TopoRig: Topology-Agnostic Facial Rigging via Multi-Source Supervision

Andrew Fleet[1,4] Soroush Mehraban[2,3,4] Vida Adeli[2,3,4] Cole Clifford[2] Babak Taati[3,4]
[1]Queen's University [2]Pickford AI [3]University of Toronto [4]Vector Institute

Abstract

*Automatic facial rigging across heterogeneous mesh topologies remains challenging because high-quality expression supervision is often tied to canonical templates, while deformation transfer to arbitrary meshes can introduce geometric artifacts and correspondence errors. We present **TopoRig**, a topology-agnostic facial rigging framework that predicts FACS-conditioned deformations directly on input mesh vertices while preserving the original topology. Starting from the ICT FaceKit expression model, we construct complementary supervision from accurate but template-biased common-topology rigs, topology-diverse but noisier transferred rigs, and targeted image-based cues for controls poorly captured by geometric transfer. TopoRig combines local surface geometry, landmark-relative semantic features, global shape context, and FACS controls to predict per-vertex displacements. We train on 3,496 generated identities using 45 non-gaze expression controls from the 53-control ICT FaceKit vocabulary. On held-out identities and unseen mesh topologies, TopoRig more faithfully reproduces the reference expression space than prior neural facial-rigging methods, while qualitative results show consistent localized deformations across diverse character geometries. Ablations demonstrate that semantic landmark features and complementary supervision improve cross-identity and cross-topology generalization. Overall, TopoRig amortizes heterogeneous and imperfect expression supervision into a single topology-preserving deformation model.*

1. Introduction

Facial rigging transforms a static neutral face into a controllable asset capable of producing a range of expressions [11, 17]. It is an important component of creating digital characters for user-controlled animation, performance retargeting, generated avatars, speech-driven animation, and stylized expression manipulation [12, 18, 21, 25]. Facial rigs commonly represent expressions using blendshapes corresponding to the Facial Action Coding System (FACS) [4, 8, 9], providing interpretable controls that can be adjusted individually or combined to form more complex expressions. However, manually constructing high-quality facial rigs requires substantial artist effort, resulting in high production costs and long production cycles [6, 14, 23].

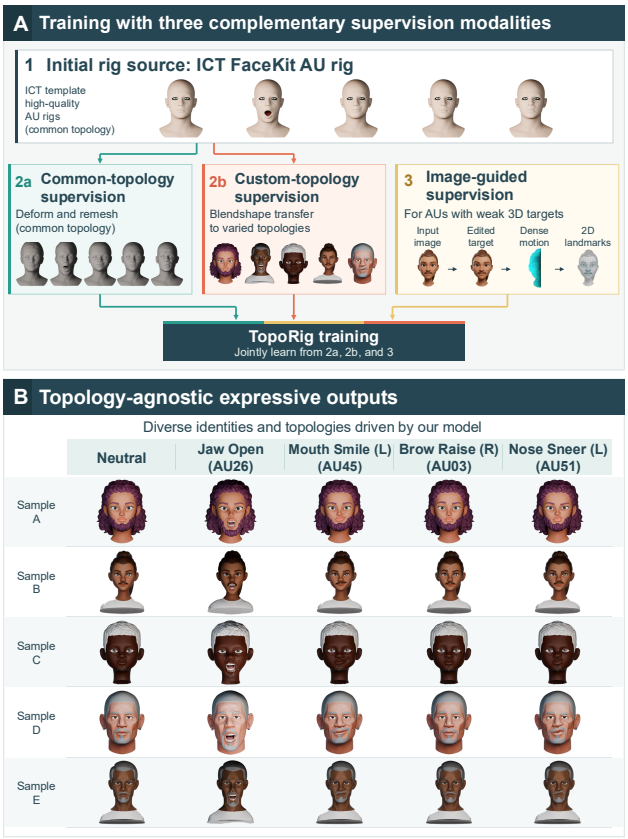

Figure 1. **Overview of TopoRig.** **(A)** Starting from the ICT FaceKit AU rig, we construct complementary supervision from common-topology fitted rigs, expression transfer to custom topologies, and targeted image-based cues for AUs with unreliable 3D targets. TopoRig jointly learns from these sources. **(B)** At inference time, it produces AU-conditioned deformations for diverse identities and previously unseen mesh topologies while preserving the input topology.

Traditional automatic methods reduce this effort by fitting a predefined facial template and transferring its blend-

shapes to a target mesh, but this process often relies on template registration, dense correspondence, or topology normalization [1, 3, 16]. Recent neural approaches instead predict facial deformations directly, with some supporting expression transfer across meshes with different connectivity and other combining global expression controls with localized deformation [2, 15, 17]. However, methods built on a canonical mesh representation, such as AU-Blendshape, remain limited to a shared topology. Template-registration approaches such as OmniFaceRig support more diverse input meshes, but construct the final facial rig through fitted template geometry and blendshape transfer [12, 22]. RigAnyFace supports varied input topology and expands training through 2D supervision, enabling facial rigging across diverse mesh structures [15].

Despite the progress, a central challenge in topology-agnostic facial rigging is obtaining supervision that is simultaneously accurate, geometrically diverse, and applicable across heterogeneous mesh connectivity. To address this challenge, we construct a multi-source training corpus using the open-source ICT FaceKit expression model [13] together with diverse neutral 3D faces generated by an off-the-shelf image-to-3D model [10]. We derive 3D expression supervision through two complementary routes. First, we fit the ICT FaceKit template to the generated identities, producing accurate and internally consistent expression targets on a shared topology, although the resulting geometry remains biased toward the template. Second, we transfer the ICT FaceKit expression deformations directly onto the generated meshes, preserving their identity-specific geometry while substantially increasing topology diversity. Although the transfer procedure includes geometric smoothing and expression-specific corrections, local correspondence ambiguities can still produce imperfect targets, particularly around closely separated facial surfaces such as the eyelids and lips. Moreover, the procedure requires substantial per-identity processing. We therefore use these transferred meshes as supervision for an amortized deformation model rather than as the final rigging procedure. For action units that are poorly captured by geometric transfer, we additionally incorporate targeted image-based supervision obtained using an off-the-shelf image editing model [5].

We propose **TopoRig**, a topology-agnostic facial rigging framework that learns a single deformation model from complementary supervision sources. Rather than relying on one automatically constructed rig representation, TopoRig combines accurate but template-biased fitted rigs, topology-diverse but noisier transferred rigs, and targeted image-based supervision for controls that are difficult to obtain reliably through geometric transfer. The network conditions each input vertex on local geometry, optional landmark-relative semantic features, global facial geometry, and the requested FACS control, and directly predicts expression-dependent displacements on the original mesh vertices. This preserves the input vertex count, connectivity, and UV parameterization while amortizing the expensive per-identity fitting and transfer process into a single forward pass.

Fig. 1 summarizes the construction of the three supervision sources and the topology-preserving inference process. In summary, our main contributions are:

- **Complementary multi-source supervision:** a training strategy that derives complementary supervision from a common AU rig through fitted common-topology targets, deformation transfer to identity-specific topologies, and targeted image-based cues for controls with unreliable 3D targets.
- **Amortized topology-preserving facial deformation:** a FACS-conditioned deformation model that jointly learns from these heterogeneous supervision sources and predicts expressions directly on previously unseen input meshes without running the fitting and transfer pipeline at inference time.
- **Large-scale training and analysis:** a training corpus containing 3,496 generated identities, using 45 non-gaze controls from the 53-control ICT FaceKit vocabulary, together with evaluations and ablations studying cross-topology generalization, semantic landmark features, complementary training sources, and data scale.

2. Related Work

Automatic facial rigging has traditionally relied on registering a canonical template and transferring blendshape or skinning deformations to a target mesh [1, 3, 16], often requiring reliable correspondence and per-character processing. Neural approaches instead learn deformation mappings across identities and mesh structures.

Neural Face Rigging (NFR) introduced a deformation autoencoder trained using interpretable controls from a linear 3D morphable model and detailed deformations from 4D facial captures [17]. This combination allows NFR to retain user-controllable expression parameters while reproducing fine-grained details from real facial performances.

Neural Face Skinning later combined a global expression representation with localized facial deformation [2]. The method predicts per-vertex skinning weights that localize the influence of a global expression representation to different facial regions. These weights are learned using facial segmentation labels, while FACS-based blendshapes provide interpretable expression supervision.

