Facial Appearance Prediction with Conditional Multi-scale Autoregressive Modeling for Orthognathic Surgical Planning

Jungwook Lee[1 [0000−0001−9483−1224]], Xuanang Xu[1], Daeseung Kim[2], Tianshu Kuang[2], Hannah H. Deng[2], Xinrui Song[1], Yasmine Soubra[2], Rohan Dharia[4], Michael A.K. Liebschner[3], Jaime Gateno[2], and Pingkun Yan[1 [0000−0002−9779−2141]]

[1] Department of Biomedical Engineering and Center for Biotechnology and Interdisciplinary Studies, Rensselaer Polytechnic Institute, Troy, NY 12180, US
yanp2@rpi.edu
[2] Department of Oral and Maxillofacial Surgery, Houston Methodist Research Institute, Houston, TX, 77030, US
[3] Department of Neurosurgey, Baylor College of Medicine, Houston, TX 77030, US
[4] Department of Biochemistry and Cell Biology, Rice University, Houston, TX 77005, US

Abstract. Craniomaxillofacial deformities often necessitate orthognathic surgery to correct jaw positions and improve both function and aesthetics. The existing patient-specific optimal face prediction for soft-tissue-driven planning struggles to accurately capture fine facial details and maintain harmonious alignment among key facial features. In this paper, we propose a novel Conditional Autoregressive Modeling for Orthognathic Surgery (CAMOS) framework that directly predicts patients' optimal 3D face from their preoperative appearance. Our approach employs a hierarchical, coarse-to-fine next-scale prediction strategy, beginning with large-scale pretraining on 44,602 control faces to construct a robust generative model that captures diverse demographic features. Subsequently, the model is fine-tuned on an in-house dataset of 86 orthognathic surgery patients, establishing a conditional path that integrates patient-specific information to form a conditional generative model. Evaluation on both public and in-house datasets demonstrates that CAMOS successfully generates patient-specific optimal face with high quality, effectively addressing the limitations of prior single-step approaches. Source code is available at https://github.com/RPIDIAL/CAMOS.

Keywords: Orthognathic Surgery · Visual AutoRegressive Modeling · Facial Landmarks · Conditional Generation

1 Introduction

Craniomaxillofacial (CMF) deformities involve congenital and acquired abnormalities affecting the skull, face, and jaw, often requiring orthognathic surgery

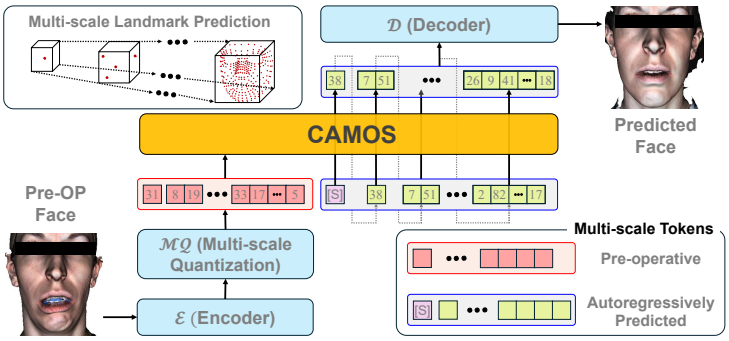

Fig. 1. Application of CAMOS to predict a patient-specific optimal 3D face from a given preoperaitve 3D face with deformity.

to reposition bones and improve functional and aesthetic outcomes [6, 16, 18]. Traditional planning workflows rely on bone-driven simulations, where experts determine bone movements and computational models predict the resultant soft tissue changes [1, 3, 5, 8, 13]. When the predicted face is unsatisfactory, surgeons must iteratively adjust the bone plan, which can be time-consuming. In response, researchers have explored soft-tissue-driven approaches that directly estimate the patient's optimal face without requiring explicit bone planning, avoiding repeated bone-planning adjustments. However, most existing work is limited to 2D lateral-view predictions [7] or using only a sparse set of anatomical landmarks [11], thereby failing to capture the full 3D facial information.

