Towards AI-Driven Radiology Education: A Self-supervised Segmentation-Based Framework for High-Precision Medical Image Editing

Kazuma Kobayashi[1,2(✉)], Lin Gu[2,3], Ryuichiro Hataya[4,2], Mototaka Miyake[5], Yasuyuki Takamizawa[5], Sono Ito[5], Hirokazu Watanabe[5], Yukihiro Yoshida[5], Hiroki Yoshimura[6], Tatsuya Harada[3,2], and Ryuji Hamamoto[1,2]

[1] National Cancer Center Research Institute, Tokyo, Japan
kazumkob@ncc.go.jp
[2] RIKEN Center for Advanced Intelligence Project, Tokyo, Japan
[3] The University of Tokyo, Tokyo, Japan
[4] RIKEN Information R&D and Strategy Headquarters, Tokyo, Japan
[5] National Cancer Center Hospital, Tokyo, Japan
[6] Hiroshima University School of Medicine, Hiroshima, Japan

Abstract. Medical education is essential for providing the best patient care in medicine, but creating educational materials using real-world data poses many challenges. For example, the diagnosis and treatment of a disease can be affected by small but significant differences in medical images; however, collecting images to highlight such differences is often costly. Therefore, medical image editing, which allows users to create their intended disease characteristics, can be useful for education. However, existing image-editing methods typically require manually annotated labels, which are labor-intensive and often challenging to represent fine-grained anatomical elements precisely. Herein, we present a novel algorithm for editing anatomical elements using segmentation labels acquired through self-supervised learning. Our self-supervised segmentation achieves pixel-wise clustering under the constraint of invariance to photometric and geometric transformations, which are assumed not to change the clinical interpretation of anatomical elements. The user then edits the segmentation map to produce a medical image with the intended detailed findings. Evaluation by five expert physicians demonstrated that the edited images appeared natural as medical images and that the disease characteristics were accurately reproduced.

Keywords: Image editing · Self-supervised segmentation · Education

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-43895-0_38.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14221, pp. 403–413, 2023.
https://doi.org/10.1007/978-3-031-43895-0_38

1 Introduction

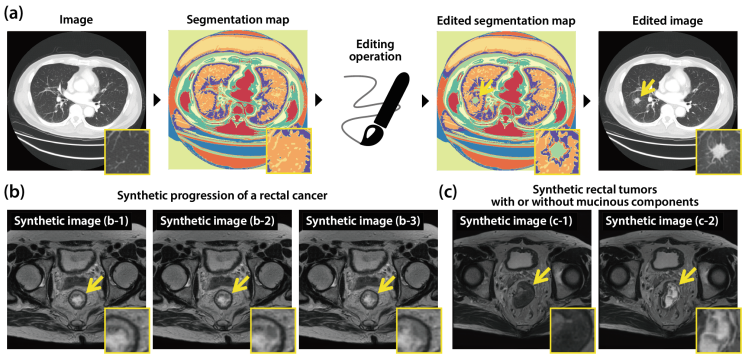

Fig. 1. Editing of anatomical elements. (a) Users can edit the segmentation map obtained from an input image to express intended fine-grained disease characteristics. A spiculated lung nodule was generated. **(b)** Synthetic disease progression showing a normal-appearing rectum (b-1), a rectal tumor extending into the submucosal layer (b-2), and the tumor extending into the muscularis propria (b-3). **(c)** A synthetic rectal tumor (c-1) and the contrasting tumor with T2 hyperintensity of extracellular mucin suspicious for mucinous adenocarcinoma (c-2).

Despite the success of artificial intelligence (AI) in aiding diagnosis, its application to medical education remains limited. Trainee physicians require several years of experience with a diverse range of clinical cases to develop sufficient skills and expertise. However, designing educational materials solely based on real-world data poses several challenges. For example, although small but significant disease characteristics (e.g., depth of cancer invasion) can sometimes alter diagnosis and treatment, collecting pairs with and without these characteristics is cumbersome. Another major challenge is longitudinal tracking of pathological progression over time (e.g., from the early stage of cancer to the advanced stage), which is difficult to understand because medical images are often snapshots. Privacy is also a concern since images of educational materials are widely distributed. Therefore, medical image editing that allows users to generate their intended disease characteristics is useful for precise medical education [3].

