

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

Su Yang^{1,2,3*}, Daeseung Kim^{1*}, Kevin Gu¹, Rohan Dharia⁴, Lois Jung^{1,5}, Jaime Gateno^{1,6,7}

¹Department of Oral and Maxillofacial Surgery, Houston Methodist Research Institute, Houston, TX, USA, ²Department of Medical Informatics, School of Medicine, Kyungpook National University, Daegu, Republic of Korea

³AI-driven Medical Innovation Center, Kyungpook National University Hospital, Daegu, Republic of Korea, ⁴Rice University, Houston, TX, USA, ⁵College of Medicine, Hallym University, Chuncheon, Republic of Korea

⁶Sylvia and James E. Norton Distinguished New Century Chair in Oral and Maxillofacial Surgery, Houston Methodist Hospital, Houston, TX, USA, ⁷Clinical Surgery, Weill Cornell Medical College, New York, NY, USA (*Equal contribution)

Abstract

Challenges

Soft-tissue-driven (face-first) orthognathic surgical planning requires reconstructing an anatomically coherent facial soft-tissue volume from a given target outer facial surface, whose inner surface serves as a patient-specific blueprint for the required skeletal configuration. However, existing implicit neural representation (INR) approaches lack explicit volumetric constraints, rely on fixed point/mesh sampling strategies, and use 3D CNN or Transformer encoders/decoders whose standard skip connections do not enforce input-conditioned volumetric consistency. These limitations lead to unrealistic soft-tissue thickness, local collapse, or irregular boundaries, undermining the clinical reliability of inverse planning.

Contributions

- Hybrid 3D encoder:** We employ a CNN-based hierarchical volumetric encoder enhanced with Transformer-based global contextual modeling to capture both local anatomical detail and long-range structural dependencies.
- Triplet skip-attention (TSA) decoder:** We design a 3D decoder with TSA that explicitly models the volumetric correspondence between the input outer surface and the reconstructed volume for input-conditioned decoding.
- Volume-constraint and projection-aware losses:** We introduce a volume-constraint loss (VRL) and projection-aware loss (PRL) to enforce voxel-level fidelity, global volumetric consistency, and anatomically coherent volume distribution via deep supervision of intermediate projections.

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

Methods

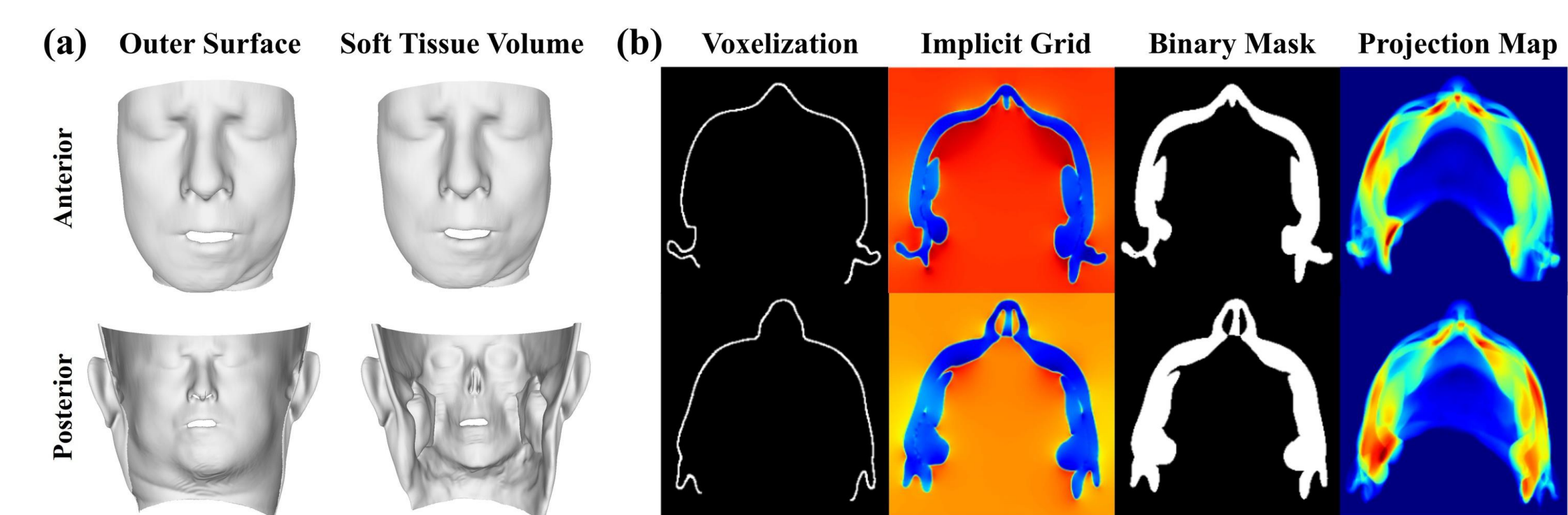


Figure 1. (a) Outer surface of the soft tissue and the soft-tissue volume consisting of the superficial layer, masseter muscles, and parotid glands. (b) Voxalization of the outer surface, the implicit grid of the soft-tissue volume, and both the binary mask and the projection map of the implicit grid.