In this paper, we introduce an innovative approach of Conditional Autoregressive Modeling for Orthognathic Surgery (CAMOS), designed to predict a patient's optimal 3D face directly from their preoperative deformed face. Other existing methods, such as [10, 11], can yield unrealistic results when predicting fine facial details (*e.g.*, the lips). In addition, those methods often fail to achieve a harmonious alignment among the lips, jaw, and overall facial structure. The root of the challenge lies in the inherent nature of patient-specific optimal face prediction, which requires both a thorough understanding of the preoperative face and the generation of a normal-looking face. When humans perceive or imagine a face, we tend to first grasp its overall structure and then refine the finer details [4]. However, existing methods predominantly rely on a single-step prediction, overlooking this natural hierarchical progression. To overcome this limitation, we propose a hierarchical multi-scale prediction strategy. By decomposing facial features into a coarse-to-fine hierarchy, we first capture the global structure and then progressively refine local details. This stepwise, multi-scale process mirrors the natural human perceptual process and ensures that each level of facial detail is accurately aligned and realistically generated.

To implement this framework, we adapted the state-of-the-art Visual AutoRegressive modeling (VAR) [19] from image generation to develop CAMOS, which progressively predicts a normal-looking face from facial landmarks of a pre-

operative deformed face as conditional input. However, directly training CAMOS requires a large amount of data with facial deformities, which is infeasible to obtain due to patient privacy concerns. To overcome this roadblock, our second key contribution lies in curating a large-scale dataset of 44,602 control subjects to represent a diverse range of normal-looking facial appearances to pretrain the model. With pretraining on this large-scale dataset followed by finetuning on a paired pre- and post-operative facial appearance dataset collected from 86 patients, we successfully trained the proposed CAMOS. Its clinical application for generating optimal face is illustrated in **Figure 1**.

For training, we first trained a multi-scale Vector Quantized Variational Autoencoder (VQ-VAE) [20] to extract latent features for each facial landmark and subsequently froze it to train a VAR generative model through next-scale token prediction. Finally, using our in-house paired pre- and post-operative dataset, we finetuned the conditional path of the generative model for patient-specific optimal facial appearance prediction. While our ultimate vision involves integrating surgical constraints to ensure the surgical feasibility of predictions, this study explicitly focuses on modeling facial outcomes as an independent yet informative step that can potentially support downstream surgical planning.

We first evaluated the pretraining performance of the CAMOS framework on public datasets and observed strong performance across multiple point-cloud generative metrics, indicating both high fidelity and diversity of the generated faces. We then evaluated CAMOS on a cohort of 86 orthognathic surgery patients, achieving superior performance in surface distance compared to other methods, particularly in the lips and jaw regions that are critical in orthognathic surgery. Moreover, qualitative evaluation confirmed that our approach generates realistic optimal faces, preserving patient-specific features while correcting preoperative deformations.

2 Datasets and Data Processing

Our study leverages both publicly available datasets and in-house patient data. The public datasets consist of 44,602 surface meshes captured using a variety of 3D scanning devices. These data were acquired from four publicly available sources, including DAD-3DHeads (N=42,152) [14], Headspace (N=1,507) [2], FaceScape (N=843) [26], and BU-3DFE (N=100) [27]. In addition, we collected our own data from 86 orthognathic surgery patients (IRB:MOD00005116). Among these, 42 cases were obtained from computed tomography (CT) scans. For these CT scans, soft tissue segmentation was performed, followed by application of the marching cubes algorithm and ambient occlusion techniques to reconstruct facial surface meshes. The remaining 44 cases in our in-house dataset were acquired using 3D stereophotogrammetry cameras, directly providing surface meshes. To focus on information related to orthognathic surgery, we extracted 282 evenly distributed dense landmarks on the facial surface meshes. Specifically, we rendered multiple 2D views of each 3D face from different angles and applied MediaPipe's face landmark detection model [12] to these 2D im-