Image editing can synthesize low- or high-level image contents [11]. Our goal is to develop high-precision medical image editing according to the fine-grained characteristics of individual diseases, rather than at the level of disease categories. For example, even if two diseases belong to the same disease category of "lung tumor," the impression of benign or malignant will differ depending on fine-grained characteristics, such as whether the margins are "smooth" or "spiculated." In this case, our approach is to edit the tumor margins to be smooth or spiculated. These fine-grained characteristics consist of low- to mid-level image

features to distinguish the substructures of organs and diseases, which we call *anatomical elements*.

Several types of image editing techniques for medical imaging have been introduced, mainly using generative adversarial networks [5] and, more recently, diffusion models [2]. Nevertheless, editing specific anatomical elements remains a challenge [1,11]. *Latent space manipulation* generates images by controlling latent feature axes [4,14], but the editable attributes are often global rather than fine-grained. *Conditional generation* can precisely edit image content by using class or segmentation labels. However, it requires manually provided labels [15] or virtual models [18], which are labor-intensive. Additionally, accurately modeling certain fine-grained characteristics, such as the textual variations of disease, can be a daunting task. *Image interpolation* [17] requires actual images with targeted content, which limits its applicability.

Here, we propose a novel framework for image editing called U3-Net that allows the generation of anatomical elements with precise conditions. The core technique is *self-supervised segmentation*, which aims to achieve pixel-wise clustering without manually annotated labels [6,7]. As shown in Fig. 1**a**, U3-Net converts an input image into a segmentation map corresponding to the anatomical elements. Once the user has completed editing, U3-Net synthesizes an image in which the targeted anatomical element has been modified. As a result, our synthesized medical images can highlight hypothetical pathological changes and significant clinical differences in a single image. For example, Fig. 1**b** shows that whether or not rectal cancer invades the muscularis propria (i.e., b-2 vs. b-3) affects cancer staging (i.e., T1 vs. T2) as well as treatment strategy (i.e., endoscopic resection vs. surgery). The distinction between mucinous and non-mucinous rectal cancers (see Fig. 1**c**) is also important to estimate the better or worse prognosis of the disease. These synthetic images can help trainees intuitively comprehend clinically significant findings and alleviate privacy concerns. Five expert physicians evaluated the edited images from a clinical perspective using two datasets: a pelvic MRI dataset and chest CT dataset.

Contributions: Our contributions are as follows:

- We propose a novel image-editing algorithm, U3-Net, to synthesize images for medical education via self-supervised segmentation.
- U3-Net can faithfully synthesize intended anatomical elements according to the editing operation on the segmentation labels.
- Evaluation by five expert physicians showed that the edited images were natural as medical images with the intended features.

2 Methodology

U3-Net consists of three neural networks: *encoder*, *decoder*, and *discriminator* (see Fig. 2). The encoder achieves self-supervised segmentation with a feature extraction (FE) module and a pixel-wise clustering (CL) module. We perform pixel-wise clustering under the constraint of invariance to photometric and

