공학박사 학위논문

Structure-Preserving Quality Improvement of Cone-Beam CT Images Using Contrastive Learning

대조 학습을 이용한 Cone-Beam CT 영상의 구조 보존 품질 향상에 관한 연구

2023 년 8 월

서울대학교 대학원

융합과학부 방사선융합의생명전공

강 세 룡

대조 학습을 이용한 Cone-Beam CT 영상의 구조 보존 품질 향상에 관한 연구

지도 교수 이 원 진

이 논문을 공학박사 학위논문으로 제출함

2023 년 5 월

서울대학교 대학원

융합과학부 방사선융합의생명전공

강 세 룡

강세룡의 공학박사 학위논문을 인준함

2023 년 6 월

위 원 장 허 민 석 (인)

부위원장 이 원 진 (인)

위 원 허 경 회 (인)

위 원 이 재 성 (인)

위 원 김 준 민 (인)

ABSTRACT

Structure-Preserving Quality Improvement of Cone-Beam CT Images Using Contrastive Learning

Se-Ryong Kang

Department of Biomedical Radiation Sciences

Graduate School of Convergence Science and Technology

Seoul National University

Cone-beam CT (CBCT) is widely used in dental clinics but exhibits limitations in assessing soft tissue pathology because of its lack of contrast resolution and low Hounsfield Units (HU)

quantification accuracy. Various techniques have been investigated to enhance the quality of CBCT images. These include analytical modeling methods, advanced iterative reconstruction methods, Monte Carlo simulation, rule-based methods with prior knowledge, and random forest. However, these approaches are time-consuming due to computational complexity and have limited effectiveness in reducing complex artifacts.

Recently, a new solution called cycle-consistent generative adversarial networks (CycleGAN) has emerged for generating CT-like images from CBCT images. The objective of this approach was to improve the quality of CBCT images so that they resembled CT images. One of the main challenges when employing this method is preserving the structure of the original CBCT image. If the CT-like generated images fail to retain the anatomical structure of the input CBCT images, they become useless for clinical applications. Therefore, it is important to consider both the similarity of the generated images to CT scans and the preservation of the anatomical structures. However, it has been discovered that the preservation of the input anatomical structures on CBCT images is limited by CycleGAN.

To address this issue, the adoption of a contrastive learning-based GAN was considered as a baseline to enhance the correspondence between inputs and outputs. The objective was to improve the image quality and HU accuracy of CBCTs to a level comparable to CTs, while simultaneously preserving anatomical

structures. The structure-preserving contrastive-learning based GAN (SPCGAN) was trained on unpaired CT and CBCT datasets with the novel combination of losses and the feature extractor pretrained on training dataset. The losses employed during training included semantic relationship consistency loss, spatial correlation loss, and reconstruction loss. The loss functions of semantic relation consistency and spatially correlative were designed to maximize the learning of information regarding semantic and spatial patterns within an image, effectively generating structure-preserved images. Also, finer information about the input CBCT images should be contained in the representations by minimizing the reconstruction loss.

The quality of the generated images was evaluated using metrics such as Frechet inception distance (FID), peak signal-to-noise ratio (PSNR), mean absolute error (MAE), and root mean square error (RMSE) over the entire image area. Additionally, the structure preservation performance was assessed by the structure score. As a result, the CT-like images generated by SPCGAN were significantly superior to those generated by various baseline models in terms of FID, PSNR, MAE, RMSE, and structure score. Therefore, it was demonstrated that the use of SPCGAN provided complementary benefits, namely preserving the anatomical structures of the input CBCT images while simultaneously enhancing the image quality to a level comparable to CT images.

Keywords: Contrastive learning, Structure-preserving, Cone-beam CT images, Deep learning, SPCGAN
Student Number: 2013-22422

CONTENTS

LIST OF FIGURES

LIST OF TABLES

INTRODUCTION

Cone-beam CT (CBCT) is widely used in dental clinics due to its lower radiation exposure compared to conventional CTs while providing a higher resolution [1, 2, 3]. Because of its high spatial resolution, CBCT is mainly used for dental implant planning, visualizing abnormal teeth, and evaluating the jaws and face [4, 5]. However, CBCT has the limitation of low contrast resolution due to various physical and technical factors [6, 7, 8]. Highly scattered radiation by cone beam geometry negatively affects the contrast in the final reconstructed images [3]. Due to these limitations, it negatively impacts the accuracy of HU quantification in CBCT images [3, 6], making them inadequate for soft tissue evaluation. Therefore, to increase the utilization of CBCT in clinical applications, it is necessary to improve the accuracy of HU quantification in CBCT images, which should be accompanied by improvements in image quality.

Various methods have been explored to improve the quality of CBCT images, such as analytical modeling methods [9, 10, 11], advanced iterative reconstruction methods [12, 13, 14, 15], Monte Carlo simulation [16, 17], rule-based methods with prior knowledge [18, 19, 20], and random forest [21, 22]. However, these methods are time-consuming due to computational

complexity and have insufficient ability to reduce complex artifacts [23]. Recently, cycle-consistent generative adversarial networks (CycleGAN) [24] emerged as a new solution for generating CT-like images from CBCT images. This approach aimed to increase the quality of CBCT images to be similar to CT images. One of the most challenging problems when adopting this method is preserving the structure of the input image. It is useless in clinical applications if the generated CT-like images lose the anatomical structure of input CBCT images. Therefore, considering both "how CT-like the generated images are" and "how well the structures are preserved" becomes imperative. However, a limitation was discovered in the preservation of the anatomical structures on the input CBCT images when utilizing CycleGAN.

To address this issue, the adoption of a contrastive learning-based GAN was employed as a baseline to enhance the correspondence between inputs and outputs. Furthermore, semantic relation consistency loss [25], spatially correlative loss [26], and reconstruction loss were incorporated to enforce the network in generating images that are both CT-like and preserve the underlying anatomical structures [27]. The loss functions of semantic relation consistency and spatially correlative were designed to maximize the learning of information regarding semantic and spatial patterns within an image, effectively

generating structure-preserved images. Also, finer information about the input CBCT images should be contained in the representations by minimizing the reconstruction loss. A hypothesis was formulated that the inclusion of these finer representations could potentially result in the generation of CT-like images while preserving the underlying anatomical structures. Therefore, this study aimed to increase the CBCT image quality while preserving the anatomical structures by using a proposed structure-preserving contrastive learning-based GAN (SPCGAN). In summary, the main contributions of the proposed work are as follows:

- Contrastive learning-based GAN was applied to unpaired image translation with the aim of enhancing the quality of CBCT images. An extensive comparison was conducted with other models to assess its performance.
- A novel combination of loss functions was devised, incorporating semantic relation consistency loss, spatially correlative loss, and reconstruction loss, in order to enhance the quality of CBCT images while simultaneously preserving anatomical structures.

LITERATURE REVIEW

Several studies have employed the U-Net CNN architecture to enhance CBCT image quality by generating synthetic CT images from CBCT [28, 29, 30]. These methods utilize deep neural networks to estimate a mapping function that transforms the input CBCT into output synthetic CT images, with the goal of minimizing dissimilarity from the corresponding ground-truth CT images, as formulated in a loss function. However, limitations arise from the fact that the overall loss function can inadvertently smooth or eliminate soft tissue contrast and anatomical features.

GAN [31] is a type of deep learning architecture composed of two components, a generator and a discriminator, trained adversarially to produce synthetic data from input data. GAN has been widely used in image-to-image translation, where the goal is to translate images from one domain to another. The difficulty in obtaining aligned paired images from different modalities in a clinical setting can limit the practical application of this approach in medical imaging. The idea behind CycleGAN [24] is to learn a mapping function from one domain to another without paired images.

Self-attention CycleGAN [32] method was developed for CBCT-to-CT conversion, resulting in synthetic CT images with accurate dose calculations and improved organ boundaries compared to CBCT. Cycle-Deblur GAN [33] improved CBCT image quality in radiotherapy, with reduced artifacts and enhanced structural details. The combination of CycleGAN and two-channel U-Net in QCBCT-NET [34] improved the contrast and uniformity of the bone image, increasing the accuracy of bone density measurement. Respatch-CycleGAN [35] was proposed to convert CBCT to synthetic CT images, reducing metal artifacts and restoring HU values to match planning CT. All previous efforts to enhance the quality of CBCT images by converting them to CT images were based on CycleGAN. However, the cycle-consistency loss assumed that the relationship between the input and output domain was a bijection, which was often too restrictive, resulting in distortion of the translated images [36].

Recently, contrastive learning-based GAN methods were introduced to improve the correspondence between inputs and outputs in one-side image translation [25, 26, 36, 37]. This approach was based on maximizing the mutual information between the same location of the input and output images through unsupervised image-to-image translation. CUT [36] was the first contrastive learning-based GAN model outperforming CycleGAN, and NEGCUT [37] included hard negative samples generated by a

negative sample generator to enhance the performance of CUT. On the other hand, several methods for preserving the structure of input images have been proposed. LSeSim [26] introduced spatially correlative loss (SC) utilizing the patchwise similarity map to preserve the structure of input images. Semantic relation consistency loss (SRC) [25] was proposed to enhance the correspondence between input and output images by regularizing diverse semantic relations.

In order to address the limitations of CycleGAN and leverage the strengths of prior research, a contrastive learning-based GAN was utilized as the foundation. To further enhance its performance, the integration of semantic relation consistency loss [25] and spatially correlative loss [26] was implemented. Additionally, to simultaneously retain quality and maintain the structure, reconstruction loss was added to provide finer representations [27].

MATERIALS AND METHODS

Data acquisition and preparation

A total of 30 patients (18 males and 12 females) ranging in age from 21 to 80 years were recruited for the study. These patients required maxillary and mandibular CT or CBCT scans for treatment at the Seoul National University Dental Hospital. All CT and CBCT images were taken from July 2021 to June 2022 and were anonymized to protect personal information. This study was performed with approval from the institutional review board of the Seoul National University Dental Hospital (no. CRI21010). The workflow of the entire investigation was completed in accordance with all applicable rules and guidelines (Declaration of Helsinki).