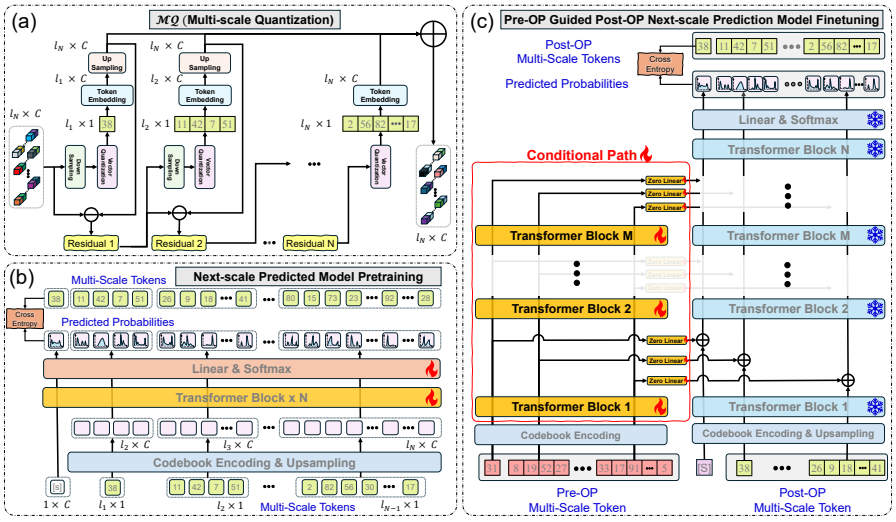

Fig. 2. Detailed architectures of networks within CAMOS. (a) $\mathcal{MQ}$ (Multi-scale Quantization) for extracting discrete tokens at multiple scales; (b) VAR network for predicting next-scale tokens; (c) Finetuning with conditional path.

ages. We then back-projected the detected landmarks onto the 3D surface using known camera parameters and ray casting [24]. By averaging the back-projected results from all views, we obtained robust and consistent 3D landmarks. We excluded landmarks around the eyes, eyebrows, and nostrils to eliminate irrelevant or noisy regions, focusing on the parts of the face most critical for surgical planning.

3 Pretraining of CAMOS

3.1 Multi-Scale Residual Token Extraction

In the hierarchical next-scale prediction approach using VAR, which involves autoregressive prediction of tokens from coarse to fine scales, multi-scale discrete tokenization is required to convert continuous facial landmark coordinates into a hierarchical set of discrete tokens. To achieve this, we train a multi-scale VQ-VAE model comprising an encoder $\mathcal{E}$, a decoder $\mathcal{D}$, and a multi-scale quantization module $\mathcal{MQ}$. Both $\mathcal{E}$ and $\mathcal{D}$ are transformers [21] in this work. The encoder $\mathcal{E}$ first embeds each landmark's coordinate (x, y, z) into a single token and then encodes them into d-dimensional latent feature vectors $\mathbf{z}$. The decoder $\mathcal{D}$ reconstructs the original coordinates from $\mathbf{z}$. **Figure 2-a** shows the details of the multi-scale quantization module $\mathcal{MQ}$. We quantize 282 facial landmarks at N scales, following a quadratic progression to ensure a smooth and gradual refinement of the facial representation from a sparse set of landmarks at the coarsest scale to the finest scale. Let l_n denote the number of landmarks at the nth scale, where

$n = 1, 2, ..., N$. We have $l_n = l_1 + (282 - l_1)\frac{(n-1)^2}{(N-1)^2}$ (with $l_1 = 1$). We then apply residual tokenization [9] across these scales, sharing a single codebook at each level to facilitate hierarchical next-scale prediction. Concretely, we first downsample the latent feature vectors $\mathbf{z}$ to the lowest scale l_1 using the farthest point sampling approach in PointNet [17]. The vectors are then mapped to the nearest code vector in a codebook via a vector quantization layer. We then upsample this quantized feature back to the full scale and calculate the residual vector. The resulting residual is passed on to the next scale. This process is repeated until the full scale is reached, yielding residual tokens at each scale. These multi-scale tokens are then used by the subsequent generative model. The upsampled full-scale features obtained at each scale are summed and passed to $\mathcal{D}$ for reconstruction.

