Intraoral Image Generation by Progressive Growing of Generative Adversarial Network and Evaluation of Generated Image Quality by Dentists

Kazuma Kokomoto
Osaka University
Rena Okawa
Osaka University
Kazuhiko Nakano
Osaka University
Kazunori Nozaki (knozaki@dent.osaka-u.ac.jp)
Osaka University Dental Hospital, 1-8 Yamada-oka, Suita, Osaka 565-0871, Japan

Research Article

Keywords: Deep Learning, Generative Adversarial Network, Pediatric dentistry

Posted Date: December 2nd, 2020

DOI: https://doi.org/10.21203/rs.3.rs-117942/v1

Abstract

Dentists need experience with plenty of clinical cases to practice specialized skills. However, the need to protect patients' private information limits the ability to utilize lots of intraoral images obtained from clinical cases. In this study, since generating realistic images could making utilizing lots of intraoral images possible, intraoral images are generated by using a progressive growing of generative adversarial network. 35,254 intraoral images were used as training data with resolutions of 128×128, 256×256, 512×512, and 1,024×1,024. The results of training datasets with and without data augmentation were compared. The sliced Wasserstein distance (SWD) was calculated to evaluate the generated images. Next, 50 real images and 50 generated images for each resolution were randomly selected and shuffled. Twelve pediatric dentists were asked to observe these images and assess whether each was real or generated. The accuracy of the assessment of the 1,024×1,024 images was significantly higher than that of the other resolutions. In conclusion, generated intraoral images with resolutions of 512×512 or lower were so realistic that the dentists could not distinguish whether they were real or generated. This implies that generated images can be used for dental education or data augmentation for deep learning free from privacy restrictions.

Introduction

Medical workers should get experience with as many clinical cases as possible to develop their professional competence [1]. Oral examination is one of the most important techniques in decision making for dentistry. Thus, dental clinicians need to study as many intraoral images as they can to increase their accuracy in oral examination. This is especially true for pediatric dentistry since the oral environment of children drastically changes as children grow. In particular, it is important to understand the difference between primary teeth and adult teeth, the timing of tooth replacement, maxillofacial bone growth, or changes in dental occlusion and alignment.

In recent years, deep learning using a convolutional neural network (CNN) has been developed prominently for computer vision and been applied to medical image analysis. GAN is one of the image generation methods based on deep learning [2], which is a class of unsupervised machine learning system and consists of a generator network and discriminator network. In some reports, generating medical images using a GAN has succeeded [3][4]. However, those studies have dealt with only grayscale X-ray images, and no evaluation of generated medical images by experts has been done.

The need to protect patient's private information restricts the use of intraoral images obtained in clinical cases. For that reason, sharing as many intraoral images as possible among different hospitals is difficult [5]. Thus, increasing the number of images of similar clinical cases is essential in clinical dental education. To solve both, non-real intraoral images are expected to be generated by using GAN for education.

In this research, we describe generation of not gray-scale but full-color intraoral images using PGGAN, and evaluation of generated intraoral images in terms of both quantity performance and visual quality assessed by pediatric dentists.

Materials And Methods

2.1. Image datasets and preprocessing

The 35,254 intraoral images were used in this study as training data and were obtained from the patients who received dental treatment between August 2008 and March 2019 at Osaka University Dental Hospital, Department of Pediatric Dentistry, Osaka, Japan. All images consisted of about 4000 x 3000 pixels and had RGB color. Those images included primary, mixed, and permanent dentitions. There were various teeth conditions: healthy, caries, stain, composite resin restoration, metal inlay, stainless steel crown, space maintainer, orthodontic appliance, hypomineralization, and so on.

All images were resized to 128x128, 256x256, 512x512, and 1,024x1,024 pixels and converted to the JPEG format. It is said that a small number of datasets can easily lead to overfitting, but the number of training data we could use in this study was limited, so data augmentation was adopted [6]. The results with and without data augmentation were compared. Each resized image was randomly augmented five times using the ImageDataGenerator function that is built in Keras [7], whose parameters were rotation_range = 20, width_shift_range = 0.1, height_shift_range = 0.1, shear_range = 0.1, zoom_range = 0.1, horizontal_flip = True, and fill_mode = 'nearest'. All computation was performed on one Ubuntu 18.04 desktop computer with one NVIDIA TITAN RTX.

2.2. Network architecture

GAN is a framework that consists of two adversarial neural networks based on deep learning. One is a generative network (*G)*, and the other is a discriminative network (*D*). *D* is trained to distinguish between true samples and generated samples, and *G* is trained to fool *D* with its generated samples. This framework corresponds to a strategy to make decisions to minimize the possible damage. As a result of the learning, *G* becomes able to output realistic images. However, many kinds of GAN have failed to generate high resolution images, mainly generating 256x256 or lower resolution images, until the Progressive Growing of Generative Adversarial Network (PGGAN) made it possible to generate 1,024x1,024 high resolution images from latent vectors with improved quality, stability, and variation [8]. The PGGAN starts training with low resolution at first and progressively increases the resolution by adding new layers to both the generator and discriminator (Fig. 1). We trained the PGGAN using resized real intraoral images of 128x128, 256x256, 512x512, and 1,024x1,024 resolution with and without data augmentation, and the generated images of each resolution generated by *G* were evaluated.

2.3. Evaluation

The visual quality of the intraoral images is considered as an important factor. 50 generated images that seemed to be real at first glance were manually selected by one pediatric dentist, and added 50 real images, then randomly shuffled and reordered. All chosen images were different for all resolutions with 128x128, 256x256, 512x512, and 1,024x1,024. Another 12 pediatric dentists were asked to observe those randomly shuffled images which consists of 50 real images and 50 generated images described above, and assess whether each image was real or generated. The differences in the results with each resolution were evaluated by the accuracy. The accuracy was calculated as follows:

$$accuracy = \frac{TP + TN}{TP+FP+TN+FN}$$

TP, FP, TN, and FN indicate the numbers of true positive answers, false positive answers, true negative answers, and false negative answers, respectively. The accuracy for each resolution was analyzed using a multiple comparison with Tukey's honestly significant differences test (TukeyHSD test) after checking the normality of each answer with Shapiro-Wilk test and checking the equal variances with Bartlett's test. All p-values < 0.05 were considered to be statistically significant.

In addition, the discriminative parts of the images at which the dentists looked to determine whether given images were true or generated were counted. These parts consisted

of four factors: tooth, alignment, soft tissues, and others.

Some metrics have been proposed to evaluate the quantitative performance of generated images. In this study, we used Sliced Wasserstein Distance (SWD) metrics, which was proposed in the original PGGAN article [8]. The SWD was calculated for each level of the Laplacian pyramids [9] of the randomly sampled 16,384 images, and the average SWD of each Laplacian pyramid was calculated. A small SWD value indicates that the true images and generated images seemed to be similar in both appearance and variation, so the weights parameters of *G* that showed the lowest SWD values ware used for generating images.

Results

3.1. Visual quality evaluation

The visual quality was compared between the real images and the generated images (Fig. 2). When compared to the original images (Fig. 2A), the boundaries of the training images with data augmentation are obviously filled with the closest pixel value, and this is repeated for all the empty values, which is the effect of the fill_mode of ImageDataGenerator (Fig. 2B). The boundaries of the generated images trained without augmented real images are clear (Fig. 2C). On the other hand, the boundaries of the trained with augmented real images are filled with the closest pixel value (Fig. 2D), which is as same as Fig. 2B.

The generated images were evaluated by twelve pediatric dentists for visual quality. The generated images by PGGAN trained with augmented real images were not evaluated because the boundaries of the images were clearly different from the boundaries of the real images as described above, so only the generated images by PGGAN trained without augmented real images were evaluated.

The accuracy was calculated for each resolution using the responses of the twelve pediatric dentists, and statistical analysis was performed. The responses of each resolution were tested for normality using the Shapiro-Wilk test. The p-value of resolutions of 128x128, 256x256, 512x512, and 1,024x1,024 were 0.9669, 0.1934, 0.2358, and 0.4393, respectively. As a result, each answer was considered to have a normal distribution. Next, Bartlett's test was performed for homogeneity of variances. The p-value was 0.9938, which means that each answer was considered to be from populations with equal variances (Table). Therefore, the TukeyHSD test was utilized for multiple comparison. The accuracy of 1,024x1,024 resolution was significantly higher than that of 128x128, 256x256, and 512x512 resolutions (p-values were 0.0013, 0.0472, and 0.0005, respectively). No statistically significant difference was observed with other combinations (Fig. 3).

What parts of the images the dentists looked at to determine whether they were true or generated were counted (Fig. 4). It was found that most of the dentists have used tooth as the assessment criteria. It was also found that the ratio for tooth increased as the resolutions increased. The other factors were not used as assessment criteria as much as tooth.

