BlendFields: Few-Shot Example-Driven Facial Modeling

Supplementary Material

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 [26], 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 [7, 32], 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 [47], users can scrutinize it with unprecedented thoroughness, thus raising the general awareness of potential misuses.

B. Concurrent Works

Gao *et al*. [14] and Xu *et al*. [71] 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 [39] globally. The interpolation weights are taken from blendshape coefficients.

Zielonka *et al*. [81] 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 [27] to model expression-dependent effects.

C. Additional results

C.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. 4. When selecting the training expressions, we aim to choose those that show all wrinkles when combined. We can see from Fig. 9 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.

# expr.	Casual Expressions			Novel Pose Synthesis		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
$K=1$	27.5834	0.9028	0.0834	28.7589	0.9147	0.0806
$K=2$	27.6783	0.9026	0.0856	29.2859	0.9186	0.0803
$K=3$	27.9137	0.9054	0.0819	29.8551	0.9279	0.0728
$K=4$	27.8140	0.9055	0.0815	30.1543	0.9336	0.0701
$K=5$	28.0254	0.9110	0.0778	30.4721	0.9372	0.0688
$K=6$	28.0517	0.9091	0.0813	–	–	–
$K=7$	28.2004	0.9115	0.0823	–	–	–
$K=8$	28.2542	0.9124	0.0830	–	–	–

Table 4. **Number of training expressions** – We ablate over the number of training expressions. We evaluate the model on the captures from the Multiface dataset [66]. We run the model for each possible expression combination for a given K and average the results. The best results are colored in 🟥 and the second best in 🟨. Increasing the number of available training expressions consistently improves the results. However, using $K=5$ expressions saturates the quality and using $K>5$ brings diminishing improvements. We do not report "Novel Pose Synthesis" for $K>5$ as we use validation expressions and poses to train those models (refer to Sec. 4.1 for more details).

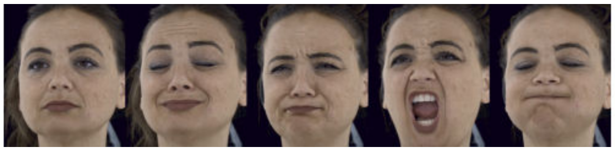

Figure 8. **Training frames** – In Sec. 4, we show results for the BlendFields trained on $K=5$ expressions. The images represent these expressions for one of the subjects. For each subject, we selected similar expressions to show all possible wrinkles when combined. Please note that we also include a "neutral" expression (the first from the left)—it is necessary to enable the learning of a face without any wrinkles.

C.2. Training frames

We present in Fig. 8 example training frames for one of the subjects. Each frame is a multi-view frame captured with ≈35 cameras (the number of available cameras varied slightly between subjects).

C.3. Quantitative results with background

We compare BlendFields and the baselines similarly to Sec. 4.1. However, in this experiment, we deliberately

Method	Casual Expressions			Novel Pose Synthesis		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
NeRF [36]	22.0060	0.6556	0.3222	23.8077	0.7448	0.2779
Conditioned NeRF [36]	21.0846	0.6280	0.3042	22.9991	0.7261	0.2362
NeRFies [42]	20.7004	0.6076	0.3579	23.0123	0.7253	0.2840
HyperNeRF-AP [43]	20.8105	0.6214	0.3504	22.8193	0.7185	0.2689
HyperNeRF-DS [43]	20.8847	0.6111	0.3656	23.0075	0.7259	0.2729
VolTeMorph$_1$ [15]	21.3265	0.7091	0.2706	22.3007	0.7795	0.2281
VolTeMorph$_{avg}$ [15]	22.0759	0.7755	0.2615	23.8974	0.8458	0.2302
BlendFields	22.8982	0.7954	0.2256	24.4432	0.8477	0.2052

Table 5. **Quantitative results without masking** – Similarly to Tab. 2, we compare BlendFields to other related approaches. However, we calculate the results over the whole image space, without removing the background. BlendFields and VolTeMorph [15] model the background as a separate NeRF-based [36] network. The points that do not fall into the tetrahedral mesh are assigned to the background. As the network overfits to sparse training views, it poorly extrapolates to novel expressions (as the new head pose or expression may reveal some unknown parts of the background) and views. At the same time, all other baselines do not have any mechanism to disambiguate the background and the foreground.

include the background in metric calculation. We show the results in Tab. 5. In all the cases, BlendFields performs best even though the method was not designed to model the background accurately. Additionally, as HyperNeRF [43], NeRFies [42], and NeRF [36] do not have any mechanism to disambiguate between the foreground and the background, the metrics are significantly worse when including the latter.