AUBlendNet instead predicts 32 personalized AU blendshape bases from a FLAME-based neutral identity mesh [12, 14]. Its accompanying AUBlendSet dataset contains artist-created AU blendshapes for 500 identities. Because all identities share the fixed FLAME topology of 5,023 vertices, AUBlendNet does not directly address facial meshes

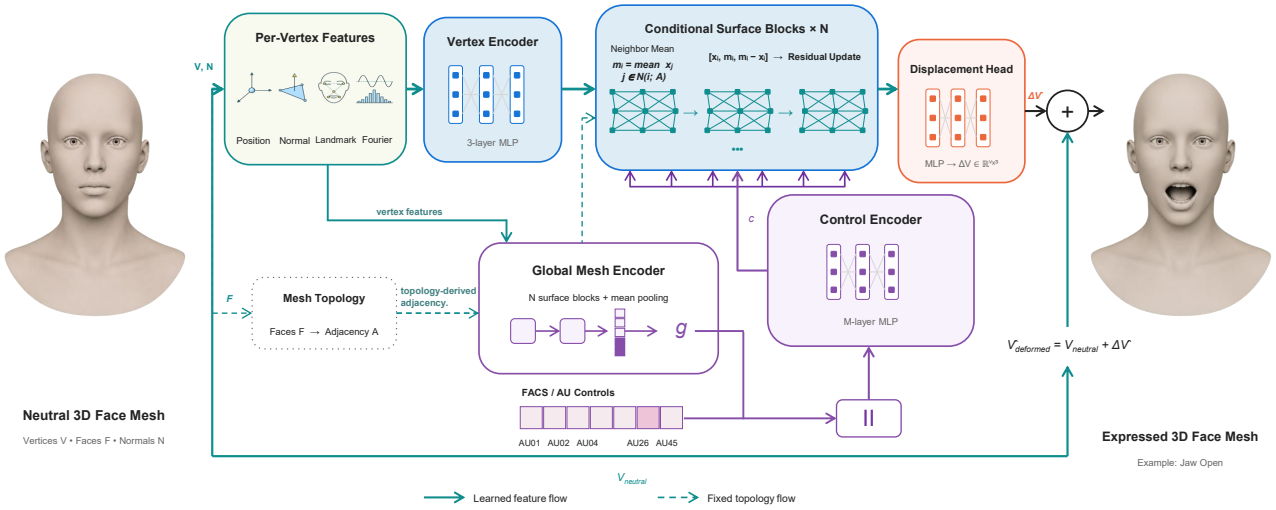

Figure 2. **TopoRig architecture.** Per-vertex geometric, landmark, and Fourier features are processed using the input mesh connectivity, while a global mesh representation and AU control jointly condition the deformation network. The model predicts a 3D displacement for each input vertex, preserving the original topology.

with arbitrary topology.

RigAnyFace directly predicts FACS-conditioned displacements on the vertices of an input neutral facial mesh using a deformation network built on DiffusionNet [15, 19]. The DiffusionNet blocks are modified to incorporate FACS conditioning, while a global encoder allows information to be shared across disconnected components such as the face and eyeballs. To reduce its dependence on expensive artist-rigged data, RigAnyFace combines 3D supervision from professionally rigged meshes with image-based supervision generated for unrigged meshes. Since its predicted displacements are applied directly to the original vertices, the input topology is preserved.

Most recently, OmniFaceRig introduced a template-based pipeline that converts a static surface-only character into an inner-mouth-aware facial rig [22]. The resulting rig contains up to 155 FACS blendshapes together with procedurally fitted teeth, gums, and tongue. Its selected facial template is rigidly and non-rigidly registered to the input before the blendshapes and inner-mouth components are transferred. Unlike RigAnyFace, which predicts displacements directly on the input vertices and therefore preserves their connectivity, OmniFaceRig constructs the rigged facial region through template fitting and fusion. Consequently, although it accepts diverse input topologies, it does not preserve the original facial connectivity.

TopoRig is most closely related to RigAnyFace in that both predict FACS-conditioned deformations directly on the vertices of the input mesh and therefore preserve its original connectivity. Our focus, however, is on constructing and combining complementary supervision from a single canonical AU rig. We derive accurate but template-biased common-topology targets, topology-diverse transferred targets on identity-specific meshes, and targeted image-based supervision for controls where geometric transfer is unreliable. TopoRig then learns a single amortized deformation model from these heterogeneous sources, avoiding the per-identity fitting and transfer procedure at inference time.

3. Method

3.1. Overview

TopoRig is a topology-preserving facial rigging framework that predicts expression-dependent deformations for a neutral mesh. Given neutral vertex positions $\mathbf{V}$, vertex normals $\mathbf{N}$, vertex landmark features $\mathbf{L}$, mesh connectivity $\mathcal{T}$, and an action-unit control vector $\mathbf{c}_{\mathrm{AU}}$, the network predicts a three-dimensional displacement, $\widehat{\Delta\mathbf{V}}$, for every input vertex:

$$f_\theta\left(\mathbf{V}, \mathbf{N}, \mathbf{L}, \mathcal{T}, \mathbf{c}_{\mathrm{AU}}\right) = \widehat{\Delta\mathbf{V}}. \tag{1}$$

The expressed mesh is obtained by adding the predicted displacements to the neutral vertices:

$$\widehat{\mathbf{V}}_{\mathrm{expr}} = \mathbf{V} + \widehat{\Delta\mathbf{V}}. \tag{2}$$

The network combines local surface geometry, facial landmark features, global facial geometry, and the requested ex-

pression. Since the predicted displacements are applied directly to the original vertices, TopoRig preserves the vertex count, faces, and connectivity of the input mesh.

Here, *topology-agnostic* means that TopoRig does not require a fixed canonical topology, vertex ordering, or correspondence across input meshes; the connectivity of each individual mesh is still used to define its local neighborhoods. The overall architecture is illustrated in Fig. 2.

3.2. Per-Vertex Input Features

Let the neutral mesh contain n_v vertices and n_f triangular faces, where both values may vary between meshes. It is represented by vertex positions $\mathbf{V} \in \mathbb{R}^{n_v \times 3}$, unit vertex normals $\mathbf{N} \in \mathbb{R}^{n_v \times 3}$, and face indices $\mathcal{T} \in \mathbb{N}^{n_f \times 3}$.

TopoRig can optionally augment each vertex with facial landmark features that provide semantic information about nearby facial regions. A landmark detector provides a set of semantically consistent 3D facial locations with fixed indices. For each mesh vertex i, we select its K_{L} nearest landmarks. For each selected landmark, we encode five values: three coordinates for the 3D offset from the vertex to the landmark, one value for their Euclidean distance, and one normalized landmark index. Concatenating these values over the K_{L} landmarks gives the feature $\mathbf{l}_i \in \mathbb{R}^{5K_{\mathrm{L}}}$. During training, we improve robustness to incomplete landmark information using two forms of landmark dropout. Modality dropout removes all landmark features, while region dropout removes landmarks from a randomly selected facial region before the nearest-landmark features are constructed. In both cases, the per-vertex input dimensionality remains fixed.

The relative offsets, distances, and landmark indices provide semantic localization cues without requiring vertex correspondence, helping the network identify facial regions across meshes with different vertex orderings and connectivity.

We normalize the vertex coordinates of each mesh by subtracting the mesh centroid and scaling by the maximum vertex distance from the centroid, yielding $\widetilde{\mathbf{v}}_i$. To provide additional spatial information, we apply a Fourier positional encoding with two frequency bands to these normalized coordinates [20]. Let $\gamma(\widetilde{\mathbf{v}}_i)$ denote this encoding. The complete per-vertex feature is then

$$\mathbf{z}_i = [\widetilde{\mathbf{v}}_i,\, \mathbf{n}_i,\, \mathbf{l}_i,\, \gamma(\widetilde{\mathbf{v}}_i)] \in \mathbb{R}^{58}, \tag{3}$$

where $\mathbf{n}_i$ is the unit surface normal and $\mathbf{l}_i$ is the landmark-relative feature defined above.

3.3. Conditional Deformation Network

TopoRig represents facial expressions using a 53-dimensional control vector, denoted by $\mathbf{c}_{\mathrm{AU}}$, whose entries correspond to the expression controls provided by ICT FaceKit. These controls follow the AU-based naming used by ICT FaceKit throughout this work. In all training and evaluation experiments, each target expression activates a single control and is therefore represented by a one-hot vector.

A vertex encoder E_{v}, implemented as an MLP, maps each per-vertex input feature $\mathbf{z}_i$ to an initial feature for the deformation network,

$$\mathbf{x}_i^{(0)} = E_{\mathrm{v}}(\mathbf{z}_i). \tag{4}$$

In parallel, inspired by the pooled global representation used in RigAnyFace [15], a separate global encoder summarizes the geometry of the complete neutral mesh into a mesh-level representation. While the deformation branch operates primarily on local vertex neighborhoods, this global representation provides identity-level shape context that is shared across the entire mesh. Its input MLP $E_{\mathrm{g,in}}$ maps the same per-vertex feature to

$$\mathbf{h}_i^{(0)} = E_{\mathrm{g,in}}(\mathbf{z}_i). \tag{5}$$

The triangular faces induce a bidirectional vertex adjacency $\mathbf{A} \in \{0,1\}^{n_v \times n_v}$, with self-edges included for all vertices.

