3D Structure-guided Network for Tooth Alignment in 2D Photograph

Yulong Dou[1]
douyl2023@shanghaitech.edu.cn

Lanzhuju Mei[1]
meilzhj2023@shanghaitech.edu.cn

Dinggang Shen[1,2,3]
dgshen@shanghaitech.edu.cn

Zhiming Cui[1]
cuizhm@shanghaitech.edu.cn

[1] School of Biomedical Engineering
ShanghaiTech University
Shanghai 201210, China

[2] Shanghai United Imaging Intelligence
Co., Ltd.
Shanghai 200230, China

[3] Shanghai Clinical Research and Trial
Center
Shanghai 201210, China

Abstract

Orthodontics focuses on rectifying misaligned teeth (i.e., malocclusions), affecting both masticatory function and aesthetics. However, orthodontic treatment often involves complex, lengthy procedures. As such, generating a 2D photograph depicting aligned teeth prior to orthodontic treatment is crucial for effective dentist-patient communication and, more importantly, for encouraging patients to accept orthodontic intervention. In this paper, we propose a 3D structure-guided tooth alignment network that takes 2D photographs as input (e.g., photos captured by smartphones) and aligns the teeth within the 2D image space to generate an orthodontic comparison photograph featuring aesthetically pleasing, aligned teeth. Notably, while the process operates within a 2D image space, our method employs 3D intra-oral scanning models collected in clinics to learn about orthodontic treatment, i.e., projecting the pre- and post-orthodontic 3D tooth structures onto 2D tooth contours, followed by a diffusion model to learn the mapping relationship. Ultimately, the aligned tooth contours are leveraged to guide the generation of a 2D photograph with aesthetically pleasing, aligned teeth and realistic textures. We evaluate our network on various facial photographs, demonstrating its exceptional performance and strong applicability within the orthodontic industry.

1 Introduction

Orthodontic treatment is an effective remedy for correcting tooth misalignment (i.e., malocclusions). It is estimated that over 90% people suffer from malocclusion problems with various degrees[2], and most of people can benefit from orthodontic intervention. This treatment not only helps prevent oral diseases at a physiological level, but also significantly boosts patients' confidence, enhancing their psychological well-being[23]. However, the complexity of orthodontic procedure, which often spans several months or even years, can deter individuals from seeking treatment. Hence, the generation and visualization of potential
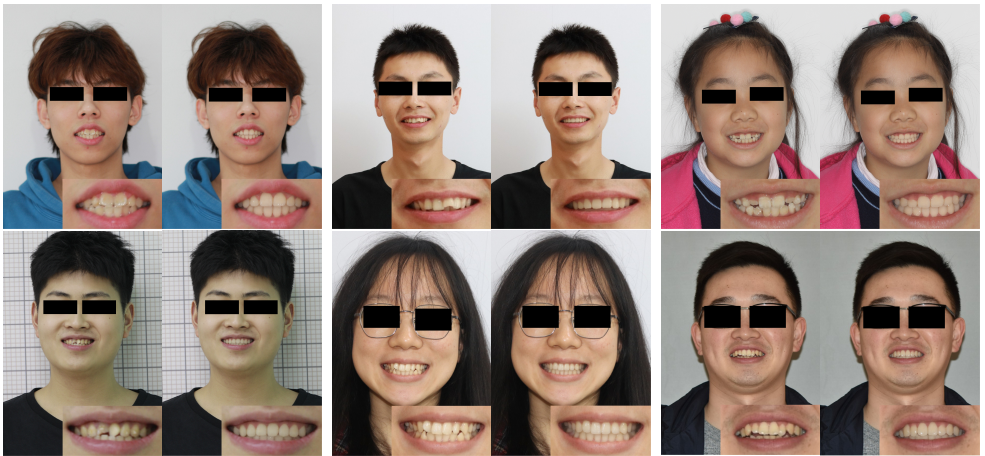
Figure 1: Orthodontic comparison photographs. For each case, we show the facial photograph with misaligned teeth (left) and the facial photograph with well-aligned teeth generated by our network (right), and the image in the lower right corner is a zoom-in of mouth region.

post-treatment facial photographs with aesthetical teeth becomes crucial. Such predictive imaging not only engages and motivates patients but also fosters more effective communication between orthodontists and their patients.

In clinical practice, visualizing patients' appearance after orthodontic treatment is referred to "Visual Treatment Objective" (VTO). This is typically performed on X-Ray images by deforming soft tissues and skeleton based on detected landmarks[21, 25]. However, this operation leaves the teeth's appearance unaltered,, making it challenging for patients to make a realistic comparison. In this study, our objective is to take 2D photograph as an input (e.g., photos captured by smartphones), and directly generate the "Orthodontic Comparison Photograph" with aligned teeth and realistic textures, as shown in Figure 1. Note that the generated photograph should follow the unique tooth alignment property of each patient in real-world treatment, instead of simplistic Photoshop approach with template teeth[32].