C.4. Additional qualitative results

We show in Fig. 10 results of baselines that do not rely on parametric models of the face [27]. 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 `index.html` file for more qualitative results.

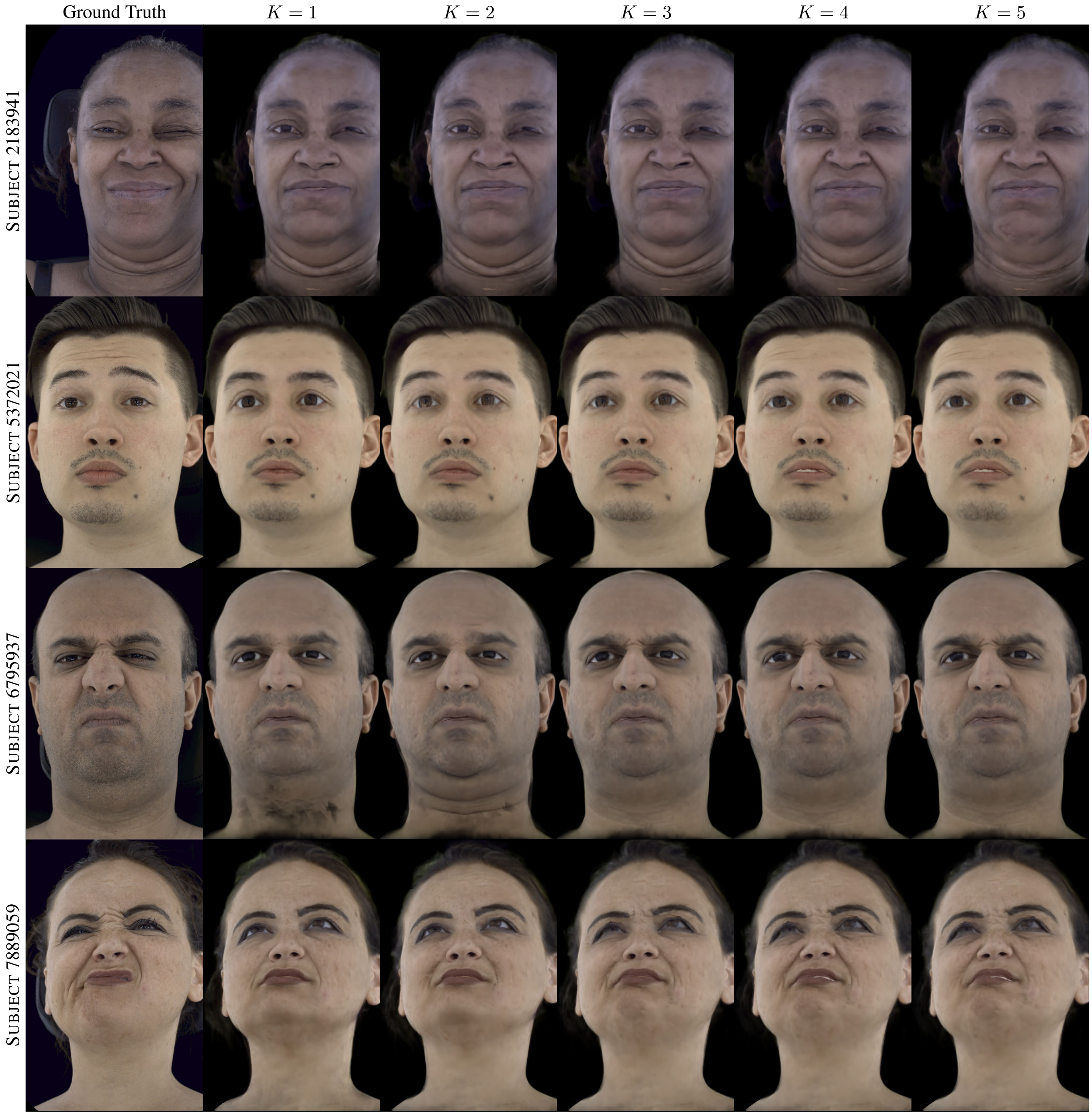

Figure 9. **Qualitative ablation over the number of training expressions** – We show qualitatively how the number of training expressions K affects the rendering quality. The first row shows the ground truth images. All other consecutive rows show the images rendered with BlendFields while increasing 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 [66]. Please refer to Tab. 4 for quantitative results.