At layer l of the global encoder, each vertex computes the mean feature over itself and its adjacent vertices,

$$\mathbf{m}_{i,\mathrm{g}}^{(l)} = \frac{\sum_{j=1}^{n_v} A_{ij}\mathbf{h}_j^{(l)}}{\sum_{j=1}^{n_v} A_{ij}}. \tag{6}$$

The current feature, neighbourhood mean, and their difference are combined by an MLP $E_{\mathrm{g}}^{(l)}$ and added residually to the current representation,

$$\mathbf{u}_i^{(l)} = E_{\mathrm{g}}^{(l)}\left([\mathbf{h}_i^{(l)}, \mathbf{m}_{i,\mathrm{g}}^{(l)}, \mathbf{m}_{i,\mathrm{g}}^{(l)} - \mathbf{h}_i^{(l)}]\right), \tag{7}$$

$$\mathbf{h}_i^{(l+1)} = \mathrm{LayerNorm}\left(\mathbf{h}_i^{(l)} + \mathbf{u}_i^{(l)}\right). \tag{8}$$

After L_{g} such layers, an output MLP $E_{\mathrm{g,out}}$ is applied to each vertex feature and the resulting features are averaged to obtain a global representation of the neutral identity,

$$\mathbf{g} = \frac{1}{n_v}\sum_{i=1}^{n_v} E_{\mathrm{g,out}}\left(\mathbf{h}_i^{(L_{\mathrm{g}})}\right). \tag{9}$$

The global representation $\mathbf{g}$ is concatenated with the AU control vector $\mathbf{c}_{\mathrm{AU}}$ and passed through a control MLP E_{ctrl} to produce the expression condition

$$\mathbf{c} = E_{\mathrm{ctrl}}([\mathbf{c}_{\mathrm{AU}}, \mathbf{g}]). \tag{10}$$

The condition $\mathbf{c}$ is then broadcast to every vertex in the deformation network. In this way, each local deformation is conditioned not only on the requested expression but also on the global geometry of the input identity.

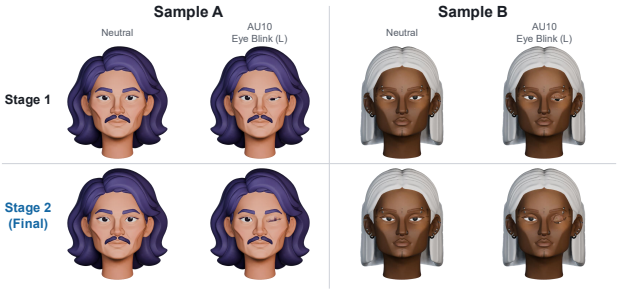

Figure 3. **Effect of Stage 2 image-guided refinement.** For left eye blink, Stage 1 inherits weak closure from unreliable transferred targets, while Stage 2 image supervision produces stronger and more complete eyelid closure on both validation identities.

The main deformation network then applies L_{d} conditional aggregation layers, starting from the vertex features $\mathbf{x}_i^{(0)}$. Using the same adjacency matrix, each layer first computes

$$\mathbf{m}_i^{(l)} = \frac{\sum_{j=1}^{n_v} A_{ij}\mathbf{x}_j^{(l)}}{\sum_{j=1}^{n_v} A_{ij}}. \tag{11}$$

The current vertex feature, its neighbourhood mean, and their difference are then combined by an MLP $E_{\mathrm{surf}}^{(l)}$ to form a local feature,

$$\mathbf{s}_i^{(l)} = E_{\mathrm{surf}}^{(l)}\left([\mathbf{x}_i^{(l)}, \mathbf{m}_i^{(l)}, \mathbf{m}_i^{(l)} - \mathbf{x}_i^{(l)}]\right). \tag{12}$$

The expression condition $\mathbf{c}$ is concatenated with this local feature and processed by a second MLP $E_{\mathrm{cond}}^{(l)}$. The resulting feature is added residually to the current vertex representation,

$$\mathbf{x}_i^{(l+1)} = \mathrm{LayerNorm}\left(\mathbf{x}_i^{(l)} + E_{\mathrm{cond}}^{(l)}\left([\mathbf{s}_i^{(l)}, \mathbf{c}]\right)\right). \tag{13}$$

This operation propagates information according to the input mesh connectivity while conditioning each vertex on the requested expression. After L_{d} conditional layers, the final vertex feature is concatenated with the expression condition and passed through a displacement MLP E_{disp},

$$\Delta\hat{\mathbf{v}}_i = E_{\mathrm{disp}}\left([\mathbf{x}_i^{(L_{\mathrm{d}})}, \mathbf{c}]\right). \tag{14}$$

The displacement head outputs a three-dimensional displacement for each input vertex. Its final layer is initialized to zero, so the untrained network initially predicts zero displacement and reproduces the neutral input mesh.

3.4. Training Objectives

TopoRig is trained in two stages. Stage 1 learns the FACS-conditioned deformation model from mesh-supervised samples. Stage 2 initializes from Stage 1 and jointly uses mesh- and image-supervised samples. For mesh-supervised samples, the primary vertex loss $\mathcal{L}_{\mathrm{vtx}}$ supervises the predicted displacement field against the target expression deformation. We additionally use a landmark loss $\mathcal{L}_{\mathrm{lmk}}$ and an active-landmark loss $\mathcal{L}_{\mathrm{active}}$ to increase supervision around semantically important facial regions and landmarks associated with the selected AU, respectively. The mesh objective is

$$\mathcal{L}_{\mathrm{mesh}} = \mathcal{L}_{\mathrm{vtx}} + \lambda_{\mathrm{lmk}}\mathcal{L}_{\mathrm{lmk}} + \lambda_{\mathrm{active}}^{\mathrm{eff}}\mathcal{L}_{\mathrm{active}} + \mathcal{L}_{\mathrm{geom}}. \tag{15}$$

The geometric regularizer $\mathcal{L}_{\mathrm{geom}}$ combines displacement smoothness, relative edge-length preservation, surface-normal consistency, and displacement-magnitude penalties:

$$\begin{aligned}\mathcal{L}_{\mathrm{geom}} = {} & \lambda_{\mathrm{smooth}}\mathcal{L}_{\mathrm{smooth}} + \lambda_{\mathrm{edge}}\mathcal{L}_{\mathrm{edge}} \\ & + \lambda_{\mathrm{normal}}\mathcal{L}_{\mathrm{normal}} + \lambda_{\mathrm{mag}}\mathcal{L}_{\mathrm{mag}}.\end{aligned} \tag{16}$$

Landmark features provide semantic localization, while the landmark losses directly supervise deformation near expression-relevant facial regions.

For image-supervised samples, target optical flow between the neutral and expressed images is estimated using WAFT [24]. We apply image-based supervision only to a subset of eight eye-region controls for which the transferred 3D blendshapes are noisy or insufficiently reliable, particularly for eyelid closure and opening.

In addition to optical flow, we use an eyelid objective $\mathcal{L}_{\mathrm{eyelid}}$ that supervises normalized projected eyelid motion and eyelid-gap reduction, together with a localized depth-anchor term. We further regularize the predicted deformation with an image-branch geometric regularizer $\mathcal{L}_{\mathrm{geom}}^{\mathrm{img}}$ and an outside-region anchor loss $\mathcal{L}_{\mathrm{anchor}}$ that suppresses motion away from the supervised facial region. The image objective is

$$\mathcal{L}_{\mathrm{image}} = \lambda_{\mathrm{flow}}\mathcal{L}_{\mathrm{flow}} + \mathcal{L}_{\mathrm{eyelid}} + \mathcal{L}_{\mathrm{geom}}^{\mathrm{img}} + \lambda_{\mathrm{anchor}}\mathcal{L}_{\mathrm{anchor}}. \tag{17}$$

During Stage 2, mesh- and image-supervised samples are trained jointly. The loss applied to each sample depends on its available supervision:

$$\mathcal{L} = \mathcal{L}_{\mathrm{mesh}} + \mathcal{L}_{\mathrm{image}}. \tag{18}$$

Detailed definitions of the individual mesh and image loss terms are provided in Sec. B of the supplementary material. At inference time, the network directly predicts a displacement field for a requested AU control on the neutral mesh. Since only the vertex positions are modified, the original vertex count, faces, and connectivity are preserved.

4. Experiments

4.1. Datasets and Metrics

Datasets. We use 3,496 facial identities generated using Nano Banana [5] and reconstruct their neutral meshes using Pixal3D [10]. The identities are divided into 2,796 training, 350 validation, and 350 test identities.

We obtain 3D supervision by fitting ICT FaceKit [13] to each reconstructed identity. This produces both a common-topology ICT mesh and an expression-transferred mesh that preserves the original Pixal3D topology. The original-topology meshes are simplified to approximately 100,000 faces (denoted **Pixal3D-100k**), and MediaPipe landmarks [7] are transferred to the simplified geometry. Further construction details are provided in Sec. A. The full ICT FaceKit control vocabulary contains 53 expression controls. For mesh-supervised experiments, we use the 45 non-gaze controls available in our fitted ICT stream. Four of these have unreliable transferred targets and are therefore omitted from the Pixal3D stream, leaving 41 controls. Image supervision is applied to eight eye-region controls and uses 17,504 expression pairs from 2,779 identities.