3.2. Quantitative evaluation

Table shows the SWD values of each resolution. With resolutions of 128x128, 256x256, and 512x512, the SWD was reduced by data augmentation of the training data, which means that the quality of the generated image was improved. On the other hand, the SWD value was increased with the resolution of 1,024x1,024, which means that data augmentation has shown negative improvement in the quality of the generated image.

Discussion

4.1. Quality of generated oral images

In this study, we have presented synthetic intraoral images by using PGGAN. When the dentists evaluated the visual quality of the generated images, only the resolution of 1,024x1,024 showed significantly higher accuracy (Fig. 3). Since 1,024x1,024 is a higher resolution, the generated image may have been noticeably rougher than the other images, which lead to dentists' easy discrimination. This is considered to be consistent with the fact that the SWD value of 1,024x1,024 is only higher than that of the other resolutions (Table). In other words, it is considered that the generated intraoral images with lower resolutions are so good that the dentist cannot distinguish whether they are real or generated images.

The reason why the SWD value of 1,024x1,024 has been increased by data augmentation is considered to be the filled pixels on boundaries. Although the quality of the image deteriorates due to the filled pixels of boundaries, the filled area is small in 512x512 or lower resolutions, and it is considered that the increase of image variation by data augmentation contributes to decrease the SWD value. On the other hand, in 1,024x1,024 resolution, since the filled area is larger than that of lower resolutions, it is considered that the influence of deterioration of image quality has been more affected than increase of image variation, which lead to the increase of the SWD value. In 512x512 or lower resolutions, data augmentation has decreased the SWD values. If the filled pixels of image boundaries were cut out, the rest of central images can be better quality than the generated images by PGGAN trained without data augmentation.

Focusing on the teeth is one of the key factors in distinguishing whether an image was real or generated (Fig. 4). Some teeth in generated images look strange compared with natural teeth; however, teeth alignment or soft tissue, such as the tongue, lips, and nose, look realistic enough (Fig. 5). In addition, the ratio for teeth used as an assessment criterion increased as resolutions increased (Fig. 4), which implies the teeth can be easily examined at higher resolutions. This is because, for primary and permanent teeth, there are many types of anatomical shapes, colors, fissures, cusps, and outlines than types of alignments or soft tissues for each person, so PGGAN cannot learn tooth features with our limited dataset.

4.2. Research limitation

The efficiency of generating realistic images by PGGAN markedly depends on the number of learned images. Even if dentists often take intraoral images in daily clinical treatment, only about 35,000 intraoral images could have been stored between 2008 and 2019. Original PGGAN has trained with CelebA dataset, which consists of about 200,000 face images [10], and has achieved significant performance. Thus, it is considered that our result may not utilize the maximum efficiency of PGGAN.

When dentists try to take intraoral images, cheek retractors and intraoral mirrors are placed in the mouth, and it is desirable that the camera axis and patient's head be parallel [11]. However, since most children cannot stay still, and cannot open their mouth enough for an intraoral mirror to be inserted, it is difficult to obtain unified intraoral images. In addition, it is impossible to correct the inappropriate intraoral images due to perspective and distortion even if photo editing software is used. For those reasons, a large-scale unified and public dental dataset need to be constructed to make the clinical application of deep learning possible with meaningful performance.

In order to challenge the generation of 1,024x1,024 with higher quality than PGGAN, another generative deep learning is needed. For example, it is demonstrated that StyleGAN or StyleGAN2 can generate better images than PGGAN [12][13]. But those new generative networks require a huge amount of machine power. It is reported that the training time of 1,024x1,024 resolution with StyleGAN is approximately one week and that of StyleGAN2 is 9 days on NVIDIA DGX-1 with 8 Tesla V100 GPUs. If we try to perform those networks on our one TITAN RTX, it is estimated that the training time takes more than 2 or 3 months. On the other hand, there is few differences between 512x512 and 1,024x1,024 resolutions in

dental diagnosis or examination because dental condition such as tooth shape, tooth color, tooth arrangement, caries, metal crowns, can be recognized enough with 512x512 resolution image. Also, the training time of 512x512 resolution with PGGAN is about 8 days on one TITAN RTX, which is more acceptable time than StyleGAN. It is considered that PGGAN has enough ability to generate 512x512 or lower resolution intraoral images with reasonable machine power and time compared with new generative methods which require a huge amount of computational resources.

4.3. Future direction and applications

There is a possibility that many kinds of images can be generated by changing the value of the latent vector. For example, image morphing can be performed by exploring the latent space [14]. In our study, morphing of intraoral images can be achieved by linearly interpolating the latent vectors that generate images of primary dentition, mixed dentition, and permanent dentition (Fig. 6). It seems that images are changing transitionally between one end and the other as if the color changes with gradation. If arbitrary images can be generated with various kinds of images, PGGAN is a useful tool for dental education or explaining materials to patients. Supplementary data which shows intraoral image generation movie can be found in the online version.

Another advantage of using generated intraoral images is its ease of use. Because the images can be generated not to include private information, researchers or educators do not need to take care about individual information and can feel free to use the generated intraoral images.

In this study, we used the latent vector as input data for the PGGAN to generate images. If the PGGAN architecture is modified and the real images can be used for input data instead of the latent vector, such as pix2pix [15] or CycleGAN [16], there is a possibility that intraoral images of the future can be generated and predicted from current intraoral images. If children could be informed that their future teeth alignment is likely to be bad, early orthodontic treatment could be recommended, which would save money and time. Also, if intraoral images of the future could be generated that show the differences between getting treatment and not getting it, dentists would be better able to recommend dental treatment to their patients. Patients could be motivated to think about their oral health and be encouraged to brush their teeth carefully at home by being shown images of a future with periodontitis or tooth loss caused by not taking care of their teeth.

Another way of applying PGGAN is to use the generated images for data augmentation of deep learning. When we perform a deep learning project, many medical images for training data are needed to achieve a good performance. It is said that a small number of training data can easily lead to overfitting in the deep learning model, so data augmentation of training images, such as translation, rotation, zoom, and contrast, is commonly used to reduce overfitting [6]. In addition to such data augmentation, it can be expected that generated realistic and varied images are used for deep learning as a method of data augmentation. In fact, it is reported that the accuracy of medical image recognition based on deep learning has been improved by adding artificial images generated by GAN to the training data [15–17]. In the future, our findings may contribute to improving the performance of deep learning related to intraoral

images by using generated images, since generated images are so realistic that pediatric dentists cannot distinguish which image is real or generated.

Conclusion

In this study, we presented intraoral images generated by PGGAN. Generated images with resolutions of 512×512 or lower were so realistic that dentists could not distinguish whether they were real or generated. This implies that generated intraoral images can be used as educational materials or for data augmentation for deep learning free from privacy restrictions.

Declarations

AUTHOR CONTRIBUTIONS

K. Kokomoto contributed to the conception, design, data acquisition, analysis, interpretation, drafting, and critical revision of the manuscript. R. Okawa, K. Nakano and K. Nozaki contributed to data analysis, interpretation, and critical revision of the manuscript. All authors gave final approval and agree to be accountable for all aspects of the work.

ACKNOWLEDGEMENTS

This research was supported by the Social Smart Dental Hospital project (S2DH, https://s2dh.org/) to get the English proof ready.

ETHICS

The Ethics Committee of Osaka University Graduate School of Dentistry approved this study (approval: R1-E29) with the need for informed consent waived, and confirmed that all methods in this study were performed in accordance with the Act on the Protection of Personal Information, and Ethical Guidelines for Medical and Health Research Involving Human Subjects.

CONFLICTS OF INTRESET

None of the contributing authors have a conflict of interest to declare.

References

[1] K.M.H.R.K.C.T.V.D.S. Wang, Radiology Education, Springer, Berlin, Heidelberg, 2012. https://doi.org/https://doi.org/10.1007/978-3-642-27600-2.

[2] I.J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative Adversarial Networks, ArXiv. (2014). https://arxiv.org/abs/1406.2661.

[3] H. Salehinejad, E. Colak, T. Dowdell, J. Barfett, S. Valaee, Synthesizing Chest X-Ray Pathology for Training Deep Convolutional Neural Networks, IEEE Trans. Med. Imaging. 38 (2019) 1197–1206. https://doi.org/10.1109/TMI.2018.2881415.

[4] D. Nie, R. Trullo, J. Lian, C. Petitjean, S. Ruan, Q. Wang, D. Shen, Medical Image Synthesis with Context-Aware Generative Adversarial Networks, in: Int. Conf. Med. Image Comput. Comput. Interv., 2017: pp. 417–425. https://doi.org/10.1007/978-3-319-66179-7_48.

