Few-Shot Example-Driven Facial Modeling with Radiance Fields

Supplementary Material

APPENDIX

A. Potential social impact

Our motivation for this work was to enable the creation of 3D avatars that could be used as communication devices in the remote working era. As our approach stems from blendshapes [20], these avatars are easily adjustable via texture coloring and may be used for entertainment. We note, however, that the potential misuse of our work includes using it as deep fakes. We highly discourage such usage. One of our future directions includes detecting fake images generated by our method. At the same time, we highlight the importance of BlendFields—in the presence of closed technologies [2], [13], it is crucial to democratize techniques for personalized avatar creation. We achieve that by limiting the required data volume to train a single model. As history shows, when given an open, readily available technology for generative modeling of images [88], users can scrutinize it with unprecedented thoroughness, thus raising the general awareness of potential misuses.

B. Concurrent Works

Gao *et al*. [12] and Xu *et al*. [89] also use an interpolation between known expressions to combine multiple neural radiance fields trained for those expressions. However, their approach interpolates between grids of latent vectors [24] globally. The interpolation weights are taken from blendshape coefficients.

Zielonka *et al*. [79] use a parametric head model to canonicalize 3D points similarly to our ends. However, instead of building a tetrahedral cage around the head, they smoothly assign each face triangle to 3D points. Then they canonicalize points using transformations that each of the assigned triangles undergoes for a given expression. They concatenate 3D points with the expression code from FLAME [17] to model expression-dependent effects.

Both of these methods fall short compared to BlendFields and InstantBlendFields in terms of data efficiency—they require more than a thousand of accurately tracked frames while our approaches can be used with just as few as $K{=}5$ frames.

C. Additional results

1) Ablating number of expressions: We ablate over the number of used expressions during the training. To evaluate the effect of the number of expressions, we add consecutive frames to the training set (starting from a single, neutral one), *i.e*., the training set has $k{<}K$ expressions. We train BlendFields for such a set for each subject separately. We then average the results for a given k across subjects. We present the results in Tab. V. When selecting the training expressions, we aim to choose those that show all wrinkles when combined. We can see from Fig. 11 that if removed, *e.g*., the expressions with eyebrows raised, then the model cannot render wrinkles on the forehead. In summary, increasing the number of expressions improves the quality results with diminishing returns when $K{>}5$, while $K{=}5$ provides a sufficient trade-off between the data capture cost and the quality.

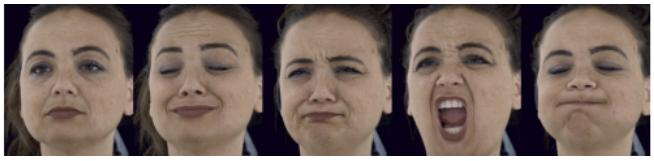

Fig. 10: **Training frames** – In Sec. IV, we show results for the BlendFields and InstantBlendFields trained on $K{=}5$ expressions. The images represent these expressions for one of the subjects. For each subject, we selected similar expressions to show all possible wrinkles when combined. Please note that we also include a "neutral" expression (the first from the left)—it is necessary to enable the learning of a face without any wrinkles.

2) Training frames: We present in Fig. 10 example training frames for one of the subjects. Each frame is a multi-view frame captured with $\approx$35 cameras (the number of available cameras varied slightly between subjects).

3) Additional qualitative results: We show in Fig. 12 results of baselines that do not rely on parametric models of the face [17]. Compared to BlendFields, they cannot render high-fidelity faces. The issue comes from the assumed data sparsity—those approaches rely on the interpolation in the training data. As we assume access to just a few frames, there is no continuity in the training data that would guide them to interpolate between known expressions. BlendFields presents superior results given novel expressions even with such a sparse dataset. See the attached video and `readme.html` file for more qualitative results.

Fig. 11: **Qualitative ablation over the number of training expressions** – We show qualitatively how the number of training expressions K affects the rendering quality. The first row shows the ground truth images. All other consecutive rows show the images rendered with InstantBlendFields while increasing the number of training expressions. The last row, K=5 corresponds to the results presented in the main part of the article. The subject's naming follows the convention introduced in the Multiface repository [83]. Please refer to Tab. V for quantitative results.

