Volumetric Feature Learning for High-Fidelity Two-Dimensional Dental Cast Image Reconstruction Using Generative Adversarial Networks (GANs)

Eman Ahmed Eldaoushy[1]*, Manal A. Abdel-Fattah[2], Nermeen Ahmed Hassan[3], Mai M. El defrawi[4]
Faculty of Computers and Artificial Intelligence-Department of Information Systems, Helwan University, Cairo, Egypt[1, 2, 4]
Faculty of Dentistry-Department of Prosthodontics, Cairo University, Giza, Egypt[3]

Abstract—**Dentistry is a medical branch that diagnoses and treats oral diseases, helps maintain oral function, and improves oral aesthetics. Dental casts are three-dimensional models of a patient's oral tissues that can be used to study oral anatomy, assess occlusal relationships, and determine tooth alignment. Traditionally, they were made of gypsum, an impression material used to pour into the patient's mouth molds. Meanwhile, digital ones are three-dimensional models generated virtually using modern digital imaging and intraoral scanners. Unlike physical models, which require a lot of manual work and ample storage space, digital models can be produced rapidly, easily modified, and stored for long-term usage. In this study, we present Denta-RecGAN, a novel approach based on Generative Adversarial Networks (GANs) that maps a two-dimensional dental cast image into a volumetric latent space and projects it back into a two-dimensional output. The proposed approach employs a 2D encoder to process dental cast images as input, enabling the extraction of spatial features. The structural depth is modelled, and noise is suppressed using volumetric 3D latent space denoising models; a 2D decoder then reconstructs a high-quality image. The model is trained under an adversarial learning approach using the IO150K dataset. The proposed architecture achieved Mean Absolute Error (MAE) of 0.0128, 0.0127, 0.0128; Structural Similarity Index Measure (SSIM) of 0.9450, 0.9452, 0.9453; and Peak Signal-to-Noise Ratio (PSNR) of 28.84, 28.85, 28.84 decibels across training, validation, and testing sets. These results demonstrate the effectiveness of volumetric feature learning in enhancing the accuracy of 2D image reconstruction and preserving fine structural details.**

Keywords—Dental image reconstruction; generative adversarial networks; latent space representation; two-dimensional to three-dimensional mapping; volumetric deep learning

I. INTRODUCTION

Orthodontists and prosthodontists use dental casts to detect oral diseases, determine the appropriate treatment for each patient, and help manufacture a precisely suitable appliance that is comfortable for patients. These casts are three-dimensional representations of the patient's oral structures, including teeth, gingiva, and other oral tissues. Dental casts can be physical or digital, depending on the method of creation [1]. These models can be used to demonstrate tooth alignment and to evaluate treatment results. The digital model can facilitate communication between dentists and dental laboratories that manufacture custom dental models for comfort [2].

In the past, dentists made dental casts using sticky materials called alginate or polyvinyl siloxane. They would place these materials in the patient's mouth to get the shape of the patient's teeth. Then they made an impression, called a stone model, by pouring plaster or resin into a mold [3]. Although widely used, this method can cause some problems. The materials may shrink or change shape, making it uncomfortable for the person taking the impression [4]. Physical models can break easily and take up a lot of space [5].

Digital technologies have replaced conventional stone models with virtual counterparts, enhancing precision and efficiency. Digital models help ensure measurements are accurate and reduce human mistakes [6]. They can use intraoral scanners, computed tomography (CT) scanners, or other imaging devices to produce detailed digital images that closely resemble real braces [7], [8]. These digital models can be seamlessly integrated into computer-aided design and manufacturing (CAD/CAM) workflows for the design of restorations, surgical guides, and orthodontic appliances [9]. However, high-end 3D scanners remain expensive and technically demanding, limiting their use in low-resource clinical environments [10].

With recent advances in artificial intelligence, particularly deep learning, there has been a growing interest in the use of neural networks for digital dental reconstruction models depending on utilizing Convolutional Neural Networks (CNNs) or employing transformers or using Generative Adversarial Networks (GANs) have shown high accuracy in reconstructing both dental and craniofacial structures from two-dimensional (2D) data [11],[12]. These methods enable the use of limited 2D inputs to infer both structural and volumetric depth, thereby reducing reliance on expensive 3D imaging systems.

This study proposes Denta-RecGAN, leveraging volumetric reasoning via a novel hybrid 2D-3D convolutional neural network designed to reconstruct 2D grayscale dental cast images. Volumetric reasoning is incorporated by combining three main components: a 2D encoder, a 3D latent space denoiser, and a 2D decoder. This combination allows us to capture both the spatial and structural relationships within the data. The spatial features were initially extracted from the input images using 2D convolutional layers, which were then reshaped into a 5D volumetric tensor, and the 3D convolutional layers were then processed together. The structural depth and inter-slice correlations were captured through the network at the volumetric processing stage. We look at flat (two-dimensional) images and

*Corresponding author.

understand their properties. Then we visualize them in three-dimensional space to enhance them. After that, we revert them to flat images, taking care to preserve their details.

This study demonstrates a new method for dentists to create better images of teeth and jaws. This method combines precise spatial reasoning (how three-dimensional objects occupy space) with two-dimensional photos (such as photographs or X-rays).

This research contributes to the growing field of AI-driven digital dentistry, without the need for complex scanning hardware. It also establishes a foundation for future applications, such as single-view 3D estimation, clinical maxillofacial reconstruction, and digital prosthesis design.

The evaluation of our proposed architecture (DentaRecGAN) is performed using three standard quantitative metrics. The average deviation between a guess and the true answer was measured by the Mean Absolute Error (MAE). It considers all guesses, determines how different they are from the real numbers, ignores whether they are too high or too low, and then calculates the average of these differences, which measures perceptual similarity and is captured by calculating the Structural Similarity Index Measure (SSIM). To understand how sharp and detailed an image is after editing or modification, we used the peak signal-to-noise ratio (PSNR). Furthermore, the proposed network does not need explicit three-dimensional supervision, reducing dependence on complex 3D ground truth datasets, and it is capable of learning volumetric cues.

The remainder of our research is divided into the following sections: Section II provides an overview of current studies on two-dimensional-to-three-dimensional and the reverse reconstruction approaches. Then Section III explains how to train a computer. It covers data preparation, cleaning, computer training, setting training rules, and verifying mastery of the learning process. Section IV compares the proposed architecture with other relevant work. Section V presents the study's conclusions, and Section VI discusses potential directions for future research to improve the clinical accuracy of three-dimensional model reconstruction.

II. Literature Review

The main advantage of deep learning in maxillofacial prosthetics is the reconstruction of three-dimensional (3D) models from two-dimensional (2D) inputs, such as images or videos. Various deep learning architectures play an essential role in the medical domain, especially the dental one, which categorizes studies into sections. Some studies use Generative Adversarial Networks (GANs), transformers, and Convolutional Neural Networks (CNNs) as basic building blocks. The first section presents studies that focus on converting two-dimensional inputs into three-dimensional outputs, and the second section presents studies that concentrate on reverse reconstruction, converting three-dimensional input data into two-dimensional outputs:

A. Two-Dimensional (2D) -To-Three-Dimensional (3D) Reconstructions

Recently, deep neural networks have led to the emergence of 3D models or meshes from single-view two-dimensional (2D) images. We use these deep neural networks to generate 3D volumes from 2D images. However, these methods often require synthetic datasets and 3D supervision. X. Zhang et al. [13] used Convolutional Neural Networks (CNNs) by introducing PX2Tooth, which is a model that uses a one panoramic image to create a three-dimensional point cloud of a tooth. The proposed architecture consists of two fundamental stages. The first stage uses a single panoramic X-ray (PX) image to segment two permanent teeth. To enhance the generation quality, particularly in the root apex region, the Tooth Generation Network (TGNet) is utilized to create 3D teeth from point clouds. The authors also used a dataset that they created themselves. This dataset consists of 499 CBCT and panoramic X-ray pairs. They split their dataset into three subsets with an 8:1:1 ratio for training, validation, and testing, respectively. They achieved an intersection over union (IoU) of 0.793 with their model but could improve results by increasing reconstruction accuracy.

