PAINT: Projection-Aware Implicit Neural Transformer for Soft-Tissue Surface Generation in Orthognathic Surgical Planning

Su Yang[1,2,3⋆], Daeseung Kim[1,⋆], Kevin Gu[1], Rohan Dharia[4], Lois Jung[1,5], and Jaime Gateno[1,6,7]

[1] Department of Oral and Maxillofacial Surgery, Houston Methodist Research Institute, 6560 Fannin St, Houston, TX 77030, USA
[2] Department of Medical Informatics, School of Medicine, Kyungpook National University, Daegu, 41944, Republic of Korea
[3] AI-driven Medical Innovation Center, Kyungpook National University Hospital, Daegu, 41944, Republic of Korea
[4] Rice University, Houston, TX 77005, USA
[5] College of Medicine, Hallym University, Chuncheon, 24252, Republic of Korea
[6] Sylvia and James E. Norton Distinguished New Century Chair in Oral and Maxillofacial Surgery, Houston Methodist Hospital, Houston, TX 77807, USA
[7] Clinical Surgery, Weill Cornell Medical College, New York, NY, USA
dkim@houstonmethodist.org

Abstract. Soft-tissue-driven orthognathic surgical planning requires reconstructing anatomically coherent facial soft-tissue volumes from a given target outer facial surface. However, existing implicit neural representation approaches often lack volumetric constraints and fail to preserve anatomical consistency. In this paper, we present a **P**rojection-**A**ware **I**mplicit **N**eural **T**ransformer (**PAINT**) for patient-specific soft-tissue volumetric reconstruction. Given a voxelized outer-surface representation, **PAINT** predicts a soft-tissue implicit grid that reconstructs the enclosed volume conditioned on the input surface. **PAINT** combines a hybrid 3D CNN encoder with transformer layers for global anatomical context, a 3D decoder with triplet skip attention to capture input-conditioned volumetric relationships, and volume-constraint and projection-aware reconstruction losses to enforce local fidelity and global volumetric consistency. Experiments demonstrate the effectiveness of key components of **PAINT** and superior performance over existing 3D networks.

Keywords: Orthognathic Surgical Planning · CT/CBCT · Implicit Neural Representation · Soft-Tissue Generation · Voxel Representation.

1 Introduction

Orthognathic surgery creates critical functional and aesthetic improvements for patients with dentofacial deformities by repositioning skeletal segments [1].

⋆ Su Yang and Daeseung Kim - These authors contributed equally

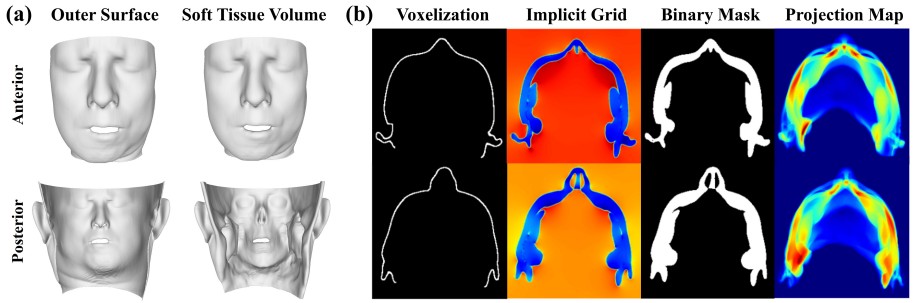

Fig. 1. (a) The outer surface of soft tissue and the soft tissue volume consisting of the superficial layer, masseter muscles, and parotid glands. (b) The voxelization of the outer surface, the implicit grid of the soft-tissue volume, and both the binary mask and the projection map of the implicit grid.

While achieving a desired postoperative facial appearance is the primary determinant of patient satisfaction, conventional surgical planning has predominantly followed a "bone-first" paradigm [1]. This approach focuses on normalizing skeletal structures under the simplified assumption that the overlying soft tissue will passively and linearly conform to bony changes. However, due to the complex non-linear biomechanical properties of facial soft tissue, including muscles, fat pads, and glands, skeletal normalization frequently fails to yield an optimal aesthetic outcome, particularly in patients with severe asymmetry or deformities [2, 3]. Consequently, a paradigm shift toward "soft-tissue-driven" (or face-first) planning has emerged as a reverse-engineering strategy: first defining the target patient-specific postoperative appearance, and then deriving the necessary skeletal movements [4–6]. To realize this inverse planning, it is insufficient to merely predict the target outer-skin surface; one must anatomically reconstruct the facial soft-tissue volume. Crucially, the inner surface of this reconstructed soft tissue volume serves as a patient-specific blueprint, delineating the exact skeletal configuration required to support the desired postoperative appearance [4–6]. However, despite its clinical importance, volumetric reconstruction of the facial soft tissue has not been adequately addressed in current orthognathic surgical-planning pipelines. This gap limits the practical adoption of truly soft-tissue-driven inverse planning. Therefore, a main challenge is to learn a patient-specific volumetric reconstruction that infers the complete enclosed volume from a given target outer facial surface, while maintaining anatomically plausible volume distribution consistent with preoperative tissue characteristics.