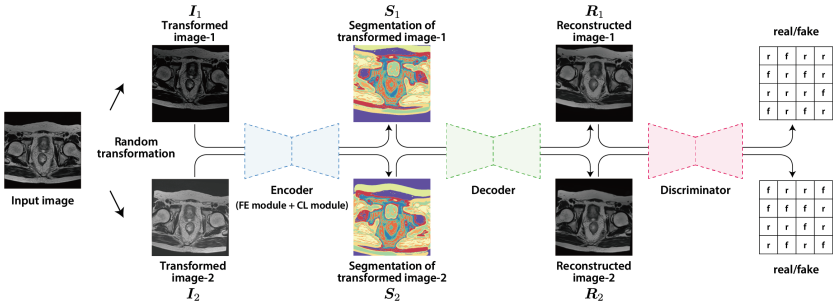

Fig. 2. Overall architecture of U3-Net. We apply two random transformations to the input image to produce images in different views, $\mathbf{I}_1$ and $\mathbf{I}_2$. The encoder converts the transformed images into quantized embedding as well as segmentation maps consisting of cluster indices, $\mathbf{S}_1$ and $\mathbf{S}_2$. Pixel-wise clustering, which should be consistent between views, is performed for the self-supervised segmentation. The decoder generates reconstructed images, $\mathbf{R}_1$ and $\mathbf{R}_2$, from the quantized embedding maps. The discriminator adversarially enhances the natural appearance by judging whether the images are real or fake on a pixel-by-pixel basis.

geometric transformations [6], with the assumption that these transformations should not change the clinical interpretation of the anatomical elements. Given a pair of differently transformed images, the FE module produces *embedding maps* corresponding to the input images. The CL module then performs K-means clustering on the embedding maps to produce two interchangeable outputs: *segmentation maps* and corresponding *quantized embedding maps*. These outputs are trained to be consistent between the two views. The decoder then estimates the corresponding images from the quantized embedding maps, while the discriminator forces the decoder to produce more realistic images.

2.1 First Training Stage for Self-supervised Segmentation

The training process for U3-Net is two-stage. First, we train the encoder and decoder (excluding the discriminator) to conduct K-class self-supervised segmentation. To achieve pixel-wise clustering that is consistent between two transformed views of the input images, we introduce four constraints: intra-cluster pull force, inter-cluster push force, cross-view consistency, and reconstruction loss.

Random Image Transformation: We consider a sequence of image transformations $[t_1, \ldots, t_n]$ specified by the type (e.g., image rotation) and magnitude (e.g., degree of rotation) of each transformation: $\mathcal{T} = t_n \circ t_{n-1} \circ \cdots \circ t_1$. Two random transformation sequences are applied to an input image $\mathbf{I} \in \mathbb{R}^{C \times H \times W}$ to produce two transformed images, $\mathcal{T}_1(\mathbf{I}) = \mathbf{I}_1$ and $\mathcal{T}_2(\mathbf{I}) = \mathbf{I}_2$. The FE module of the encoder produces two embedding maps $f(\mathbf{I}) = \mathbf{E} \in \mathbb{R}^{D \times H \times W}$, $\mathbf{E}_1$ and $\mathbf{E}_2$, which are then fed into the CL module.

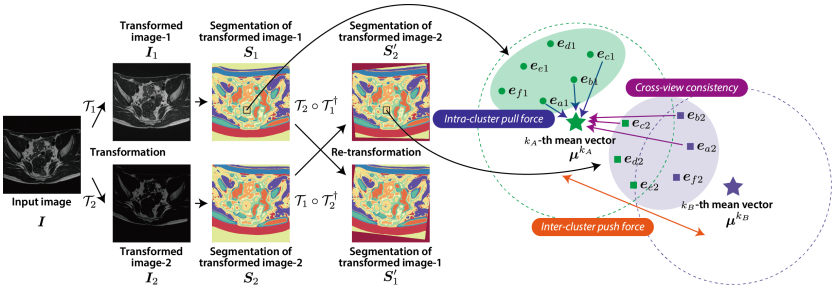

Fig. 3. Transformation-invariant pixel-wise clustering. Suppose that the majority of pixels inside the black box in $\mathbf{S}_1$ are assigned to the k_A-th cluster. The intra-cluster pull force causes the embedding vectors $\mathbf{e}_{a1}, \ldots \mathbf{e}_{f1}$ to adhere to the mean vector $\boldsymbol{\mu}^{k_A}$. From the other viewpoint, some of the same pixels, $\mathbf{e}_{a2}$, $\mathbf{e}_{b2}$, and $\mathbf{e}_{f2}$, are assigned to the k_B-th cluster, which can be assessed by re-transforming $\mathbf{S}_2$ into the coordination of $\mathbf{S}_1$. *Cross-view consistency loss* forces the embedding vectors of one view, $\mathbf{e}_{a2}, \ldots \mathbf{e}_{f2}$, to match the mean vector of the other view $\boldsymbol{\mu}^{k_A}$. The inter-cluster push force maintains the distance between the mean vectors.