The overall training objective $\mathcal{L}$ of our proposed multi-scale VQ-VAE is:

$$\mathcal{L} = \|x - \hat{x}\|_2^2 + \frac{1}{N}\sum_{i=1}^{N} \|\text{sg}[S_i(x)] - z_q\|_2^2 + \beta\frac{1}{N}\sum_{i=1}^{N} \|\text{sg}[z_q] - S_i(x)\|_2^2, \quad (1)$$

where $\hat{x} = \mathcal{D}(\mathcal{MQ}(\mathcal{E}(x)))$ is the reconstructed output, and $S_i(x)$ is the residual quantized token at the i-th scale and z_q is the codebook vector closest to $S_i(x)$. β is a weighting coefficient, and $\text{sg}[\cdot]$ indicates the stop-gradient operation.

3.2 Next-Scale Token Prediction

Next, we utilize the multi-scale discrete tokens obtained earlier to progressively predict facial landmarks from sparse to dense, thereby generating high-quality, realistic faces. This hierarchical next-scale prediction is implemented using a VAR as shown in **Figure 2b**. To reflect diverse demographics, we employed a large-scale public dataset of control subjects for pretraining. Since these datasets contain only a single neutral face per subject and are not related to orthognathic surgery, we used them to pretrain an unconditional generative model. Given tokens at one scale, the model autoregressively predicts tokens at the next finer scale, similar to GPT-style language modeling except that the tokens at one scale are predicted in a batch as in VAR [19]. Concretely, each multi-scale discrete token is first mapped to its corresponding vector from the pretrained VQ-VAE codebook V, then passed through an upsampling layer to match the number of next-scale tokens, and finally fed into a transformer. The output of transformer is then passed through a softmax layer to produce a probability distribution, and the transformer is trained with cross-entropy loss between this probability distribution and the ground-truth next-scale discrete tokens.

4 Finetuning CAMOS for Conditional Generation

Finally, we incorporate preoperative information to generate patient-specific optimal faces as shown in **Figure 2c**. To accomplish this, we finetune previously pretrained generative model, ensuring that the knowledge gained from a large

and diverse demographic population remains intact. Specifically, we divide the model into two paths: a generative path and a conditional path. The generative path preserves the ability to produce high-quality, normal-looking faces learned from the large-scale public dataset, whereas the conditional path injects patient-specific information from the preoperative face. To leverage the high-quality generation learned during pretraining, we freeze the generative path. We initialize the transformer blocks in the conditional path with the same pretrained weights from VAR, but allow them to be updated during finetuning. Inspired by ControlNet [28], we fuse the outputs of the corresponding Transformer blocks from the generative and the conditional path by adding tokens of the same scale and landmark location. Additionally, we insert a zero-initialized linear layer before the fusion step to gradually inject conditional information. During finetuning, we employ autoregressive modeling conditioned on the preoperative tokens to guide the generation process. The generation process is expressed as: $p(x_1, x_2, \ldots, x_N) = \prod_{n=1}^{N} p\big(x_n \mid \{c_i\}_{i=1}^{N}, x_1, x_2, \ldots, x_{n-1}\big)$, where $x_n, c_n \in [V]$ are tokens at the n-th scale of postoperative and preoperative landmarks, respectively. Each token in x_n and c_n is an index from the VQ-VAE codebook V, which was trained and shared across all scales.

Surface Reconstruction from Facial Dense Landmark In clinical settings, a full 3D facial surface is required rather than a set of landmarks even though they are dense. Therefore, after finetuning the network, we first predict the patient-specific optimal facial landmarks from the patient's preoperative landmarks. We then apply Thin-Plate Spline (TPS) interpolation between the preoperative and predicted landmarks to generate a deformation field, which is applied to the preoperative facial surface. This process yields a complete 3D facial surface for surgical planning and visualization.