[5] T. Ching, D.S. Himmelstein, B.K. Beaulieu-Jones, A.A. Kalinin, B.T. Do, G.P. Way, E. Ferrero, P.M. Agapow, M. Zietz, M.M. Hoffman, W. Xie, G.L. Rosen, B.J. Lengerich, J. Israeli, J. Lanchantin, S. Woloszynek, A.E. Carpenter, A. Shrikumar, J. Xu, E.M. Cofer, C.A. Lavender, S.C. Turaga, A.M. Alexandari, Z. Lu, D.J. Harris, D. Decaprio, Y. Qi, A. Kundaje, Y. Peng, L.K. Wiley, M.H.S. Segler, S.M. Boca, S.J. Swamidass, A. Huang, A. Gitter, C.S. Greene, Opportunities and obstacles for deep learning in biology and medicine, Journal of the Royal Society Interface, 2018. https://doi.org/10.1098/rsif.2017.0387.

[6] A. Krizhevsky, I. Sutskever, G.E. Hinton, ImageNet Classification with Deep Convolutional Neural Networks, Commun. ACM. 60 (2017) 84–90. https://doi.org/https://doi.org/10.1145/3065386.

[7] F. and others Chollet, keras, GitHub. (2015). https://keras.io (accessed April 10, 2020).

[8] T. Karras, T. Aila, S. Laine, J. Lehtinen, Progressive Growing of GANs for Improved Quality, Stability, and Variation, ArXiv. (2017). http://arxiv.org/abs/1710.10196.

[9] P.J. Burt, E.H. Adelson, The Laplacian Pyramid as a Compact Image Code BT - Fundamental Papers in Wavelet Theory, Fundam. Pap. Wavelet Theory. C (1983) 532–540.

[10] Z. Liu, P. Luo, X. Wang, X. Tang, Deep Learning Face Attributes in the Wild, ArXiv. (2014). http://arxiv.org/abs/1411.7766.

[11] I. Ahmad, Digital dental photography, Br. Dent. J. 207 (2009) 151–157. https://doi.org/10.1038/sj.bdj.2009.715.

[12] T. Karras, S. Laine, T. Aila, A Style-Based Generator Architecture for Generative Adversarial Networks, ArXiv. (2018). http://arxiv.org/abs/1812.04948.

[13] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, T. Aila, Analyzing and Improving the Image Quality of StyleGAN, ArXiv. (2019). http://arxiv.org/abs/1912.04958 (accessed May 22, 2020).

[14] A. Radford, L. Metz, S. Chintala, Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks, (2015) 1–16. http://arxiv.org/abs/1511.06434.

[15] P. Isola, J.Y. Zhu, T. Zhou, A.A. Efros, Image-to-image translation with conditional adversarial networks, Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017. 2017-Janua (2017) 5967–5976. https://doi.org/10.1109/CVPR.2017.632.

[16] J.Y. Zhu, T. Park, P. Isola, A.A. Efros, Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks, Proc. IEEE Int. Conf. Comput. Vis. 2017-Octob (2017) 2242–2251. https://doi.org/10.1109/ICCV.2017.244.

[17] H. Salehinejad, S. Valaee, T. Dowdell, E. Colak, J. Barfett, Generalization of Deep Neural Networks for Chest Pathology Classification in X-Rays Using Generative Adversarial Networks, ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc. 2018-April (2018) 990–994. https://doi.org/10.1109/ICASSP.2018.8461430.

[18] R. Togo, K. Ishihara, T. Ogawa, M. Haseyama, Anonymous Gastritis Image Generation via Adversarial Learning from Gastric X-Ray Images, Proc. - Int. Conf. Image Process. ICIP. (2018) 2082–2086. https://doi.org/10.1109/ICIP.2018.8451019.

[19] R. Togo, T. Ogawa, M. Haseyama, Synthetic Gastritis Image Generation via Loss Function-Based Conditional PGGAN, IEEE Access. 7 (2019) 87448–87457. https://doi.org/10.1109/access.2019.2925863.

Tables

Due to technical limitations, table 1-2 is only available as a download in the Supplemental Files section.

Figures

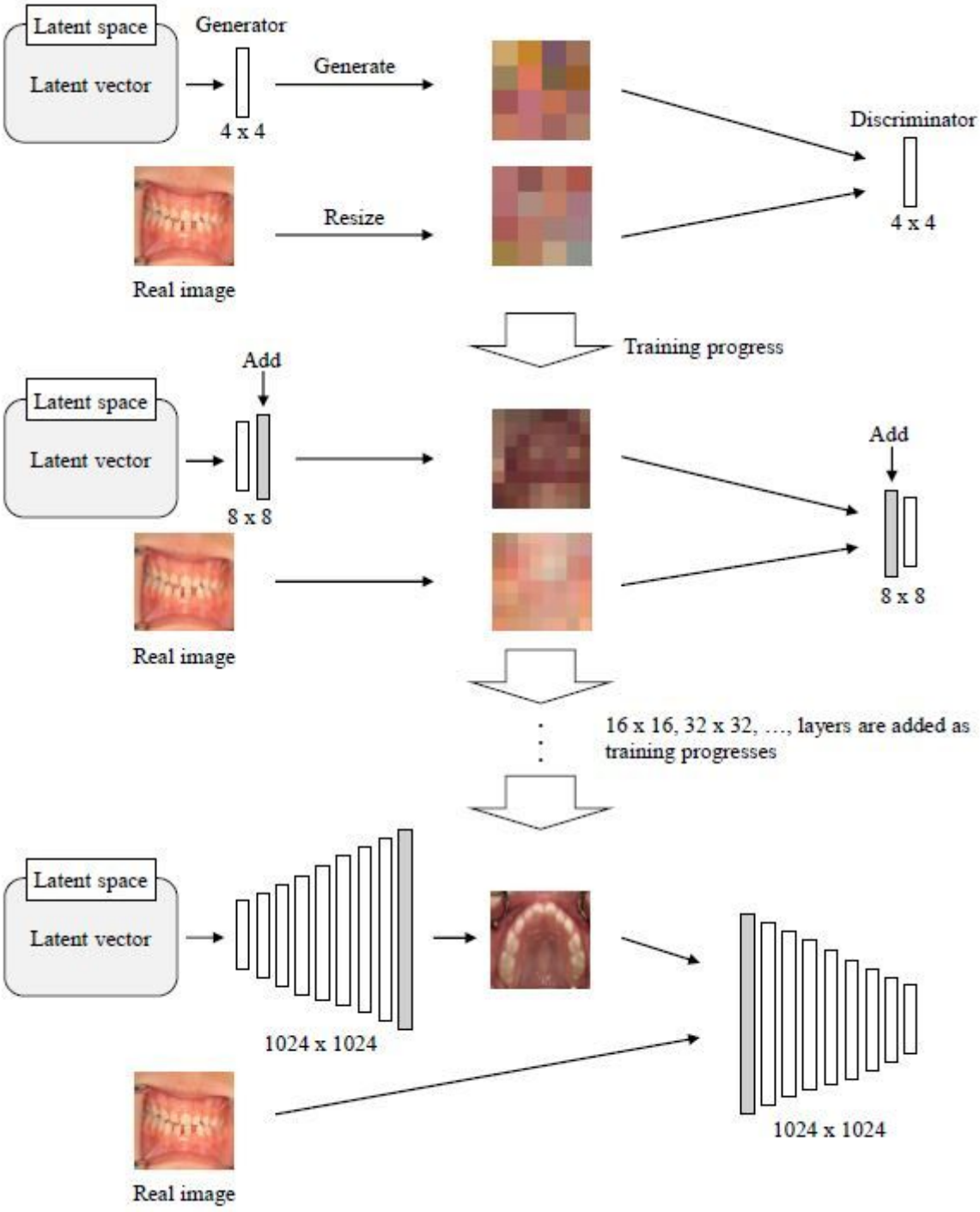

Figure 1

Architecture of progressive growing generative adversarial networks (PGGAN).