Due to the inherent limitations of point-based reconstruction in representing continuous and watertight surfaces, implicit neural representations (INR), such as occupancy fields [7, 8], signed distance functions [13, 9, 10], and indicator grids [11, 16, 12], have emerged for continuous shape modeling in an end-to-end manner. Despite their promising performance, existing INR-based approaches exhibit several limitations when applied to soft-tissue–driven inverse planning. **First**, most INR-based frameworks lack explicit volumetric constraints that enforce global anatomical fidelity and volumetric preservation [11, 16, 10]. In our

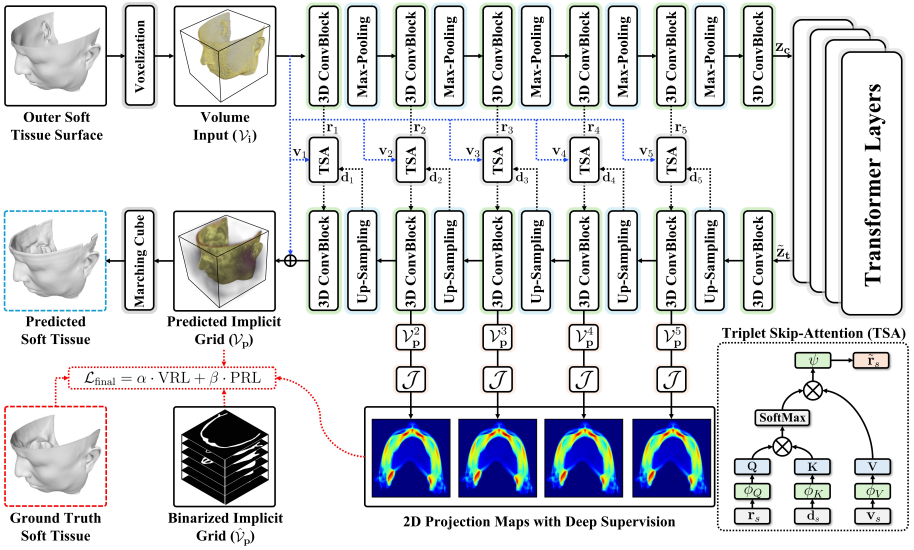

Fig. 2. Illustration of our **PAINT** architecture. **PAINT** takes a voxelized outer-surface representation and reconstructs the soft-tissue volume via the predicted implicit grid. The corresponding surface is extracted for visualization.

setting, the reconstructed soft-tissue volume must simultaneously remain consistent with the given outer facial geometry and maintain anatomically plausible internal structure. Without volume-aware regularization, INR-based approaches [11, 16, 10] may lead to unrealistic soft-tissue thicknesses, local collapse, or irregular boundaries, directly undermining the clinical reliability of inverse planning. **Second**, many INR-based approaches rely on fixed point/mesh sampling strategies or reconstruct the surface from a predetermined number of point clouds [16, 12]. Such sampling strategies may under-represent anatomically complex regions and be sensitive to patient-specific variations in soft-tissue volume and geometric complexity [16]. **Finally**, many INR-based frameworks utilize an implicit grid using voxel-based encoders and decoders built on 3D CNNs or Transformers [11, 16, 12]. While 3D CNNs [20–22] are effective for voxel-to-voxel learning, their inductive bias is primarily local and may miss long-range dependencies. Transformer-based models capture long-range context [14, 15, 23, 24], but naive self-attention is computationally expensive, and standard skip connections do not enforce input-conditioned volumetric consistency.

