# A Semantic-guided and Knowledge-based Generative Framework for Orthodontic Visual Outcome Preview

Yizhou Chen[1] and Xiaojun Chen[1,2](✉)

[1] Institute of Biomedical Manufacturing and Life Quality Engineering, School of Mechanical Engineering, Shanghai Jiao Tong University, Shanghai, China
[2] Institute of Medical Robotics, Shanghai Jiao Tong University, Shanghai, China
xiaojunchen@sjtu.edu.cn

**Abstract.** Orthodontic treatment typically lasts for two years, and its outcome cannot be predicted intuitively in advance. In this paper, we propose a semantic-guided and knowledge-based generative framework to predict the visual outcome of orthodontic treatment from a single frontal photo. The framework involves four steps. Firstly, we perform tooth semantic segmentation and mouth cavity segmentation and extract category-specific teeth contours from frontal images. Secondly, we deform the established tooth-row templates to match the projected contours with the detected ones to reconstruct 3D teeth models. Thirdly, we apply a teeth alignment algorithm to simulate the orthodontic treatment. Finally, we train a semantic-guided generative adversarial network to predict the visual outcome of teeth alignment. Quantitative tests are conducted to evaluate the proposed framework, and the results are as follows: the tooth semantic segmentation model achieves a mean intersection of union of 0.834 for the anterior teeth, the average symmetric surface distance error of our 3D teeth reconstruction method is 0.626 mm on the test cases, and the image generation model has an average Fréchet inception distance of 6.847 over all the test images. These evaluation results demonstrate the practicality of our framework in orthodontics.

**Keywords:** Semantic segmentation · 3D teeth reconstruction · Generative adversarial network (GAN).

## 1 Introduction

Orthodontic treatment aims to correct misaligned teeth and restore normal occlusion. Patients are required to wear dental braces or clear aligners for a duration of one to three years, reported by [22], with only a vague expectation of the treatment result. Therefore, a generative framework is needed to enable patients to preview treatment outcomes and assist those considering orthodontic treatment in making decisions. Such framework may involve multiple research fields, such as semantic segmentation, 3D reconstruction, and image generation.

Deep learning methods have achieved great success in image-related tasks. In the field of tooth semantic segmentation, there exist plenty of studies targeting on different data modalities, such as dental mesh scanning [32], point cloud [27, 31], cone beam CT image [5, 6], panoramic dental X-ray image [29], and 2D natural image [33]. Regarding image generation, the family of generative adversarial networks (GAN) [8, 14, 16, 24, 13] and the emerging diffusion models [11, 23, 26] can generate diverse high-fidelity images. Although the diffusion models can overcome the mode collapse problem and and excel in image diversity compared to GAN [7], their repeated reverse process at inference stage prolongs the execution time excessively, limiting their application in real-time situations.

When it comes to 3D teeth reconstruction, both template-based and deep-learning-based frameworks offer unique benefits. Wu et al. employed a template-based approach by adapting their pre-designed teeth template to match teeth contours extracted from a set of images [30]. Similarly, Wirtz et al. proposed an optimization-based pipeline that uses five intra-oral photos to restore the 3D arrangement of teeth [28]. Liang et al. restored 3D teeth using convolution neural networks (CNN) from a single panoramic radiograph [19]. However, while deep CNNs have a strong generalization ability compared to template-based methods, it often struggles to precisely and reasonably restore occluded objects.

Predicting the smiling portrait after orthodontic treatment has recently gained much attention. Yang et al. developed three deep neural networks to extract teeth contours from smiling images, arrange 3D teeth models, and generate images of post-treatment teeth arrangement, respectively [20]. However, their framework requires a single frontal smiling image and the corresponding unaligned 3D teeth model from dental scanning, which may be difficult for general users to obtain. In contrast, Chen et al. proposed a StyleGAN generator with a latent space editing method that utilizes GAN inversion to discover the optimal aligned teeth appearance from a single image [3]. Although their method takes only a frontal image as input and manipulates the teeth structure and appearance implicitly in image space, it may overestimate the treatment effect and result in inaccurate visual outcomes.