Each patient underwent a CT scan followed by a consecutive CBCT scan using a CT scanner (Somatom Definition Edge, Siemens AG, Erlangen, Germany) and a CBCT scanner (CS 9300, Carestream Health, Inc., Rochester, US). The CT images were taken at 120 kVp and 120 mA with voxel sizes of 0.37 × 0.37 × 0.6 mm^3, dimensions of 512 × 512 pixels, and 12-bit depth. The CBCT images were acquired at 80 kVp and 8 mA with voxel sizes of 0.25 × 0.25 × 0.25 mm^3, dimensions of 669 × 669 pixels, and 16-bit depth. The CT and CBCT scans were re-scaled to 256 × 256 × 250 pixels, yielding 250 slices per patient. To avoid the influence

of non-anatomical structures in the training process, a binary mask was applied to the CT and CBCT images to separate the maxillomandibular region from the non-anatomical regions [33], [34]. Voxel values outside the masked region were replaced with the minimum HU value.

For the unpaired and unaligned training datasets, 6000 CT slices and 6000 CBCT slices from 24 patients were prepared. To conduct the test and evaluation, 1500 CT and 1500 CBCT slices acquired from the remaining six patients were matched by paired-point registration using software (3D Slicer, MIT, Cambridge, MA, USA). The manually selected eight landmarks for registration were the vertex on the lateral incisors, the buccal cusps of the first premolars, the distobuccal cusps of the first molars, the anterior nasal spine, and the pogonion [34]. Although paired-point registration was performed, it was difficult to perfectly match the anatomical structures in CTs to CBCTs. Therefore, for more accurate registrations, a non-rigid registration method was additionally applied using MATLAB software (Ver. R2018a, MathWorks, Natick, MA, USA).

The minimum sample size required for statistically significant evaluations of various deep-learning methods was estimated. A sample size of 290 was obtained, considering a significance level of 0.05 for determining the confidence level of the assessment, a statistical power of 0.95 for determining the

tolerance, and an effect size of 0.5 for determining the strength of the relationship between the two groups (G* Power for Windows 10, version 3.1.9.7, University of Dusseldorf, Dusseldorf, Germany). In this study, the number of 2D CBCT slices used for testing was set to 1500, which significantly exceeded the minimum requirement.

Contrastive learning-based CBCT-to-CT generation

The objective of this study was to perform image translation from the input CBCT slice domain to the output CT slice domain while preserving the structural characteristics of the input images. To achieve this, the following approach was employed: (1) Initially, a contrastive learning-based method, such as CUT adopted by Park [36], was adopted as a starting point to establish the baseline performance. (2) Subsequently, three additional losses were incorporated to assist in generating images that closely resembled CT scans while preserving the anatomical structural details of the input CBCT images. These losses included semantic relation consistency loss [25], spatially correlative loss [26], and reconstruction loss. The details are presented in the following subsections.

The patchwise contrastive learning-based GAN [25, 26, 36, 37] was employed in this study due to its effectiveness in preserving the structure of the input images. This approach maximized the mutual information between the input patches and their corresponding output patches, resulting in improved preservation of structural details. Formally, $X = \{\mathbf{x} \in \mathcal{X}\}$ and $Y = \{\mathbf{y} \in \mathcal{Y}\}$ are CBCT dataset and CT dataset, respectively, both of which are unpaired.

Following the notation used in a previous study [36], the generator G can be decomposed into two components: an

encoder G_{enc} and decoder G_{dec}. These components work together to generate the output image $\hat{\mathbf{y}} = G_{dec}(G_{enc}(\mathbf{x})) \in \mathbb{R}^{H \times W \times 1}$, with $\mathbf{x}$ representing the input image. Using the input image $\mathbf{x}$ and its corresponding output image $\hat{\mathbf{y}}$, embedded vectors can be extracted as $\mathbf{z}_k = H(G_{enc}(\mathbf{x})_k) \in \mathbb{R}^{C'}$ and $\mathbf{w}_k = H(G_{enc}(\hat{\mathbf{y}})_k) \in \mathbb{R}^{C'}$. Here, $H: \mathbb{R}^C \to \mathbb{R}^{C'}$ is a Multi-Layer Perceptron (MLP) network, and the index k refers to the spatial location in the encoded feature map $G_{enc}(\mathbf{x}) \in \mathbb{R}^{S \times C}$. It should be noted that for notational convenience, the feature map $G_{enc}(\mathbf{x})$ is described as a flattened feature map, where S represents the number of spatial locations. These embedding vectors $\mathbf{z}_k$ and $\mathbf{w}_k$ correspond to patches in the raw image space, which is why this method is referred to as 'patchwise' contrastive learning. More specifically, the Patchwise Contrastive Learning loss used is defined as:

$$L_{PCL} = \sum_{k \in \mathcal{J}} \left[-\log \frac{\exp(\mathbf{z}_k^\top \mathbf{w}_k / \tau)}{\sum_{j | j \neq k} \exp\left(\mathbf{z}_j^\top \mathbf{w}_k / \tau\right)} \right]$$

where $\mathcal{J}$ is the location indices set of size K sampled from $\{1,2,\ldots,S\}$ without replacement and τ is the temperature parameter; the higher the τ, the softer the distribution will be. To encourage the output image to be similar to the real CT images, an adversarial loss [31] is used as follows:

$$L_{GAN} = \log D(\mathbf{y}) + \log(1 - D(G(\mathbf{x})))$$

where D is a discriminator trained by maximizing L_{GAN}.

To generate CT-like and structure-preserved images, I combined three losses with the contrastive loss: semantic relation consistency loss [25], spatially correlative loss [26], and reconstruction loss inspired by CAiD [27].

Semantic relation consistency loss. For a given location index $k \in \mathcal{J}$, I can compute the patchwise semantic relation in the input image as follows:

$$Z_k(i) = \frac{\exp(\mathbf{z}_k^\top \mathbf{z}_i)}{\sum_{j \in \mathcal{J}} \exp\left(\mathbf{z}_k^\top \mathbf{z}_j\right)}$$

where $i \in \mathcal{J}$ is another location index. Note that $Z_k(i)$ can be interpreted as the distribution to capture the semantic closeness between the i-th and k-th location patches of the input images [25]. In a similar way, the patchwise semantic relation in the output image is defined as:

$$W_k(i) = \frac{\exp(\mathbf{w}_k^\top \mathbf{w}_i)}{\sum_{j \in \mathcal{J}} \exp\left(\mathbf{w}_k^\top \mathbf{w}_j\right)}.$$

Note that $W_k(i)$ is defined with the same location indices set $\mathcal{J}$ used in $Z_k(i)$. Semantic relation consistency loss is defined to preserve the structure during image translation as:

$$L_{SRC} = \sum_{k \in \mathcal{J}} JSD(Z_k \parallel W_k)$$

where $JSD(\cdot \parallel \cdot): \mathbb{R}^K \times \mathbb{R}^K \to [0, \infty)$ denotes Jensen-Shannon Divergence.

Spatially correlative loss. Let F_{ext} denote a pretrained external feature extractor that uses images $\mathbf{x}$ and $\mathbf{y} \in \mathbb{R}^{H \times W \times 1}$ as input, and produces feature maps $F_{ext}(\mathbf{x})$ and $F_{ext}(\mathbf{y}) \in \mathbb{R}^{H' \times W' \times C'}$. For a given spatial location index $(i, j) \in [1, H'] \times [1, W']$, let $F_{ext}(\mathbf{x})_{(i,j)} \in \mathbb{R}^{1 \times C'}$ denote a feature of $F_{ext}(\mathbf{x})$ whose spatial coordinate is (i, j), while $F_{ext}(\mathbf{x})^*_{(i,j)} \in \mathbb{R}^{N^2 \times C'}$ denotes the 'neighboring features' including $F_{ext}(\mathbf{x})_{(i,j)}$ in the center, where N represents the height and width of the neighboring features. The self-similarity, also known as the spatially correlative map, can be computed as follows:

$$S_{\mathbf{x}}^{(i,j)} = \left(F_{ext}(\mathbf{x})_{(i,j)}\right)\left(F_{ext}(\mathbf{x})^*_{(i,j)}\right)^T$$

Note that the spatially correlative map $S_{\mathbf{x}}^{(i,j)} \in \mathbb{R}^{1 \times N^2}$ captures the correlation between the query feature $(F_{ext}(\mathbf{x})_{(i,j)})$ and the neighboring features $(F_{ext}(\mathbf{x})^*_{(i,j)})$ within the input image $\mathbf{x}$. Similarly, a spatially correlative map within the output image $\hat{\mathbf{y}}$ with the same query point (i, j) is computed as:

$$S_{\hat{\mathbf{y}}}^{(i,j)} = \left(F_{ext}(\hat{\mathbf{y}})_{(i,j)}\right)\left(F_{ext}(\hat{\mathbf{y}})^*_{(i,j)}\right)^T$$

To preserve the structure, spatially correlative loss [27] is defined as follows:

$$L_{SC} = \sum_{(i,j)\in\mathcal{J}'} 1 - \frac{S_{\mathbf{x}}^{(i,j)} \cdot S_{\hat{\mathbf{y}}}^{(i,j)}}{\| S_{\mathbf{x}}^{(i,j)} \| \| S_{\hat{\mathbf{y}}}^{(i,j)} \|}$$

where $\mathcal{J}'$ denotes the location indices set sampled from $[1, H'] \times [1, W']$ without replacement. On a final note, a previous study [26] adopted VGG16 [38] pretrained on ImageNet [39] as the feature extractor F_{ext}. However, an ImageNet pretrained feature extractor such as VGG16 might be ineffective because CBCT and CT images differ quite from the general domain images. Therefore, the feature extractor F_{ext} was pretrained using the training set by employing an autoencoder structure. In this case, F_{ext} refers to the encoder component of the pretrained autoencoder model. The ablation study section provides the performance comparison between VGG16 and the customized one as the feature extractor F_{ext}.

Reconstruction loss. Hosseinzadeh Taher et al. [27] integrated a reconstruction loss with contrastive loss to encode finer information into the representations. Inspired by this, the feature map$G_{enc}(\mathbf{x})$ was encouraged to have more information about the input images by adding a reconstruction loss. The reconstruction loss was defined as follows:

$$L_{recon} = \| \mathbf{x} - G_{rec}(G_{enc}(\mathbf{x})) \|_1$$

where G_{rec} denotes an auxiliary decoder for reconstruction with the same structure as G_{dec}.

Final objective. Overall, generator G, which consists of G_{enc} and G_{dec}, is trained by minimizing the following final loss:
$L_{final} = \lambda_{PCL} L_{PCL} + \lambda_{SRC} L_{SRC} + \lambda_{SC} L_{SC} + \lambda_{recon} L_{recon} + L_{GAN}$.