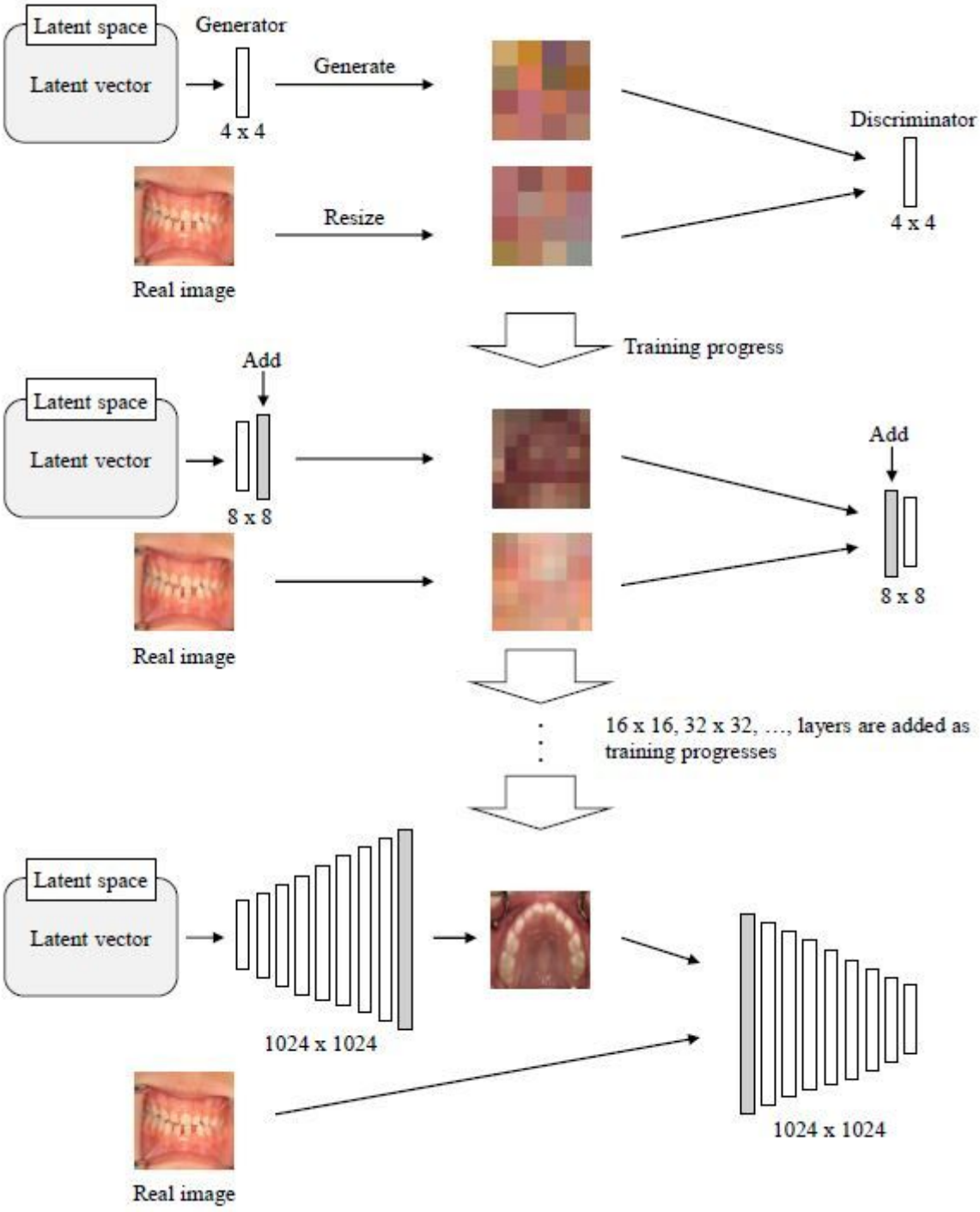

Figure 1

Architecture of progressive growing generative adversarial networks (PGGAN).

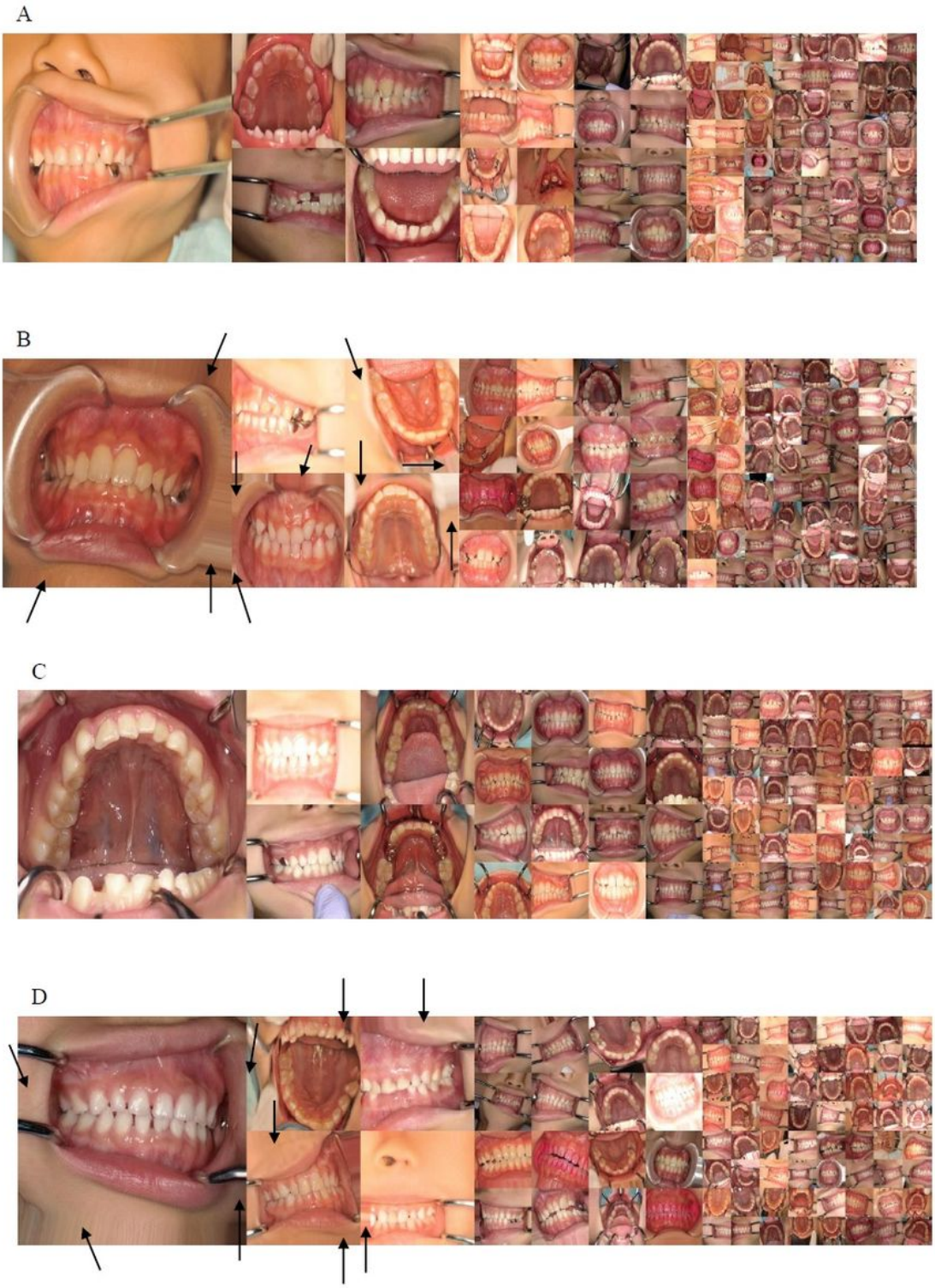

Figure 2

Examples of images used in this study with resolution of 1,024x1,024, 512x512, 256x256, and 128x128. Arrows indicate the examples where the boundaries of the training images with data augmentation are obviously filled with the closest pixel value and repeated for all the empty values, which is the effect of the fill_mode of ImageDataGenerator. (A) Real images used for training without data augmentation. (B)

Real images used for training with data augmentation. (C) Generated images trained with real images that were not augmented. (D) Generated images trained with real images that were augmented.