In this study, we propose an explainable generative framework, which is semantic-guided and knowledge-based, to predict teeth alignment after orthodontic treatment. Previous works have either required 3D teeth model as additional input and predicted its alignment using neural networks [20], or directly utilized an end-to-end StyleGAN to predict the final orthodontic outcome [3]. In contrast, our approach requires only a single frontal image as input, restores the 3D teeth model through a template-based algorithm, and explicitly incorporates orthodontists' experience, resulting in a more observable and explainable process. Our contributions are therefore three-fold: 1) we introduce a region-boundary feature fusion module to enhance the 2D tooth semantic segmentation results; 2) we employ statistical priors and reconstruct 3D teeth models from teeth semantic boundaries extracted in a single frontal image; 3) by incorporating an orthodontic simulation algorithm and a pSpGAN style encoder [24], we can yield more realistic and explainable visualization of the post-treatment teeth appearance.

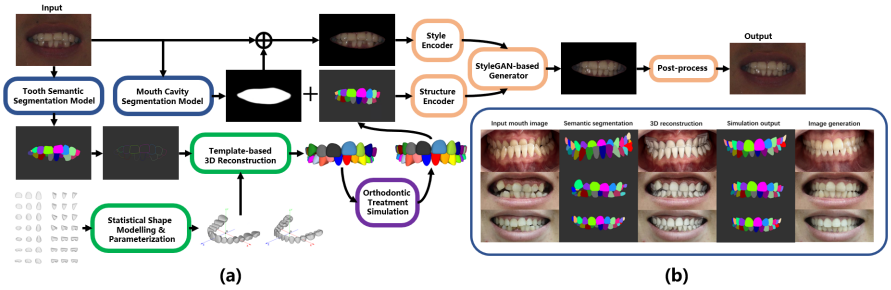


**Fig. 1.** An overview of (a) the proposed generative framework for orthodontic treatment outcome prediction, and (b) the framework's outputs at each stage with images from left to right: input mouth image, tooth semantic segmentation map, input image overlaid with semi-transparent 3D reconstruction teeth mesh, projection of orthodontic treatment simulation output, and orthodontic visual outcome prediction.

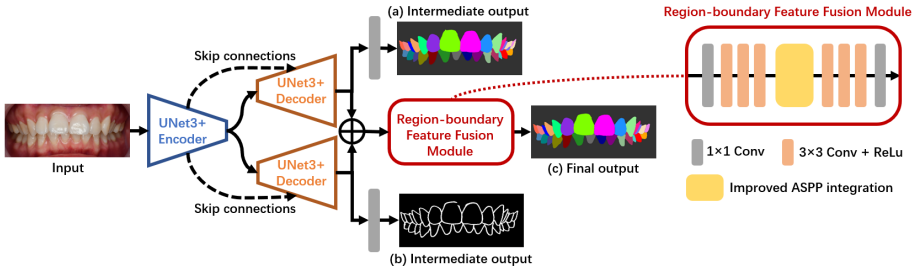


**Fig. 2.** The proposed tooth semantic segmentation model with (a) coarse region segmentation, (b) auxiliary boundary segmentation, and (c) final region segmentation generated by the region-boundary feature fusion module.

## 2 Method

The proposed generative framework (Fig.1) consists of four parts: semantic segmentation in frontal images, template-based 3D teeth reconstruction, orthodontic treatment simulation, and semantic-guided image generation of mouth cavity.

### 2.1 Semantic Segmentation in Frontal Images

The tooth areas and mouth cavity are our region of interest for semantic segmentation in each frontal image. As a preprocessing step, the rectangle mouth area is firstly extracted from frontal images by the minimal bounding box that encloses the facial key points around the mouth, which are detected by dlib toolbox [18]. We then use two separate segmentation models to handle these mouth images, considering one-hot encoding and the integrity of mouth cavity mask. A standard U-Net [25] is trained with soft dice loss [21] to predict mouth cavity.