Cluster Assignment and Update: In the CL module, K-means clustering in the first iteration initializes K mean vectors $\boldsymbol{\mu}^k \in \mathbb{R}^D$. Then, the embedding vector of the i-th pixel $\mathbf{e}_{i\in\{1,\ldots,H\times W\}} \in \mathbb{R}^D$ in the embedding maps, $\mathbf{E}_1$ and $\mathbf{E}_2$, is assigned to the cluster with the nearest mean vector as follows: $y_i = \text{argmin}_{k\in\{1,\ldots,K\}} \|\mu^k - \mathbf{e}_i\|^2$, where y_i is the cluster index of the i-th pixel. By replacing embedding vectors with their respective mean vectors, quantized embedding maps, $\mathbf{E}_{\text{q}1}$ and $\mathbf{E}_{\text{q}2}$, are generated $g(\mathbf{E}) = \mathbf{E}_{\text{q}} = [\boldsymbol{\mu}^{y_1}, \ldots, \boldsymbol{\mu}^{y_{H\times W}}] \in \mathbb{R}^{D\times H\times W}$. The cluster indices form the segmentation maps $\mathbf{S} = [y_1, \ldots, y_{H\times W}] \in \mathbb{R}^{H\times W}$, $\mathbf{S}_1$ and $\mathbf{S}_2$. The mean vectors $\boldsymbol{\mu}^k$ are updated by using the exponential moving average [9].

Intra-cluster Pull Force: For transformation-invariant pixel-wise clustering, we define four loss terms. The first term, *cluster loss*, forces the embedding vectors to adhere to the associated mean vector (see Fig. 3), as defined: $L_{\text{cluster}} = \sum_{i\in H\times W} \|\boldsymbol{\mu}^{y_i} - \mathbf{e}_i\|^2$.

Inter-cluster Push Force: The second term, *distance loss*, pushes the distance between the mean vectors above a margin parameter m (see Fig. 3), as defined: $L_{\text{dist}} = \frac{1}{K(K-1)} \sum_{k_A=1}^{K} \sum_{k_B=1, k_B\neq k_A}^{K} [2m - \|\boldsymbol{\mu}^{k_A} - \boldsymbol{\mu}^{k_B}\|]_+^2$, where k_A and k_B indicate two different cluster indices.

Cross-view Consistency: The segmentation maps from the different views, $\mathbf{S}_1$ and $\mathbf{S}_2$, should overlap after re-transforming to align the coordinates. Such a re-transform is composed of inverse and forward geometric transformations: $\mathcal{T}_2(\mathcal{T}_1^\dagger(\mathbf{S}_1)) = \mathbf{S}'_1$ and $\mathcal{T}_1(\mathcal{T}_2^\dagger(\mathbf{S}_2)) = \mathbf{S}'_2$. The inverse transformations of the

photometric transformations are not considered. Using the re-transformed segmentation maps, we impose a third term, *cross-view consistency loss*, which forces the embedding vectors of one view to match the mean vector of the other (see Fig. 3), as defined: $L_{\text{cross}} = \sum_{i \in H \times W} \|\boldsymbol{\mu}^{y_{i2}} - \mathbf{e}_{i1}\|^2 + \sum_{i \in H \times W} \|\boldsymbol{\mu}^{y_{i1}} - \mathbf{e}_{i2}\|^2$.

Reconstruction Loss: Without user editing, the decoder reconstructs the input images from quantized embedding maps $h(\mathbf{E}_{\text{q}}) = \mathbf{R} \in \mathbb{R}^{C \times H \times W}$. We thus employ *reconstruction loss*, which minimizes the mean squared error between the reconstructed and input images.