Many studies have reconstructed three-dimensional (3D) models using single or multiple 2D images. Fathallah et al. [14] used a model architecture based on Graph Convolutional Networks (GCNs) as a key component. A lightweight GCN-based discriminator was used to improve the accuracy. The authors used the 300 W-LP and AFLW2000-3D datasets for the evaluation. Their architecture is divided into two fundamental stages: preprocessing and three-dimensional reconstruction. First, the preprocessing stage reduces Noise and augments the data. Then, the three-dimensional reconstruction stage performs the following main functions: triangulation, running the GCN-IGAN model, and outputting the final three-dimensional facial mesh. Their model achieved 0.0075 and 0.120, representing the chamfer distance and earth mover's distance, respectively. Their model can be improved by adding additional features and using evolutionary algorithms. The need to employ evolutionary algorithms and integrate additional face features still limits the applicability of their model.

Chenfan Xu et al. [15] introduced a framework using Convolutional Neural Networks (CNNs) to reconstruct both the upper and lower teeth using five intraoral photographs per case. They used 3,200 cases for their approach, as they used 3,000,100 and 100 for training, validation, and testing, respectively. Their model achieved 18.85, 0.8347, 0.0114, 2.1126, 0.1670 and 0.4122 representing PSNR, SSIM, LPIPS, Hausdorff distance, chamfer distance, and Intersection over Union (IOU) respectively. Their model showed promising quantitative results and can be further improved by extracting low-level features, such as edges. Their model suffered from reduced precision and a long reconstruction process time.

X. Wang et al. [16] used Convolutional Neural Networks (CNNs) to focus on maxillary segmentation and defect refinement. They also created their own dataset using CBCT scans of 60 patients, including 39 males and 21 females, with an average age ranging between 11 and 52 years.

Their model achieved 0.92 ±0.01 and 0.77 ±0.06, representing the Dice Similarity Coefficients for the maxilla and the Dice Similarity Coefficients for the defect, respectively. These results show strong segmentation performance and a correlation between the defect parameters and the maxillary cleft side. This automatic segmentation requires orthodontic refinement, as it takes approximately 5 minutes per CBCT image.

Marek Wodzinski et al. [17] used Convolutional Neural Networks (CNNs) to develop a 3D printing pipeline for model cranial implants using a U-Net architecture to reconstruct defects and refine implants through iterative procedures. Their pipeline consisted of five stages: data loading, preprocessing, defect reconstruction, defect refinement, and 3D printing preparation. They used the SkullBreak, real cranial defect, and SkullFix datasets for evaluation. Their model achieved 0.91, 0.94, and 1.53 mm representing Dice Coefficient, Boundary Dice Coefficient, and 85th percentile Hausdorff Distance, respectively. Their model accurately reconstructed cranial defects, but they could further enhance it by integrating mixed reality, real defect data, and multiple implant reconstructions.

T. C. Niño-Sandoval et al. [18] employed a Procrustes fit on 55 tomographs to obtain the 3D mandibular shape, using convolutional neural networks (CNNs) to analyze the images. The software developers also collected 629 X-ray images to train the computer. When tested, the software performed exceptionally well at matching the photos, with error rates ranging from 0.0033 to 0.0059. These results showed that the model could infer the shape of the lower jaw. It has high accuracy, but it still faces challenges, including difficulty handling significant bone defects and extensive mandibular deformities.

Y. Liang et al. [19] designed an oral viewer as an educational tool with interactive 3D visualizations, using Convolutional Neural Networks (CNNs) and two-dimensional dental panoramic X-rays to reconstruct three-dimensional models of teeth, gums, and the jawbone. Their dataset was collected from patients at an orthodontic hospital and consisted of panoramic X-rays and cone-beam CT (CBCT) scans. Their model achieved an Intersection Over Union (IOU) of 0.771, representing reconstruction accuracy. The authors proposed several future methods to enhance their tool, including two approaches for modeling the root canal: augmenting existing solid tooth models with artificial canals and incorporating canal structures into the convolutional network training process, and adding additional virtual instruments to improve surgical simulation.

B. Three-Dimensional (3D) -To-Two-Dimensional (2D) Reconstructions

Autoencoders are used in image processing tasks, including image denoising, image compression, and generation. Variants of autoencoders, such as U-Nets and convolutional autoencoders, can preserve spatial details useful for medical imaging.

Some methods use volumetric or geometric reasoning for a three-dimensional (3D) shape. Our idea is to develop this technique by adding a special three-dimensional twisted block to a simple two-dimensional encryption and decryption process, making it more secure. Our approach builds on this idea, embedding a 3D convolutional block within a 2D encoder-decoder pipeline.

Melas-Kyriazi et al. [20] presented a novel method, called Projection-Conditioned Point Cloud Diffusion (PC2), for single-image three-dimensional reconstruction. The shapes can be represented as point clouds by their framework. They started with a collection of tiny dots floating in space, which came together to form a three-dimensional shape, much like a jigsaw puzzle. First, they cleaned the dots to ensure their accuracy. Then, they examined small parts of the image to make sure everything looked correct, which helped them build a better three-dimensional model. They also guessed the proper colors for the dots to make the model look more realistic. Sometimes, they created several possible shapes and used a special method to choose the best one. Their method proved more successful than others, and their model showed qualitative improvements on real-world datasets such as Co3D. However, their method depends on ground-truth point clouds for training. The reconstruction quality can be affected because multi-view methods, like COLMAP, can be noisy and incomplete in real-world scenarios.

Peng et al. [21] introduced a graph-based framework for detecting changes in buildings using bitemporal remote sensing images. The spatial dependencies between neighboring buildings and the temporal relationships between image pairs are modelled using spatial–temporal graph neural networks (ST-GNNs). They constructed a graph by representing building instances as nodes and the contextual relationships as edges. Their method of tracking building changes was superior to other methods that relied on analyzing small image fragments. It was more accurate when using numerous city images and more effective in complex urban areas where buildings overlap or are constructed in unusual ways. However, their model suffers from a significant limitation: dependence on accurate building footprint extraction. A preprocessing step error can negatively affect detection performance, and any mistake in footprint delineation can be propagated during graph construction.

C. Generative Adversarial Networks-Based Architecture in Dental Imaging

Toscano et al. [22] proposed a hybrid point cloud completion framework for dental molds that integrates symmetry-based data augmentation, iterative latent-space GANs, and a hybrid AE-RL GAN completion strategy. The dataset consisted of 45-point clouds of real lower-jaw teeth. This dataset is downsampled to 2048 points. These training data were expanded using mirroring and point cloud recombination. It also expanded using iterative IGAN augmentation, yielding 49 additional high-quality samples. They used Chamfer Distance (CD) as a metric. This metric showed the importance of each module, as after removing data filtering, the average Chamfer Distance increased by 38.34%. The Chamfer Distance increases by 43.20% after removing iterative I-GAN augmentation. Then, it increased by 13.42% after removing the RL-GAN module, and then increased by 5.34% after removing hybrid selection. The increasing Chamfer Distance indicates a reduction in geometric error, particularly in high-missing-rate scenarios. This approach still suffers from an inability to generalize to non-symmetric anatomical structures. It also suffers from being computationally expensive and still depends on bilateral symmetry assumptions.

Kim et al. [23] proposed a GAN-based framework. It can enhance the accuracy of tooth segmentation, especially in full-arch intraoral scans that are affected by occlusal artifacts. Their framework contains 3 main steps. The first step is to manually remove the occluded interdental regions from the 3D scan data. The second step is to slice the cleaned scan at 0.1 mm intervals, then complete the 2D image using an Edge Connect-based GAN. The third step is to reconstruct the missing 3D geometry by stacking and remeshing the completed slices. They used a dataset of intraoral scans from 10 orthodontic patients acquired with a Trios 3 scanner. The ground truth is generated by a technician. The dataset consists of 10,000 cropped 256 X 256 images. They achieved 0.921 and 26.68 Db representing SSIM and PSNR, respectively. In tooth segmentation evaluation, their model achieved 0.027 ± 0.007 mm, representing the average mean surface distance. Previous boundary- and region-based segmentation methods suffer from inaccuracies due to occlusions. However, their approach reconstructs the missing interdental geometry while reducing operator dependency. However, their model incurs a high computational cost due to manual mask detection. Generalization is limited, especially in severe occlusion patterns.

