OrthoFlow: Decoupled Landmark Prediction and Image Synthesis for Orthodontic Treatment Outcome Visualization

Leyuan Wang[1*], Yulong Dou[1*], Jiancheng Yang[2,3], Guangying Song[4], and Zhiming Cui[1(✉)]

[1] School of Biomedical Engineering & State Key Laboratory of Advanced Medical Materials and Devices, ShanghaiTech University, Shanghai, China
cuizhm@shanghaitech.edu.cn
[2] ELLIS Institute Finland, Espoo, Finland
[3] Aalto University, Espoo, Finland
[4] Peking University School and Hospital of Stomatology, Beijing, China

Abstract. Predicting post-treatment facial profiles from lateral cephalograms is clinically valuable for orthodontic planning and patient communication, yet remains challenging due to the complex interplay between skeletal, dental, and soft-tissue changes. Conventional approaches either predict discrete clinical indices without visual output, or apply direct image-to-image translation that fails to preserve anatomical consistency. We propose OrthoFlow, a two-stage decoupled generative framework that separates geometric prediction from image synthesis. In the first stage, a Flow Matching model with a Graph-Transformer hybrid backbone predicts post-treatment landmark displacements conditioned on textual treatment plans. In the second stage, a conditional Flow Matching model synthesizes high-fidelity lateral cephalograms guided by the predicted anatomical contours. Experiments on a large-scale orthodontic dataset demonstrate that OrthoFlow achieves superior geometric accuracy and visual fidelity compared to existing methods, validated by both quantitative metrics and clinical evaluation by orthodontists. Our code and dataset are available at https://github.com/ShanghaiTech-IMPACT/OrthoFlow.

Keywords: Orthodontics · Cephalometric Synthesis · Flow Matching.

1 Introduction

Orthodontic treatment planning relies heavily on cephalometric analysis, the measurement and interpretation of anatomical landmarks from lateral cephalograms [6]. Because treatment typically spans months to years, generating accurate and visually realistic predictions of post-treatment outcomes, termed Visual Treatment Objectives (VTOs), is essential. Such predictions assist clinicians in

* indicates equal contribution.

formulating precise treatment plans and serve as intuitive communication tools that reduce patient anxiety [14]. Currently, orthodontists rely on rule-based tools such as Ricketts VTO [14] and Holdaway analysis [6], or subjective clinical estimation. These approaches are limited to linear and angular measurements, depend heavily on practitioner expertise, and cannot produce visual simulations of the post-treatment facial profile. Moreover, different treatment strategies (e.g., with or without extraction) yield substantially different outcomes, necessitating treatment-plan-aware prediction. There is therefore a clear clinical demand for automated systems that generate geometrically accurate and visually realistic predictions conditioned on the specific clinical plan.

Conventional machine learning methods [1] predict discrete clinical indices (e.g., ANB angle or treatment duration) but fail to model complex nonlinear anatomical changes, offering limited clinical utility. Pure landmark prediction methods [20] provide geometric precision yet lack visual output, restricting their value in patient communication. More recently, image-to-image translation has been applied to medical image synthesis. GAN-based approaches, including pix2pixHD [16] and UGATIT [22,7], produce high-resolution outputs but frequently suffer from mode collapse and training instability. Diffusion models [4,13] offer improved stability through iterative denoising but remain computationally expensive. However, directly applying either paradigm to orthodontic prediction presents fundamental challenges: multi-center data inevitably introduces device-specific domain shifts, slight posture deviations hinder precise pixel-level alignment, and the absence of explicit structural constraints causes models to overfit pixel-level differences while ignoring clinically meaningful geometric changes.

A key insight of our work is that the orthodontic prediction problem can be naturally decomposed into two distinct sub-problems with fundamentally different characteristics: predicting *where* anatomical structures move (a low-dimensional geometric problem well-suited to structured generative models) and synthesizing *what* the resulting image looks like (a high-dimensional appearance problem). This decomposition brings two advantages. First, geometric prediction in landmark space is inherently low-dimensional and topology-constrained, making it amenable to Flow Matching [8,9], whose deterministic optimal transport trajectories require fewer integration steps than DDPMs [4] and naturally respect geometric structure. Second, by conditioning image synthesis on explicit structural priors rather than learning pixel-level correspondences end-to-end, the synthesis stage can focus entirely on texture fidelity without the confounding burden of simultaneously inferring spatial deformations.