To address these challenges, we propose a **P**rojection-**A**ware **I**mplicit **N**eural **T**ransformer (**PAINT**) for patient-specific soft-tissue volumetric reconstruction in soft-tissue–driven orthognathic surgical planning. Given a voxelized representation of the target outer facial surface, **PAINT** predicts a soft-tissue implicit grid that reconstructs the enclosed soft-tissue volume while remaining consistent with the input outer geometry. The corresponding soft-tissue surface is extracted from the implicit grid using the Marching Cubes algorithm [17]. **PAINT** consists of three key components. (i) **Hybrid 3D encoder**. To capture both local

anatomical detail and long-range structural dependencies, we employ a CNN-based hierarchical volumetric encoder enhanced with Transformer-based global contextual modeling. (ii) **Triplet skip-attention decoder**. A 3D decoder with triplet skip-attention (TSA) predicts the soft-tissue implicit grid by explicitly modeling the volumetric correspondence between the input outer surface and the reconstructed volume for input-conditioned decoding. (iii) **Volume-constraint and projection-aware losses**. We introduce volume-constraint reconstruction loss (VRL) and projection-aware reconstruction loss (PRL) to enforce voxel-level fidelity, global volumetric consistency, and anatomically coherent volume distribution through deep supervision of intermediate projections. In this study, we validate **PAINT** in a surface-to-volume reconstruction setting, which serves as a necessary step toward full inverse-planning scenarios.

2 Methods

Network overview. The overview of our **PAINT** is shown in Fig. 2. Given the outer surface of a soft tissue (Fig. 1a), we voxelize [11] the outer surface to obtain the voxelized outer-surface volume $\mathcal{V}_{\mathbf{i}} \in \mathbb{R}^{C\times D\times H\times W}$, where C, D, H and W denote the channel, depth, height and width of the input volume (Fig. 1b). Then $\mathcal{V}_{\mathbf{i}}$ is fed to the encoder of **PAINT** to produce high-level feature representations $\tilde{\mathbf{z}}_{\mathbf{t}}$ with 3D CNN and Transformer. The decoder of **PAINT** takes $\tilde{\mathbf{z}}_{\mathbf{t}}$ to generate the soft tissue implicit grid $\mathcal{V}_{\mathbf{p}} \in \mathbb{R}^{C\times D\times H\times W}$ with TSA. After predicting the soft-tissue implicit grid $\mathcal{V}_{\mathbf{p}}$, we extract the corresponding surface using the Marching Cubes algorithm [17] for visualization and evaluation.
3D encoder with Transformer. The encoder of **PAINT** consists of CNN and Transformer layers. The CNN backbone contains five resolution levels implemented via repeated 3D convolution blocks. At each level, we apply two consecutive $3 \times 3 \times 3$ convolutions followed by instance normalization (IN) and leaky rectified linear unit (LReLU) activation (3D ConvBlock in Fig. 2), and then downsample the feature map using a $2 \times 2 \times 2$ max-pooling layer. The CNN backbone outputs an encoded bottleneck feature map $\mathbf{z}_{\mathbf{c}}$, which is fed into the Transformer [14, 15] to capture long-range anatomical dependencies. Given $\mathbf{z}_{\mathbf{c}} \in \mathbb{R}^{C'\times D'\times H'\times W'}$, we perform 3D tokenization by flattening the spatial dimensions into a token sequence $\mathbf{X} = [\mathbf{x}_1; \mathbf{x}_2; \cdots; \mathbf{x}_N] \in \mathbb{R}^{N\times C'}$, where $N = D'H'W'$ and each token $\mathbf{x}_i \in \mathbb{R}^{C'}$ corresponds to a spatial location in the $D' \times H' \times W'$ grid (i.e., $\mathbf{X} = \text{Flatten}(\mathbf{z}_{\mathbf{c}})$) [14, 15]. To retain spatial information, we add a 3D sinusoidal positional encoding $\mathbf{PE}$ [14, 15] to the token sequence, resulting in $\mathbf{z}_0 = \mathbf{X} + \mathbf{PE}, \quad \mathbf{PE} \in \mathbb{R}^{N\times C'}$.

