate latent space W, which is more disentangled semantically, further boosting latent vector investigations. SeFa [25] provided an unsupervised algorithm to identify dominant directions in the latent space.

The present study uses the GAN inversion method for generative representation learning on a relatively small-sized image dataset. GAN inversion discovers a code in the GAN-trained latent space, generating the best reconstruction of a given query image. The most direct and intuitive GAN inversion method is the optimization-based method proposed by Abdal et al. [26]. This optimization-based method tries to find $z,^*$ which satisfies the following Eq (1):

$$z^* = \text{argmin}_z \, \ell(\text{G}(\text{z}), \, \text{x}), \tag{1}$$

where ℓ(image, image) is a predefined similarity metric, x is the given query image, and G(z) is the latent vector generated image. $z^*$ corresponds to the latent manifold vector that resembles the given query image the most. In this sense, we propose an alternative novel method utilizing features embedded in an intermediate manifold. Specifically, the vector acquired using the generative adversarial network (GAN) inversion [26] is equivalent to the discriminative method' extracted feature as the inherent information about the given images is embedded in the acquired vector. This feature extraction differs from conventional models that utilize CNNs as CNN models craft features by accumulating information acquired from convolution operations on image patches. In contrast, our method extracts more general, large-scale features from the whole image.

## Materials and methods

### 1 Dataset and preparation

This retrospective study followed the Declaration of Helsinki principles [27]. The study protocol was approved by our institution's (Center 1) Institutional Review Board Committees and other institutions (Center 2), which waived the need for written informed patient consent.

**1.1 Dataset.** Fig 1 depicts a data collection and split schematic. We chronologically split the training and validation datasets by using a portion of the data obtained between 1997 and

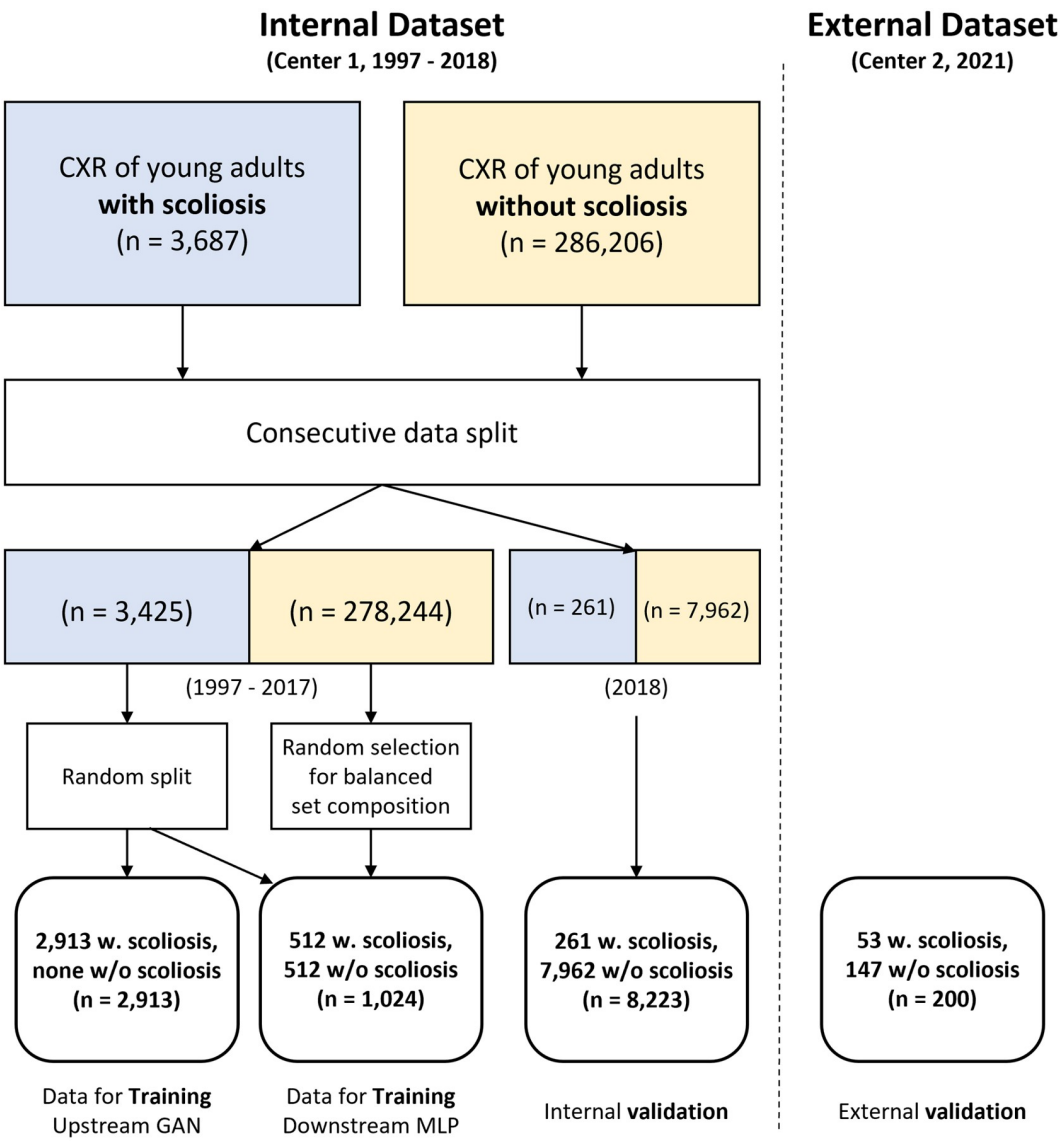


**Fig 1. A schematic diagram of collected data.**

https://doi.org/10.1371/journal.pone.0285489.g001

2017 as the training data and all data obtained in 2018 as the validation data. When composing the training dataset, all chest X-rays (CXRs) with AIS were used, and some normal counterpart CXRs were randomly selected to form a balanced dataset. As our model requires two-stage training, we randomly split the collected CXRs with AIS into two datasets: upstream GAN and downstream multi-layer perceptron (MLP). The diagnostic threshold was set as 20 degrees considering that close observation is generally recommended for patients with an initial Cobb's angle of fewer than 20 degrees [28]. An orthopedic surgeon with eight years of clinical experience did the angle measurement. The exclusion criteria were cases post-spine surgery or of younger children. As a result, we obtained three datasets.

