Profile-Specific 3DMM Regression from a Single Lateral Face Image

Supplementary Material

Table 1. ProfileSynth generation setup.

Item	Setting
Face model	FLAME2020 [1]
Requested sample count	100,000
Shape dimensionality	300
Expression	fixed to zero
Yaw range	$[85^\circ, 95^\circ]$
Other pose components	clipped Gaussian sampling
Camera	fixed perspective, distance 0.8, FoV 20°
Render resolution	1024×1024
Input resolution	512×512
Conditioning signals	depth + normal
Diffusion backbone	Stable Diffusion v1.5 [2]
ControlNet branches	depth + normal [6]
ControlNet model IDs	`lllyasviel/control_v11f1p_sd15_depth` `lllyasviel/control_v11p_sd15_normalbae`
Diffusion settings	25 steps, guidance 7.5, cond. scales $0.7/0.4$
Saved assets	RGB, silhouette, depth, normal, 2D/3D landmarks, FLAME parameters, and camera parameters

1. ProfileSynth Construction

The main paper introduces *ProfileSynth* as the source of paired supervision for extreme-profile FLAME regression. This supplement records the concrete generation settings relevant to reproducibility. Table 1 summarizes the fixed data-generation configuration used to create the synthetic profile regime studied in the paper.

Each sample is generated by first sampling FLAME shape and pose parameters in the extreme-profile regime, rendering depth maps, normal maps, silhouettes, and landmarks under a fixed camera, and then synthesizing RGB appearance with ControlNet conditioned on the rendered depth and normal maps. The prompt family is intentionally narrow: it requests photorealistic side-profile portraits with neutral expression, straight-ahead gaze, visible ear and jawline, tied-back hair, no glasses or mask, no hands near the face, and soft even lighting on a plain background, while the negative prompt suppresses low-resolution, non-photographic styles, frontal or three-quarter views, strong occlusions, harsh lighting, and obvious facial distortions. This narrow prompting is deliberate: the paper does not aim to model the full diversity of in-the-wild portrait photography, but rather to create a controlled profile-only supervision source.

Table 2. Geometry–appearance consistency on the ProfileSynth test split (10,000 samples). Lower is better for distances; higher is better for coverage.

Setting	Mean	Cov.@2px	Sym. Chamfer
Render vs. GT	0.565	–	–
Synth vs. matched	10.047	0.567	5.722
Synth vs. shifted	32.214	0.307	17.250

2. Geometry–Appearance Consistency

Because the training RGB images are diffusion-generated, a key concern is label misalignment between the conditioning geometry and the synthesized appearance. To test this, we measure whether strong edges in the generated RGB remain near the conditioning silhouette boundary. As in the main paper, we compare the matched silhouette against a shifted-control silhouette obtained by translating the ground-truth mask.

The matched setting is substantially better than the shifted control on all reported measures. This does not imply that the synthetic RGB is perfectly realistic; rather, it supports the narrower and more important claim that the generated profile appearance remains sufficiently tied to the conditioning geometry to serve as supervised training data.

3. Profile-Aware Supervision

The training loss used in the paper is

$$\mathcal{L} = w_p \mathcal{L}_{param} + w_l \mathcal{L}_{lm3d} + w_j \mathcal{L}_{jaw}, \qquad (1)$$

with $(w_p, w_l, w_j) = (1, 100, 10)$. The parameter term supervises FLAME shape and pose, the landmark term supervises 51 static non-contour landmarks, and the jawline term supervises a fixed FLAME jawline-band vertex set (65 vertices in our implementation).

Two design choices are important in the extreme-profile setting. First, we exclude the 17 contour landmarks from the standard 68-point configuration because contour definitions become unstable near complete profile views and FLAME dynamic landmarks are not designed for this regime. Second, the jawline loss is applied only to the *visible* subset of jawline-band vertices, where visibility is determined by rasterizing the ground-truth mesh under the ground-truth camera. This prevents the supervision from over-penalizing

Table 3. Ablation study on ProfileSynth (test split, no rigid alignment). Lower is better for errors and boundary Chamfer; higher is better for IoU.

Variant	3D LM L2↓	E_{vis} ↓	$E_{\text{jaw,vis}}$ ↓	IoU↑	B-Chamfer↓
Baseline (param + 3D lm + jaw)	**0.003782**	**0.003558**	**0.003420**	0.979422	0.002114
w/o 3D landmark loss	0.004019	0.003653	0.003649	0.979562	0.002096
w/o jawline loss	0.003786	**0.003558**	0.003436	0.979400	0.002112
Param-only supervision	0.003908	0.003591	0.003562	**0.979639**	**0.002083**
High jawline weight ($\times$30)	0.003835	0.003583	0.003484	0.979490	0.002099

self-occluded geometry that is not directly supported by the input image.

The main paper already reports this compact ablation; here we emphasize a cautious interpretation. Table 3 shows that 3D landmark supervision contributes most clearly: removing it degrades landmark and visible-region geometry, indicating that sparse 3D anchors remain important even when contour cues dominate. Param-only supervision is also worse than the baseline on landmark and jawline-band errors, even though it slightly improves the silhouette metrics. By contrast, the standalone effect of the jawline term is modest in this summary: removing it changes the reported metrics only slightly, and increasing its weight does not improve the jawline-band error monotonically. We therefore view the jawline term as a visibility-aware profile regularizer rather than as the dominant source of the gain. This is consistent with the central message of the paper: profile reconstruction benefits from profile-specific supervision, yet contour-aware reconstruction is not solved by simply upweighting a local 3D vertex loss.