5 Experiments and Results

5.1 Evaluation Metrics

We evaluated our pretrained model on a large-scale dataset of control subjects using four point-cloud generative model metrics, including Minimum Matching Distance (MMD), Coverage score (COV), 1-Nearest Neighbor Accuracy (1-NNA), and Jensen-Shannon Divergence (JSD). MMD measures average proximity to real samples, COV and 1-NNA quantify coverage and distributional similarity, and JSD evaluates overlap of marginal distributions [25]. In addition, we also assessed our finetuned model on in-house patient data. For the quantitative evaluation, we measured the Chamfer Distance (CD) and Hausdorff Distance (HD) between the predicted and the actual postoperative facial surface. For more detailed analysis, we divided the face into four regions relevant to orthognathic surgery (nose, lips, cheeks, and chin) and calculated these metrics for each region separately. To assess the statistical significance of the observed differences, we performed a Wilcoxon signed-rank test [22]. For the qualitative assessment, we randomly selected one patient from three jaw deformity type (asymmetry,

Table 1. Performance of the facial landmarks generative models pretrained with large-scale dataset. ↑/↓: higher/lower is better. Best results are in bold.

	MMD (↓)	COV (↑)	1-NN (↓)	JSD×10^{-3} (↓)
DiT (Diffusion) [15]	**0.3252**	0.3818	0.6624	1.0063
CAMOS (VAR)	0.3351	**0.4612**	**0.6412**	**0.6360**

Table 2. Comparison of surface distance between different prediction methods. Best results are in bold; asterisks indicate p-values (<0.05) compared to the best result.

	Chamfer Distance (mm)				Hausdorff Distance (mm)			
	Nose	Lips	Cheeks	Chin	Nose	Lips	Cheeks	Chin
SR [23]	2.52 ± 1.16*	4.31 ± 2.10*	3.49 ± 0.97*	5.58 ± 2.72*	4.76 ± 2.30*	6.74 ± 2.62*	6.97 ± 2.40*	8.62 ± 3.72*
AnaLand [11]	1.75 ± 0.51*	2.24 ± 0.78*	2.63 ± 0.54*	3.66 ± 1.53*	3.29 ± 1.12*	4.03 ± 1.33*	4.74 ± 1.52*	6.32 ± 2.39*
PCNet [10]	1.53 ± 0.38	2.17 ± 0.75*	2.59 ± 0.62*	3.06 ± 1.20*	2.71 ± 0.80	3.91 ± 1.36*	4.65 ± 1.35*	5.38 ± 2.19*
CAMOS	$\mathbf{1.47 \pm 0.29}$	$\mathbf{1.91 \pm 0.55}$	$\mathbf{2.32 \pm 0.33}$	$\mathbf{2.37 \pm 0.74}$	$\mathbf{2.57 \pm 0.57}$	$\mathbf{3.29 \pm 1.06}$	$\mathbf{4.03 \pm 0.83}$	$\mathbf{3.76 \pm 1.12}$

retruded, and protruded) for visualization and compared the facial surfaces generated by each method.

5.2 Implementation Details and Results

For $\mathcal{E}$ and $\mathcal{D}$, we used a transformer with 12 layers, a hidden dimensional size $d = 192$, and 3 attention heads. The codebook size V was set to 256. We empirically determined the scale N=8 and the number of landmarks ({1, 7, 24, 53, 93, 144, 207, 282}) at each scale in VAR (see details below). During finetuning, we stacked 10 transformer blocks to form the conditional path. For better reproducibility, we release our source code at https://github.com/RPIDIAL/CAMOS.

Generation Performance: As shown in **Table 1**, our hierarchical multi-scale approach using VAR outperformed the large-scale pretrained facial landmarks generative model. Compared to the diffusion-based DiT [15], which achieved a lower MMD, our model demonstrated superior COV and 1-NNA scores. This suggests that our coarse-to-fine hierarchical multi-scale approach captures diverse demographics more effectively, reinforcing our claim that coarse-to-fine prediction enhances the realism and harmony of generated faces.