Minhas et al. [24] proposed a deep learning-based framework for 3D reconstruction from a single 2D panoramic X-ray to assess maxillary impacted canines. They proposed a GAN-based architecture (Pan2CBCT) derived from X2CT-GAN, which expands 2D panoramic images into pseudo-3D volumetric images. They used a dataset comprising 123 pre-treatment CBCT scans of individuals aged 11-18 years. The 2D panoramic X-rays with their pseudo-3D images. The distribution of impacted canines is divided into 36, 12, 26, 65, and 9, representing buccal, middle, lingual, mesial, and distal cases, respectively. Their model achieved 0.71, 41%, and 55% for mean SSIM, accuracy of buccal/middle/lingual position, and accuracy of mesial/distal position, respectively. The previous related work clinically ignores complex cases, as impacted canines are evaluated. Their SSIM values indicate insufficient reliability for orthodontic diagnosis. However, their model suffers from several limitations. The first limitation is the use of a small, imbalanced dataset. Another limitation is that it depends on a single image modality. The third limitation is the decrease in performance in lingual positions.

Galba et al. [25] proposed HoloDent3D, a dental imaging system that uses single-view panoramic radiographs for 3D reconstruction. The first stage is to acquire and preprocess the standard orthopantomogram (OPG) to optimize the input image. In the second stage, the reconstruction module is trained on large datasets of paired 2D radiographs and corresponding 3D jaw models, inferring a volumetric mesh of teeth, roots, and bone. The third stage involves rendering a high-speed LED holographic fan display with gesture control, enabling visualization of multi-angle anatomy. Their model achieved volumetric Intersection Over Union (IOU) values ranging from 0.65 to 0.79, showing an n improvement over earlier voxel-based approaches such as X2Teeth. However, HolodENT3D remains at a theoretical stage and faces challenges, including a lack of paired 2D-3D training datasets, variability in generalization across patients, and no clinical validation.

III. Methodology

This section explains the full methodology implemented in TensorFlow/Keras for grayscale image reconstruction using a 2D encoder, a 3D latent space denoiser, and a 2D decoder, trained under an adversarial framework. The methodology of this study was divided into six main phases: dataset preparation, data pipeline, model architecture, training strategy, and evaluation metrics.

A. Data Preparation

Using the IO150K dataset [22], the dataset consisted of single-channel (grayscale) images. It is a publicly available dataset of 2D intraoral images that can be used for different purposes, such as instance segmentation and semantic labeling. The dataset includes over 150,000 2D intraoral images. It consists of 3 subsets: challenge80k, plaster 70k images, and rendered 2D images generated from 1800 3D intraoral scans. Its main source is the 3D Teeth Segmentation and Labeling Challenge 2023. The data acquisition methods for the challenge, which generate 3D dental scans, project them into multiple 2D views. Plaster 70k images are collected by photographing real dental plaster casts, thereby capturing realistic tooth morphology, including spacing, crowding, and missing teeth. Clinical RGB photography is acquired using DCLR or mobile cameras, with reflections. Preprocessing steps are performed on the dataset, including standardization, patch embedding preparation, foreground-background separation, localization, segmentation, and labeling. All the photos were set to a fixed spatial resolution of 128 * 128 pixels. Image files were collected recursively with valid extensions (.jpg, .jpeg, .png) and then shuffled. The dataset was divided into 70%, 10%, and 20% for the training, validation, and test sets, respectively. The weights are updated using the training subset, while the validation subset is used to refine hyperparameters and determine when to stop training. The test subset is then used to assess the model's final performance. Table I describes the dataset's distribution.

TABLE I. Dataset Splits for Training, Validation, and Testing

Subset	Number of images	Percentage
Training Set	49612	70%
Validation Set	7087	10%
Testing Set	14176	20%

B. Data Pipeline

Images were read from the disk and decoded into single-channel tensors; each image was then resized to 128 x 128 pixels. Normalization is applied to stabilize gradients during optimization. Normalization was also used, as pixel intensities are scaled to [0,1]. The dataset then maps each image to an (input, target) tuple. The training data were shuffled, and batched prefetching is used as an asynchronous prefetch that overlaps I/O with GPU computing.

C. Model Architecture

The proposed framework is a hybrid encoder–decoder system that integrates 2D and 3D learning in a generative adversarial setting. Our architecture comprises three main components for several reasons. The procedure starts with the encoder receiving a black-and-white photograph of teeth, which is examined to pinpoint key details, such as edges and lines.

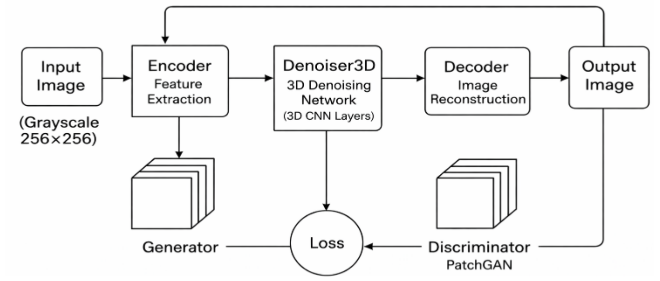


Fig. 1. This is a general overview of the proposed model.

Fig. 1 shows the overall architecture, including the 2D encoder, 3D latent-space denoiser, and 2D decoder, as well as the adversarial training loop that improves reconstruction accuracy while preserving dental structural details. Fig. 2 to Fig. 4 illustrate the detailed network architectures of the proposed framework. Fig. 2 presents the 2D encoder, which progressively downsamples the input grayscale dental image to a compact latent representation. Fig. 3 depicts the 3D latent-space denoiser, which operates on the reshaped volumetric latent representation to model inter-slice dependencies and suppress latent noise. Fig. 4 shows the 2D decoder, which reconstructs the final image by gradually restoring spatial resolution through transposed convolutions. Together, these components form an end-to-end pipeline for robust dental image reconstruction. Let $x \in R^{H\,x\,W}$ denote an input grayscale dental image with H = W = 128. The 2D encoder $E_\theta(.)$, implemented using convolutional layers, extracts hierarchical spatial features from the input image and maps it to a compact latent representation, as in:

$$Z_{2D}{=}E_\theta(x) \qquad ,Z_{2D} \in R^{h\,x\,w\,x\,c} \tag{1}$$

To enable volumetric reasoning without requiring explicit 3D input data, the 2D latent representation is reshaped and expanded along a depth dimension to form a volumetric latent tensor, as in:

$$Z_{3D} = \mathcal{R}(Z_{2D}),Z_{3D} \in R^{d\,x\,h\,x\,w\,x\,\acute{c}} \tag{2}$$

A 3D latent-space denoiser $N_\varphi(.)$, composed of 3D convolutional operations, is applied to the volumetric latent tensor to suppress noise and enforce inter-slice consistency, as in:

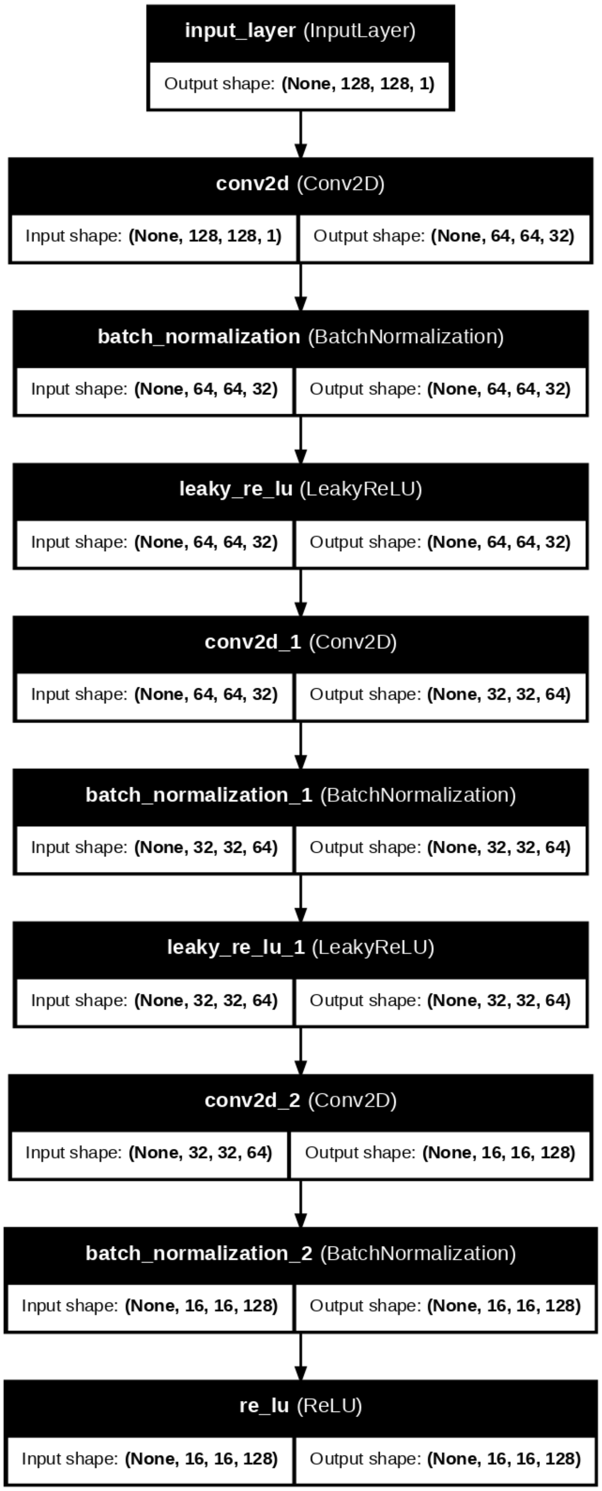
$$\tilde{z}_{3D} = N_\psi(Z_{3D}) \tag{3}$$

The denoised volumetric representation is then collapsed and decoded back into the image domain using a 2D decoder $G_{w(.)}$, as in:

$$\hat{x} = G_w(\tilde{z}_{3D}) \tag{4}$$

where, $\hat{x} \in R^{H\,x\,W}$ denotes the reconstructed dental image. The discriminator receives both real dental images and reconstructed outputs, providing adversarial feedback that encourages perceptually realistic and structurally accurate reconstructions.

We enhanced our architecture, discarding any noise and unwanted distortion. The decoder module further improves image quality. Both authentic dental images and the generator's outputs serve as inputs to the discriminator architecture. The generator is motivated to produce increasingly compelling visual and numerical reconstructions by receiving adversarial feedback. To create more transparent, more realistic images, the denoiser module works in tandem with the main architecture to remove noise and minor imperfections.



input_layer (InputLayer)
Output shape: **(None, 128, 128, 1)**

conv2d (Conv2D)
Input shape: **(None, 128, 128, 1)** Output shape: **(None, 64, 64, 32)**

batch_normalization (BatchNormalization)
Input shape: **(None, 64, 64, 32)** Output shape: **(None, 64, 64, 32)**

leaky_re_lu (LeakyReLU)
Input shape: **(None, 64, 64, 32)** Output shape: **(None, 64, 64, 32)**

conv2d_1 (Conv2D)
Input shape: **(None, 64, 64, 32)** Output shape: **(None, 32, 32, 64)**

batch_normalization_1 (BatchNormalization)
Input shape: **(None, 32, 32, 64)** Output shape: **(None, 32, 32, 64)**

leaky_re_lu_1 (LeakyReLU)
Input shape: **(None, 32, 32, 64)** Output shape: **(None, 32, 32, 64)**

conv2d_2 (Conv2D)
Input shape: **(None, 32, 32, 64)** Output shape: **(None, 16, 16, 128)**

batch_normalization_2 (BatchNormalization)
Input shape: **(None, 16, 16, 128)** Output shape: **(None, 16, 16, 128)**

re_lu (ReLU)
Input shape: **(None, 16, 16, 128)** Output shape: **(None, 16, 16, 128)**

Fig. 2. This is a 2D encoder architecture for extracting a compact latent representation from a grayscale input image.

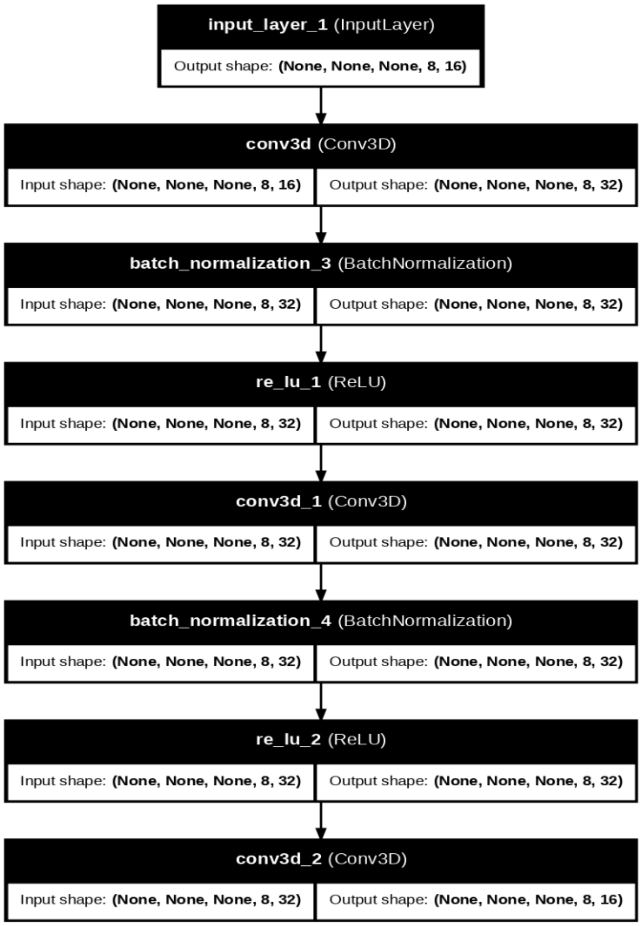

input_layer_1 (InputLayer)
Output shape: (None, None, None, 8, 16)
conv3d (Conv3D)
Input shape: (None, None, None, 8, 16)
Output shape: (None, None, None, 8, 32)
batch_normalization_3 (BatchNormalization)
Input shape: (None, None, None, 8, 32)
Output shape: (None, None, None, 8, 32)
re_lu_1 (ReLU)
Input shape: (None, None, None, 8, 32)
Output shape: (None, None, None, 8, 32)
conv3d_1 (Conv3D)
Input shape: (None, None, None, 8, 32)
Output shape: (None, None, None, 8, 32)
batch_normalization_4 (BatchNormalization)
Input shape: (None, None, None, 8, 32)
Output shape: (None, None, None, 8, 32)
re_lu_2 (ReLU)
Input shape: (None, None, None, 8, 32)
Output shape: (None, None, None, 8, 32)
conv3d_2 (Conv3D)
Input shape: (None, None, None, 8, 32)
Output shape: (None, None, None, 8, 16)

Fig. 3. This is a 3D latent-space denoiser operating on the reshaped volumetric latent representation.

D. Training Strategy

All experiments were conducted on the Kaggle platform. Training the proposed model on the IO150K dataset for 50 epochs with a batch size of 8 required approximately 2 hours, 58 minutes, and 58 seconds using two NVIDIA Tesla T4 GPUs. Training time and inference speed were empirically measured on the same hardware. Inference was evaluated by timing end-to-end forward passes with a batch size of 1, yielding an average latency of 3.41 ms per image (or 27.29 ms per batch), demonstrating near–real-time performance suitable for practical deployment.