An overview of the proposed method is provided in Figure 1.

The investigation focused on evaluating the performance of SPCGAN in generating CT-like images while simultaneously preserving the anatomical structures present in the input CBCT images. Furthermore, an ablation study was conducted to determine the loss function that had the most significant impact on both the preservation of anatomical structures and the improvement of image quality. Existing models, including CycleGAN [24], CUT [36], NEGCUT [37], LSeSim [26], and SRC [25], were compared to assess the effectiveness of SPCGAN in achieving these objectives. The purpose of this comparison was to establish the superiority of SPCGAN in terms of generating CT-like images with preserved anatomical structures when compared to other established models.

All models shown in this section were trained using the unpaired CBCT and CT images. Each domain had 6,000 axial slice images from 24 patients. Then, evaluations were conducted with the paired and aligned test sets composed of 1,500 axial slices of CBCTs and CTs from 6 patients. HU values of the CT images ranged from -1,024 to 3,071, whereas CBCT images in the training set ranged from -1,000 to 17,983. Therefore, the CT and CBCT images were normalized to the range of [0, 1] using the following procedure:

$$I'_{CT} = (I_{CT} + 1024)/4095$$
$$I'_{CBCT} = (\text{clip}(I_{CBCT}, [-1000, 3071]) + 1000)/4071$$

where $\text{clip}(I, [min, max])$ clamps all elements in input I into the range $[min,\ max]$. All images were resized to 256×256, and a random horizontal flip was applied in the training models. The Adam optimizer [40] ($\beta_1 = 0.5,\ \beta_2 = 0.999$) with a learning rate of 0.0002 was used. Following the original papers, the values of λ_{PCL}, λ_{SRC}, λ_{SC}, and λ_{recon} were set to 0.1, 0.05, 10.0, and 1.0, respectively. These values were chosen to balance the contribution of each loss term and align with the objectives of the study.

Quality of generated CT-like image

First, the Fréchet inception distance (FID) [41] was computed as a metric to quantify the quality of images generated by the generative model. FID is commonly employed to evaluate the similarity between the generated images and real images based on their feature representations. FID score compared two groups' feature vectors extracted from the Inception V3 model pre-trained on the Imagenet dataset. Therefore, FID score did not require CBCT-paired ground truth CT images. More precisely, the FID score between two image groups, A and B, was calculated as follows:

$$\mathrm{FID} = \|\mathbf{m}_A - \mathbf{m}_B\|_2^2 + \mathrm{Tr}\left(\mathbf{C}_A + \mathbf{C}_B - 2(\mathbf{C}_A\mathbf{C}_B)^{1/2}\right)$$

where $\mathbf{m}_A, \mathbf{m}_B \in \mathbb{R}^{2048}$ were the feature-wise means of the feature vectors, and $\mathbf{C}_A, \mathbf{C}_B \in \mathbb{R}^{2048\times2048}$ were covariance matrices of the feature vectors.

Unlike the setting for obtaining FID scores, the quality of the generated CT-like images was also evaluated using a paired test setting. In this setting, each CBCT image in the test set was paired with its corresponding CT image. This evaluation approach allowed for a direct comparison between the generated images and their ground truth counterparts, enabling a more comprehensive assessment of the image quality.

In this setting, several metrics were computed to evaluate the quality of the generated CT-like images. These metrics included the Peak Signal-to-Noise Ratio (PSNR), Mean Absolute Error (MAE), and Root Mean Square Error (RMSE):

$$PSNR = 10 \cdot \log_{10}\left(\frac{MAX^2}{RMSE^2}\right)$$

$$MAE = \frac{1}{mn}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}|I(i,j) - K(i,j)|$$

$$RMSE = \sqrt{\frac{1}{mn}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}[I(i,j) - K(i,j)]^2}$$

where I and K were a generated CT-like image and the ground truth CT image, respectively, while MAX denoted the maximum possible value of the image (i.e., 4,095 in case when the values were represented using 12 bits). Additionally, the assessment of the uniformity of the generated CT-like images was performed by computing the Spatial Non-Uniformity (SNU):

$$SNU = \frac{\overline{\mathrm{HU}}_{\max} - \overline{\mathrm{HU}}_{\min}}{1000} \times 100(\%)$$

where $\overline{\mathrm{HU}}_{\max}$ and $\overline{\mathrm{HU}}_{\min}$ were the maximum and the minimum of the mean of the selected ROIs, respectively. Six rectangular ROIs of soft tissue for each patient were selected around the mandible and maxilla.

Preservation of anatomical structures

To evaluate how well a model preserved the structures of an input, the structure score in SSIM [42] was used:

$$s(x, y) = \frac{\sigma_{xy} + C}{\sigma_x \sigma_y + C}$$

where C, σ_x^2, σ_y^2, and σ_{xy} denote a constant for computational stability, the variance of an image x, the variance of an image y, and the covariance of x and y, respsctively. It is important to note that the structure score was calculated specifically between the input CBCT image ($\mathbf{x}$) and the generated CT-like image ($\hat{\mathbf{y}}$) obtained from different models. This comparison allowed for a quantitative assessment of how well each model preserved the structural information present in the original CBCT image while generating the CT-like image. By evaluating the structure score between the input and generated images, the effectiveness of each model in preserving the anatomical structures could be determined.

Ablation study

The final loss of SPCGAN was composed of reconstruction (Rec), spatially correlative (SC), and semantic relation consistency (SRC) losses. To better understand the influence of different components on the proposed loss, a thorough analysis was conducted by training models under various conditions using the same datasets. This involved systematically adding the components of SPCGAN in a progressive manner and observing the corresponding performance. By examining the quantitative performance of the models at each stage of component addition, the individual contributions and impacts of these components on the overall performance of SPCGAN could be comprehended and evaluated. In addition, SC had two variants in terms of the external feature extractor F_{ext}; one used the feature extractor pretrained on training set (customized SC or cmSC), while the other used the VGG16 pretrained on ImageNet (vggSC) [26].

Evaluation of image quality by CBCT z-axis slice position

The cone-shaped beam geometry of CBCT causes asymmetry in the X-ray path, leading to position-dependent in beam hardening and endo/exo-mase effects [43-45]. As a result, image quality degradation and artifacts may occur more at the top and bottom of the field of view (FOV), and the contrast and HU quantification accuracy of the image are affected depending on the z-axis position of the slice image [46]. To evaluate the performance of SPCGAN with respect to the z-axis position of CBCT images, a quantitative performance of various models was conducted. The CBCT slices were categorized into three groups (upper, middle, and lower) based on their z-axis position, as depicted in Figure 8 alongside the entire field of view (FOV).

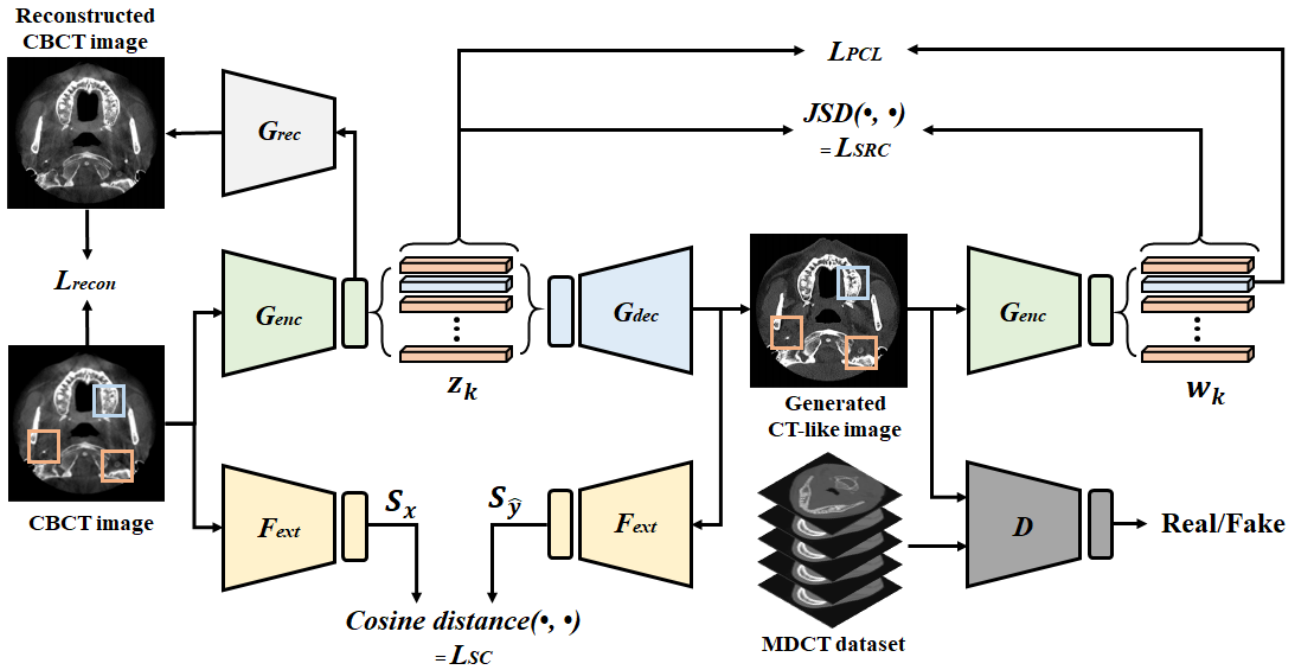

Figure 1. An overview of the proposed structure preserving contrastive-learning based GAN (SPCGAN). SPCGAN is constructed by adding G_{enc} for L_{PCL} and L_{SRC}, F_{ext} for L_{SC}, and G_{rec} for L_{recon} based on a GAN consisting of a generator and a discriminator. G_{enc} encoder of generator, G_{dec} decoder of generator, D discriminator, F_{ext} feature extractor, G_{rec} generator for image reconstruction, *GAN* generator adversarial network, L_{PCL} patchwise contrastive learning loss, L_{SRC} semantic relation consistency loss, L_{SC} spatially correlative loss, L_{recon} reconstruction loss.

RESULTS

Table 1 shows FID scores between CT images and CT-like images generated by various models. The SPCGAN performed the best, achieving an FID score of 42.126, which indicates that SPCGAN generated CT-like images most similarly with original CT images.