Currently, significant advancements in deep learning, particularly in generative networks, have achieved promising results in computer vision community. However, most of these models heavily rely on paired images, that is not suitable for our task. This is primarily because collecting paired pre- and post-orthodontic facial photographs is challenging due to the long-term orthodontic procedure and changes in facial appearance over time. Furthermore, the 2D photograph does not provide the 3D structure of teeth. Thus, how to learn the clinical knowledge of tooth alignment, defined on 3D tooth models, from 2D photographs is also a significant challenge.

In this paper, we propose a 3D structure-guided network for tooth alignment in 2D photograph. The key idea is to learn the clinical tooth alignment knowledge defined on the 3D intra-oral scanning models[17], and apply the learned property to guide the 2D post-orthodontic photograph generation. Specifically, we begin by collecting a set of paired pre- and post-orthodontic intra-oral scanning tooth models in clinics, and render[24] them onto the oral area of a 2D facial photograph. In this way, we can obtain the paired pre- and post-orthodontic tooth contours in 2D photograph (as shown in Figure 2). Then, a Diffusion Model[11] is applied to learn tooth alignment knowledge, i.e., generating post-orthodontic tooth contours with the input of pre-orthodontic tooth contours, derived from 3D tooth models. Note that only the tooth structures are captured, without any texture information. In the

inference process, we can directly take the tooth contours segmented from 2D facial photograph. Finally, guided by the aligned tooth contours, we employ another Diffusion Model to generate a realistic 2D photograph with aligned teeth. In particular, To enhance similarity with the patient's original appearance, we incorporate facial skin color and intra-oral highlights into the generation process, accounting for texture and lighting information. In the experiment, we collect a large number of photographs from patients suffering from malocclusion problems with various degrees, and achieve superior performance compared to the state-of-the-art methods, including GAN[8] and Diffusion Models. Furthermore, we also conduct a user study to validate the alignment and authenticity performance of our algorithm, demonstrating its potential applicability within orthodontic industry.

2 Related Work

Digital Orthodontics. Digital orthodontics employs digital imaging technologies such as intra-oral scanning[17], CBCT[2], and panoramic radiograph[1] to provide dentists with information about the structure and occlusion of patients' teeth. This helps dentists with pre-treatment diagnosis and orthodontic treatment planning. A variety of emerging techniques have been introduced in related fields, including tooth segmentation[5, 6], 3D tooth reconstruction[46, 47], and 3D tooth arrangement[35]. In terms of orthodontic comparison photographs, Lingchen et al. [20] have introduced iOrthoPredictor which can synthesize an image of well-aligned teeth based on a patient's facial photograph and an additional input of the patient's 3D dental model. Chen et al. [4] have introduced OrthoAligner which needs only a facial photograph but no 3D dental model as input, by introducing the concept of StyleGAN inversion. But OrthoAligner is limited in that it only uses facial photographs to learn tooth transformation, without utilizing information from 3D dental models.

Image Generation. Image generation is a field of research in computer vision that aims to generate new digital images by using algorithms or models from scratch or by modifying existing images. Several models have been proposed for image generation, including GAN[8], VAE[18], Diffusion Model[11]. Specifically, GAN simultaneously trains the generator and discriminator to generate more realistic images. Many models based on GAN have been proposed, such as unsupervised StyleGAN[15], and supervised Pix2pix GAN[13]. VAE is a generative model that uses variational inference for sampling from probability distributions. Ho et al. [11] propose Diffusion Model based on Score Matching[12] and Denoising Autoencoder[34], and elaborate on its mathematical principles. Diffusion Model is a generative model that utilizes a forward process of step-by-step adding noise and a backward denoising process to generate high-quality images. Choi et al. [4] propose a reference-guided conditional Diffusion Model that fine-tunes the backward denoising process. Singh et al. [31] introduce condition noise to navigate Diffusion Model. Saharia et al. [29] propose an image-to-image Diffusion Model guided by condition image.

3 Method

3.1 Overview

Overall, our goal is to design a tooth alignment network that incorporates 3D structural information derived from intra-oral scanning models, which is essential for clinical orthodontic treatment, and guide the orthodontic comparison photograph generation.