The Pixal3D expression meshes are automatically constructed by transferring ICT expression deformations to the original Pixal3D topology. We therefore treat them as *transferred reference targets*, rather than independent ground-truth rigs. Evaluation on this set measures how well a method preserves the canonical expression space while generalizing to previously unseen identities and mesh topologies. Importantly, lower error does not necessarily imply higher perceptual rig quality, since the transferred references may contain local correspondence artifacts. We therefore use this evaluation as a measure of transfer fidelity and topology generalization, complemented by qualitative results on diverse character meshes.

The held-out evaluation set contains 350 identities, yielding 15,750 ICT and 14,350 Pixal3D mesh–expression samples. All methods use the same identities, controls, preprocessing, scale normalization, and evaluation masks.

Evaluation Metrics. Predictions are evaluated as displacement fields rather than absolute vertex positions. Each neutral head is standardized to a height of 240 mm, and evaluation is restricted to vertices whose reference displacement exceeds 0.01 mm. We report coordinate-wise MAE and the 95th percentile (Q95) of per-vertex Euclidean displacement error, averaged equally across identity–control pairs. MAE captures average deformation error, while Q95 emphasizes larger localized deviations.

4.2. Implementation Details

We use $K_L = 8$ nearest landmarks, two Fourier frequency bands, and two aggregation layers in both the global encoder and deformation network. Training uses AdamW with a global batch size of 16. Stage 1 is trained for 20 epochs with mesh supervision, while Stage 2 is initialized from Stage 1 and refined for 10 epochs using joint mesh and image supervision. Image supervision is applied to eight eye-region controls, and Stage 2 additionally includes mesh-only refinement for selected higher-error controls. Full optimization, loss-weight, augmentation, and refinement settings are provided in the supplementary material.

Table 1. **Quantitative comparison with prior facial-rigging methods.** Pixal3D-100k errors measure agreement with automatically transferred ICT reference expressions on held-out identities and topologies, rather than with independently authored ground-truth rigs.

Method	ICT		Pixal3D-100k	
	MAE (mm) ↓	Q95 (mm) ↓	MAE (mm) ↓	Q95 (mm) ↓
NFS [2]	1.476	4.358	1.416	4.260
NFR [17]	1.928	5.480	1.273	4.145
TopoRig (Ours)	**0.400**	**1.257**	**0.274**	**0.975**

4.3. Comparison with baselines

We compare TopoRig with Neural Face Rigging (NFR) [17] and Neural Face Skinning (NFS) [2] using their publicly released pretrained checkpoints. RigAnyFace [15], although closely related, is not included because neither its implementation nor pretrained models are publicly available. For each control, we provide the corresponding ICT expression together with the neutral test mesh to each baseline. Since both baselines also predict a neutral reconstruction, expression displacements are measured relative to their predicted neutral output. All methods use the same test identities, shared controls, preprocessing, evaluation masks, scale normalization, and metrics described in Sec. 4.1.

As shown in Tab. 1, TopoRig substantially outperforms both prior methods on common-topology ICT meshes and on identity-specific Pixal3D meshes. Relative to the strongest baseline, TopoRig reduces MAE by 72.9% on ICT and 78.5% on Pixal3D-100k, with similarly large reductions in Q95. On Pixal3D-100k, the lower errors indicate that TopoRig more faithfully reproduces the transferred reference expressions across nonuniform, identity-specific topologies; they should not be interpreted as an independent measure of perceptual rig quality. Overall, TopoRig more accurately reproduces the reference expression targets on previously unseen identities and mesh structures.

Fig. 4 provides a qualitative comparison for jaw opening, showing more consistent deformation across the ICT template and identity-specific Pixal3D topologies than NFR and NFS. Fig. 5 further demonstrates generalization to stylized in-the-wild characters whose facial geometry and appearance differ substantially from the training identities. Additional qualitative results across held-out valida-

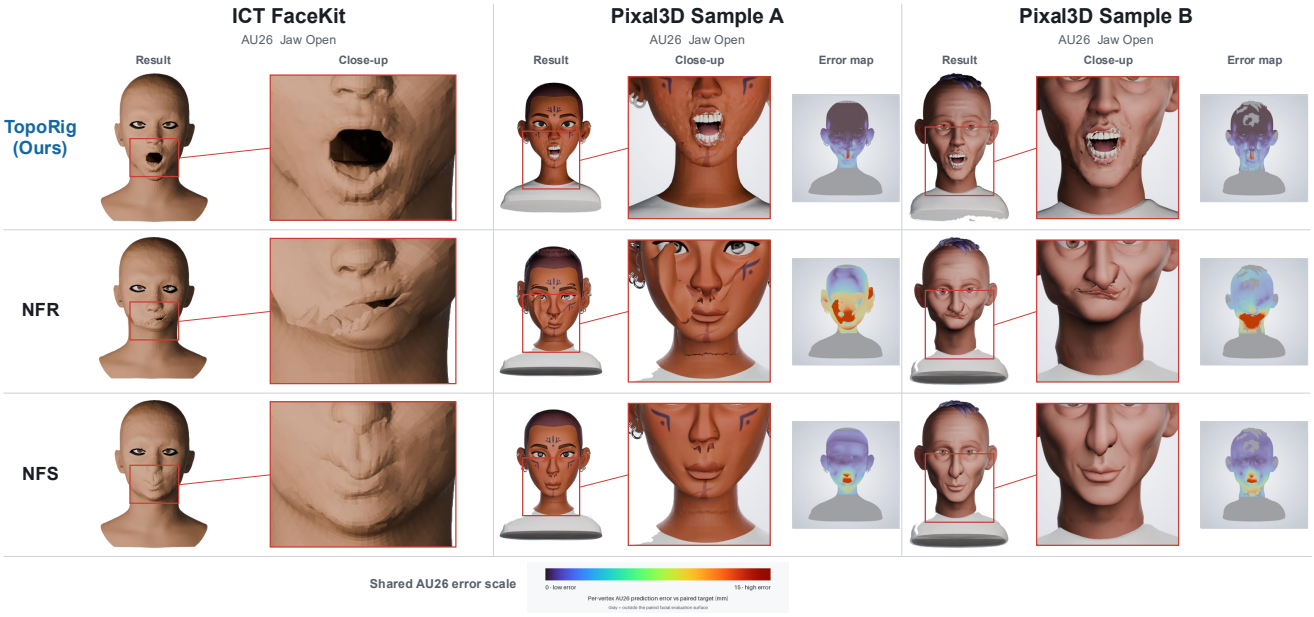

Figure 4. **Qualitative comparison with NFR and NFS for jaw opening.** Columns show the ICT FaceKit topology and two Pixal3D identities with custom topologies. For Pixal3D, error maps show per-vertex Euclidean distance to the transferred reference target on a shared 0–15 mm scale; gray denotes vertices outside the evaluated region. TopoRig produces more consistent deformation across common and custom topologies.

Table 2. **Ablation and training analysis.** Effects of input features, supervision sources, training-set size, and Stage 2 refinement. Errors are computed on moving vertices after normalization to a 240 mm head height relative to the held-out reference expression targets.

Model	ICT		Pixal3D-100k	
	MAE (mm) ↓	Q95 (mm) ↓	MAE (mm) ↓	Q95 (mm) ↓
No landmark input	0.641	1.940	0.373	1.456
No Fourier encoding	0.446	1.438	0.282	1.008
ICT only	**0.372**	**1.089**	1.086	4.153
Pixal3D only	1.165	3.974	0.286	1.039
25% data	0.768	2.260	0.351	1.255
50% data	0.512	1.737	0.308	1.115
75% data	0.494	1.624	0.291	1.044
Full Stage 1	0.416	1.333	0.275	0.983
Full Stage 2 (TopoRig)	0.400	1.257	**0.274**	**0.975**

tion identities and facial controls are provided in Sec. C of the supplementary material.

4.4. Ablation and Training Analysis

Tab. 2 analyzes the input representation, the value of complementary supervision and data diversity, and the effect of Stage 2 refinement.

Landmark-relative features are the dominant input cue. Relative to Full Stage 1, removing them increases MAE by 54% on ICT and 36% on Pixal3D-100k, whereas removing Fourier encoding causes only a small degradation. This indicates that semantic localization is particularly important for cross-topology generalization, with Fourier features providing a complementary benefit.

Training on either the ICT or transferred Pixal3D supervision alone produces clear domain-specific behavior, while joint training performs well across both common and identity-specific topologies. Performance also improves consistently with additional training identities, indicating that both complementary mesh supervision and identity diversity support generalization.

Stage 2 changes aggregate MAE and Q95 only slightly because it targets a small subset of controls with unreliable transferred 3D supervision. Its benefit is more evident qualitatively: as shown in Fig. 3, image-guided refinement produces stronger and more complete eyelid closure than mesh supervision alone.

