ALIGN BEFORE YOU CORRECT: SURFACE-ALIGNED EVIDENCE FOR 3D FACE REFINEMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Refining an existing 3D face reconstruction requires choosing where corrective evidence should be represented. Camera-ray corrections vary with viewpoint, while coefficient updates depend on a specific face model. We therefore express the correction on the reconstructed surface. We project each surface point into the calibrated views, where visible appearance and viewing geometry are fused to predict a signed displacement along the local normal. A robust parametric fit converts these displacements into face-model coefficients. Surface displacements are easier to predict than both ray-space targets. On held-out FaceScape identities, the advantage is smallest for frontal views and largest for oblique ones. Residual correlation is 0.39 for ray projection and 0.48 for ray intersection, compared with 0.52 on the surface using one view and 0.74 using five. With five views, symmetric Chamfer error drops by 4.5% from a strong start and by 8.5% from an image-predicted start. This recovers half of the reduction achieved by ground-truth displacements through the same fit and roughly twice the gain from averaging one-view predictions. Direct coefficient regression from the same evidence produces about one fifth of the gain from the image-predicted start, while population-mean, constant, and random displacement fields provide no comparable improvement. Without retraining, the same predictor reduces error by 9.5–11.7% for four published reconstructors and a multi-view landmark fit on FaceScape and by 4.3% for Pixel3DMM on Headspace. Expressing correction on the reconstructed surface therefore allows the same image evidence to transfer across calibrated rigs, face models, and starting reconstructors.

1 INTRODUCTION

Refining an existing 3D face reconstruction raises a basic representation question. Parametric face reconstruction began with the morphable model of Blanz & Vetter (1999). Current methods pair richer face models such as FLAME (Li et al., 2017) with networks that recover plausible geometry from one or more images (Zhu et al., 2016; Feng et al., 2021; Zielonka et al., 2022; Giebenhain et al., 2026), yet even strong estimates retain local surface error. In calibrated multi-view capture, multiple observations provide complementary geometric evidence about the same underlying surface (Yao et al., 2018; Bai et al., 2020; Li et al., 2021). The challenge is to represent this residual evidence in a coordinate system shared across calibrated views and usable by different face models.

Image and screen space preserve a direct connection to visual observations (Li et al., 2018; Giebenhain et al., 2026). Their predictions, however, remain indexed by the observing view. Methods that regress or iteratively update model parameters express correction in a compact structured space (Kanazawa et al., 2018; Kolotouros et al., 2019; Zhang et al., 2021), where the update depends on the chosen parameterization. We use the current surface itself as the coordinate system for correction. A surface location provides a shared geometric anchor across calibrated views, and its local normal defines a compact direction for residual shape error. The central idea is to align image evidence on the current surface before predicting how the geometry should change.

We introduce a surface-aligned residual representation for multi-view face refinement. Projecting image features onto mesh vertices or 3D surface locations has proved effective for connecting visual evidence with geometry (Wang et al., 2018; Wen et al., 2019; Zhang et al., 2021). These methods use such projections for reconstruction or feature aggregation, but do not address how residual correction should be represented once an initial geometry is already available. We use the current surface itself

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

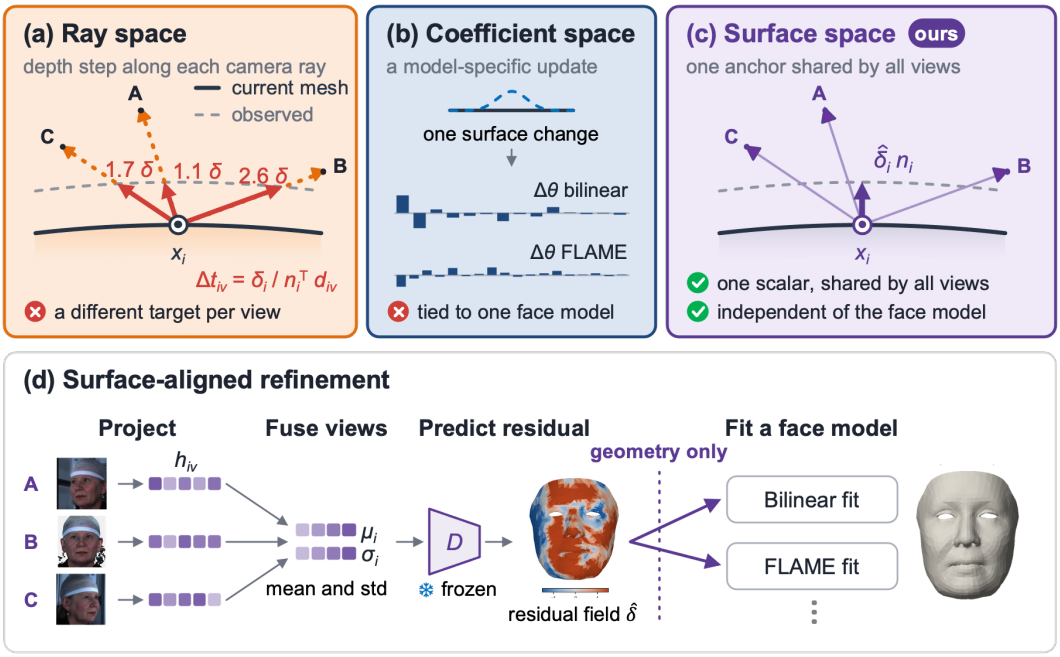


Figure 1: **Three coordinate systems for corrective evidence.** (a) Along camera rays, a surface error δ becomes a view-dependent depth step $\Delta t = \delta / n^\top d$ that diverges at oblique views. (b) A coefficient update is tied to one face model. (c) On the surface, all views share one normal displacement $\hat{\delta}_i$. (d) Per-view evidence is fused by its mean and standard deviation, and a frozen decoder D predicts $\hat{\delta}$ for a bilinear or FLAME fit.

to define that correction space. Each sampled point on the current mesh is projected into the calibrated images. Its visible features and viewing geometry are fused, and a shared decoder predicts a signed displacement along the local normal, independent of camera coordinates and face-model coefficients. A robust, regularized fit then maps this geometric evidence to the manifold of a chosen parametric face model (Vlasic et al., 2005; Li et al., 2017). This keeps the learned predictor separate from the model update. Figure 1 illustrates the formulation.

Our experiments establish three main findings. First, surface-relative correction is better conditioned than correction along a camera ray, whether the ray target is represented by projection or by a depth step along the ray. On held-out FaceScape identities (Yang et al., 2020; Zhu et al., 2023), the gap is smallest for frontal views and largest for oblique ones, following the viewing geometry. Second, aligned evidence becomes more informative as calibrated views are added. Both residual prediction and the geometry fitted from it improve with additional views, and controlled experiments show that the gain comes from the subject's own images. Third, one frozen predictor remains effective across starting quality, face-model parameterizations, and four published reconstructors and a multi-view landmark fit on FaceScape, and it transfers to Pixel3DMM (Giebenhain et al., 2026) on Headspace (Dai et al., 2017; 2020) without retraining.

Our contributions are

- We introduce a surface-aligned correction interface that represents residual evidence directly on the current geometry. This separates learned correction from both camera coordinates and a particular face-model parameterization, allowing the same predictor to support different downstream geometric models.
- We establish the geometric and empirical case for this representation. Surface-normal residuals provide a better-conditioned prediction target than camera-ray corrections, while alignment on the current surface gives multiple calibrated views a common coordinate system in which evidence can be accumulated.

- We show that surface-aligned evidence forms a reusable refinement signal across reconstruction quality, face-model parameterizations, and external reconstruction methods, including transfer across datasets without retraining the residual predictor.

2 Related work

Single image 3D face reconstruction. Morphable face models represent facial geometry in a compact parametric space (Blanz & Vetter, 1999; Vlasic et al., 2005; Li et al., 2017). Monocular methods estimate morphable face parameters from landmark, photometric, or scan supervision (Zhu et al., 2016; Sanyal et al., 2019; Deng et al., 2019). Recent methods improve animatable detail, expression, and metric identity shape (Feng et al., 2021; Daněček et al., 2022; Zielonka et al., 2022). Other approaches predict intermediate geometric representations such as depth, correspondence maps, UV position maps, surface normals, or canonical coordinates before fitting a face model (Sela et al., 2017; Feng et al., 2018; Giebenhain et al., 2026). Our setting starts from an existing reconstruction and focuses on residual refinement.

Iterative and residual correction. Many refinement methods condition on a current estimate and predict an update. Iterative error feedback, body-model regression, optimization-in-the-loop methods, and mesh-aligned feedback update structured model parameters or geometry (Carreira et al., 2016; Kanazawa et al., 2018; Kolotouros et al., 2019; Zhang et al., 2021). Render-and-compare systems express mismatch in observation space and use it to update geometry or pose (Li et al., 2018). Related systems refine correspondences before optimization (Teed & Deng, 2021), while differentiable rendering propagates image-space objectives to 3D parameters (Kato et al., 2018; Liu et al., 2019; Ravi et al., 2020).

Multi-view geometric feature alignment. Projecting image features onto 3D points or mesh vertices is widely used to connect visual observations with geometry. Pixel2Mesh and Pixel2Mesh++ sample features at projected vertices to deform meshes (Wang et al., 2018; Wen et al., 2019), while pixel-aligned implicit models condition 3D predictions on projected image features (Saito et al., 2019; Yu et al., 2021). Multi-view stereo and volumetric reconstruction aggregate calibrated observations in shared 3D coordinates (Kar et al., 2017; Yao et al., 2018; Murez et al., 2020), with related constructions used for multi-view face capture (Bai et al., 2020; Li et al., 2021). We use projected features to represent correction relative to an existing parametric estimate.

