Supplementary Material for "High-Quality Mesh Blendshape Generation from Face Videos via Neural Inverse Rendering"

A Discussion of Viewpoint Numbers

Our technique gives users the flexibility to use different numbers of cameras to achieve different levels of geometry quality. As shown in Fig. 1, we qualitatively demonstrate the geometric reconstruction results under single-view and four-view inputs, where under four-view input, we achieve more accurate geometric reconstruction of the cheek-puffing expression compared with the single-view setting. Additionally, in Tab. 1, we conducted a quantitative evaluation of reconstruction accuracy on the Multiface and NeRSemble datasets. We found that utilizing four-view inputs further reduces geometric errors compared to single-view inputs, indicating that four views provide a more accurate reconstruction of facial shape. We choose a four-view setup following [4] and believe that our method is also applicable to other sparse-view setups. Please also note that when compared with existing SOTA techniques in Fig. 2, our method demonstrates visually superior results in the single-view scenario, indicating that our method also works well with single-view input.

B Evaluation of the Neural Regressor

In our multi-phone captured data, we cannot precisely synchronize the recording initiation time across devices, so each device records a frame at a slightly different moment. As frames from distinct devices are not captured simultaneously, they record varying shapes in motion, necessitating different motion parameters. The temporal misalignment can be up to $1/60$ second since the frame rate is 30 FPS. Methods like dynamic time warping could be applied here to match the frames from different views. However, even if it can always find the temporally closest ones as matches, errors are still there as they are naturally recorded at different times (shown by Fig. 3). The incorporation of the neural regressor allows us to implicitly achieve synchronization across cameras under different views, ensuring temporal accuracy in tracking both expression coefficients and head pose. Without the regressor, temporal nearest frames across devices are erroneously assumed to depict an identical shape, resulting in diminished reconstruction accuracy. To assess the impact of the neural regressor, we simply "synchronize" videos from different views by finding the temporally nearest frame, and use this as a baseline to compare with our method. As depicted in Fig. 4, our method achieves more accurate geometric details, particularly around the ears. Furthermore, we quantitatively evaluate the regressor on a subject with ground truth

Fig. 1: Comparison of geometric reconstruction of the puff expression under single-view and four-view inputs.

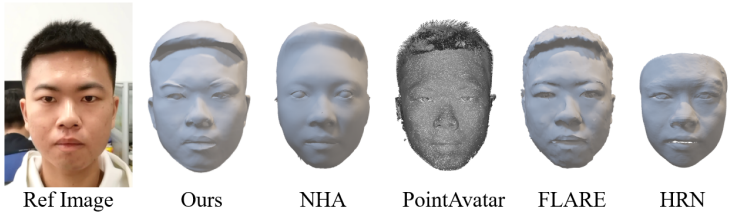

Fig. 2: Comparisons of geometry reconstruction between our method and other baselines under single-view input.

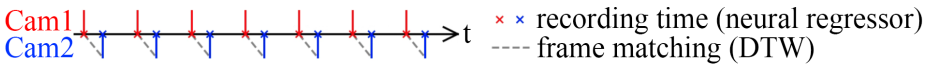

Fig. 3: Visualizations of the timeline corresponding to the unsynchronized multi-view inputs.

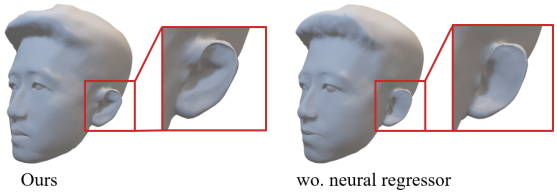

Fig. 4: Evaluating the effectiveness of the neural regressor.

Error(mm)	Multiface		NeRSemble	
	Mean	Std	Mean	Std
Single view	3.02	0.16	3.58	0.37
Four views	**2.31**	**0.05**	**2.73**	**0.26**

Table 1: Quantitative comparison of point-to-plane errors between four-view and single-view inputs.

Fig. 5: The quantitative reconstruction enhancement obtained by the neural regressor. The second and fourth columns are the reconstruction results, and the third and fifth columns are the point-to-plane error heatmaps compared to the ground truth scan.

scanned by light stage and observe that including the regressor yields enhanced accuracy, especially in foreheads, as shown in Fig. 5.