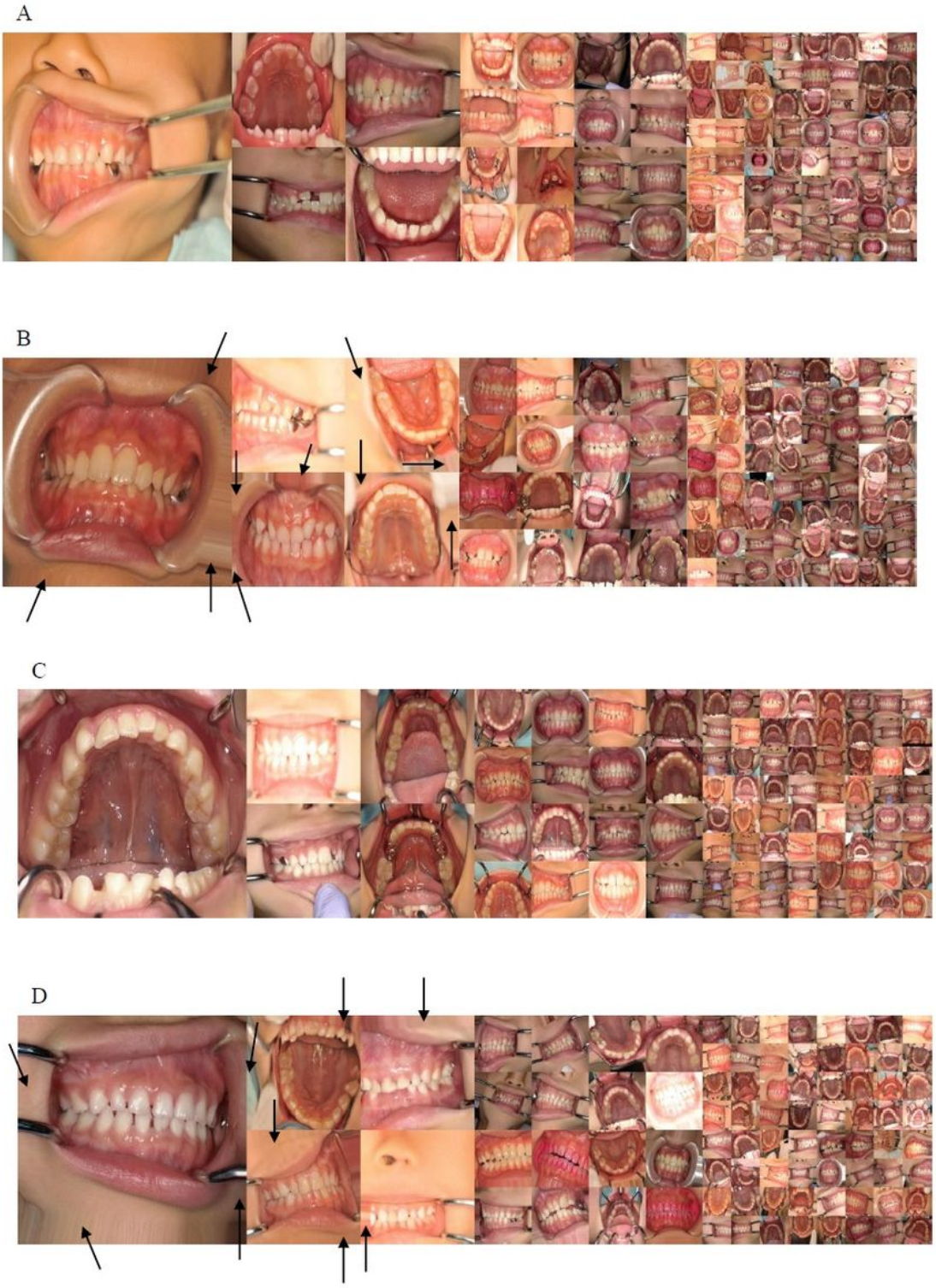

Figure 2

Examples of images used in this study with resolution of 1,024x1,024, 512x512, 256x256, and 128x128. Arrows indicate the examples where the boundaries of the training images with data augmentation are obviously filled with the closest pixel value and repeated for all the empty values, which is the effect of

the fill_mode of ImageDataGenerator. (A) Real images used for training without data augmentation. (B) Real images used for training with data augmentation. (C) Generated images trained with real images that were not augmented. (D) Generated images trained with real images that were augmented.

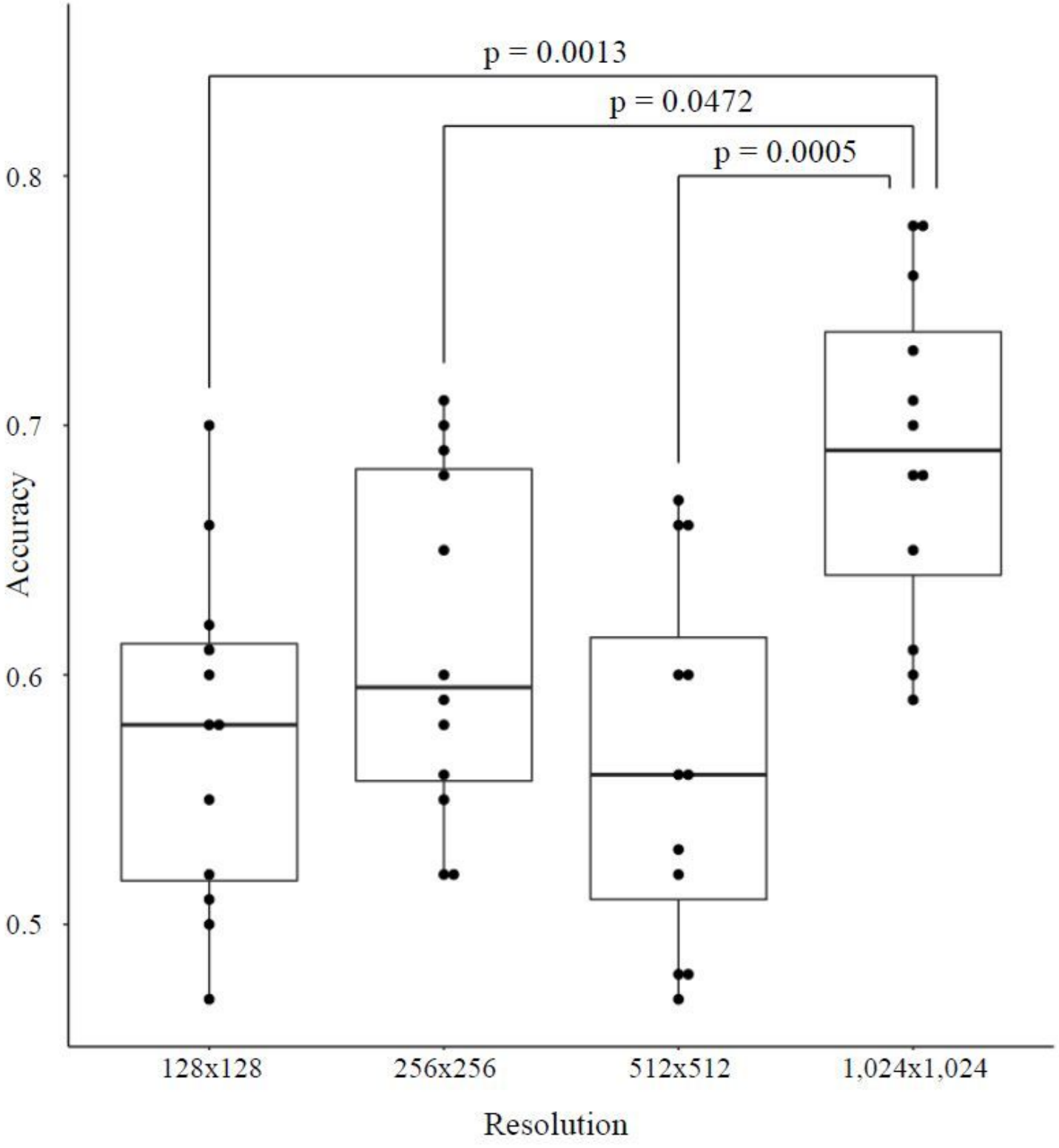

Figure 3

Boxplot of accuracy for each resolution. The black dots indicate the raw data of the twelve dentists. The accuracy of 1,024x1,024 was significantly higher (TukeyHSD test).