After tokenization and positional encoding, we employ a Transformer [14, 15] to capture long-range dependencies among 3D tokens. Specifically, the token sequence $\mathbf{z}_0 \in \mathbb{R}^{N\times C'}$ is processed by L stacked Transformer blocks, where we set L as 4. Each Transformer block consists of multi-head self-attention (MSA) and a multi-layer perceptron (MLP) [14]. The ℓ-th block is defined as:

$$\mathbf{z}'_\ell = \text{MSA}(\text{LN}(\mathbf{z}_{\ell-1})) + \mathbf{z}_{\ell-1}, \quad \mathbf{z}_\ell = \text{MLP}(\text{LN}(\mathbf{z}'_\ell)) + \mathbf{z}'_\ell, \tag{1}$$

where $\mathrm{LN}(\cdot)$ denotes layer normalization, and $\mathrm{MSA}(\cdot)$ is multi-head self-attention with h heads. $\mathrm{MLP}(\cdot)$ consists of two fully connected layers with a Gaussian error linear unit activation and dropout. Residual connections are applied after each sub-layer. After L blocks, the output tokens are reshaped back to encoded feature maps to initialize the decoding:

$$\tilde{\mathbf{z}}_{\mathbf{t}} = \mathrm{Unflatten}(\mathrm{LN}(\mathbf{z}_L)) \in \mathbb{R}^{C' \times D' \times H' \times W'}. \tag{2}$$

3D decoder with triplet skip-attention. The 3D decoder of **PAINT** progressively upsamples the Transformer-encoded feature maps and restores the target voxel volume at the input resolution. Starting from the Transformer-encoded feature map $\tilde{\mathbf{z}}_{\mathbf{t}} \in \mathbb{R}^{C' \times D' \times H' \times W'}$, we apply a sequence of four upsampling stages. At each stage, the decoder feature is upsampled by a factor of 2 using trilinear interpolation, followed by two $3 \times 3 \times 3$ convolutions with LN and LReLU.

To effectively model the volumetric correspondence between the voxelized outer-surface input $\mathcal{V}_{\mathbf{i}}$ and the reconstructed soft-tissue volume, we introduce a triplet skip-attention (TSA) module at each decoding stage to adaptively inject input-conditioned contextual information into the skip connections. Let $\mathbf{d}_s$ denote the upsampled decoder feature at stage s, and let $\mathbf{r}_s$ be the corresponding encoder feature maps from the CNN encoder. In addition, we resize the voxel input $\mathcal{V}_{\mathbf{i}}$ to the same spatial resolution, obtaining $\mathbf{v}_s$. TSA forms query $\mathbf{Q}_s = \phi_Q(\mathbf{r}_s)$, key $\mathbf{K}_s = \phi_K(\mathbf{d}_s)$, and value $\mathbf{V}_s = \phi_V(\mathbf{v}_s)$, where $\phi_Q(\cdot)$, $\phi_K(\cdot)$, and $\phi_V(\cdot)$ are $1 \times 1 \times 1$ convolutions projecting features into a shared embedding space. Here, $\mathbf{r}_s$ provide anatomical context, $\mathbf{d}_s$ encode reconstruction states, and $\mathbf{v}_s$ inject explicit outer-surface conditioning. We then compute multi-head attention using $\mathbf{Q}_s$, $\mathbf{K}_s$, and $\mathbf{V}_s$ to obtain triplet attention features:

$$\tilde{\mathbf{r}}_s = \psi(\mathrm{MSA}(\mathbf{Q}_s, \mathbf{K}_s, \mathbf{V}_s)), \quad \mathbf{u}_s = \vartheta_s(\mathrm{Concat}(\mathbf{d}_s, \tilde{\mathbf{r}}_s)), \tag{3}$$

where $\psi(\cdot)$ denotes a layer normalization and $1 \times 1 \times 1$ convolution that maps the triplet attention embedding dimension to C_e. The attentive skip feature maps $\tilde{\mathbf{r}}_s$ are concatenated with the upsampled decoder feature and passed through 3D ConvBlock. $\vartheta_s(\cdot)$ denotes 3D ConvBlock consisting of two consecutive $3 \times 3 \times 3$ convolutions, each followed by IN and LReLU. The refined feature maps $\mathbf{u}_s$ are forwarded to the next decoding stage, and the last output layer consists of a $1 \times 1 \times 1$ convolution with hyperbolic tangent activation to predict a soft tissue implicit grid $\mathcal{V}_{\mathbf{p}}$. To explicitly preserve the outer-surface geometry in the reconstructed volume, we add the voxelized outer-surface input $\mathcal{V}_{\mathbf{i}}$ to the predicted implicit grid $\mathcal{V}_{\mathbf{p}}$. This residual addition explicitly enforces hard geometric consistency at the boundary voxels.

