*bioengineering*

MDPI

*Article*

# Volumetric Imitation Generative Adversarial Networks for Anatomical Human Body Modeling

**Jion Kim, Yan Li and Byeong-Seok Shin ***

Department of Electrical and Computer Engineering, Inha University, Incheon 22212, Republic of Korea; 3161508@gmail.com (J.K.); leeyeon@inha.ac.kr (Y.L.)
* Correspondence: bsshin@inha.ac.kr

**Abstract:** Volumetric representation is a technique used to express 3D objects in various fields, such as medical applications. On the other hand, tomography images for reconstructing volumetric data have limited utilization because they contain personal information. Existing GAN-based medical image generation techniques can produce virtual tomographic images for volume reconstruction while preserving the patient's privacy. Nevertheless, these images often do not consider vertical correlations between the adjacent slices, leading to erroneous results in 3D reconstruction. Furthermore, while volume generation techniques have been introduced, they often focus on surface modeling, making it challenging to represent the internal anatomical features accurately. This paper proposes volumetric imitation GAN (VI-GAN), which imitates a human anatomical model to generate volumetric data. The primary goal of this model is to capture the attributes and 3D structure, including the external shape, internal slices, and the relationship between the vertical slices of the human anatomical model. The proposed network consists of a generator for feature extraction and up-sampling based on a 3D U-Net and ResNet structure and a 3D-convolution-based LFFB (local feature fusion block). In addition, a discriminator utilizes 3D convolution to evaluate the authenticity of the generated volume compared to the ground truth. VI-GAN also devises reconstruction loss, including feature and similarity losses, to converge the generated volumetric data into a human anatomical model. In this experiment, the CT data of 234 people were used to assess the reliability of the results. When using volume evaluation metrics to measure similarity, VI-GAN generated a volume that realistically represented the human anatomical model compared to existing volume generation methods.

**Keywords:** GAN; imitation; 3D reconstruction; volumetric representation; human body; deep learning



**Citation:** Kim, J.; Li, Y.; Shin, B.-S. Volumetric Imitation Generative Adversarial Networks for Anatomical Human Body Modeling. *Bioengineering* **2024**, *11*, 163. https://doi.org/10.3390/bioengineering11020163

Academic Editor: Alan Wang

Received: 9 January 2024
Revised: 2 February 2024
Accepted: 6 February 2024
Published: 7 February 2024



## 1. Introduction

Volumetric representation [1,2] is a popular technique to express 3D objects, such as surface modeling [3,4]. Volumetric data are generated mainly by reconstructing from tomographic images, such as computed tomography (CT) and magnetic resonance imaging (MRI) [5,6]. On the other hand, it is challenging to adopt tomographic images for purposes other than medical, such as diagnosis and surgery, because they contain personal information. GAN-based medical image generation techniques [7,8] can produce anatomically meaningful virtual tomographic images applicable to volume reconstruction. On the other hand, the image generation process does not consider the relationship between the adjacent slices of a volume because these methods account for 2D correlations exclusively. This can lead to erroneous results when reconstructing volumes from generated images, particularly when maintaining the 3D structural coherence between adjacent slices. While techniques have been proposed to generate 3D volumes [9,10], these approaches have been limited to generating surface representations and have failed to capture the internal characteristics. Consequently, a volume generation method is needed to reflect the entire 3D human anatomical model, including its internal portion.

This paper introduces the volumetric imitation GAN (VI-GAN), a novel approach that aims to imitate human anatomical models to generate volumetric data. The primary

goal of this approach is to generate 3D models that faithfully capture the attributes and 3D structure (external shape, internal slice, and the relationship between vertical slices) within the human anatomical model. The proposed network comprises two main components: a generator to obtain the volumetric data and a discriminator to evaluate the authenticity between the generated volume and the ground truth. The generator performs feature extraction and up-sampling to produce a volume based on the 3D U-Net [11] and ResNet [12] structures. Moreover, the initial feature extraction process from the input image set uses a 3D-convolution-based LFFB (local feature fusion block) [13] to incorporate features at various scales. On the other hand, the discriminator uses 3D convolution to extract the authenticity of the generated volume and ground truth. Using the proposed network structure makes it possible to account for vertical correlations in the volume generation process compared to existing 2D image generation techniques. Moreover, it provides a more realistic representation compared to previous methods focused solely on surface generation because it can faithfully intimate the internal features of the anatomical model. The volume comprises 3D data with higher dimensions than the image, so converging in a specific shape is difficult. Therefore, if the basic distance loss alone is applied to generate the volume, it barely converges in the form of a human anatomical model [14]. Thus, the proposed method devises a reconstruction loss so VI-GAN can generate the volume converging to the human anatomical model. Reconstruction loss includes feature loss and similarity loss. Feature loss is calculated using an overlapping region between the generated volume and the ground truth. Similarity loss is also calculated as an internal similarity based on a structural similarity index map (SSIM) [15].

The spine data from the Digital Korean dataset [16] provided by the Korean Institute of Science and Technology Information (KISTI) and the liver data from CT volumes with multiple organ segmentations (CT-ORG) [17] were applied during the experiment to validate the proposed technique. Evaluation metrics, such as F1-score, dice coefficient, peak signal-to-noise ratio (PSNR), and universal image quality index (UQI), were used to measure the resemblance between the generated volume and the human anatomical model. The VI-GAN outperformed existing methods by producing volumes closely representing the human anatomical model.

The volumetric data generated by the VI-GAN included the external shape and internal structure of the human anatomical model. Therefore, it can be used in various fields, such as diagnosis [18,19] and surgical simulation [20,21]. VI-GAN can produce a 3D human anatomical model that can enhance training efficiency and immersion for medical professionals. Moreover, virtual tomographic images can be produced by decomposing the volumetric data generated by the VI-GAN. Compared to existing medical image generation methods, these images have fewer errors in the relationship between neighboring slices. The tomographic images produced by the VI-GAN can enhance the capabilities of medical professionals to distinguish diseases effectively during the diagnostic process.

The contributions of this paper are as follows. (1) This paper proposes a VI-GAN to create 3D volumetric data that capture the attributes and 3D structure of a human anatomical model (external shape, internal slice, and vertical slice relationships). (2) This paper introduces reconstruction loss that encompasses feature loss and similarity loss to enhance the convergence rate of VI-GAN in volume generation. The feature loss measures the overlapping regions, while the similarity loss quantifies the resemblance between the generated volume and ground truth.

## 2. Related Works

Several studies have proposed generating the surface of a human anatomical model in the volume format from tomographic images. Balashova et al. [9] proposed a method to reconstruct the surface of the liver using a single X-ray image. They used mask data and images in the training process to generate liver data closer to the ground truth. Henzler et al. [22] generated the surface of animal bones using multiple X-ray images. Their study combined the volumes generated from images of multiple viewpoints. It

allowed the production of a high-quality volume with a complete reconstruction of the parts that could be obscured easily from a single viewpoint. Kasten et al. [10] devised a network to reconstruct knee bones from bi-planar X-ray images. Their study synthesized the volumes produced by duplicating axial, coronal, and sagittal images in the $z$-axis direction in the training process. On the other hand, the generated data could not preserve the characteristics of the entire portion of that model because these methods generated only the surface of the human anatomical model. Furthermore, many of these studies have been proposed in the form of CNNs. The proposed technique requires generating virtual volume data. Hence, applying a GAN specialized for generation is necessary.