Face Prediction Performance: As shown in **Table 2**, when evaluating optimal face predictions on our in-house patient dataset, the fine-tuned CAMOS model demonstrated superior performance, particularly in regions critical for orthognathic surgery, such as the lips and chin. Qualitative results in **Figure 3** illustrates that CAMOS generates faces that are both high-quality and closely resembles the postoperative outcome. In contrast, other methods either fail to retain patient-specific details or produce faces with residual deformity along with low surface quality. These results confirm that even after finetuning, CAMOS can maintain high quality generation and effectively preserve patient information through hierarchical multi-scale prediction. The color-coded surface distance maps further highlight the superior performance of CAMOS.

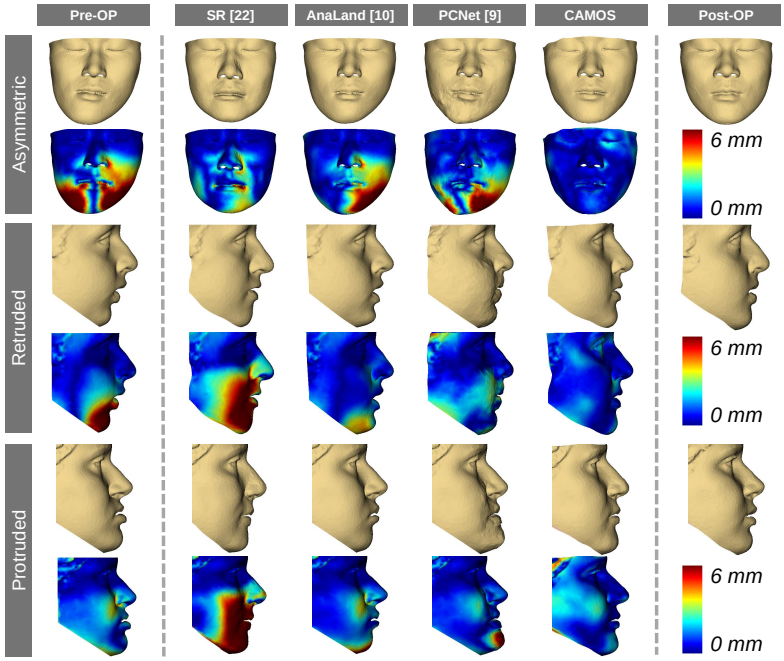

Fig. 3. Visualizations for qualitative evaluation comparing different approaches against our proposed CAMOS framework.

Scale as a hyperparameter: We also conducted experiments to analyze the impact of key hyperparameters on multi-scale prediction. As shown in **Table 3**, our experiments showed that a codebook size of 256 outperformed the settings of 128 and 512. Similarily, dividing the landmarks into 8 scales produced better results than using 6 or 10 scales. The codebook size represents the number of discrete class labels and the number of scales corresponds to the number of prediction steps. Although increasing these values can capture a broader range of features, an excessive number may make the model overly complex.

6 Conclusions

This paper presents the CAMOS framework for patient-specific optimal face prediction in orthognathic surgery. By adopting a hierarchical, coarse-to-fine prediction with large-scale pretraining, CAMOS overcomes previous limitations in capturing fine facial details and overall harmony. Evaluations on both large-scale public dataset and an in-house clinical cohort showed that CAMOS consistently achieves superior generative performance while delivering accurate postoperative predictions. While this study focuses exclusively on predicting patients' optimal facial outcomes, the results can inform subsequent stages of surgical planning, including 1) the estimation of required bone movements and 2) the assessment

Table 3. Ablation study on the impact of the codebook size and scale number. Best results are in bold; asterisks indicate p-values (<0.05) compared to the best result.