Positioning. Existing refinement methods typically express learned updates in parameter space (Carreira et al., 2016; Kanazawa et al., 2018; Kolotouros et al., 2019) or observation space (Li et al., 2018; Teed & Deng, 2021; Giebenhain et al., 2026). Parameter-space updates inherit the chosen model parameterization, while observation-space updates inherit the measuring camera. We attach the learned correction to the reconstructed surface. This gives calibrated views a shared geometric reference, supports evidence accumulation at common surface points, and allows the same predictor to feed different parametric face models. The implementation combines established components including features sampled at projected vertices (Wang et al., 2018; Zhang et al., 2021), mean and variance pooling across views (Yao et al., 2018; Wang et al., 2021), normal-direction displacements (Feng et al., 2021), and point-to-plane error (Chen & Medioni, 1992; Rusinkiewicz & Levoy, 2001).

3 Surface-Aligned Multi-View Correction

Multi-view images reveal local discrepancies between a current face estimate and the observed subject, while a parametric face model couples geometry globally through a compact coefficient space. We connect the two with a surface-attached residual field. Given an initial mesh and K calibrated images $\{I_v\}_{v=1}^{K}$, where v indexes views and i indexes sampled surface locations, the predictor estimates one scalar correction per sample. A separate parametric fit converts these local corrections into a coherent face.

This factorization gives the predictor a simple geometric target while leaving global shape consistency to the face model. It also separates residual prediction from a specific coefficient parameterization. The residual predictor is trained on FaceScape and transferred to Headspace without retraining. Implementation details are given in Appendix E.

3.1 SURFACE CORRECTION TARGET

A useful correction target should describe surface error consistently across views. For a sampled vertex x_i with unit normal n_i, we use the signed normal displacement

$$\delta_i^\star = (y_i - x_i)^\top n_i, \tag{1}$$

where y_i is a locally compatible point on the reference scan. To avoid correspondences across surface sheets or dominated by tangential mismatch, we filter matches by distance, normal compatibility, and tangential offset. Samples without a valid correspondence are excluded. Details are given in Appendix B.

For comparison, let d_{iv} be the unit camera ray from view v through x_i. Under a local tangent-plane approximation, the ray depth step required to realize the same normal displacement is

$$\Delta t_{iv} \approx \frac{\delta_i^\star}{n_i^\top d_{iv}}, \tag{2}$$

so the target becomes view dependent and is amplified at oblique views as $|n_i^\top d_{iv}|$ decreases. We also consider the bounded ray projection $(y_i - x_i)^\top d_{iv}$, whose normal component shrinks as $\delta_i^\star n_i^\top d_{iv}$ at oblique views. We compare against both ray parameterizations and floor $|n_i^\top d_{iv}|$ at 0.1 for the ray-intersection target.

Scan correspondences are used only to construct training targets. Inference requires only the initial mesh and calibrated images.

3.2 SURFACE-ALIGNED IMAGE FEATURES

The calibrated cameras and current mesh provide an explicit correspondence between each surface location and its image observations. For projection function Π_v,

$$p_{iv} = \Pi_v(x_i), \qquad f_{iv} = F(I_v)[p_{iv}], \tag{3}$$

where p_{iv} is the image projection, F is an image encoder held fixed during residual-predictor training, and $[\cdot]$ denotes bilinear sampling. Our main model uses DINOv2 ViT-B/14 (Oquab et al., 2024) at 448 pixels with features from blocks 8 and 12. Section 5.1 compares three additional encoder settings, including ResNet-50 (He et al., 2016).

Visibility $m_{iv} \in \{0, 1\}$ is obtained by depth-buffer projection. We combine the sampled image feature with viewing geometry g_{iv} and compute

$$h_{iv} = m_{iv}\, \phi([f_{iv}, g_{iv}]), \tag{4}$$

where ϕ is a shared per-view projection network. The geometric features encode orientation, relative depth, and visibility; details are in Appendix E.

3.3 MULTI-VIEW FUSION

Because all observations are attached to the same surface location, fusion should be permutation invariant. Let $V_i = \{v : m_{iv} = 1\}$ be the visible views for sample i. For $|V_i| > 0$, we summarize their evidence by

$$\mu_i = \mathrm{mean}_{v \in V_i}\, h_{iv}, \qquad \sigma_i = \mathrm{std}_{v \in V_i}\, h_{iv}, \tag{5}$$

and predict

$$\hat{\delta}_i = D(\mu_i, \sigma_i, \gamma_i), \tag{6}$$

where D is the shared residual decoder and γ_i summarizes local surface geometry and view support. The mean captures cross-view agreement, while the standard deviation captures disagreement. For $V_i = \emptyset$, we set $\hat{\delta}_i = 0$.

The decoder is trained with an L_1 loss on $\delta_i^\star$ over valid visible samples. Implementation details are given in Appendix E, and Table 3 compares moment-based fusion with learned attention.

3.4 PARAMETRIC FITTING STAGE

The predicted field is local and may contain noise or request deformations outside the capacity of the chosen face model. We therefore treat it as geometric evidence and fit it robustly in parameter space. For each sample,

$$\tilde{x}_i = x_i + m_i \hat{\delta}_i n_i, \tag{7}$$

where $m_i = \max_v m_{iv}$ indicates visibility in at least one selected view.

Let $X_i(\theta)$ denote sample i generated by the parametric model with coefficient vector θ, with $X_i(\theta_0) = x_i$ for the initial coefficients θ_0. We parameterize the update as $\theta = \theta_0 + s \odot u$, where u is the normalized coefficient update, s is a fixed per-coefficient scaling vector, and $\odot$ denotes element-wise multiplication. We solve

$$\hat{u} = \arg\min_u \frac{1}{|S|} \sum_{i \in S} H(\|X_i(\theta_0 + s \odot u) - \tilde{x}_i\|) + \lambda \|u\|_2^2, \tag{8}$$

where H is a Huber loss, S is the fixed face-region fitting support, and λ is the regularization weight. The scaling vector s, loss parameters, and solver settings are given in Appendix E.

The robust data term limits isolated residual errors, while regularization stabilizes weakly supported parameter directions. Unseen samples satisfy $\tilde{x}_i = x_i$ and remain anchored to the initial surface. Because the predictor communicates with the fitting stage through surface geometry, the same residual field can be consumed by different parametric face models.

4 EXPERIMENTAL SETUP

We first test whether surface residuals are easier to predict than ray corrections and whether aligned evidence improves with additional views on FaceScape. We then evaluate transfer across starting reconstructions, datasets, and face models on FaceScape and Headspace.

Controlled evaluation on FaceScape. FaceScape provides registered scans with 20 expressions per identity (Yang et al., 2020; Zhu et al., 2023). We use disjoint identity splits for training, validation, and test. To vary initialization quality while preserving the same underlying target, we interpolate between a per-subject scan fit θ_{fit} from Appendix C.3 and an image-predicted estimate θ_{pred},

$$\theta_\alpha = \theta_{\text{fit}} + \alpha \left(\theta_{\text{pred}} - \theta_{\text{fit}}\right). \tag{9}$$

We evaluate strong starts at $\alpha = 0.25$ and 0.5 and the natural image-predicted start at $\alpha = 1$. All methods share the official bilinear basis, built from all 847 FaceScape subjects, and the learned models use identity-disjoint splits.

Capacity-matched one-view ray and surface predictors isolate the effect of the correction coordinates. Within the surface representation, one, three, and five input views vary the amount of aligned evidence while keeping the target and fitting stage fixed. All model selection, including encoder choice, uses only training and validation identities. The 40 test identities are reserved for final reporting. Further protocol and implementation details are in Appendix C.3 and Appendix D.

Cross-dataset and cross-parameterization transfer. Headspace (Dai et al., 2017; 2020) provides 3dMD head scans with five calibrated color camera pods. We evaluate 776 subjects from an 806-subject held-out cohort after fixed exclusions described in Appendix F.4. Pixel3DMM (Giebenhain et al., 2026) provides the starting reconstruction, and FLAME (Li et al., 2017) fits the predicted field. The predictor is trained only on FaceScape, so this experiment changes the dataset, starting reconstruction, and face model simultaneously.

On FaceScape, we also fit the same frozen predictor with bilinear and FLAME models from starts with matched headroom. The FLAME start is biased toward the training mean.

Evaluation. Residual prediction is measured by Pearson correlation $\rho(\hat{\delta}, \delta^\star)$, pooled over all admissible samples visible in the reference view and averaged over three seeds. Reconstruction quality is measured by the paired reduction in Chamfer distance after fitting, averaged over each

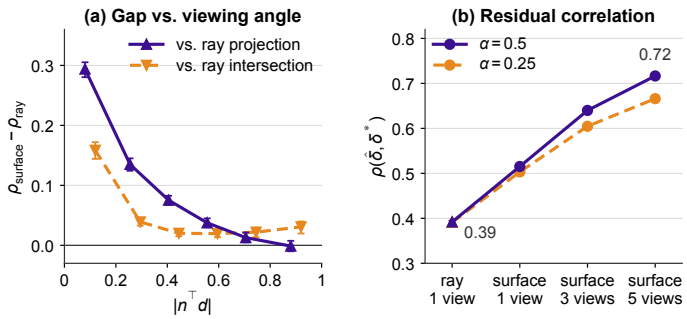


Figure 2: **Surface-coordinate advantage across viewing geometry.** Residual correlation of DINOv2 predictors. Triangles are ray targets and circles surface targets. (a) Surface minus ray correlation of one-view predictors on 40 test identities at $\alpha = 0.5$, binned by $|n^\top d|$ (small is oblique). The two comparators are dodged horizontally. 95% bootstrap intervals. (b) Pooled correlation on validation with seed variation.

identity's scans and then over identities, together with subject-level bootstrap intervals and the number of identities or subjects improved. FaceScape Chamfer is computed on the face region between model vertices and scan points within $5\,\text{c.u.}$ of the official registration's facial surface, excluding cap, hair, and neck regions.