Several studies have proposed producing 2D medical data using GANs. These studies have focused mainly on the synthesis and reconstruction of images. Synthesis techniques include changing the style and modality [23,24] and adding characteristics, such as nodules and tumors [25]. Reconstruction techniques cover the super-resolution [26] process. Among these techniques, medical image generation [7,27] can produce the virtual tomographic images required for volume reconstruction. Chuquicusma et al. [28] and Frid-Adar et al. [29] proposed techniques for generating images representing lung modules and liver lesions using a deep convolution GAN (DCGAN). Beer et al. [8] devised a method for generating tomographic images through the progressive growing of GAN (PGGAN) [30] to express skin lesions realistically. These methods generated tomographic images to improve the classification and segmentation performance during the training process. Nevertheless, these medical image generation techniques considered only the 2D correlation within the image. Therefore, erroneous results can be obtained when reconstructing volumes from these generated images because the vertical correlation with the adjacent slices is not considered. Hence, the volume generation technique using a GAN should be proposed to prevent such erroneous results.

Some studies adopted the GAN structure for volume generation by expanding various image generation techniques into 3D space. Wu et al. [31] proposed 3D-GAN and 3D variational autoencoder GAN (3D-VAE-GAN) structures to produce the surface of an object as a volume using a single 2D image. Smith et al. [32] applied the Wasserstein distance to 3D-GAN structures, which improved the volume quality. Volume-based GANs are applied in the medical field, such as classification [33], segmentation [34], denoising [35], and detection [36]. Vox2Vox [37] is one of the volume-based GAN techniques used in the medical field for segmenting brain tumors. Nevertheless, few studies have applied GAN structures in volume generation, particularly for human anatomical models. Thus, it is essential to devise a method for generating volume data similar to a human anatomical model using a GAN.

## 3. Methods

### *3.1. Training Process*

Figure 1 presents the overall network structure and components of the VI-GAN. The VI-GAN aims to generate volumetric data similar to the human anatomical model. This generation process is achieved using the volume generator **G**, which incorporates a 3D-convolution-based LFFB (local feature fusion block) [13], denoted as the 3D LFFB. The discriminator **D** is used to assess the authenticity of the generated volume and the ground truth. In addition, the VI-GAN devises reconstruction loss, including distance, feature, and similarity losses, to converge the generated volumes as a human anatomical model.

The proposed method should reconstruct all the voxels in the generated volume similar to the corresponding voxels in the ground truth volume. The difference between any two corresponding values should be close to zero. Equation (1) represents the training objective for a volumetric dataset. $\mathbf{V}$ represents the ground truth volume given during training, and $\hat{\mathbf{V}}$ denotes the corresponding generated volume. $\mathbf{V}[m]$ and $\hat{\mathbf{V}}[m]$ are the $m$th voxel of the ground truth volume $\mathbf{V}$ and generated volume $\hat{\mathbf{V}}$, respectively. $l$, $w$, and $h$ represent $x$, $y$, and $z$-axis resolutions of the volume, respectively.

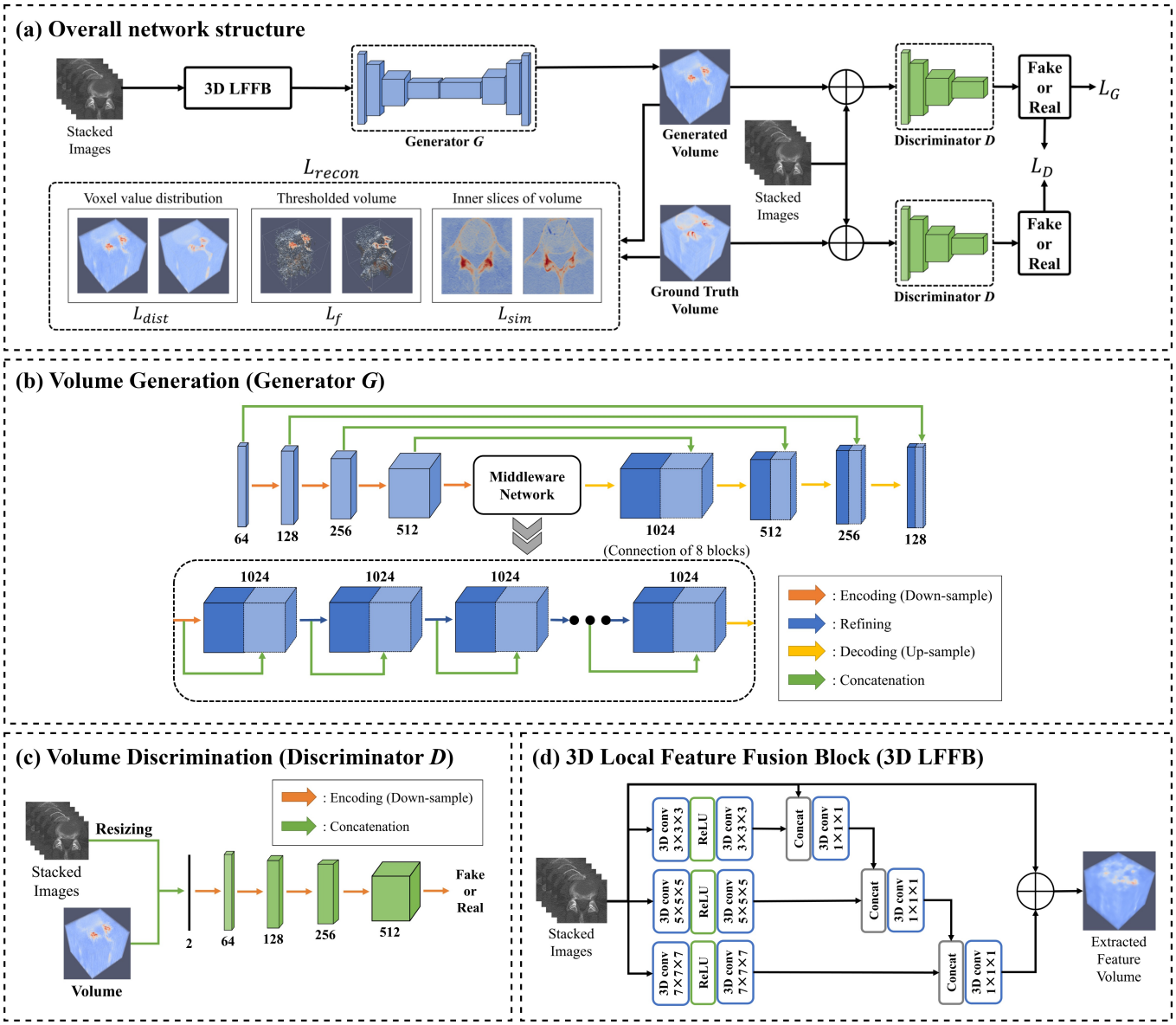


**Figure 1.** Overall proposed framework and components of VI-GAN. The blue cubes represent features extracted by the generator, while the green cubes depict features extracted by the discriminator. (**a**) Overview of the proposed network structure for generating a volume. (**b**,**c**) Detailed structure of the volume generator and discriminator. Three dots means repeating the preceding refining and concatenation process in the same manner. (**d**) 3D local feature fusion block (3D LFFB) that initially extracts the features from an image set before the volume generator.