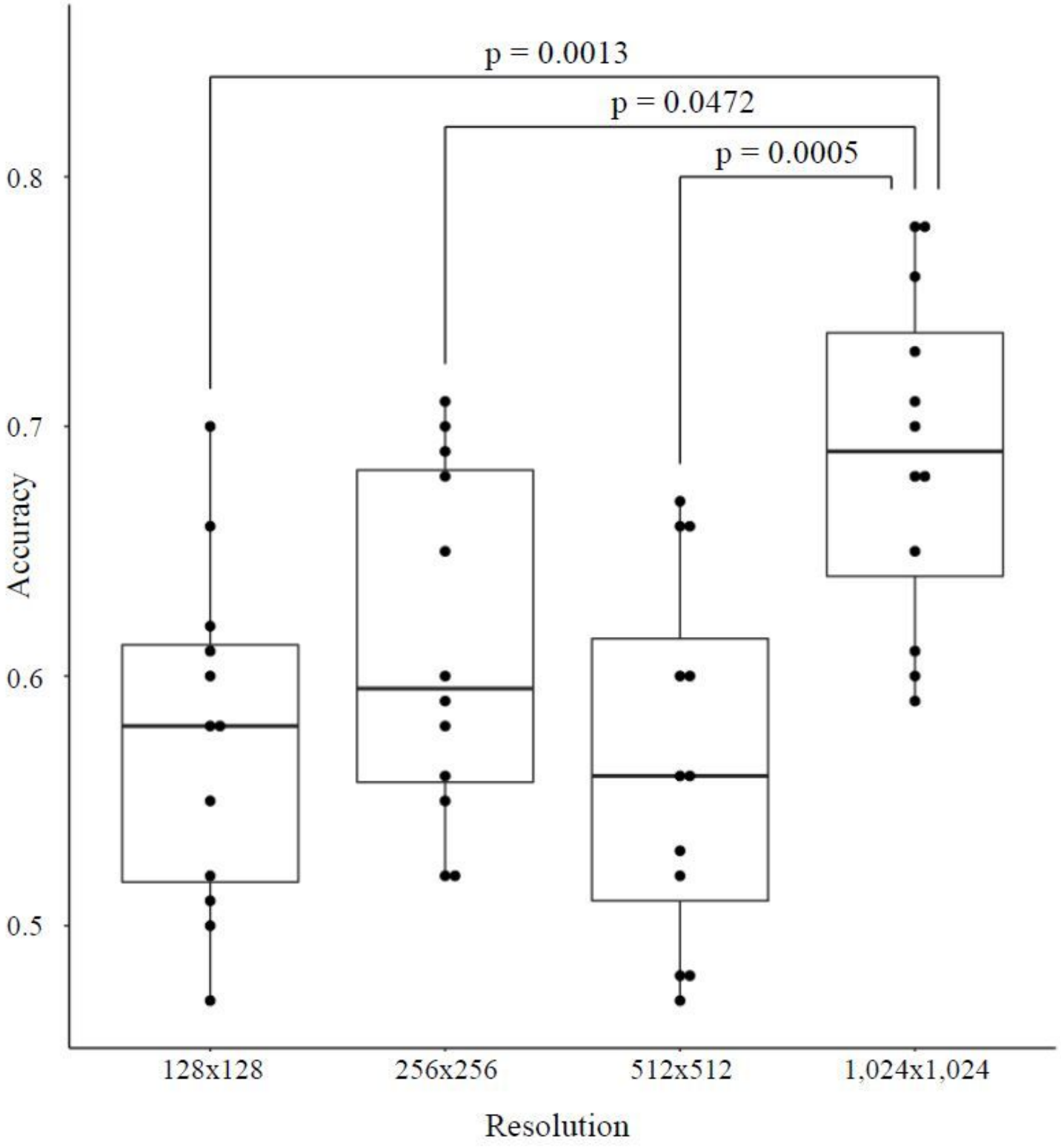

Figure 3

Boxplot of accuracy for each resolution. The black dots indicate the raw data of the twelve dentists. The accuracy of 1,024x1,024 was significantly higher (TukeyHSD test).

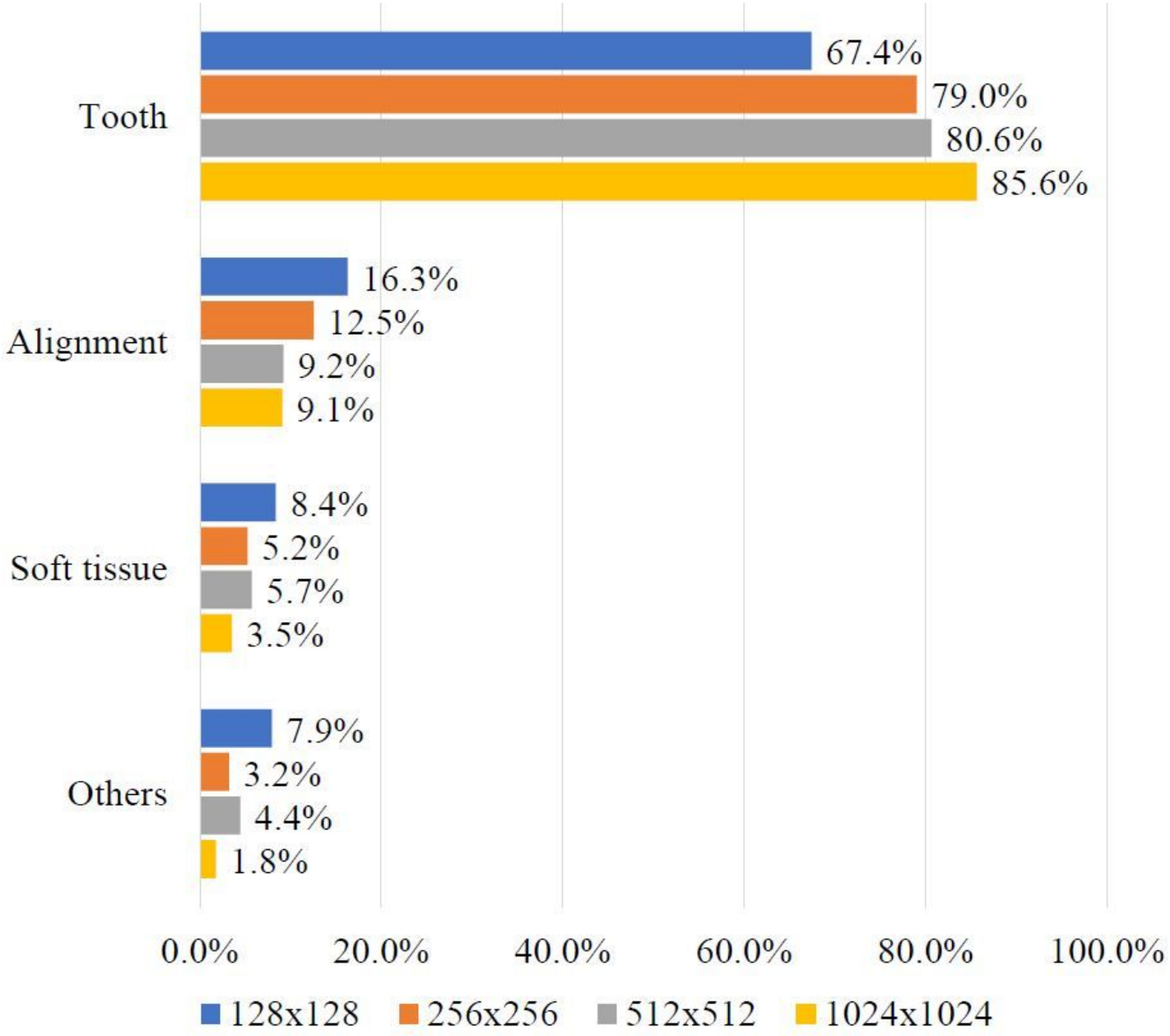

Figure 4

Ratio of each element used for criteria to assess whether the given image is real or generated. Most of dentists used tooth to discriminate images.