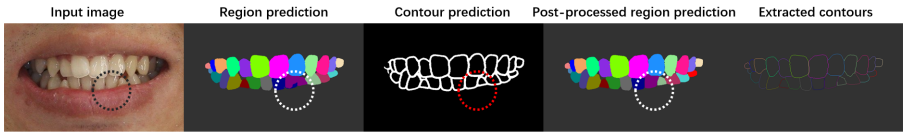


**Fig. 3.** Illustration of the effect of the segmentation post-process algorithm.

The tooth semantic segmentation model (Fig.2) is a dual-branch U-Net3+ based network that predicts tooth regions and contours simultaneously. We employ a standard U-Net3+ [12] encoder and two identical U-Net3+ decoders for tooth region and contour segmentation. Such inter-related multi-task learning can enhance the performance of each task and mitigate overfitting. The teeth are manually labeled using FDI World Dental Federation notation, resulting in a total of 33 classes, including background.

To generate a more precise tooth segmentation map, we introduce the region-boundary feature fusion module, which merges the tooth region and boundary information, i.e., the last hidden feature maps of the two decoders. The module is constructed as a stack of convolutional layers, which incorporates an improved atrous spatial pyramid pooling (ASPP) module [4]. This ASPP module employs atrous separable convolution and global pooling to capture long-range information. The integration of ASPP has a dilation rate of 6, and the number of filters in each convolutional layer, except the output layer, is set to 64. These three outputs are supervised by soft dice loss [21] during training.

Some post-process techniques are added to filter segmented regions and obtain smoother tooth contours. The predicted binary contours are dilated and used to divide the semantic segmentation map into multiple connected regions. The pixels in each region are classified by its dominant tooth label. Background components are ignored and small ones are removed. Once two regions have duplicate labels, a drop-or-relabel strategy is performed on the region away from the center of central incisors. The connected regions are processed sequentially from central incisors to the third molars. Fig.3 shows the changes of tooth region prediction before and after the post process.

### 2.2 Template-based 3D Teeth Reconstruction

To achieve 3D tooth reconstruction from a single frontal image, we deform the parametric templates of the upper and lower tooth rows to match the projected contours with the extracted teeth contours in semantic segmentation.

The parametric tooth-row template is a statistical model that characterizes the shape, scale, and pose of each tooth in a tooth row. To describe the shape of each tooth, we construct a group of morphable shape models [1]. We model the pose (orientation and position) of each tooth as a multivariate normal distribution as Wu et al. [30] did. Additionally, we suppose that the scales of all teeth follows a multivariate normal distribution. The mean shape and average of the tooth scales and poses are used to generate a standard tooth-row template.

The optimization-based 3D teeth reconstruction following [30] is an iterative process alternating between searching for point correspondences between the projected and segmented teeth contours and updating the parameters of the tooth-row templates. The parameters that require estimation are the camera parameters, the relative pose between the upper and lower tooth rows, and the scales, poses, and shape parameters of each tooth.

The point correspondences are established by Equ.(1) considering the semantic information in teeth contours, where $c_i^\tau$ and $n_i^\tau$ are the position and normal of the detected contour point $i$ of tooth $\tau$ in image space, $\hat{c_j}^\tau$ and $\hat{n_j}^\tau$ are those of the projected contour point $j$, $<\cdot,\cdot>$ denotes inner product, and $\sigma_{angle} = 0.3$ is a fine-tuned hyper parameter in [30].

$$\hat{c_i}^\tau = \underset{\hat{c_j}^\tau}{\arg\min} ||c_i^\tau - \hat{c_j}^\tau||_2^2 \cdot \exp\left[-\left(\frac{<n_i^\tau, \hat{n_j}^\tau>}{\sigma_{angle}}\right)^2\right] \tag{1}$$

We use $\mathcal{L}$ as the objective function to minimize, expressed in Equ.(2), which comprises an image-space contour loss [30] and a regularization term $\mathcal{L}_{prior}$ described by Mahalanobis distance in probability space, where $N$ is the number of detected contour points, $\lambda_n = 50$ and $\lambda_p = 25$ are the fine-tuned weights.

$$\mathcal{L} = \frac{1}{N}\sum_\tau \sum_i \left(||c_i^\tau - \hat{c_i}^\tau||_2^2 + \lambda_n <c_i^\tau - \hat{c_i}^\tau, \hat{n_i}^\tau>^2\right) + \lambda_p \, \mathcal{L}_{prior} \tag{2}$$