We have pre-orthodontic intra-oral scanning models $S = \{S_1, S_2, ..., S_N\}$, post-orthodontic intra-oral scanning models $\hat{S} = \{\hat{S}_1, \hat{S}_2, ..., \hat{S}_N\}$ of same patients collected in clinics, and unpaired facial photographs $I = \{I_1, I_2, ..., I_M\}$ collected by smartphones. Given that the facial photographs I and 3D intra-oral scanning models $S, \hat{S}$ in our dataset are not paired, we design a module, named *Align-Mod*, for tooth alignment that can still incorporate 3D structural information from intra-oral scanning models as guidance. This module randomly selects the pre- and post-orthodontic intra-oral scanning models (i.e., $S_r \in S$ and $\hat{S}_r \in \hat{S}$) for an unpaired facial photograph (i.e., $I_r \in I$), and makes a coarse 2D-3D registration between $S_r, \hat{S}_r$ and I_r, respectively. Then, the 3D tooth structures are projected onto the 2D facial photograph to obtain pre-orthodontic tooth contours $C_r \in \mathbb{R}^{128 \times 256}$ and post-orthodontic tooth contours $\hat{C}_r \in \mathbb{R}^{128 \times 256}$. **In this way, our *Align-Mod* module can learn the tooth transformation $\mathcal{T}(\cdot)$, which represents the clinical orthodontic knowledge derived from the 3D intra-oral scanning models.**

In addition to the pre-trained tooth alignment module, we also design a **segmentation module, named *Segm-Mod*, to locate the mouth region and segment tooth contours C from facial photographs I, and a generation module, named *Gen-Mod*, to generate the facial image with aesthetically pleasing teeth.**

In summary, the three modules designed in this framework are shown in Figure 2.
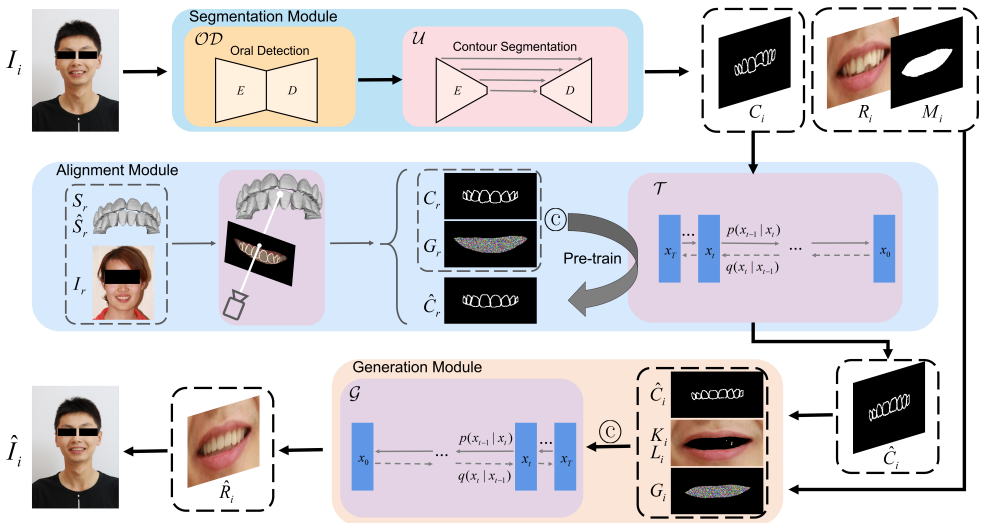

Figure 2: Overall pipeline. When a facial photograph is input into our network, it first goes through *Segm-Mod* to obtain oral mask, mouth region and tooth contours. Then it enters Pre-trained *Align-Mod* to predict well-aligned tooth contours, and finally goes through *Gen-Mod* to generate a facial photograph with well-aligned teeth.

3.2 Segmentation Module

To begin with, *Segm-Mod* needs to detect the position of face[16, 30] and obtain a standardized face $F_i \in \mathbb{R}^{512 \times 512}$ from any given facial photograph $I_i \in I \subseteq \mathbb{R}$. As shown in Figure 2, to accurately locate the mouth, we propose an oral detection network $\mathcal{OD}(\cdot)$[19, 38] to segment the oral mask $M_i \in \mathbb{R}^{128 \times 256}$ and crop the mouth region $R_i \in \mathbb{R}^{128 \times 256}$ from the standardized face F_i. Then, to obtain tooth contours $C_i \in \mathbb{R}^{128 \times 256}$ which contains struc-

tural information of teeth, we employ a commonly-used supervised segmentation network, U-Net[27], for segmenting tooth contours C_i from the mouth region R_i. The formulation of process of *Segm-Mod* is as:

$$
\begin{aligned}
R_i, M_i &= \mathcal{OD}(I_i), \\
C_i &= \mathcal{U}(R_i, M_i), \quad \forall i = 1, 2, ..., N,
\end{aligned}
\tag{1}
$$

where $\mathcal{OD}(\cdot)$ is the oral detection network performed in 2D facial photograph I_i, $\mathcal{U}(\cdot)$ denotes the U-Net-based contour segmentation network, and C_i, R_i, M_i are the pre-orthodontic tooth contours, mouth region and oral mask obtained by *Segm-Mod*, respectively.