The first dataset collected 2,913 CXRs of young adults with scoliosis to train the upstream GAN, which was used for feature extraction. The second dataset comprised 1,024 CXRs of young adults collected at Center 1 with a 1:1 normal-to-scoliosis ratio for training the

**Table 1. Detailed demographic information of the collected datasets.**

| | Center 1 | | | Center 2 |
|---|---|---|---|---|
| | Upstream GAN | Downstream MLP | Internal validation | External validation |
| Radiograph quantity | 2,913 | 1,024 | 8,223 | 200 |
| Scoliosis | 2,913 | 512 | 261 | 53 |
| Normal | - | 512 | 7,962 | 147 |
| Age | 14.40 ± 2.25 | 14.90 ± 2.26 | 14.40 ± 2.68 | 18.64 ± 4.17 |
| Scoliosis | 14.40 ± 2.25 | 14.45 ± 2.23 | 14.59 ± 2.01 | 22.42 ± 3.29 |
| Normal | - | 15.34 ± 2.21 | 14.39 ± 2.70 | 17.27 ± 3.58 |
| Sex (M/F/O) | 754/1473/686 | 354/407/263 | 4720/3481/22 | 69/131/0 |
| Scoliosis | 754/1473/686 | 135/262/115 | 63/197/1 | 10/43/0 |
| Normal | - | 219/145/148 | 4657/3284/21 | 59/88/0 |
| Collection year(s) | 1997–2017 | 1997–2017 | 2018 | 2021 |

M, Male; F, Female; O, Others (anonymized data). Age is presented as mean ± standard deviation.

https://doi.org/10.1371/journal.pone.0285489.t001

downstream MLP to detect scoliosis. The final dataset constituted 8,223 CXRs of young adults collected at Center 1; 261 had scoliosis, and 7,962 had no noticeable disease. For external validation, we collected 53 CXRs with scoliosis and 147 CXRs without scoliosis from Center 2 in 2021. In addition, after the study began, the most recent 200 cases were selected retrospectively and externally validated. More detailed demographic information about the collected dataset is provided in Table 1.

**1.2 Image preprocessing.** As our task focused on global skeletal orientation and not fine features on soft tissues, image preprocessing was to enhance the bone area and preserve the aspect ratio. First, we checked the image spacing information to preserve the image aspect ratio if the image was distorted due to different vertical/horizontal spacing ratios. Then, we applied CLAHE [29, 30], an equalizing histogram technique to improve contrast on image patches, with a 2.0 clipLimit and (8,8) tileGridSize on input images to differentiate vertebral bones from soft tissues, especially in the lumbar vertebral area. To introduce the CXR into the model, resizing to 512 × 512 resolution was required. However, naive resizing would change the aspect ratio, unintendedly altering Cobb's angle, which is the gold standard for indicating scoliosis. Therefore, we added padding to the image so the original image's aspect ratio would not change, even when resized. Next, we stacked the processed black-and-white image channel-wise to mirror the same shape as natural RGB images.

## 2 Methods

**2.1 Training GAN with the diseased dataset.** Fig 2 illustrates our proposed method. The training dataset usually includes a natural distribution for maximum generated-image diversity for the training GAN. However, data scarcity results in low-quality images with minimal intra-class variation. Shahbazi et al. noted that this tendency depreciates the conditional training [31]. Considering our objective and dataset size, we trained our network using only CXRs expressing some scoliosis degree. Using the scoliosis classification criteria set by Goldstein [32] and Cruickshank [33], curve patterns were determined by observing where the curve apex exists. From this standard, we noted that even in severe scoliosis cases, some parts could be diagnosed as normal in focal view. For example, local thoracic spine observation could determine little difference in a CXR with severe scoliosis on the lumbar spine from the thoracic spine without scoliosis. Therefore, we hypothesized that we could more effectively embed the

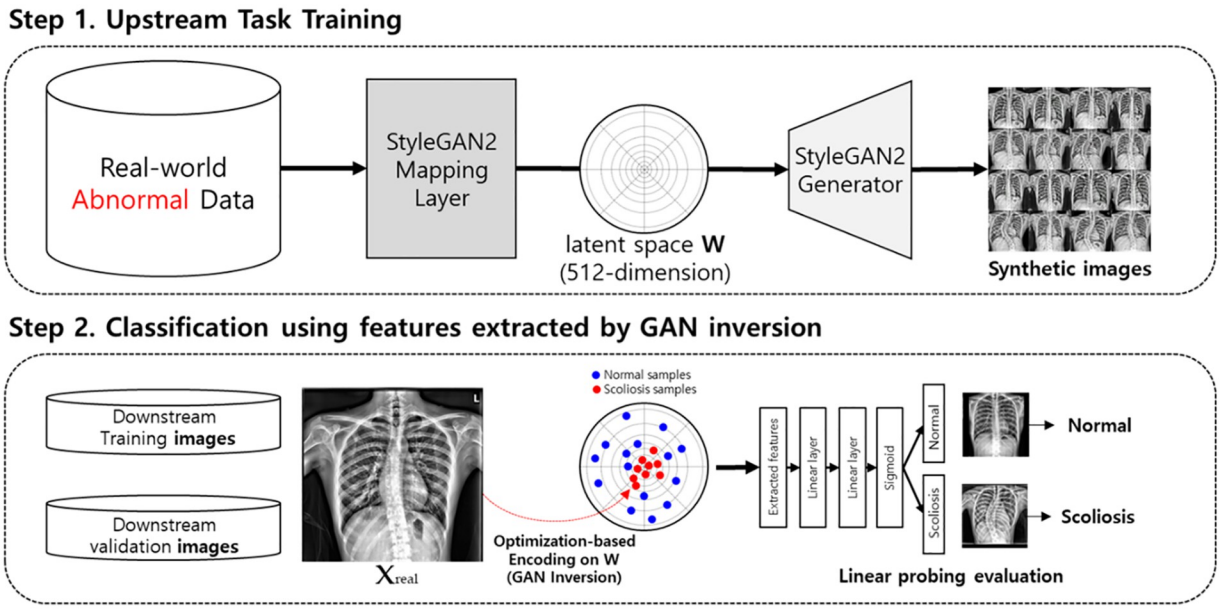


**Fig 2. Training strategy for developing sensitive classification networks using GAN inversion as a feature-extracting method.** The first step is an upstream task that trains the generative adversarial network (GAN). The second step classifies through linear probing and projection combination using GAN inversion.

https://doi.org/10.1371/journal.pone.0285489.g002

disease axis into the model feature space without degrading the normal distribution too much by using an imbalanced dataset. An empirical analysis justifying our method is provided in the discussion section.