Learning Objective: The weighted sum of the loss functions is set to be minimized: $L_{\text{total}} = w_{\text{cluster}} L_{\text{cluster}} + w_{\text{dist}} L_{\text{dist}} + w_{\text{cross}} L_{\text{cross}} + w_{\text{recon}} L_{\text{recon}}$.

2.2 Second Training Stage for Faithful Image Synthesis

In the second stage, we train the decoder and discriminator (excluding the encoder) to produce naturally appearing images from the quantized embedding maps. The decoder, initially optimized in the first training stage, undergoes further training to enhance its image generation capabilities. We impose adversarial learning with an extended reconstruction loss term, called *appearance loss*. The training is performed only in a single view.

Appearance Loss: Appearance loss combines *mean squared loss* L_{mse}, *focal frequency loss* L_{ffl} [8], *perceptual loss* L_{lpips} [19], and *intermediate loss* L_{int}, as follows: $L_{\text{app}} = w_{\text{mse}} L_{\text{mse}} + w_{\text{ffl}} L_{\text{ffl}} + w_{\text{lpips}} L_{\text{lpips}} + w_{\text{int}} L_{\text{int}}$, where *intermediate loss* L_{int} refers to the L2 distance of the intermediate features of the discriminator between the reconstructed and input images.

Learning Objective: We impose *generator loss* L_{gen} for the decoder to produce more faithful images by deceiving the discriminator, and *discriminator loss* L_{dis} for the discriminator to judge the real or fake of the images as the per-pixel feedback [16]. We also add *cutmix augmentation* L_{cutmix} and *consistency regularization* L_{cons} to the latter [16]. In this stage, the decoder and discriminator are trained by alternately minimizing the following competing objectives: $L_{\text{Dec}} = L_{\text{app}} + w_{\text{gen}} L_{\text{gen}}$ and $L_{\text{Dis}} = w_{\text{dis}} L_{\text{dis}} + w_{\text{cutmix}} L_{\text{cutmix}} + w_{\text{cons}} L_{\text{cons}}$.

2.3 Inference Stage for Medical Image Editing

After training, the encoder can output a segmentation map from a testing image. As shown in Fig. 1**a**, when a user edits the segmentation map $\mathbf{S} \rightarrow \mathbf{S}'$ by changing the cluster indices $y_i \rightarrow y_i'$, the quantized embedding map is subsequently updated $\mathbf{E}_{\text{q}} \rightarrow \mathbf{E}'_{\text{q}}$ by reassigning the mean vectors according to the edited indices $\boldsymbol{\mu}^{y_i} \rightarrow \boldsymbol{\mu}^{y_i'}$. Finally, the decoder converts the quantized embedding map into a synthetic image with the intended disease characteristics $h(\mathbf{E}'_{\text{q}}) = \mathbf{R} \in \mathbb{R}^{C \times H \times W}$.

3 Experiments and Results

Implementation and Datasets: All neural networks were implemented in Python 3.8 using the PyTorch library 1.10.0 [12] on an NVIDIA Tesla A100 GPU running CUDA 10.2. The encoder, decoder, and discriminator were implemented based on U-Net [13] (see **Supplementary Information** for details). The *pelvic MRI dataset* with rectal cancer contained 289 image series for training and 100 image series for testing. For each image series, the min-max normalization converted the pixel values to $[-1, 1]$. The *chest CT dataset* with lung cancer contained 500 image series for training and 100 image series for testing. The CT values in the range $[-2048, 2048]$ were normalized to $[-1, 1]$. Both were in-house datasets collected from a single hospital. Every image series comprises two-dimensional (2D) consecutive slices, and we applied our algorithm on a per 2D slice basis.