The regularization term $\mathcal{L}_{prior}$ in Equ.(3) is the negative log likelihood of the distributions of the vector of tooth scales, denoted by $\boldsymbol{s}$, the pose vector of tooth $\tau$, denoted by $\boldsymbol{p}^\tau$, and the shape vector of tooth $\tau$, denoted by $\boldsymbol{b}^\tau$. The covariance matrices $\Sigma_s$ and $\Sigma_p^\tau$ are obtained in building tooth-row templates.

$$\mathcal{L}_{prior} = (\boldsymbol{s} - \overline{\boldsymbol{s}})^T \Sigma_s^{-1} (\boldsymbol{s} - \overline{\boldsymbol{s}}) + \sum_\tau \left[(\boldsymbol{p}^\tau - \overline{\boldsymbol{p}}^\tau)^T \Sigma_p^{\tau\,-1} (\boldsymbol{p}^\tau - \overline{\boldsymbol{p}}^\tau) + ||\boldsymbol{b}^\tau||_2^2\right] \tag{3}$$

During optimization, we first optimize the camera parameters and the relative pose of tooth rows for 10 iterations and optimize all parameters for 20 iterations. Afterward, we use Poisson surface reconstruction [17] to transform the surface point clouds into 3D meshes.

### 2.3 Orthodontic Treatment Simulation

We implement a naive teeth alignment algorithm to mimic orthodontic treatment. The symmetrical beta function (Equ.(4)) is used to approximate the dental arch curve of a tooth row [2]. Its parameters $W$ and $D$ can be estimated through linear regression by fitting the positions of tooth landmarks [2].

$$\beta(x; D, W) = 3.0314 * D * \left[\frac{1}{2} + \frac{x}{W}\right]^{0.8} \left[\frac{1}{2} - \frac{x}{W}\right]^{0.8} \tag{4}$$

We assume that the established dental arch curves are parallel to the occlusal plane. Each reconstructed tooth is translated towards its expected position in the

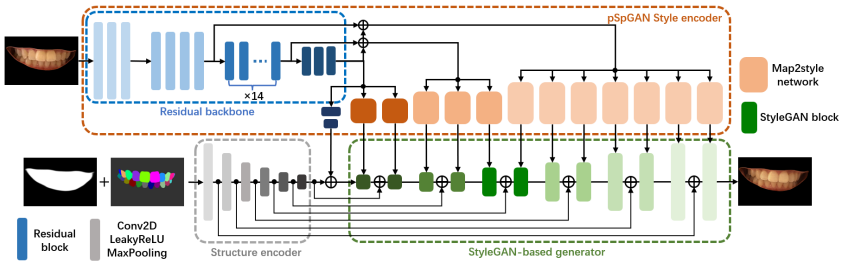


**Fig. 4.** Architecture of the semantic-guided image generation model. The multi-level style feature maps are extracted from the residual backbone and encoded into twelve 512-dim style vectors through the map2style networks [24], structural information are compressed and skip connected through the structure encoder, and structure and style features are entangled in the StyleGAN-based generator with weight modulation [16].

dental arch and rotated to its standard orientation while preserving its shape. The teeth gaps are then reduced along the dental arch curve with collision detection performed. The relative pose between tooth rows is re-calculated to achieve a normal occlusion. Finally, the aligned 3D teeth models are projected with the same camera parameters to generate the semantic image output of the simulation.

### 2.4 Semantic-Guided Image Generation

The idea behind the semantic-guided image generation model is to decompose an image into an orthogonal representation of its style and structure. By manipulating the structural or style information, we can control the characteristics of the generated image. Improving upon [20], we replace the naive style encoder in [20] with the PixelStylePixel style encoder [24] to capture the multi-level style features and use semantic teeth image instead of teeth contours as input to better guide the generation process. The architecture is illustrated in Fig.4. At training stage, the model learns the style and structural encoding of teeth images and attempts to restore the original image. Gradient penalty [9] and path length regularization [16] are applied to stabilize the training process. We use the same loss function as [20] did and take a standard UNet encoder connected with dense layers as the discriminator. At inference stage, the semantic teeth image output from orthodontic simulation is used to control the generated teeth structure. To remove the boundary artifacts, we dilate the input mouth cavity map and use Gaussian filtering to smooth the sharp edges after image patching.