Volume-constraint reconstruction loss. To explicitly enforce volumetric consistency, we propose VRL that enforces volumetric consistency between $\mathcal{V}_{\mathbf{p}}$ and the ground truth implicit grid $\mathcal{V}_{\mathbf{g}}$. Specifically, we combine a voxel-wise regression term $\mathcal{L}_{\mathrm{MSE}}$ with a volume-consistency term $\mathcal{L}_{\mathrm{DSC}}$ to jointly constrain occupancy fidelity and volume overlap in the implicit grid:

$$\mathcal{L}_{\mathrm{MSE}} = \|\mathcal{V}_{\mathbf{p}} - \mathcal{V}_{\mathbf{g}}\|_2^2, \quad \mathcal{L}_{\mathrm{DSC}} = 1 - \frac{2\sum \hat{\mathcal{V}}_{\mathbf{p}} \hat{\mathcal{V}}_{\mathbf{g}} + \epsilon}{\sum \hat{\mathcal{V}}_{\mathbf{p}} + \sum \hat{\mathcal{V}}_{\mathbf{g}} + \epsilon}, \tag{4}$$

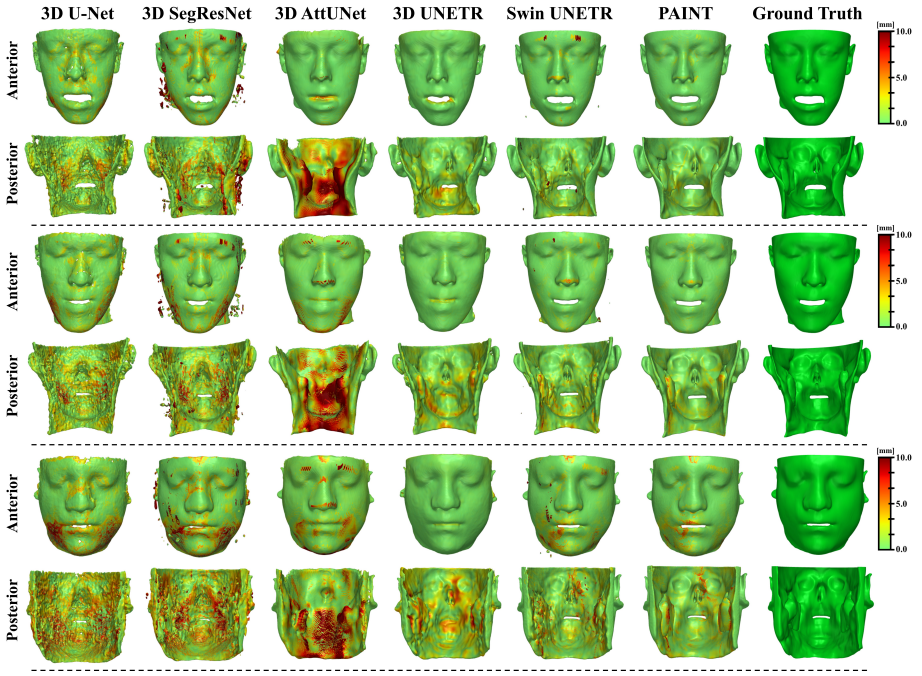

Fig. 3. Visual comparison of different methods for inner surface generation of the soft tissue. The color map denotes surface distance error.

where $\hat{\mathcal{V}}_{\mathbf{p}}$ and $\hat{\mathcal{V}}_{\mathbf{g}}$ are the binarized implicit grids obtained by zero-level thresholding in the implicit grids. ϵ is a small constant for numerical stability. The overall volume-constraint reconstruction loss is defined as $\text{VRL} = \mathcal{L}_{\text{MSE}} + \mathcal{L}_{\text{DSC}}$.

Projection-aware reconstruction loss. To regularize the global volume distribution of the soft tissue in an implicit grid, we introduce PRL with deep supervision. The 3D decoder produces intermediate implicit grids $\{\mathcal{V}_{\mathbf{p}}^s\}_{s\in\{2,3,4,5\}}$ size of C, D, H, and W at multiple decoding stages s with deep supervision [19]. We enforce global consistency by matching the 2D depth-wise mean projection of each intermediate prediction to that of $\mathcal{V}_{\mathbf{g}}$, where the 2D mean projection is defined as $\mathcal{J}(\mathcal{V}) = \frac{1}{D}\sum_{d=1}^{D}\mathcal{V}(:,d,:,:)$. PRL penalizes discrepancy in global volume distribution between the projected $\mathcal{V}_{\mathbf{p}}$ and $\mathcal{V}_{\mathbf{g}}$ with deep supervision at decoding stages:

$$\text{PRL} = \sum_{s\in\{2,3,4,5\}} \left\| \mathcal{J}(\mathcal{V}_{\mathbf{p}}^s) - \mathcal{J}(\mathcal{V}_{\mathbf{g}}) \right\|_2^2 . \tag{5}$$