FaceScape distances use the canonical units of the registered dataset and are reported as c.u. rather than calibrated millimeters. Headspace distances are reported in millimeters after similarity alignment as described in Appendix F.3. Absolute magnitudes are therefore not directly comparable across the two datasets.

5 Results

5.1 Surface coordinates improve residual prediction

Surface and ray targets diverge most at oblique views, where ray-depth correction becomes poorly conditioned. We train three one-view predictors with the same encoder, decoder, training data, and seeds. They differ only in the target. The targets are the surface residual, the ray-intersection depth step from Eq. (2), and the ray projection $(y_i - x_i)^\top d_{iv}$. Ray predictions are mapped back to the surface normal for evaluation. The ray-intersection target contains the same underlying residual as $\delta^\star$, so its performance isolates the effect of target conditioning.

On the 40 held-out FaceScape identities at $\alpha = 0.5$, the surface predictor exceeds ray intersection by 0.16 ($[0.14, 0.17]$) in residual correlation for oblique views with $|n_i^\top d_{iv}| < 0.2$ and by 0.02 to 0.04 in the remaining bins (Figure 2a). The small residual gap outside the oblique bin is mainly explained by the loss weighting used for the ray-depth target (Appendix A). The gap to ray projection grows monotonically from essentially zero for frontal views, -0.001 with a 95% interval of $[-0.010, +0.008]$ at $|n_i^\top d_{iv}| > 0.8$, to 0.29 for oblique views. Ground-truth ray projection itself correlates with $\delta^\star$ at only 0.23 in the oblique bin, compared with 0.96 in frontal views. The same surface advantage appears with all four image encoders tested (Table 4).

The same five-view evidence is also more effective when expressed as surface correction than as a direct coefficient update. An iterative-error-feedback-style refiner (Carreira et al., 2016; Kanazawa et al., 2018) reads the same surface-aligned evidence and regresses a coefficient update. It improves Chamfer by $0.075\,\text{c.u.}$ at $\alpha = 0.5$ (Table 1), produces about one fifth of the gain achieved by the surface field from the image-predicted start, and degrades the strongest start. The trend remains when the coefficient refiner is trained on out-of-fold starts (Appendix C.2).

5.2 Surface-aligned evidence accumulates across views

Every calibrated view indexes the same surface locations, allowing evidence to be fused before the model update. On test, residual correlation rises from 0.52 with one view to 0.66 with three views

Table 1: **Main result on FaceScape.** Symmetric Chamfer reduction after fitting on 40 test identities from the $\alpha = 0.5$ start (1.9953 c.u.), with 95% bootstrap intervals. *Realized* is the fraction of the ground-truth-offset reduction. Averaged rows average one-view predictions.

Evidence	ρ	Reduction [95% CI]	Improved	Realized
Learned evidence				
ray projection, K=1	0.394	+0.0097 [+0.0069, +0.0126]	32	0.05
ray intersection, K=1	0.481	+0.0214 [+0.0162, +0.0267]	37	0.12
surface, K=1	0.521	+0.0238 [+0.0180, +0.0298]	37	0.13
ray intersection, 5 views averaged	0.565	+0.0458 [+0.0400, +0.0516]	40	0.26
surface, 5 views averaged	0.597	+0.0509 [+0.0439, +0.0576]	40	0.28
surface, K=3	0.663	+0.0586 [+0.0468, +0.0696]	37	0.33
surface, K=5	**0.738**	**+0.0902 [+0.0776, +0.1021]**	**39**	**0.50**
Same evidence, coefficient update				
coefficient refiner, K=5	–	+0.0752 [+0.0660, +0.0848]	40	0.42
Controls				
ground-truth offsets	1	+0.1794 [+0.1662, +0.1933]	40	1
population mean field	–	+0.0021 [−0.0013, +0.0054]	23	–
another subject's images	–	−0.0059 [−0.0141, +0.0022]	17	–

and 0.74 with five. The fitted geometry improves in parallel. Learned multi-view fusion accounts for most of this gain. Averaging one-view predictions over the same five views reaches correlations of only 0.60 for surface residuals and 0.57 for ray intersection, and yields 0.039 c.u. less Chamfer reduction than the five-view surface predictor (Table 1).

At the $\alpha = 0.5$ start, five views reduce symmetric Chamfer by 0.090 c.u. or 4.5% on 39 of 40 identities. This is half of the reduction achieved by ground-truth offsets through the same fitting stage. Control fields based on other identities, population means, constants, or random values provide no comparable gain (Appendix E.2), so the improvement on these starts comes from subject-specific image evidence.

The gain increases as the starting reconstruction becomes weaker. At the natural image-predicted start $\alpha = 1$, five views reduce symmetric Chamfer by 0.21 c.u. or 8.5% after fitting and improve all 40 identities (Figure 3a). Averaged one-view predictions use the same images, yet the five-view predictor gains an additional 0.115 c.u. at $\alpha = 1$. Figure 4 shows predicted and ground-truth residuals on real scans.

5.3 The same predictor transfers across models and reconstructors

Across face models. On FaceScape validation, with the ResNet-50 encoder and matched headroom at $\alpha = 0.5$, the five-view predictor with a ResNet-50 encoder reduces common-region forward Chamfer by 0.054 c.u. with the bilinear fit and by 0.042 c.u. with FLAME. Each fit improves 38 of 39 identities (Appendix F.1).

Refining existing reconstructors. On FaceScape test, five calibrated views reduce face-region error by 9.5–11.7% when starting from a calibrated multi-view landmark fit or from the single-image reconstructors DECA, EMOCA, MICA, and Pixel3DMM. All 40 identities improve for every starting method (Table 2). A FLAME fit driven only by landmarks detected in the same views reduces error by 0.27–0.58 c.u. and improves 26 to 29 identities. Relative to this landmark-only refinement, our field adds 0.24–0.35 c.u. of reduction for the landmark start, DECA, and EMOCA (Appendix F.2).

On Headspace, the same FaceScape-trained predictor reduces Pixel3DMM error by 4.3% or 0.12 mm and improves 698 of 776 subjects (Figure 3c). The mean ground-truth offset of the 39 quality-check and development subjects reduces the error by 0.25 mm through the same fitting stage, and the predictor recovers about half of this correction without Headspace ground truth (Table 12).

Cost. On one H200 GPU, encoding five 448-pixel views takes 34 ms, the decoder takes 2 ms, and the fit takes 0.14 s per scan in batches of 16.

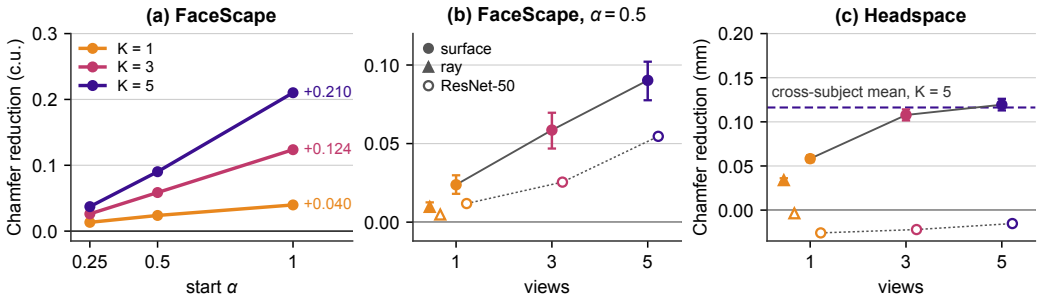


Figure 3: **Geometric gains across starting quality, views, and datasets.** Chamfer reduction after fitting. Colour gives the number of views K. Circles are surface predictors, triangles one-view ray projection, filled markers DINOv2 and hollow ones ResNet-50. (a,b) 40 FaceScape test identities, with $\alpha = 1$ the image-predicted start. (c) 776 Headspace subjects from Pixel3DMM starts in mm. The dashed cross-subject mean is the average of the other subjects' K=5 fields. 95% bootstrap intervals.

Table 2: **Refining existing reconstructors.** Face-region symmetric Chamfer (c.u.) on 695 FaceScape test scans from 40 identities. *Landmarks* is a FLAME fit to landmarks from the same views, and *Ours* uses the same fit. †Landmarks from four frontal photographs.

Start	Alone	Landmarks	Ours	Gain [95% CI]	Improved
multi-view landmark fit†	5.291	5.024	**4.674**	**+0.617** [+0.555, +0.682]	40
DECA (Feng et al., 2021)	5.369	5.040	**4.759**	**+0.610** [+0.563, +0.660]	40
Pixel3DMM (Giebenhain et al., 2026)	5.486	4.992	**4.941**	**+0.546** [+0.492, +0.601]	40
EMOCA (Daněček et al., 2022)	5.538	5.141	**4.901**	**+0.637** [+0.587, +0.691]	40
MICA (Zielonka et al., 2022)	5.653	5.071	**5.118**	**+0.535** [+0.473, +0.598]	40

6 Limitations

The proposed method relies on calibrated views and a well-aligned initial reconstruction. At $\alpha = 0.5$, it retains 87% of its performance gain under reprojection errors of up to 1.5 input pixels. When the error exceeds approximately 4 pixels, the fit begins to worsen the most accurate initializations (Appendix E.3). The face model also limits the range of corrections that can be recovered during fitting. On FaceScape test at $\alpha = 0.5$, the five-view configuration reduces symmetric Chamfer distance by 0.13 c.u. with free-form displacement and by 0.09 c.u. after fitting. The fitted surface preserves the smoothness of the initialization, while additional smoothing of the post-fit residual recovers part of the remaining discrepancy (Appendix E.2).