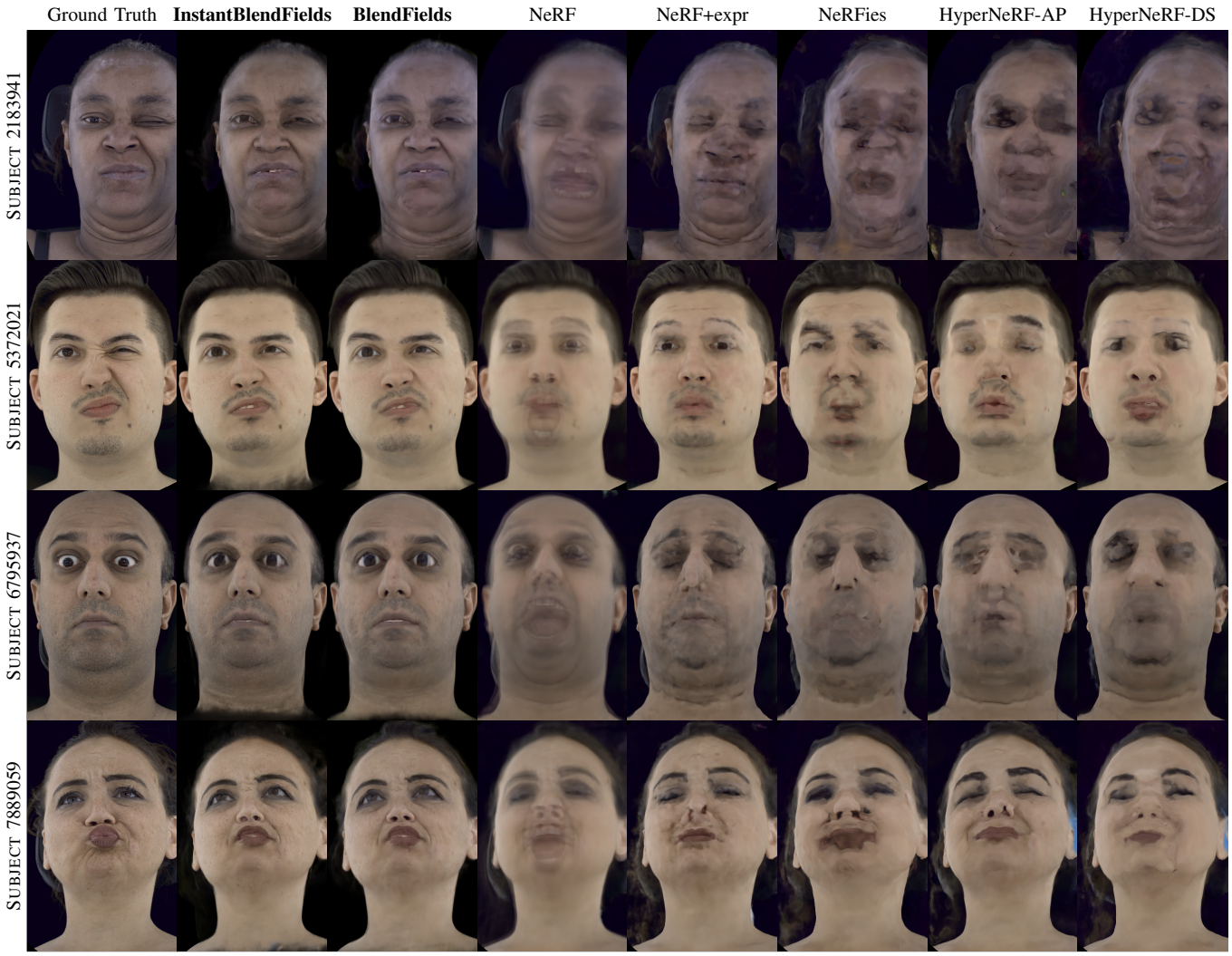

Fig. 12: **Comparison to strictly data-driven approaches** – We compare BlendFields to other baselines that do not rely on mesh-driven rendering: NeRF [3], NeRF conditioned on the expression code (NeRF+expr) [3], NeRFies [4], and HyperNeRF-AP/DS [5]. As a static model, NeRF converges to an average face from available (K=5) expressions. All other baselines exhibit severe artifacts compared to BlendFields and InstantBlendFields. Those baselines rely on the data continuity in the training set (*e.g.*, from a video), and cannot generalize to any other expression. Please see the supplemented video for the animations.