As shown in Table 2, SPCGAN outperformed all other models in terms of PSNR, MAE, and RMSE, achieving a PSNR of 26.735, an MAE of 63.044, and an RMSE of 156.020. However, SPCGAN showed SNU of 29.921, which was inferior to that of CUT achieving the best score of 28.546. All measures of SPCGAN exhibited significant differences from all other models ($p < 0.05$) except for the SNU from CycleGAN, NEGCUT, and SRC. Therefore, SPCGAN resulted in greater image quality improvement in all aspects except for uniformity compared to the other methods.

Figure 2 shows the HU accuracy by visualizing the difference between generated CT-like images and ground truth CT images. In the soft tissue region of the subtraction images, it was evident that the generated CT-like images exhibited significantly fewer discrepancies with the ground truth CT

images compared to the input CBCT images. In particular, the differences around the teeth, dense bone, and airway were more reduced in SPCGAN than in other networks. However, there were relatively large differences in the spinal and occipital bone regions at the maxilla. This results from the registration limitations due to differences in patient scan positions during the acquisition of the CBCT and CT datasets.

Figure 3 shows the horizontal and vertical HU-line profiles of the generated CT-like images and the ground truth CT images. The Pearson correlation coefficient between the ground truth CT images and the CT-like images generated by SPCGAN was 0.975 for the horizontal profile and 0.951 for the vertical profile, outperforming all other models' images. In other words, SPCGAN better reflected the boundaries and fine details of the images than others. Figure 4 shows the HU histogram plots between -300 HU and 300 HU in ground truth CT, input CBCT, and generated CT-like images. Compared with the histogram of the other methods' images, the histogram of SPCGAN's images was closer to that of ground truth CT images, demonstrating improvement in HU fidelity.

Figure 5 shows the Bland–Altman plots between the ground truth CT and SPCGAN, SRC, LSeSim, NEGCUT, CUT, CycleGAN, and input CBCT images. The Bland-Altman plot between the ground truth CT and SPCGAN's images demonstrated a lower bias of mean difference and a better level

of agreement than the plots between the ground truth CT and CT-like images generated by other models. In other words, the CT-like images generated by SPCGAN exhibited statistically significant similarity with the ground truth CT compared to other methods. Figure 6 shows the linear regression curves between the ground truth CT and generated CT-like images. The slope between the ground truth CT and SPCGAN ‘s images was closer to 1 than that between the ground truth CT and other methods’ images, demonstrating the superiority of SPCGAN in terms of HU accuracy.

SPCGAN had a structural score of 0.880 ± 0.043, which outperformed LSeSim (0.868 ± 0.047), SRC (0.866 ± 0.043), NEGUT (0.862 ± 0.045), CUT (0.865 ± 0.046), and CycleGAN (0.862 ± 0.042). The statistical analysis using t-tests confirmed a significant difference between the result of SPCGAN and those from the other deep-learning models.

Figure 7 shows an input CBCT image and CT-like images generated by various models. SPCGAN retained the finely detailed structures of the input CBCT image, whereas the other models tended to deform these structures. In particular, distortions were prominent in the airway area (indicated by the cyan box) and vertebrae area (indicated by the pink box) of the images generated by other deep-learning models. In addition, as shown in the soft tissue area (indicated by the green box), SPCGAN best preserved

the boundaries and distribution of fat and muscle. CycleGAN, a non-contrastive learning model, suffered the greatest structure collapse.

The results presented in Table 3 indicate how each loss influenced the evaluation metrics. When cmSC was added to SRC, there was no significant improvement in the quality evaluation factor, but the structure score improved. When Rec was further applied, the results showed the best performance not only in structure scores, but also in FID, PSNR, MAE, and RMSE. This demonstrated that SC was advantageous for preserving anatomical structures and that adding Rec improved the overall image quality. In addition, cmSC exhibited more effective improvement in anatomical preservation and image quality than vggSC.

The results in Table 4 indicate how the evaluation metrics vary in slice images based on the z-axis position. Notably, when evaluating the slice located in the middle of the z-axis, the best performances were obtained for FID, PSNR, MAE, RMSE, SNU, and structure score. However, the overall results did not match those of the middle group when examining the upper and lower slice images. Despite this, the difference was not statistically significant. The findings demonstrated that the deep learning models, including SPCGAN, effectively improved the quality of CBCT images as much as CT images across the entire FOV of CBCT.

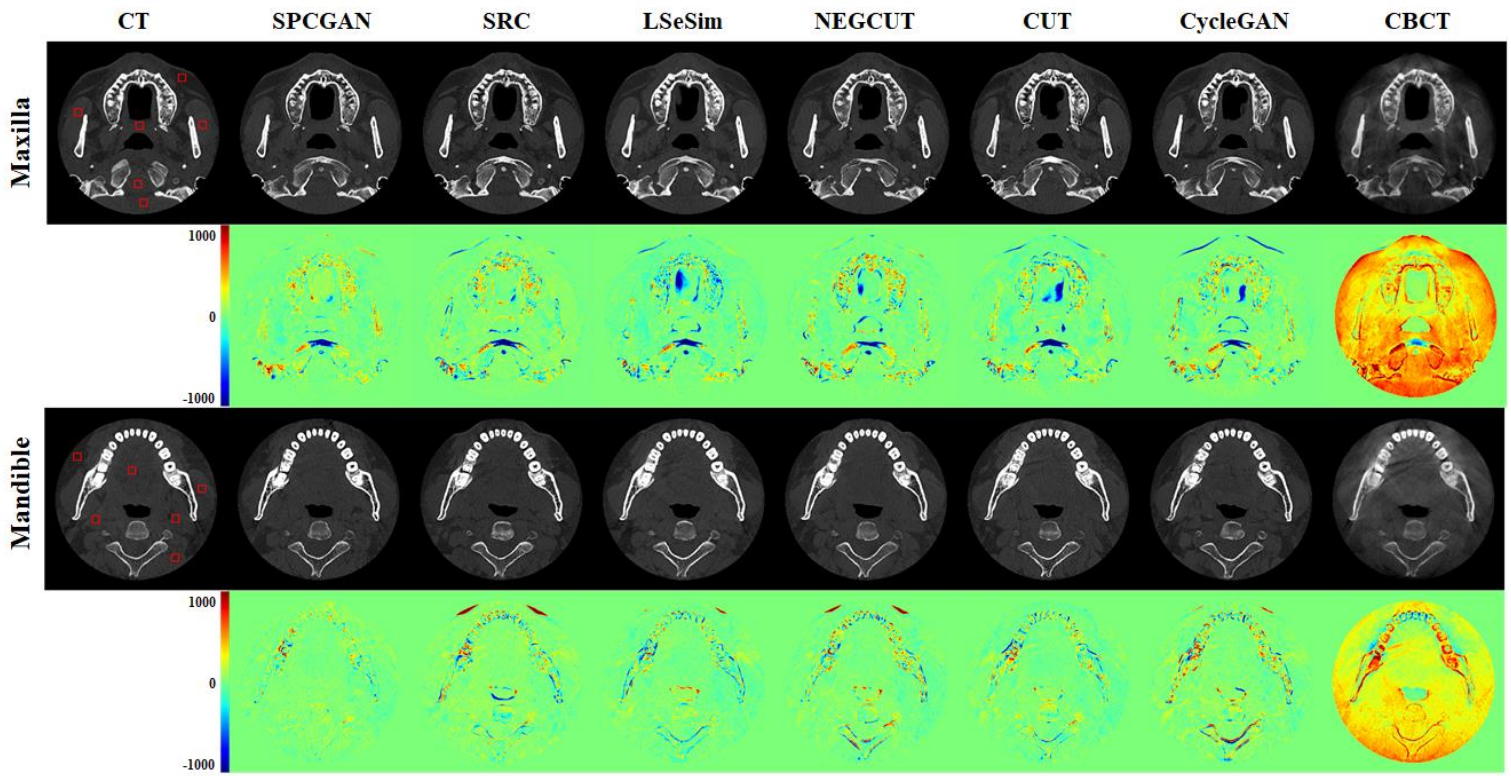

Figure. 2. The ground truth CT images, the CT-like images generated by deep learning methods and input CBCT images (the first and third rows) and their subtractions from the ground truth CT images (the second and fourth rows) at the maxilla and the mandible. The red squares shown in the CT images are the ROIs for calculation of spatial non-uniformity (SNU).