Building on this insight, we present OrthoFlow, a decoupled two-stage framework for adult orthodontic outcome prediction, where patients possess fully mature craniofacial skeletons and profile improvements depend entirely on intervention rather than natural growth. OrthoFlow contributes a text-conditioned structural prediction module powered by Flow Matching with a Graph-Transformer hybrid backbone [15], which models both local anatomical topology via dual-path graph networks and global cross-structure interactions via Transformer decoders, enabling accurate landmark displacement prediction that adapts to

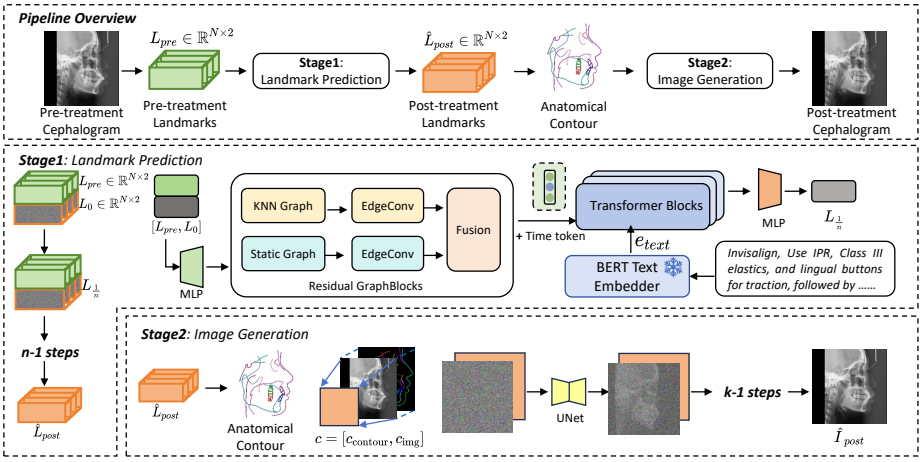

Fig. 1. Overview of OrthoFlow. Stage 1 predicts post-treatment landmarks via text-conditioned Flow Matching with a Graph-Transformer backbone. Stage 2 synthesizes the post-treatment cephalogram conditioned on the predicted anatomical contour and the pre-treatment image.

diverse clinical plans. OrthoFlow further contributes a structure-guided image synthesis module that leverages the predicted landmarks as explicit geometric constraints, employing a conditional Flow Matching model to synthesize high-fidelity cephalograms that are both visually realistic and clinically meaningful.

2 Method

The overall pipeline of OrthoFlow is illustrated in Fig. 1. Given a pre-treatment lateral cephalogram, a U-Net-based landmark detector [18] extracts the pre-treatment landmark coordinates L_{pre}. In Stage 1, we predict the post-treatment landmarks $\hat{L}_{\text{post}}$ from L_{pre} conditioned on a textual treatment plan. The predicted landmarks are rendered as an anatomical contour map. In Stage 2, this contour map, together with the pre-treatment cephalogram, serves as the conditioning signal for synthesizing the post-treatment cephalogram.

2.1 Stage 1: Text-Conditioned Landmark Flow Matching

Flow Matching Formulation. Unlike traditional single-step regression models, Flow Matching [8,9], as a generative method, holds a natural advantage in producing diverse outcomes. Furthermore, compared to denoising models such as Diffusion [4], Flow Matching constructs deterministic optimal transport trajectories, thereby providing better determinism and higher efficiency.

Concretely, the model learns a linear trajectory from a Gaussian noise distribution $L_{\text{noise}} \sim \mathcal{N}(0, \mathbf{I})$ to the ground-truth post-treatment landmarks L_{post}.

At timestep $t \in [0, 1]$, the intermediate state is:

$$L_t = t\, L_{\text{post}} + (1 - t)\, L_{\text{noise}} \tag{1}$$

The patient's pre-treatment anatomy L_{pre} is injected as a spatial condition by channel-wise concatenation with L_t at each step. It learns a conditional velocity field $\hat{v}_t = v_\theta(L_t, t, L_{\text{pre}}, \text{text})$. At inference, generation proceeds as an ODE from $L_0 \sim \mathcal{N}(0, \mathbf{I})$, iteratively updating $L_{t+\mathrm{d}t} = L_t + \hat{v}_t \cdot \mathrm{d}t$ until t=1 to obtain $\hat{L}_{\text{post}}$.

Graph-Transformer Hybrid Backbone. We design a Graph-Transformer model as the velocity predictor v_θ. Each landmark corresponds to a specific anatomical position within a particular tissue type. A Graph Neural Network (GNN) extracts local topological relationships among these points, while a Transformer models global cross-structure interactions.