C Evaluation of Joint Optimizing Blendshapes and Expression Coefficients

Both blendshapes and expression coefficients affect facial expressions. To address the underdetermined nature of joint optimization, previous works [2, 3, 10] often adopt a two-stage approach, which stabilizes optimization but does not achieve optimal convergence. In the first stage, a generic expression model is used to fit the expression coefficients [8]. Subsequently, in the second stage, the expression coefficients are fixed and not optimized, focusing solely on optimizing the expression bases. We propose constraints on the semantic preservation of blendshapes, enabling joint optimization. To assess the necessity of jointly optimizing expression coefficients and blendshapes, we compare the results obtained by optimizing blendshapes alone with those achieved through joint optimization. When optimizing blendshapes alone, we derive estimates of the expression coefficients and head poses by fitting the original blendshapes of the ICT face model. Subsequently, we fix the expression coefficients and exclusively optimize the blendshapes. We illustrate the resulting neutral face and a "cheek puffing" blendshape in Fig. 6. In the baseline, artifacts manifest at the corners of the mouth due to inaccurate tracking of the cheek puffing expression during the coefficient estimation stage. Additionally, geometric details of the hair are inaccurately embedded into blendshapes.

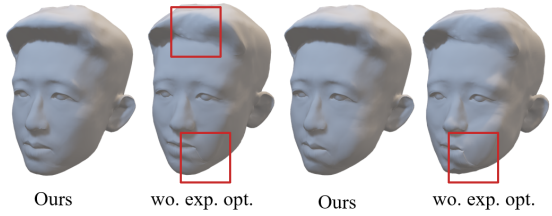

Fig. 6: Updated blendshapes for neutral and cheek-puffing expressions. The first and third columns show the results of joint optimization, while the second and fourth columns demonstrate the results using a two-stage optimization approach.

D Baseline Implementation Details

State-of-the-art facial avatar methods such as Neural Head Avatar [3], PointAvatar [9] and FLARE [1] inherently operate with monocular videos as input. To facilitate a fair comparison with our approach, we enhanced these frameworks to accommodate sparse multi-view videos as input.

Specifically, in the data processing phase of these methods, we utilize multi-view videos with calibrated camera parameters as input. During the tracking and optimization stages, we ensure consistency of facial parameters such as shape, poses, and expressions across different viewpoints and optimize these parameters by calculating losses based on images rendered from multiple viewpoints. After the modification, these methods demonstrated enhanced performance when utilizing multi-view inputs compared to their previous performance with monocular inputs, thus serving as an improved baseline for comparison against our method.

E Tetrahedral Connections Establishment

The ICT model features detailed cavities for the mouth, nose, and eyes, with significant facial expressions potentially leading to mesh interpenetration issues. To fully utilize the Laplacian constraint in surface deformation, we establish a connection between the internal cavities and the surface vertices. Specifically, we use tetgen [5] to preprocess the ICT's blendshapes, constructing a small closed volume space between the surface and corresponding internal sockets, where mesh interpenetration is likely to occur.

As illustrated in Fig. 7, we first manually specify the region in the head where we wish to fill tetrahedral connections. Tetgen fills this watertight enclosed area with tetrahedrons, whose edges form the connections we intend to establish between the internal cavity and surface vertices. Note that even for filling the same enclosed mesh, tetgen's filling results cannot be guaranteed to be topologically consistent each time. Since different blendshapes must share the same topology in our per-vertex deformation technique, we first fill the neutral face and then use the ARAP (As-Rigid-As-Possible) [6] algorithm to deform the tetrahedral

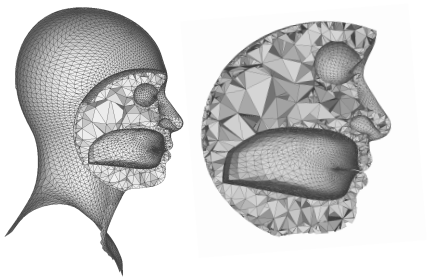

Fig. 7: The tetrahedral connections between the cavities and surface.

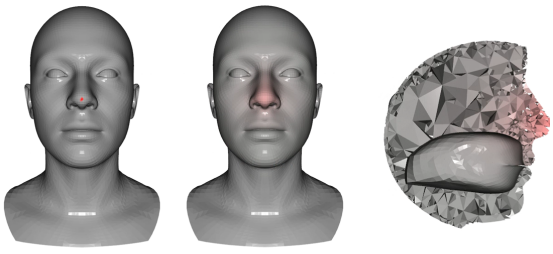

Fig. 8: Effectiveness of differential coordinates. The first column shows the gradient obtained at the tip of the nose without using differential coordinates (the gradients are indicated in red). The second column displays the propagated gradients with differential coordinates. The third column presents the gradients for the internal vertices.

vertices to match other blendshapes, creating a set of blendshapes with augmented and consistent topology. Moreover, we avoid filling the entire head to reduce computational complexity.