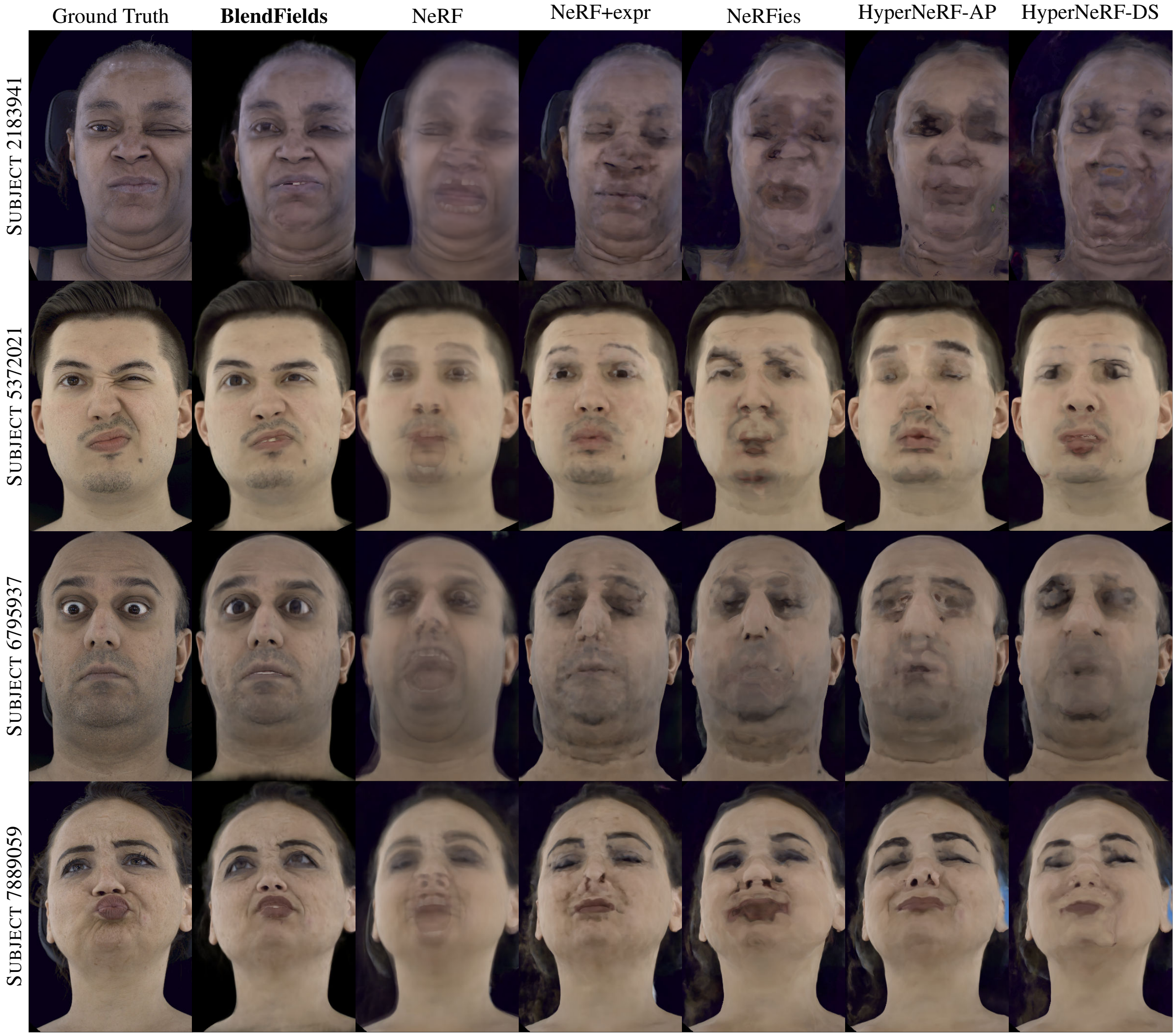

Figure 10. **Comparison to strictly data-driven approaches** – We compare BlendFields to other baselines that do not rely on mesh-driven rendering: NeRF [36], NeRF conditioned on the expression code (NeRF+expr) [36], NeRFies [42], and HyperNeRF-AP/DS [43]. As a static model, NeRF converges to an average face from available ($K{=}5$) expressions. All other baselines exhibit severe artifacts compared to BlendFields. 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.