$$\begin{aligned} &\forall m \left| \mathbf{V}[m] - \hat{\mathbf{V}}[m] \right| \approx 0 \\ &0 \le m \le l \times w \times h, \forall m(\mathbf{V}[m], \hat{\mathbf{V}}[m] \in \mathbb{R}^{l \times w \times h}) \end{aligned} \tag{1}$$

Volume generator **G** produces a volume using the input image set $\mathbf{I}^{\mathrm{G}}$. The structure of **G** is represented in Figure 1b using Equation (2). $\mathbf{X}_k$ represents the intermediate result in the $k$th layer. The input image set is reconstructed into a volume via each $f$ layer. In the overall volume generation process, the network structures based on 3D U-Net [11] and ResNet [12] are categorized into the encoding, refining, and decoding parts. The encoding, refining, and decoding parts comprise $f_k^{\mathrm{down}}$, $f_k^{\mathrm{mid}}$, and $f_k^{\mathrm{up}}$ layers, respectively. The final layer is formed as the $f^{'}$ layer; $k$ is the index of those layers. $\theta_k^{\mathrm{G}}$ is the learning parameter of the $k$th layer in the generator. Before passing to the $f$ layers, 3D LFFB is performed to

extract the essential features of the image set. Figure 1d presents the network structure of 3D LFFB.

$f_k^{\text{down}}$ is the convolution layer that extracts the features from the input through down-sampling. $f_k^{\text{mid}}$ is the convolution layer that refines features using the ResNet architecture. $f_k^{\text{up}}$ is the deconvolution layer that reconstructs a volume from features by up-sampling. The final layer $f^{'}$ is a convolution layer that generates a volume $\hat{\mathbf{V}}$ in which all voxel values are normalized from 0.0 to 1.0. The kernel sizes of all layers are $4 \times 4 \times 4$, and the stride sets were assigned for $f_k^{\text{down}}$, $f_k^{\text{up}}$, and $f^{'}$ as two and $f_k^{\text{mid}}$ as one. $n_{\text{G}}$ indicates the total number of layers of the generator; $d$, $r$, and $u$ are the layer indices of the encoding, refining, and decoding parts, respectively.

$$\begin{aligned} &\mathbf{X}_0 = \mathbf{I}^{\text{G}}, \mathbf{X}_d = f_{d-1}^{\text{down}}(\mathbf{X}_{d-1}; \theta_{d-1}^{\text{G}}), \quad \mathbf{X}_r = f_{r-1}^{\text{mid}}(\mathbf{X}_{r-1}; \theta_{r-1}^{\text{G}}), \\ &\mathbf{X}_u = f_{u-1}^{\text{up}}(\mathbf{X}_{u-1}; \theta_{u-1}^{\text{G}}) \quad \hat{\mathbf{V}} = f^{'}(\mathbf{X}_{n_{\text{G}}-1}; \theta_{n_{\text{G}}-1}^{\text{G}}), \\ &0 < d < r < u < n_{\text{G}} \end{aligned} \tag{2}$$

Volume discriminator $\mathbf{D}$ distinguishes the authenticity between the generated and ground truth volume. The structure of $\mathbf{D}$ is represented in Figure 1c with Equation (3). $\mathbf{I}^{\text{D}}$ is the input volume. $\mathbf{Y}_k$ represents the intermediate result in the $k$th layer. The probability $p$ indicating the authenticity of the input volume is calculated through each $g$ layer. $\theta_k^{\text{D}}$ is the learning parameter of the $k$th layer in the discriminator. $g_k$ is a convolution layer that extracts the features from $\mathbf{I}^{\text{D}}$ by down-sampling. The final layer, $g^{'}$, is a fully connected layer to generate a probability normalized from 0.0 to 1.0. The kernel sizes and stride sets of all layers are $4 \times 4 \times 4$ and 2, respectively. $n_{\text{D}}$ represents the total number of layers in the discriminator.

$$\begin{aligned} &\mathbf{Y}_0 = \mathbf{I}^{\text{D}}, \quad \mathbf{Y}_k = g_{k-1}(\mathbf{Y}_{k-1}; \theta_{k-1}^{\text{D}}), \\ &p = g^{'}(\mathbf{Y}_{n_{\text{D}}-1}; \theta_{n_{\text{D}}-1}^{\text{D}}), \quad 0 < k < n_{\text{D}} \end{aligned} \tag{3}$$

### *3.2. Loss Function*

The loss function of the generator and discriminator was designed, as shown in Equation (4). $L_{\text{G}}$ and $L_{\text{D}}$ are the generator and discriminator loss, respectively, of the ground truth volumes $\mathbf{V}$, generated volumes $\hat{\mathbf{V}}$, and the input image set of the generator $\mathbf{I}^{\text{G}}$. $L_{\text{recon}}$ is the reconstruction loss, and $\alpha$ is a constant weight assigned to $L_{\text{recon}}$.

$$\begin{aligned} L_{\text{G}} &= \mathbf{E}_{x \sim \mathbf{I}^{\text{G}}, \hat{v} \sim \hat{\mathbf{V}}} \| D(x, \hat{v}) - \mathbf{1} \|_2^2 + \alpha L_{\text{recon}} \\ L_{\text{D}} &= \mathbf{E}_{x \sim \mathbf{I}^{\text{G}}, v \sim \mathbf{V}, \hat{v} \sim \hat{\mathbf{V}}} [\| D(x, v) - \mathbf{1} \|_2^2 + \| D(x, \hat{v}) \|_2^2] \end{aligned} \tag{4}$$

The reconstruction loss $L_{\text{recon}}$ calculates the discrepancy between the generated and the ground truth volumes. Reconstruction loss consists of the distance, feature, and similarity loss. Among these losses, the distance loss $L_{\text{dist}}$ is defined using Equation (5). The L1 loss is used for the distance loss.

$$L_{\text{dist}} = \mathbf{E}_{v \sim \mathbf{V}, \hat{v} \sim \hat{\mathbf{V}}} \| v - \hat{v} \|_1 \tag{5}$$

The voxel values in the generated volume indicate the predicted density of the corresponding area of the human anatomical model. The ground truth volume contains many undefined parts outside the human anatomical model, represented by low-value voxels. Therefore, many voxels in the ground truth volume have low values. The proposed method should reconstruct the entire part of the human anatomical model. Nevertheless, when only the distance loss is used in the regularization term, all voxels are averaged to minimize the distance loss. This process reduces the value of the high-value voxels [14]. The generated volume barely converges with the target model because the shape and characteristics of the human anatomical model are composed mainly of a high-value area. This paper proposes

a reconstruction loss that consists of distance loss with feature loss and similarity loss to solve this problem.