Local Topology Modeling. We extract local anatomical features via a dual-path residual Graph Neural Network [21] G_d based on EdgeConv [17]. The *Static Anatomical Graph* encodes the regional affiliation (e.g., tooth, jaw) of each landmark through predefined fixed edges, capturing anatomical priors that are invariant across patients. The *Dynamic KNN Graph* complements this by constructing a k-nearest-neighbor graph per sample in feature space, adaptively discovering cross-tissue correlations (e.g., between distant teeth undergoing coordinated movement) that do not follow predefined adjacency. Outputs from both paths are fused through a pointwise convolutional layer with residual connections.

Global Context Aggregation. Transformer [15] decoder layers are applied to the graph features to model long-range interactions that the local graph cannot capture, such as the propagation of mandibular repositioning to the soft-tissue profile. A dedicated time token is prepended to the landmark sequence to condition the velocity prediction on the continuous flow timestep t.

Text-Conditioned Generation. A central clinical requirement is that predictions should vary with the treatment plan: teeth extraction v.s non-extraction, for instance, produces markedly different dental retraction and soft-tissue outcomes. To incorporate this plan-level guidance, textual treatment descriptions are encoded via BERT [3] into an embedding $\mathbf{e}_{\text{text}}$, linearly projected to the model's hidden dimension D. Cross-attention in each Transformer decoder layer enables the structural features to selectively attend to the relevant treatment semantics. Classifier-Free Guidance [5] is applied by randomly replacing 15% of text inputs with zero vectors during training, strengthening text-conditioning at inference.

Training Objective with Anatomical Constraints. Predicting independent landmarks often disrupts the continuous anatomical topology. To preserve structural integrity, we introduce an Edge Consistency Loss ($\mathcal{L}_{\text{edge}}$). Rather than merely constraining individual point coordinates, this loss enforces topological

coherence within key anatomical regions (e.g., mandible, teeth, and soft tissues). Specifically, it penalizes the discrepancies in Euclidean distances between interconnected landmark pairs (edges) across the predicted and ground-truth shapes.

The implied post-treatment landmarks are derived from the predicted velocity via $\hat{L}_{\text{post}} = L_t + (1-t)\hat{v}_t$ (see Eq. 1). The overall loss function comprises a weighted MSE loss ($\mathcal{L}_{\text{mse}}$) and an edge consistency loss, which are computed between this predicted $\hat{L}_{\text{post}}$ and the ground truth L_{post}:

$$\mathcal{L}_{\text{stage1}} = \mathcal{L}_{\text{mse}} + \lambda\,\mathcal{L}_{\text{edge}}, \quad \lambda = 0.1 \tag{2}$$

2.2 Stage 2: Landmark-Conditioned Cephalogram Synthesis

The predicted coordinates $\hat{L}_{\text{post}}$ are rendered into a 512×512 color-coded anatomical contour map, where distinct structures (skeletal contours, soft-tissue profiles, and dental boundaries) are assigned specific colors. Non-dental structures are rendered using centripetal Catmull-Rom splines [2] for smooth interpolation, while dental structures are rendered using fixed templates. This contour map serves as both an explicit structural constraint for synthesis and an intermediate output for clinical review.

Composite Conditioning. The anatomical contour provides an explicit geometric boundary prior c_{contour}. To supply texture context for untreated regions, we apply Gaussian blurring to the treatment-affected areas in the pre-treatment cephalogram I_{pre}, yielding a contextual prior c_{img} that retains the patient's native anatomy outside the treatment zone. The final conditioning signal is $c = [c_{\text{contour}}, c_{\text{img}}]$.

Masked Flow Matching. Given c and a binary mask M indicating the treatment region, a U-Net backbone v_θ takes the concatenation of intermediate state x_t and condition c as input, predicting the velocity field from noise $x_0 \sim \mathcal{N}(0, \mathbf{I})$ to the target I_{post}. The mask M is automatically derived as the axis-aligned bounding box of all landmarks belonging to the treatment-affected structures (teeth, mandible, and lower soft tissue), dilated by a fixed margin to accommodate soft-tissue deformation. Restricting the loss to the masked region ensures that the model focuses on synthesizing treatment-induced changes while preserving unaffected anatomy:

$$\mathcal{L}_{\text{stage2}} = \mathbb{E}_{t,x_0,I_{\text{post}}}\left[\left\| M \odot \Big(v_\theta([x_t, c], t) - (I_{\text{post}} - x_0) \Big) \right\|_2^2\right] \tag{3}$$