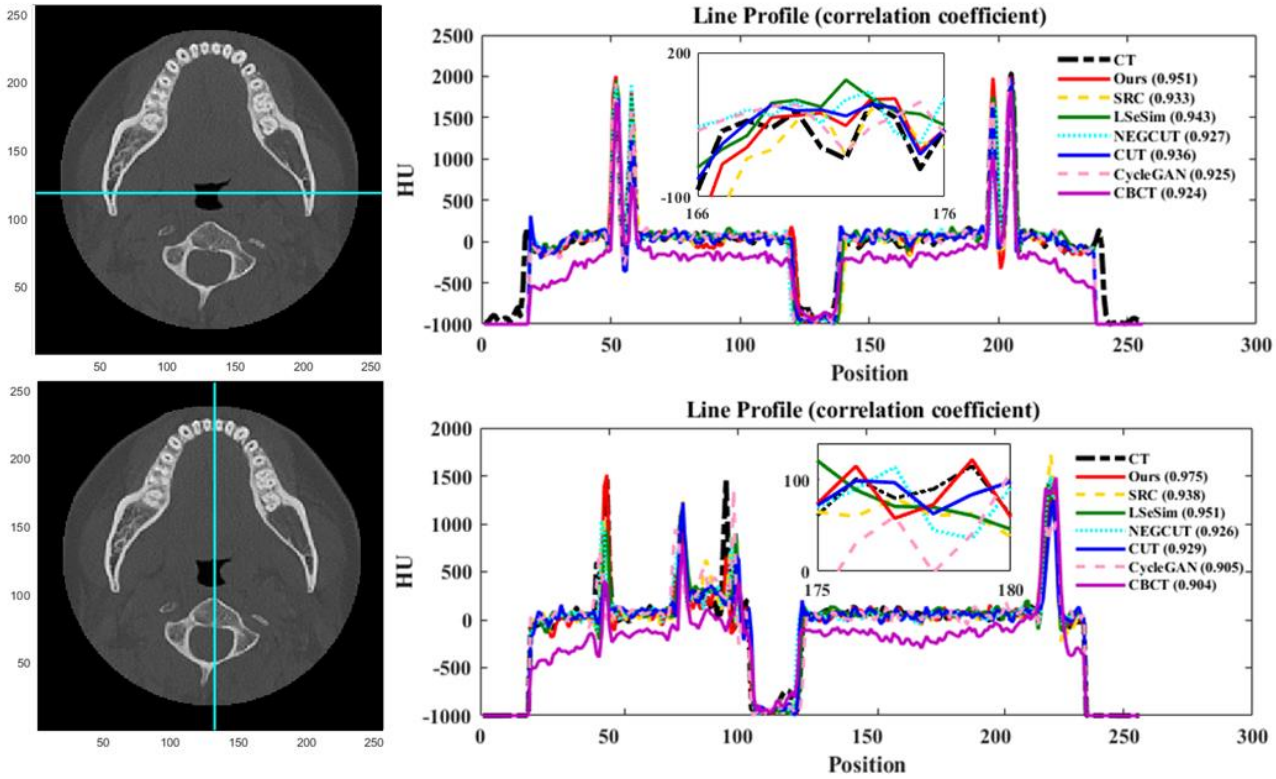

Figure. 3. The line profiles of HU values for CT, SPCGAN, LSeSim, NEGCUT, CUT, CycleGAN, and input CBCT image in the horizontal (top) and vertical (bottom) directions. The images shown on the left are CT-like images generated by SPCGAN. Pearson correlation coefficients of SPCGAN, SRC, LSeSim, NEGCUT, CUT, CycleGAN, and input CBCT images with the ground truth CT images are 0.975, 0.938, 0.951, 0.926, 0.929, 0.905, and 0.904, respectively, for the horizontal profile and 0.951, 0.933, 0.943, 0.927, 0.936, 0.925, and 0.924, respectively, for the vertical profile.

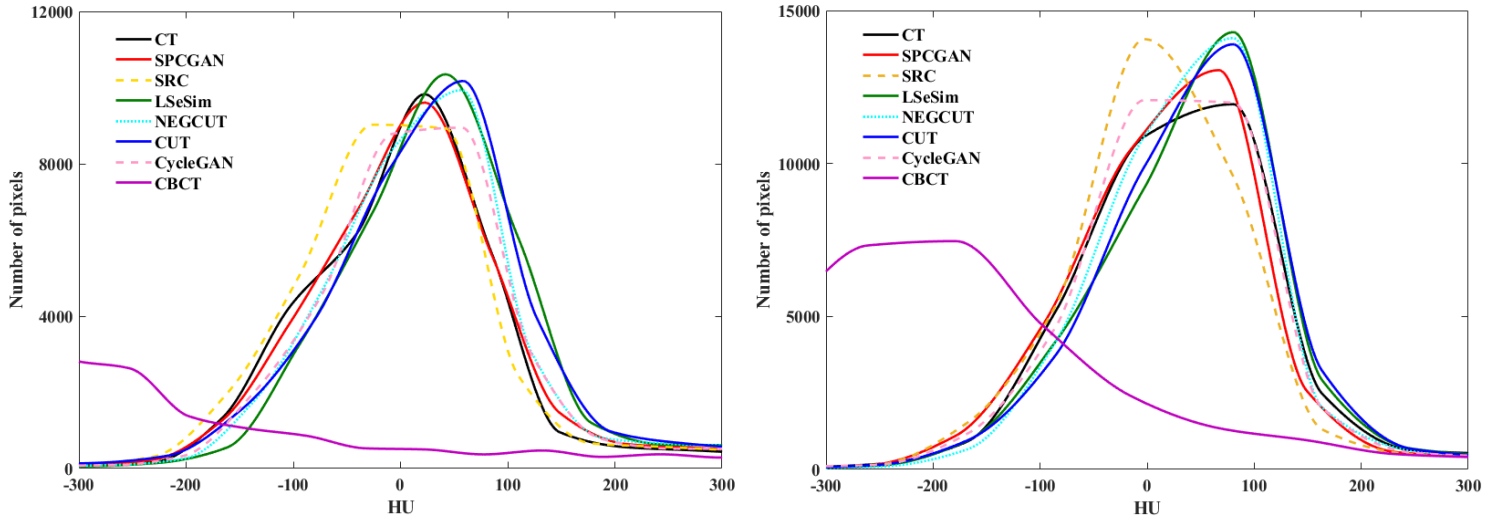

Figure. 4. The histogram shows the range of HU values ranging from -300 to 300 within the circular regions of interest (ROIs) in the maxilla (left) and mandible (right) areas, excluding the non-anatomical regions, in the ground truth CT, SPCGAN, SRC, LSeSim, NEGCUT, CUT, CycleGAN, and input CBCT images shown in Figure 2.

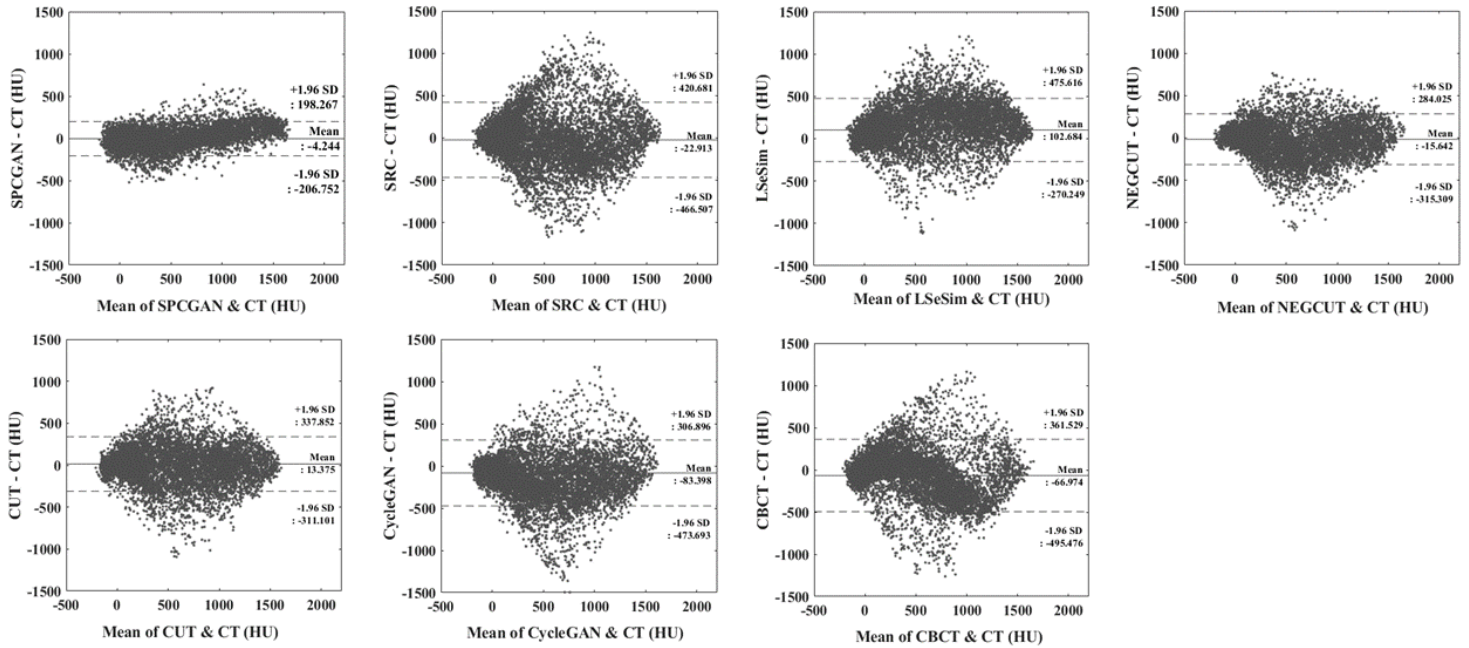

Figure 5. The Bland–Altman plots of HU between the ground truth CT and SPCGAN, SRC, LSeSim, NEGCUT, CUT, CycleGAN, and input CBCT images.

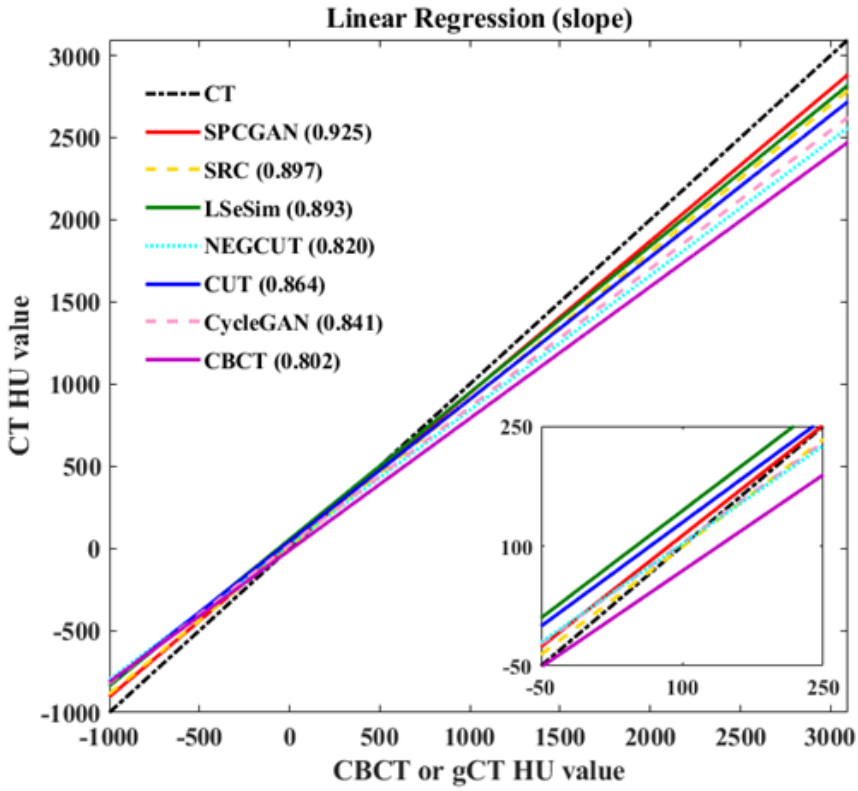

Figure 6. The linear regressions between the ground truth CT and the generated CT-like images. The slope of SPCGAN is 0.925 and the intercept is 20.330. The slope of SRC is 0.897 and the intercept is 9.164. The slope of LSeSim is 0.893 and the intercept is 54.622. The slope of NEGCUT is 0.820 and the intercept is 20.536. The slope of CUT is 0.864 and the intercept is 42.857. The slope of CycleGAN is 0.841 and the intercept is 19.421. The slope of input CBCT is 0.802 and the intercept is 69.069.