Feature loss represents how many high-value voxels overlap between two volumes. Feature loss can be used to emphasize and preferentially reconstruct high-value voxels during the volume-reconstruction process. Feature loss $L_{\mathrm{f}}$ is expressed as Equation (6). This loss uses $N_{\mathrm{r}}$ number of thresholds; $t_s$ is the $s$th threshold; $m$ is the voxel index in the volume; $l$, $w$, and $h$ are the $x$, $y$, and $z$-axis resolutions of the volume, respectively. $\mathbf{I}(\cdot)$ is the indicator function.

$$L_{\mathrm{f}} = 1 - \mathbf{E}_{v\sim\mathbf{V},\hat{v}\sim\hat{\mathbf{V}}}\left[\sum_{s}^{N_{\mathrm{r}}} \frac{\sum_{m}^{l\times w\times h} \mathrm{I}(v[m] \geq t_s)\mathrm{I}(\hat{v}[m] \geq t_s)}{\sum_{m}^{l\times w\times h} \mathrm{I}(\mathrm{I}(v[m] \geq t_s) + \mathrm{I}(\hat{v}[m] \geq t_s))}\right] \quad (6)$$
$$0 \leq t_s \leq 1$$

Feature loss can be used to generate the characteristics of a human anatomical model, composed mainly of drastic changes in density. Such characteristics are often represented primarily by a high value. The high-value area within the human anatomical model can be fully reconstructed in the output volume by assigning feature loss during the training process.

Similarity loss $L_{\mathrm{sim}}$ reflects the difference in image quality of each internal image slice between the generated and ground truth volumes, which is defined in Equation (7). The SSIM, $SSIM(\cdot,\cdot)$ [15], measures the difference in image quality. Similarity loss computes the difference in volume quality calculated based on the SSIM between the internal image slices in the generated and ground truth volume for all z-values. $S(\cdot,k)$ is the selector that extracts the $k$th slice in the volume, and $h$ is the resolution in the $z$-axis direction of the volume. Using the similarity loss, the internal voxel values can be reconstructed like those of the ground truth volume.

$$L_{\mathrm{sim}} = 1 - \mathbf{E}_{v\sim\mathbf{V},\hat{v}\sim\hat{\mathbf{V}}}\{\frac{1}{h}\sum_{k=0}^{h} SSIM(S(v,k), S(\hat{v},k))\} \quad (7)$$

The final reconstruction loss is represented as Equation (8).

$$L_{\mathrm{recon}} = L_{\mathrm{dist}} + L_{\mathrm{f}} + L_{\mathrm{sim}} \quad (8)$$

### *3.3. Experimental Setting*

This study used CT images of the spine (fifth lumbar vertebra and right hip bone) from the Digital Korean dataset provided by the KISTI [16] and the liver from CT-ORG [17]. The CT data for 94 people for the spine and 140 people for the liver were used. Among them, the CT data of the following were applied: 70% for training, 20% for testing, and 10% for validation. A total of 3117 slices for the fifth lumbar vertebra (average of 33 slices per person), 4579 slices for the right hip bone (average of 49 slices per person), and 19,314 slices for the liver (average of 483 slices per person) were used to generate volumetric data for training. NVIDIA GeForce RTX 3090 Ti with 24,268 MB GPU memory was applied for training. The Adam optimizer [38] was used with a learning rate of $2 \times 10^{-4}$. The dropout rates of the middle-ware network blocks in Figure 1b were set to 0.2. The constant $\alpha$ was set to 33.0 when implementing Equation (4).

## 4. Results

The generated volumes were evaluated using a confusion matrix [39]. Each voxel in the generated volume was classified as positive if it had a high value and negative if it had a low value. In addition, each voxel in the ground truth volume was also categorized as true if it had a high value and false if it had a low value. The states of the voxels were judged by a comparison with the threshold value, whether each voxel value was high or low. Equation (9) expresses the TP (true positive), FP (false positive), and FN (false negative) used for the evaluation metric. $t$ is the threshold; $v$ is the ground truth volume;

$\hat{v}$ is the corresponding generated volume; $l, w$, and $h$ are the $x, y$, and $z$-axis resolutions of the volume, respectively. The thresholds were used to evaluate how well the voxels that formulate a shape and internal characteristics of a human anatomical model were reconstructed. TP, FP, and FN were used to calculate the precision, recall, F1-score, and Dice coefficient [40].

$$\begin{aligned} \mathrm{TP} &= \frac{1}{l \times w \times h} \sum_{m=0}^{l \times w \times h} \mathrm{I}(v[m] > t)\mathrm{I}(\hat{v}[m] > t) \\ \mathrm{FP} &= \frac{1}{l \times w \times h} \sum_{m=0}^{l \times w \times h} \mathrm{I}(v[m] \leq t)\mathrm{I}(\hat{v}[m] > t) \\ \mathrm{FN} &= \frac{1}{l \times w \times h} \sum_{m=0}^{l \times w \times h} \mathrm{I}(v[m] \leq t)\mathrm{I}(\hat{v}[m] \leq t) \end{aligned} \tag{9}$$

Figure 2 compares the generated volumes between the proposed VI-GAN and existing methods. The quality of the generated volume was evaluated using three criteria: volumetric shape, thresholding result, and internal image slice. The volumetric shape describes how accurately the generated volume represents the external shape of the ground truth volume. The thresholding result describes the internal structure that represents how much high-value voxels overlap between generated and ground truth volume. The internal image slice expresses how similar the internal area is between the generated and ground truth volume. The value of each voxel is a floating point between 0.0 and 1.0. The positions with a lower or higher voxel value are blue or red, respectively. The thresholds were determined for thresholding results by analyzing the voxel value distributions within the ground truth volumes. Specifically, thresholds of 0.4, 0.3, and 0.28 were applied to the fifth lumbar vertebra, right hip bone, and liver, respectively. For the fifth lumber vertebra and right hip bone, the Q3 (third quantile, 75% of data points) [41] values were adopted as thresholds to emphasize the rigid areas within the skeletal system. For the liver, the Q1 (first quantile, 25% of data points) values were used as thresholds to visualize the soft tissue density. For the internal image slice, the slice position is the center of the volume. The slices correspond to the $xy$ plane exactly, the $xy$ plane leaning at $-45°$, and the $yz$ plane in the cases of the fifth lumbar vertebra, right hip bone, and liver, respectively. All slices in the CT volume were used as input data in Pix2Vox. The real CT volume was applied as input for the end-to-end CNN instead of the synthesized volume.