## 3 Experiments and Results

### 3.1 Dataset and Implementation Details

We collected 225 digital dental scans with labelled teeth and their intra-oral photos, as well as 5610 frontal intra-oral images, of which 3300 were labelled,

**Table 1.** Segmentation accuracy on the test data measured by mean intersection of union for different groups of tooth labels and for different network architectures where DB denotes dual-branch architecture and RBFF denotes region-boundary feature fusion. Group A has 32 tooth classes, group B has 28 classes with the third molars excluded, and group C has 24 classes with the second and third molars excluded.

| Settings | UNet | | | UNet3+ | | |
|---|---|---|---|---|---|---|
| | Baseline | DB | RBFF+DB | Baseline | DB | RBFF+DB |
| Group A | 0.679 | 0.697 | 0.708 | 0.686 | 0.699 | 0.730 |
| Group B | 0.764 | 0.774 | 0.789 | 0.766 | 0.780 | 0.803 |
| Group C | 0.800 | 0.809 | 0.820 | 0.800 | 0.816 | 0.834 |

**Table 2.** Teeth reconstruction error (avg.± std.) on all the teeth of the 95 test cases (ASSD: average symmetric surface distance, HD: Hausdorff distance, CD: Chamfer distance, DSC: Dice similarity coefficient).

| Methods | ASSD(mm)↓ | HD(mm)↓ | CD($mm^2$)↓ | DSC↑ |
|---|---|---|---|---|
| Wirtz et al. [28] | $0.848_{\pm 0.379}$[28] | $2.627_{\pm 0.915}$[28] | — | $0.659_{\pm 0.140}$[28] |
| Nearest retrieval | $0.802_{\pm 0.355}$ | $2.213_{\pm 0.891}$ | $2.140_{\pm 2.219}$ | $0.653_{\pm 0.158}$ |
| Ours | $0.626_{\pm 0.265}$ | $1.776_{\pm 0.723}$ | $1.272_{\pm 1.364}$ | $0.732_{\pm 0.125}$ |

and 4330 smiling images, of which 2000 were labelled, from our partner hospitals. The digital dental scans were divided into two groups, 130 scans for building morphable shape models and tooth-row templates and the remaining 95 scans for 3D teeth reconstruction evaluation. The labelled 3300 intra-oral images and 2000 smiling images were randomly split into training (90%) and labelled test (10%) datasets. The segmentation accuracy was computed on the labelled test data, and the synthetic image quality was evaluated on the unlabelled test data.

All the models were trained and evaluated on an NVIDIA GeForce RTX 3090 GPU. We trained the segmentation models for 100 epochs with a batch size of 4, and trained the image generation models for 300 epochs with a batch size of 8. The input and output size of the image segmentation and generation models are $256 \times 256$. The training was started from scratch and the learning rate was set to $10^{-4}$. We saved the models with the minimal loss on the labelled test data. At the inference stage, our method takes approximately 15 seconds to run a single case on average, with the 3D reconstruction stage accounting for the majority of the execution time, tests performed solely on an Intel 12700H CPU.

### 3.2 Evaluation

**Ablation study of Tooth Segmentation Model** We conduct an ablation study to explore the improvement of segmentation accuracy brought by the dual-branch network architecture and the region-boundary feature fusion module. Segmentation accuracy is measured by the mean intersection over union (mIoU) metric for different groups of tooth labels. The results listed in Table 1 show that the proposed region-boundary feature fusion module assisted with dual-branch architecture can further enhance the segmentation accuracy for UNet and its

**Table 3.** Average Fréchet inception distance of different generators on the test data.

| Model | TSynNet | Contour-guided pSpGAN | Semantic-guided pSpGAN |
|---|---|---|---|
| Test smiling images | 11.343 | 7.292 | 6.501 |
| All test images | 20.133 | 7.832 | 6.847 |

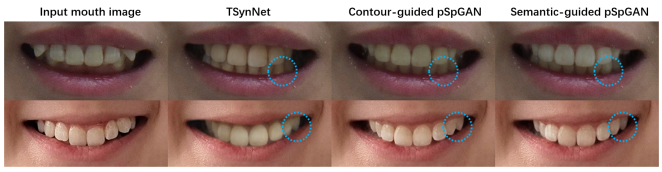


**Fig. 5.** Comparison of the orthodontic treatment outcome predictions generated by different models: TSynNet[20], contour-guided pSpGAN, and semantic-guided pSpGAN.