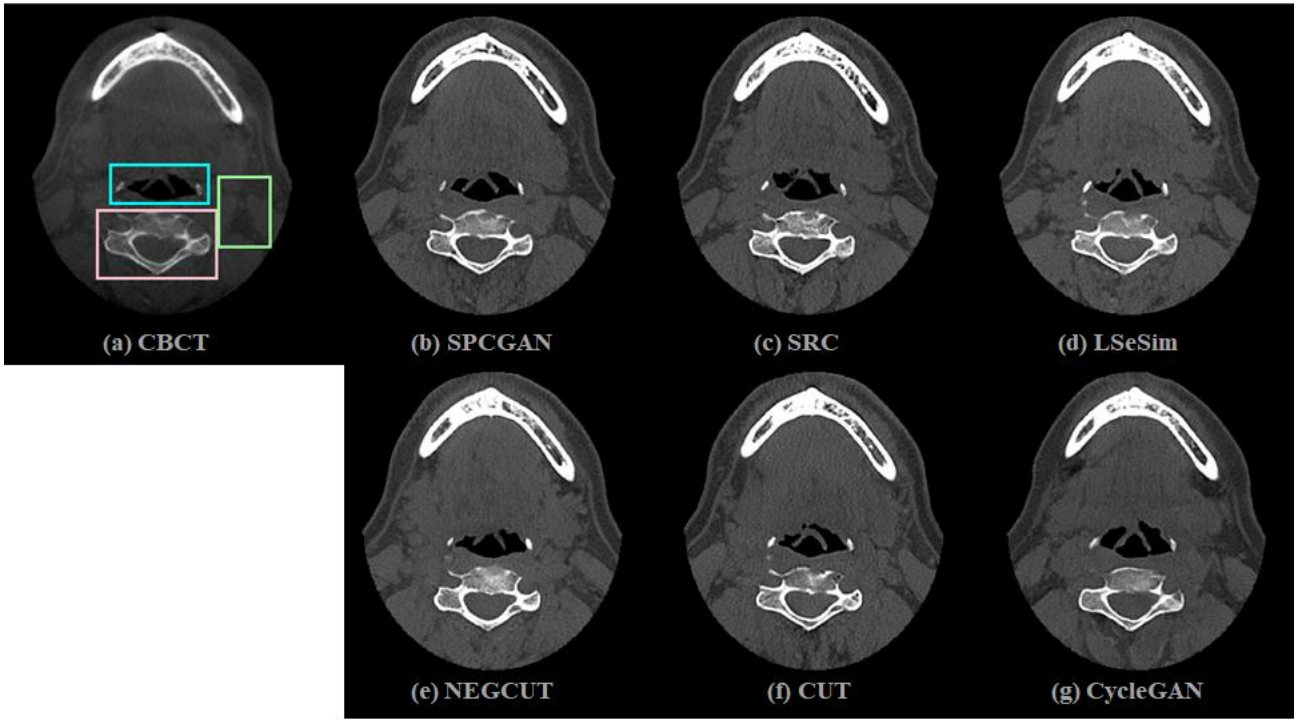

Figure 7. Visual comparison between the input CBCT image from the test datasets and CT-like images generated by deep learning methods. The cyan, pink, and green squares shown in the input CBCT image represent ROIs with significant differences compared to the generated images. The display window is set equal to [−300, 900] HU.

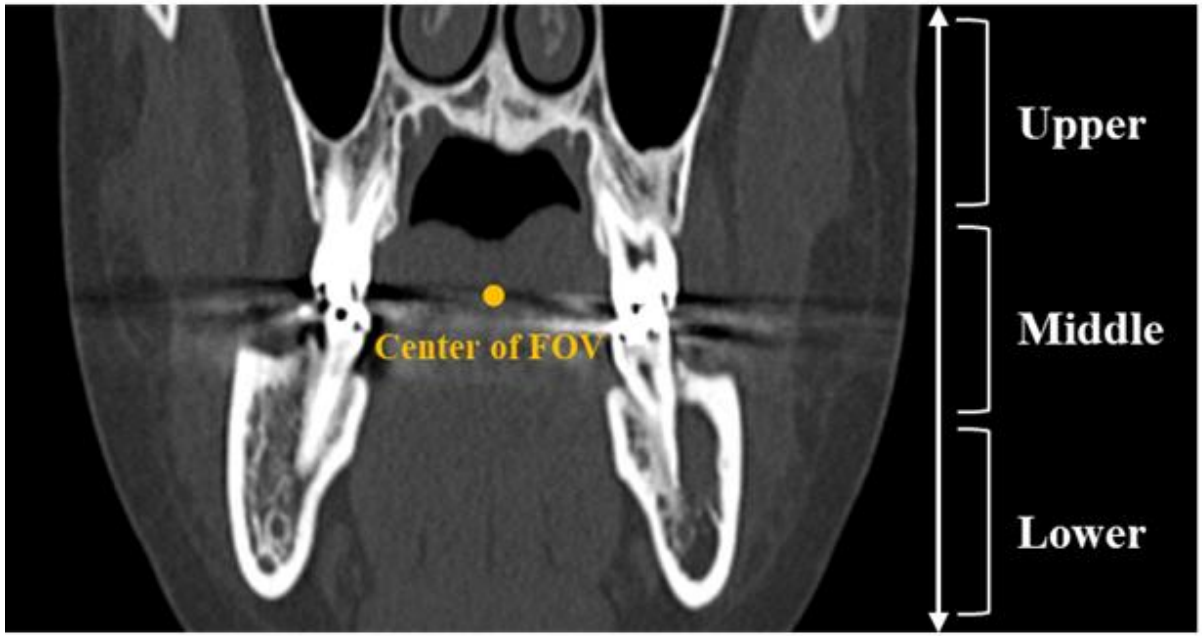

Figure. 8. The slices according to the z-axis position of the CBCT were divided into three groups (upper, middle, and lower) along with the entire field of view (FOV).

Table 1. Fréchet Inception Distance (FID) score between CT and generated CT-like images in the test set.

	CycleGAN	CUT	NEGCUT	LSeSim	SRC	SPCGAN
FID ↓	50.418	47.201	43.929	43.410	42.913	**42.126**

Note: The lower the score, the more similar the two groups of images are.

Table 2. Quantitative analysis results for assessing the quality of CBCT and CT-like images generated by CycleGAN, CUT, NEGCUT, LSeSim, SRC, and SPCGAN compared to the ground truth CT images. *PSNR* peak signal to noise ratio, *MAE* mean absolute error, *RMSE* root mean square error, *SNU* spatial non-uniformity.

	PSNR ↑	MAE(HU) ↓	RMSE ↓	SNU(%) ↓
CBCT	22.169 ± 1.430*	156.915 ± 3.735*	246.325 ± 44.973*	38.477 ± 7.698*
CycleGAN	25.389 ± 1.813*	68.551 ± 3.656*	171.636 ± 41.240*	30.282 ± 8.736
CUT	25.062 ± 1.900*	71.938 ± 4.367*	178.638 ± 44.974*	**28.546** ± 5.733*
NEGCUT	25.861 ± 1.972*	66.486 ± 3.502*	163.342 ± 43.711*	30.913 ± 6.958
LSeSim	25.653 ± 1.963*	64.511 ± 3.017*	167.190 ± 43.558*	31.447 ± 5.470*
SRC	25.863 ± 2.073*	65.763 ± 3.312*	163.782 ± 45.821*	30.319 ± 6.154
SPCGAN	**26.735** ± 1.794	**63.044** ± 3.086	**156.020** ± 40.992	29.921 ± 4.958

Note: PSNR, MAE, and RMSE compare generated CT-likes and ground truth CTs, whereas SNU is computed using only generated CTs. * indicates a statistically significant difference ($p < 0.05$) when conducting an independent t-test between SPCGAN and each other model.

Table 3. Quantitative results of ablation studies over the adding components of the loss function of SPCGAN. *FID* Fréchet inception distance, *PSNR* peak signal to noise ratio, *MAE* mean absolute error, *RMSE* root mean square error, *SNU* spatial non-uniformity.

Settings				Structure score ↑	FID ↓	PSNR ↑	MAE (HU) ↓	RMSE ↓	SNU (%) ↓
SRC	cmSC	vggSC	Rec						
✓				0.866 ± 0.043* †	42.913	25.863 ± 2.073*	65.763 ± 3.312*	163.782 ± 45.821*	30.319 ± 6.154 †
✓		✓		0.869 ± 0.038* †	43.605	25.927 ± 1.854*	65.164 ± 3.711*	163.038 ± 42.931*	31.540 ± 6.063*
✓		✓	✓	0.870 ± 0.046* †	43.583	26.176 ± 1.923* †	63.485 ± 3.406* †	162.608 ± 42.596*	30.217 ± 5.027 †
✓	✓			0.878 ± 0.044	42.959	25.892 ± 2.104*	65.373 ± 3.204*	165.405 ± 44.639*	31.934 ± 5.832*
✓	✓		✓	**0.880** ± 0.043	**42.126**	**26.735** ± 1.794	**63.044** ± 3.086	**156.020** ± 40.992	**29.921** ± 4.958

Note: $*$ indicates a statistically significant difference $(p < 0.05)$ when conducting an independent t-test between SPCGAN (SRC + scmSC + Rec) and each other model. † indicates a statistically significant difference $(p < 0.05)$ when conducting an independent t-test between cmSC +SRC and each other model.

Table 4. Quantitative results of image quality evaluation by the CBCT z-axis slice position (upper, middle, lower) of the CBCT and the generated CT-like images. *FID* Fréchet inception distance, *PSNR* peak signal to noise ratio, *MAE* mean absolute error, *RMSE* root mean square error, *SNU* spatial non-uniformity.