On Headspace, the dominant correction is a shared error pattern in Pixel3DMM rather than subject-specific geometry. The population mean of the other subjects' predicted fields performs similarly to each subject's own field, and offsets from another subject's images lose only 0.030 mm of gain (Table 12). On FaceScape, refinement driven only by landmarks in the same views reaches similar final accuracy from the Pixel3DMM and MICA starts (Table 2). Our evaluation focuses on sparse calibrated views and therefore does not cover multi-view stereo or learned head-capture systems (Bai et al., 2020; Wang et al., 2025; Li et al., 2021; Bolkart et al., 2023), which typically use denser overlapping observations and different capture settings. The five FaceScape views are sampled from a 360° rig, with the face absent from 31% of validation scans, while only two Headspace pods directly observe the face.

7 Conclusion

The current surface provides calibrated views with a shared reference for corrective image evidence and keeps the learned predictor independent of the face model's coefficient parameterization. Surface-aligned residuals are better conditioned than ray-space corrections, become more informative as

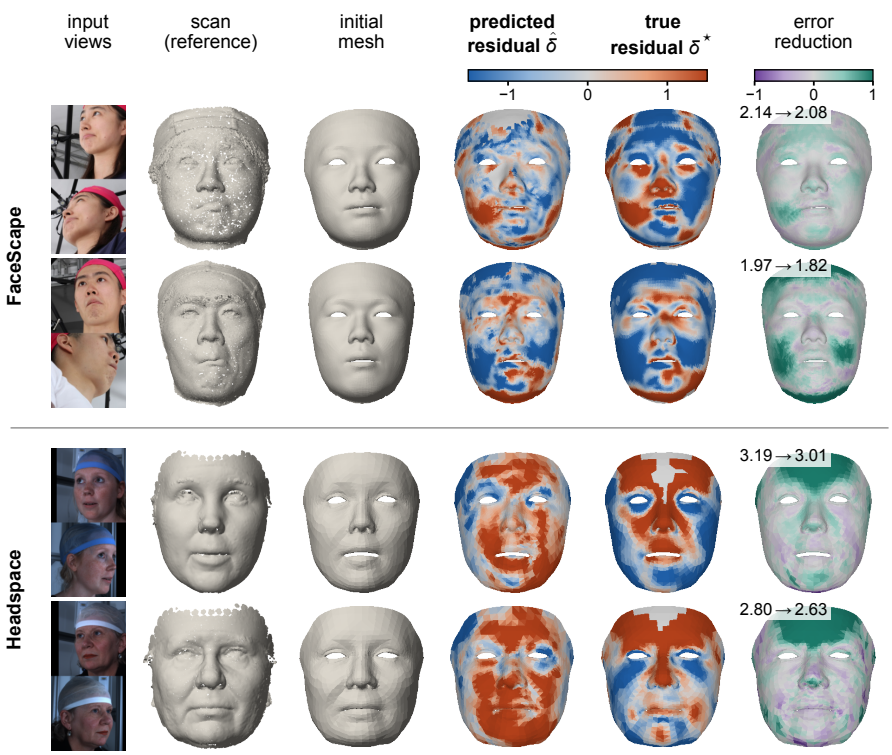


Figure 4: **Surface residuals on real scans** with DINOv2 and five views, two shown. $\hat{\delta}$ and $\delta^\star$ are predicted and ground-truth normal residuals (c.u. for FaceScape, mm for Headspace), grey without correspondence. *Error reduction* is the per-vertex decrease in scan distance after fitting, with Chamfer before→after in the corner. Appendix G shows more cases.

views are added, and can be fitted by different face models. The same frozen predictor also refines several external reconstructions across FaceScape and Headspace. Expressing residual evidence on the surface makes local correction easier to predict, combine, and reuse.

AI Use Statement

We used generative AI tools to aid and polish the writing of this paper, and for retrieval and discovery, such as finding related work. The authors reviewed all AI-assisted content and take responsibility for the final manuscript.

Ethics Statement

Face reconstruction supports telepresence, accessibility, and digital production, but can also enable impersonation and biometric surveillance. The experiments use existing face datasets and introduce no new human-subject collection. Any deployment should respect dataset licenses and consent restrictions and assess demographic differences in reconstruction error.

Reproducibility Statement

FaceScape training, validation, and test identities are disjoint. All surface correction models were trained on training identities, and all development and model selection, including the encoder and the post-fit residual, used validation identities and 24 Headspace development subjects. All development and model selection use FaceScape validation identities and Headspace subjects outside the evaluation cohort. The FaceScape test identities and the 776 scored Headspace evaluation subjects are used only for final reporting. The Headspace FLAME coefficient scale is fixed using non-evaluation subjects as

described in Appendix E. Section 3 and Appendix E give the architecture, training schedule, fitting objective and solver settings. Appendix B gives the correspondence rule, Appendix C.3 the data splits, scan fits and training starts, Appendix F.4 the Headspace cohorts, and Appendix F.2 the external reconstructor starts. All intervals use one subject bootstrap per dataset (10,000 resamples of identities or subjects, RandomState(0)), and the learned surface and ray conditions use three seeds. Code, configuration files and per-scan results will be released with the paper.

REFERENCES

Ziqian Bai, Zhaopeng Cui, Jamal Ahmed Rahim, Xiaoming Liu, and Ping Tan. Deep facial non-rigid multi-view stereo. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5849–5859, 2020.

Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3D faces. In *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques*, pp. 187–194, 1999.

Timo Bolkart, Tianye Li, and Michael J. Black. Instant multi-view head capture through learnable registration. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 768–779, 2023.

Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2D & 3D face alignment problem? (and a dataset of 230,000 3D facial landmarks). In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 1021–1030, 2017.

João Carreira, Pulkit Agrawal, Katerina Fragkiadaki, and Jitendra Malik. Human pose estimation with iterative error feedback. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4733–4742, 2016.

Yang Chen and Gérard Medioni. Object modelling by registration of multiple range images. *Image and Vision Computing*, 10(3):145–155, 1992.

Hang Dai, Nick Pears, William A. P. Smith, and Christian Duncan. A 3D morphable model of craniofacial shape and texture variation. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3085–3093, 2017.

Hang Dai, Nick Pears, William Smith, and Christian Duncan. Statistical modeling of craniofacial shape and texture. *International Journal of Computer Vision*, 128(2):547–571, 2020.

Radek Daněček, Michael J. Black, and Timo Bolkart. EMOCA: Emotion driven monocular face capture and animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20311–20322, 2022.

Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3D face reconstruction with weakly-supervised learning: From single image to image set. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 285–295, 2019.

Yao Feng, Fan Wu, Xiaohu Shao, Yanfeng Wang, and Xi Zhou. Joint 3D face reconstruction and dense alignment with position map regression network. In *European Conference on Computer Vision*, pp. 557–574, 2018.

Yao Feng, Haiwen Feng, Michael J. Black, and Timo Bolkart. Learning an animatable detailed 3D face model from in-the-wild images. *ACM Trans. Graph.*, 40(4):88:1–88:13, 2021.

Simon Giebenhain, Tobias Kirschstein, Martin Rünz, Lourdes Agapito, and Matthias Nießner. Pixel3DMM: Versatile screen-space priors for single-image 3D face reconstruction. In *International Conference on Learning Representations*, volume 2026, pp. 50353–50376, 2026.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7122–7131, 2018.

Abhishek Kar, Christian Häne, and Jitendra Malik. Learning a multi-view stereo machine. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Hiroharu Kato, Yoshitaka Ushiku, and Tatsuya Harada. Neural 3D mesh renderer. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3907–3916, 2018.

Nikos Kolotouros, Georgios Pavlakos, Michael J. Black, and Kostas Daniilidis. Learning to reconstruct 3D human pose and shape via model-fitting in the loop. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2252–2261, 2019.

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):194:1–194:17, 2017.

Tianye Li, Shichen Liu, Timo Bolkart, Jiayi Liu, Hao Li, and Yajie Zhao. Topologically consistent multi-view face inference using volumetric sampling. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3804–3814, 2021.

Yi Li, Gu Wang, Xiangyang Ji, Yu Xiang, and Dieter Fox. DeepIM: Deep iterative matching for 6D pose estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 683–698, 2018.

Shichen Liu, Tianye Li, Weikai Chen, and Hao Li. Soft rasterizer: A differentiable renderer for image-based 3D reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7708–7717, 2019.

Zak Murez, Tarrence van As, James Bartolozzi, Ayan Sinha, Vijay Badrinarayanan, and Andrew Rabinovich. Atlas: End-to-end 3D scene reconstruction from posed images. In *European Conference on Computer Vision*, pp. 414–431, 2020.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, pp. 1–31, 2024.

Nikhila Ravi, Jeremy Reizenstein, David Novotny, Taylor Gordon, Wan-Yen Lo, Justin Johnson, and Georgia Gkioxari. Accelerating 3D deep learning with PyTorch3D. *arXiv preprint arXiv:2007.08501*, 2020.

Szymon Rusinkiewicz and Marc Levoy. Efficient variants of the ICP algorithm. In *Proceedings of the Third International Conference on 3-D Digital Imaging and Modeling*, pp. 145–152, 2001.

Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. PIFu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 2304–2314, 2019.

Soubhik Sanyal, Timo Bolkart, Haiwen Feng, and Michael J. Black. Learning to regress 3D face shape and expression from an image without 3D supervision. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7755–7764, 2019.

Matan Sela, Elad Richardson, and Ron Kimmel. Unrestricted facial geometry reconstruction using image-to-image translation. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 1585–1594, 2017.

Zachary Teed and Jia Deng. DROID-SLAM: Deep visual SLAM for monocular, stereo, and RGB-D cameras. In *Advances in Neural Information Processing Systems*, volume 34, pp. 16558–16569, 2021.

Daniel Vlasic, Matthew Brand, Hanspeter Pfister, and Jovan Popović. Face transfer with multilinear models. *ACM Trans. Graph.*, 24(3):426–433, 2005.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual geometry grounded transformer. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5294–5306, 2025.