In the Pix2Vox and end-to-end CNN model, the volumetric shape had an ambiguous form that cannot express the human anatomical model in detail. In addition, the high-value voxels were scattered and could not typically converge to the shape of the human anatomical model. The generated volumes of the Vox2Vox model converged more to the human anatomical model than those of the Pix2Vox and end-to-end CNN models. The volumetric shapes and internal slices depicted as the form of the liver and bone represent this convergence. In the case of the fifth lumber vertebra and right hip bone, however, the high-value voxels in some areas of the volume were not fully reconstructed, which could not make the shape of the human anatomical model clear. In the case of the liver, the volumetric shape did not resemble the ground truth volume, and the organ shape did not appear in the center, which is the correct position for the internal slice. The VI-GAN generated a volumetric shape, high-value voxels, and internal parts that were more similar to the ground truth volume than the other methods. In conclusion, in a qualitative comparison, the VI-GAN generated volume data that are more similar to the human anatomical model than other existing methods.

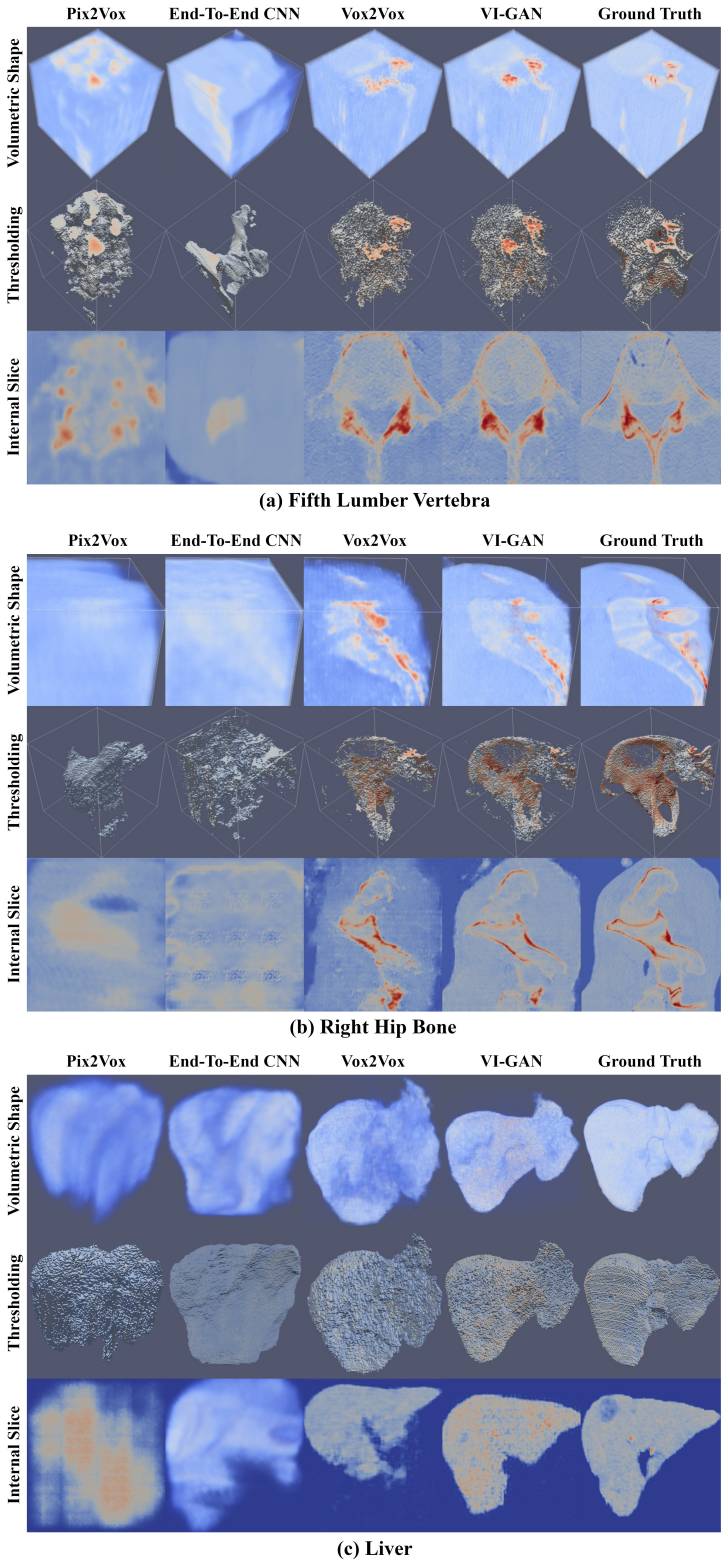


**Figure 2.** Qualitative comparison of the generated volumes in the proposed VI-GAN and existing method. The (**a**) fifth lumber vertebra, (**b**) right hip bone, and (**c**) liver volume data are represented. The images consist of a volumetric shape (top row), thresholding result (center row), and internal image slice (bottom row). The volumes generated by the Pix2Vox [42] (first column), end-to-end CNN [10] (second column), Vox2Vox [37] (third column), and VI-GAN (fourth column) models were compared.

Table 1 lists the quality measurement of the generated volume between the proposed VI-GAN and other existing methods. The quality was calculated using evaluation factors to measure the disparity between the generated and ground truth volume. The intersection over union (IoU), F1-score (F1), and Dice coefficient (DC) [43] represent the rate of overlapping voxels between thresholded results. The threshold values for calculating the IoU, F1, and DC metrics were 2.5, 1.9, and 2.8, respectively, corresponding to the fifth lumbar vertebra, right hip bone, and liver. These values represent the Q1 (25% of data points) value of the voxel distribution. The L1 error (L1) describes the voxel-wise difference between volumes. The peak signal-to-noise ratio (PSNR) [15], universal quality index (UQI) [44], visual saliency-induced index (VSI) [45], and structural similarity index map (SSIM) showed the similarity between the generated and ground truth volumes calculated using all slices.

A comparison of the VI-GAN with Pix2Vox and end-to-end CNN revealed that the VI-GAN had the highest result for all evaluation factors in all cases (the fifth lumber vertebra, right hip bone, and liver). As a result, the proposed method produces a volume that closely resembles the ground truth, displaying a higher rate of overlap among high-value voxels compared to Pix2Vox and end-to-end CNN. A comparison of the VI-GAN with Vox2Vox using the IoU, F1-score, and Dice coefficient showed that the VI-GAN produced better results in all cases than Vox2Vox. When comparing the metrics of L1, PSNR, UQI, VSI, and SSIM, it is difficult to definitively conclude whether the VI-GAN or Vox2Vox exhibited superior performance on similarity. Furthermore, the disparities in results were relatively minor in most cases. Based on these findings, the volumes produced by Vox2Vox and the VI-GAN showed similar degrees of resemblance to the ground truth. In summary, as listed in Table 1, the VI-GAN effectively generated a volume by accurately capturing high-value voxels compared to other methods while preserving the similarity to the human anatomical model.