Self-supervised Medical Image Segmentation: We began by optimizing the hyperparameters to achieve self-supervised segmentation. Appropriate transformations were selected from six candidate functions: t_1, `RandomHorizontalFlip`, t_2, `RandomAffine`, t_3, `ColorJitter`, t_4, `RandomGaussianBlur`, t_5, `RandomPosterize`, t_6, `RandomGaussianNoise`. Because anatomical elements, including the substructures of organs and diseases, are too detailed for human annotators to segment, it was difficult to create ground-truth labels. Therefore, the training configuration was selected based on the consensus of two expert radiologists with domain knowledge. By comparing different settings on the pelvic MRI training dataset (see **Supplementary Information**), the number of segmentation classes of 10, the combination of t_1, t_2, and t_3 with moderate magnitude, the weakly imposed reconstruction loss, and a certain value of the margin parameter were considered suitable for self-supervised segmentation. In particular, we found that reconstruction loss is essential for obtaining segmentation maps corresponding to anatomical elements, although such a loss term was not included in previous studies [6,7]. A similar configuration was applied to the chest CT training dataset. The resultant segmentation maps are shown in Fig. 4**ab**. The anatomical substructures, including the histological structure of the colorectal wall and subregions within the lung, corresponded well with the segmentation maps in both the pelvic MRI and chest CT testing datasets. Because our self-supervised segmentation extracts low- to mid-level image content, a semantic object (e.g., rectum or lung cancer) typically consists of multiple segmentation classes shared with other objects (see the magnified images in Fig. 4**ab**). These anatomical elements may be too detailed for humans to annotate, demonstrating the necessity of self-supervised segmentation for high-precision medical-image editing.

Evaluation of the Synthesized Images: We measured the quality of image reconstruction using mean square error (MSE), structural similarity (SSIM), and peak signal-to-noise ratio (PSNR). The mean ± standard deviations of MSE,

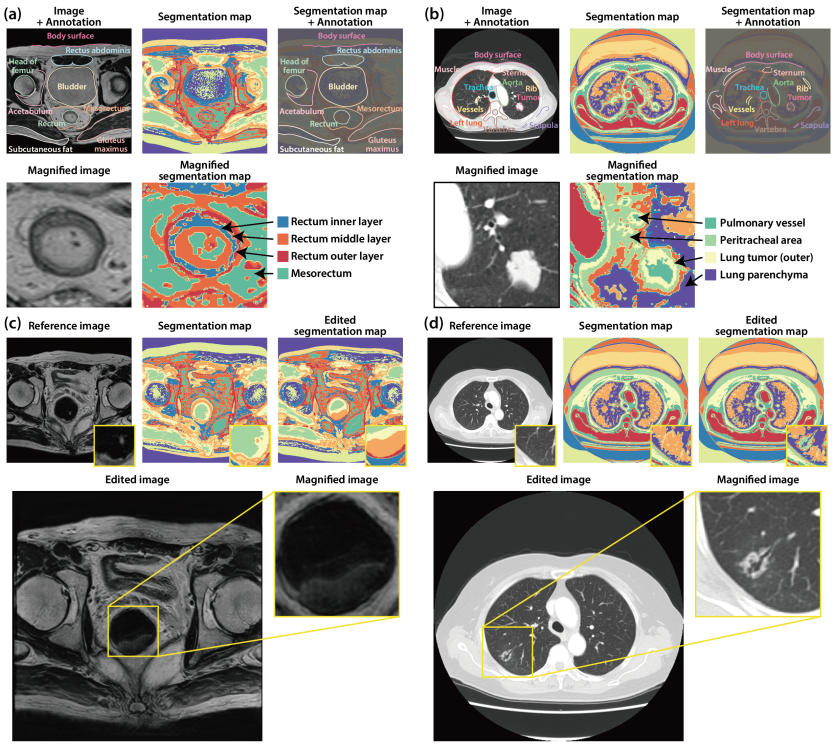

Fig. 4. Results of the image segmentation and editing. The segmentation maps were well aligned with the anatomical elements in both **(a)** the pelvic MRI and **(b)** the chest CT testing datasets. **(c)** A synthetic image generated by editing the testing image with the caption, *"Axial T2-weighted MR image shows a tumor approximately 4 cm in size on the dorsal wall of the rectum. The deepest structure of the rectal wall was intact, indicating no infiltration beyond the muscularis propria."* **(d)** A synthetic image with the caption, *"Axial CT image showing a pulmonary nodule with a length of 2–3 cm and a cavity on the dorsal side of the right upper lobe of the lung."*