**2.2 GAN inversion for feature extraction.** We conducted optimization-based method experiments considering data scarcity and embedded our vector into the original W space instead of the extended W+ space. Karras et al. [24] and Yang et al. [34] established that different semantics are hierarchically determined from different resolution layers. Therefore, embedding vectors in a fine-grained manner would benefit fine feature reconstruction, which was not the primary goal of this study. Next, we suspected embedding into W+ space would bring excessive computational cost as it has a much higher dimension than W space.

**2.3 Evaluation.** Supervised learning combined latent codes (through GAN inversion-extracted features) and scoliosis presence based on the original image. The projection head evaluated GAN's latent space discriminative ability [35], similar to the widely used linear probing method [36] in self-supervised learning evaluations. The projection head is a 2-layer MLP; hidden MLP layers were 512 dimensions and used ReLU [37], and output MLP layers were one dimension with a sigmoid. None of the MLP layers contained batch normalization [38].

First, each performance was analyzed using a likelihood value threshold between 0 and 1 extracted by the MLP layers. We evaluated the 95% confidence interval (CI) and the area under the receiver operating characteristics curve (AUROC) to determine whether the model performance was significantly better. Since this study aimed to screen AIS in a real-world setting, we fixed the sensitivity at 0.9 and calculated the true-positive (TP), false-positive (FP), true-negative (TN), and false-negative (FN) values. In addition, the quantitative classification assessment included accuracy, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV). Finally, an orthopedic surgeon with eight years of clinical experience visually analyzed the FP/FN cases to evaluate the method's performance. We calculated

the accuracy, sensitivity, specificity, PPV, and NPV to evaluate the method's efficiency in a clinical AIS diagnosis scenario. We compared performance by adjusting our model's threshold relative to screening purposes.

Second, we conducted an ablation study to classify performance relative to MLP stack numbers. In addition, a data stress test was conducted to confirm classification performance relative to latent code numbers used for the downstream training phase.

Third, we investigated whether different pre-trained weights affected the feature extraction performance using the GAN inversion method by evaluating the downstream classification performances corresponding to different weights.

Fourth, while we hypothesized that normal images could be embedded well into the latent space formed by learning the abnormal image distribution, it is against AnoGAN's widely-accepted philosophy [39]. Therefore, to determine that our trained latent space has expanded enough to generate normal images, we conducted qualitative and quantitative analyses on our approach's ability to embed normal images.

**2.4 Training configurations.** We utilized StyleGAN-ADA [40] architecture for training the upstream GAN network. We inherited implementation details concerning Pytorch [41] implementation without modifications from the study's authors [40]. The training data was preprocessed following the method mentioned in *Section 3.1.2*. The input data augmentation to the discriminator during styleGAN2-ADA training was performed with maximal provided pipeline combinations. However, we excluded some augmentations not applicable to medical deep learning, such as horizontal flip or cutout. We used a non-saturating loss [19] with R1 regularization [42], utilizing 6.5536 as the coefficient value for the loss function. Finally, we used the ADAM [43] optimizer with a 0.002 learning rate. Next, we used the Frechet inception distance (FID) [44] on the total training dataset to evaluate the upstream network's convergence. We selected epochs with the lowest FID value after training for 1 million iterations, so the FID conveys convergence.

We used a simple loss for feature extraction that minimizes the L2 norm of the given query and generated images with a noise regularization term. Then, we iterated 1000 times to extract the final vector without weight update to the GAN generator. For classifying the extracted vectors, we used binary cross-entropy as a loss to train the binary classifier. The downstream training set's normal-to-abnormal ratio was set to 1:1, and the training data in each dataset started from 32 and doubled up to 1024. The model was trained for 200 epochs with complete batch learning, and the learning rate was set to 0.001 in the Adam optimizer.

## Results

### 1. Classification result

The threshold was set at a 0.9 sensitivity in the screening setting, and our model's internal and external validation results are organized in Table 2. When the sensitivity was set to 0.9 in the internal and external datasets, the specificity was 0.697 and 0.646, respectively. There were 25 FN cases of our model in the internal test dataset and 5 FN cases in the external test dataset, which are shown in Fig 3.

### 2. Ablation study

The AUROC evaluation results relative to the layer numbers in the projection head and downstream training samples are summarized in Fig 4. This evaluation incorporated a linear protocol to evaluate the classification performance. As sample quantities increased, the AUROC tended to improve. When the projection head layer number was 2 and the downstream

**Table 2. Our final model's scoliosis classification performance.**

| Test dataset | Performance measure | | | | | |
|---|---|---|---|---|---|---|
| | AUROC | ACC | SEN | SPE | PPV | NPV |
| **Internal** | 0.850 | 0.704 | 0.9* | 0.697 | 0.096 | 0.995 |
| **External** | 0.847 | 0.715 | | 0.646 | 0.480 | 0.950 |