PRL regularizes global volume distribution via single-axis projection, preventing collapse to average shapes while avoiding excessive volumetric supervision. The final loss $\mathcal{L}_{\text{final}}$ is defined as $\mathcal{L}_{\text{final}} = \alpha \cdot \text{VRL} + \beta \cdot \text{PRL}$, where we heuristically set α and β as 0.7 and 0.3.

Table 1. Performance comparison of **PAINT** with existing 3D deep networks.

Methods	MAE(↓)	R^2(↑)	IoU(↑)	HD(↓)	SDE(↓)	NC(↑)
3D U-Net	77.947±16.324	0.701±0.311	0.384±0.061	8.673±1.841	2.407±0.600	0.876±0.009
3D SegResNet	80.795±11.647	0.369±0.106	0.503±0.050	7.595±1.104	2.122±0.358	0.862±0.017
3D AttUNet	94.562±8.972	0.614±0.057	0.639±0.061	9.637±1.170	2.826±0.295	0.889±0.010
3D UNETR	35.163±8.266	0.836±0.049	0.815±0.032	5.171±0.077	1.453±0.252	0.933±0.018
Swin UNETR	35.714±8.255	0.830±0.042	0.837±0.030	4.892±0.718	1.247±0.223	0.932±0.019
PAINT	**30.120±9.675**	**0.866±0.040**	**0.866±0.025**	**4.576±0.868**	**1.105±0.242**	**0.957±0.014**

Table 2. Ablation study results of key components and loss functions in **PAINT**. TRF, TSA, VRL, and PRL indicate Transformer, triplet skip-attention, volume-constraint reconstruction loss, and projection-aware reconstruction loss, respectively.

(a) Ablation study on key components						
Components	MAE(↓)	R^2(↑)	IoU(↑)	HD(↓)	SDE(↓)	NC(↑)
Baseline	35.734±9.474	0.870±0.039	0.862±0.025	5.052±0.968	1.207±0.271	0.951±0.014
+ TRF	31.067±9.689	**0.872±0.032**	**0.869±0.024**	4.788±0.869	1.133±0.238	0.957±0.015
+ TSA	33.297±9.878	0.846±0.040	0.865±0.026	4.794±0.886	1.139±0.248	0.956±0.015
+ TRF + TSA	**30.120±9.675**	0.866±0.040	0.866±0.025	**4.576±0.868**	**1.105±0.242**	**0.957±0.014**
(b) Ablation study on VRL and PRL						
Baseline	34.302±8.279	0.869±0.037	0.860±0.023	5.074±0.927	1.212±0.240	0.953±0.014
+ VRL	31.956±8.856	0.871±0.039	**0.869±0.023**	4.964±0.959	1.164±0.254	0.957±0.015
+ PRL	31.566±8.799	**0.874±0.038**	0.868±0.023	4.947±0.970	1.152±0.053	0.957±0.014
+ VRL + PRL	**30.120±9.675**	0.866±0.040	0.866±0.025	**4.576±0.868**	**1.105±0.242**	**0.957±0.014**

3 Experiments

3.1 Datasets and Implementation Details

Dataset. We collected a total of 100 scans, including 73 CT scans and 27 CBCT scans acquired from an anonymized dataset (IRB# Pro00009723). The soft tissues and corresponding outer surfaces were annotated by an experienced oral and maxillofacial surgeon using our in-house software and MeshLab [25] (Fig. 1a). The dataset was randomly split into 80 training and 20 test cases. Due to the difficulty of acquiring paired pre- and post-deformation volumetric data, our study adopts a single-surface reconstruction setting, where the outer surface serves as input and the corresponding soft-tissue volume provides supervision. The volume itself is never used as input.

Pre-processing. We rigidly registered each subject to a randomly selected soft-tissue template to mitigate inter-subject variability in head position. Rigid registration was performed using eight anatomical landmarks (left and right tragus, endocanthion, exocanthion, menton, and glabella) [18] to the soft-tissue template. The estimated rigid transform was applied to both the soft-tissue volume (ground truth) and the corresponding outer surface (input). To estimate the ground truth implicit grid of the soft tissue, we apply DPSR [11] to the points and normals of the soft-tissue surface. The outer surface was voxelized to obtain the voxelized outer-surface volume used as network input.