**Table 1.** Quality comparison between the volumes of proposed VI-GAN and existing methods using the evaluation factors. The best results are represented in bold.

| Method | Fifth Lumber Vertebra | | | | Right Hip Bone | | | | Liver | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **IoU** | **DC** | **F1** | **L1** | **IoU** | **DC** | **F1** | **L1** | **IoU** | **DC** | **F1** | **L1** |
| Pix2Vox | 0.501 | 0.670 | 0.662 | 0.121 | 0.381 | 0.553 | 0.551 | 0.131 | 0.264 | 0.488 | 0.402 | 0.139 |
| End-to-end CNN | 0.524 | 0.681 | 0.686 | 0.121 | 0.372 | 0.544 | 0.541 | 0.114 | 0.231 | 0.484 | 0.346 | 0.155 |
| Vox2Vox | 0.596 | 0.743 | 0.742 | **0.096** | 0.457 | 0.628 | 0.626 | **0.075** | 0.393 | 0.612 | 0.555 | **0.053** |
| **VI-GAN** | **0.712** | **0.832** | **0.831** | 0.102 | **0.865** | **0.929** | **0.927** | **0.075** | **0.552** | **0.741** | **0.685** | 0.056 |
| **Method** | **Fifth Lumber Vertebra** | | | | **Right Hip Bone** | | | | **Liver** | | | |
| | **PSNR** | **UQI** | **VSI** | **SSIM** | **PSNR** | **UQI** | **VSI** | **SSIM** | **PSNR** | **UQI** | **VSI** | **SSIM** |
| Pix2Vox | 16.226 | 0.839 | 0.827 | 0.329 | 15.764 | 0.729 | 0.829 | 0.279 | 15.486 | 0.884 | 0.831 | 0.284 |
| End-to-end CNN | 16.573 | 0.849 | 0.836 | 0.244 | 16.676 | **0.817** | **0.872** | **0.613** | 14.804 | 0.839 | 0.776 | 0.409 |
| Vox2Vox | **18.249** | 0.876 | **0.875** | 0.483 | **24.615** | **0.817** | **0.872** | **0.613** | 20.942 | **0.900** | **0.880** | 0.652 |
| **VI-GAN** | 18.189 | **0.883** | 0.874 | **0.505** | 24.518 | 0.816 | 0.865 | 0.600 | **26.661** | 0.886 | 0.874 | **0.669** |

Figure 3 presents the assessment results using the intersection over union, F1-score, and Dice coefficient across various threshold values. The objective of this experiment was to evaluate the accuracy in reconstructing high-value voxels, which holds significance in accurately representing the essential information within the volume. This information encompasses the overall appearance of soft tissue or rigid structures within the skeletal system. Furthermore, achieving precise reconstructions of high-value voxels is challenging because of the limited occurrence of such areas in the ground truth volume. The thresholds

were selected within the range of Q1 (25% of data points) to Q3 (75% of data points), corresponding to 25%, 37.5%, 50%, 62.5%, and 75% within the voxel distribution. In this experiment, the following thresholds were applied: 0.2500, 0.2875, 0.3250, 0.3625, and 0.4000 for the fifth lumbar vertebra; 0.1900, 0.2175, 0.2450, 0.2725, and 0.3000 for the right hip bone; and 0.2800, 0.3250, 0.3700, 0.4150, and 0.4600 for the liver. Figure 3 shows a consistent trend across all methods in most cases; the results of the evaluation factor tended to decrease as the threshold value increased. This trend shows the difficulty of reconstructing high-value voxels within the volume accurately. Furthermore, the proposed VI-GAN consistently achieved superior results in most cases. Consequently, the VI-GAN demonstrated higher performance in reconstructing high-value voxels than the existing methods.

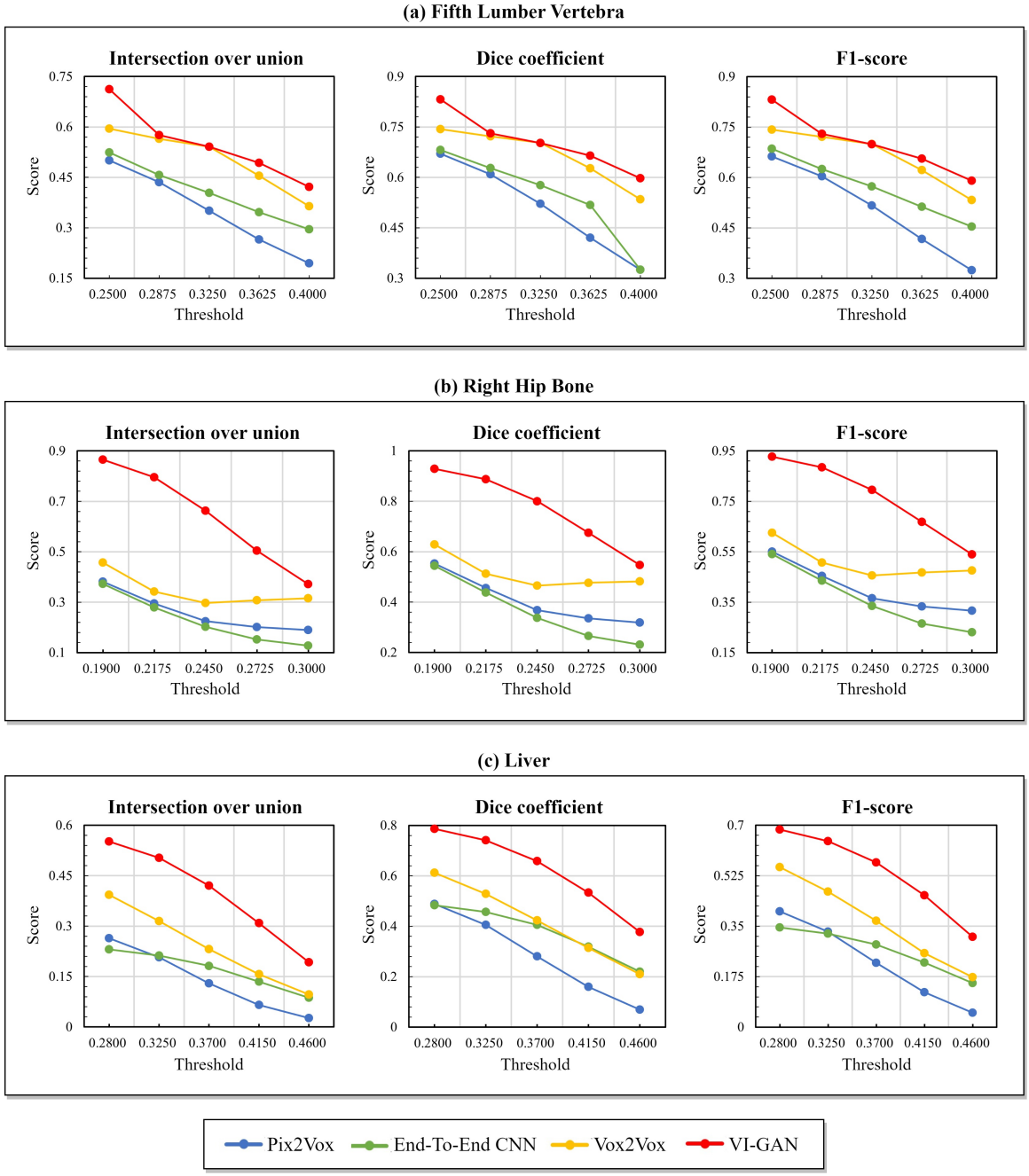


**Figure 3.** Quality comparison between the volumes of the proposed VI-GAN and existing methods across a range of thresholds.