Materials

A total of 100 scans (73 CT and 27 CBCT) were collected from an anonymized dataset (IRB# Pro00009723). Soft tissues and corresponding outer surfaces were annotated by an experienced oral and maxillofacial surgeon using in-house software and MeshLab (Figure 1a). The dataset was randomly split into 80 training and 20 test cases. Each subject was rigidly registered to a soft-tissue template using eight anatomical landmarks (left/right tragus, endocanthion, exocanthion, menton, and glabella). The ground-truth implicit grid was estimated by DPSR from the points and normals of the soft-tissue surface at a resolution of 256^3 , and the outer surface was voxelized to obtain the network input. All networks were trained with the Adam optimizer for 2000 epochs (learning rate 10^{-4} , batch size 2) on an NVIDIA RTX A6000 GPU.

Network Architecture

Given the voxelized outer-surface volume V_i , a five-level 3D CNN encoder ($3 \times 3 \times 3$ convolutions, IN, LReLU, $2 \times 2 \times 2$ max-pooling) produces a bottleneck feature z_c , which is tokenized with 3D sinusoidal positional encoding and processed by $L = 4$ Transformer blocks (Figure 2). The decoder upsamples the encoded features in four stages, and at each stage TSA forms the query, key, and value from the encoder feature r_s , decoder feature d_s , and resized voxel input v_s , respectively, to adaptively inject input-conditioned context into the skip connections. The voxelized input V_i is added to the predicted implicit grid V_p to enforce hard geometric consistency at boundary voxels, and the surface is extracted by Marching Cubes.

The volume-constraint reconstruction loss (VRL) combines a voxel-wise regression term (L_{MSE}) and a volume-consistency term (L_{DSC}) on the binarized grids, and the projection-aware reconstruction loss (PRL) matches the 2D depth-wise mean projection $J(V)$ of intermediate predictions to that of the ground truth V_g with deep supervision.

$$L_{MSE} = \|V_p - V_g\|_2^2, \quad L_{DSC} = 1 - \frac{2 \sum \hat{V}_p \hat{V}_g + \epsilon}{\sum \hat{V}_p + \sum \hat{V}_g + \epsilon},$$

$$PRL = \sum_{s \in \{2,3,4,5\}} \|\mathcal{J}(V_p^s) - \mathcal{J}(V_g)\|_2^2.$$