3 Experiments

3.1 Dataset and Setup

We construct a large-scale orthodontic dataset comprising 44,809 lateral cephalograms from 17,622 patients. For Stage 1, we select 8,000 adult patients (age >

19) whose treatment exceeded two years, forming pre-/post-treatment pairs. Treatment records are extracted from electronic medical records (EMRs) and summarized into concise, standardized textual plans using Qwen [19], prompted to extract key clinical decisions (e.g., extraction strategy, appliance type, and anchorage protocol) while discarding irrelevant narrative; the extracted plans were subsequently reviewed by an orthodontist on a random 5% subset, confirming >95% factual accuracy. Because pre- and post-treatment cephalograms are acquired 1–3 years apart, differences in imaging equipment and patient positioning across sessions can introduce spatial misalignment unrelated to treatment. Alignment errors are therefore computed based on the stable cranial base, and the 10% of pairs with the largest residual registration error are discarded as unreliable. The remaining pairs are split 9:1 into training and test sets. For Stage 2, the test set contains all images from Stage 1 test individuals; the remaining 40,000+ images form the training set.

In Stage 1, the model is trained with batch size 32 and learning rate 1×10^{-4}. In Stage 2, the synthesis model operates at 512^2 resolution with batch size 4 and learning rate 5×10^{-5}. All experiments run on one NVIDIA A100 80GB GPU.

3.2 Evaluation Metrics

For Stage 1, we report Upper/Lower Soft-tissue error (mm) and Upper/Lower Incisor Angle Error (°). For Stage 2, we adopt FID and Precision. We additionally conduct a double-blinded clinical evaluation: two senior orthodontists (>10 years experience) and one resident independently assessed 100 randomly sampled image sets, where real and generated cephalograms were presented in randomized order without source labels. Each rater scored (1) *Image Quality* (radiographic realism and absence of artifacts) and (2) *Prediction Accuracy* (clinical plausibility relative to the pre-treatment input and treatment plan) on a 0–10 scale. Inter-rater agreement (ICC) was 0.82 and 0.79 respectively, indicating good to excellent agreement.

3.3 Results

Table 1 compares our landmark prediction against a large residual MLP [11], a Transformer-based diffusion model (DiT) [10,12], and a vanilla Flow Matching network [8]. The MLP baseline produces the highest errors, confirming that one-step regression struggles with the inherent multi-modality of treatment outcomes. Generative frameworks (DiT and Flow Matching) substantially improve soft-tissue prediction through iterative refinement but remain limited in dental accuracy, as they lack explicit anatomical topology modeling. OrthoFlow achieves the best overall performance: the lowest Upper soft-tissue error (1.32 mm) and the best Upper/Lower Incisor Angle errors (3.90°/3.74°), demonstrating that combining graph-based topology with text-conditioned flow matching effectively balances soft-tissue and dental prediction.

We compare the synthesis module against Pix2PixHD [16], UGATIT [7], and conditional DDPM (Palette) [13]. As shown in Table 2, OrthoFlow achieves the

Table 1. Quantitative comparison and ablation study for landmark prediction. Bold values indicate the best result; lower is better.

Method	S-up (mm)	S-low (mm)	UT-angle (°)	LT-angle (°)
MLP [11]	2.32 ± 0.89	2.83 ± 1.03	4.27 ± 3.13	4.03 ± 3.08
DiT [10,12]	1.44 ± 0.63	1.80 ± 0.77	4.51 ± 3.36	4.16 ± 2.97
Flow Matching [8]	1.37 ± 0.60	$\mathbf{1.69 \pm 0.74}$	4.39 ± 3.27	4.12 ± 3.02
w/o Flow	1.34 ± 0.65	1.80 ± 0.81	3.96 ± 2.96	$\mathbf{3.66 \pm 2.74}$
w/o Text	1.38 ± 0.64	1.76 ± 0.84	4.08 ± 2.98	3.86 ± 2.93
w/o Graph	1.49 ± 0.73	1.82 ± 0.78	4.09 ± 2.96	4.11 ± 3.11
OrthoFlow (Ours)	$\mathbf{1.32 \pm 0.64}$	1.74 ± 0.87	$\mathbf{3.90 \pm 2.82}$	3.74 ± 2.81