4.5. Discussion and Limitations

The results highlight a useful trade-off between the different supervision sources used by TopoRig. Common-topology ICT targets provide consistent expression supervision, but constrain the training geometry to a canonical mesh structure. In contrast, transferred Pixal3D targets expose the model to substantially more varied identity-specific geometry and connectivity, at the cost of correspondence noise introduced by deformation transfer. The single-source ablations in Tab. 2 reflect this trade-off: training on either source alone produces strong domain-specific behavior, whereas joint training achieves substantially more balanced perfor-

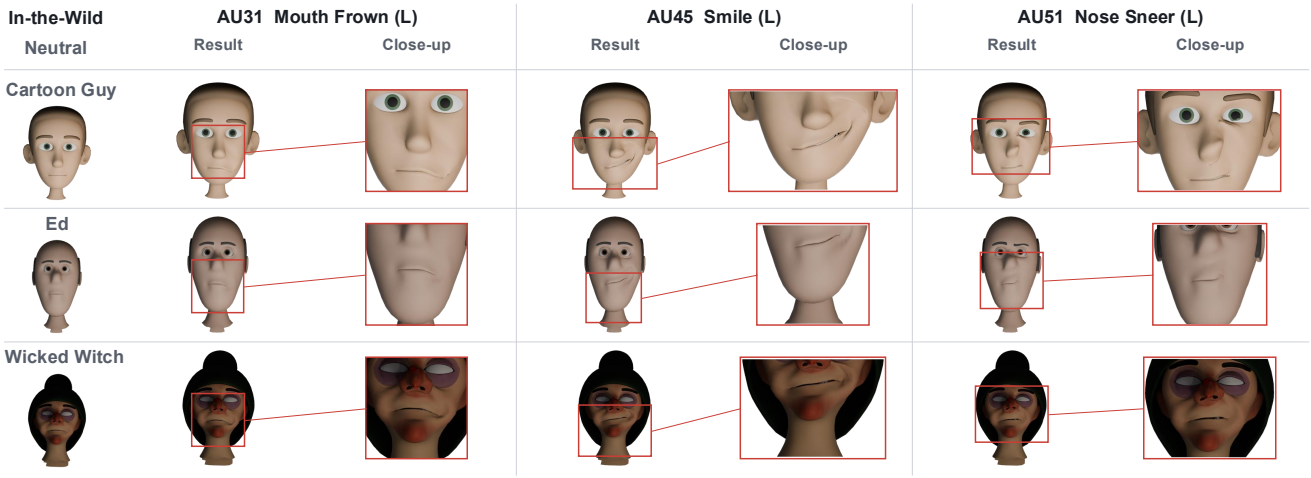

Figure 5. **Generalization to in-the-wild character meshes.** TopoRig produces localized AU-conditioned deformations on three stylized characters with geometry and appearance distinct from the training identities, while preserving their character-specific shape. The character meshes are publicly available CC0 assets from BlendSwap: *Cartoon Guy V1.2* by mjedewaard, *Ed* from *Cartoon Character Pack 1* by VMComix, and *Wicked Witch* by squibblejack. These meshes are used only for qualitative evaluation and are not part of the TopoRig training data.

mance across the two mesh domains. These results are consistent with the two supervision sources providing complementary geometric biases, although the ablation does not fully disentangle this effect from the additional supervision available under joint training.

The evaluation should nevertheless be interpreted with several limitations. Because the Pixal3D expression targets are themselves constructed by automatic deformation transfer, they are not independently authored ground-truth rigs. Quantitative performance on Pixal3D-100k therefore measures agreement with the transferred canonical expression space and generalization across unseen mesh structures, rather than perceptual rig quality in isolation. The in-the-wild examples in Fig. 5 provide additional qualitative evidence that the learned deformation model can operate on character geometries outside the generated training distribution, but these examples likewise lack artist-authored target rigs and are therefore not used for quantitative evaluation.

Our current expression supervision also focuses on individual AU controls: all training and quantitative evaluation targets activate a single control. Although the network accepts an AU control vector, compound expressions have not been systematically supervised or evaluated. In addition, image-guided refinement currently targets eight eye-region controls for which transferred 3D supervision is particularly unreliable. Extending this supervision to other difficult facial regions, evaluating combinations of controls, and collecting artist-authored rigs or perceptual judgments would provide stronger measures of animation quality beyond deformation transfer fidelity.

5. Conclusion

We presented TopoRig, a topology-agnostic facial rigging framework that predicts AU-conditioned deformations directly on the vertices of an input mesh while preserving its original topology. TopoRig learns from complementary supervision sources comprising accurate common-topology targets, topology-diverse transferred targets, and targeted image-based cues for controls with unreliable 3D supervision.

Across held-out identities, TopoRig achieves substantially lower deformation error than the evaluated neural baselines on both common-topology ICT meshes and identity-specific Pixal3D meshes. Qualitative results demonstrate consistent localized deformations across diverse identities and stylized in-the-wild character meshes. Our ablations show that landmark-relative semantic features, complementary mesh supervision, and increased identity diversity contribute to cross-topology generalization, while image-guided refinement improves localized deformations such as eyelid closure.

The supervision sources play complementary roles: canonical-topology targets provide consistent expression supervision, transferred meshes introduce geometric and topological diversity, and targeted image cues address regions where geometric transfer is unreliable. Overall, TopoRig amortizes this heterogeneous supervision into a single topology-preserving model, avoiding per-identity fitting and transfer while preserving the connectivity and identity-specific structure of unseen character meshes.

References

[1] Emma Carrigan, Eduard Zell, Cédric Guiard, and Rachel McDonnell. Expression packing: As-few-as-possible training expressions for blendshape transfer. In *Computer Graphics Forum*, pages 219–233. Wiley Online Library, 2020. 2

[2] Sihun Cha, Serin Yoon, Kwanggyoon Seo, and Junyong Noh. Neural face skinning for mesh-agnostic facial expression cloning. *Computer Graphics Forum*, 44(2):e70009, 2025. 2, 6

[3] Ludovic Dutreve, Alexandre Meyer, Veronica Orvalho, and Saida Bouakaz. Easy rigging of face by automatic registration and transfer of skinning parameters. In *International Conference on Computer Vision and Graphics*, pages 333–341. Springer, 2010. 2

[4] Paul Ekman and Wallace V. Friesen. *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Consulting Psychologists Press, Palo Alto, California, 1978. 1

[5] Michael Gerstenhaber. Building on the bananas momentum of generative media models on google cloud. Google Cloud Blog, 2025. 2, 6

[6] Jing Hou, Dongdong Weng, Zhihe Zhao, Ying Li, and Jixiang Zhou. Neutral facial rigging from limited spatiotemporal meshes. *Electronics*, 13(13):2445, 2024. 1

[7] Yury Kartynnik, Artsiom Ablavatski, Ivan Grishchenko, and Matthias Grundmann. Real-time facial surface geometry from monocular video on mobile GPUs. *arXiv preprint arXiv:1907.06724*, 2019. 6

[8] John P Lewis and Ken-ichi Anjyo. Direct manipulation blendshapes. *IEEE Computer Graphics and Applications*, 30(4):42–50, 2010. 1

[9] John P Lewis, Ken Anjyo, Taehyun Rhee, Mengjie Zhang, Frederic H Pighin, and Zhigang Deng. Practice and theory of blendshape facial models. *Eurographics (State of the Art Reports)*, 1(8):2, 2014. 1

[10] Dong-Yang Li, Wang Zhao, Yuxin Chen, Wenbo Hu, Meng-Hao Guo, Fang-Lue Zhang, Ying Shan, and Shi-Min Hu. Pixal3D: Pixel-aligned 3d generation from images. In *ACM SIGGRAPH 2026 Conference Papers*, 2026. 2, 6

[11] Hao Li, Thibaut Weise, and Mark Pauly. Example-based facial rigging. *Acm transactions on graphics (tog)*, 29(4): 1–6, 2010. 1

[12] Hao Li, Ju Dai, Feng Zhou, Kaida Ning, Lei Li, and Junjun Pan. AU-Blendshape for fine-grained stylized 3d facial expression manipulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. 1, 2

[13] Ruilong Li, Karl Bladin, Yajie Zhao, Chinmay Chinara, Owen Ingraham, Pengda Xiang, Xinglei Ren, Pratusha Prasad, Bipin Kishore, Jun Xing, and Hao Li. Learning formation of physically-based face attributes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3410–3419, 2020. 2, 6, 1

[14] Tianye Li, Timo Bolkart, Michael J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4d scans. *ACM Transactions on Graphics*, 36 (6), 2017. 1, 2

[15] Wenchao Ma, Dario Kneubuehler, Maurice Chu, Ian Sachs, Haomiao Jiang, and Sharon X. Huang. RigAnyFace: Scaling neural facial mesh auto-rigging with unlabeled data. In *Advances in Neural Information Processing Systems*, 2025. 2, 3, 4, 6

[16] Hayato Onizuka, Diego Thomas, Hideaki Uchiyama, and Rin-ichiro Taniguchi. Landmark-guided deformation transfer of template facial expressions for automatic generation of avatar blendshapes. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 2100–2108. IEEE, 2019. 2