	CBCT			**CycleGAN**			**CUT**			**NEGCUT**			**LSeSim**			**SRC**			**SPCGAN**		
	Upper	**Middle**	**Lower**	**Upper**	**Middle**	**Lower**	**Upper**	**Middle**	**Lower**	**Upper**	**Middle**	**Lower**	**Upper**	**Middle**	**Lower**	**Upper**	**Middle**	**Lower**	**Upper**	**Middle**	**Lower**
FID	64.721	62.513	63.812	51.231	49.428	50.968	47.867	47.101	48.023	44.523	43.464	44.731	44.110	42.957	43.552	43.216	42.857	43.164	43.158	**42.012**	42.878
PSNR	22.334 ± 1.250	23.234 ± 1.703	23.149 ± 2.210	24.823 ± 2.679	25.761 ± 2.337	24.897 ± 2.963	25.090 ± 2.183	25.328 ± 2.067	25.274 ± 2.960	26.091 ± 3.336	26.416 ± 2.828	25.431 ± 3.600	25.080 ± 2.543	26.064 ± 2.471	25.885 ± 2.566	25.099 ± 3.475	25.884 ± 2.774	25.796 ± 3.336	27.006 ± 2.959	**27.342** ± 2.705	27.072 ± 2.876
MAE (HU)	158.312 ± 3.142	156.221 ± 3.291	156.615 ± 3.946	68.412 ± 4.184	67.990 ± 3.730	68.928 ± 4.243	71.272 ± 5.443	71.108 ± 5.310	71.989 ± 5.335	66.315 ± 4.396	65.624 ± 4.000	65.960 ± 4.340	64.797 ± 4.100	64.043 ± 3.907	64.774 ± 4.787	65.766 ± 4.521	64.881 ± 4.261	65.397 ± 4.960	63.035 ± 4.965	**63.023** ± 4.051	63.794 ± 4.748
RMSE	244.699 ± 47.121	243.300 ± 45.036	245.598 ± 45.183	170.086 ± 42.477	169.486 ± 41.744	170.164 ± 42.584	179.976 ± 46.817	178.520 ± 45.951	179.043 ± 46.573	164.017 ± 44.844	163.128 ± 44.631	163.620 ± 44.905	165.101 ± 44.608	164.937 ± 44.387	165.411 ± 44.619	163.037 ± 47.265	162.374 ± 46.372	162.705 ± 47.050	158.508 ± 42.369	**155.610** ± 41.726	158.436 ± 42.574
SNU (%)	39.187 ± 6.582	37.508 ± 7.698	39.513 ± 6.913	29.918 ± 9.683	29.852 ± 8.739	30.708 ± 8.912	29.030 ± 7.108	**28.186** ± 6.418	28.769 ± 7.322	31.445 ± 7.855	30.878 ± 7.518	31.418 ± 7.792	31.202 ± 5.709	31.061 ± 5.500	31.947 ± 5.521	30.757 ± 7.926	30.030 ± 7.111	30.700 ± 7.584	30.530 ± 6.227	29.105 ± 5.879	29.453 ± 6.191
Structure score				0.859 ± 0.043	0.866 ± 0.043	0.861 ± 0.046	0.854 ± 0.045	0.864 ± 0.046	0.862 ± 0.047	0.860 ± 0.046	0.865 ± 0.044	0.859 ± 0.047	0.864 ± 0.045	0.873 ± 0.047	0.867 ± 0.044	0.867 ± 0.045	0.870 ± 0.043	0.863 ± 0.044	0.884 ± 0.046	**0.890** ± 0.044	0.886 ± 0.047

DISCUSSION

CT-like images were generated from CBCT images using the SPCGAN with unpaired datasets. SPCGAN was trained by a novel combination of semantic relation consistency, spatially correlative, and reconstruction losses. The loss functions of semantic relation consistency and spatially correlative were designed to maximize the learning of information regarding semantic and spatial patterns within an image, effectively generating structure-preserved images. Also, finer information about the input CBCT images should be contained in the representations by minimizing the reconstruction loss. As a result, the CT-like images generated by SPCGAN were significantly superior to those generated by various baseline models in terms of the quality improvement of CBCT images (Table 1 and Table 2) and structure preservation (Table 3). This clearly showed that SPCGAN, which maximized the mutual information between inputs and corresponding outputs, was more effective in the CBCT-to-CT task than the other methods.

As shown in the ablation study (Table 4), adding reconstruction loss improved the overall quality of the images. It was conjectured that the inclusion of the reconstruction loss contributed to enriching the feature map $G_{enc}(\mathbf{x})$ with input

information, leading to improved quality in CBCT images. One trade-off for the quality improvement of CBCT images was training time. When adding the reconstruction loss to SRC, the number of learnable parameters increased from 14.7M to 19.8M, and the training time increased from 0.098 to 0.119 (s/image). Note that even if the reconstruction loss caused the training process to slow slightly, testing time remained the same, which was more important to end users such as practitioners.

Someone may interpret the reconstruction loss as the cycle consistency loss in CycleGAN. However, there is a difference in that the cycle consistency loss regulates both G_{enc} and G_{dec} by reversing $\hat{\mathbf{y}}$ to $\mathbf{x}$, and the reconstruction loss regulates only G_{enc} by reversing $G_{enc}(\mathbf{x})$ to $\mathbf{x}$. Moreover, the cycle consistency loss required another generator (F_{enc}, F_{dec}) and discriminator (D'), whereas the reconstruction loss required only another decoder (G_{rec}). Therefore, it may be reasonable to interpret the reconstruction loss as a 'light' cycle consistency loss.

It is also worth mentioning that using customized F_{ext} pretrained on training set was more effective for the structure preservation than was using VGG16 pretrained on ImageNet, even if the customized structure was a simple encoder composed of ResBlocks and trained by autoencoding with an auxiliary decoder. This demonstrated that the CT domain was quite different from the general domain (e.g., ImageNet), which implied room for improvement by fine-tuning the model.

There were some limitations in this study. The model developed in this work used a 2D network architecture to generate CT-like images slice by slice. Considering that CT data are 3D volumes, a 3D network that takes an entire 3D CBCT volume as input and generates an entire CT-like volume at a time can capture the relationship between the upper and lower slices. However, when implementing a 3D model, the required GPU memory and model size is much larger, requiring more training datasets than the current model. To address these limitations, the initial focus of this study was on the 2D model to demonstrate the performance of the contrast learning method. Further studies are required to enhance its performance by transitioning to a 3D model. Additionally, it's important to note that the CBCT datasets used in this study were obtained from a single device. Therefore, the generality of the method to other devices may be limited. Future investigations using CBCT datasets from multiple devices would provide a more comprehensive evaluation and enhance the applicability of the proposed method in different settings.

In conclusion, CT-like images were successfully generated from CBCT images using SPCGAN model. The model incorporated a novel combination of losses and a pretrained feature extractor, enhancing its performance. The generated CT-like images outperformed those produced by various baseline models, as indicated by significant improvements in FID, PSNR, MAE, RMSE, and structure score. The results of this study

demonstrated the complementary benefits of SPCGAN. It effectively preserved the anatomical structures present in the input CBCT images while simultaneously improving the image quality to closely resemble that of CT images. This achievement had important implications, as accurate quantification of Hounsfield Units (HU) could be achieved using CBCT, expanding its potential utility in clinical settings. The success of this proposed method highlighted its efficacy in enhancing the quality and fidelity of CBCT images, bringing them closer to the standards set by CT images. This breakthrough opened up new possibilities for utilizing CBCT in a wider range of clinical scenarios, providing valuable insights and facilitating improved patient care.

REFERENCES

1. W. C. Scarfe, A. G. Farman, P. Sukovic, Clinical applications of cone-beam computed tomography in dental practice, Journal-Canadian Dental Association 72 (2006) 75-80.

2. J. B. Ludlow, M. Ivanovic, Comparative dosimetry of dental CBCT devices and 64-slice CT for oral and maxillofacial radiology, Oral Surgery, Oral Medicine, Oral Pathology, Oral Radiology, and Endodontology 106 (2008) 106-114.

3. T. Kiljunen, T. Kaasalainen, A. Suomalainen, M. Kortesniemi, Dental cone beam CT: A review, Physica Medica 31 (2015) 844-860.

4. H. Watanabe, E. Honda, A. Tetsumura, T. Kurabayashi, A comparative study for spatial resolution and subjective image characteristics of a multi-slice CT and a cone-beam CT for dental use, European Journal of Radiology 77 (2011) 397-402.

5. J.-W. Choi, S.-S. Lee, S.-C. Choi, M.-S. Heo, K.-H. Huh, W.-J. Yi, S.-R. Kang, D.-H. Han, E.-K. Kim, Relationship between physical factors and subjective image quality of cone-beam computed tomography images according to diagnostic task, Oral Surgery, Oral Medicine, Oral Pathology and Oral Radiology 119 (2015) 357-365.

6. R. Pauwels, K. Araki, J. Siewerdsen, S. S. Thongvigitmanee, Technical aspects of dental CBCT: State of the art,

Dentomaxillofacial Radiology 44 (2015) 20140224.
7. W. C. Scarfe, Z. Li, W. Aboelmaaty, S. Scott, A. Farman, Maxillofacial cone beam computed tomography: Essence, elements and steps to interpretation, Australian Dental Journal 57 (2012) 46-60.
8. R. Pauwels, R. Jacobs, S. R. Singer, M. Mupparapu, CBCT-based bone quality assessment: are Hounsfield units applicable?, Dentomaxillofacial Radiology 44 (2015) 20140238.
9. S. Naimuddin, B. Hasegawa, C. A. Mistretta, Scatter-glare correction using a convolution algorithm with variable weighting, Medical Physics 14 (1987) 330-334.
10. E.-P. Ruhrnschopf, K. Klingenbeck, A general framework and review of scatter correction methods in x-ray cone-beam computerized tomography. Part 1: Scatter compensation approaches, Medical Physics 38 (2011) 4296-4311.
11. E.-P. Ruhrnschopf and, K. Klingenbeck, A general framework and review of scatter correction methods in cone beam CT. Part 2: Scatter estimation approaches, Medical Physics 38 (2011) 5186-5199.
12. E. Y. Sidky, C.-M. Kao, X. Pan, Accurate image reconstruction from few-views and limited-angle data in divergent-beam CT, Journal of X-ray Science and Technology 14 (2006) 119-139.
13. Z. Tian, X. Jia, K. Yuan, T. Pan, S. B. Jiang, Low-dose CT reconstruction via edge-preserving total variation regularization,

Physics in Medicine & Biology 56 (2011) 5949-5967.

14. H. Yu, G. Wang, A soft-threshold filtering approach for reconstruction from a limited number of projections, Physics in Medicine & Biology 55 (2010) 3905-3916.