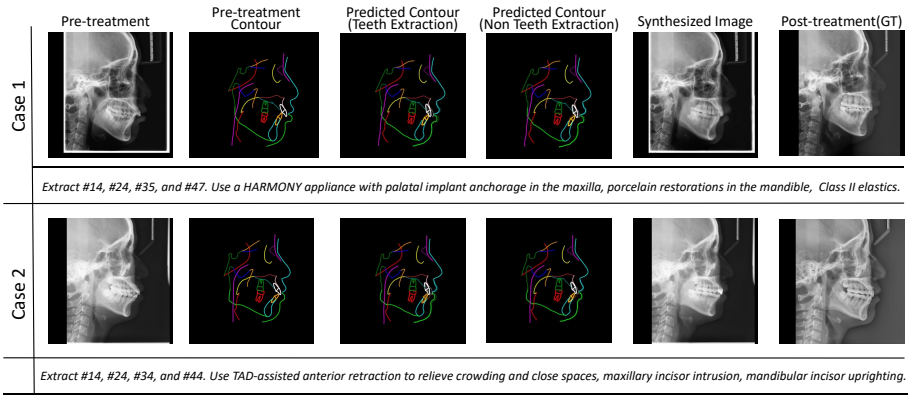

Fig. 2. Workflow of OrthoFlow. Predicted contours and generated cephalograms are shown. Results with and without teeth extraction demonstrate plan-specific prediction.

best FID (22.89) and Precision (0.51). In clinical evaluation, our method receives the highest scores for image quality (8.50) and prediction accuracy (7.02), outperforming all baselines by a substantial margin. Fig. 2 shows representative predictions, and Fig. 3 provides cross-method comparisons confirming that OrthoFlow produces sharper anatomical boundaries and more plausible results. We attribute this advantage to the decoupled design: because Stage 2 receives explicit geometric constraints from Stage 1, it can focus on texture synthesis without the confounding task of simultaneously inferring spatial deformations.

3.4 Ablation Studies

The lower section of Table 1 ablates the Stage 1 module. Removing text conditioning (w/o Text) increases the Upper Incisor Angle error from 3.90° to 4.08°, confirming that treatment plans provide critical guidance for fine-grained dental movements where different clinical strategies (e.g., extraction vs. non-extraction) dictate distinct displacement patterns. Removing the graph module (w/o Graph)

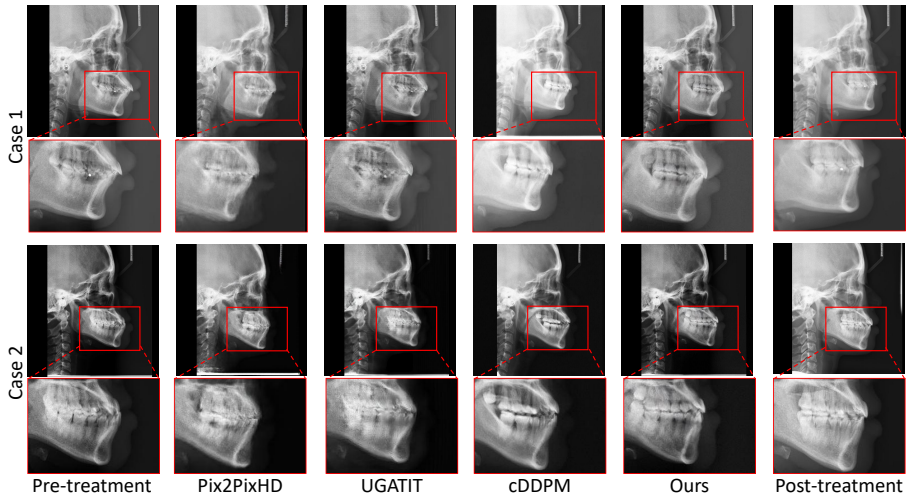

Fig. 3. Qualitative comparison of synthesized post-treatment cephalograms across methods. The red bounding boxes highlight the same regions.

Table 2. Quantitative and clinical comparison of synthesis methods.

Method	Quantitative		Clinical Evaluation	
	FID ↓	Prec. ↑	Img. Qual. ↑	Pred. Acc. ↑
Pix2PixHD [16]	50.21	0.12	3.65	4.22
UGATIT [7]	33.42	0.46	6.85	5.89
cDDPM [13]	31.38	0.41	7.09	5.56
OrthoFlow (Ours)	**22.89**	**0.51**	**8.50**	**7.02**

degrades all metrics, most notably Upper Soft-tissue error (1.32 mm → 1.49 mm), demonstrating that explicitly modeling anatomical adjacency is essential for coherent soft-tissue deformation. Converting to direct one-step regression (w/o Flow) achieves competitive dental angle errors but degrades soft-tissue prediction, confirming that iterative flow-based generation suppresses local point drift and maintains anatomically smooth contours.