## 5. Conclusions

This paper proposed a VI-GAN for generating a volumetric model to describe the human anatomical model using a GAN-based volume generator and discriminator with 3D LFFB. This paper also proposed reconstruction loss, including feature loss and similarity loss, to reconstruct high-value areas and describe the essential characteristics of the model accurately. The experimental result showed that the generated volume of the VI-GAN represents the shape, high-value areas, and internal part better than the other existing methods evaluated. Furthermore, the experimental results with varying threshold values showed that the VI-GAN accurately generates high-value areas compared to existing methods. Based on these results, VI-GANs may enhance the availability of medical data and improve the training efficiency of medical professionals by generating high-quality volumetric data.

**Author Contributions:** Conceptualization, B.-S.S.; methodology, J.K., Y.L. and B.-S.S.; software, J.K.; validation, J.K. and Y.L.; formal analysis, J.K., Y.L. and B.-S.S.; investigation, J.K. and Y.L.; resources, J.K. and B.-S.S.; data curation, J.K.; writing—original draft preparation, J.K.; writing—review and editing, J.K., Y.L., and B.-S.S.; visualization, J.K.; supervision, Y.L. and B.-S.S.; project administration, B.-S.S.; funding acquisition, B.-S.S. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was supported by National Research Foundation of Korea (NRF) grants funded by the Korean government (No. NRF-2022R1A2B5B01001553 and No. NRF-2022R1A4A1033549) and Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korean government (MSIT) (No. RS-2022-00155915, Artificial Intelligence Convergence Innovation Human Resources Development (Inha University)).

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The spine CT images from the Digital Korean dataset are not accessible to the public. Please contact the Korea Institute of Science & Technology Information (KISTI) to obtain the data. The liver CT images from CT-ORG are publicly accessible and are available at https://wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=61080890 (accessed on 30 January 2024).

**Acknowledgments:** The authors gratefully acknowledge the human data support provided by the Korea Institute of Science & Technology Information (KISTI), which produced these data with the Catholic University of Medicine.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## References

1. Li, R.; Huang, T.; Zhang, X.; Liao, H. 4: Interactive Volume Rendering Method Using Dynamic Ray Casting for Autostereoscopic Display. In *SID Symposium Digest of Technical Papers*; Wiley Online Library: Hoboken, NJ, USA, 2021; Volume 52, pp. 26–29.
2. Fang, C.; An, J.; Bruno, A.; Cai, X.; Fan, J.; Fujimoto, J.; Golfieri, R.; Hao, X.; Jiang, H.; Jiao, L.R.; et al. Consensus recommendations of three-dimensional visualization for diagnosis and management of liver diseases. *Hepatol. Int.* **2020**, *14*, 437–453. [CrossRef]
3. Nakao, M.; Nakamura, M.; Mizowaki, T.; Matsuda, T. Statistical deformation reconstruction using multi-organ shape features for pancreatic cancer localization. *Med. Image Anal.* **2021**, *67*, 101829. [CrossRef]
4. Kavur, A.E.; Gezer, N.S.; Barış, M.; Aslan, S.; Conze, P.H.; Groza, V.; Pham, D.D.; Chatterjee, S.; Ernst, P.; Özkan, S.; et al. CHAOS challenge-combined (CT-MR) healthy abdominal organ segmentation. *Med. Image Anal.* **2021**, *69*, 101950. [CrossRef]
5. Zhao, J.; Zhang, Y.; He, X.; Xie, P. Covid-ct-Dataset: A ct Scan Dataset about COVID-19. *arXiv* **2020**, arXiv:2003.13865.
6. Wisse, L.E.; Chételat, G.; Daugherty, A.M.; de Flores, R.; la Joie, R.; Mueller, S.G.; Stark, C.E.; Wang, L.; Yushkevich, P.A.; Berron, D.; et al. Hippocampal subfield volumetry from structural isotropic 1 mm3 MRI scans: A note of caution. *Hum. Brain Mapp.* **2021**, *42*, 539–550. [CrossRef]
7. Costa, P.; Galdran, A.; Meyer, M.I.; Niemeijer, M.; Abràmoff, M.; Mendonça, A.M.; Campilho, A. End-to-end adversarial retinal image synthesis. *IEEE Trans. Med. Imaging* **2017**, *37*, 781–791. [CrossRef]
8. Beers, A.; Brown, J.; Chang, K.; Campbell, J.P.; Ostmo, S.; Chiang, M.F.; Kalpathy-Cramer, J. High-resolution medical image synthesis using progressively grown generative adversarial networks. *arXiv* **2018**, arXiv:1805.03144.