To train the network, we employ the Dice Loss[22] and Weighted Cross-Entropy Loss[28]. Given the imbalanced area between foreground (tooth contours) to background, Dice Loss has excellent performance in situations with severe imbalance and focuses on learning the foreground area. Furthermore, Weighted Cross-Entropy Loss can address the imbalance problem by adjusting the weighted proportion, thus making it a suitable complement to Dice Loss. Our designed loss function is defined as:

$$
\mathcal{L} = w_{dice} \cdot \mathcal{L}_{dice} + w_{ce} \cdot \mathcal{L}_{ce}.
\tag{2}
$$

3.3 Alignment Module

One of the most innovative aspects of our method is that we can incorporate structural information from 3D intra-oral scanning models $S, \hat{S}$ into *Align-Mod*, which are essential for clinical orthodontic treatment. We employ a 3D-to-2D Render to project the 3D intra-oral scanning models $S, \hat{S}$ into the oral area of the 2D facial photographs, as opposed to the approach used by Wirtz et al. [36], Zheng et al. [40] to reconstruct 3D dental models from multi-view tooth photographs. Furthermore, we design a conditional Diffusion Model-based network for learning the clinical orthodontic knowledge $\mathcal{T}(\cdot)$ in the space of tooth contours $C_r, \hat{C}_r$ obtained by Render.

3.3.1 Render

Since the 3D intra-oral scanning models $S, \hat{S}$ and facial photographs I are collected from different environments and sources, i.e., one is in clinics and another is from smartphones in daily life, we cannot perform precise 3D-2D registration through rigid transformation. Fortunately, precise registration is not necessary for our task as our purpose is to create paired tooth contours from the intra-oral scanning models. Therefore, to obtain tooth contours $C_r, \hat{C}_r \in \mathbb{R}^{128 \times 256}$, we perform coarse registration based on the landmarks between the coordinates of tooth cusp points of central and lateral incisors in both render-used intra-oral scanning models $S_r \in S, \hat{S}_r \in \hat{S}$ and facial photograph $I_r \in I$.

We use Numerical Optimization[33] to perform coarse registration and then obtain 2D tooth contours by projection. The essential principle of projecting 3D to 2D is to solve the camera parameters. As shown in Equation 3,

$$
\rho m = (KR^T | - KR^T C) \begin{pmatrix} M \\ 1 \end{pmatrix},
\tag{3}
$$

where M represents the coordinate of the point in world coordinate system, denoted as $M = (X, Y, Z)^T$, and m represents the coordinate of the corresponding point in pixel coordinate

system, denoted as $m = (u, v, 1)^T$. ρ is the projection depth, C is the position of camera, R is the rotation matrix that represents camera pose, and K is the matrix of intrinsic parameters[9]. Specifically, we use the four paired tooth cusp points mentioned above on both the facial photo and the intra-oral scanning model to derive their coordinates, represented as m and M. Subsequently, we can resolve camera parameters, primarily the unknown variables in K and C, given the other known parameters.

3.3.2   Tooth Transformation

Once we render 3D intra-oral scanning models $S_r, \hat{S}_r$ onto the 2D facial photograph I_r and obtain pre- and post-orthodontic tooth contours $C_r, \hat{C}_r$ as mentioned above, we can then learn the $\mathcal{T}(\cdot)$, i.e., clinical orthodontic knowledge. We employ a network based on image-to-image Diffusion Model[29], as shown in Figure 2. To emphasize the generation of oral regions, we introduce Gaussian noise $G_r \in \mathbb{R}^{128 \times 256}$ generated within the oral mask M_r. We concatenate the pre-orthodontic tooth contours C_r with the Gaussian noise G_r to form the condition information, which serves as guidance for our diffusion model. Therefore, the formula for $\mathcal{T}(\cdot)$ during the pre-trained process is given as:

$$\hat{C}_r = \mathcal{T}(C_r \copyright G_r), \tag{4}$$

where $\copyright$ denotes channel-wise concatenation.

    Since we have pre-learned the $\mathcal{T}(\cdot)$, which represents the clinical orthodontic knowledge of tooth transformation, we can apply the learned knowledge to process the 2D tooth contours C_i derived from *Segm-Mod*. Hence, we concatenate the tooth contours C_i, together with intra-oral Gaussian noise G_i, and then feed them into our diffusion model, expecting a reasonable prediction for well-aligned tooth contours $\hat{C}_i$. The inference process is formulated as:

$$\hat{C}_i = \mathcal{T}(C_i \copyright G_i), \quad \forall i = 1, 2, ..., N. \tag{5}$$