AUROC, area under the receiver operating characteristic curve; ACC, accuracy; SEN, sensitivity; SPE, specificity; PPV, positive predictive value; NPV, negative predictive value

https://doi.org/10.1371/journal.pone.0285489.t002


**Fig 3. False-negative case examples in the (A) internal test and (B) external validation datasets.**

https://doi.org/10.1371/journal.pone.0285489.g003

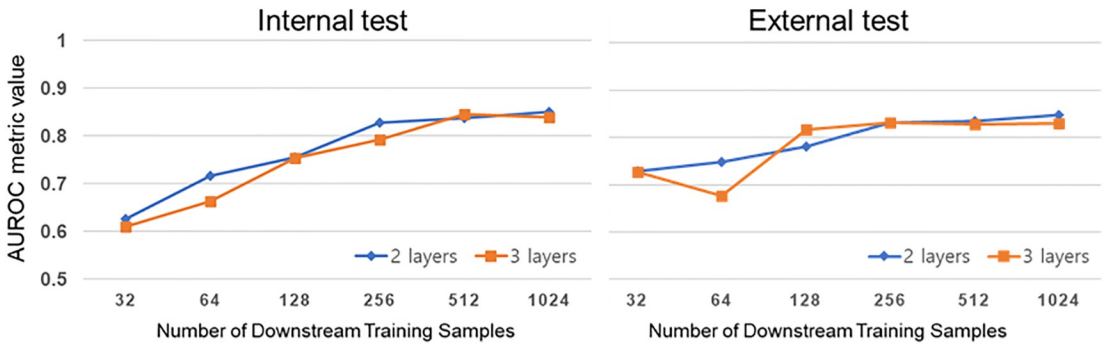


**Fig 4. Scoliosis classification performance within the number of projection head layers and downstream training samples.**

https://doi.org/10.1371/journal.pone.0285489.g004

training sample was 1024, internal and external AUROCs were 0.850 and 0.847, respectively, indicating the highest classification performance.

## 3. Pre-trained weight effect of the upstream task

We evaluated performance relative to pre-trained weights, and the results are summarized in Fig 5 and Table 3. Internal and external AUROCs were 0.850 and 0.847 for scratch training, respectively; AUROCs with Flickr-Faces-HQ (FFHQ) [24] pre-training weights were 0.868 and 0.828, and medical transfers were 0.858 and 0.845, respectively. When considering scratch and FFHQ pre-trained weights, the downstream classification result differences were not statistically significant. For FFHQ, internal AUROC increased by 0.028 compared with scratch, whereas external AUROC decreased by 0.019. Lastly, scratch training and medical transfer pre-trained weights were used to analyze the downstream AUROC. Internal AUROC increased by 0.008, and external AUROC decreased by 0.002. Similarly, there was no statistically significant difference.

## 4. Quantitative analysis for method validation

We calculated query and projected image peak signal-to-noise ratios (PSNR), structural similarity index measures (SSIM) [45], and root mean square errors (RMSE) to quantitatively compare how well images embedded into the latent space. Based on the ground truth, we calculated these three metrics on every sample in the downstream training set. Table 4 shows the average image-reconstruction quality metric values measured on scoliosis and the normal

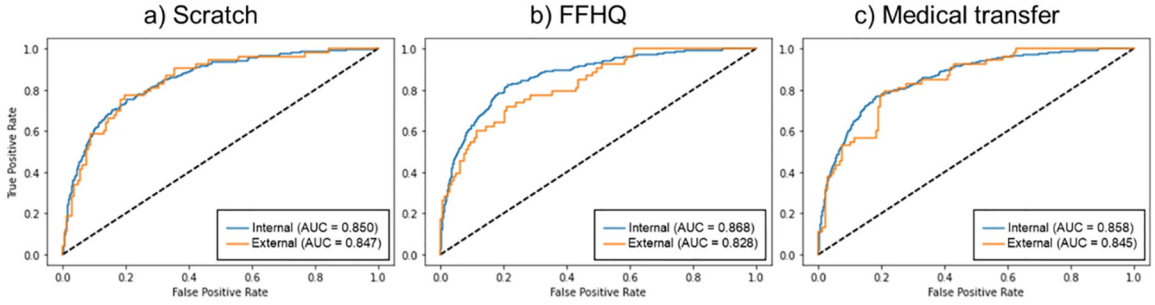


**Fig 5. Downstream validations relative to upstream pre-trained weight type.**

https://doi.org/10.1371/journal.pone.0285489.g005

**Table 3. Performance metric comparison under a fixed 0.9 sensitivity and statistical analysis using independent t-test among different feature extractor weights.**

| | Scratch | | FFHQ | | Medical | |
|---|---|---|---|---|---|---|
| | Internal | External | Internal | External | Internal | External |
| **95% CI** | 0.842–0.858 | 0.789–0.894 | 0.833–0.849 | 0.768–0.878 | 0.825–0.841 | 0.787–0.892 |
| **AUROC** | 0.850 | 0.847 | 0.868 | 0.828 | 0.858 | 0.845 |
| **Sensitivity (fixed)** | 0.903 | 0.906 | 0.904 | 0.906 | 0.897 | 0.906 |
| **Specificity** | 0.585 | 0.646 | 0.581 | 0.503 | 0.580 | 0.585 |
| **PPV** | 0.066 | 0.480 | 0.066 | 0.397 | 0.065 | 0.440 |
| **NPV** | 0.995 | 0.950 | 0.995 | 0.937 | 0.994 | 0.945 |
| Statistical analysis (*P* values of independent t-test) | | | | | | |
| | **Scratch vs. FFHQ** | | **Scratch vs. Medical** | | **FFHQ vs. Medical** | |
| **Internal** | 0.636 | | 0.383 | | 0.447 | |
| **External** | 0.534 | | 0.950 | | 0.588 | |

CI: confidence interval, AUROC: area under the receiver operating characteristic curve, PPV = positive predictive value. NPV = negative predictive value.
Scratch refers to training without pre-trained weight. FFHQ refers to training pre-trained weight using Flickr Faces HQ Dataset. Medical refers to training pre-trained weight using chest radiograph.

https://doi.org/10.1371/journal.pone.0285489.t003

downstream training set. Images with scoliosis were better reconstructed according to SSIM and RMSE metrics, whereas images without scoliosis were better reconstructed with PSNR metrics. Finally, we selected a well-reconstructed normal image example using vectors from the GAN inversion method to further corroborate these results (Fig 6).

## Discussion

This study evaluated a novel deep learning model's AIS diagnosing accuracy using latent vectors acquired from query images using GAN inversion as features. Based on the best performance in the dataset, internal and external dataset AUROCs were 0.850 and 0.847, respectively. Furthermore, we provide internal and external dataset ROC curves in Fig 5a. Our method indicated good generalizability because the AUROC value did not degrade drastically when tested on an external dataset. Therefore, our model is a potential tool in practical AIS screening. In the results depicted in Table 2, the sensitivity was fixed at 0.9 because this model's primary purpose was for AIS screening. Despite the specificity result being inevitably lower than sensitivity, the internal and external dataset specificities were 0.697 and 0.646, respectively. Thus, we believe this model may be preferable for real-world use.