Nanyang Wang, Yinda Zhang, Zhuwen Li, Yanwei Fu, Wei Liu, and Yu-Gang Jiang. Pixel2Mesh: Generating 3D mesh models from single RGB images. In *European Conference on Computer Vision*, pp. 55–71, 2018.

Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul P. Srinivasan, Howard Zhou, Jonathan T. Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas Funkhouser. IBRNet: Learning multi-view image-based rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4690–4699, 2021.

Chao Wen, Yinda Zhang, Zhuwen Li, and Yanwei Fu. Pixel2Mesh++: Multi-view 3D mesh generation via deformation. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1042–1051, 2019.

Haotian Yang, Hao Zhu, Yanru Wang, Mingkai Huang, Qiu Shen, Ruigang Yang, and Xun Cao. FaceScape: A large-scale high quality 3D face dataset and detailed riggable 3D face prediction. In *2020 IEEE/CVF Conference on computer vision and pattern recognition (CVPR)*, pp. 598–607, 2020.

Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. MVSNet: Depth inference for unstructured multi-view stereo. In *European Conference on Computer Vision*, pp. 785–801, 2018.

Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelNeRF: Neural radiance fields from one or few images. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4576–4585, 2021.

Hongwen Zhang, Yating Tian, Xinchi Zhou, Wanli Ouyang, Yebin Liu, Limin Wang, and Zhenan Sun. PyMAF: 3D human pose and shape regression with pyramidal mesh alignment feedback loop. In *2021 IEEE/CVF international conference on computer vision (ICCV)*, pp. 11426–11436, 2021.

Hao Zhu, Haotian Yang, Longwei Guo, Yidi Zhang, Yanru Wang, Mingkai Huang, Menghua Wu, Qiu Shen, Ruigang Yang, and Xun Cao. FaceScape: 3D facial dataset and benchmark for single-view 3D face reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(12): 14528–14545, 2023.

Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, and Stan Z. Li. Face alignment across large poses: A 3D solution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 146–155, 2016.

Wojciech Zielonka, Timo Bolkart, and Justus Thies. Towards metrical reconstruction of human faces. In *European Conference on Computer Vision*, pp. 250–269, 2022.

Appendix results use FaceScape validation identities unless marked as test or Headspace. Gains *after fitting* are measured on the geometry produced by the parametric fit of Section 3.4.

A CAPACITY-MATCHED RAY COMPARATORS

The ray rows of Tables 1 and 3 are point-sampled decoders that share the frozen encoder, the $526{,}593$-parameter decoder, the training cells, the sample support, and the fitting stage with the surface predictors. The ray-projection decoder regresses $(y_i - x_i)^\top d_i$, and its prediction is projected onto the normal, which attenuates a pure normal correction to $\delta_i^\star (n_i^\top d_i)^2$. The ray-intersection decoder regresses the depth step of Eq. (2). This target is heavy-tailed at oblique views, and an L_1 loss on it weights each sample by $1/|\tilde{c}_i|$ relative to one on δ. With this weighting matched in the surface predictor, the surface advantage in residual correlation is not significant at frontal views and remains 0.17 at oblique views, as Eq. (2) predicts.

After the fit, the surface predictor gains $0.0024\,\text{c.u.}$ more than the ray-intersection prediction mapped to the normal at $K = 1$ ($[0.0013, 0.0035]$, 30 of 40 test identities). The averaged rows of Table 1 run each one-view predictor on all five views, map each ray prediction to the normal, and average the offsets of the visible views per sample. At matched averaging the surface target gains $0.0051\,\text{c.u.}$ more at $\alpha = 0.5$ ($[0.0030, 0.0069]$, 35 of 40) and $0.0104\,\text{c.u.}$ more at $\alpha = 1$ ($[0.0075, 0.0132]$, 37 of 40). Both averaged predictors remain well below learned five-view fusion.

B SURFACE-TARGET CONSTRUCTION

Correspondence and visibility. The correspondence rule of Section 3.1 takes the $k = 16$ nearest scan points from a build half of the scan as candidates and keeps the nearest one within $5\,\text{c.u.}$, with normal agreement $n_i^\top n_y \geq 0.5$ and tangential offset at most $2\,\text{c.u.}$. Samples without such a candidate are excluded from supervision. Chamfer is evaluated on the other half of the scan. Visibility is determined by a z-buffer at 224×224 with a $1\,\text{c.u.}$ tolerance. Headspace uses the same thresholds in millimeters, with correspondences computed after the similarity alignment of Appendix F.3.

Ground-truth normal displacements. Normal displacement is an expressive enough oracle for evaluating the correction interface. Applied directly to the initial surface without predictor or fit, the ground-truth normal displacements take the $\alpha = 0.25$ validation start from 1.844 to $1.554\,\text{c.u.}$ (per image). Applied along the view ray instead, the same correspondences keep 41 to 43% of this gain at the two strong starts, and the unconstrained displacement $y_i - x_i$ is no better than the normal one.

C CONTROLLED ABLATIONS

C.1 MULTI-VIEW FUSION AND ENCODER CHOICE

Table 3 gives residual prediction and the gain after fitting for each representation and view count with the ResNet-50 encoder, and Table 4 repeats the comparison across encoders. The ordering is the same for every encoder. The surface predictor improves over the one-view ray predictor, and additional aligned views improve it further. Learned softmax pooling over visible views (510,338 parameters) in place of Eq. 5 does not improve over moment-based fusion after fitting.

C.2 COMPARISON WITH COEFFICIENT-SPACE UPDATES

A coefficient-update refiner in the style of iterative error feedback (Carreira et al., 2016; Kanazawa et al., 2018) and mesh-aligned feedback (Zhang et al., 2021) reads the same evidence as the five-view surface predictor, namely the frozen DINOv2 features at the projected vertices, the viewing-geometry channels, and mean and standard-deviation fusion over visible views. It pools the per-vertex descriptors with learned attention, appends the current standardized coefficients, and regresses a coefficient update in the fit's own scale, applied three times with the features re-sampled after each update (611,393 parameters). It is supervised by the mean vertex distance to the scan-fit mesh after every step, with the training cells, epochs, optimizer and seeds of the surface decoder. At the strong and image-predicted starts surface correction is more reliable (Table 5). The coefficient refiner

Table 3: Residual prediction from strong starts on FaceScape validation with the ResNet-50 encoder, three seed means. ρ is measured on admissible samples visible in the reference view. The remaining columns, all at $\alpha = 0.5$, give MAE reduction relative to predicting zero, free-form symmetric Chamfer gain, and gain after the bilinear fit with its 95% interval and identities improved. Ground-truth offsets with the 5 c.u. correspondence cap gain $+0.3055$ free form and $+0.1976$ c.u. after fitting. The ray row is the ray-projection decoder (Appendix A), and attention rows replace Eq. 5 with learned softmax pooling.

		$\rho(\hat{\delta}, \delta^\star)$		$\alpha = 0.5$			
Representation	K	α=0.25	α=0.5	MAE red.	free form	fitted [95% CI]	ids
ray	1	0.290	0.292	3.9%	+0.0084	+0.0045 [+0.0029, +0.0062]	31/39
surface	1	0.361	0.375	6.8%	+0.0158	+0.0115 [+0.0079, +0.0152]	32/39
surface	3	0.440	0.469	11.0%	+0.0389	+0.0322 [+0.0245, +0.0404]	36/39
surface	5	0.507	0.555	16.4%	+0.0642	+0.0601 [+0.0498, +0.0712]	38/39
surface, attention	3	0.445	0.468	10.9%	+0.0386	+0.0320	36/39
surface, attention	5	0.485	0.520	13.8%	+0.0559	+0.0532	38/39

Table 4: Encoder ablation at $\alpha = 0.5$, with input pixels in parentheses. The hybrid encoder combines ResNet-50 layer 2 with DINOv2 block 12, ray is the one-view ray-projection predictor, and the gap is the one-view surface minus ray reduction after fitting (c.u.).

	$\rho(\hat{\delta}, \delta^\star)$			
Encoder	ray 1	surf. 1	surf. 5	Gap [95% CI]
Validation				
ResNet-50 (224)	0.292	0.375	**0.555**	+0.0070 [+0.0048, +0.0093]
DINOv2 (224)	0.332	0.419	**0.576**	+0.0092 [+0.0064, +0.0121]
hybrid (224)	0.349	0.443	**0.619**	+0.0101 [+0.0070, +0.0132]
DINOv2 (448)	0.392	0.515	**0.717**	+0.0147 [+0.0108, +0.0187]
Test				
ResNet-50 (224)	0.282	0.366	**0.565**	+0.0070 [+0.0048, +0.0091]
DINOv2 (448)	0.394	0.521	**0.738**	+0.0141 [+0.0108, +0.0174]

degrades the strongest start and realizes substantially less of the gain from the image-predicted start, and this holds when both are trained on out-of-fold starts. At $\alpha = 2$, twice as far from the scan fit as the image prediction and outside the residual-refinement setting, the coefficient refiner gains more.

Table 5: Coefficient updates from the same surface-aligned evidence on FaceScape test, five views, DINOv2 encoder, three seeds. Symmetric Chamfer reduction (c.u.) with 95% subject bootstrap intervals and identities improved of 40. *Out-of-fold* rows are trained on the starts of Appendix C.3. The $\alpha = 2$ start lies twice as far from the scan fit as the image prediction.