**Fig. 6.** Teeth alignment predictions of the proposed generative framework on frontal images in the Flickr-Faces-HQ data set.

variant. Our tooth segmentation model can predict quite accurately the region of the frontal teeth with a mIoU of 0.834.

**Accuracy of 3D Teeth Reconstruction** We reconstruct the 3D teeth models of the 95 test cases from their intra-oral photos. The restored teeth models are aligned with their ground truth by global similarity registration. We compare the reconstruction error using different metrics, shown in Table 2, with the method of [28] that reconstructs teeth models from five intra-oral photos and the nearest retrieval that selects the most similar teeth mesh in the 135 teeth meshes for building tooth-row templates. The results show that our teeth reconstruction method significantly outperforms the method of [28] and nearest retrieval.

**Image Generation Quality** We use Fréchet inception distance (FID) [10] to evaluate the quality of images generated different generators on the unlabelled test data, results listed in Table 3. The multi-level style features captured by pSpGAN improve greatly the image quality from the quantitative comparison (Table 3) and the visual perception (Fig.5). Our semantic-guided pSpGAN that takes semantic teeth image as input can further increase the constrast of different

teeth and yield sharper boundaries. We test our framework on some images in Flickr-Faces-HQ dataset [15] to visualize virtual teeth alignment, shown in Fig.6.

## 4 Conclusion

In conclusion, we develop a semantic-guided generative framework to predict the orthodontic treatment visual outcome. It comprises tooth semantic segmentation, template-based 3D teeth reconstruction, orthodontic treatment simulation, and semantic-guided mouth cavity generation. The results of quantitative tests show that the proposed framework has a potential for orthodontic application.

## Acknowledgement

This work was supported by grants from the National Natural Science Foundation of China (81971709; M-0019; 82011530141), the Foundation of Science and Technology Commission of Shanghai Municipality (20490740700; 22Y11911700), Shanghai Pudong Science and Technology Development Fund (PKX2021-R04), Shanghai Jiao Tong University Foundation on Medical and Technological Joint Science Research (YG2021ZD21; YG2021QN72; YG2022QN056; YG2023ZD19; YG2023ZD15)

## References

1. Blanz, V., Vetter, T.: A morphable model for the synthesis of 3d faces. In: Proceedings of the 26th annual conference on Computer graphics and interactive techniques. pp. 187–194 (1999)
2. Braun, S., Hnat, W.P., Fender, D.E., Legan, H.L.: The form of the human dental arch. The Angle Orthodontist **68**(1), 29–36 (1998)
3. Chen, B., Fu, H., Zhou, K., Zheng, Y.: Orthoaligner: Image-based teeth alignment prediction via latent style manipulation. IEEE Transactions on Visualization and Computer Graphics (2022)
4. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European conference on computer vision (ECCV). pp. 801–818 (2018)
5. Chen, Y., Du, H., Yun, Z., Yang, S., Dai, Z., Zhong, L., Feng, Q., Yang, W.: Automatic segmentation of individual tooth in dental cbct images from tooth surface map by a multi-task fcn. IEEE Access **8**, 97296–97309 (2020)
6. Chung, M., Lee, M., Hong, J., Park, S., Lee, J., Lee, J., Yang, I.H., Lee, J., Shin, Y.G.: Pose-aware instance segmentation framework from cone beam ct images for tooth segmentation. Computers in Biology and Medicine **120**, 103720 (2020)
7. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in Neural Information Processing Systems **34**, 8780–8794 (2021)
8. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. Communications of the ACM **63**(11), 139–144 (2020)

9. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.C.: Improved training of wasserstein gans. Advances in neural information processing systems **30** (2017)
10. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017)
11. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems **33**, 6840–6851 (2020)
12. Huang, H., Lin, L., Tong, R., Hu, H., Zhang, Q., Iwamoto, Y., Han, X., Chen, Y.W., Wu, J.: Unet 3+: A full-scale connected unet for medical image segmentation. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1055–1059 (2020)
13. Karras, T., Aittala, M., Laine, S., Härkönen, E., Hellsten, J., Lehtinen, J., Aila, T.: Alias-free generative adversarial networks. Advances in Neural Information Processing Systems **34**, 852–863 (2021)
14. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4401–4410 (2019)
15. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4401–4410 (2019)
16. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8110–8119 (2020)
17. Kazhdan, M., Bolitho, M., Hoppe, H.: Poisson surface reconstruction. In: Proceedings of the fourth Eurographics symposium on Geometry processing. vol. 7 (2006)
18. King, D.E.: Dlib-ml: A machine learning toolkit. The Journal of Machine Learning Research **10**, 1755–1758 (2009)
19. Liang, Y., Song, W., Yang, J., Qiu, L., Wang, K., He, L.: X2teeth: 3d teeth reconstruction from a single panoramic radiograph. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part II 23. pp. 400–409 (2020)
20. Lingchen, Y., Zefeng, S., Yiqian, W., Xiang, L., Kun, Z., Hongbo, F., Zheng, Y.: iorthopredictor: model-guided deep prediction of teeth alignment. ACM Transactions on Graphics **39**(6), 216 (2020)
21. Ma, J., Chen, J., Ng, M., Huang, R., Li, Y., Li, C., Yang, X., Martel, A.L.: Loss odyssey in medical image segmentation. Medical Image Analysis **71**, 102035 (2021)
22. Mavreas, D., Athanasiou, A.E.: Factors affecting the duration of orthodontic treatment: a systematic review. European journal of orthodontics **30**(4), 386–395 (2008)
23. Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models. arXiv preprint arXiv:2112.10741 (2021)
24. Richardson, E., Alaluf, Y., Patashnik, O., Nitzan, Y., Azar, Y., Shapiro, S., Cohen-Or, D.: Encoding in style: a stylegan encoder for image-to-image translation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2287–2296 (2021)
25. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. pp. 234–241 (2015)

26. Saharia, C., Chan, W., Chang, H., Lee, C., Ho, J., Salimans, T., Fleet, D., Norouzi, M.: Palette: Image-to-image diffusion models. In: ACM SIGGRAPH 2022 Conference Proceedings. pp. 1–10 (2022)
27. Tian, Y., Zhang, Y., Chen, W.G., Liu, D., Wang, H., Xu, H., Han, J., Ge, Y.: 3d tooth instance segmentation learning objectness and affinity in point cloud. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) **18**(4), 1–16 (2022)
28. Wirtz, A., Jung, F., Noll, M., Wang, A., Wesarg, S.: Automatic model-based 3-d reconstruction of the teeth from five photographs with predefined viewing directions. In: Medical Imaging 2021: Image Processing. vol. 11596, pp. 198–212 (2021)
29. Wirtz, A., Mirashi, S.G., Wesarg, S.: Automatic teeth segmentation in panoramic x-ray images using a coupled shape model in combination with a neural network. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part IV 11. pp. 712–719 (2018)
30. Wu, C., Bradley, D., Garrido, P., Zollhöfer, M., Theobalt, C., Gross, M.H., Beeler, T.: Model-based teeth reconstruction. ACM Trans. Graph. **35**(6), 220–1 (2016)
31. Zanjani, F.G., Moin, D.A., Verheij, B., Claessen, F., Cherici, T., Tan, T., et al.: Deep learning approach to semantic segmentation in 3d point cloud intra-oral scans of teeth. In: International Conference on Medical Imaging with Deep Learning. pp. 557–571 (2019)
32. Zhao, Q., Wu, X., Zhu, F., Liu, J., Wei, M., Peng, J., Lu, Y.: Automatic 3d teeth semantic segmentation with mesh augmentation network. In: 2022 3rd International Conference on Pattern Recognition and Machine Learning. pp. 136–142 (2022)
33. Zhu, G., Piao, Z., Kim, S.C.: Tooth detection and segmentation with mask r-cnn. In: 2020 International Conference on Artificial Intelligence in Information and Communication (ICAIIC). pp. 070–072 (2020)