We used a toy comparison experiment using a balanced dataset to justify using an imbalanced dataset as the training dataset. Fig 7 illustrates the generated CXR images that only differ in the training dataset composition, indicating that samples generated from the model trained

**Table 4. Image-reconstruction quality metrics of scoliosis and normal downstream training sets.**

| Metrics / Data | PSNR | SSIM | RMSE |
|---|---|---|---|
| **Scoliosis** | 19.048 ± 2.110 | 0.464 ± 0.108* | 8.659 ± 0.742* |
| **Normal** | 19.398 ± 1.313* | 0.432 ± 0.059 | 8.887 ± 0.505 |
| †***P* value** | 0.002 | <0.001 | <0.001 |

†Paired t-test for comparing image reconstruction quality metric among disease presence.

*superior data among scoliosis or normal dataset.

PSNR, peak signal to noise ratio; SSIM, structural similarity index measure; RMSE, root mean square error.

https://doi.org/10.1371/journal.pone.0285489.t004

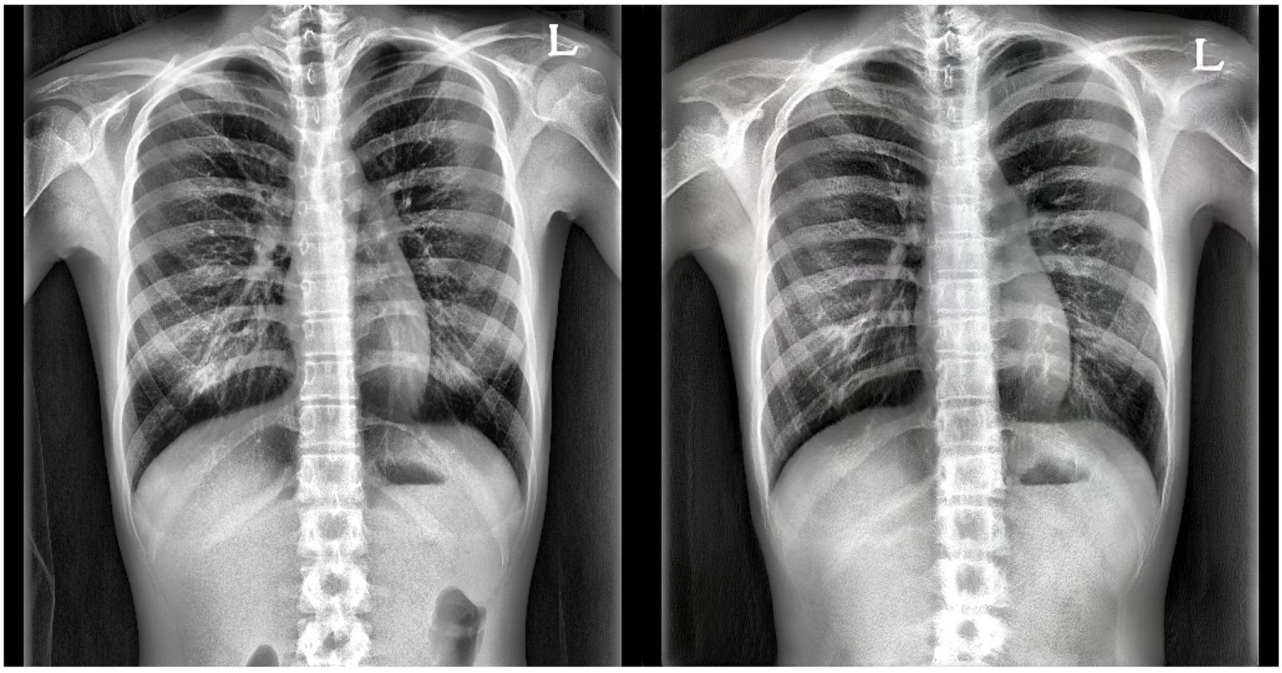

**Fig 6. Example of a well-reconstructed normal sample.** (Left) original chest X-ray. (Right) Reconstructed image using the GAN inversion technique. The encoding generator was only trained with chest X-ray images with scoliosis.

https://doi.org/10.1371/journal.pone.0285489.g006

with an imbalanced dataset have much more diversity regarding scoliosis severity and location. Furthermore, a sample that can be diagnosed as normal (red box in Fig 7) was also included in generated samples, further supporting our hypothesis that the latent space can be expanded to generate normal samples even when trained on a scoliosis-only dataset. On the other hand,

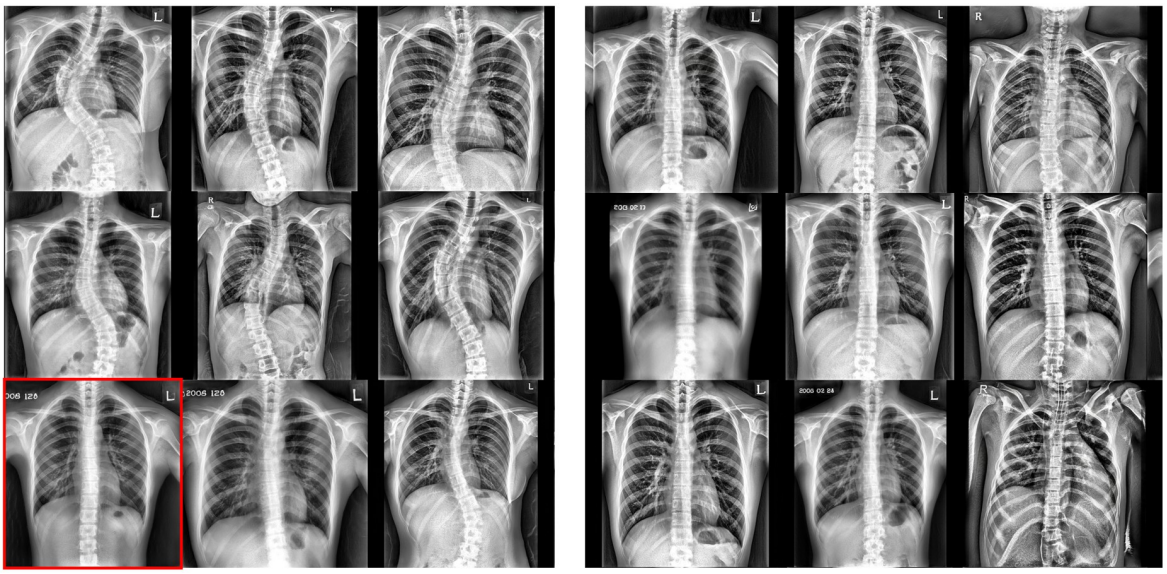

**Fig 7. Generated CXR samples.** (Left) Samples from a model trained on an imbalanced dataset (our method). (Right) Samples from a model trained on a balanced dataset. Red box: a sample radiograph that can be diagnosed as a normal spine in generated samples.