Both accuracy and efficiency are balanced during training. Overfitting is also prevented during training. Table II lists the training hyperparameters used in our work.

The primary reconstruction loss function is the Mean Absolute Error (MAE). It was chosen for its robustness in penalizing deviations between the result produced by the architecture and the actual target output.

We used the Adam optimizer in our architecture because it can adapt the learning rate. A dynamically adaptable learning rate. Batch size carefully balances GPU memory constraints and convergence speed; smaller batch sizes enhance generalization capabilities, while larger batches may speed up training but sometimes compromise model robustness.

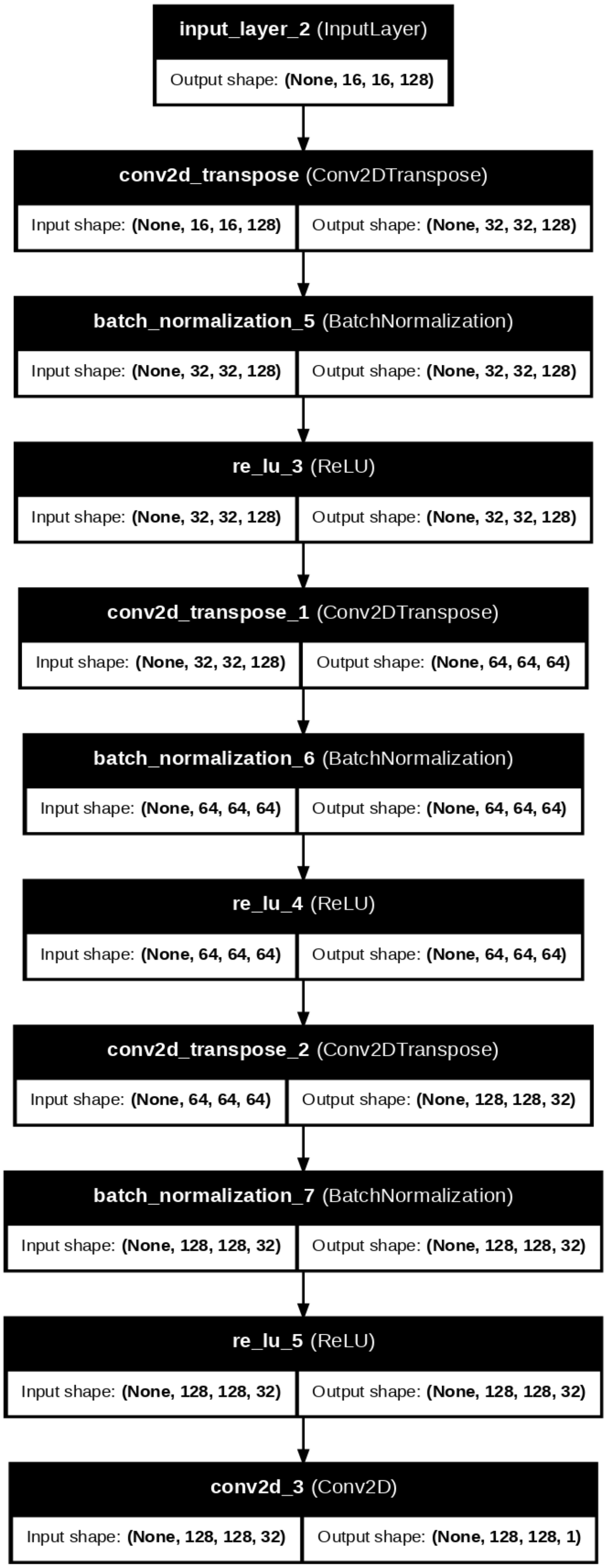

input_layer_2 (InputLayer)
Output shape: (None, 16, 16, 128)
conv2d_transpose (Conv2DTranspose)
Input shape: (None, 16, 16, 128)
Output shape: (None, 32, 32, 128)
batch_normalization_5 (BatchNormalization)
Input shape: (None, 32, 32, 128)
Output shape: (None, 32, 32, 128)
re_lu_3 (ReLU)
Input shape: (None, 32, 32, 128)
Output shape: (None, 32, 32, 128)
conv2d_transpose_1 (Conv2DTranspose)
Input shape: (None, 32, 32, 128)
Output shape: (None, 64, 64, 64)
batch_normalization_6 (BatchNormalization)
Input shape: (None, 64, 64, 64)
Output shape: (None, 64, 64, 64)
re_lu_4 (ReLU)
Input shape: (None, 64, 64, 64)
Output shape: (None, 64, 64, 64)
conv2d_transpose_2 (Conv2DTranspose)
Input shape: (None, 64, 64, 64)
Output shape: (None, 128, 128, 32)
batch_normalization_7 (BatchNormalization)
Input shape: (None, 128, 128, 32)
Output shape: (None, 128, 128, 32)
re_lu_5 (ReLU)
Input shape: (None, 128, 128, 32)
Output shape: (None, 128, 128, 32)
conv2d_3 (Conv2D)
Input shape: (None, 128, 128, 32)
Output shape: (None, 128, 128, 1)

Fig. 4. This is a 2D decoder architecture for reconstructing the output image from the denoised latent features.

TABLE II. HYPERPARAMETERS USED FOR TRAINING THE GAN-BASED DENTAL IMAGE RECONSTRUCTION MODEL STYLES

Hyperparameter	Value
Input image size	128x128(grayscale)
Batch size	8
Epochs(max)	50
Optimizer	Adam
Initial learning rate	2×10^{-4}
Learning Rate Reduction	Factor 0.5 on plateau
Min Learning Rate	1×10^{-6}
Early Stopping Patience	8 epochs (val MAE)
Train/Val/Test split	70%/10%/20%
Latent volume Depth	8 slices
Random seed	42
Base channels(encoder/decoder)	32

Early stopping was performed to avoid overfitting. The training was halted once the validation loss stopped improving after a predefined number of epochs. This strategy prevents unnecessary over-training and ensures computational efficiency.

IV. DISCUSSION AND EVALUATION

In the standard model, the choice of evaluation metrics is crucial for assessing model effectiveness by measuring pixel-level accuracy and perceptual visual similarity, thereby ensuring a comprehensive evaluation of quantitative performance and visual quality.

The first evaluation metric quantifies the average difference between the reconstructed and original images. This difference is measured by the Mean Absolute Error (MAE) metric. This metric helps capture the overall pixel-level accuracy. The second metric is the Structural Similarity Index (SSIM), which measures perceptual similarity by considering the contrast, luminance, and structural information between two images. The SSIM aligns more closely with human visual perception. Peak Signal-to-Noise Ratio (PSNR), serving as the third evaluation metric, quantifies reconstruction fidelity as the ratio of the maximum signal power to the generated noise power during reconstruction. This measures the ratio between the maximum possible signal power and the power of the reconstruction noise, expressed in decibels.

Table III summarizes the three key performance indicators. The results are presented in Table IV, which lists the metrics for training, validation, and testing. Table IV shows results that may appear similar at first glance but, upon closer inspection of the numerical values, reveal small but consistent differences across the three splits. Specifically, MAE values vary between 0.0127 and 0.0128, the SSIM values increase slightly from 0.9450 (training) to 0.9453 (testing), and PSNR values range from 28.8430 dB to 28.8595 dB, so the variations occur at the third and fourth decimal places and may appear identical when rounded, but they confirm that the results are not exactly the same. The dataset was partitioned into non-overlapping subsets prior to training, so no data leakage occurred during either training or evaluation. Moreover, the close alignment between the training and validation MAE curves in Fig. 5 indicates stable learning behavior without divergence, as the model generalizes well, rather than overfitting to the training data. Fig. 5 shows the mean absolute error curves obtained during training and validation. These curves show how the model's prediction error evolves over epochs. The decreasing trend in the training curve indicates the model's ability to learn from the training data, while the validation curve shows the model’s generalization performance on unseen data. These two curves are compared to assess the potential overfitting or underfitting.