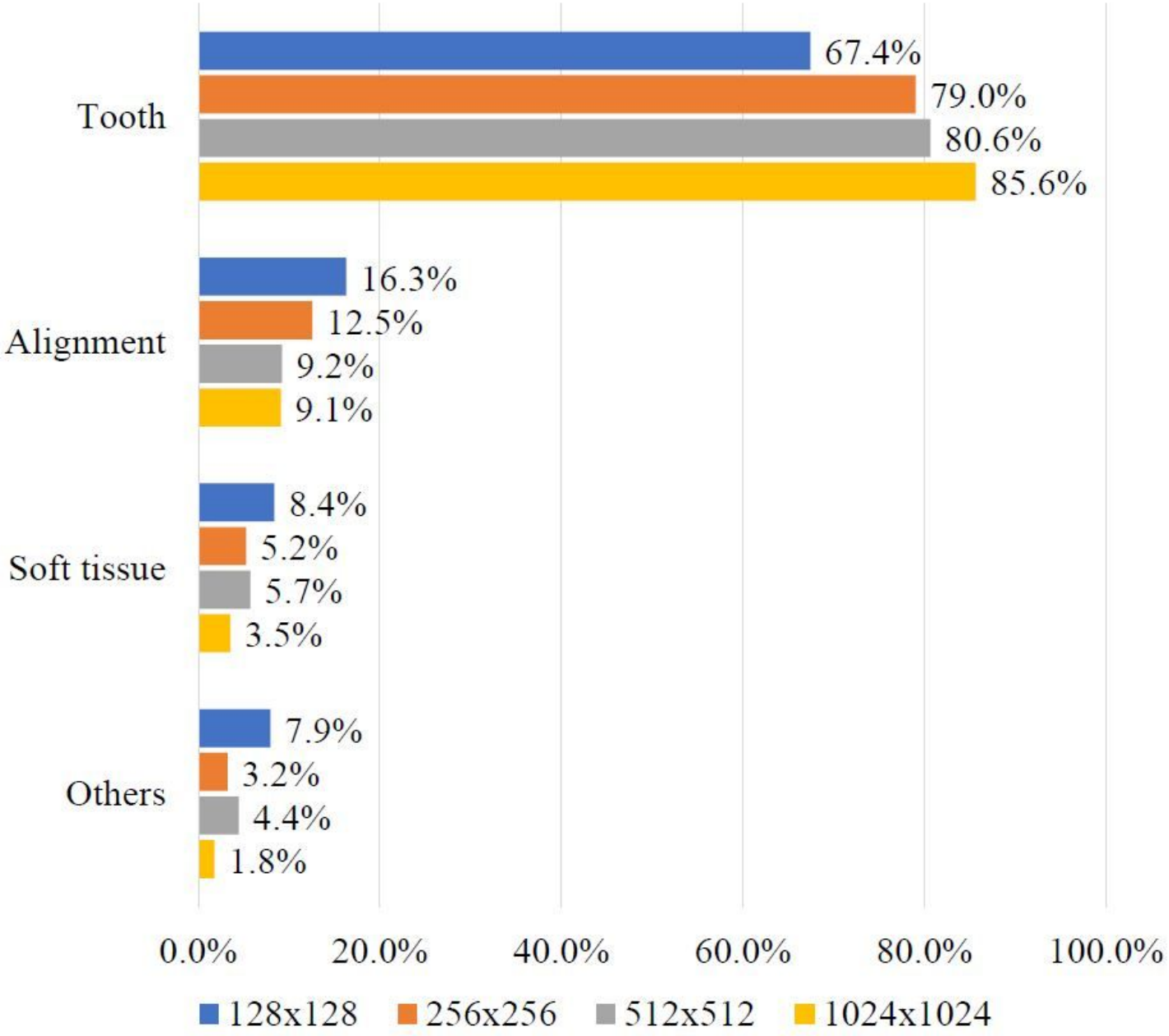

Figure 4

Ratio of each element used for criteria to assess whether the given image is real or generated. Most of dentists used tooth to discriminate images.

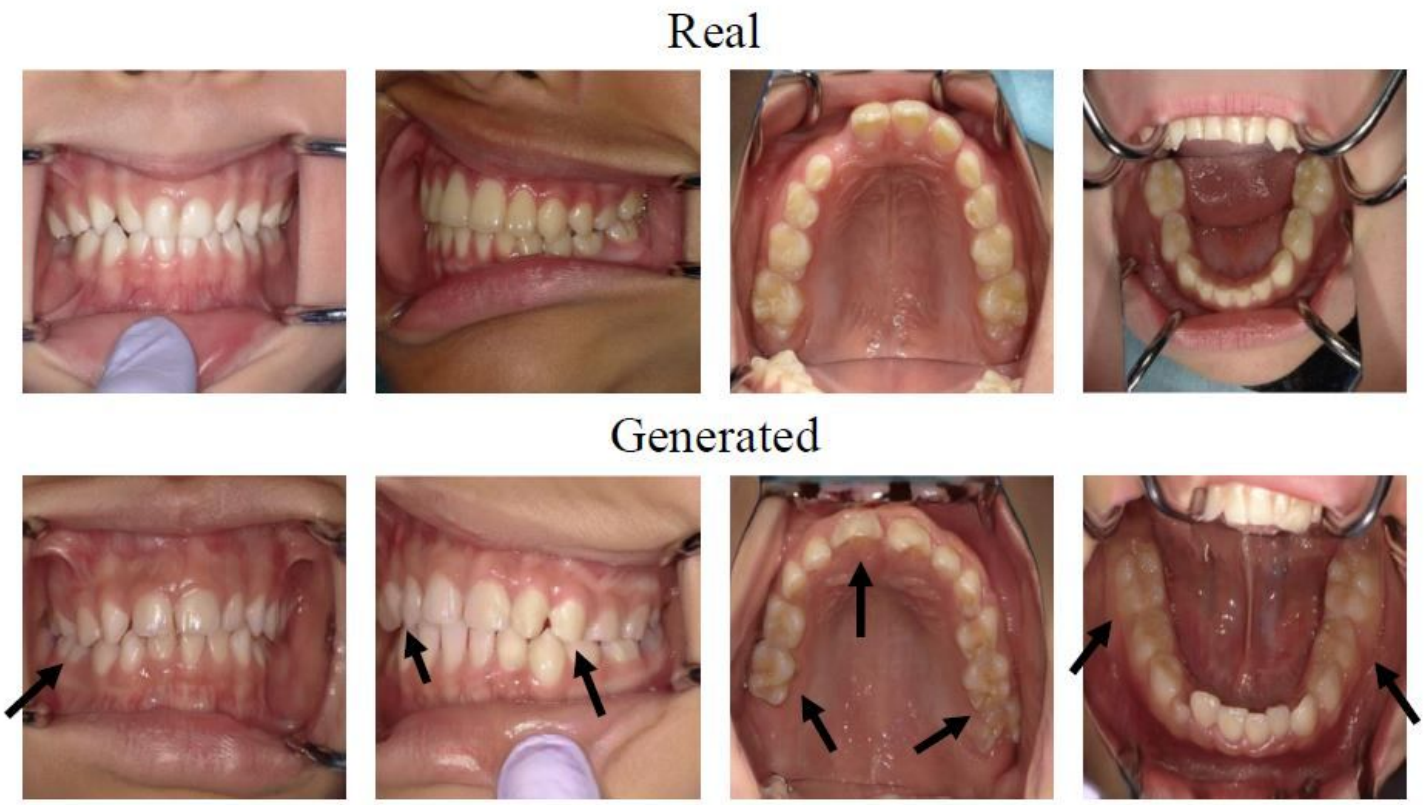

Figure 5

Comparison between real and generated intraoral images. Arrows indicate points of strange teeth. Tooth alignment and soft tissues in both images are similar.

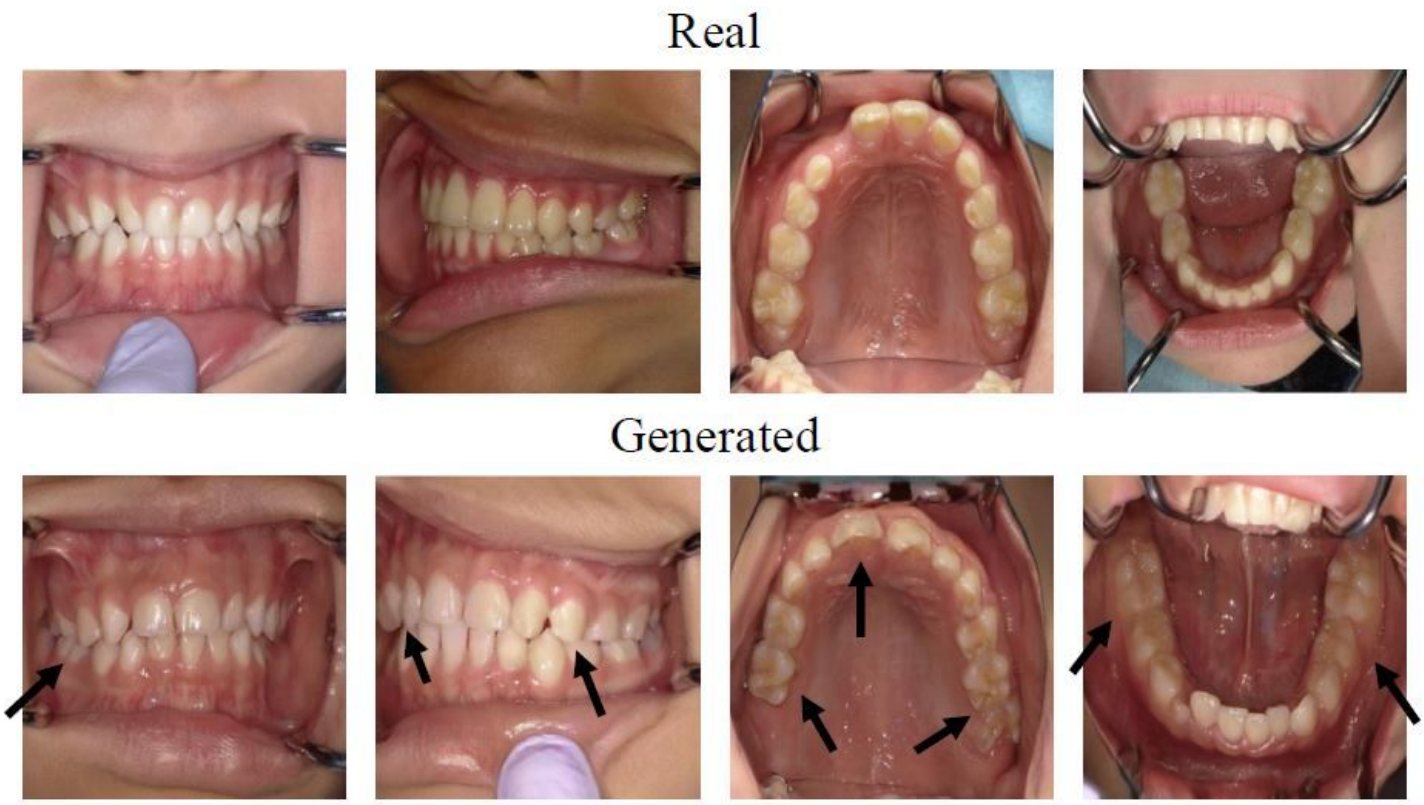

Figure 5

Comparison between real and generated intraoral images. Arrows indicate points of strange teeth. Tooth alignment and soft tissues in both images are similar.