Number of Parameters		Chamfer Distance (mm)				
Codebook	Scale	Nose	Lips	Cheeks	Chin	Lower Face
256	6	1.48±0.26	1.87±0.48	2.36±0.37	2.47±1.13	2.12±0.36
128	8	1.56±0.32	2.02±0.72*	2.47±0.55	2.75±1.08*	2.25±0.51*
256	**8**	1.52±0.33	**1.82±0.54**	**2.32±0.39**	**2.30±0.78**	**2.07±0.33**
512	8	**1.46±0.29**	1.87±0.44	2.41±0.53	2.85±1.22*	2.19±0.43*
256	10	1.52±0.38	1.91±0.63	2.40±0.51	2.71±1.12*	2.18±0.47*

of surgical feasibility. A complete soft-tissue-driven workflow would additionally involve simulating the postoperative result based on those predicted skeletal adjustments. One limitation of this work is the absence of formal clinical validation; future research will include expert evaluation to assess the practical utility and surgical relevance of the generated outcomes. Taken together, CAMOS establishes a foundation for facial outcome-guided orthognathic planning with potential to support more efficient and patient-specific surgical decision-making.

Acknowledgments. This work was supported by the National Institutes of Health (NIH) under awards R01DE027251 and R01DE021863.

Disclosure of Interests. No competing interests declared.

References

1. Chabanas, M., Marecaux, C., Payan, Y., Boutault, F.: Models for planning and simulation in computer assisted orthognatic surgery. In: Medical Image Computing and Computer-Assisted Intervention—MICCAI 2002: 5th International Conference Tokyo, Japan, September 25–28, 2002 Proceedings, Part II 5. pp. 315–322. Springer (2002)
2. Dai, H., Pears, N., Smith, W., Duncan, C.: Statistical modeling of craniofacial shape and texture. International Journal of Computer Vision **128**(2), 547–571 (2020)
3. Fang, X., Kim, D., Xu, X., Kuang, T., Lampen, N., Lee, J., Deng, H.H., Liebschner, M.A., Xia, J.J., Gateno, J., Yan, P.: Correspondence attention for facial appearance simulation. Medical Image Analysis **93**, 103094 (2024)
4. Goffaux, V., Peters, J., Haubrechts, J., Schiltz, C., Jansma, B., Goebel, R.: From coarse to fine? spatial and temporal dynamics of cortical face processing. Cerebral Cortex **21**(2), 467–476 (2011)
5. Huang, X., He, D., Li, Z., Zhang, X., Wang, X.: Maxillofacial bone movements-aware dual graph convolution approach for postoperative facial appearance prediction. Medical Image Analysis **99**, 103350 (2025)
6. Khechoyan, D.Y.: Orthognathic surgery: general considerations. In: Seminars in plastic surgery. vol. 27, pp. 133–136. Thieme Medical Publishers (2013)