F Effective of Differential Coordinates

The utilization of differential coordinates propagates vertex gradients to neighboring vertices according to mesh connectivity, thus ensuring the smoothness of per-vertex deformation. As illustrated in Fig. 8, when the tip of the nose receives a gradient, our method propagates the gradient to the surrounding vertices, including the internally filled vertices. The smoothed gradients facilitate smooth geometry updates while preventing self-intersections.

G More Comparisons of Reconstruction Results

In Fig. 9, we show more reconstruction results to qualitatively compare our method and baselines. As illustrated in the main text, all methods accept inputs of four views.

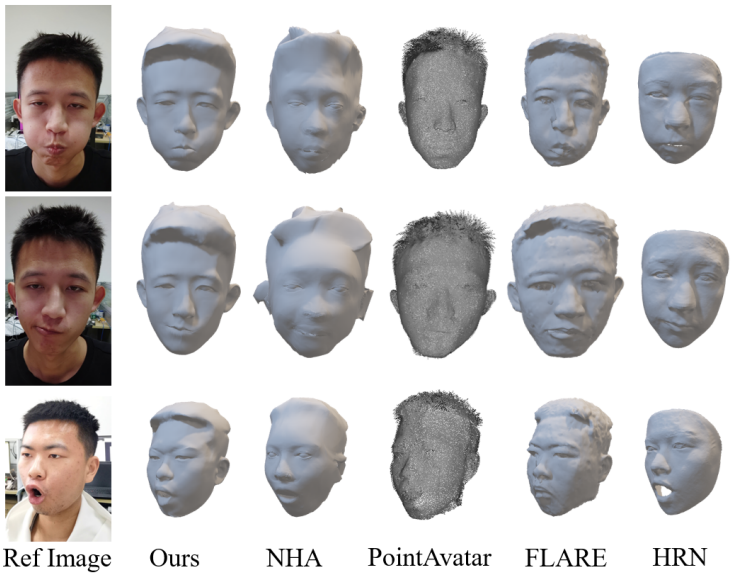

Fig. 9: More qualitative comparisons between our method and baselines. All methods accept inputs of four views.

H More Comparisons of Extracted Blendshapes

Here, we present a comparison between the blendshapes extracted by our method and deformation transfer [7] from the neutral. As shown in Fig. 10, our updated blendshape (fourth column) matches the reference image (third column) more closely than deformation transfer (fifth column), which only deforms details in neutral (second column) to the blendshape, lacking expression-specific details such as the wrinkles at the corners of the eyes and mouth (highlighted by red boxes).

I Additional Discussion on Limitations

The level of details is constrained by the topology and resolution of the ICT model. Increasing the mesh resolution could help cope this problem, but it comes with increasing training time. As shown in Fig. 11, subdividing the mesh results in better forehead wrinkles and nasolabial folds (second column to third column). Failure cases may occur when the movement is rapid enough to cause motion blur, as illustrated in Fig. 12. Global illumination effects are challenging for our neural radiance methods, such as moving shadows caused by varying head poses. As shown in Fig. 13, the nasal wing in the shadow (second column) is blurrier than the other side (third column).

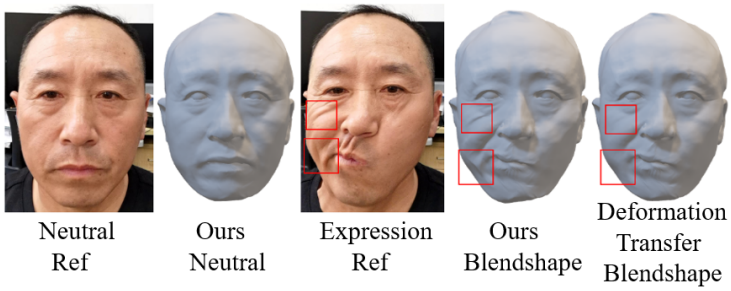

Fig. 10: Comparisons of extracted blendshapes with deformation transfer.