Method	$\alpha = 0.25$	$\alpha = 0.5$	$\alpha = 1$	$\alpha = 2$
coefficient refiner, 3 steps	−0.0278 [−0.0384, −0.0183], 7	+0.0752 [+0.0660, +0.0848], 40	+0.0414 [+0.0235, +0.0599], 31	+0.5116 [+0.4471, +0.5763], 40
coefficient refiner, 3 steps, out-of-fold	−0.0984 [−0.1240, −0.0754], 1	+0.0649 [+0.0398, +0.0878], 34	+0.0690 [+0.0360, +0.1029], 28	+1.1446 [+1.0343, +1.2594], 40
surface predictor + fit	+0.0372 [+0.0303, +0.0436], 38	+0.0902 [+0.0776, +0.1021], 39	+0.2102 [+0.1840, +0.2351], 40	+0.4125 [+0.3655, +0.4570], 40
surface predictor + fit, out-of-fold	+0.0311 [+0.0203, +0.0404], 34	+0.0905 [+0.0724, +0.1067], 37	+0.2336 [+0.1987, +0.2657], 39	+0.4873 [+0.4255, +0.5451], 39

C.3 Training protocol and out-of-fold controls

Scan fits and starts. The scan fit θ_{fit} shares one identity vector across a subject's 20 expressions and has one expression vector and similarity placement per scan. It minimizes a 10%-trimmed symmetric Chamfer distance between the model's face-region vertices and the scan points within 5 c.u. of the official registration's facial surface, with ridge penalties $\lambda_{\text{id}} = 10^{-1}$ and $\lambda_{\text{exp}} = 1$, so cap, hair, ears and neck never pull on the fit. All methods use the official fixed FaceScape bilinear model (v1.6) with 50 identity and 51 expression coefficients. The image-predicted start ($\alpha = 1$) comes from a coefficient regressor, a fine-tuned ResNet-50 that maps one image to the 101 coefficients through a globally pooled feature vector and is supervised with θ_{fit}.

Splits and conditions. FaceScape splits contain 280 training, 39 validation, and 40 test identities with 20 expressions and five calibrated views each. The training grid has 22,400 scan–start cells ($\alpha \in \{0.25, 0.5, 1, 2\}$), and validation and test have 3,120 and 3,200. The one-view ray decoders and the one-, three-, and five-view surface decoders share the frozen encoder, the 526,593-parameter architecture, the training cells, four epochs and three seeds. For each reference view, the K consecutive views from it form the input, and predictions are scored on samples visible in the reference view.

Out-of-fold control for initialization construction. The coefficient regressor that produces the training starts is trained on the training identities. As a control, Table 6 retrains the DINOv2 decoders on starts from five cross-fitted regressors, each predicting only identities it never saw. The primary trends are unchanged, so the gains do not depend on in-sample construction of the training starts.

Table 6: Out-of-fold versus in-sample training starts for the five-view DINOv2 decoder, FaceScape validation, three seeds. Residual correlation averaged per image, unlike the pooled ρ elsewhere, and symmetric Chamfer reduction after fitting (c.u.), with 95% intervals for out-of-fold.

	Training starts	$\alpha = 0.25$	$\alpha = 0.5$	$\alpha = 1$	$\alpha = 2$
correlation	in-sample	0.615	0.655	0.697	0.709
	out-of-fold	0.582	0.635	0.698	0.735
reduction	in-sample	+0.0413	+0.0982	+0.2238	+0.4219
	out-of-fold	+0.0387	+0.1055	+0.2597	+0.5130
		[+0.0261, +0.0517]	[+0.0838, +0.1285]	[+0.2166, +0.3056]	[+0.4351, +0.5967]

D FULL HELD-OUT TEST RESULTS

All training, development and model selection used training and validation identities, and the 40 FaceScape test identities are used only for reporting. Table 7 gives the full test grid for both encoders. The three primary comparisons at $\alpha = 0.5$ with DINOv2 have 95% subject bootstrap intervals above zero. Surface over ray at $K = 1$ gains $+0.0141$ c.u. ($[+0.0108, +0.0174]$) after fitting, five views over one gain $+0.0664$ c.u. ($[+0.0575, +0.0746]$), and the subject's own evidence over another subject's images gains $+0.0961$ c.u. ($[+0.0810, +0.1101]$, 38 of 40). The ground-truth offsets of Table 1 use a wider correspondence cap (10 c.u.) to cover all starts.

Table 7: Full test grids, 40 FaceScape identities, three seed means. Ray is the ray-projection predictor. Residual correlation uses reference-view support, and gains are symmetric Chamfer reductions after the bilinear fit (c.u.), followed by the number of identities improved. The DINOv2 rows at $\alpha = 0.5$ are those of Table 1.

	$\rho(\hat{\delta}, \delta^\star)$				gain after fitting (c.u.), ids improved			
Condition	0.25	0.5	1	2	0.25	0.5	1	2
ResNet-50 encoder								
ray, $K = 1$	0.294	0.282	0.239	0.166	+0.0032, 31	+0.0048, 32	+0.0068, 30	+0.0088, 30
surface, $K = 1$	0.366	0.366	0.344	0.297	+0.0063, 32	+0.0118, 35	+0.0208, 36	+0.0337, 36
surface, $K = 3$	0.446	0.465	0.479	0.461	+0.0089, 29	+0.0255, 35	+0.0565, 37	+0.1047, 39
surface, $K = 5$	**0.520**	**0.565**	**0.622**	**0.644**	**+0.0194**, 32	**+0.0546**, 38	**+0.1282**, 40	**+0.2498**, 40
DINOv2 encoder								
ray, $K = 1$	0.399	0.394	0.346	0.256	+0.0065, 33	+0.0097, 32	+0.0134, 33	+0.0166, 33
surface, $K = 1$	0.515	0.521	0.492	0.419	+0.0133, 33	+0.0238, 37	+0.0398, 37	+0.0588, 36
surface, $K = 3$	0.627	0.663	0.694	0.689	+0.0262, 33	+0.0586, 37	+0.1237, 38	+0.2202, 38
surface, $K = 5$	**0.688**	**0.738**	**0.792**	**0.813**	**+0.0372**, 38	**+0.0902**, 39	**+0.2102**, 40	**+0.4125**, 40

E MODEL FIT AND IMPLEMENTATION

E.1 PREDICTOR AND PARAMETRIC FITTING

The decoder architecture is $\phi : 1539 \to 256 \to 256$ and $D : 518 \to 128 \to 1$ with GELU activations. The encoder features are the patch tokens of blocks 8 and 12 of a frozen DINOv2 ViT-B/14 at 448 pixels (two 768-channel 32×32 maps after the final layer norm, the same crop region as the 224-pixel input at twice the resolution), sampled bilinearly. They have the same 1536 channels as the 512- and 1024-channel layer-2 and layer-3 maps of the frozen ResNet-50 of the encoder comparison, so the decoder is unchanged. The hybrid encoder of Table 4 has 1280 channels, and only the width of the decoder's input layer changes. The output is scaled by the training standard deviation of $\delta^\star$, and σ_i is the population standard deviation over V_i, with 10^{-6} added to the variance inside the square root. Training uses AdamW with learning rate 10^{-3}, cosine decay, weight decay 10^{-4}, four epochs, and batches of four cells with 2,048 samples each. The viewing geometry is $g_{iv} = [n_i^\top d_{iv},\, (z_{iv} - \bar{z}_v)/50\,\text{c.u.},\, m_{iv}]$, where z_{iv} is the depth of x_i in view v and $\bar{z}_v$ its mean over the samples visible in that view. The geometric summary is $\gamma_i = [n_i,\, |V_i|/5,\, \text{mean}_{v \in V_i}\, n_i^\top d_{iv},\, \max_{v \in V_i}\, n_i^\top d_{iv}]$, so D reads $256 + 256 + 6$ inputs. The one-view ray conditions use the ray d_i of the input view. In the ray-projection condition the target is $(y_i - x_i)^\top d_i$, and the prediction $\hat{p}_i$ enters the fit as its normal component $\hat{\delta}_i = \hat{p}_i\, n_i^\top d_i$. In the ray-intersection condition the target is the depth step $\Delta t_i^\star = \delta_i^\star / \tilde{c}_i$ of Eq. (2), with $\tilde{c}_i = \text{sign}(n_i^\top d_i)\, \max(|n_i^\top d_i|, 0.1)$, scaled by its training standard deviation. Its prediction enters the fit as $\hat{\delta}_i = \widehat{\Delta t}_i \tilde{c}_i$.

Eq. 8 is minimized over standardized coefficients with $|u_j| \le 6$ by 200 Adam steps at learning rate 2×10^{-2} followed by four L-BFGS rounds of 50 iterations with strong Wolfe line search. The Huber loss is $H(r) = r^2/2$ for $r \le \eta$ and $\eta(r - \eta/2)$ beyond, with $\eta = 2\,\text{c.u.}$ ($2\,\text{mm}$ on Headspace), and $\lambda = 10^{-2}$. The support S holds the face-region samples visible in at least one of the five calibrated views, the same set for every K. For the bilinear model, s is the per-coefficient standard deviation of $\theta_{\text{fit}} - \theta_{\text{pred}}$ over the training images. For FLAME it is the per-coefficient standard deviation of $\psi_{\text{fit}} - \bar{\psi}_{\text{train}}$ over the training scans in Appendix F.1, and on Headspace and for the reconstructor starts of Appendix F.2 it is a single constant, 0.998, set on Headspace subjects outside the evaluation cohort and held fixed for all evaluation subjects. Correspondence is by vertex identity, and pose and placement are not optimized.

E.2 SURFACE QUALITY AND CONTROL FIELDS

Geometry quality under free-form and model-constrained updates. Symmetric Chamfer does not penalize a rough surface, so Table 8 adds the mean angle between each vertex normal and the normal of its nearest scan point (vertices within 2 c.u. of the scan), the mean uniform-Laplacian magnitude $\|x_i - \bar{x}_{\mathcal{N}(i)}\|$ as roughness, and region Chamfer on the nose, lips, eye region and chin. Free-form displacement reaches a slightly lower Chamfer at the cost of poorer geometric regularity, with an about 30% rougher surface and a larger normal error than the start. The parametric fit preserves the start's surface quality while improving every evaluated facial region.

Table 8: Surface quality at $\alpha = 0.5$, FaceScape validation (780 scans), five views, DINOv2 encoder, one predictor seed. Chamfer, region errors and roughness are in c.u. and normal error in degrees. Lower is better everywhere.