7. Kim, I.H., Kim, J.S., Jeong, J., Park, J.W., Park, K., Cho, J.H., Hong, M., Kang, K.H., Kim, M., Kim, S.J., Kim, Y.J., Sung, S.J., Kim, Y.H., Lim, S.H., Baek, S.H., Kim, N.: Orthognathic surgical planning using graph cnn with dual embedding module: External validations with multi-hospital datasets. Computer Methods and Programs in Biomedicine **242**, 107853 (2023)
8. Lampen, N., Kim, D., Fang, X., Xu, X., Kuang, T., Deng, H.H., Barber, J.C., Gateno, J., Xia, J., Yan, P.: Deep learning for biomechanical modeling of facial tissue deformation in orthognathic surgical planning. International journal of computer assisted radiology and surgery **17**(5), 945–952 (2022)
9. Lee, D., Kim, C., Kim, S., Cho, M., Han, W.S.: Autoregressive image generation using residual quantization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11523–11532 (2022)
10. Lee, J., Kim, D., Xu, X., Fang, X., Kuang, T., Lampen, N., Deng, H.H., Liebschner, M.A., Xia, J.J., Gateno, J., Yan, P.: A feasibility study on estimating desired postoperative face using deep learning for patients with craniomaxillofacial deformities. In: Proceedings of the 37th International Congress and Exhibition on Computer Assisted Radiology and Surgery (CARS 2023). Munich, Germany (June 20–23 2023)
11. Lee, J., Kim, D., Xu, X., Kuang, T., Gateno, J., Yan, P.: Predicting optimal patient-specific postoperative facial landmarks for patients with craniomaxillofacial deformities. International Journal of Oral and Maxillofacial Surgery (2024)
12. Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., Zhang, F., Chang, C.L., Yong, M.G., Lee, J., Chang, W.T., Hua, W., Georg, M., Grundmann, M.: Mediapipe: A framework for building perception pipelines. arXiv preprint arXiv:1906.08172 (2019)
13. Ma, L., Xiao, D., Kim, D., Lian, C., Kuang, T., Liu, Q., Deng, H., Yang, E., Liebschner, M.A., Gateno, J., Xia, J.J., Yap, P.T.: Simulation of postoperative facial appearances via geometric deep learning for efficient orthognathic surgical planning. IEEE transactions on medical imaging **42**(2), 336–345 (2022)
14. Martyniuk, T., Kupyn, O., Kurlyak, Y., Krashenyi, I., Matas, J., Sharmanska, V.: Dad-3dheads: A large-scale dense, accurate and diverse dataset for 3d head alignment from a single image. In: Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition. pp. 20942–20952 (2022)
15. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4195–4205 (2023)
16. Proffit, W., Turvey, T., Phillips, C.: Orthognathic surgery: a hierarchy of stability. The International journal of adult orthodontics and orthognathic surgery **11**(3), 191–204 (1996)
17. Qi, C.R., Su, H., Mo, K., Guibas, L.J.: PointNet: Deep learning on point sets for 3D classification and segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 652–660 (2017)
18. Shafi, M., Ayoub, A., Ju, X., Khambay, B.: The accuracy of three-dimensional prediction planning for the surgical correction of facial deformities using maxilim. International journal of oral and maxillofacial surgery **42**(7), 801–806 (2013)
19. Tian, K., Jiang, Y., Yuan, Z., Peng, B., Wang, L.: Visual autoregressive modeling: Scalable image generation via next-scale prediction. arXiv preprint arXiv:2404.02905 (2024)
20. Van Den Oord, A., Vinyals, O., Kavukcuoglu, K.: Neural discrete representation learning. Advances in neural information processing systems **30** (2017)

21. 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)
22. Wilcoxon, F.: Individual comparisons by ranking methods. In: Breakthroughs in statistics: Methodology and distribution, pp. 196–202. Springer (1992)
23. Xiao, D., Wang, L., Deng, H., Thung, K.H., Zhu, J., Yuan, P., Rodrigues, Y.L., Perez, L., Crecelius, C.E., Gateno, J., Kuang, T., Shen, S.G., Kim, D., Alfi, D.M., Yap, P.T., Xia, J.J., Shen, D.: Estimating reference bony shape model for personalized surgical reconstruction of posttraumatic facial defects. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part V 22. pp. 327–335. Springer (2019)
24. Xu, X., Lee, J., Lampen, N., Kim, D., Kuang, T., Deng, H.H., Liebschner, M.A., Gateno, J., Yan, P.: DiRecT: Diagnosis and reconstruction transformer for mandibular deformity assessment. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 141–151. Springer (2024)
25. Yang, G., Huang, X., Hao, Z., Liu, M.Y., Belongie, S., Hariharan, B.: Pointflow: 3D point cloud generation with continuous normalizing flows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4541–4550 (2019)
26. Yang, H., Zhu, H., Wang, Y., Huang, M., Shen, Q., Yang, R., Cao, X.: Facescape: a large-scale high quality 3D face dataset and detailed riggable 3D face prediction. In: Proceedings of the ieee/cvf conference on computer vision and pattern recognition. pp. 601–610 (2020)
27. Yin, L., Wei, X., Sun, Y., Wang, J., Rosato, M.J.: A 3d facial expression database for facial behavior research. In: 7th international conference on automatic face and gesture recognition (FGR06). pp. 211–216. IEEE (2006)
28. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3836–3847 (2023)