15. X. Jia, B. Dong, Y. Lou, S. B. Jiang, GPU-based iterative cone-beam CT reconstruction using tight frame regularization, Physics in Medicine & Biology 56 (2011) 3787-3807.

16. E. Mainegra-Hing, I. Kawrakow, Variance reduction techniques for fast Monte Carlo CBCT scatter correction calculations, Physics in Medicine & Biology 55 (2010) 4495-2507.

17. C. Zollner, S. Rit, C. Kurz, G. Vilches-Freixas, F. Kamp, G. Dedes, C. Belka, K. Parodi, G. Landry, Decomposing a prior-CT-based cone-beam CT projection correction algorithm into scatter and beam hardening components, Physics and Imaging in Radiation Oncology 3 (2017) 49-52.

18. S. Rit, J. W. Wolthaus, M. van Herk, J.-J. Sonke, On-the-fly motion-compensated cone-beam CT using an a priori model of the respiratory motion, Medical Physics 36 (2009) 2283-2296.

19. Y. Zhang, F.-F. Yin, W. P. Segars, L. Ren, A technique for estimating 4D-CBCT using prior knowledge and limited-angle projections, Medical Physics 40 (2013) 121701.

20. C. Mory, G. Janssens, S. Rit, Motion-aware temporal regularization for improved 4D cone-beam computed tomography, Physics in Medicine & Biology 61 (2016) 6856-6877.

21. L. Wang, Y. Gao, F. Shi, G. Li, K.-C. Chen, Z. Tang, J. J. Xia,

D. Shen, Automated segmentation of dental CBCT image with prior-guided sequential random forests, Medical Physics 43 (2016) 336-346.
22. Y. Lei, X. Tang, K. Higgins, J. Lin, J. Jeong, T. Liu, A. Dhabaan, T. Wang, X. Dong, R. Press, W. J. Curran, X. Yang, Learning-based CBCT correction using alternating random forest based on auto-context model, Medical Physics 46 (2019) 601-618.
23. N. Yuan, S. Rao, Q. Chen, L. Sensoy, J. Qi, Y. Rong, Head and neck synthetic CT generated from ultra-low-dose cone-beam CT following image gently protocol using deep neural network, Medical Physics 49 (2022) 3263-3277.
24. J.-Y. Zhu, T. Park, P. Isola, A. A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, in: Proceedings of the IEEE International Conference on Computer Vision (2017) 2223-2232.
25. C. Jung, G. Kwon, J. C. Ye, Exploring patch-wise semantic relation for contrastive learning in image-to-image translation tasks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2022) 18260-18269.
26. C. Zheng, T.-J. Cham, J. Cai, The spatially-correlative loss for various image translation tasks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2021) 16407-16417.
27. M. R. Hosseinzadeh Taher, F. Haghighi, M. B. Gotway, J. Liang, CAiD: Context-aware instance discrimination for self-

supervised learning in medical imaging, in: Medical Imaging with Deep Learning 172 (2022) 535-551.
28. X. Wang, W. Jian, B. Zhang, L. Zhu, Q. He, H. Jin, G. Yang, C. Cai, H. Meng, X. Tan, F. Li, Z. Dai, Synthetic CT generation from cone-beam CT using deep-learning for breast adaptive radiotherapy, Journal of Radiation Research and Applied Sciences 15 (2022) 275-282.
29. L. Chen, X. Liang, C. Shen, S. Jiang, J. Wang, Synthetic CT generation from CBCT images via deep learning, Medical Physics 47 (2020) 1115-1125.
30. N. Yuan, S. Rao, Q. Chen, L. Sensoy, J. Qi, Y. Rong, Head and neck synthetic CT generated from ultra-low-dose cone-beam CT following Image Gently Protocol using deep neural network, Medical Physics 49 (2022) 3263-3277.
31. I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: Advances in Neural Information Processing Systems (2014) 2672-2680.
32. Y. Liu, Y. Lei, T. Wang, Y. Fu, X. Tang, W. J. Curran, T. Liu, P. Patel, X. Yang, CBCT-based synthetic CT generation using deep-attention cycleGAN for pancreatic adaptive radiotherapy, Medical Physics 47 (2020) 2472-2483.
33. H.-J. Tien, H.-C. Yang, P.-W. Shueng, J.-C. Chen, Cone-beam CT image quality improvement using Cycle-Deblur consistent adversarial networks (Cycle-Deblur GAN) for chest CT imaging

in breast cancer patients, Scientific Reports 11 (2021) 1133.
34. T.-H. Yong, S. Yang, S.-J. Lee, C. Park, J.-E. Kim, K.-H. Huh, S.-S. Lee, M.-S. Heo, W.-J. Yi, QCBCT-NET for direct measurement of bone mineral density from quantitative cone-beam CT: A human skull phantom study, Scientific Reports 11 (2021) 15083.
35. L. Deng, J. Hu, J. Wang, S. Huang, X. Yang, Synthetic CT generation based on CBCT using respath-cycleGAN, Medical Physics 49 (2022) 5317-5329.
36. T. Park, A. A. Efros, R. Zhang, J.-Y. Zhu, Contrastive learning for unpaired image-to-image translation, in: European Conference on Computer Vision (2020) 319-345.
37. W. Wang, W. Zhou, J. Bao, D. Chen, H. Li, Instance-wise hard negative example generation for contrastive learning in unpaired image-to-image translation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (2021) 14020-14029.
38. K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint (2014) arXiv:1409.1556.
39. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition (2009) 248-255.
40. D. P. Kingma, J. Ba, Adam: A method for stochastic

optimization, in: International Conference on Learning Representations (2015).

41. M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, S. Hochreiter, GANs trained by a two time-scale update rule converge to a local Nash equilibrium, in: Advances in Neural Information Processing Systems (2017) 6629-6640.

42. Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE Transactions on Image Processing 13 (2004) 600-612.

43. J. Bamba, K. Araki, A. Endo, T. Okano, Image quality assessment of three cone beam CT machines using the SEDENTEXCT CT phantom, Dentomaxillofacial Radiology 42 (2013) 20120445.

44. A. Katsumata, A. Hirukawa, S. Okumura, M. Naitoh, M. Fujishita, E. Ariji, R. P. Langlais, Effects of image artifacts on gray-value density in limited-volume cone-beam computerized tomography, Oral Surgery, Oral Medicine, Oral Pathology, Oral Radiology, and Endodontology 104 (2007) 829-836.

45. M. Naitoh, H. Aimiya, K. Nakata, K. Gotoh, E. Ariji, Stability of voxel values in cone-beam computed tomography. Oral Radiology 30 (2014) 147-152.

46. R. Molteni, Prospects and challenges of rendering tissue density in Hounsfield units for cone beam computed tomography, Oral Surgery, Oral Medicine, Oral Pathology and Oral Radiology 116 (2013) 105-119.

국 문 초 록

대조 학습을 이용한 Cone-Beam CT 영상의 구조 보존 품질 향상에 관한 연구

강세룡

융합과학부 방사선융합의생명전공

서울대학교 융합과학기술대학원

Cone-beam CT (CBCT)는 공간해상도가 높고 CT보다 촬영이 용이하여 그 활용도가 높다. 그러나, CBCT는 영상의 연조직 대조도와 Hounsfield Units (HU) 정량화 정확도가 CT 영상에 비해 부족하여 연조직 진단에는

한계가 있다. CBCT 영상의 품질을 향상시키기 위해 분석 모델링 방법, 반복 재구성 방법, Monte Carlo 시뮬레이션, 사전 지식 기반 규칙 기반 방법, 그리고 랜덤 포레스트 등을 포함한 다양한 기술들이 연구되었다. 그러나 이러한 방법들은 계산 복잡성으로 인해 시간이 많이 소요되며, 복잡한 아티팩트 감소에 있어서 CT 영상의 품질만큼 향상시키기에는 한계가 있다.

최근에는 CycleGAN이라는 새로운 해결책이 등장하여 CBCT 영상에서 CT와 유사한 영상을 생성하는 데 사용되고 있다. 이 방법의 목표는 CBCT 영상의 품질을 개선하여 CT 영상과 유사하게 만드는 것이다. 이 방법을 적용할 때 주요한 과제 중 하나는 원본 CBCT 영상의 구조를 보존하는 것이다. 그러나 CycleGAN이 입력 CBCT 영상의 구조를 보존하는 데 한계를 가지고 있다.

이 문제를 해결하기 위해, 대조 학습 기반 GAN을 기본 모델로 사용하여, 입력과 출력 사이의 대응 관계를 개선한 뒤, 입력 CBCT 영상의 구조를 보존하면서, CT와 유사한 영상 품질로 CBCT 영상의 품질을 향상시키고자 했다. Structure-preserving contrastive-learning based GAN (SPCGAN)을 설계한 뒤, CBCT와 CT 데이터셋을

활용하여 훈련시켰다. 이 과정에서는 새로운 조합의 손실 함수와 훈련 데이터셋에서 사전 훈련된 특징 추출기를 사용했다. 훈련에 사용된 손실 함수는 의미론적 관계 일관성 손실, 공간 상관 관계 손실, 그리고 재구성 손실로 구성했다. 의미론적 관계 일관성 손실과 공간 상관 관계 손실은 이미지 내 의미론적 및 공간적 패턴을 포착하여 구조를 보존하는 이미지를 효과적으로 생성하는 데 사용했고, 재구성 손실을 최소화함으로써 입력 CBCT 영상에 대한 더 세부적인 정보를 표현에 포함시켰다.

FID (Frechet Inception Distance), PSNR (Peak Signal-to-Noise Ratio), MAE (Mean Absolute Error), RMSE (Root Mean Square Error), 및 structure score를 지표로 SPCGAN이 생성한 영상의 품질을 평가했다. 환자 데이터를 이용한 실험을 통해 SPCGAN이 생성한 영상이 다른 기존 모델이 생성한 영상보다 FID, PSNR, MAE, RMSE, structure score 측면에서 우수하다는 것을 입증했다. 본 연구는 CBCT 영상의 해부학적 구조를 보존하면서 영상의 품질을 CT 영상과 유사하게 향상시키는 딥러닝 방법을 제안하였다. 이를 통해, CBCT의 정확한 HU 정량화를 가능하게 하여, 다양한 임상에서 CBCT를 사용할 수

있게 할 것이다.

주요어 : 대조 학습, 구조 보존, Cone-beam CT 영상, 딥러닝, SPCGAN

학 번 : 2013-22422