Output	Chamfer	normal (°)	roughness	nose	lips	eye region	chin
initial mesh	2.031	22.21	0.244	1.287	1.996	1.402	3.457
fit	1.932	21.62	0.245	1.210	1.833	1.317	3.216
free-form displacement	1.909	23.37	0.316	1.174	1.766	1.279	3.233
scan fit (reference)	1.745	20.70	0.248	1.058	1.524	1.227	2.879

Fields without subject-specific evidence. Table 9 passes fields that carry no subject-specific image evidence through the same fit. On FaceScape, constant shifts along the normal (magnitude b equal to the mean $|\hat{\delta}|$), random fields and the population mean of the other identities' fields leave the start unchanged or make it worse, while the subject's own field minus the population mean keeps almost

all of the gain. The FaceScape gain therefore comes from subject-specific evidence. Headspace controls are reported in Table 12.

Table 9: Fields through the fit. FaceScape validation (39 identities) at each start and FaceScape test (40 identities) at $\alpha = 0.5$, in c.u., five views, DINOv2 encoder. Gains are after fitting, averaged over three predictor seeds for the first three rows and from one seed otherwise. Population means are leave-one-subject-out.

	Validation, α				Test
Field	0.25	0.5	1	2	0.5
subject's own field	$+0.0413$	$+0.0982$	$+0.2238$	$+0.4219$	$+0.0902$
own field minus population mean	$+0.0393$	$+0.0939$	$+0.2165$	$+0.4153$	$+0.0861$
population mean of predicted fields	-0.0003	$+0.0012$	$+0.0026$	$+0.0006$	$+0.0021$
population mean of ground-truth offsets	-0.0013	-0.0013	-0.0020	-0.0038	$+0.0022$
constant shift $+b$	-0.0502	-0.0458	-0.0405	-0.0300	-0.0444
constant shift $-b$	-0.0616	-0.0713	-0.0953	-0.1365	-0.0744
random, matched variance	-0.0006	-0.0003	-0.0003	$+0.0004$	-0.0006
random, matched spectrum	-0.0289	-0.0317	-0.0398	-0.0453	–
sign-flipped field	-0.0918	-0.1506	-0.2786	-0.4730	–

E.3 Camera perturbation robustness

Every camera is perturbed independently at inference only, by a focal-length error, a rotation about its centre or a shift of its centre, while the start, the fit and the metric are unchanged. The gain degrades smoothly with the mean image displacement of the visible vertices, whatever its cause (Table 10). It remains positive under moderate calibration error and keeps about 87% of its five-view gain at $\alpha = 0.5$ up to 1.5 pixels of input-crop displacement.

Table 10: Camera error at inference, FaceScape validation, five views, DINOv2 encoder, one predictor seed. Gains are after fitting (c.u.). Displacement is the mean reprojection shift of visible vertices, in input-crop and full-resolution pixels.

	displacement (px)		gain after fitting, α			
Perturbation	crop	full	0.25	0.5	1	2
calibrated cameras	0.0	0	$+0.0413$	$+0.0982$	$+0.2238$	$+0.4219$
centre shift 1 c.u.	0.7	6	$+0.0382$	$+0.0956$	$+0.2209$	$+0.4168$
centre shift 2 c.u.	1.5	12	$+0.0277$	$+0.0851$	$+0.2091$	$+0.4015$
rotation 0.1°	2.4	19	$+0.0056$	$+0.0642$	$+0.1883$	$+0.3771$
centre shift 5 c.u.	3.7	29	-0.0309	$+0.0263$	$+0.1457$	$+0.3239$
focal length $\pm 5\%$	4.1	32	-0.0846	-0.0242	$+0.1004$	$+0.2816$
rotation 0.25°	6.0	47	-0.0997	-0.0458	$+0.0637$	$+0.2293$
focal length $\pm 10\%$	8.2	64	-0.2096	-0.1628	-0.0704	$+0.0726$
rotation 1°	24.2	189	-0.1171	-0.0949	-0.0549	-0.0004

F Cross-model and cross-dataset evaluation

F.1 Matched cross-parameterization on FaceScape

This experiment tests whether the same predicted surface evidence can be used by different parametric models. It is not a comparison of the two models. The bilinear start is θ_α of Eq. (9). The FLAME start interpolates the scan's FLAME fit toward the training mean, $\psi_\beta = \psi_{\text{fit}} + \beta(\bar{\psi}_{\text{train}} - \psi_{\text{fit}})$, with one β per α chosen so that both models have the same mean headroom over their scan fits. The same predicted field is then fitted by each model and scored on the face region common to both (Table 11).

Table 11: Cross-parameterization at matched FaceScape headroom, five views, ResNet-50 encoder, common-region forward Chamfer (c.u.). *Specific* is own minus other-subject evidence, and *realized* is the own gain divided by that of ground-truth offsets. The headroom at $\alpha = 0.25$ is $0.0704\,\mathrm{c.u.}$ for both models.

Start	Fit	Own [95% CI]	Other subj.	Specific [95% CI]	GT offsets	Realized	Improved
$\alpha = 0.5$	bilinear	$+0.054$ $[+0.044, +0.065]$	-0.010	$+0.064$ $[+0.051, +0.078]$	$+0.162$	0.34	38/39
$\alpha = 0.5$, matched	FLAME	$+0.042$ $[+0.032, +0.052]$	-0.024	$+0.066$ $[+0.054, +0.079]$	$+0.141$	0.30	38/39
$\alpha = 0.25$	bilinear	$+0.0216$ $[+0.0139, +0.0287]$	-0.0091	$+0.0308$ $[+0.0223, +0.0390]$	$+0.0766$	0.28	32/39
$\alpha = 0.25$, matched	FLAME	$+0.0109$ $[+0.0050, +0.0166]$	-0.0277	$+0.0386$ $[+0.0310, +0.0465]$	$+0.0576$	0.19	28/39

F.2 Reconstructor starts on FaceScape

This experiment isolates shape refinement from pose and scale estimation. Every reconstructor start is therefore rigidly aligned to the scan's FLAME fit before refinement, the placement every FaceScape start uses, so only the shape comes from the reconstructor. Each start is re-expressed losslessly as FLAME plus a fixed per-vertex offset and refined by the FLAME fit used on Headspace.

Pixel3DMM. Each FaceScape scan is started from the official Pixel3DMM (Giebenhain et al., 2026) reconstruction of its most frontal photograph, run with the released default single-image configuration. Evaluation uses the 695 test scans that every reconstructor processes. On these scans Pixel3DMM exhibits a substantial shared bias. The population mean of the other identities' fields gains $+0.3958\,\mathrm{c.u.}$, and the subject's own field provides an additional reduction of $+0.1500\,\mathrm{c.u.}$ ($[+0.1044, +0.1948]$, 33 of 40) over it.

Other reconstructors. DECA (Feng et al., 2021), EMOCA v2 (Daněček et al., 2022) and MICA (Zielonka et al., 2022) start from the same photograph with their released weights and the authors' face crops. MICA predicts identity only, with a neutral expression. The multi-view landmark fit uses the scan's four most frontal photographs and their calibrated cameras. It fits FLAME with 100 shape and 50 expression coefficients and jaw and head pose at metric scale to the 68 FAN landmarks (Bulat & Tzimiropoulos, 2017) of every photograph, with a Huber reprojection loss normalized by face size, from a triangulated initialization. Its two prior weights are chosen on validation by its own start error. Every start is re-expressed, placed and fitted as above. Five aligned views improve every start on all 40 identities (Table 2).

Same-view landmark refinement. This control receives the placed start and the same five calibrated photographs. FAN landmarks detected in these photographs take the place of the residual field in the same FLAME fit, with the same coefficients, prior and solver, and a Huber reprojection loss weighted by landmark confidence. Its weight against the prior is chosen for each start on validation. Scans whose views show no detectable face (25% of test scans) keep their start. The landmark-only control is weaker for the landmark, DECA and EMOCA starts, by $+0.350$ $[+0.283, +0.410]$, $+0.282$ $[+0.199, +0.362]$ and $+0.240$ $[+0.154, +0.326]\,\mathrm{c.u.}$. From Pixel3DMM and MICA its final accuracy is comparable, with differences of $+0.052$ $[-0.094, +0.191]$ and -0.047 $[-0.192, +0.091]$.

F.3 Cross-dataset transfer to Headspace

Headspace evaluates whether a FaceScape-trained surface predictor can be reused across a different dataset, capture rig, starting reconstructor, and downstream face model without retraining. Headspace Chamfer is the symmetric Chamfer distance between the mesh's face-region vertices and the held-out scan points within $5\,\mathrm{mm}$ of the subject's FLAME registration. Each mesh, refined or not, is first aligned to the scan by a similarity, initialized from the registration and refined by 30 rounds of 10%-trimmed ICP. Pods 1C and 2C see 73% and 75% of the face region and pods 3C to 5C under 5%, so views beyond the second add little. Five views reduce error by $0.1194\,\mathrm{mm}$ and improve 698 of 776 subjects (Table 12). The population-mean control gains nearly as much, so most of this gain is a correction shared across the cohort. The subject's own evidence still exceeds another subject's evidence by $0.0299\,\mathrm{mm}$.

Table 12: Refining Pixel3DMM on Headspace, 776 subjects, Pixel3DMM start ($2.7669\,\mathrm{mm}$ similarity-aligned symmetric Chamfer), DINOv2 encoder, FLAME fit, with gains in mm and 95% subject bootstrap intervals. *Population mean* is the leave-one-out mean of the other 775 subjects' predicted five-view fields, and *mean ground-truth offset* is the mean ground-truth residual of the 39 quality-check and development subjects (Appendix F.4), through the same fit. The last two rows are paired contrasts.