SSIM, and PSNR were $1.41 \times 10^{-2} \pm 1.04 \times 10^{-2}$, $7.40 \times 10^{-1} \pm 0.57 \times 10^{-1}$, and 22.5 ± 2.7 in the pelvic MRI testing dataset and $5.03 \times 10^{-4} \pm 3.03 \times 10^{-4}$, $9.08 \times 10^{-1} \pm 0.34 \times 10^{-1}$, and 38.6 ± 1.7 in the chest CT testing dataset. Subsequently, segmentation maps from the testing images were edited to generate images with the intended characteristics (see Fig. 4**cd**). Five expert physicians (two diagnostic radiologists, two colorectal surgeons, and one thoracic surgeon) assessed them from a clinical perspective. First, we tested whether the evaluators could identify real or synthesized images from 20 images, which include ten real images and ten synthesized images. The accuracies (i.e., the ratio of images correctly identified as real or synthetic) were 0.69 ± 0.11 and 0.65 ± 0.11, for the pelvic MRI and chest CT testing datasets, respectively. Note that when the synthetic images cannot be distinguished at all, the accuracy should be 0.5. Second, we presented

image captions explaining the radiological features, which also represented the editing intention for the synthetic images. We asked the evaluators to rate each presented image from A to C. ***A****: The image is natural as a medical image, and the caption is consistent with the image.* ***B****: The image is natural as a medical image, but the caption is NOT consistent with the image.* ***C****: The image is NOT natural as a medical image.* This test was conducted after informing the evaluators of the assumption that all 20 images could be synthetic, without indicating which image was real or synthetic. As a result, the ratio of synthetic images (vs. that of real images) categorized as A, B, and C were 0.80 ± 0.15 (vs. 0.78 ± 0.20), 0.02 ± 0.04 (vs. 0.08 ± 0.07), and 0.18 ± 0.11 (vs. 0.14 ± 0.13) for the pelvic MRI testing dataset, and 0.74 ± 0.28 (vs. 0.76 ± 0.30), 0.08 ± 0.09 (vs. 0.12 ± 0.15), and 0.18 ± 0.21 (vs. 0.12 ± 0.14) for the chest CT testing dataset. There were no significant differences between real and synthetic images (t-test: $p > 0.05$). Consequently, the majority of the edited images were natural-looking medical images with accurately reproduced disease features.

4 Conclusion

In this study, we propose a medical image-editing framework to edit fine-grained anatomical elements. The self-supervised segmentation extracted low- to mid-level content of medical images, which corresponded well to the clinically meaningful substructures of organs and diseases. The majority of the edited images with intended characteristics were perceived as natural medical images by several expert physicians. Our medical image editing method can be applied to medical education, which has been overlooked as an application of AI. Future challenges include improving scalability with fewer manual operations, validating segmentation maps from a more objective perspective, and comparing our proposed algorithm with existing methods, such as those based on superpixels [10].

Data use declaration and acknowledgment: The pelvic MRI and chest CT datasets were collected from the National Cancer Center Hospital. The study, data use, and data protection procedures were approved by the Ethics Committee of the National Cancer Center, Tokyo, Japan (protocol number 2016-496). Our implementation and all synthesized images will be available here: https://github.com/Kaz-K/medical-image-editing.

References

1. Chen, Y., et al.: Generative adversarial networks in medical image augmentation: a review. Comput. Biol. Med. **144**, 105382 (2022). https://doi.org/10.1016/j.compbiomed.2022.105382
2. Dhariwal, P., Nichol, A.: Diffusion Models Beat GANs on Image Synthesis. In: Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., Vaughan, J.W. (eds.) Advances in Neural Information Processing Systems (NeurIPS). vol. 34, pp. 8780–8794. Curran Associates, Inc. (2021). https://proceedings.neurips.cc/paper_files/paper/2021/file/49ad23d1ec9fa4bd8d77d02681df5cfa-Paper.pdf