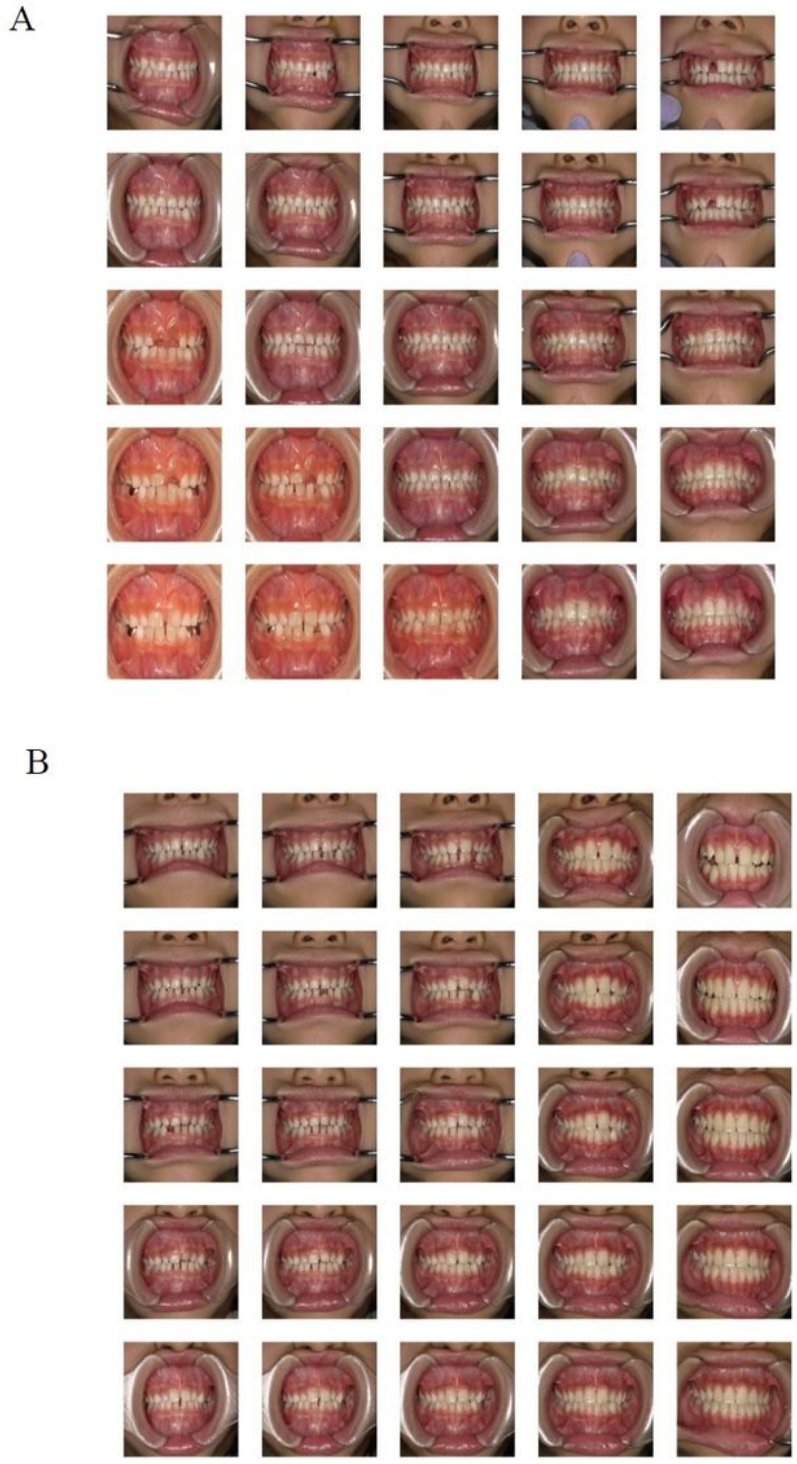

Figure 6

An example of exploring latent space. The images generated by PGGAN are gradually chang-ing between one end and the other, from primary dentition to permanent dentition.

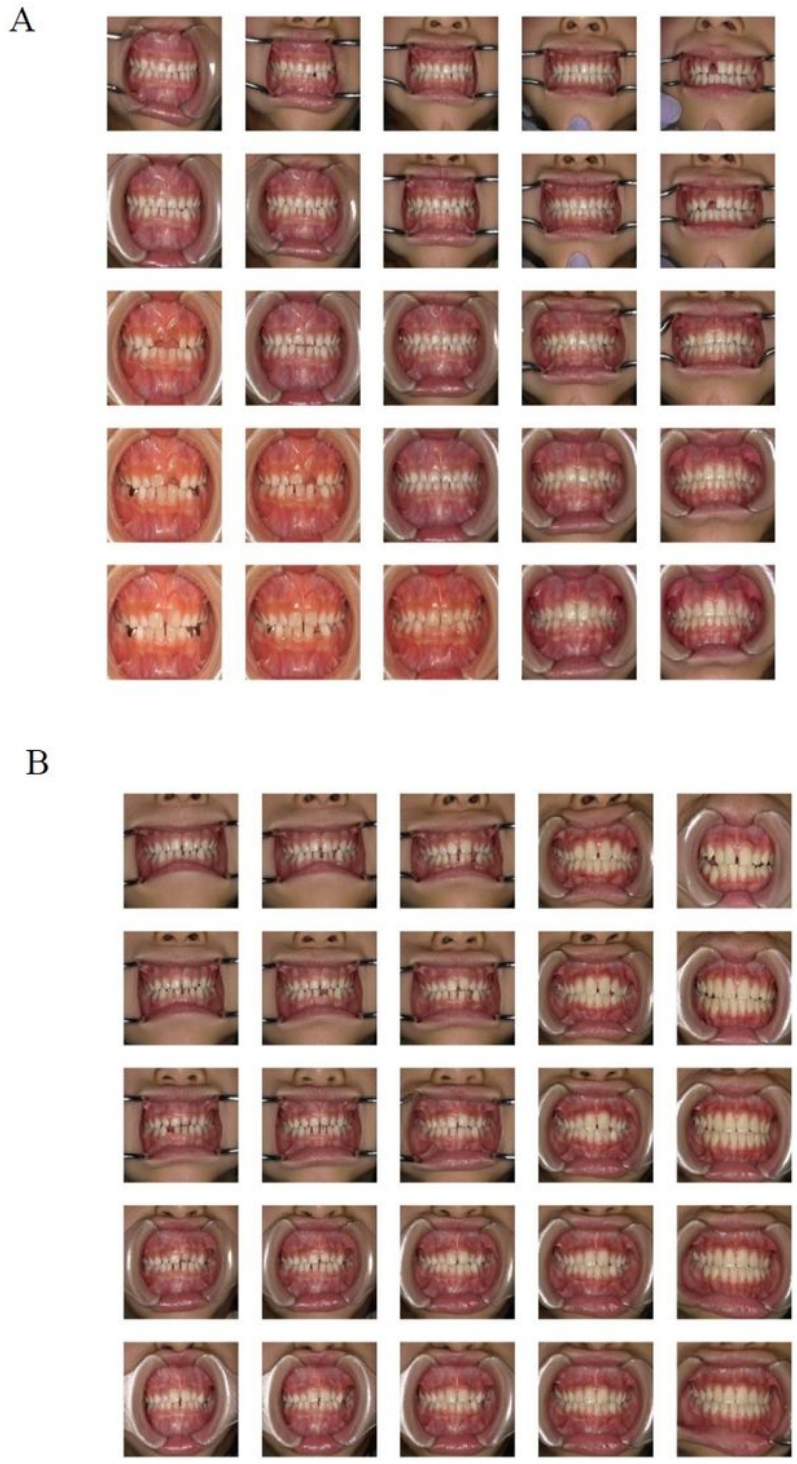

Figure 6

An example of exploring latent space. The images generated by PGGAN are gradually chang-ing between one end and the other, from primary dentition to permanent dentition.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- TableI.pdf
- TableI.pdf
- TableII.pdf
- TableII.pdf
- supplementarycTrim.mp4
- supplementarycTrim.mp4