https://doi.org/10.1371/journal.pone.0285489.g007

samples generated from the model trained on a balanced dataset express little to no diversity in scoliosis scope, where all generated samples have "straight" spines.

The internal 25 FN out of 261 scoliosis case validation analysis demonstrated that 36% of cases were levoscoliosis (Fig 3). In addition, an American Academy of Family Physicians review estimated that 10 to 15% of adolescents with scoliosis had left curves or levoscoliosis [46]. Therefore, the FN cases are assumed as incorrect because levoscoliosis radiographs were rarely included during training. Since levoscoliosis can be recognized as a negative Cobb's angle from the SeFa factorization perspective [25], where we theorize dextroscoliosis progression as the prominent data distribution variation, they occupy a part of the supernormal, not both scoliosis and normal. However, it is postulated that these false cases occurred because dextroscoliosis and levoscoliosis were not distinguished.

We also checked FP cases and found that some external devices were visible in 1,838 of the 3,297 cases, including Hickman catheters, chemo ports, vital sign monitor lines, and cardiovascular devices, such as cardiac pacemakers or implantable cardioverter-defibrillators. We assume that tube curves were incorrectly precepted as a bent spine since we designed our model to be sensitive to structural changes. Therefore, we manually excluded all cases with external devices from the test dataset and calculated the metrics again. As a result, the AUROC increased from 0.850 to 0.894in the internal test dataset, and the specificity increased from 0.585 to 0.697 when the sensitivity was fixed at 0.9. However, these metrics' significance could not be verified as data sample quantities differed.

According to Fig 4, the best classification performance was achieved when the projection head layers were two instead of three, and the downstream training sample quantity was 1024. Previous studies did not have a promised projection head structure [35, 47, 48]. However, Chen et al. conducted a data stress test with projection head layers ranging from two to four and noted that a larger layer quantity was associated with higher representation performance [49]. This trend seems stronger with a smaller downstream data volume, but the above characteristic did not appear as our downstream data set is a very small scale of only 1024 samples. From the data stress test results in Fig 4, when linear probing was performed with 256 labeled samples, the internal and external AUROCs expressed sufficiently high performance within a 0.828 and 0.831 data limitation, respectively. Although the labeled data amount was reduced by a quarter, the best-performing internal and external AUROC difference was only 0.022 and 0.016, a notable advantage of our method.

As for the ablation study represented by Table 3, we examined whether providing prior knowledge on training upstream GAN boosts model performance, specifically on training upstream GAN. We evaluated fine-tuning pre-trained weights effects [50] compared with upstream training networks from randomly initialized settings. We used FFHQ pre-trained weights [40] and a trained weight on a private 200,000 CXR images dataset from Center 1. Since FID demonstrated the best delegate diversity measure in the generated image set [44], better quality FID-generated images from fine-tuned GAN could not be applied to our task.

Furthermore, to demonstrate that even normal images are embedded in the latent space, we manipulated images using the "scoliosis direction" found by navigating the latent space in an unsupervised manner [25]. As a result, we confirmed normal images generated from vectors in the latent space. Fig 8 shows plausible normal image examples generated from image manipulation from scoliosis images.

## Conclusion

We developed a classifier for AIS through generative representation learning. Our model shows good AUROC under screening chest radiographs in both the internal and external