4 Conclusion

We presented OrthoFlow, a two-stage decoupled framework for predicting post-orthodontic lateral cephalograms. The core insight is that separating geometric landmark prediction from image synthesis allows each stage to leverage its optimal inductive bias: structured flow matching with anatomical graph priors for geometry, and conditional generation with explicit structural constraints for appearance. Experiments validate that OrthoFlow outperforms existing approaches

across quantitative metrics and clinical evaluations by orthodontists. We hope this framework serves as a reproducible baseline for the orthodontic outcome prediction community. Future work will explore extension to 3D cephalometric prediction and longitudinal treatment monitoring.

Acknowledgments. The work was supported by the Capital's Funds for Health Improvement and Research under Grant CFH2024-1-4101, the National Natural Science Foundation of China (NSFC) under Grant Nos. 6230012077 and 62531024, the Shanghai Municipal Central Guided Local Science and Technology Development Fund under Project No. YDZX20233100001001, the Beijing Natural Science Foundation under Grant No. L242133, and the HPC Platform of ShanghaiTech University.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Auconi, P., Scazzocchio, M., Cozza, P., McNamara Jr, J.A., Franchi, L.: Prediction of class iii treatment outcomes through orthodontic data mining. European journal of orthodontics **37**(3), 257–267 (2015)
2. DeRose, T.D., Barsky, B.A.: Geometric continuity, shape parameters, and geometric constructions for catmull-rom splines. ACM Transactions on Graphics (TOG) **7**(1), 1–41 (1988)
3. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers). pp. 4171–4186 (2019)
4. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
5. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)
6. Holdaway, R.A.: A soft-tissue cephalometric analysis and its use in orthodontic treatment planning. part i. American journal of orthodontics **84**(1), 1–28 (1983)
7. Kim, J., Kim, M., Kang, H., Lee, K.: U-gat-it: Unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation. arXiv preprint arXiv:1907.10830 (2019)
8. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. arXiv preprint arXiv:2210.02747 (2022)
9. Liu, X., Gong, C., Liu, Q.: Flow straight and fast: Learning to generate and transfer data with rectified flow. arXiv preprint arXiv:2209.03003 (2022)
10. Luo, S., Hu, W.: Diffusion probabilistic models for 3d point cloud generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2837–2845 (2021)
11. Ma, X., Qin, C., You, H., Ran, H., Fu, Y.: Rethinking network design and local geometry in point cloud: A simple residual mlp framework. arXiv preprint arXiv:2202.07123 (2022)
12. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4195–4205 (2023)

13. 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)
14. Toepel-Sievers, C., Fischer-Brandies, H.: Validity of the computer-assisted cephalometric growth prognosis vto (visual treatment objective) according to ricketts. Journal of Orofacial Orthopedics/Fortschritte der Kieferorthopädie **60**(3), 185–194 (1999)
15. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
16. Wang, T.C., Liu, M.Y., Zhu, J.Y., Tao, A., Kautz, J., Catanzaro, B.: High-resolution image synthesis and semantic manipulation with conditional gans. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8798–8807 (2018)
17. Wang, Y., Sun, Y., Liu, Z., Sarma, S.E., Bronstein, M.M., Solomon, J.M.: Dynamic graph cnn for learning on point clouds. ACM Transactions on Graphics (tog) **38**(5), 1–12 (2019)
18. Wu, H., Wang, C., Mei, L., Yang, T., Zhu, M., Shen, D., Cui, Z.: Cephalometric landmark detection across ages with prototypical network. In: International conference on medical image computing and computer-assisted intervention. pp. 155–165. Springer (2024)
19. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025)
20. Zhang, X., Mei, L., Yan, X., Wei, J., Li, Y., Li, H., Li, Z., Zheng, W., Li, Y.: Accuracy of computer-aided prediction in soft tissue changes after orthodontic treatment. American Journal of Orthodontics and Dentofacial Orthopedics **156**(6), 823–831 (2019)
21. Zhou, J., Cui, G., Hu, S., Zhang, Z., Yang, C., Liu, Z., Wang, L., Li, C., Sun, M.: Graph neural networks: A review of methods and applications. AI open **1**, 57–81 (2020)
22. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision. pp. 2223–2232 (2017)