3.4   Generation Module

After obtaining well-aligned tooth contours $\hat{C}_i$ through the tooth transformation $\mathcal{T}(\cdot)$ of our *Align-Mod*, we aim to generate a mouth region with realistic teeth $\hat{R}_i$ guided by $\hat{C}_i$. To achieve this, we adapt a conditional Diffusion Model-based generative network $\mathcal{G}(\cdot)$ in our *Gen-Mod*. We still introduce Gaussian noise $G_i \in \mathbb{R}^{128 \times 256}$ generated within the oral mask M_i to emphasize the generation of oral regions. Besides well-aligned tooth contours $\hat{C}_i$ and intra-oral Gaussian noise G_i mentioned above, we additionally introduce intra-oral highlights $L_i \in \mathbb{R}^{128 \times 256}$ and facial skin color $K_i \in \mathbb{R}^{128 \times 256}$, which are helpful for generating more realistic tooth color and environmental lighting. Then, four of them are concatenated together as the condition information to guide our generation model.

    In terms of intra-oral highlights, we employ Contrast Limited Adaptive Histogram Equalization (CLAHE) [41] and Thresholding, aimed to enhance image contrast and detect highlights within the oral region. Specifically, we utilize CLAHE as in Equation 6 to improve local contrast in mouth region R_i by using a histogram equalization approach with a specified contrast limit of 5 in 20×20 local window, thus preventing over-amplification of noise in flat areas while enhancing contrast in textured areas, defined as:

$$g_{xy} = \frac{L-1}{S_{xy}} \sum_{z=0}^{f_{xy}} \frac{h(x,z)}{S_{xy}}, \quad f'_{xy} = g_{xy} \times (L-1), \quad \text{CLAHE}_{xy} = \begin{cases} f'_{xy} & \text{if } f'_{xy} < L-1 \\ L-1 & \text{if } f'_{xy} \geq L-1 \end{cases}, \tag{6}$$

where S_{xy} denotes size of the local window, f_{xy} denotes pixel intensity at pixel (x,y), $h(x,z)$ is histogram of pixel intensities in local window, L means number of intensity level, g_{xy} is gain factor, and f'_{xy} is transformed pixel intensity.

Once we generate a mouth region with realistic teeth $\hat{R}_i$, we use the face and mouth position stored in *Segm-Mod* to replace the oral region in the initial facial photograph, thereby obtaining a facial photograph with well-aligned and aesthetically pleasing teeth $\hat{I}_i$ for VTO (see Figure 2). The formulation of process of *Gen-Mod* is shown as:

$$\hat{I}_i = \mathcal{G}(\hat{C}_i \copyright G_i \copyright L_i \copyright K_i), \ \ \forall i = 1, 2, ..., N, \tag{7}$$

where $\copyright$ denotes channel-wise concatenation of segmented tooth contours $\hat{C}_i$, Gaussian noise G_i, intra-oral highlights L_i and facial skin color K_i. $\hat{I}_i$ is the predicted facial photograph with well-aligned and aesthetically pleasing teeth through the Diffusion Model-based generative network $\mathcal{G}(\cdot)$ mentioned in *Gen-Mod*.

4 Experiments

4.1 Experiments Settings

Dataset. Our dataset comprises 1367 facial photographs I, of which 1129 are used to train *Segm-Mod* and *Gen-Mod*, 138 are used to create datasets through Render in *Align-Mod*, and the remaining 100 are reserved for testing our overall pipeline. For the 138 render-used facial photographs, we manually annotate the coordinates of tooth cusp points of central and lateral incisors in the upper jaw. Moreover, We have 1257 3D intra-oral scanning models S collected in dental clinics, along with their corresponding orthodontic treatment plans provided by dentists. In this way, we can also obtain corresponding 1257 post-orthodontic intra-oral scanning models $\hat{S}$ accordingly. Note that for each of the 138 render-used facial photographs, we randomly select 10 from the pool of 1257 intra-oral scanning models to perform the Render process as mentioned in 3.3.1. Thus we have 1380 pre- and post-orthodontic tooth contours $C_r, \hat{C}_r$, respectively, for training *Align-Mod*.

Implementation Details. The proposed method is implemented in PyTorch with 2 NVIDIA A100 80GB GPU. By iteratively tuning and training *Segm-Mod*, we utimately choose to assign 0.8 to w_{dice} and 0.2 to w_{ce}, along with a weight of 20 for the foreground and a weight of 1 for the background in $\mathcal{L}_{ce}$. In terms of *Align-Mod* and *Gen-Mod*, we set the batch-size of our diffusion model to 60. The learning rate is $5e$-5 and we use Exponential Moving Average[10] with $\beta = 0.9999$ to update parameters of diffusion model. Lastly regarding the parameters in Render, we set focal length in camera intrinsic parameters to 213.33, and we use SGD[26] optimizer with an initial learning rate of 0.01 and a learning rate scheduler, which reduces the learning rate by a factor of 0.9 every 500 steps.