[17] Dafei Qin, Jun Saito, Noam Aigerman, Thibault Groueix, and Taku Komura. Neural face rigging for animating and retargeting facial meshes in the wild. In *ACM SIGGRAPH 2023 Conference Proceedings*, 2023. 1, 2, 6

[18] Alexander Richard, Michael Zollhöfer, Yandong Wen, Fernando De la Torre, and Yaser Sheikh. Meshtalk: 3d face animation from speech using cross-modality disentanglement. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1153–1162. IEEE, 2021. 1

[19] Nicholas Sharp, Souhaib Attaiki, Keenan Crane, and Maks Ovsjanikov. DiffusionNet: Discretization agnostic learning on surfaces. *ACM Transactions on Graphics*, 41(3), 2022. 3

[20] Matthew Tancik, Pratul P. Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan T. Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. In *Advances in Neural Information Processing Systems*, pages 7537–7547, 2020. 4

[21] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2387–2395, 2016. 1

[22] Chao Wang, Guangyao Ma, John Doublestein, Junming Chen, Yiming Lin, Zhaoen Su, Xiaomin Luo, Shiyang Cheng, Jie Shen, Doug Roble, Dilin Wang, Yilei Li, and Rakesh Ranjan. OmniFaceRig: Fully automatic inner-mouth-aware face rigging across diverse 3d character topologies. *arXiv preprint arXiv:2606.08043*, 2026. 2, 3

[23] Haoyu Wang, Haozhe Wu, Junliang Xing, and Jia Jia. Versatile face animator: Driving arbitrary 3d facial avatar in rgbd space. In *Proceedings of the 31st ACM international conference on multimedia*, pages 7776–7784, 2023. 1

[24] Yihan Wang and Jia Deng. WAFT: Warping-alone field transforms for optical flow. In *International Conference on Learning Representations (ICLR)*, 2026. 5

[25] Feng Xu, Jinxiang Chai, Yilong Liu, and Xin Tong. Controllable high-fidelity facial performance transfer. *ACM Transactions on Graphics (TOG)*, 33(4):1–11, 2014. 1

TopoRig: Topology-Agnostic Facial Rigging via Multi-Source Supervision

Supplementary Material

A. Template Fitting and Blendshape Transfer

To generate expression-capable training meshes while preserving the topology of each target identity, we use the ICT FaceKit Light model [13] as a canonical facial deformation template. ICT FaceKit provides a common facial topology together with 53 expression blendshapes, allowing a consistent expression space to be transferred across identities with different geometry and connectivity.

Given a neutral target mesh and the neutral ICT template, we first align the two meshes using corresponding facial landmarks. The ICT template is then non-rigidly deformed to match the target identity while retaining its original connectivity. The resulting identity deformation is applied consistently to the ICT neutral mesh and its expression shapes, preserving the expression displacement fields. Finally, the fitted ICT blendshape displacements are spatially transferred onto the vertices of the original target topology.

The overall procedure consists of four stages: (i) landmark-based similarity alignment, (ii) non-rigid ICT template fitting, (iii) blendshape-preserving identity deformation, and (iv) deformation-field transfer to the target topology.

A.1. ICT FaceKit Expression Template

We use ICT FaceKit Light [13], which provides a canonical neutral facial mesh and 53 expression blendshapes registered to a common topology. The model includes localized deformations of the brows, eyelids, cheeks, jaw, and mouth, with asymmetric expressions represented independently for the left and right sides of the face.

The ICT expression meshes share the same vertex ordering and connectivity as the neutral template. We therefore represent each expression k as a per-vertex displacement field relative to the neutral mesh. Let $\mathbf{v}_i^0$ denote vertex i of the neutral ICT mesh and $\mathbf{v}_i^k$ denote the corresponding vertex under expression k. The expression displacement is

$$\boldsymbol{\delta}_i^k = \mathbf{v}_i^k - \mathbf{v}_i^0. \tag{19}$$

These deformation fields form the canonical expression space that is subsequently adapted to each target identity.

A.2. Landmark-Based Similarity Alignment

Let

$$\mathcal{M}_{\mathrm{ICT}} = (\mathbf{V}_{\mathrm{ICT}}, \mathbf{F}_{\mathrm{ICT}}) \tag{20}$$

denote the neutral ICT template and

$$\mathcal{M}_{\mathrm{T}} = (\mathbf{V}_{\mathrm{T}}, \mathbf{F}_{\mathrm{T}}) \tag{21}$$

denote the neutral target mesh.

Before non-rigid fitting, the target mesh is placed into the coordinate frame of the ICT template using corresponding facial landmarks. Given shared landmark positions $\{\mathbf{p}_l^{\mathrm{ICT}}\}_{l=1}^{L}$ and $\{\mathbf{p}_l^{\mathrm{T}}\}_{l=1}^{L}$, we estimate a similarity transformation consisting of a scale s, rotation $\mathbf{R}$, and translation $\mathbf{t}$:

$$\hat{\mathbf{p}}_l^{\mathrm{T}} = s\mathbf{p}_l^{\mathrm{T}}\mathbf{R} + \mathbf{t}. \tag{22}$$

To improve robustness to inaccurate landmark correspondences, landmarks with large residual errors are removed using percentile-based trimming and the transformation is re-estimated using the remaining correspondences. The resulting transformation is then applied to the complete target mesh.

A.3. Non-Rigid ICT Template Fitting

Following global alignment, the ICT template is non-rigidly deformed to match the geometry of the target identity while retaining the original ICT mesh connectivity. Let $\mathbf{v}_i^0$ denote the original position of ICT vertex i and $\mathbf{v}_i$ its current fitted position. The corresponding identity displacement is

$$\mathbf{d}_i = \mathbf{v}_i - \mathbf{v}_i^0. \tag{23}$$

The fitting procedure combines surface correspondence, semantic landmark constraints, and topology-aware displacement smoothing.

Surface Correspondence. Each ICT vertex is associated with its nearest point on the aligned target surface. The resulting correspondence term can be interpreted as

$$\mathcal{L}_{\mathrm{surf}} = \sum_i \|\mathbf{v}_i - \mathrm{NN}_{\mathrm{T}}(\mathbf{v}_i)\|_2^2, \tag{24}$$

where $\mathrm{NN}_{\mathrm{T}}(\mathbf{v}_i)$ denotes the nearest target surface sample. Large-distance correspondences are removed using percentile-based trimming. A lower-weight reverse correspondence from the target surface to the ICT template is additionally used to encourage coverage of the target geometry.

Landmark Constraints. Nearest-surface correspondence alone does not enforce semantic alignment between facial structures. We therefore constrain ICT landmark vertices to their corresponding target landmark positions:

$$\mathcal{L}_{\mathrm{lm}} = \sum_{l=1}^{L} w_l \left\| \mathbf{v}_{i_l} - \hat{\mathbf{p}}_l^{\mathrm{T}} \right\|_2^2. \tag{25}$$

Landmark influence is propagated to nearby vertices using spatially decaying weights, providing smooth semantic guidance around the eyes, mouth, nose, and brows.

Topology-Aware Smoothing. To maintain a coherent surface during fitting, neighboring ICT vertices are encouraged to undergo similar identity displacements. For the one-ring neighborhood $\mathcal{N}(i)$ of vertex i, the mean neighboring displacement is

$$\bar{\mathbf{d}}_i = \frac{1}{|\mathcal{N}(i)|} \sum_{j \in \mathcal{N}(i)} \mathbf{d}_j. \tag{26}$$

The corresponding regularization can be written as

$$\mathcal{L}_{\mathrm{smooth}} = \sum_i \left\| \mathbf{d}_i - \bar{\mathbf{d}}_i \right\|_2^2. \tag{27}$$

The fitting process can therefore be interpreted as balancing

$$\mathcal{L}_{\mathrm{fit}} = \lambda_{\mathrm{surf}} \mathcal{L}_{\mathrm{surf}} + \lambda_{\mathrm{lm}} \mathcal{L}_{\mathrm{lm}} + \lambda_{\mathrm{smooth}} \mathcal{L}_{\mathrm{smooth}}, \tag{28}$$

although the implementation performs iterative weighted displacement updates rather than explicitly optimizing a scalar objective through gradient descent.

A.4. Geometric Refinement

After the initial non-rigid registration, the fitted displacement field is further refined to improve local surface correspondence. A higher-resolution sampling of the target surface is used for local detail projection, while additional smoothing and outlier removal suppress isolated artifacts.

Region-specific constraints are applied around sensitive structures such as the eyelids and lips, where closely separated surfaces can otherwise result in incorrect nearest-neighbor correspondences.