TABLE III. EVALUATION METRICS USED FOR ASSESSING THE GAN-BASED DENTAL IMAGE RECONSTRUCTION

Metric	Range	Interpretation
MAE	$[0, \infty)$	Lower is better (0 = perfect reconstruction)
SSIM	$[0, 1]$	Higher is better (1 = perfect perceptual match)
PSNR	$(0, \infty)$ decibels	Higher is better (≥30 decibels = high-quality image)

TABLE IV. EVALUATING PERFORMANCE METRICS OF THE GAN-BASED DENTAL IMAGE RECONSTRUCTION MODEL ON TRAIN, VALIDATION, AND TEST SETS

Set	MAE	SSIM	PSNR
Train	0.0128	0.9450	28.844700 decibels
Validation	0.0127	0.9452	28.859501 decibels
Test	0.0128	0.9453	28.843000 decibels

Fig. 6 to Fig. 10 illustrate various key aspects of the study. It presents quantitative evaluation metrics, including the Structural Similarity Index Measure (SSIM) and Peak Signal-to-Noise Ratio (PSNR), used to assess the quality and fidelity of the reconstructed images. These figures highlight interactions and contributions of the main components of the proposed architecture involving the generator, discriminator, and noise module, so by combining these architectural elements and performance metrics. These figures offer a comprehensive overview of the system's effectiveness in achieving high-quality image reconstruction.

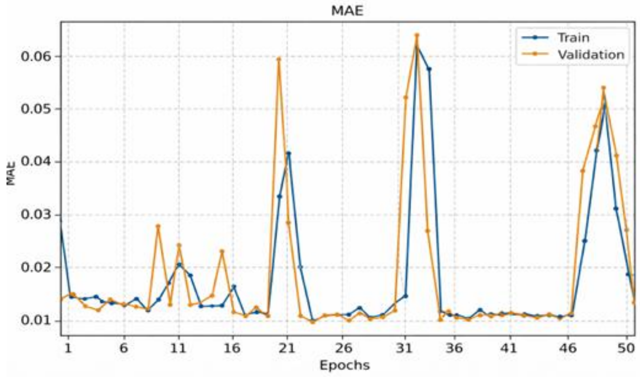


Fig. 5. The two curves represent mean absolute error results during training and validation.

The proposed 2D-to-3D denoising GAN achieved strong reconstruction performance, with average scores of MAE=0.0128, SSIM=0.9453, and PSNR=28.84 dB on the test set. The reconstruction of grayscale dental casts demonstrates accuracy and high structural similarity. There are three figures, namely Fig. 11, Fig. 12, and Fig. 13, representing the input image, the generated output image, and the latent 3D mesh, respectively, at the last epoch (50) after training. Specifically, Fig. 11 shows the original input image used by the model, providing a reference for comparison. Fig. 12 shows the model's output image, demonstrating its reconstruction effectiveness. Finally, Fig. 13 shows the network's learned latent 3D mesh representation, which encodes internal 3D features. These figures present a visual summary of the system's performance, showing the quality of the generated images and the underlying structure of the latent 3D representations after full training.

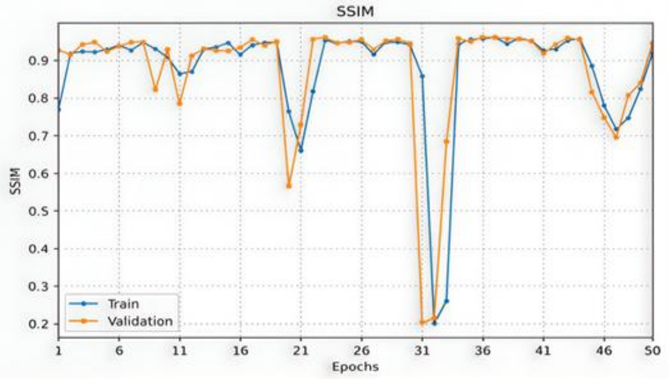


Fig. 6. The two curves represent structural similarity index results during training and validation.

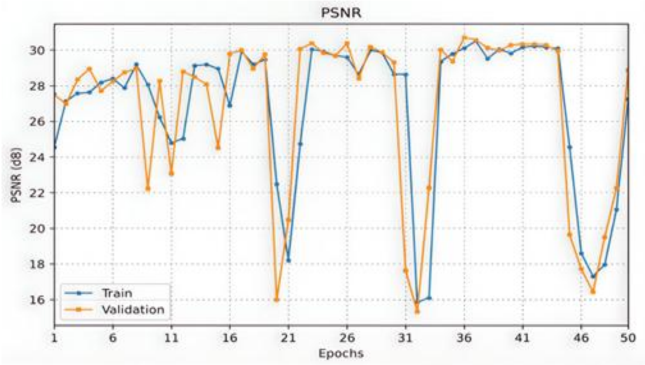


Fig. 7. The two curves represent peak signal-to-noise ratio results during training and validation.

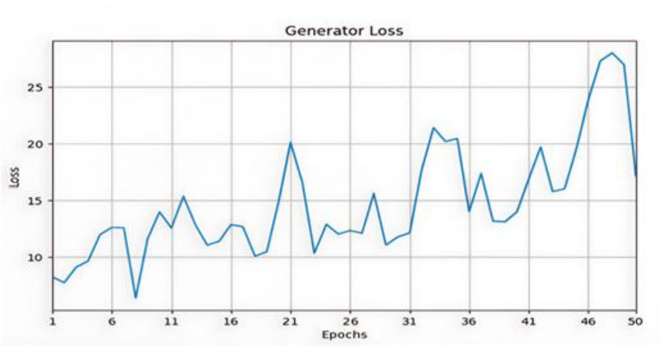


Fig. 8. The curve represents the generator loss.

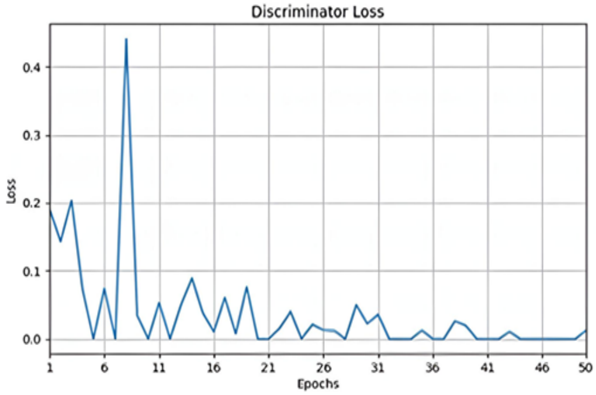


Fig. 9. The curve represents the discriminator loss.

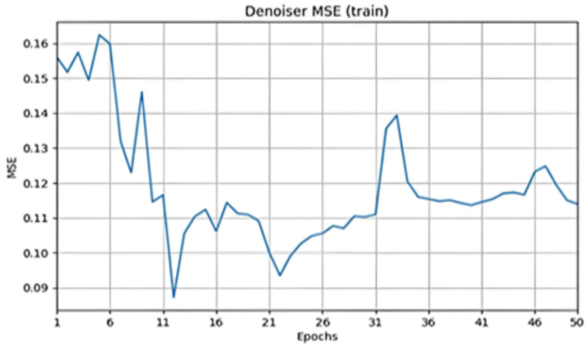


Fig. 10. The curve represents the denoiser.

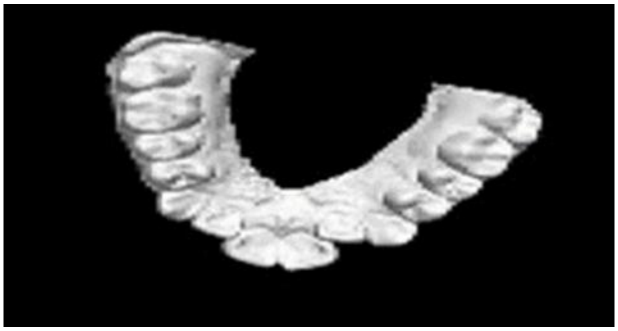


Fig. 11. An example of the input image.