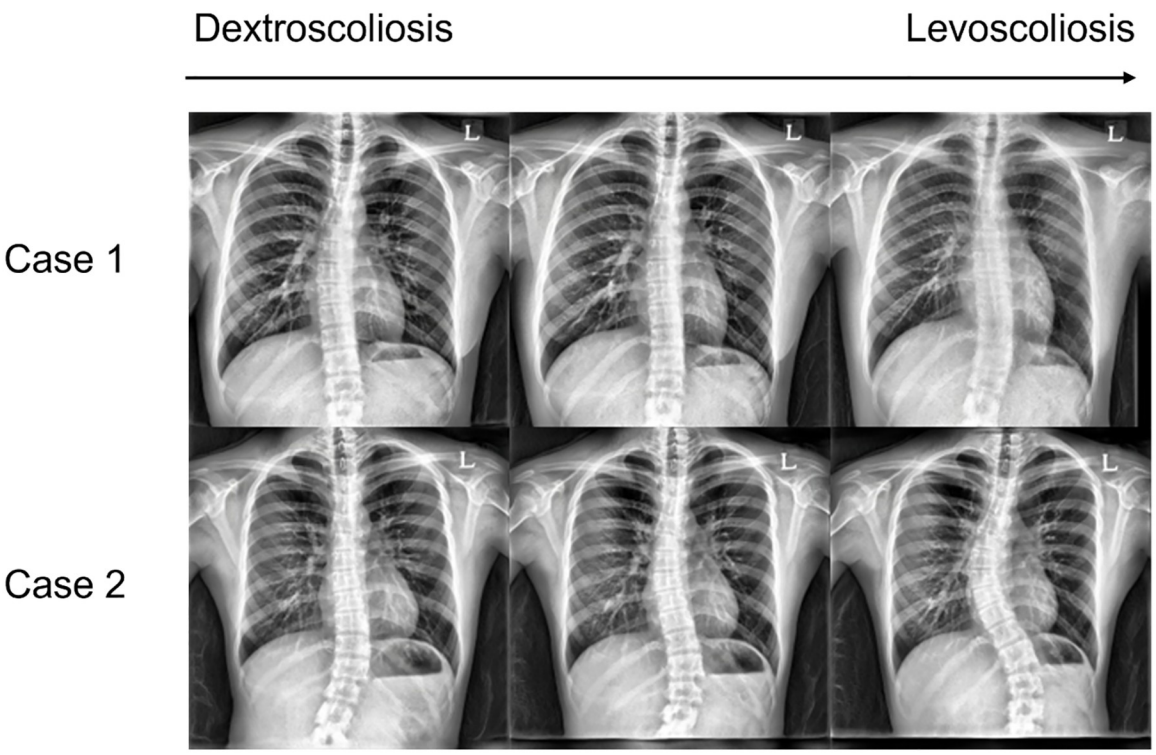


**Fig 8. Two manipulated image examples.** Note that scoliosis forms traverse between levoscoliosis and dextroscoliosis.

https://doi.org/10.1371/journal.pone.0285489.g008

datasets. Our model has learned the spectral severity of AIS, enabling it to generate normal images even when trained solely on scoliosis radiographs.

## Supporting information

**S1 File.**
(DOCX)

## Author Contributions

**Conceptualization:** Jun Soo Lee, Keewon Shin, Seung Min Ryu, Namkug Kim.

**Data curation:** Jun Soo Lee, Keewon Shin, Seung Min Ryu, Seong Gyu Jegal, Woojin Lee, Min A. Yoon, Gil-Sun Hong.

**Funding acquisition:** Namkug Kim.

**Investigation:** Seung Min Ryu, Min A. Yoon, Gil-Sun Hong, Sanghyun Paik, Namkug Kim.

**Methodology:** Jun Soo Lee, Keewon Shin, Sanghyun Paik.

**Project administration:** Gil-Sun Hong, Namkug Kim.

**Resources:** Keewon Shin, Woojin Lee, Min A. Yoon, Gil-Sun Hong, Sanghyun Paik, Namkug Kim.

**Software:** Seong Gyu Jegal.

**Validation:** Seong Gyu Jegal, Woojin Lee, Min A. Yoon, Gil-Sun Hong, Sanghyun Paik.

**Visualization:** Jun Soo Lee.

**Writing – original draft:** Jun Soo Lee, Keewon Shin, Seung Min Ryu.

**Writing – review & editing:** Jun Soo Lee, Keewon Shin, Seung Min Ryu, Seong Gyu Jegal, Woojin Lee, Namkug Kim.

## References

1. Lonstein JE. Adolescent idiopathic scoliosis. Lancet. 1994; 344(8934):1407–12. Epub 1994/11/19. https://doi.org/10.1016/s0140-6736(94)90572-x PMID: 7968079.
2. Yang J, Zhang K, Fan H, Huang Z, Xiang Y, Yang J, et al. Development and validation of deep learning algorithms for scoliosis screening using back images. Commun Biol. 2019; 2:390. Epub 2019/11/02. https://doi.org/10.1038/s42003-019-0635-8 PMID: 31667364.
3. Konieczny MR, Senyurt H, Krauspe R. Epidemiology of adolescent idiopathic scoliosis. J Child Orthop. 2013; 7(1):3–9. Epub 2014/01/17. https://doi.org/10.1007/s11832-012-0457-4 PMID: 24432052.
4. Suh SW, Modi HN, Yang JH, Hong JY. Idiopathic scoliosis in Korean schoolchildren: a prospective screening study of over 1 million children. Eur Spine J. 2011; 20(7):1087–94. Epub 2011/01/29. https://doi.org/10.1007/s00586-011-1695-8 PMID: 21274729.
5. Geijer H, Beckman K, Jonsson B, Andersson T, Persliden J. Digital radiography of scoliosis with a scanning method: initial evaluation. Radiology. 2001; 218(2):402–10. Epub 2001/02/13. https://doi.org/10.1148/radiology.218.2.r01ja32402 PMID: 11161153.
6. Geijer H, Verdonck B, Beckman KW, Andersson T, Persliden J. Digital radiography of scoliosis with a scanning method: radiation dose optimization. Eur Radiol. 2003; 13(3):543–51. Epub 2003/02/21. https://doi.org/10.1007/s00330-002-1476-1 PMID: 12594558.
7. Pruijs JE, Hageman MA, Keessen W, van der Meer R, van Wieringen JC. Variation in Cobb angle measurements in scoliosis. Skeletal Radiol. 1994; 23(7):517–20. Epub 1994/10/01. https://doi.org/10.1007/BF00223081 PMID: 7824978.
8. Zhang J, Lou E, Shi X, Wang Y, Hill DL, Raso JV, et al. A computer-aided Cobb angle measurement method and its reliability. J Spinal Disord Tech. 2010; 23(6):383–7. Epub 2010/02/04. https://doi.org/10.1097/BSD.0b013e3181bb9a3c PMID: 20124919.
9. Basak H, Kundu R, Singh PK, Ijaz MF, Wozniak M, Sarkar R. A union of deep learning and swarm-based optimization for 3D human action recognition. Sci Rep. 2022; 12(1):5494. Epub 2022/04/02. https://doi.org/10.1038/s41598-022-09293-8 PMID: 35361804.
10. Tu YC, Wang N, Tong F, Chen HM. Automatic measurement algorithm of scoliosis Cobb angle based on deep learning. J Phys Conf Ser. 2019; 1187. Artn 042100 https://doi.org/10.1088/1742-6596/1187/4/042100
11. Caesarendra W, Rahmaniar W, Mathew J, Thien A. Automated Cobb Angle Measurement for Adolescent Idiopathic Scoliosis Using Convolutional Neural Network. Diagnostics. 2022; 12(2). ARTN 396 https://doi.org/10.3390/diagnostics12020396 PMID: 35204487
12. Fu XL, Yang GS, Zhang KL, Xu NF, Wu J. An automated estimator for Cobb angle measurement using multi-task networks. Neural Comput Appl. 2021; 33(10):4755–61. https://doi.org/10.1007/s00521-020-05533-y
13. Geirhos R, Rubisch P, Michaelis C, Bethge M, Wichmann FA, Brendel WJapa. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. 2018.
14. Doersch C, Zisserman A, editors. Multi-task self-supervised visual learning. Proceedings of the IEEE International Conference on Computer Vision; 2017.
15. Zhang R, Isola P, Efros AA, editors. Colorful image colorization. European conference on computer vision; 2016: Springer.
16. Noroozi M, Favaro P, editors. Unsupervised learning of visual representations by solving jigsaw puzzles. European conference on computer vision; 2016: Springer.
17. Hinton GE, Osindero S, Teh Y-W. A fast learning algorithm for deep belief nets. Neural computation. 2006; 18(7):1527–54. https://doi.org/10.1162/neco.2006.18.7.1527 PMID: 16764513
18. Kingma DP, Welling M. Auto-encoding variational bayes. arXiv preprint arXiv:13126114. 2013.
19. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial nets. Advances in neural information processing systems. 2014;27.