A.5. Blendshape-Preserving Identity Deformation

Let $\mathbf{v}_i^0$ denote the neutral ICT vertex and $\mathbf{v}_i^k$ the same vertex under expression k. The canonical expression displacement is

$$\boldsymbol{\delta}_i^k = \mathbf{v}_i^k - \mathbf{v}_i^0. \tag{29}$$

After fitting the neutral ICT template to the target identity, let the identity deformation be

$$\boldsymbol{\Delta}_i = \mathbf{v}_i^{\mathrm{fit}} - \mathbf{v}_i^0. \tag{30}$$

The same identity deformation is applied to both the neutral template and every expression shape:

$$\hat{\mathbf{v}}_i^0 = \mathbf{v}_i^0 + \boldsymbol{\Delta}_i, \tag{31}$$

$$\hat{\mathbf{v}}_i^k = \mathbf{v}_i^k + \boldsymbol{\Delta}_i. \tag{32}$$

Therefore,

$$\hat{\mathbf{v}}_i^k - \hat{\mathbf{v}}_i^0 = \left(\mathbf{v}_i^k + \boldsymbol{\Delta}_i\right) - \left(\mathbf{v}_i^0 + \boldsymbol{\Delta}_i\right) \tag{33}$$

$$= \mathbf{v}_i^k - \mathbf{v}_i^0 \tag{34}$$

$$= \boldsymbol{\delta}_i^k. \tag{35}$$

Thus, fitting changes the identity represented by the ICT template while retaining its canonical expression deformation fields.

A.6. Blendshape Transfer to the Target Topology

The fitted ICT mesh acts as an intermediate deformation template. To retain the connectivity of the original target mesh, the ICT expression displacement fields are transferred onto the target vertices.

For target vertex $\mathbf{q}_j$, a local set of neighboring fitted ICT vertices $\mathcal{N}_{\mathrm{ICT}}(j)$ is identified. The transferred displacement for expression k is

$$\boldsymbol{\delta}_{j,\mathrm{T}}^k = \sum_{i \in \mathcal{N}_{\mathrm{ICT}}(j)} w_{ji} \boldsymbol{\delta}_i^k. \tag{36}$$

The unnormalized interpolation weights are defined using a Gaussian kernel,

$$\tilde{w}_{ji} = \exp\left(-\frac{1}{2} \frac{d_{ji}^2}{\sigma^2} \right), \tag{37}$$

where d_{ji} denotes the Euclidean distance between target vertex j and fitted ICT vertex i. The normalized weights are

$$w_{ji} = \frac{\tilde{w}_{ji}}{\sum_{m \in \mathcal{N}_{\mathrm{ICT}}(j)} \tilde{w}_{jm}}. \tag{38}$$

Correspondences exceeding a maximum transfer distance are rejected. The resulting target expression is then

$$\mathbf{q}_j^k = \mathbf{q}_j^0 + \boldsymbol{\delta}_{j,\mathrm{T}}^k. \tag{39}$$

This allows the canonical ICT expression deformation space to be transferred to target meshes with different vertex counts and connectivity.

A.7. Semantic Expression Corrections

Spatial interpolation can produce ambiguous correspondences in regions containing closely separated surfaces, particularly around the eyes and mouth. We therefore apply expression-specific semantic constraints following deformation transfer.

For eye-related expressions, landmark-derived masks restrict transferred motion to the relevant eyelid and surrounding facial regions. Additional constraints suppress unintended deformation of nearby interior eye geometry, while blink expressions use a localized eyelid closure correction.

Mouth-related expressions are similarly constrained using lip landmarks. In particular, jaw-opening deformation is stabilized near the upper and lower lip boundaries to reduce incorrect correspondence between nearby mouth surfaces.

A.8. Resulting Training Meshes

The final representation preserves the geometry and connectivity of the generated target identity while inheriting the canonical ICT expression space:

$$\begin{aligned}&\text{Target identity} + \text{Target topology}\\ &\quad + \text{transferred ICT expression blendshapes.}\end{aligned} \tag{40}$$

These expression-specific target meshes are subsequently used as mesh supervision during training.

B. Training Loss Details

We provide the definitions of the mesh- and image-supervised loss terms introduced in Sec. 3.4. Let $\Delta\hat{\mathbf{v}}_i$ and $\Delta\mathbf{v}_i$ denote the predicted and target displacement of vertex i, respectively, and let $\hat{\mathbf{v}}_i = \mathbf{v}_i + \Delta\hat{\mathbf{v}}_i$.

B.1. Mesh-Supervised Losses

Vertex and landmark losses. We divide vertices into active and inactive sets according to the target displacement,

$$\mathcal{A} = \{i : \|\Delta\mathbf{v}_i\|_2 \geq \tau\}, \qquad \mathcal{I} = \{i : \|\Delta\mathbf{v}_i\|_2 < \tau\}, \tag{41}$$

with $\tau = 10^{-3}$ in normalized coordinates. For a vertex set $\mathcal{S}$, define

$$E(\mathcal{S}) = \frac{1}{3|\mathcal{S}|}\sum_{i\in\mathcal{S}} \|\Delta\hat{\mathbf{v}}_i - \Delta\mathbf{v}_i\|_1. \tag{42}$$

The primary vertex loss balances active and inactive regions and normalizes by the target expression magnitude,

$$\mathcal{L}_{\text{vtx}} = \frac{\frac{1}{2}\left[E(\mathcal{A}) + E(\mathcal{I})\right]}{\max\left(\frac{1}{|\mathcal{A}|}\sum_{i\in\mathcal{A}} \|\Delta\mathbf{v}_i\|_2, 10^{-3}\right)}, \tag{43}$$

where an empty region is omitted from the average.

Let $\mathcal{K}$ denote the MediaPipe landmark vertices and $\mathcal{K}_{\text{act}} = \mathcal{K} \cap \mathcal{A}$. The landmark terms are

$$\mathcal{L}_{\text{lmk}} = E(\mathcal{K}), \qquad \mathcal{L}_{\text{active}} = E(\mathcal{K}_{\text{act}}). \tag{44}$$

The active-landmark coefficient is adaptively balanced as

$$\lambda^{\text{eff}}_{\text{active}} = \min\left(20, \frac{5\,\mathcal{L}_{\text{vtx}}}{\mathcal{L}_{\text{active}} + \epsilon}\right), \tag{45}$$

while $\lambda_{\text{lmk}} = 5$.

Geometric regularization. Let $\mathcal{E}$ and $\mathcal{F}$ denote the mesh edges and triangular faces. We use

$$\mathcal{L}_{\text{smooth}} = \frac{1}{3|\mathcal{E}|}\sum_{(i,j)\in\mathcal{E}} \|\Delta\hat{\mathbf{v}}_i - \Delta\hat{\mathbf{v}}_j\|_2^2, \tag{46}$$

$$\mathcal{L}_{\text{edge}} = \frac{1}{|\mathcal{E}|}\sum_{(i,j)\in\mathcal{E}} \left(\frac{\hat{\ell}_{ij} - \ell_{ij}}{\max(\ell_{ij}, 10^{-4})}\right)^2, \tag{47}$$

$$\mathcal{L}_{\text{normal}} = \frac{1}{|\mathcal{F}|}\sum_{f\in\mathcal{F}} \left(1 - \mathbf{n}_f^\top \hat{\mathbf{n}}_f\right), \tag{48}$$

$$\mathcal{L}_{\text{mag}} = \frac{1}{n_v}\sum_i \left[\max\left(\|\Delta\hat{\mathbf{v}}_i\|_2 - 0.12, 0\right)\right]^2, \tag{49}$$

where ℓ_{ij} and $\hat{\ell}_{ij}$ are the neutral and deformed edge lengths, and $\mathbf{n}_f$ and $\hat{\mathbf{n}}_f$ are the corresponding unit face normals. These terms form $\mathcal{L}_{\text{geom}}$ as defined in Eq. (17); their weights are reported in Sec. 4.2.

B.2. Image-Supervised Losses

Optical-flow loss. WAFT provides a target flow $\mathbf{F}^{\text{tgt}}$, while differentiable projection of the neutral and predicted meshes gives $\mathbf{F}^{\text{pred}}$. For supervision mask M, we use the masked Charbonnier loss

$$\mathcal{L}_{\text{flow}} = \frac{\sum_p M_p \sum_{c\in\{x,y\}} \left(\sqrt{(F^{\text{pred}}_{p,c} - F^{\text{tgt}}_{p,c})^2 + \epsilon^2} - \epsilon\right)}{2\sum_p M_p}, \tag{50}$$

with $\epsilon = 0.01$.

Eyelid objective. Let $\mathbf{p}_k^0, \mathbf{p}_k^1$ denote the neutral and deformed projected mesh landmarks and $\mathbf{q}_k^0, \mathbf{q}_k^1$ the corresponding target landmarks. For image-derived targets, $\mathbf{q}_k^0$ and $\mathbf{q}_k^1$ are detected in the neutral and expressed images, respectively. Each domain is normalized by its neutral eye width,

$$\hat{\mathbf{m}}_k = \frac{\mathbf{p}_k^1 - \mathbf{p}_k^0}{w_{\text{mesh}}}, \qquad \mathbf{m}_k = \frac{\mathbf{q}_k^1 - \mathbf{q}_k^0}{w_{\text{img}}}. \tag{51}$$

We subtract the mean motion of the two eye corners from both quantities before comparison.