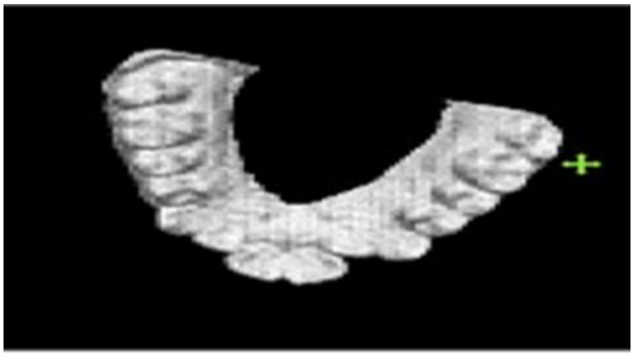


Fig. 12. An example of the generated output image.

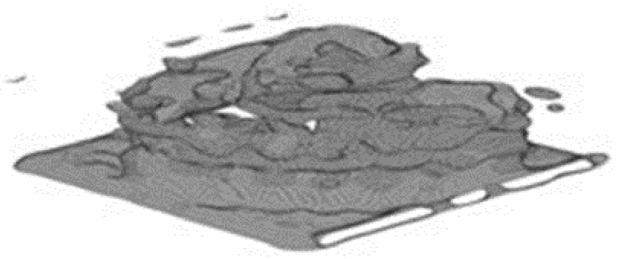


Fig. 13. An example of the three-dimensional latent mesh.

Compared to existing studies, the proposed framework exhibits competitive advantages. Li et al. [27] developed a U-CPML-Net for 3D lung image reconstruction based on CT-pixel-matrix learning and electrical impedance tomography. Their method achieved SSIM values between 0.80 and 0.85, depending on the case complexity. In contrast, the proposed framework produced higher SSIM values and lower MAE values, suggesting that volumetric reasoning integrated into a 2D-to-3D pipeline can outperform conventional matrix learning–based strategies in terms of structural fidelity.

Tan et al. [28] presented an Edge-Aware Reconstruction (EAR) network for reconstructing 3D vertebral structures from biplanar X-ray images. Their method improved edge and local structural detail detection by integrating two modules into the autoencoder, which serves as the backbone of their architecture. A combination of four loss terms—reconstruction, edge, frequency, and projection losses were used to guide the training process. The EAR was evaluated on three public datasets and compared against four cutting-edge methods, demonstrating improvements of 25.32%, 15.32%, 86.44%, 80.13%, 23.76%, and 0.30% in MSE, MAE, Dice, SSIM, PSNR, and frequency distance, respectively. Although this approach effectively improved edge preservation, challenges remain due to information loss from X-ray projection processes, particularly in maintaining. Asymmetrical vertebral structures. Compared with EAR, the proposed denoising GAN focuses less on edge-aware reconstruction but achieves a higher SSIM (0.943) and stable PSNR. Volumetric features effectively compensate for the limitations inherent in projection-based approaches in imaging and analysis. Tables V and VI present the results of other papers and compare them with those of our study, respectively.

TABLE V. SUMMARY OF RELATED WORKS AND THEIR RESULTS

Paper	Work	Results
Li et al. [27]	Developed U-CPML-Net for 3D lung image reconstruction based on CT pixel matrices learning with electrical impedance tomography.	Their method achieved SSIM values between 0.80 and 0.85, depending on case complexity.
Tan et al. [28]	Presented the Edge-Aware Reconstruction (EAR) network for reconstructing 3D vertebrae structures from bi-planar X-ray images.	EAR was evaluated on three publicly available datasets and compared against four state-of-the-art methods, demonstrating improvements of 25.32%, 15.32%, 86.44%, 80.13%, 23.76%, and 0.30% with respect to MSE, MAE, Dice, SSIM, PSNR, and frequency distance, respectively.

TABLE VI. COMPARISON OF THE PROPOSED GAN-BASED DENTAL IMAGE RECONSTRUCTION MODEL WITH RESULTS FROM RELATED WORKS

Paper	Work Description	MAE	SSIM	PSNR
Proposed 2D-to-3D Denoising GAN	Reconstruction of grayscale dental cast images using volumetric feature reasoning (Encoder → 3D Denoiser → Decoder + Discriminator).	0.0128	0.9453	28.84
Li et al. [27]	Developed U-CPML-Net for 3D lung image reconstruction based on CT pixel matrix learning with electrical impedance tomography.	---------------	0.80:0.85	--------------
Tan et al. [28]	Introduced Edge-Aware Reconstruction (EAR) network for 3D vertebrae reconstruction from bi-planar X-ray images using edge and frequency enhancement modules.	---------	≈0.94	≈28.0

V. CONCLUSION

This study presented a two-dimensional-to-three-dimensional denoising generative adversarial network designed to improve the quality of reconstruction, to address the limitations in grayscale image reconstruction of traditional 2D models by incorporating volumetric reasoning within the latent space. We capture both spatial and structural relationships by integrating three main components: a 2D encoder, a 3D denoising block, and a 2D decoder. These spatial and structural relationships exist across the different feature depths. The key advantages of this architecture are a computationally lightweight design and the ability to learn complex spatial dependencies. These advantages make it suitable for medical applications, especially in dentistry. Another key contribution of this study is its adaptable training strategy, which balances between three perspectives: reconstruction accuracy, efficiency, and stability. The model's design enables easy integration with other architectures, allowing extensions to multimodal data. This method can also be widely used, such as for larger image resolutions and domain-specific adaptations. In addition, the adversarial training paradigm can generalize to unseen samples by using a generator that produces structurally consistent outputs under discriminator supervision.

Overall, this study highlights directions for the digital reconstruction of dental cast images in clinical imaging. By integrating advanced volumetric and two-dimensional techniques, it offers promising applications in diagnostic procedures, tissue repair, and various medical interventions. These innovative approaches aim to enhance accuracy, improve patient outcomes, and advance the overall capabilities of modern medical imaging technologies.

VI. FUTURE WORK

Several directions can be pursued to expand the two-dimensional-to-three-dimensional generative adversarial network further. A future perspective is to use multiple input datasets simultaneously to improve resolution and achieve generalization by training across a variety of datasets.

Another promising enhancement is to integrate different ways of focusing attention, such as visual space or channels, which also works well with communication systems. This improves overall accuracy and helps build stronger, more efficient structures.

Another direction is to explore more appropriate loss functions. Perceptual or feature-based losses can be combined with conventional objectives. This combination can lead to a better balance between structural accuracy and perceptual realism.

These directions highlight a future pathway for advancing image reconstruction by pushing the boundaries of hybrid 2D-to-3D-to-2D learning strategies.

Future studies can use adaptive strategies to determine the depth. This depth may be adapted to the complexity of the input, which leads to more effective feature representations while maintaining computational efficiency. The precision of the anatomical structures and texture representation can be improved by refining the denoising process.

This approach shows promise for advancing image restoration methods in fields such as medical imaging, dental diagnosis, industrial inspection, and others where both accuracy and perceptual quality are critical. It can also be extended to real-world applications.

ACKNOWLEDGMENT

The authors acknowledge the use of the IO150K intraoral image dataset, introduced in the Teeth-SEG study, for supporting the development and evaluation of this work. We thank the dataset creators for making this resource openly available [26].

REFERENCES

[1] R. Hendi, M. Moharrami, H. Siadat, H. Hajmiragha, and M. Alikhasi, “The effect of conventional, half-digital, and full-digital fabrication techniques on the retention and apical gap of post and core restorations,” J. Prosthet. Dent., vol. 121, no. 2, pp. 364.e1–364.e6, Dec. 2018, doi: 10.1016/j.prosdent.2018.09.014.

[2] S. E. Barros, E. Ferreira, C. Rösing, G. Janson, and K. Chiqueto, “Expanding torque possibilities: A skeletally anchored torqued cantilever for uprighting ‘kissing molars,” Am. J. Orthod. Dentofacial Orthop., vol. 153, no. 4, pp. 588–598, Apr. 2018, doi: 10.1016/j.ajodo.2017.12.006.