4.2 Results

Based on the pre-trained *Align-Mod*, our three-stage network can infer a facial photograph with well-aligned and aesthetically pleasing teeth $\hat{I}_i$ according to patient's previous photograph I_i, without any 3D intra-oral scanning model as input, while still benefiting from the guidance of clinical orthodontic knowledge within 3D structure of intra-oral scanning models. To demonstrate the outstanding results of our method and provide a more detailed view

of the inference process and confidence level, we present some testing cases in Figure 1 and Figure 3.

Based on the visual results and inference process we presented, it's evident that our *Segm-Mod* has excellent segmentation ability, even with misaligned teeth. Our *Align-Mod* also performs great reliability on predicting well-aligned tooth contours, closely based on the pre-trained transformation $\mathcal{T}(\cdot)$ in the image space, which is crucial for clinical orthodontic treatment. Besides, our *Gen-Mod* can infer reasonable realistic teeth with similar color and lighting compared with patient's previous teeth photograph and its shooting environment.
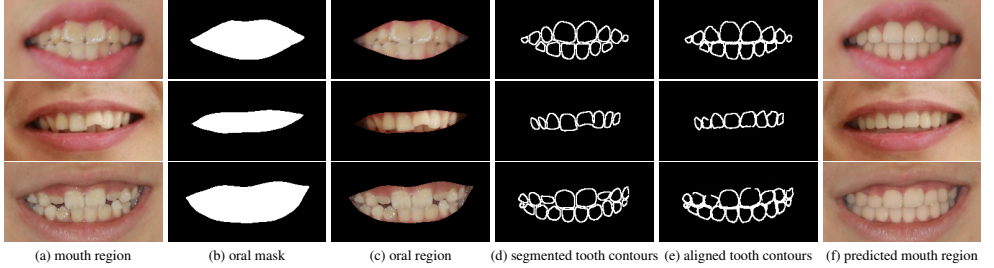

(a) mouth region   (b) oral mask   (c) oral region   (d) segmented tooth contours   (e) aligned tooth contours   (f) predicted mouth region

Figure 3: Inference process. For each detected mouth region R_i (a), we segment to obtain the oral mask M_i (b) and oral region (c). We further obtain tooth contours C_i (d) from our *Segm-Mod* and input it into our *Align-Mod* to yield well-aligned tooth contours $\hat{C}_i$ (e). We finally predict a mouth region with well-aligned teeth $\hat{R}_i$ (f) through our *Gen-Mod*.

4.3   Comparison

Additionally, we qualitatively compare our tooth alignment network with Pix2pix GAN[8, 13], especially for *Align-Mod* and *Gen-Mod*, and the comparison results are visually shown in Figure 4. It is shown that our Diffusion Model-based methods are more capable than Pix2pix GAN-based methods, with more reasonable alignment prediction and more realistic tooth color and lighting. As shown in Table 1, we quantitatively evaluate our proposed *Align-Mod* and *Gen-Mod* to make comparisons with pix2pix GAN using pixel-wise $\mathcal{L}1$, $\mathcal{L}2$ and LPIPS error[14, 39]. $\mathcal{L}1$ and $\mathcal{L}2$ are commonly used pixel-wise metrics for quantifying discrepancies between generated results and the target, whereas LPIPS is a perceptual metric which calculates the perceptual distance and visual similarity between images. It is shown that our method are consistently better than Pix2pix GAN-based methods considering pixel-wise metrics.

4.4   Ablation Study

We have mentioned in Subsection 3.4 that, in order to make our *Gen-Mod* yield more realistic tooth color and environmental lighting, we introduce intra-oral highlights L_i and facial skin color K_i and concatenate them together with well-aligned tooth contours $\hat{C}_i$ and intra-oral Gaussian noise G_i as guidance. To validate the effectiveness of these two condition images guiding *Gen-Mod*, we conduct an ablation study to demonstrate their effectiveness. We design four groups of ablation experiments as shown in Table 2, all using Diffusion Model-based generative network and the same datasets for training. Visual results of four ablation experiments are shown in Figure 5.

Specifically, Ablation I includes neither of the two condition images, resulting in the poorest generation performance. Ablation II includes facial skin color, resulting in better
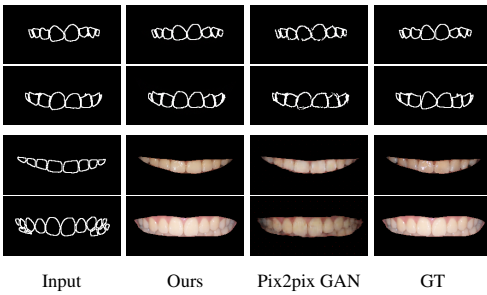
	$\mathcal{L}1$	$\mathcal{L}2$	LPIPS
Diffusion Model -based *Align-Mod*	**0.053**	**0.319**	**0.057**
Pix2pix GAN -based *Align-Mod*	0.061	0.343	0.104
Diffusion Model -based *Gen-Mod*	**0.029**	**0.093**	**0.038**
Pix2pix GAN -based *Gen-Mod*	0.047	0.148	0.107

Figure 4: Qualitative comparisons. The upper two rows are two testing cases of *Align-Mod*, and the lower two rows are two testing cases of *Gen-Mod*.