4. Evaluation Scope and Real-Image Transfer

The synthetic benchmark remains the primary evidence of the paper because it provides exact FLAME ground truth and exact silhouettes under strict profile views. For cross-method comparison on ProfileSynth, we evaluate after rigid alignment to isolate shape quality from differences in predicted pose or camera parameterization across methods. We then report both visible-region 3D errors and silhouette-based measures, since contour fidelity is the dominant observable cue under near-90° yaw.

The NoW profile-subset experiment serves a different purpose. It is intended only as a preliminary transfer study showing whether a profile-specific synthetic prior can generalize beyond the synthetic domain. Because the subset is small and the official NoW metric is a global scan-based error after alignment, it should not be interpreted as a contour-specific benchmark. For this reason, the supplementary clinical proxy evaluation below is useful: it is still not a substitute for exact 3D ground truth, but it probes the profile contour directly rather than through a global scan distance.

5. Supplementary Clinical Contour-Proxy Evaluation

Because the paper is motivated by radiation-free cephalometric analysis from lateral facial photographs, we additionally conducted a supplementary evaluation on real clinical profile photographs from the same institutional image collection used in recent cephalometric photograph studies [4, 5]. The images were made available to us for research use under the same data-governance framework as those studies. This experiment is not used as the main benchmark of the paper, but it is informative as a real-domain check of whether the learned profile prior improves clinically relevant visible contour reconstruction.

Since these clinical photographs do not provide mesh-level 3D ground truth, we evaluate with a silhouette-proxy metric focused on the anterior jawline contour. Predicted FLAME silhouettes are aligned to the clinical mask using isotropic scale and translation estimated from an anatomy-focused upper-profile bounding box, and the primary metric is computed on the front-side jawline ROI rather than on the full silhouette. We keep the older full-ROI contour metric only as a diagnostic because it is noticeably more sensitive to posterior-neck and torso contamination.

For mask extraction, we first obtain a person mask and then apply a profile-specific trimming heuristic that suppresses lower-neck and shoulder leakage; GrabCut [3] is used as a fallback when the initial mask is insufficient. Images are excluded when the mask is invalid, the predicted mesh is missing or malformed, the aligned render fails, the jawline ROI is too small to evaluate, or the image is obviously top-cropped. In the validated run, 2,362 images are discovered, 2,277 are valid, and 85 are skipped (76 non-profile or otherwise unusable images, 7 invalid bounding boxes, and 2 missing or malformed predictions).

The primary metric on the valid subset has mean 60.20 px, median 57.72 px, and standard deviation 27.55 px. A bootstrap analysis gives a 95% confidence interval of [59.06, 61.35] for the mean and [56.19, 59.73] for the median, indicating that the central tendency of the proxy metric is stable under resampling. We summarize the protocol-stability checks in Table 4.

Table 4 shows that the chosen alignment protocol is not

Table 4. Clinical protocol stability checks on the supplementary contour-proxy evaluation. Alignment sensitivity is measured on the 512-sample subset with ROI start fixed at 0.55. Lower is better.

Check	Setting	Mean
Alignment	none	120.05
Alignment	bbox	**59.79**
Alignment	bbox+pca	125.00
Sanity	predicted shape	**59.79**
Sanity	mean shape	61.29
Sanity	random shape	62.40

Table 5. Supplementary clinical contour-proxy comparison on the common valid subset ($n = 2{,}277$). Lower is better.

Method	Mean ↓	Median ↓	Std. ↓
DECA	83.03	82.93	31.33
EMOCAv2	82.46	82.57	31.26
Ours	**60.20**	**57.72**	**27.55**

arbitrary: using the upper-profile bounding box reduces the contour error by roughly half relative to no alignment, while adding PCA-based orientation is harmful in this dataset. The same sampled subset also provides a sanity check for the learned prior: the predicted-shape contour is better than both a mean-shape baseline and random-shape samples (paired Wilcoxon $p = 1.38 \times 10^{-9}$ and 1.05×10^{-14}, respectively). The diagnostic full-ROI metric is at least 20 px worse than the front-side ROI metric on 14.10% of valid samples and at least 40 px worse on 2.33%, which supports the decision to focus the evaluation on the clinically relevant visible jawline.

Table 5 shows that our method improves substantially over DECA and EMOCAv2 on this contour-focused proxy metric. On the common validated subset, the paired mean improvement is 22.83 px over DECA and 22.26 px over EMOCAv2; our method is better on 79.49% and 78.92% of images, respectively, with paired Wilcoxon $p = 3.83 \times 10^{-164}$ and 6.64×10^{-161}. We emphasize that this does not replace exact 3D evaluation: the metric is a carefully designed proxy on visible profile structure. However, it is useful supplementary evidence that the profile-specific prior learned from *ProfileSynth* transfers to real clinical photographs in a way that is consistent with the medical motivation of the paper.

References

[1] Tianye Li, Timo Bolkart, Michael J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4d scans. *ACM Trans. Graph.*, 36(6), 2017.

[2] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022.

[3] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. Grabcut: Interactive foreground extraction using iterated graph cuts. In *ACM SIGGRAPH 2004 Papers*, pages 309–314, 2004.

[4] Yui Shimamura, Chie Tachiki, Kaisei Takahashi, Satoru Matsunaga, Takashi Takaki, Masafumi Hagiwara, and Yasushi Nishii. Accuracy of cephalometric landmark and cephalometric analysis from lateral facial photograph by using cnn-based algorithm. *Scientific Reports*, 14(31089), 2024.

[5] Kaisei Takahashi, Yui Shimamura, Chie Tachiki, Yasushi Nishii, and Masafumi Hagiwara. Cephalometric landmark detection without x-rays combining coordinate regression and heatmap regression. *Scientific Reports*, 13(20011), 2023.

[6] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3836–3847, 2023.