PAINT Architecture

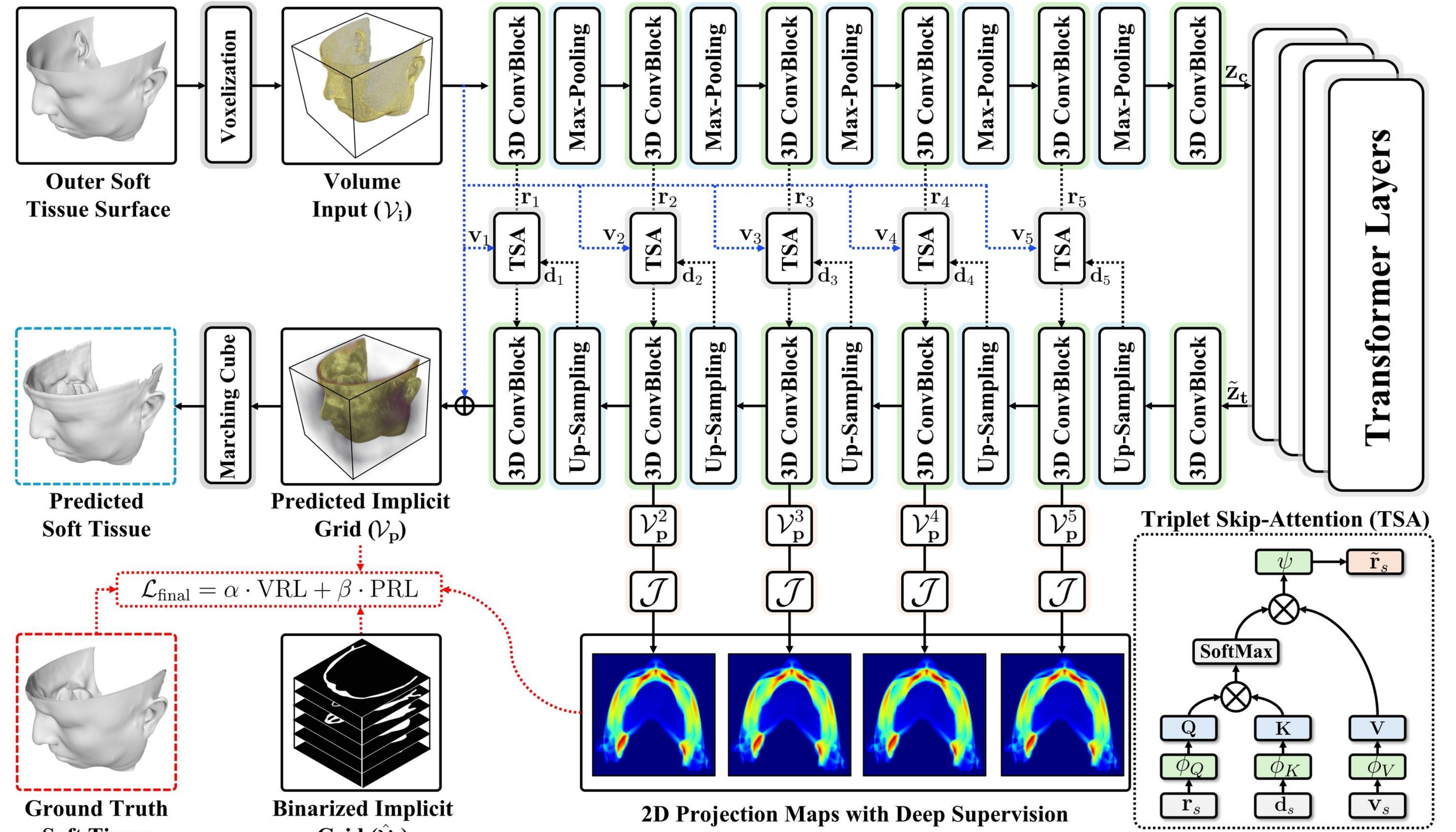


Figure 2. Illustration of the PAINT architecture. PAINT takes a voxelized outer-surface representation and reconstructs the soft-tissue volume via the predicted implicit grid with triplet skip-attention (TSA), volume-constraint reconstruction loss (VRL), and projection-aware reconstruction loss (PRL) under deep supervision.