Training setup. All networks were trained using the Adam optimizer for 2000 epochs with an initial learning rate of 10^{-4}. The batch size was set to 2. Training was conducted on an NVIDIA GeForce RTX A6000 GPU (48 GB memory). All models were implemented in Python3 using the PyTorch framework under identical computational environments. The implicit grid resolution was set to

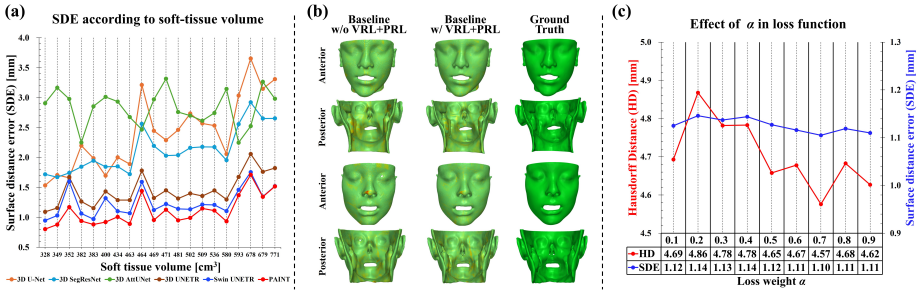

Fig. 4. (a) Surface distance errors according to the soft-tissue volume of patients. (b) Visual comparison of the effectiveness of the volume-constraint and projection-aware reconstruction losses. (c) Ablation study results on the effect of loss weight α in $\mathcal{L}_{\text{final}}$.

256^3 using DPSR [11].

Evaluation metrics. Volumetric reconstruction is evaluated using mean absolute error (MAE) and the coefficient of determination (R^2) between the predicted and ground truth implicit grids [16]. For binary volumes (zero threshold), we compute the intersection-over-union (IoU). Normal consistency (NC) measures surface normal similarity [11], while Surface distance error (SDE) and Hausdorff distance (HD) quantify geometric discrepancy between extracted and ground truth surfaces (mm) [16].

3.2 Experimental Results

Comparison with existing methods. We compare **PAINT** with the five encoder-decoder networks [20–24]. As shown in Table 1, **PAINT** achieves the best performance across all evaluation metrics, outperforming the strongest competing model (UNETR) in MAE (30.120 ± 9.675), R^2 (0.866 ± 0.040), IoU (0.866 ± 0.025), HD (4.576 ± 0.868), SDE (1.105 ± 0.242), and NC (0.957 ± 0.014). These results highlight the benefit of volumetric constraints and input-conditioned decoding. Fig. 3 shows that **PAINT** produces reconstructions closely aligned with the ground-truth, while competing methods exhibit local collapse or irregular boundaries. Fig. 4a further demonstrates consistently low SDE across subjects with varying soft-tissue volumes, indicating robust reconstruction under anatomical variation.

Key component analysis. We conduct ablation studies on TRF, TSA, and reconstruction losses (VRL and PRL) in **PAINT**. As shown in Table 2(a), combining TRF and TSA (i.e., **PAINT**) yields consistent improvements over the baseline and individual components, indicating complementary effects between global contextual modeling and input-conditioned decoding. Table 2(b) shows that combining VRL and PRL improves performance over either loss alone, indicating complementary benefits of local voxel fidelity and global volume constraints (Fig. 4b).

Effect of α and β in loss function. We evaluate the effects of α and β in $\mathcal{L}_{\text{final}}$ (Fig. 4c). Underweighting α (e.g., $\alpha = 0.1$), degrades HD (4.69 ± 0.88) and SDE

(1.12 ± 0.24), while increasing α improves both metrics. The best performance is achieved at $\alpha = 0.7$ and $\beta = 0.3$, (SDE 1.10 ± 0.24, HD 4.57 ± 0.86), which we adopt in all experiments.

4 Conclusion

In this paper, we presented **PAINT** for patient-specific soft-tissue volumetric reconstruction in soft-tissue–driven orthognathic surgical planning. By reconstructing anatomically coherent volumes from a given target outer facial surface, **PAINT** enables volumetric inverse planning beyond surface-level reconstruction. Experimental results demonstrate consistent improvements over state-of-the-art 3D networks. Future work will focus on multi-center validation and integration into unified inverse planning frameworks.