Table 1: Quantitative comparisons between different methods on Testing Dataset.

	facial skin color	intra-oral highlights
Ablation I		
Ablation II	✓	
Ablation III		✓
Ablation IV	✓	✓

Table 2: Four groups of ablation experiments.

Figure 5: Visual results of four ablation experiments.

tooth color but lighting information is lost compared to the original image. Ablation III adds intra-oral highlights, significantly restoring the environmental lighting but less realistic tooth color. Ablation IV has intuitively the best generation performance, with both realistic facial skin color and intra-oral highlights.

4.5 User Study

To further demonstrate the reliability and credibility of our method, we conduct a user study that we invite 30 individuals to vote for assessing the alignment and authenticity of photographs (only concentrated on mouth region) generated by our method. Specifically, for assessing alignment, we randomly select 10 generated facial photographs and 10 photographs from patients who have received orthodontic treatments. Participants are asked to rate the alignment of teeth in the photographs on a scale of 1 to 5, with higher scores indicating better alignment. Similarly, for assessing authenticity, we randomly select 10 generated facial photographs and 10 real ones, and ask participants to vote on whether they are real or fake. Table 3 has shown the average alignment scores and average percentage of photographs being classified to "real". Compared with well-aligned or real photographs, photographs generated by our method achieve high scores in terms of both alignment and authenticity, only slightly lower than scores of the well-aligned teeth on real photographs.

5 Discussion

In this work, we propose a 3D structure-guided tooth alignment network to effectively generate orthodontic comparison photographs. According to the experimental results above, our

	Well-aligned photos	Real photos	Ours
Alignment	3.84		3.82
Authenticity		72.67%	65.00%

Table 3: Voting results of user study. First row represents the average score regarding alignment, second row represents the average percentage of photographs being classified as "real".

method utilizes 3D dental models to learn the orthodontic knowledge in the image space. The 3D structure successfully guides the learning and prediction of our network, giving our method practical clinical significance. Additionally, we introduce Diffusion Model into the task of orthodontic comparison photograph generation and have shown the great power of Diffusion Model in our *Align-Mod* and *Gen-Mod*. Importantly, different from state-of-the-art methods[3, 20] in the field, our method can still incorporate clinical orthodontic knowledge into the network without requiring additional input of dental models. This demonstrates that our method is much more clinically practical, user-friendly and applicable within orthodontic industry.

Our method, however, is not without limitations. For example, our network cannot handle with several cases, such as that teeth are too misaligned and patients smile too widely. Moreover, our method cannot take collision and occlusal relationship into consideration since our method is just performed in the image space. In future, we plan to first reconstruct 3D tooth models from 2D photograph and then make the tooth alignment.

6 Conclusion

In this paper, we have designed a 3D structure-guided network to infer a facial photograph with well-aligned and aesthetically pleasing teeth based on the patient's previous facial photograph. Our method stands out from existing approaches as it can learn the clinical orthodontic knowledge based on 3D intra-oral scanning models, making our method highly reliable and potentially applicable in clinical practice.

Acknowledgements

References

[1] Christos Angelopoulos, Aurelija Bedard, Jerald O Katz, Stelios Karamanis, and Nikos Parissis. Digital panoramic radiography: An overview. In *Seminars in Orthodontics*, volume 10, pages 194–203. Elsevier, 2004.

[2] Olaf Bernhardt, Karl-Friedrich Krey, Amro Daboul, Henry Voelzke, Stefan Kindler, Thomas Kocher, and Christian Schwahn. New insights in the link between malocclusion and periodontal disease. *Journal of clinical periodontology*, 46(2):144–159, 2019.

[3] Beijia Chen, Hongbo Fu, Kun Zhou, and Youyi Zheng. Orthoaligner: Image-based teeth alignment prediction via latent style manipulation. *IEEE Transactions on Visualization and Computer Graphics*, 2022.

[4] Jooyoung Choi, Sungwon Kim, Yonghyun Jeong, Youngjune Gwon, and Sungroh Yoon. Ilvr: Conditioning method for denoising diffusion probabilistic models. *arXiv preprint arXiv:2108.02938*, 2021.

[5] Zhiming Cui, Changjian Li, and Wenping Wang. Toothnet: automatic tooth instance segmentation and identification from cone beam ct images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6368–6377, 2019.