Evidence	Gain [95% CI] (mm)	Improved (of 776)
matched ray, $K{=}1$	$+0.0339$ $[+0.0316, +0.0361]$	677
surface, $K{=}1$	$+0.0583$ $[+0.0544, +0.0621]$	682
surface, $K{=}3$	$+0.1078$ $[+0.1016, +0.1141]$	695
surface, $K{=}5$	$\mathbf{+0.1194}$ $[\mathbf{+0.1129}, \mathbf{+0.1261}]$	**698**
Controls		
population mean	$+0.1163$ $[+0.1126, +0.1200]$	764
mean ground-truth offset	$+0.2517$ $[+0.2427, +0.2608]$	753
other-subject evidence	$+0.0895$ $[+0.0828, +0.0964]$	653
surface, $K{=}5$ − other-subject evidence	$+0.0299$ $[+0.0227, +0.0372]$	485
surface, $K{=}5$ − population mean	$+0.0032$ $[-0.0018, +0.0082]$	441

F.4 Headspace cohorts

Of the Headspace subjects with complete scans, landmarks and color images, 946 have a FLAME registration and a usable frontal view. A fixed split reserves 140 of them for development and leaves an evaluation cohort of 806. Rules fixed before evaluation remove 3 Pixel3DMM crop failures and 27 gross landmark failures for all conditions, so 776 subjects are scored. The mean ground-truth offset of Table 12 uses the 39 quality-check and development subjects that have a Pixel3DMM start. Ground-truth control fields and the fixed FLAME coefficient scale use only subjects outside the evaluation cohort.

G Qualitative results

Figures 5 and 6 extend Figure 4 with rows chosen by a fixed rule, in the same columns and colour scales. FaceScape identities shown are limited to those the dataset releases for publication, which are the three publishable identities in our training split. Their rows therefore use out-of-fold starts and decoder. On Headspace the rows span the distribution of the 776 subject gains.

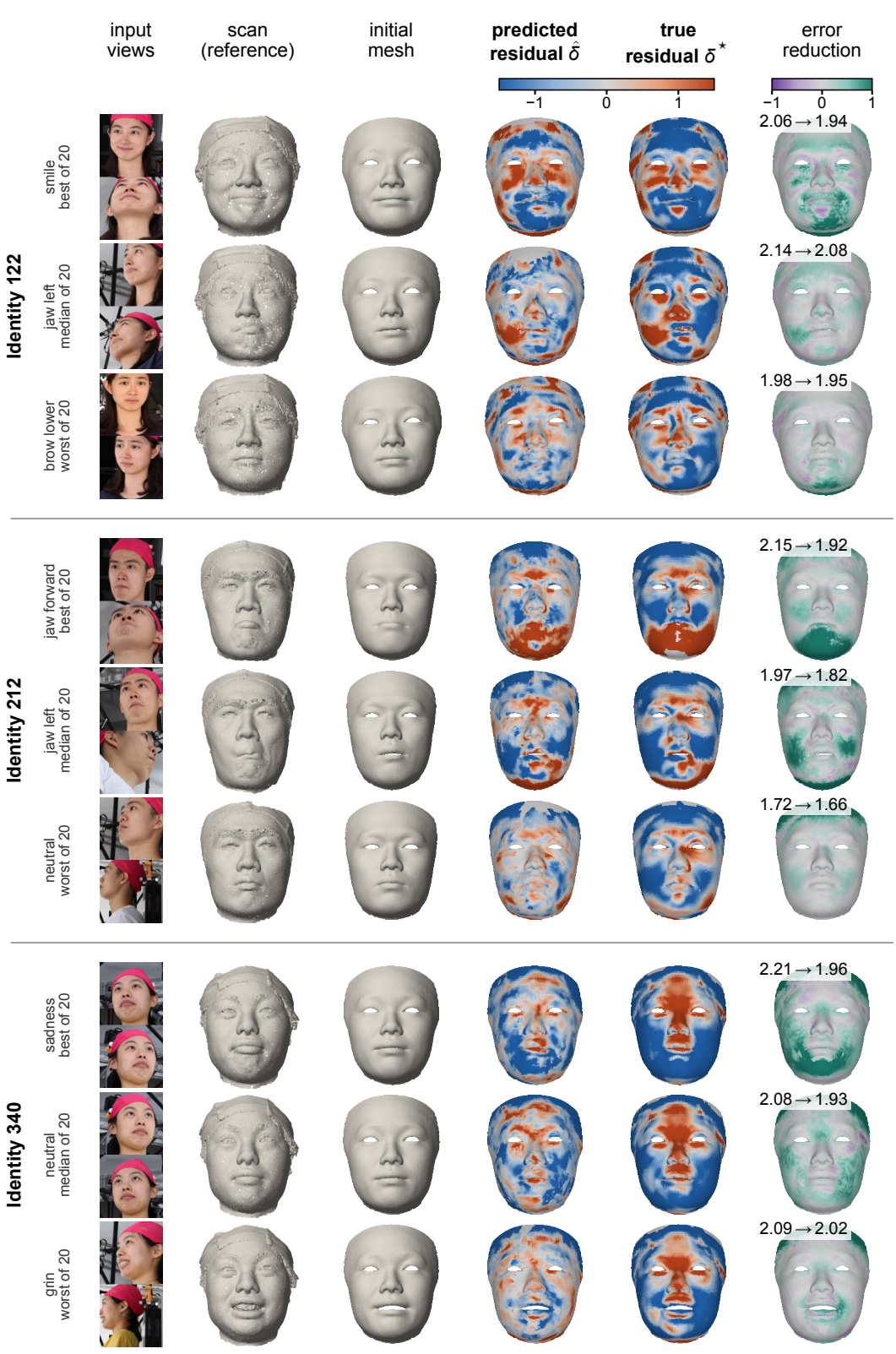


Figure 5: **FaceScape gallery.** Representative successful, median and challenging $\alpha = 0.5$ scans of each of the three identities, out-of-fold start and decoder (c.u.). All 60 scans of these identities improve.

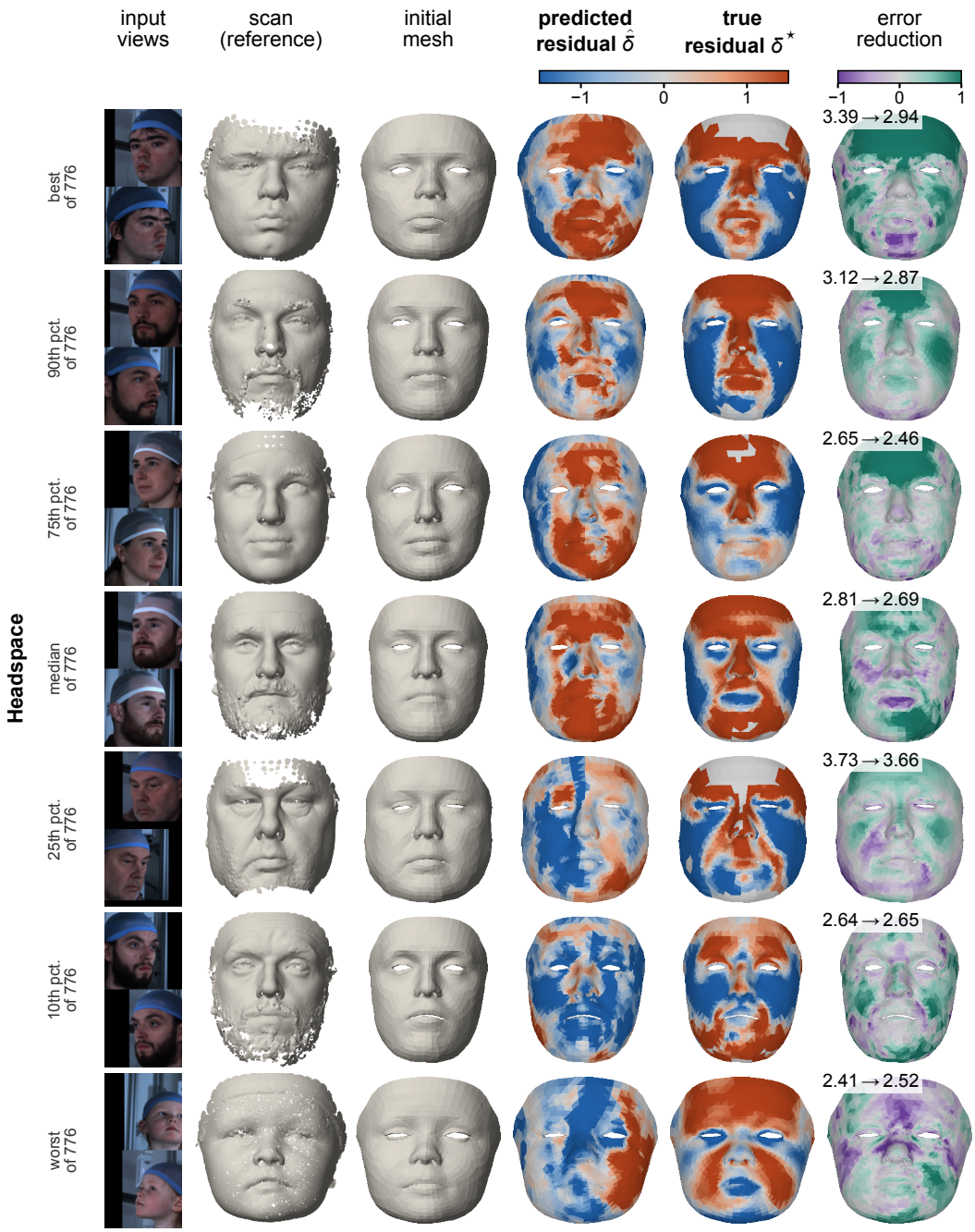


Figure 6: **Headspace gallery.** Pixel3DMM start, FLAME fit (mm). Cases spanning the distribution of subject-level gains, from best (top) to worst (bottom).