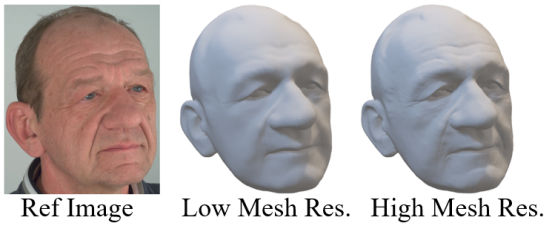

Fig. 11: Comparisons of reconstruction results with different mesh resolutions.

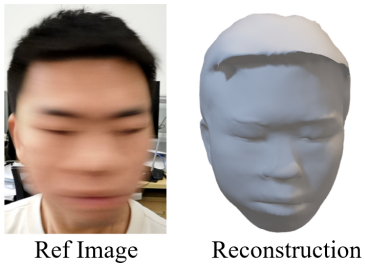

Fig. 12: Fast motion leads to quality degradation as blur occurs in the input.

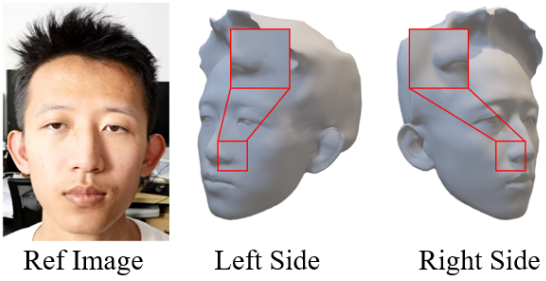

Fig. 13: Reconstruction results under strong shadows.

J More Details on Capture Protocols

The input videos contain around 800 frames with a resolution 504×896. Subjects are allowed to perform arbitrary expressions, with no need to follow a predefined expression sequence.

References

1. Bharadwaj, S., Zheng, Y., Hilliges, O., Black, M.J., Abrevaya, V.F.: FLARE: fast learning of animatable and relightable mesh avatars. CoRR **abs/2310.17519** (2023). https://doi.org/10.48550/ARXIV.2310.17519, https://doi.org/10.48550/arXiv.2310.17519
2. Gafni, G., Thies, J., Zollhöfer, M., Nießner, M.: Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8649–8658 (June 2021)
3. Grassal, P., Prinzler, M., Leistner, T., Rother, C., Nießner, M., Thies, J.: Neural head avatars from monocular RGB videos. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022. pp. 18632–18643. IEEE (2022). https://doi.org/10.1109/CVPR52688.2022.01810, https://doi.org/10.1109/CVPR52688.2022.01810
4. Shao, R., Zheng, Z., Tu, H., Liu, B., Zhang, H., Liu, Y.: Tensor4d: Efficient neural 4d decomposition for high-fidelity dynamic reconstruction and rendering. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023. pp. 16632–16642. IEEE (2023). https://doi.org/10.1109/CVPR52729.2023.01596, https://doi.org/10.1109/CVPR52729.2023.01596
5. Si, H.: Tetgen, a delaunay-based quality tetrahedral mesh generator. ACM Trans. Math. Softw. **41**(2) (feb 2015). https://doi.org/10.1145/2629697, https://doi.org/10.1145/2629697
6. Sorkine, O., Alexa, M.: As-rigid-as-possible surface modeling. In: Proceedings of the Fifth Eurographics Symposium on Geometry Processing. p. 109–116. SGP '07, Eurographics Association, Goslar, DEU (2007)
7. Sumner, R.W., Popović, J.: Deformation transfer for triangle meshes. ACM Transactions on graphics (TOG) **23**(3), 399–405 (2004)
8. Thies, J., Zollhöfer, M., Stamminger, M., Theobalt, C., Nießner, M.: Face2face: Real-time face capture and reenactment of RGB videos. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016. pp. 2387–2395. IEEE Computer Society (2016). https://doi.org/10.1109/CVPR.2016.262, https://doi.org/10.1109/CVPR.2016.262
9. Zheng, Y., Yifan, W., Wetzstein, G., Black, M.J., Hilliges, O.: Pointavatar: Deformable point-based head avatars from videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
10. Zielonka, W., Bolkart, T., Thies, J.: Instant volumetric head avatars. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023. pp. 4574–4584. IEEE (2023). https://doi.org/10.1109/CVPR52729.2023.00444, https://doi.org/10.1109/CVPR52729.2023.00444