[6] Zhiming Cui, Yu Fang, Lanzhuju Mei, Bojun Zhang, Bo Yu, Jiameng Liu, Caiwen Jiang, Yuhang Sun, Lei Ma, Jiawei Huang, et al. A fully automatic ai system for tooth and alveolar bone segmentation from cone-beam ct images. *Nature communications*, 13(1):2096, 2022.

[7] W De Vos, Jan Casselman, and GRJ19464146 Swennen. Cone-beam computerized tomography (cbct) imaging of the oral and maxillofacial region: a systematic review of the literature. *International journal of oral and maxillofacial surgery*, 38(6):609–625, 2009.

[8] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[9] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003.

[10] David Haynes, Steven Corns, and Ganesh Kumar Venayagamoorthy. An exponential moving average algorithm. In *2012 IEEE Congress on Evolutionary Computation*, pages 1–8. IEEE, 2012.

[11] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

[12] Aapo Hyvärinen and Peter Dayan. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(4), 2005.

[13] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[14] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 694–711. Springer, 2016.

[15] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019.

[16] Vahid Kazemi and Josephine Sullivan. One millisecond face alignment with an ensemble of regression trees. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1867–1874, 2014.

[17] Hidemichi Kihara, Wataru Hatakeyama, Futoshi Komine, Kyoko Takafuji, Toshiyuki Takahashi, Jun Yokota, Kenta Oriso, and Hisatomo Kondo. Accuracy and practicality of intraoral scanner in dentistry: A literature review. *Journal of prosthodontic research*, 64(2):109–113, 2020.

[18] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[19] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5549–5558, 2020.

[20] YANG Lingchen, SHI Zefeng, WU Yiqian, LI Xiang, ZHOU Kun, FU Hongbo, and Youyi Zheng. iorthopredictor: model-guided deep prediction of teeth alignment. *ACM Transactions on Graphics*, 39(6):216, 2020.

[21] Richard P McLaughlin and John C Bennett. The dental vto: an analysis of orthodontic tooth movement. *Journal of Clinical Orthodontics: JCO*, 33(7):394–403, 1999.

[22] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. Ieee, 2016.

[23] Minghui Peng, Jing Kang, and Xiao Deng. The role of body image in orthodontic treatment for adolescents. *West China Journal of Stomatology*, 35(5):489, 2017.

[24] Matt Pharr, Wenzel Jakob, and Greg Humphreys. *Physically based rendering: From theory to implementation*. Morgan Kaufmann, 2016.

[25] G Power, J Breckon, M Sherriff, and F McDonald. Dolphin imaging software: an analysis of the accuracy of cephalometric digitization and orthognathic prediction. *International journal of oral and maxillofacial surgery*, 34(6):619–626, 2005.

[26] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

[27] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 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*, pages 234–241. Springer, 2015.

[28] Reuven Rubinstein. The cross-entropy method for combinatorial and continuous optimization. *Methodology and computing in applied probability*, 1:127–190, 1999.

[29] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–10, 2022.

[30] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[31] Vedant Singh, Surgan Jandial, Ayush Chopra, Siddharth Ramesh, Balaji Krishna-murthy, and Vineeth N Balasubramanian. On conditioning the input noise for controlled image generation with diffusion models. *arXiv preprint arXiv:2205.03859*, 2022.

[32] Manoj Kumar Sundar and BDS Venkataraman Chelliah. Ten steps to create virtual smile design templates with adobe photoshop cs6. *Compendium*, 39(3), 2018.

[33] Bo D Tapley and JM Lewallen. Comparison of several numerical optimization methods. *Journal of Optimization Theory and Applications*, 1:1–32, 1967.

[34] Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural computation*, 23(7):1661–1674, 2011.

[35] Guodong Wei, Zhiming Cui, Yumeng Liu, Nenglun Chen, Runnan Chen, Guiqing Li, and Wenping Wang. Tanet: towards fully automatic tooth arrangement. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16*, pages 481–497. Springer, 2020.

[36] Andreas Wirtz, Florian Jung, Matthias Noll, Anqi Wang, and Stefan Wesarg. Automatic model-based 3-d reconstruction of the teeth from five photographs with predefined viewing directions. In *Medical Imaging 2021: Image Processing*, volume 11596, pages 198–212. SPIE, 2021.

[37] Chenglei Wu, Derek Bradley, Pablo Garrido, Michael Zollhöfer, Christian Theobalt, Markus H Gross, and Thabo Beeler. Model-based teeth reconstruction. *ACM Trans. Graph.*, 35(6):220–1, 2016.

[38] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 325–341, 2018.

[39] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

[40] TX Zheng, Shuai Huang, YF Li, and MC Feng. Key techniques for vision based 3d reconstruction: a review. *Zidonghua Xuebao*, 46:631–652, 2020.

[41] Karel Zuiderveld. Contrast limited adaptive histogram equalization. *Graphics gems*, pages 474–485, 1994.