20. Arora S, Risteski A, Zhang Y, editors. Do GANs learn the distribution? some theory and empirics. International Conference on Learning Representations; 2018.

21. Salimans T, Goodfellow I, Zaremba W, Cheung V, Radford A, Chen X. Improved techniques for training gans. Advances in neural information processing systems. 2016; 29:2234–42.

22. Naga Srinivasu P, Krishna TB, Ahmed S, Almusallam N, Khaled Alarfaj F, Allheeib N. Variational Autoencoders-BasedSelf-Learning Model for Tumor Identification and Impact Analysis from 2-D MRI Images. J Healthc Eng. 2023; 2023:1566123. Epub 2023/01/28. https://doi.org/10.1155/2023/1566123 PMID: 36704578.

23. He KM, Chen XL, Xie SN, Li YH, Dollar P, Girshick R. Masked Autoencoders Are Scalable Vision Learners. Proc Cvpr Ieee. 2022:15979–88.

24. Karras T, Laine S, Aila T, editors. A style-based generator architecture for generative adversarial networks. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2019.

25. Shen Y, Zhou B, editors. Closed-form factorization of latent semantics in gans. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2021.

26. Abdal R, Qin Y, Wonka P, editors. Image2stylegan: How to embed images into the stylegan latent space? Proceedings of the IEEE/CVF International Conference on Computer Vision; 2019.

27. Association GAotWM. World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects. The Journal of the American College of Dentists. 2014; 81(3):14–8. PMID: 25951678

28. Force USPST, Grossman DC, Curry SJ, Owens DK, Barry MJ, Davidson KW, et al. Screening for Adolescent Idiopathic Scoliosis: US Preventive Services Task Force Recommendation Statement. JAMA. 2018; 319(2):165–72. Epub 2018/01/11. https://doi.org/10.1001/jama.2017.19342 PMID: 29318284.

29. Pizer SM, Amburn EP, Austin JD, Cromartie R, Geselowitz A, Greer T, et al. Adaptive histogram equalization and its variations. Computer vision, graphics, and image processing. 1987; 39(3):355–68.

30. Zuiderveld K. Contrast limited adaptive histogram equalization. Graphics gems. 1994:474–85.

31. Shahbazi M, Danelljan M, Paudel DP, Van Gool L. Collapse by Conditioning: Training Class-conditional GANs with Limited Data. arXiv preprint arXiv:220106578. 2022.

32. Goldstein L, Waugh T. Classification and terminology of scoliosis. Clinical Orthopaedics and Related Research®. 1973; 93:10–22. https://doi.org/10.1097/00003086-197306000-00003 PMID: 4722939

33. Cruickshank J, Koike M, Dickson R. Curve patterns in idiopathic scoliosis. A clinical and radiographic study. The Journal of bone and joint surgery British volume. 1989; 71(2):259–63. https://doi.org/10.1302/0301-620X.71B2.2925744 PMID: 2925744

34. Yang C, Shen Y, Zhou B. Semantic hierarchy emerges in deep generative representations for scene synthesis. International Journal of Computer Vision. 2021; 129(5):1451–66.

35. Chen T, Kornblith S, Norouzi M, Hinton G, editors. A simple framework for contrastive learning of visual representations. International conference on machine learning; 2020: PMLR.

36. Zhang R, Isola P, Efros AA, editors. Split-brain autoencoders: Unsupervised learning by cross-channel prediction. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2017.

37. Nair V, Hinton GE, editors. Rectified linear units improve restricted boltzmann machines. Icml; 2010.

38. Ioffe S, Szegedy C, editors. Batch normalization: Accelerating deep network training by reducing internal covariate shift. International conference on machine learning; 2015: PMLR.

39. Schlegl T, Seeböck P, Waldstein SM, Schmidt-Erfurth U, Langs G, editors. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. International conference on information processing in medical imaging; 2017: Springer.

40. Karras T, Aittala M, Hellsten J, Laine S, Lehtinen J, Aila T. Training generative adversarial networks with limited data. arXiv preprint arXiv:200606676. 2020.

41. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems. 2019; 32:8026–37.

42. Mescheder L, Geiger A, Nowozin S, editors. Which training methods for GANs do actually converge? International conference on machine learning; 2018: PMLR.

43. Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv preprint arXiv:14126980. 2014.

44. Heusel M, Ramsauer H, Unterthiner T, Nessler B, Hochreiter S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems. 2017;30.

45. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP. Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing. 2004; 13(4):600–12. https://doi.org/10.1109/tip.2003.819861 PMID: 15376593

46. Horne JP, Flannery R, Usman S. Adolescent Idiopathic Scoliosis: Diagnosis and Management. Am Fam Physician. 2014; 89(3):193–8. PMID: 24506121

47. Chen X, Xie S, He K, editors. An empirical study of training self-supervised vision transformers. Proceedings of the IEEE/CVF International Conference on Computer Vision; 2021.

48. Grill J-B, Strub F, Altché F, Tallec C, Richemond P, Buchatskaya E, et al. Bootstrap your own latent-a new approach to self-supervised learning. Advances in Neural Information Processing Systems. 2020; 33:21271–84.

49. Chen T, Kornblith S, Swersky K, Norouzi M, Hinton GE. Big self-supervised models are strong semi-supervised learners. Advances in neural information processing systems. 2020; 33:22243–55.

50. Grigoryev T, Voynov A, Babenko A, editors. When, Why, and Which Pretrained GANs Are Useful? International Conference on Learning Representations; 2021.