9. Balashova, E.; Wang, J.; Singh, V.; Georgescu, B.; Teixeira, B.; Kapoor, A. 3D Organ Shape Reconstruction from Topogram Images. In Proceedings of the International Conference on Information Processing in Medical Imaging, Hong Kong, China, 2–7 June 2019; Springer: Berlin/Heidelberg, Germany, 2019; pp. 347–359.
10. Kasten, Y.; Doktofsky, D.; Kovler, I. End-to-end convolutional neural network for 3D reconstruction of knee bones from bi-planar X-ray images. In Proceedings of the International Workshop on Machine Learning for Medical Image Reconstruction, Lima, Peru, 8 October 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 123–133.
11. Çiçek, Ö.; Abdulkadir, A.; Lienkamp, S.S.; Brox, T.; Ronneberger, O. 3D U-Net: Learning dense volumetric segmentation from sparse annotation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Athens, Greece, 17–21 October 2016; Springer: Berlin/Heidelberg, Germany, 2016; pp. 424–432.
12. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
13. Jiang, M.; Zhi, M.; Wei, L.; Yang, X.; Zhang, J.; Li, Y.; Wang, P.; Huang, J.; Yang, G. FA-GAN: Fused attentive generative adversarial networks for MRI image super-resolution. *Comput. Med. Imaging Graph.* **2021**, *92*, 101969. [CrossRef] [PubMed]
14. Odena, A.; Olah, C.; Shlens, J. Conditional image synthesis with auxiliary classifier gans. In Proceedings of the International Conference on Machine Learning, PMLR, Sydney, Australia, 6–11 August 2017; pp. 2642–2651.
15. Setiadi, D.R.I.M. PSNR vs. SSIM: Imperceptibility quality assessment for image steganography. *Multimed. Tools Appl.* **2021**, *80*, 8423–8444. [CrossRef]
16. Lee, S.H.; Lee, S.B. Production and usage of Korean human information in KISTI. *J. Korea Contents Assoc.* **2010**, *10*, 416–421. [CrossRef]
17. Rister, B.; Yi, D.; Shivakumar, K.; Nobashi, T.; Rubin, D.L. CT-ORG, a new dataset for multiple organ segmentation in computed tomography. *Sci. Data* **2020**, *7*, 381. [CrossRef]
18. Dai, W.C.; Zhang, H.W.; Yu, J.; Xu, H.J.; Chen, H.; Luo, S.P.; Zhang, H.; Liang, L.H.; Wu, X.L.; Lei, Y.; et al. CT imaging and differential diagnosis of COVID-19. *Can. Assoc. Radiol. J.* **2020**, *71*, 195–200. [CrossRef] [PubMed]
19. Byl, J.; Nelliadi, S.S.; Samuel, B.; Vettukattil, J. True 3D Viewer facilitates accurate diagnosis of lung infarction. *Vasc. Dis. Manag.* **2021**, *18*, E267–E268.
20. Shi, W.; Liu, P.X.; Zheng, M. Cutting procedures with improved visual effects and haptic interaction for surgical simulation systems. *Comput. Methods Programs Biomed.* **2020**, *184*, 105270. [CrossRef] [PubMed]
21. Munawar, A.; Li, Z.; Nagururu, N.; Trakimas, D.; Kazanzides, P.; Taylor, R.H.; Creighton, F.X. Fully Immersive Virtual Reality for Skull-base Surgery: Surgical Training and Beyond. *arXiv* **2023**, arXiv:2302.13878.
22. Henzler, P.; Rasche, V.; Ropinski, T.; Ritschel, T. Single-image Tomography: 3D Volumes from 2D Cranial X-Rays. In *Computer Graphics Forum*; Wiley Online Library: Hoboken, NJ, USA, 2018; Volume 37, pp. 377–388.
23. Xu, Y.; Li, Y.; Shin, B.S. Medical image processing with contextual style transfer. *Hum.-Centric Comput. Inf. Sci.* **2020**, *10*, 46. [CrossRef]
24. Qiao, Z.; Qian, Z.; Tang, H.; Gong, G.; Yin, Y.; Huang, C.; Fan, W. CorGAN: Context aware Recurrent Generative Adversarial Network for Medical Image Generation. In Proceedings of the 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Seoul, Republic of Korea, 16–19 December 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 1100–1103.
25. Wang, Q.; Zhang, X.; Zhang, W.; Gao, M.; Huang, S.; Wang, J.; Zhang, J.; Yang, D.; Liu, C. Realistic lung nodule synthesis with multi-target co-guided adversarial mechanism. *IEEE Trans. Med. Imaging* **2021**, *40*, 2343–2353. [CrossRef] [PubMed]
26. Masutani, E.M.; Bahrami, N.; Hsiao, A. Deep learning single-frame and multiframe super-resolution for cardiac MRI. *Radiology* **2020**, *295*, 552–561. [CrossRef] [PubMed]
27. Kitchen, A.; Seah, J. Deep generative adversarial neural networks for realistic prostate lesion MRI synthesis. *arXiv* **2017**, arXiv:1708.00129.
28. Chuquicusma, M.J.; Hussein, S.; Burt, J.; Bagci, U. How to fool radiologists with generative adversarial networks? A visual turing test for lung cancer diagnosis. In Proceedings of the 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), Washington, DC, USA, 4–7 April 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 240–244.
29. Frid-Adar, M.; Diamant, I.; Klang, E.; Amitai, M.; Goldberger, J.; Greenspan, H. GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification. *Neurocomputing* **2018**, *321*, 321–331. [CrossRef]
30. Karras, T.; Aila, T.; Laine, S.; Lehtinen, J. Progressive growing of gans for improved quality, stability, and variation. *arXiv* **2017**, arXiv:1710.10196.
31. Wu, J.; Zhang, C.; Xue, T.; Freeman, B.; Tenenbaum, J. Learning a probabilistic latent space of object shapes via 3D generative-adversarial modeling. *Adv. Neural Inf. Process. Syst.* **2016**, *29*, 2–9.
32. Smith, E.J.; Meger, D. Improved adversarial systems for 3D object generation and reconstruction. In Proceedings of the Conference on Robot Learning, PMLR, Mountain View, CA, USA, 13–15 November 2017; pp. 87–96.
33. Kruthika, K.; Maheshappa, H.; Initiative, A.D.N. CBIR system using Capsule Networks and 3D CNN for Alzheimer's disease diagnosis. *Inform. Med. Unlocked* **2019**, *14*, 59–68. [CrossRef]
34. Xu, C.; Xu, L.; Ohorodnyk, P.; Roth, M.; Chen, B.; Li, S. Contrast agent-free synthesis and segmentation of ischemic heart disease images using progressive sequential causal GANs. *Med. Image Anal.* **2020**, *62*, 101668. [CrossRef]
35. Ran, M.; Hu, J.; Chen, Y.; Chen, H.; Sun, H.; Zhou, J.; Zhang, Y. Denoising of 3D magnetic resonance images using a residual encoder—Decoder Wasserstein generative adversarial network. *Med. Image Anal.* **2019**, *55*, 165–180. [CrossRef] [PubMed]

36. Han, C.; Rundo, L.; Murao, K.; Noguchi, T.; Shimahara, Y.; Milacski, Z.Á.; Koshino, S.; Sala, E.; Nakayama, H.; Satoh, S. MADGAN: Unsupervised medical anomaly detection GAN using multiple adjacent brain MRI slice reconstruction. *BMC Bioinform.* **2021**, *22*, 31. [CrossRef]
37. Cirillo, M.D.; Abramian, D.; Eklund, A. Vox2Vox: 3D-GAN for brain tumour segmentation. In Proceedings of the International MICCAI Brainlesion Workshop, Lima, Peru, 4–8 October 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 274–284.
38. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
39. Visa, S.; Ramsay, B.; Ralescu, A.L.; Van Der Knaap, E. Confusion matrix-based feature selection. *Maics* **2011**, *710*, 120–127.
40. Taha, A.A.; Hanbury, A. Metrics for evaluating 3D medical image segmentation: Analysis, selection, and tool. *BMC Med. Imaging* **2015**, *15*, 29. [CrossRef]
41. Roschger, P.; Gupta, H.; Berzlanovich, A.; Ittner, G.; Dempster, D.; Fratzl, P.; Cosman, F.; Parisien, M.; Lindsay, R.; Nieves, J.; et al. Constant mineralization density distribution in cancellous human bone. *Bone* **2003**, *32*, 316–323. [CrossRef]
42. Xie, H.; Yao, H.; Sun, X.; Zhou, S.; Zhang, S. Pix2vox: Context-aware 3d reconstruction from single and multi-view images. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 2690–2698.
43. Bertels, J.; Eelbode, T.; Berman, M.; Vandermeulen, D.; Maes, F.; Bisschops, R.; Blaschko, M.B. Optimizing the dice score and jaccard index for medical image segmentation: Theory and practice. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Shenzhen, China, 13–17 October 2019; Springer: Berlin/Heidelberg, Germany, 2019; pp. 92–100.
44. Wang, Z.; Bovik, A.C. A universal image quality index. *IEEE Signal Process. Lett.* **2002**, *9*, 81–84. [CrossRef]
45. Zhang, L.; Shen, Y.; Li, H. VSI: A visual saliency-induced index for perceptual image quality assessment. *IEEE Trans. Image Process.* **2014**, *23*, 4270–4281. [CrossRef] [PubMed]