Visual Comparison

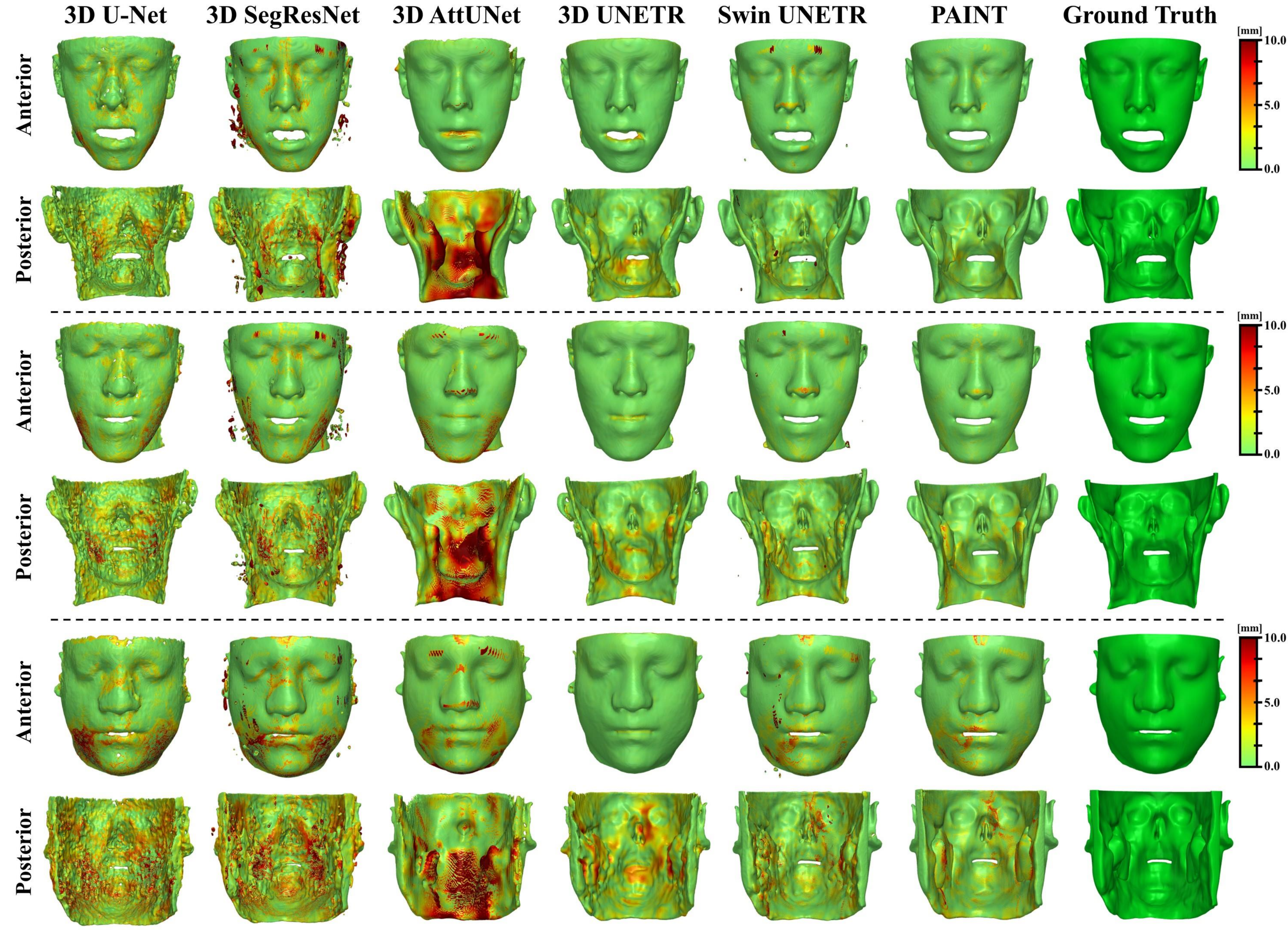


Figure 3. Visual comparison of different methods for inner surface generation of soft tissue. The color map denotes the surface distance error.

Experimental Results

Table 1. Performance comparison of PAINT with existing 3D deep networks (mean $\pm$ SD; HD and SDE in mm).

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

Table 2. Ablation study of key components (TRF, Transformer; TSA, triplet skip-attention) and loss functions (VRL, PRL) in PAINT.

Components	MAE (↓)	R ² (↑)	IoU (↑)	HD (↓)	SDE (↓)	NC (↑)
(a) Key components						
Baseline	35.734 $\pm$ 9.474	0.870 $\pm$ 0.039	0.862 $\pm$ 0.025	5.052 $\pm$ 0.968	1.207 $\pm$ 0.271	0.951 $\pm$ 0.014
+ TRF	31.067 $\pm$ 9.689	0.872 $\pm$ 0.032	0.869 $\pm$ 0.024	4.788 $\pm$ 0.869	1.133 $\pm$ 0.238	0.957 $\pm$ 0.015
+ TSA	33.297 $\pm$ 9.878	0.846 $\pm$ 0.040	0.865 $\pm$ 0.026	4.794 $\pm$ 0.886	1.139 $\pm$ 0.248	0.956 $\pm$ 0.015
+ TRF + TSA (PAINT)	30.120 $\pm$ 9.675	0.866 $\pm$ 0.040	0.866 $\pm$ 0.025	4.576 $\pm$ 0.868	1.105 $\pm$ 0.242	0.957 $\pm$ 0.014
(b) VRL and PRL						
Baseline	34.302 $\pm$ 8.279	0.869 $\pm$ 0.037	0.860 $\pm$ 0.023	5.074 $\pm$ 0.927	1.212 $\pm$ 0.240	0.953 $\pm$ 0.014
+ VRL	31.956 $\pm$ 8.856	0.871 $\pm$ 0.039	0.864 $\pm$ 0.023	4.964 $\pm$ 0.959	1.164 $\pm$ 0.254	0.957 $\pm$ 0.015
+ PRL	31.566 $\pm$ 8.799	0.874 $\pm$ 0.038	0.868 $\pm$ 0.023	4.947 $\pm$ 0.970	1.152 $\pm$ 0.053	0.957 $\pm$ 0.014
+ VRL + PRL (PAINT)	30.120 $\pm$ 9.675	0.866 $\pm$ 0.040	0.866 $\pm$ 0.025	4.576 $\pm$ 0.868	1.105 $\pm$ 0.242	0.957 $\pm$ 0.014