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

Acknowledgements. This study was supported in part by the National Institutes of Health (NIH) under Grant R01DE021863.

References

1. Naran, Sanjay, Derek M. Steinbacher, and Jesse A. Taylor. "Current concepts in orthognathic surgery." Plastic and reconstructive surgery 141.6, 925e-936e, 2018.
2. Xia, James, et al. Three-dimensional virtual-reality surgical planning and soft-tissue prediction for orthognathic surgery. IEEE transactions on information technology in biomedicine 5.2, 97-107, 2001.
3. Ma, Lei, et al. Simulation of postoperative facial appearances via geometric deep learning for efficient orthognathic surgical planning. IEEE transactions on medical imaging 42.2, 336-345, 2022.
4. Lee, Jungwook, et al. Predicting optimal patient-specific postoperative facial landmarks for patients with craniomaxillofacial deformities. International Journal of Oral and Maxillofacial Surgery 53.11, 934-941, 2024.
5. Lee, Jungwook, et al. Facial Appearance Prediction with Conditional Multi-scale Autoregressive Modeling for Orthognathic Surgical Planning. International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham: Springer Nature Switzerland, 2025.
6. Lee, Jungwook, et al. Facial Appearance Prediction for Orthognathic Surgery with Diffusion Models. Medical Image Analysis, 103934, 2026.
7. Mescheder, Lars, et al. Occupancy networks: Learning 3d reconstruction in function space. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019.
8. Peng, Songyou, et al. Convolutional occupancy networks. European Conference on Computer Vision. Cham: Springer International Publishing, 2020.
9. Shen, Xinze, et al. TranSDFNet: Transformer-based truncated signed distance fields for the shape design of removable partial denture clasps. IEEE Journal of Biomedical and Health Informatics 27.10: 4950-4960, 2023.

10. Chen, Hongbo, et al. Neural implicit surface reconstruction of freehand 3D ultrasound volume with geometric constraints. Medical Image Analysis 98: 103305, 2024.
11. Peng, Songyou, et al.: Shape as points: A differentiable Poisson solver. In: Advances in Neural Information Processing Systems, vol. 34, pp. 13032-13044, 2021.
12. Wei, Linda, et al. VBCD: A Voxel-Based Framework for Personalized Dental Crown Design. International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham: Springer Nature Switzerland, 2025.
13. Park, Jeong Joon, et al. Deepsdf: Learning continuous signed distance functions for shape representation. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019.
14. Dosovitskiy, Alexey. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929, 2020.
15. Devlin, Jacob, et al. 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, pp. 4171-4186, 2019.
16. Yang, Su, et al. DCrownFormer+: Morphology-aware mesh generation and refinement transformer for dental crown prosthesis from 3D scan data of preparation and antagonist teeth. Medical Image Analysis, 103717, 2025.
17. Lorensen, William E., and Harvey E. Cline. Marching cubes: A high resolution 3D surface construction algorithm. Seminal graphics: pioneering efforts that shaped the field. 347-353, 1998.
18. Liu, Qin, et al. SkullEngine: a multi-stage CNN framework for collaborative CBCT image segmentation and landmark detection. In: International workshop on machine learning in medical imaging. Cham: Springer International Publishing, pp. 606-614, 2021.
19. Lee CY, Xie S, Gallagher P, Zhang Z, Tu Z. Deeply-supervised nets. In Artificial Intelligence and Statistics, Pmlr, pp. 562-570, 2015.
20. Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. International Conference on Medical image computing and computer-assisted intervention. Cham: Springer international publishing, 2015.
21. Myronenko, Andriy. 3D MRI brain tumor segmentation using autoencoder regularization. International MICCAI brain lesion workshop. Cham: Springer International Publishing, 2018.
22. Oktay, Ozan, et al. Attention u-net: Learning where to look for the pancreas. arXiv preprint arXiv:1804.03999, 2018.
23. Hatamizadeh, Ali, et al. Unetr: Transformers for 3d medical image segmentation. Proceedings of the IEEE/CVF winter conference on applications of computer vision, 2022.
24. Hatamizadeh, Ali, et al. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. International MICCAI brain lesion workshop. Cham: Springer International Publishing, 2021.
25. Cignoni, Paolo, et al. Meshlab: an open-source mesh processing tool. Eurographics Italian chapter conference. Vol. 2008. 2008.