For paired upper–lower eyelid landmarks, we additionally measure the reduction in eyelid gap. Let g_k^0, g_k^1 denote the neutral and deformed projected mesh gaps and $\tilde{g}_k^0, \tilde{g}_k^1$ the corresponding target gaps. We use

$$\hat{r}_k = \frac{g_k^0 - g_k^1}{w_{\text{mesh}}}, \qquad r_k = \frac{\tilde{g}_k^0 - \tilde{g}_k^1}{w_{\text{img}}}. \tag{52}$$

Let ρ_β denote the Smooth-L1 loss with transition parameter β. The projected-motion and gap terms are

$$\mathcal{L}_{\text{motion}} = \rho_{0.05}(\hat{\mathbf{m}}, \mathbf{m}), \qquad \mathcal{L}_{\text{gap}} = \rho_{0.05}(\hat{\mathbf{r}}, \mathbf{r}). \tag{53}$$

For the two blink controls, we replace the detected expressed landmarks with a synthetic full-closure target constructed from the neutral projected mesh. Each paired upper eyelid landmark is moved 0.8 of the way toward its corresponding lower landmark, while the lower landmark is moved 0.2 of the way toward the upper landmark.

Depth anchor. Because projected supervision alone does not constrain motion along the depth direction, we penalize depth displacement within a localized eyelid region. Let $\mathcal{R}_{\text{depth}}$ denote the corresponding mesh-hop region and let z_i^0, z_i^1 be the neutral and deformed depth coordinates. We define

$$\mathcal{L}_{\text{depth}} = \frac{1}{|\mathcal{R}_{\text{depth}}|} \sum_{i \in \mathcal{R}_{\text{depth}}} \rho_{0.01}\left(\frac{z_i^1 - z_i^0}{w_{\text{mesh}}}, 0 \right). \tag{54}$$

The complete eyelid objective is

$$\mathcal{L}_{\text{eyelid}} = \lambda_{\text{eyelid}} \left(\mathcal{L}_{\text{motion}} + 2\mathcal{L}_{\text{gap}} + 2\mathcal{L}_{\text{depth}} \right), \tag{55}$$

where $\lambda_{\text{eyelid}} = 3$ for image-derived targets and $\lambda_{\text{eyelid}} = 5$ for the two blink controls.

Image-space geometric regularization. We additionally penalize overall displacement magnitude,

$$\mathcal{L}_{\text{disp}} = \frac{1}{3n_v} \sum_i \|\Delta\hat{\mathbf{v}}_i\|_2^2, \tag{56}$$

and reuse the displacement-smoothness, relative edge-length, and normal-consistency terms defined in Sec. B.1:

$$\begin{aligned} \mathcal{L}_{\text{geom}}^{\text{img}} = \lambda_{\text{disp}}\mathcal{L}_{\text{disp}} + \lambda_{\text{smooth}}\mathcal{L}_{\text{smooth}} \\ + \lambda_{\text{edge}}\mathcal{L}_{\text{edge}} + \lambda_{\text{normal}}\mathcal{L}_{\text{normal}}. \end{aligned} \tag{57}$$

Finally, letting $\mathcal{R}$ denote the supervised facial region, the outside-region anchor is

$$\mathcal{L}_{\text{anchor}} = \frac{1}{|\overline{\mathcal{R}}|} \sum_{i \notin \mathcal{R}} \|\Delta\hat{\mathbf{v}}_i\|_2^2. \tag{58}$$

The complete image objective is given in Eq. (18) of the main paper.

B.3. Training and Implementation Settings

Architecture and optimization. The per-vertex input has 58 channels: normalized position, surface normal, 40 landmark-relative features from the $K_L = 8$ nearest MediaPipe landmarks, and a 12-dimensional Fourier encoding with two frequency bands. Landmark indices are normalized by 467. Both the global encoder and deformation network use two aggregation layers. We use LayerNorm and SiLU activations with dropout 0.3, and initialize the final displacement layer to zero.

Training uses AdamW with zero weight decay, a global batch size of 16, and gradient clipping at 1.0. Stage 1 is trained for 20 epochs with initial learning rate 5×10^{-4}. Stage 2 initializes from the Stage 1 checkpoint and is trained for 10 epochs with initial learning rate 10^{-4}. Both stages use cosine learning-rate decay to 10^{-6}. Landmark modality dropout and landmark-region dropout are applied in both stages with probabilities 0.15 and 0.10, respectively; the latter randomly drops the left-eye, right-eye, or mouth landmarks.

Mesh-supervised losses. The active-motion threshold is 10^{-3} in normalized mesh coordinates. The landmark loss weight is $\lambda_{\text{lmk}} = 5$. The active-landmark term is adaptively balanced relative to $\mathcal{L}_{\text{vtx}}$ with a target loss ratio of 5 and a maximum effective weight of 20. The displacement-smoothness, relative edge-length, and normal-consistency weights are 0.05, 0.001, and 0.0005, respectively. The displacement-magnitude term has weight 5 and penalizes displacement exceeding 0.12 normalized units.

Image-supervised losses. Image supervision is applied to eight eye-region controls. WAFT target flow is computed at 512×512, while predicted mesh motion is rendered at 128×128 using nvdiffrast. The flow loss has weight 0.1.

For the image-branch geometric regularizer, the displacement, displacement-smoothness, relative edge-length, and normal-consistency weights are 0.05, 1.5, 0.05, and 0.05, respectively. The outside-region anchor has weight 500. The supervised region extends eight mesh hops around the relevant landmarks and twelve hops for cheek-squint controls. The eyelid depth anchor uses an eight-hop region. Before image-space supervision is evaluated, predicted displacements at coincident vertices are synchronized using a tolerance of 10^{-6}.

Stage 2 refinement. The mesh and image objectives have equal outer weights. Stage 2 additionally includes mesh-only refinement streams from both 3D supervision sources for AUs 24, 25, 26, 27, 28, 33, 34, and 40, with the same unit outer loss weight. These streams provide additional training exposure to controls with comparatively high

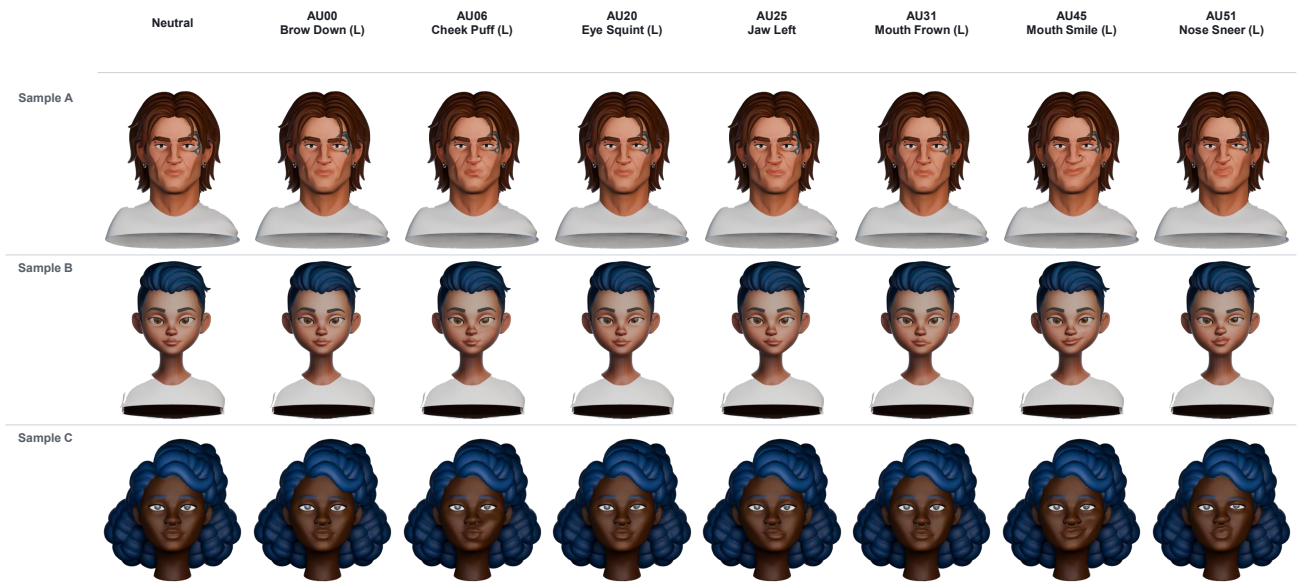

Figure S1. **Additional AU-conditioned expression results.** Rows show three held-out validation identities and columns show different unilateral and asymmetric facial controls. TopoRig produces localized deformations across diverse facial regions while preserving identity-specific geometry.

Stage 1 validation error and do not receive image supervision.

The checkpoint with the lowest mesh-only validation MAE is retained; image losses are excluded from checkpoint selection.

C. Additional Qualitative Results

Fig. S1 provides additional qualitative results on held-out validation identities. We visualize several unilateral and asymmetric facial controls spanning the brows, cheeks, eyes, jaw, mouth, and nose. Across identities with substantially different facial geometry and appearance, TopoRig produces localized AU-conditioned deformations while preserving the characteristic shape of each input identity.