[3] J. S. Park, Y. F. A. Alshehri, E. Kruger, and L. Villata, “Accuracy of digital versus conventional implant impressions in partially dentate patients: A systematic review and meta-analysis,” J. Dent., vol. 160, p. 105918, Sep. 2025, doi: 10.1016/j.jdent.2025.105918.

[4] E. Yuzbasioglu, H. Kurt, R. Turunc, and H. Bilir, “Comparison of digital and conventional impression techniques: Evaluation of patients’ perception, treatment comfort, effectiveness and clinical outcomes,” BMC Oral Health, vol. 14, no. 1, Jan. 2014, doi: 10.1186/1472-6831-14-10.

[5] A. Marghalani, H.-P. Weber, M. Finkelman, Y. Kudara, K. El Rafie, and P. Papaspyridakos, “Digital versus conventional implant impres sions for partially edentulous arches: An evaluation of accuracy,” J. Prosthet Dent., vol. 119, no. 4, pp. 574–579, Sep. 2017, doi: 10.1016/j.prosdent.2017.07.002.

[6] F. Mangano, J. A. Shibli, and T. Fortin, “Digital Dentistry: New Materials and Techniques,” Int. J. Dent., vol. 2016, pp. 1–2, Jan. 2016, doi: 10.1155/2016/5261247.

[7] A. Lennartz, A. Dohmen, S. Bishti, H. Fischer, and S. Wolfart, “Retrievability of implant-supported zirconia restorations cemented on zirconia abutments,” J. Prosthet. Dent., vol. 120, no. 5, pp. 740–746, May 2018, doi: 10.1016/j.prosdent.2018.01.011.

[8] A. Uyar, B. Piskin, B. Senel, H. Avsever, O. Karakoc, and C. Tasci, “Effects of nocturnal complete denture usage on cardiorespiratory parameters: A pilot study,” J. Prosthet. Dent., vol. 128, no. 5, pp. 964–969, Feb. 2021, doi: 10.1016/j.prosdent.2021.01.008.

[9] K. P. Shirley, L. J. Windsor, G. J. Eckert, and R. L. Gregory, “In vitro effects of Plantago major extract, aucubin, and baicalein on Candida albicans biofilm formation, metabolic activity, and cell surface hydrophobicity,” J. Prosthodont., vol. 26, no. 6, pp. 508–515, Nov. 2015, doi: 10.1111/jopr.12411.

[10] K. Hung, A. W. K. Yeung, R. Tanaka, and M. M. Bornstein, “Current applications, opportunities, and limitations of AI for 3D imaging in dental research and practice,” Int. J. Environ. Res. Public Health, vol. 17, no. 12, p. 4424, Jun. 2020, doi: 10.3390/ijerph17124424.

[11] Z. Zhou, J. Zhu, Y. Zhang, X. Guan, P. Wang, and T. Li, “Deep learning in dental image analysis: A systematic review of datasets, methodologies, and emerging challenges,” arXiv, Oct. 23, 2025. doi: 10.48550/arxiv.2510.20634.

[12] S. Ren and X. Li, “HResFormer: Hybrid residual transformer for volumetric medical image segmentation,” IEEE Trans. Neural Netw. Learn. Syst., vol. 36, no. 6, pp. 10558–10566, Jun. 2025, doi: 10.1109/tnnls.2024.3519634.

[13] W. Ma, H. Wu, Z. Xiao, Y. Feng, J. Wu, and Z. Liu, “PX2Tooth: Reconstructing the 3D point cloud teeth from a single panoramic X-ray,” arXiv, Nov. 06, 2024. doi: 10.48550/arxiv.2411.03725.

[14] M. Fathallah, S. Eletriby, M. Alsabaan, M. I. Ibrahem, and G. Farok, “Advanced 3D face reconstruction from single 2D images using enhanced adversarial neural networks and graph neural networks,” Sensors (Basel), vol. 24, no. 19, p. 6280, Sep. 2024, doi: 10.3390/s24196280.

[15] Y. Chen, X. Chen, S. Gao, and P. Tu, “Automatic 3D teeth reconstruction from five intra-oral photos using parametric teeth model,” IEEE Trans. Vis. Comput. Graph., vol. 30, no. 8, pp. 4780–4791, Aug. 2024, doi: 10.1109/tvcg.2023.3277914.

[16] E. Wood et al., “3D face reconstruction with dense landmarks,” arXiv, Apr. 06, 2022. doi: 10.48550/arxiv.2204.02776.

[17] S. Saito, T. Li, and H. Li, “Real-time facial segmentation and performance capture from RGB input,” arXiv, Apr. 10, 2016. doi: 10.48550/arxiv.1604.02647.

[18] T. C. Niño-Sandoval, R. A. Jaque, F. A. González, and B. C. E. Vasconcelos, “Mandibular shape prediction model using machine learning techniques,” Clin. Oral Investig., vol. 26, no. 3, pp. 3085–3096, Jan. 2022, doi: 10.1007/s00784-021-04291-y.

[19] Y. Liang et al., “OralViewer: 3D demonstration of dental surgeries for patient education with oral cavity reconstruction from a 2D panoramic X-ray,” ACM, Apr. 2021, pp. 553–563. doi: 10.1145/3397481.3450695.

[20] L. Melas-Kyriazi, C. Rupprecht, and A. Vedaldi, “PC^2: Projection conditioned point cloud diffusion for single-image 3D reconstruction,” arXiv, Feb. 21, 2023. doi: 10.48550/arxiv.2302.10668.

[21] Y. Mao and K. D. Splinter, “Application of SAR-optical fusion to extract shoreline position from cloud-contaminated satellite images,” ISPRS J. Photogramm. Remote Sens., vol. 220, pp. 563–579, Feb. 2025, doi: 10.1016/j.isprsjprs.2025.01.013.

[22] J. D. Toscano, C. Zuniga-Navarrete, W. D. Siu, L. J. Segura, and H. Sun, “Teeth mold point cloud completion via data augmentation and hybrid RL-GAN,” J. Comput. Inf. Sci. Eng., vol. 23, no. 4, 2023, doi: 10.1115/1.4056566.

[23] T. Kim, Y. Cho, D. Kim, M. Chang, and Y.-J. Kim, “Tooth segmentation of 3D scan data using generative adversarial networks,” Appl. Sci., vol. 10, no. 2, p. 490, 2020, doi: 10.3390/app10020490.

[24] S. Minhas, T.-H. Wu, D.-G. Kim, S. Chen, Y.-C. Wu, and C.-C. Ko, “Artificial intelligence for 3D reconstruction from 2D panoramic X-rays to assess maxillary impacted canines,” Diagnostics, vol. 14, no. 2, p. 196, 2024, doi: 10.3390/diagnostics14020196.

[25] T. Galba, Č. Livada, and A. Baumgartner, “Interactive holographic reconstruction of dental structures: A review and preliminary design of the HoloDent3D concept,” Appl. Sci., vol. 16, no. 1, p. 433, 2025, doi: 10.3390/app16010433.

[26] B. Zou, S. Wang, H. Liu, G. Sun, Y. Wang, F. Zuo, C. Quan, and Y. Zhao, “IO150K: Large-Scale Intraoral Image Dataset for TeethSEG,” Version 1, 2024. Available at: https://openreview.net/forum?id=P6tNXNycrT.

[27] Z. Li et al., “A fast 3D lung image reconstruction method based on CT pixel matrices learning with electrical impedance to mography,” Measurement, vol. 251, p. 117176, Jun. 2025, doi: 10.1016/j.measurement.2025.117176.

[28] L. Tan, S. Song, Y. He, K. Zhou, T. Lu, and R. Xiao, “EAR: Edge-aware reconstruction of 3-D vertebrae structures from bi-planar X-ray images,” arXiv, Jul. 30, 2024. doi: 10.48550/arxiv.2407.20937.